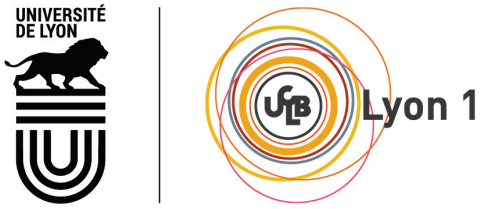
Models and algorithms applied to metabolism : from revealing the responses to perturbations towards the design of microbial consortia

Alice Julien-Laferriere



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 E2M2
ED-341

Spécialité de doctorat : BioInformatique

Soutenue publiquement le 08/12/2016, par :

Alice Julien-Laferrière

Models and algorithms applied to metabolism: From revealing the responses to perturbations towards the design of microbial consortia

---

Devant le jury composé de :

Brochier-Armanet Céline, Professeur des universités, UCBL	Examinatrice
Förster Jochen, Professor, Carlsberg Group	Rapporteur
Sauer Uwe, Professor, ETH Zurich	Rapporteur
Van Helden Jacques, Professeur des universités, AMU	Rapporteur
Sagot Marie-France, Directrice de Recherche, INRIA	Directrice de thèse
Vinga Susana, Principal Investigator, CSI/IDMEC	Co-directrice de thèse
Lacroix Vincent, Maître de Conférence Universitaire, UCBL	Co-encadrant

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique 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 des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. 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 GEP	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
Polytech Lyon	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

Après l'obtention de mon diplôme d'ingénieur, j'ai travaillé au LBBE, laboratoire qui aura été mon labo d'accueil pendant 5 ans, d'abord en tant qu'ingénieure, puis en tant que doctorante. Je suis ici aujourd'hui grâce à multitudes de personnes tant d'un point de vue personnel que professionnel et j'espère n'oublier personne. Si c'est le cas, je m'en excuse.

J'aimerais remercier les personnes qui m'ont fait confiance pendant ces 5 ans, et qui m'ont toujours traité avec respect et intérêt. Tout d'abord Stéphane Dray qui m'a proposé mon premier contrat au sein du Laboratoire de Biométrie et Biologie Évolutive et Vincent Lacroix qui m'a engagée comme ingénieure et qui a ensuite accepté de me co-encadrer sur ce sujet qui s'éloignait pourtant déjà ses problématiques actuelles de recherches. Dans ce début de vie professionnelle, j'aimerais aussi remercier Vincent Miele qui m'a introduit le HPC, m'a supervisé tout particulièrement lors de mon premier contrat, a accepté de faire partie de mon comité de pilotage et qui aura donc été un interlocuteur de choix pendant ces 5 ans.

Enfin merci Marie-France de m'avoir proposé cette thèse et d'avoir su me faire confiance sur ce sujet.

J'ai été encadrée lors de cette thèse par trois personnes exceptionnelles et complémentaires, que ce soit dans leurs domaines de compétences que dans leurs qualités humaines. Merci Marie-France pour cette chouette introduction au monde de la recherche et pour partager avec nous tes idées toujours plus "folles". Merci à Susana pour ta disponibilité, par Skype ou lors de mon séjour au Tecnico de Lisbonne et pour ton partage et tes cours sur l'optimisation, le Lasso et la FBA. Enfin, Vincent, discuter avec toi est toujours un plaisir, tes qualités humaines et ton esprit scientifique font que les discussions permettent toujours de dépoussiérer des idées, de les améliorer et d'éclaircir les futurs développements. Je tiens aussi à remercier Philippe Lejeune, Delphine Ropers, Vincent Miele et Ludovic Cottret pour avoir accepté de faire partie de mon comité de pilotage en ayant toujours mes intérêts à cœur lors de discussions fructueuses. Merci aussi à Celine Brochier-Armanet, Jochen Förster, Uwe Sauer et Jacques Van Helden pour leur participation à mon jury de thèse et pour leurs retours sur ce manuscript.

L'équipe Erable/Bamboo/Baobab a été composée (et l'est toujours) de personnes qui permettent à tous de travailler et d'évoluer dans des conditions chaleureuses. J'y ai passé 4 superbes années et tient pour cela à remercier ses membres permanents, temporaires, honoraires, présents et passés : Caio, Cecilia, Christian B., Christian G, Susan, Gustavo, Alex, Carole, Mariana, Sheila, Leandro, David, Laurent, Delphine, Marc, Taneli, Mattia, Florence, Marina, Camille, Hélène, Xavier, Clara, Louis, David, Catherine et Audric et plus particulièrement Martin, Laura, Arnaud, Ricardo, Mariana et Blerina qui se sont efforcés d'être présents particulièrement lors de la rédaction.

J'ai pu aussi au cours de cette thèse apprécier l'hospitalité portugaise lors de mon séjour au Tecnico en compagnie d'André, Andras, Mahdi, Ana.

Je remercie mes différents collaborateurs qui ont toujours permis de mener des débats et discussions respectueuses et productifs : Arnaud, Laurent, Ricardo, Andras, André, Louis, Delphine, Leen et Alberto.

Enfin ce travail de longue haleine n'aurait pas été possible sans le soutien du pôle administratif et informatique du LBBE : Bruno, Simon, Lionel, Stéphane, Philippe, Aurélie, Adil, Nathalie, Laeticia, Odile, Florence, Florence et Marina.

Lors de cette thèse, j'ai eu l'opportunité et le plaisir de donner des cours à l'université Lyon 1 et cela m'a été extrêmement facilité par l'équipe pédagogique du LBBE ainsi que l'implication et la motivation des enseignants des UEs.

Sur une note plus personnelle, merci aux amis pour des weeks-ends toujours détendus et parfois sportifs. Tout d'abord aux BIM et insaliens qui étaient déjà passés par là pour la plupart : Elsa, Elza, Hugo, Sacha, Bérénice, Audrey, Léa, Romain, Raphaëlle, Samuel, Camille, Mélaine, etc. Aussi les poules/poulets, parti(e)s pour la plupart vers d'autre horizons, ont permis de vivre de grands moments, sur le campus, en extérieur et parfois en luge : Léo, Aurélie, Marine, Marion, Cécile, Anne, Nicolas. Finalement les week-ends escalade entre araignées m'ont toujours été précieux.

Dans cette vie hors labo, je dois enfin te remercier Grégoire, pour tes encouragements, ta gentillesse, ton soutien indéfectible et tes blagues (souvent célèbres). Ta présence à mes cotés depuis quelques années maintenant m'aide chaque jour à me dépasser.

Enfin pour leur soutien, leur compréhension et leur intérêt pour mon travail j'aimerais remercier ma famille, plus particulièrement mes parents Hélène et Gabriel et mon frère Gaspard trop souvent absent mais toujours ancré dans mes pensées ainsi que celle que j'ai intégrée au fur et à mesures de années : Marie, Jacques-Henri, Odile, Christine, Gaspard, Adrien, François, Charlotte, Otto et Adélie.

Résumé en français

De nos jours, avec l'avènement des techniques dites "omiques" (métabolomique, transcriptomique, génomique, etc.), nous pouvons obtenir de plus en plus de données sur un système biologique (ici les micro-organismes) dans son ensemble. Il nous faut cependant ensuite arriver à les interpréter de manière raisonnable. Une manière d'y arriver est au travers de la modélisation des mécanismes biologiques qui nous intéressent.

Lors de cette thèse, je me suis intéressée à la modélisation du métabolisme des micro-organismes. Le métabolisme est l'ensemble des réactions chimiques qui permettent à un organisme de se développer au sein d'un environnement en consommant des substrats présents qui lui permettent ensuite de générer des composés qui lui sont utiles mais aussi de l'énergie. Dans ce travail, nous nous sommes focalisés sur le métabolisme des petites molécules, c'est-à-dire qui ne prend pas en compte des procédés biologiques tels que la synthèse des protéines ou la formation d'autres macromolécules.

Le métabolisme n'est pas déconnecté des autres processus des cellules, est n'est pas simple à appréhender. Cependant, je crois que des méthodes de modélisation efficaces peuvent économiser une quantité considérable de temps et élucider des processus qui ne sont pas forcément évidents à comprendre. En *biologie synthétique* par exemple, l'obtention d'une souche productrice stable n'est pas toujours simple et la modélisation peut aider les expérimentateurs à prédire les conséquences de certaines modifications génétiques, et / ou proposer de nouveaux plans expérimentaux (Carneiro et al., 2013). Par ailleurs, l'étude des sous-parties d'un système, par exemple dans le cas d'un réseau métabolique, d'une voie spécifique ou d'un sous-ensemble de réactions, peut être limitante. Aussi lors de ce travail, nous nous sommes efforcés de travailler avec les réseaux métaboliques dans leur globalité, c'est-à-dire avec les connaissances complètes des capacités métaboliques d'un organisme. Cette modélisation du métabolisme dans son ensemble peut nous permettre de proposer des solutions qui ne sont pas intuitives. Dans chaque approche de modélisation, que ce soit pour le métabolisme ou d'autres processus, il est nécessaire d'extraire de nos connaissances biologiques de simples hypothèses génériques qui décrivent l'objet que nous voulons modéliser. Les approches de modélisation ne prétendent pas être une réplique exacte et *in silico* de la cellule, mais plutôt une simplification éventuelle de celle-ci, sur la base d'hypothèses explicites. Un modèle peut ainsi proposer et découvrir de nouveaux mécanismes. En effet, si les mécanismes en jeu ont été correctement modélisés, il sera possible d'en déduire de nouvelles expériences et de valider les réponses données. Dans le cas contraire, si les solutions apportées par le cadre de modélisation ne sont pas cohérentes ou validables, alors les hypothèses de modélisation étaient (au moins partiellement) mal posées et doivent donc être nuancées ou réfutées. Ainsi, avec une démarche itérative, de nouvelles hypothèses peuvent être faites afin de concevoir un modèle plus adapté. Je crois donc que le processus de modélisation doit être basé

sur des hypothèses (*hypothesis-driven*) plutôt que sur les données disponibles (*data-driven*).

Dans ce travail de thèse, j'ai ainsi adopté une vision systémique, en essayant de décrire génériquement le métabolisme, avec des modèles applicables à différentes conditions ou à différents micro-organismes. Étant donné que ces modèles ne sont pas spécifiques, ils peuvent sembler offrir moins de pouvoir prédictif que des modèles plus complexes. Néanmoins, ces modèles génériques sont fondés sur des hypothèses simples et ma conviction est qu'ils sont utiles dans des conditions inconnues ou pour de nouveaux organismes.

Ce manuscrit est divisé en quatre parties qui contiennent une partie du travail réalisé pendant ces trois ans.

Tout d'abord, dans le Chapitre 1, je présente le métabolisme des petites molécules et les formalismes de modélisation généralement utilisés. J'évoque aussi les mécanismes de régulation du métabolisme, qui agissent sur la synthèse des protéines (transcriptomiques) ou sur leur activité (allostériques). Enfin, je présente des méthodes développées non seulement pour la compréhension des micro-organismes, mais aussi pour leurs modifications qui permettent actuellement de produire des molécules d'intérêt. Cet état de l'art se concentre tout d'abord sur les micro-organismes individuellement, puis sur les communautés bactériennes et leurs possibles modélisations.

Ensuite, nous avons utilisé différents formalismes de modélisations qui permettent de répondre aux questions que nous nous posées au cours de ce travail.

Tout d'abord, nous avons proposé des modèles afin de comprendre les réponses métaboliques d'organismes aux perturbations de leurs milieux extérieurs. Ainsi, en utilisant des hypergraphes dirigés et aussi de la modélisation sous contraintes, nous avons proposé d'identifier les réactions impactées par un changement de condition pour un organisme, à savoir comment son métabolisme réagit à une ou plusieurs perturbations. Il y a plusieurs motivations pour ce travail. En effet, bien que les micro-organismes soient souvent étudiés en isolation et dans des conditions stables (c'est-à-dire en laboratoire), ils évoluent naturellement dans un environnement changeant. Nous pouvons donc dire qu'ils sont soumis en continu à des variations dudit environnement et changent leur métabolisme en fonction. Or ces différents états métaboliques peuvent amener par exemple à une émergence de pathogénicité chez des organismes qui font partie de notre microbiome et qui sont habituellement commensaux. Savoir réguler les changements d'état stables chez un organisme peut ainsi permettre de proposer un retour à l'état asymptomatique. Comprendre comment les organismes s'adaptent à des perturbations permet également leur utilisation pour dépolluer des environnements. Par exemple, la levure peut emprisonner le cadmium, un métal, dans sa vacuole, faisant diminuer les concentrations de ce métal dans l'environnement.

Dans le Chapitre 2, nous présentons la méthode nommée TOTORO (*TOpological analysis of Transient metabOlic RespOnse*). L'idée ici est d'arriver à inférer les réactions impactées par un changement de conditions. Pour cela, nous disposons de mesures qualitatives des concentration des métabolites. Nous connaissons ainsi la variation possible de concentrations de certains métabolites entre deux états stationnaires: augmentation, décroissance ou identique. Mon équipe avait déjà proposé une méthode qui obtenait des résultats intéressants (Acuña et al., 2012a; Milreu et al., 2014). Cependant nous avons ici développé une nouvelle méthode qui améliore les résultats précédents. Tout d'abord, nous pouvons maintenant nous passer d'une étape de compression du réseau métabolique qui était lourde et pouvait perdre des solutions. De plus, une contrainte

était que les solutions recherchées devaient être acycliques dans le graphe des composés. Cette acyclicité n'était pas basée sur un *a priori* biologique et créait des problèmes d'énumération. Nous avons pu nous libérer de ces deux limitations (la compression et l'acyclicité) et proposer une toute nouvelle définition pour ce que nous appelons les *hyper-histoires métaboliques*. De plus, dans cette nouvelle modélisation, nous représentons les réseaux métaboliques avec des hypergraphes dirigés (la méthode précédente utilisait le graphe des composés). Cette représentation permet de retranscrire topologiquement le couplage des substrats et produits lors d'une réaction chimique. La nécessité d'avoir par exemple deux substrats pour produire un certain composé est présente, ce qui n'est pas le cas dans le graphe des composés qui était utilisé précédemment. Nous présentons tout d'abord la précédente méthode, et la démarche de modélisation suivie pour définir les *hyper-histoires métaboliques*. Nous proposons dans ce cas que la transition d'un état stationnaire à l'autre se fasse avec une réorganisation des flux minimaux, c'est-à-dire qu'un sous-ensemble restreint de réactions participe au changement d'états. Nous montrons comment, en utilisant la programmation par ensembles réponses (ASP: *Answer Set Programming*), nous pouvons énumérer les solutions recherchées. Ces solutions permettent d'expliquer les flux de matière qui ont du participer au changement d'état. En effet, l'idée est que la cellule étant de volume constant, l'augmentation de la concentration d'un métabolite se doit d'être liée à la diminution de la concentration d'un autre composé. TOTORO propose de trouver les réactions liant donc un métabolite dont la concentration a décru avec un métabolite dont la concentration a accru. De plus, nous inférons aussi la direction du changement des réactions. En effet, nous proposons si une réaction aurait du être activée, c'est-à-dire eu son flux augmenté, ou inhibée (diminution du flux). Nous ne pouvons pour l'heure, avec les données utilisées, proposer comment une telle régulation a eu lieu, mais nous pouvons pointer le sous-réseau impacté par un changement de condition. TOTORO a ensuite été appliqué à un jeu de données sur la levure en présence de cadmium (Madalinski et al., 2008). Nos solutions proposent une redirection des flux vers la synthèse d'un composé : la glutathione réduite. Or il est connu que la glutathione peut se lier au métal afin de le capturer et de le transférer vers la vacuole. Nous avons donc pu montrer que nous retrouvions les mécanismes connus de désintoxication mis en place par la levure. TOTORO permet de proposer de nouveaux mécanismes. Nous discutons aussi dans ce chapitre du fait que certains métabolites de cellules ne sont pas mesurés lors des expériences de métabolomiques. Nous proposons aussi avec cette nouvelle méthode de pointer sur certains métabolites non mesurés qui devraient voir leurs concentrations changées. Nous proposons ainsi de nouvelles expériences et mesures qui permettront d'expliciter les solutions ambiguës.

Ensuite dans le Chapitre 3, en utilisant une méthode de modélisation par contraintes, nous discutons d'un développement en cours, qui propose d'utiliser les mesures de concentration des métabolites entre différentes conditions pour inférer de manière quantitative les possibles asynchronies des réactions lors du passage d'un état stable à un autre. Pour cette méthode nommée KOTOURA (*Kantitative analysis Of Transient metabOlic and regUlatory Response And control*), nous connaissons exactement les variations de concentration entre deux états-stationnaires et proposons d'utiliser la matrice de stoichiométrie afin de quantifier les réactions qui ont été limitantes ou qui ont transféré un excès de matière entre ces deux états. Il s'agit donc, comme pour TOTORO, d'identifier le sous-réseau impacté, mais ici, grâce à des mesures quantitatives, nous pouvons inférer précisément à quel point les réactions ont eu un effet sur les variations de concentrations. Dans ce chapitre, nous présentons tout d'abord la formulation utilisée qui est basée

sur de l'optimisation linéaire en nombres entiers. Nous faisons l'hypothèse d'une ré-organisation parcimonieuse des flux et testons la méthode sur des données simulées. Nous nous proposons aussi d'énumérer les différents ensembles de réactions impactées. Tout d'abord, nous utilisons un petit exemple afin d'illustrer notre méthode et nos hypothèses puis nous utilisons un modèle cinétique publié par (Chassagnole et al., 2002) afin de pouvoir simuler l'état transitoire qui répond à une augmentation soudaine de glucose extracellulaire chez *Escherichia coli* et montrons que nous retrouvons bien les réactions ayant participé à cet état transitoire. Nous discutons aussi de la reproductibilité d'un modèle cinétique de la littérature.

Enfin, nous avons également développé un modèle permettant la sélection d'un consortium et de voies métaboliques pour la production de composés d'intérêts. En effet, les micro-organismes sont de plus en plus utilisés par l'industrie. Dans l'agroalimentaire, des procédés de fermentation permettent depuis longtemps de produire du pain, de la bière, des yaourts et fromages, etc. De plus, il est maintenant possible de modifier génétiquement des micro-organismes afin de produire des molécules qu'ils ne possèdent pas naturellement. Par exemple, en pharmacologie la production de nouvelles molécules est possible grâce à l'ADN recombinant. Les organismes sont souvent utilisés en culture pure, mais de plus en plus de recherches s'intéressent à l'usage de communautés bactériennes.

Dans le Chapitre 4, nous proposons MultiPus (*MULTIple species for the synthetic Production of Useful biochemical Substances*), une méthode qui permet d'inférer les voies métaboliques d'intérêt pour la production de composés chimiques au sein d'une communauté de micro-organismes.

MultiPus permet aussi de sélectionner les espèces faisant partie du consortium considéré. Les réseaux métaboliques sont représentés sous la forme d'hypergraphes dirigés et pondérés. Les réseaux métaboliques des différentes souches ou espèces pouvant prendre part au consortium sont considérés conjointement, et nous proposons de permettre un possible transport impliquant tous les composés possédés par au moins deux membres du consortium. Enfin, des réactions non présentes nativement mais annotées dans la nature (c'est-à-dire dans d'autres organismes) peuvent être insérées dans le réseau puisque l'on considère que les gènes codants pour les enzymes responsables de la catalyse peuvent être exprimés dans les organismes producteurs par ingénierie génétique. Les pondérations des hyperarcs représentent trois catégories de réactions : les réactions endogènes, les transports, et les réactions insérées (exogènes). Leur poids correspondent alors à la difficulté d'expression des enzymes ou de transport. Nous cherchons ensuite les hyperchemins de poids minimum tels que depuis un (ou plusieurs) substrat(s) prédéfini(s), il soit possible topologiquement de produire un (ou plusieurs) composé(s).

Nous proposons de résoudre ce problème d'énumération d'hyper-arbres de Steiner dirigés soit avec un algorithme de programmation dynamique paramétré, soit par une approche de programmation par ensemble réponse (ASP). Ces deux propositions sont ensuite utilisées dans deux cas d'applications. Nous considérons tout d'abord une communauté synthétique pour la production de deux antibiotiques (la pénicilline et la céphalosporine C) à partir d'une seule source de carbone, la cellulose qui est un substrat peu coûteux et facilement disponible. Nous avons sélectionné comme espèces pouvant faire partie de la communauté trois actinobacteries et une archée méthanogène. Les réactions insérées venaient de deux organismes, un champignon et une actinobactérie. La solution obtenue montre que le meilleur consortium microbien pour la production des deux bêta-lactamines est constitué de deux micro-organismes uniquement : *Streptomyces cattleya* et *Methanosarcina barkeri*. MultiPus nous permet donc d'obtenir les voies métaboliques

nécessaires à la synthèse des produits mais aussi de sélectionner le meilleur consortium possible dans un plus large ensemble d'espèces. Nous avons ensuite testé un consortium artificiel constitué de *Klebsiella pneumoniae*, qui produit naturellement du 1,3-propanediol (PDO) et une Archaea méthanogène *Methanosarcina mazei*. Le 1,3-propanediol est un polymère d'intérêt pour l'industrie, il est utilisé pour la fabrication de nombreux produits (peintures, composites, etc.). Lors de la production de PDO, de l'acétate est aussi synthétisé. Or l'acétate a une action inhibitrice sur la production de PDO et la croissance de *K. pneumoniae* car il est toxique lorsque présent en grandes concentrations. Cependant *M. mazei* peut pousser sur de l'acétate et produire du méthane qui peut être récupéré pour créer du biogaz. La source de carbone proposée était du glycérol, un sous-produit de la production de biodiesel et donc un substrat de choix pour des procédés biotechnologiques. Dans un premier temps, les poids des réactions (endogène, exogène et transport) ont été définis uniformément dans chacune des catégories. Puis nous avons proposé de diminuer le poids du transport de l'acétate en supposant qu'il existe un processus d'excrétion puisque celui-ci est toxique. Dans ce cas, nous obtenons un ensemble de réactions métaboliques permettant effectivement de produire à la fois du PDO et du méthane en partant uniquement du glycérol grâce à la consommation de l'acétate par *M. mazei*. Nous avons donc pu montrer que MultiPus retrouvait bien les voies métaboliques connues pour la production de PDO associé avec un organisme méthanogène.

Au cours de ces trois ans, nous avons ainsi proposé, implémenté et testé trois modèles qui permettent de mieux appréhender le métabolisme de micro-organismes. Pour cela, nous avons utilisé différents formalismes et avons travaillé en collaboration avec des informaticiens et biologistes afin de pouvoir poser des questions cohérentes biologiquement et de développer des méthodes calculables sur des réseaux métaboliques complets.

Les méthodes développées sont disponibles en ligne aux adresses :

- `http://hyperstories.gforge.inria.fr/` pour Totoro et Kotoura,
- `http://multipus.gforge.inria.fr/` pour MultiPus.

MultiPus a de plus fait l'objet d'une publication (Julien-Laferrière et al., 2016) et deux manuscrits sont en préparation pour les deux autres méthodes (Totoro et Kotoura).

Contents

Introduction

Metabolism can be defined as the ensemble of chemical reactions inside an organism. It allows such organisms (as a single individual, or as communities) to live in various environments, importing from them resources, in the form of molecules, that will be transformed into energy and biomass (for growth and maintenance). It is an essential process of the cells, that furthermore enables to rapidly react to a changing environment, including possibly act on it to allow for a better survival. As an example, Yeast (*Saccharomyces cerevisiae*) can cope with a high level of metals (Wysocki and Tamás, 2010) and can sequestrate cadmium in one of its organelles, namely the vacuole (Volesky et al., 1993). The bio-accumulation of cadmium, a toxic metal, which is captured inside the organisms thus decreases the extracellular concentration.

The initial studies of metabolism focused on an understanding of small subparts of it, such as, for instance, related to obtaining a dynamic characterisation of glycolysis which is the ensemble of reactions degrading glucose into pyruvate while releasing energy that is important for the organism. With the advent of new techniques, such as genomics, transcriptomics and metabolomics, it then became possible to obtain more data, and eventually, to infer *in vivo* measurements instead of *in vitro* ones. For example, metabolic network reconstructions and fluxomics can provide kinetic information on multiple enzymes of a pathway whereas before the characterization of an enzyme was an extensive work.

Moreover, genome-scale metabolic networks (GEMs) are now being reconstructed using the functional annotations of an organism together with manual curation that remains extensive. Such reconstructions can be considered as a representation of the global capacities of the metabolism of an organism. Together, the new *in vivo measurements* and the availability of accurate metabolic reconstructions are thus providing a more systematic and global view of metabolism.

Metabolism is at the core of this PhD. The motivations for studying it can be multiple. Besides natural curiosity, understanding how organisms, including ours, function may lead to a better knowledge of the current world and of its organisation since metabolism participates in the transformation of the chemical molecules that surround us. For this, I decided to focus on microorganisms. Indeed, it appears as a necessary step for the development of new methods, as less processes have to be taken into account (hopefully) and we have already a substantial knowledge of some model organisms (such as *Escherichia coli* and *Saccharomyces cerevisiae*). More data are thus available to formulate hypotheses and to test them.

One clear interest is to discover novel treatments against pathogens. Indeed, microorganisms are present everywhere, including in our body. One can model the pathogenic action of a bacterium, for example to identify possible drug targets (Mazumdar et al., 2009). Moreover, some diseases can result from an organism that is usually commensal (Henriques-Normark and Normark, 2010). For example, *Candida albicans* is a commensal fungus of the human gut flora.

However, it is also involved in candidiasis, in particular in immunodepressive individuals. In response to an external signal (probably of an interplay between the organisms and its host), the fungus becomes pathogenic (Hube, 2004). Understanding why and how a commensal organism can thus turn pathogenic is important. Indeed, knowing what triggered pathogenicity could indicate how to cope with symptoms, and eventually how to return to an asymptotic state.

Another reason for studying metabolism is to discover novel compounds produced by organisms growing in extremely stressful conditions. Microorganisms are disseminated all around the globe, including in environments that appear inhabitable. One can find them at high and low temperatures, pressures, salinity and even radiation levels. Already just apprehending the mechanisms in play is challenging, but understanding evolutionary pressures and defences to such hostile environments can lead to the discovery of new industrial processes and products (Shivlata and Satyanarayana, 2015). Moreover, once it is understood how bacteria for example, can cope with some toxic compounds, it is possible to use them directly for decontamination. Microorganisms are thus used for the decontamination of polluted water or soil as they can capture heavy metal and other compounds. (Shetty and Jespersen, 2006) also reviewed the possibility to employ lactic acid bacteria and Yeast to decontaminate food products that can present high levels of mycotoxins which cause intoxication and may lead to the death of animals (including humans).

Microorganisms have been part of our industrial processes since a long time. Indeed, fermentation processes are now well understood and the food industry produces beer, wine, yoghurt or bread using yeast and other organisms. Improving the taste, yield and other controlled variables with better fermentors, culture media etc. is thus a hot topic in a highly competitive environment. Such improvements require obtaining novel strains that are in some way "better".

Furthermore, with the increasing development of *synthetic biology*, microorganisms are now being used to produce compounds that they do not possess naturally (Rollié et al., 2012). A major breakthrough was insulin, that is nowadays produced by Yeast or *Escherichia coli*. Increasingly more biopharmaceuticals can thus be extracted from different organisms after the discovery of recombinant DNA (Huang et al., 2012), including new molecules (usually proteins) that are difficult to synthetise chemically. Non pharmaceutical molecules that are also of interest for human health or other industries (such as involved in the development of colours) can be extracted from organisms that produce molecules naturally, or after gene-insertions and other modifications. Usually the objective is to obtain simpler, more productive and controlled processes. My PhD was funded by the project **BacHBerry**, BACterial Hosts for production of Bioactive phenolics from bERRY fruits (FP7-613793), whose main objective was to produce certain polyphenols from plants (more particularly several species and varieties of berries) using mainly two organisms, *Lactococcus lactis* and *Corynebacterium glutamicum*. Another common application is the production of the components used for polymers (Nakamura and Whited, 2003). Lately, the production of biofuels has also gained more audience despite the difficulties in scaling up the processes (Hollinshead et al., 2014).

Finally, it is now possible to biologically reproduce metabolic processes happening in human inside microorganisms, such as the baker yeast, using genetic technologies (Franssens et al., 2013). Such biological models (*i.e.* the microorganism) thus reproduce some of the symptoms observed. Through measurements of the metabolome and of the fluxome, one can study the disease in a control condition. This is particularly important for diseases such as neurodegenerative ones (*e.g.* Parkinson disease) where a biopsy would be destructive. Apprehending how the metabolic roads

are impacted therefore provides insights on the development of the diseases and could pave the way for possible biomarkers obtainable in a non invasive way (in the blood or urine for example).

The study of the metabolism of single organisms has already proven useful. However, as mentioned, microorganisms are ubiquitous and do not live in isolation but rather in interaction with their environment including other organisms. In the past few years, increasingly more work has been done towards the understanding of natural communities, their composition and the interactions between the different actors composing them. Those studies have the same motivations as for single organisms but raise even bigger challenges. In particular, in the case of animal diseases (including humans), pathogenesis may be due to the appearance of a desequilibrium in commensal communities. Increasingly more focus is thus put on the microbiome and its impact on health (Donaldson and Mabbott, 2016; Wu and Ross, 2016).

Deciphering metabolism in general is hence of great interest. I believe that understanding better microorganisms in specific conditions could thus lead to more biological control, and to novel industrial productions.

Of course, metabolism is not disconnected from the other processes of the cells, nor is it simple to apprehend. Furthermore, in the case of recombinant strains, the exact repercussion of modifications at the genomic level is not always known at the metabolic or transcriptomic level.

However, I believe that efficient modelling methods can save a considerable amount of time and unravel processes that may not be obvious. In *synthetic biology* for example, obtaining a stable, high producing strain remains not straightforward and computational methods can help experimentalists to predict the consequences of certain genetic modifications, or/and propose new experimental plans (Carneiro et al., 2013). Moreover, the study of the individual processes, such a specific pathway or a subset of reactions, can be limiting and thus in our case we try to place ourselves at the level of the metabolism.

Modelling metabolism as a whole can propose solutions that were not intuitive. In every modelling approach, whether this is for metabolism or for some other processes, there is the need to extract from our generic biological knowledge simple hypotheses that will describe the object that we want to model. Modelling approaches do not pretend to be an exact *in silico* replicate of the cell, but rather a possible simplification of it, based on explicit hypotheses. A model can thus propose and discover novel mechanisms. Indeed, if the mechanisms in play have been correctly modelled, it will be possible to infer new experiments and validate the given responses. Otherwise, if the solutions provided by the modelling framework are not possible, then the initial modelling process was (at least partially) wrong and the initial hypotheses should be nuanced or refuted. Moreover, in a iterative process, new hypotheses can be made and a novel model designed and tested. I therefore believe that modelling should be more *hypothesis-driven* than *data-driven* to start.

In this PhD work, I adopted a systemic view of modelling. I tried to generically describe possible organism behaviours, with models that are applicable to different conditions or to different microorganisms. Since those models are not specific, they may appear to offer less predictive power than more complex and better fitted ones. Nevertheless, those generic models are based on simple hypotheses and my belief is that they will prove useful in unknown conditions or for new organisms.

Modelling metabolism can be done using different types of mathematical objects, formalisms and techniques as will be discussed in Chapter 1. It thus requires knowledge in mathematics/computer science, software engineering/programming and biology. The work performed here

has been at the intersection of those areas with the support of many collaborators of such disciplines. Adopting the right formalism and right definition to address a problem, obtaining a rigorous and tractable approach to compute the results and propose a biological solution that if satisfiable provides new possibilities, was the constant preoccupation of all this work.

In this manuscript, I will present the methods developed during these three years. Each uses a different formalism.

First, I focused on inferring the behaviour of an organism in terms of its metabolism when the organism is subjected to a change in conditions. In this case, one can infer the reactions impacted when the changes are controlled and known (*e.g.* exposition to toxic compounds, changes in culture conditions). However, understanding how the metabolism of an organism changes of equilibrium is also of interest to infer the processes related for example to a transition between a commensal or beneficial bacterium to a pathogenic one as aforementioned.

This question led to two different methods that will be presented sequentially. First, in Chapter 2, we will discuss Totoro (TOpological analysis of Transient metabOlic RespOnse), a method based on the topology of metabolic networks to infer the reactions involved in a transient state, when an organism goes from one state of growth to another. We proposed a novel definition using the directed hypergraph representation and discuss its application on a dataset of Yeast exposed to cadmium. We show that this method proposes more complete solutions of the reactions impacted during the metabolic shift.

In Chapter 3, such problem is then treated from a constraint-based perspective in a more quantitative approach. For now, we applied this method, called Kotoura (Kantitative analysis Of Transient metabOlic and regUlatory Response And control), to a simulated dataset and we are currently trying to infer the possible quantitative responses to mutations with a more complete kinetic model. An image used by (Klein et al., 2012) is that condition-specific models provide a snapshot of the metabolism of an organism, whether it is at the evolutionnary-time scale or at the scale of a specific environment condition describing a physiological process. Our idea here is thus to infer the transitions between those snapshots.

A paper is in preparation for each of the works presented in Chapters 2 and 3.

Finally, in Chapter 4, I will discuss the search for microorganism consortia adapted to the production of compounds of interest from a topological perspective. The associated method, MultiPus (MULTIple species for the synthetic Production of Useful biochemical Substances), uses a weighted directed hypergraph modelling of the metabolic network. We developed two implementations of this method, one using an algorithm inspired by the Directed Steiner Tree problem, and the other using disjunctive logic programming. In both cases, we proposed a possible consortium for the production of target compounds. First, we proposed to select the members of such consortium to produce two antibiotics from a low cost substrate. The solutions of our method also include the metabolic roads to produce the antibiotics and the necessary heterologous reactions to use. Second, we showed that is was possible to produce 1,3-propanediol (used in polymers synthesis) and methane using an enterobacteria (*Clostridium butyricum*) and a methagenic archae. Here, we inferred the possible metabolic exchanges of the community. This method has been published (Julien-Laferrière et al., 2016).

Besides the above, I also participated during this PhD and in the context of the FP7-BacHBerry project that funded it, in two other works with manuscripts in preparation.

The first one is the use of a multiobjective mixed integer optimisation. In particular, this framework, called OptMulti, proposes genetic modifications (gene knockouts) while optimising

simultaneously several objectives, in particular several cellular functions. The second, called MassBlast, is an automated workflow that combines different tools to align multiple transcriptomes (or genomes) onto target gene-sequences and aims to accelerate multi-transcriptome analyses.

We also developed a dynamic viewer of hypergraphs called Dinghy (Dynamic Interactive Navigator for General Hypergraphs in biologY) available at http://dinghy.gforge.inria.fr/. Dinghy was presented during the French Conference in computational biology and bioinformatics, JOBIM (for Journées Ouvertes en Biologie, Informatique & Mathématiques) (Bulteau et al., 2015).

The methods discussed in the manuscript are available at:

- http://hyperstories.gforge.inria.fr/ for Totoro and Kotoura (website in construction),
- http://multipus.gforge.inria.fr/ for MultiPus.

Chapter 1

Concepts: Modelling metabolism

Contents

1.1 Introduction

The work presented in this thesis belongs to the area of *Systems Biology* and *Synthetic Biology* that arose because of the increasing number and type of data available (genome, metabolome, proteome, transcriptome). This amount of data on various scales allowed to start describing systems in a global manner, *i.e.* by understanding how organisms work and respond to environmental or genetic variations.

Studies in Systems Biology can be done at different levels using the known building blocks of biology: DNA, RNA, proteins. Such building blocks are used as entities for the system modelling and are represented in the form of networks, that can, through their analysis, unravel part of the behaviour of an organism.

In the literature, the networks representing the interaction between the building blocks have been divided into three classes based on their biological function (Machado et al., 2011; Klamt et al., 2014):

- *Cell signalling networks*: the components of such a system can be proteins, metabolites, lipids, etc. that can interact among them, carrying control information, either an inhibition or an activation (for example of the transcription machinery). These can have an impact on the transcription and/or translation and are the result of an external signal.

- *Regulatory networks*: the components of such a system are genes, RNA and proteins. The interactions represent the regulations that one component exercises on another (inhibition/activation).

- *Metabolic networks*: the main components are metabolites and the interactions (arcs) represent the chemical transformations of one set of components into another. Those reactions are cat alysed by enzymes, which are proteins.

In this work, I focused on metabolic networks, although the different levels presented above are intertwined biologically. More in particular, I focused on the metabolism of unicellular organisms.

The purpose of this chapter is to introduce the notions in graph theory, in metabolism and on how metabolism has been modelled that are at the basis of this thesis.

I will furthermore provide some background on metabolism regulation and on current applications in synthetic biology, and also on the modelling and design of microbial communities.

1.2 Preliminaries

1.2.1 Graph theory

Here, I will briefly introduce directed graphs and directed hypergraphs as well as the associated notations that I will adopt throughout the thesis. Indeed, such mathematical objects are used to represent biological networks and their analysis has allowed to discover interesting properties.

A **directed graph** (also called *digraph*) is composed of two disjoint sets:

- the vertices (also called nodes);

- the arcs representing the interaction between the vertices.

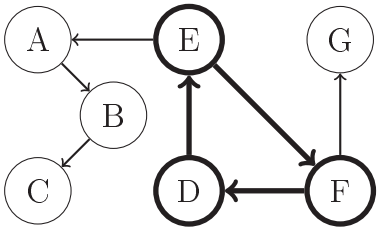
An arc is represented as an ordered pair of vertices (u, v), indicating a relation $u \to v$ that is not necessarily symmetrical.

We can denote a **directed graph** $\mathcal{G}$ by a pair of sets $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ where $\mathcal{V}$ are the vertices and $\mathcal{A}$ are the arcs.

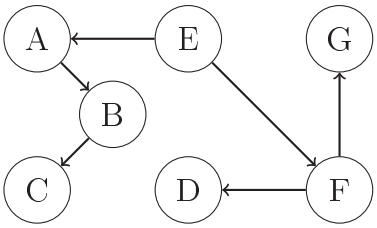
The **in-degree** of a vertex is the number of arcs entering the vertex, and is denoted by d^-. The **out-degree** is the number of arcs exiting such vertex and is denoted by d^+. A **source** is a vertex with an in-degree null ($d^- = 0$) while a **sink** is a vertex with an out-degree null ($d^+ = 0$).

In a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, it is possible to define a **directed path** between two vertices u and v such that there is a set of arcs linking them. More formally, a directed path of size n between two vertices u and v (denoted by $u \rightsquigarrow v$) is defined as a succession of vertices $\{p_0, p_1, ...p_n\}$ such that $p_0 = u$, $p_n = v$, and for all $k = 0, 1, ...n-1$, $(p_k, p_{k+1}) \in \mathcal{A}$. A path is elementary if it does not contain the same vertex twice. Finally, a **cycle** in a digraph is a path starting and ending at the same vertex ($u \rightsquigarrow u$).

In Figure 1.1, there are several paths. For example, it is possible to go from A to C passing through B. In Figure 1.1a, there is also a cycle F $\rightsquigarrow$ F passing through D and E.



(a) A digraph containing a cycle {D, E, F} (in bold).



(b) A directed acyclic graph.

Figure 1.1: Examples of directed graphs.

One can define also directed acyclic graphs (**DAG**), which are directed graphs with no cycle (for example in Figure 1.1b).

Finally, it is possible to add to the graph representation a mapping function either on the vertices or on arcs. It is usually assimilated to a cost/weight function, for example on the arcs: $w : \mathcal{A} \mapsto \mathbb{R}$.

Digraphs have been studied extensively in graph theory and algorithms have been developed to solve various problems in applied areas such as biology.

In this thesis, I worked **with directed hypergraphs**. A directed hypergraph is also defined through two sets: the set of the vertices and the set of the hyperarcs.

In the case of a directed hypergraph, a hyperarc a represents the relation between two sets of vertices X and Y that are called the **head** ($Y = head(a)$) and the **tail** ($X = tail(a)$) as defined by (Gallo et al., 1993; Ausiello et al., 2001). For example, in Figure 1.2, X =(A,B) represents the tail vertices of the hyperarc H_1 and Y = (C,D) its head vertices.

A *sub-hypergraph* of a directed hypergraph is defined as follows:

Definition 1. *A* sub-hypergraph $\mathcal{H}'$ *is a hypergraph* $\mathcal{H}' = (\mathcal{V}', \mathcal{A}')$ *with* $\mathcal{V}' \subseteq \mathcal{V}$ *and* $\mathcal{A}' \subseteq \mathcal{A}$.

As previously, it is possible to define paths and cycles in directed hypergraphs. However, here it is necessary to choose in the definition whether all the tail vertices of a hyperarc have to be

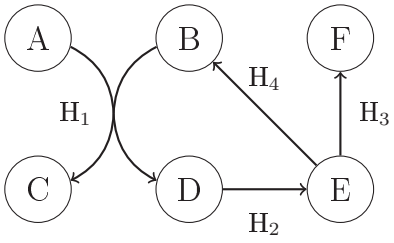


Figure 1.2: Example of directed hypergraph, where $\mathcal{A} = \{H_1, H_2, H_3, H_4\}$ and $\mathcal{V} = \{\text{A, B, C, D, E, F}\}$.

reached by the path in order to use such hyperarc. There are at least two existing definitions of a hyperpath in the literature. (Gallo et al., 1993) defined a hyperpath $\mathcal{P}_{st}$ of a directed hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{A})$ as in Definition 2, with a succession of vertices and hyperarcs.

Definition 2. *A hyperpath $\mathcal{P}_{st}$, of length q, in the hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{A})$ is a sequence of vertices and hyperarcs $\mathcal{P}_{st} = (v_1 = s, a_{i_1}, v_2, a_{i_2}, ..., a_{i_q}, v_{q+1} = t)$, where $s \in tail(a_{i_1}), t \in head(a_{i_q})$, and $v_j \in head(a_{i_{j-1}}) \cap tail(a_{i_j})$, $j = 2, ..., q$.*

In Definition 2, it is not necessary to reach all the vertices of the tail of a hyperarc to include such hyperarc in the path.

(Ausiello et al., 2001) proposed a more stringent definition of a hyperpath that is the generalisation of the simple path in a directed graph (see Definition 3).

Definition 3. *A hyperpath in $\mathcal{H}$ from a set of vertices $S \subset \mathcal{V}$,with $S \neq \emptyset$, called source, to a target vertex $t \in \mathcal{V}$ is a sub-hypergraph $\Pi_{S,t} = (\mathcal{V}_{\Pi_{S,t}}, \mathcal{A}_{\Pi_{S,t}})$ of $\mathcal{H}$ having the following property: if $t \in S$, then $\mathcal{A}_{\Pi_{S,t}} = \emptyset$, otherwise its $k \geqslant 1$ hyperarcs can be ordered in a sequence $(a_1, ..., a_k)$ such that:*

1. *$\forall a_i \in \mathcal{A}_{\Pi_{S,t}}$, $tail(a_i) \subseteq S \cup \{head(a_1), ..., head(a_{i-1})\}$;*
2. *$t \in head(a_k)$;*
3. *No proper sub-hypergraph of $\Pi_{S,t}$ is a hyperpath from S to t in $\mathcal{H}$.*

For example, in Figure 1.2 according to Definition 2, there is a hyperpath $\mathcal{P}_{\text{A,F}}$ in the hypergraph. We have that: $\mathcal{P}_{\text{A,F}} = (\text{A}, H_1, \text{D}, H_2, \text{E}, H_3, \text{F})$. This path is shown Figure 1.3a.

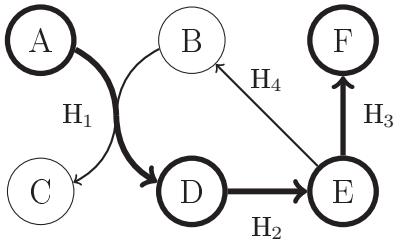
There are no hyperpaths $\Pi_{S=\text{A},t=\text{F}}$ using Definition 3. However, there exists one starting from the set of sources composed of A and B ($\Pi_{S=\{\text{A,B}\},t=\text{F}}$). It is presented in Figure 1.3b.

A hyperpath is a cycle if $s = t$ or $t \in S$.

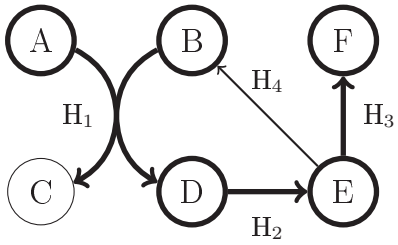
These notions will be used later on, more particularly in Chapter 2 and Chapter 4. They are required to topologically model the metabolism. However, there are many different ways of modelling metabolism as will be presented in Section 1.3. First, I introduce the metabolism itself in the next section.

1.2.2 Metabolism

A cell needs to create energy that will be consumed afterwards for maintenance and growth. Hence, metabolism can be divided into:



(a) The hyperpath $\mathcal{P}_{\mathrm{A,F}}$ is shown in bold, it is possible to reach F only with the starting vertex A using Definition 2.



(b) The hyperpath $\Pi_{S=\{\mathrm{A,B}\},t=\mathrm{F}}$ is shown in bold. Both A and B are required for H_1 to be part of the path in the Definition 3.

Figure 1.3: Example of hyperpaths in a hypergraph using the two definitions proposed.

- *catabolism* that uses the substrates available in the environment to form smaller molecules and create energy (exergonic processes);
- *anabolism* that forms bigger molecules to sustain growth (*e.g.* fatty-acids, nucleic acids); such reactions require energy (endergonic processes).

Metabolism is the set of chemical reactions happening in a living cell and can be described as an ensemble of enzymatic conversions.

Chemical reactions are catalysed by enzymes, which are usually proteins. Some reactions may be catalysed by different enzymes called isoenzymes that can have different substrate affinities and kinetic parameters. Finally, a reaction can be catalysed by an enzymatic complex that is composed of several proteins bonded together.

Reactions can be grouped into *metabolic pathways*. Whereas the notion of pathway is commonly used, there is no consensus on a strict definition of what exactly is a metabolic pathway, but one can say that a pathway is a set of reactions that are involved in a common task. For example, glycolysis is the metabolic pathway that converts glucose into pyruvate. However, with the evolution of the metabolic networks that are now genome-scale and thus increasingly more branching and complex, the discovery of novel pathways and their precise boundaries is subject to controversies as discussed by (Faust et al., 2011).

In this work, we focused on the small molecule metabolism, that is the set of reactions that involve only small molecules; reactions involving macromolecules are discarded. Macromolecules are polymers and are built of small molecules (monomers). Usually carbohydrates such as starch or branched glycogen, but also proteins and nucleic acids are considered as macromolecules. When working with the small molecule metabolism, part of the limitations in the predictions made will thus be related to the protein and DNA synthesis or the polysaccharides storage (Campbell et al., 2015). Another limitation is that there is no precise reaction to predict how the organisms maintain themselves and eventually duplicate. Usually, a biomass reaction that is a proxy for maintenance and growth is then added to the models (Feist and Palsson, 2010). More particularly, in quantitative models such as those described in Section 1.3.4, this biomass reaction is used to compute the growth rate in a specific condition and helps to constrain the solution space.

It is also possible to make a distinction between metabolites and what are commonly called cofactors (or currency metabolites (Schilling et al., 2000)) involved in redox balance and/or energy production/consumption (such as ATP).

I now define the notations that will be used throughout this thesis.

Metabolism can be summarised as a set of reactions $\mathcal{R}$ and a set of metabolites $\mathcal{C}$. A reaction is the conversion of the substrates $subs \subseteq \mathcal{C}$ into the products $prods \subseteq \mathcal{C}$.

I will illustrate the next notions using the simple reaction presented in Equation (1.1).

In this reaction, the substrates are 2-acetolactate and CO_2 (carbon dioxide) that are transformed into one product that is pyruvate. We can notice that we have here an additional information, namely the *stoichiometry*. Indeed, one molecule of 2-acetolactate and one molecule of carbon dioxyde will create **two** molecules of pyruvate. This is based on the fact that the mass of the substrates consumed should be equal to the mass of the molecules produced. Such masses are usually represented by integers. In Equation (1.1), the stoichiometric number 1 was indicated but this is often implied (that is, when the stoichiometry is one, nothing is written).

$$R1: \ 1\ 2\text{-}Acetolactate + 1\ CO_2 \Longleftrightarrow 2\ Pyruvate \tag{1.1}$$

One way of representing such balanced association is through the so-called *stoichiometric matrix* $\mathcal{S}$ of dimension $|\mathcal{C}|\text{x}|\mathcal{R}|$. In $\mathcal{S}$, the lines correspond to the metabolites whereas the columns are the reactions. The stoichiometric coefficient s_{ij} of the metabolite i in the reaction j is negative if i is consumed by j or positive if i is produced. Furthermore s_{ij} is null (0) if i is not involved in j. Since most metabolites are involved in a small number of reactions, $\mathcal{S}$ is sparse. An example of stoichiometric matrix $\mathcal{S}$ for the reactions in (1.1) and in (1.2) is given in Equation (1.3), where $\mathcal{S}$ denotes the stoichiometric matrix for the reactions $R1$ and $R2$.

$$R2: \ Oxaloacetate \Longleftrightarrow Pyruvate + CO_2 \tag{1.2}$$

$$\mathcal{S} = \begin{array}{r} \\ \text{2-Acetolactate} \\ \text{CO}_2 \\ \text{Pyruvate} \\ \text{Oxaloacetate} \end{array} \begin{array}{c} \begin{array}{cc} R1 & R2 \end{array} \\ \begin{pmatrix} -1 & 0 \\ -1 & 1 \\ 2 & 1 \\ 0 & -1 \end{pmatrix} \end{array} \tag{1.3}$$

The reaction (1.1) and the reaction (1.2) are reversible (as indicated by the sign: $\Longleftrightarrow$). We presented them from left to right but here, according to the notations, we could also say that for $R1$, two molecules of pyruvate can be split into one molecule of 2-Acetolactate and one molecule of carbon dioxide.

In theory, any reaction could be reversible. However, the energy required for the reaction to take place in one of the directions may be high and hence will not be favoured. Some reactions might be spontaneous, that is *exorgenic*, whereas others will require an input of energy (consuming energy); in this latter case, they are called *endergonic*. Such concept is translated into the one of *Gibbs energy* denoted by ΔG, where $\Delta G > 0$ indicates an endergonic reaction, whereas a reaction with $\Delta G < 0$ indicates an exorgenic one. The higher the Gibbs energy, the more consuming is the reaction in terms of energy. Enzymes are catalysts, hence they allow reactions to happen more easily (and faster) by lowering the energy barrier needed for the reaction.

Metabolism can be modelled in different ways as presented in the next section.

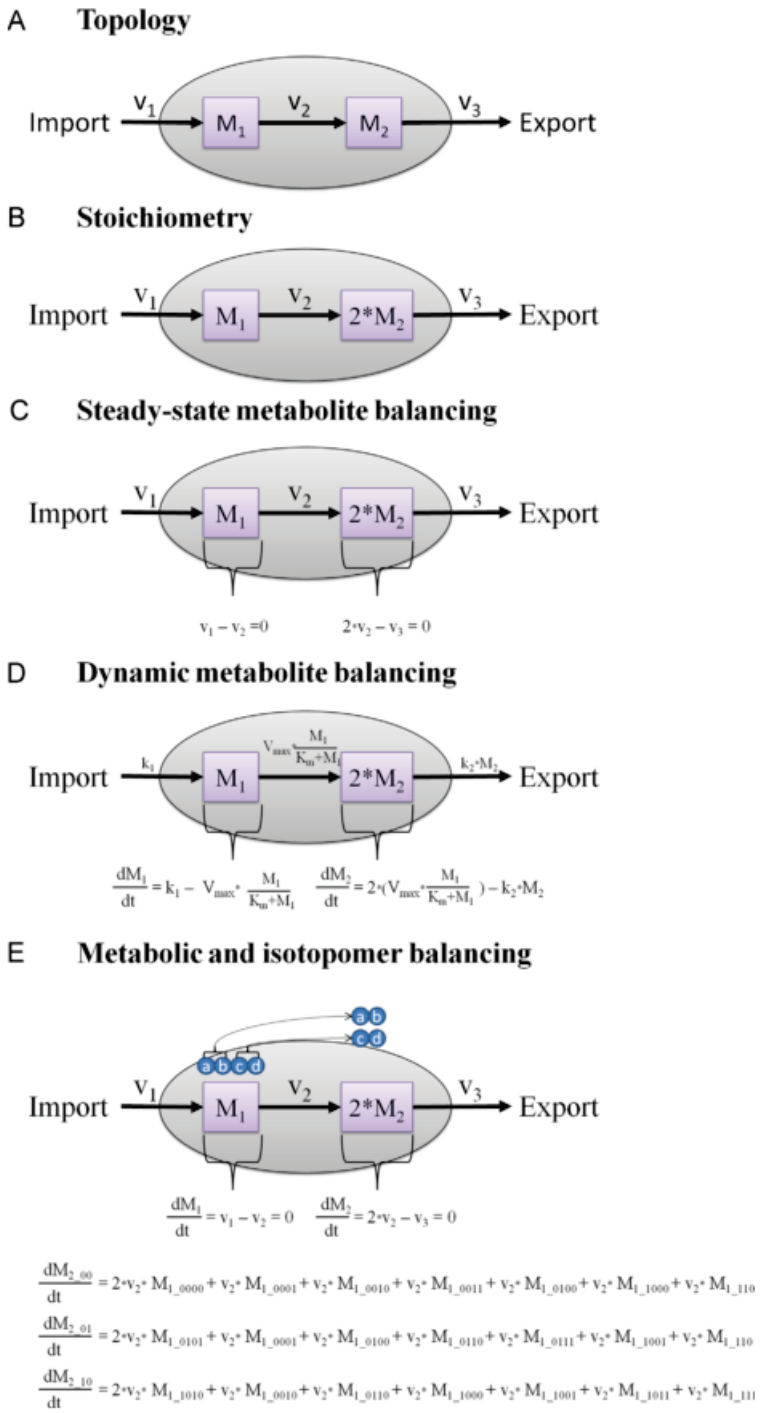


Figure 1.4: Figure from (Dersch et al., 2015). Different ways of modelling the metabolism. **A) Topological:** the focus is on the conversion of a metabolite (called substrate) to another metabolite (called product). **B) Quantitative:** the information is not only the transformation of M1 into M2 (with v2), but also the quantities required, that is the stoichiometry. One mole of M1 will create 2 moles of M2. **C) Steady-state metabolite balancing:** here, an important hypothesis is added, which is that no metabolite is accumulated nor depleted. **D) Dynamic metabolic balancing:** using kinetic laws and parameters, variations in concentration and reaction fluxes along time can be inferred. **E) Metabolic and isotopomer balancing:** here, under the steady-state condition, it is possible to infer intracellular fluxes precisely using ^{13}C labelled substrate.

1.3 Modelling metabolism

In Figure 1.4, different models that have been used for representing metabolism are shown with an increasing level of detail. It is important to observe that the models not only require information about the possible chemical transformations but also that an important hypothesis is verified which is whether the system is working in steady-state (or in quasi-steady-state QSS).

The *steady-state* is a perfect dynamic equilibrium such that the system maintains constant its variables. In the case of an organism uptaking substrates and excreting end-products, the hypothesis is that the concentration of metabolites and their fluxes are stable. Such phenomenon can be reproduced experimentally using chemostats where the extracellular concentrations of the substrates (sources) and end-products (sinks) are kept constant by continuously replacing the culture medium. External parameters (temperature, pH, cell density, etc.) are also controlled. It is usually assumed that the system is in quasi-steady-state, that is, any possible variation will happen at a time-scale that is larger than the one of the observations.

I next discuss more in detail the models presented in Figure 1.4 to the exception of Figure 1.4 E), since ^{13}C metabolic flux analysis has not been used in this PhD. However, it is important to know that this method can quantify intracellular fluxes experimentally. Cells are grown on ^{13}C labelled substrates. Once equilibrium is reached in terms of metabolites and of isotopes (the isotopomer fractions are constant over time), it is possible to measure the enrichment of the intracellular metabolites in labelled carbon and to infer the fluxes. Such methods are however not genome-scale, but instead usually focus on the central carbon metabolism. For more information on this method, the reader can refer to (Zamboni et al., 2009).

I will first introduce the differences between the dynamic modelling and the steady-state assumption, then I will talk about the topological (also called structural) model of metabolism and finally about the constraint-based model that allows to compute the fluxes in an organism.

1.3.1 The steady-state assumption or dynamic modelling

In the quasi-steady-state assumption (QSS), internal metabolites are not accumulated nor depleted. This can be translated into Equation (1.4), where $[X]$ is the concentration of metabolite X and s_j are the stoichiometric coefficients of the reactions producing or consuming X (where $j = 1, ..., n$ is the index of the reactions with $n = |\mathcal{R}|$). If reaction j is producing (resp. consuming) X, then $s_j > 0$ (resp. < 0). If reaction j is not using X, then $s_j = 0$. Finally, v_j is the flux going through reaction j per time unit.

Here the incoming and outgoing fluxes are balanced in such a way that the variation of the concentration of X along time, namely $\frac{d[X]}{dt}$, is null.

$$\frac{d[X]}{dt} = \sum_{j=1}^{n} s_j \cdot v_j = 0 \quad \texttt{or} \quad \frac{d[X]}{dt} = \sum v_{production} - \sum v_{consumption} = 0. \tag{1.4}$$

Such assumption will be essential in the methods described later on. It is represented in Figure 1.4 C).

In the model presented in Figure 1.4 D), concentrations and fluxes are not constant. The goal of such models is to infer their changes over time, that is to compute the variations of the metabolite concentrations and the reaction fluxes. Hence the system does not verify the QSS

assumption.

However, to obtain such variations, it is necessary to infer the kinetic parameters and laws. For example, a classical law is the so–called Michaelis-Menten. It is shown in Equation (1.5) where V_{max} is the *maximum velocity* and K_m is the Michaelis constant:

$$v = \frac{[S].V_{max}}{[S] + K_m}. \tag{1.5}$$

We will not go here into the details of kinetic laws. It is however important to observe that even though Equation (1.5) can reproduce some known enzymatic behaviours (*e.g.* an increase in substrate concentration leads to a larger flux rate, or the enzyme saturates at high substrate concentration), other laws are more adapted in various situations (for example, the Michaelis-Menten law does not take into account product inhibition or the concentration of more than one substrate). There are several other equations to describe the kinetics of reactions.

Estimating the kinetic parameters such as V_{max} and K_m is also not easy. They can be inferred quickly from the equation but experimental measurements of substrate concentrations and flux velocities can be difficult to obtain.

Furthermore, the possible variability in the *in vivo* data obtained can lead to some errors that can drastically change the value of the estimated parameters. Usually, the kinetic parameters are provided over intervals of values that can span several orders of magnitude. For more details, classical books of enzymology and/or metabolic control analysis can be consulted (Fell and Cornish-Bowden, 1997a).

The development of kinetic models has therefore been slowed down by the difficulty in obtaining precise kinetic parameters for the organisms.

The Michaelis-Menten law was derived under certain assumptions. Reactions transforming a substrate S into a product P can be decomposed into 3 reactions (shown in (1.6)).

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} SE \overset{k_2}{\rightarrow} P + E. \tag{1.6}$$

Those elementary reactions show first the binding of the substrate (S) to the enzyme (E) to form a complex enzyme-substrate ($S.E$) that will then transform the bound substrate into a product (P) that will be released, thereby freeing also the enzyme. Those elementary reactions follow the mass action kinetic laws ((1.7)):

$$v_1 = k_1 \cdot [S].[E] \ ; \ v_{-1} = k_{-1} \cdot [SE] \text{ and } v_2 = k_2 \cdot [SE], \tag{1.7}$$

where v_i represents the rate of reaction i, $[S]$ the concentration of the substrate, $[E]$ the concentration of the enzyme and $[SE]$ the concentration of the complex enzyme-substrate. Several groups proposed the same law under different assumptions. For Michaelis and Menten, the main hypothesis was that the first elementary reaction was at equilibrium and faster than the transformation of ES to $E + P$. Van Slyke and Cullen made the assumption that k_{-1} was much lower than k_2, meaning that once the complex enzyme-substrate is formed, the substrate cannot be released (making the substrate binding irreversible). Finally, Brigg and Haldane assumed steady-state (Fell and Cornish-Bowden, 1997a; Storey, 2005). It is possible to estimate the kinetic parameters from the values of k_i, with some variations due to these different hypotheses.

1.3.2 Dynamic analysis and Ensemble Modelling

Decomposing the reactions into smaller blocks respecting mass action laws is a procedure that has been used by (Tran et al., 2008) and then (Khodayari et al., 2014). Both proposed methods that were tested in *Escherichia coli* to infer kinetic parameters using a procedure called Ensemble Modelling.

The reactions of the chosen model were decomposed into *elementary reactions*:

$$X_i + E_i \underset{v_{i,2}}{\overset{v_{i,1}}{\rightleftarrows}} X_iE_i \underset{v_{i,4}}{\overset{v_{i,3}}{\rightleftarrows}} X_{i+1}E_i \underset{v_{i,6}}{\overset{v_{i,5}}{\rightleftarrows}} X_{i+1} + E_i. \tag{1.8}$$

We adopt here the notation of (Tran et al., 2008), using odd (resp. even) indexes for forward (resp. reverse) elementary reactions.

Each elementary reaction rate follows the mass action principle (presented in Equation (1.7)).

After the decomposition of the reactions, the fluxes of the model depend on the reaction reversibility (Equation (1.9)) and on the enzyme fraction (that is the concentration of free enzyme over the overall concentration of the enzyme).

The reversibility of reaction i is computed as:

$$R_{ij} = \frac{\min(v_{i,2j-1}, v_{i,2j})}{\max(v_{i,2j}, v_{i,2j}),} \tag{1.9}$$

where $2j-1$ and $2j$ are the forward and backward directions of step j of the reaction.

The parameter space of such model is large due to the multiple parameters arising from the reaction decomposition. Hence, there are several parameter sets that will fit the experimentally obtained data. However, those parameters might not be biologically correct, *i.e.* might not correspond to the known kinetics of the enzyme.

One can perturb such models (for example by knocking out or changing the expression level of an enzyme) and see if, in these new conditions, the models remain satisfiable. The resulting models are satisfiable if the steady-state flux distributions are similar to the one measured independently *in vivo* (where the similarity can be more or less stringent).

(Tran et al., 2008) modelled the central metabolism starting from the model of (Chassagnole et al., 2002), which included glycolysis, the pentose phosphate pathway and the phosphotransferase system (PTS). The network in (Tran et al., 2008) contains 25 metabolites and 29 reactions. (Khodayari et al., 2014) then improved the procedure by modelling up to 138 reactions and 91 metabolites. This model was tested for eight conditions (knock-outs) using measurements obtained *in vivo* by (Ishii et al., 2007). Refinements of the models were done using two genetic algorithms, trying to minimise the difference between the computed and the measured fluxes. Although the model published can still be improved since some fluxes were not accurately predicted, it remains a starting point for further development.

Obtaining increasingly larger kinetic models is useful for a global approach of the metabolism of an organism. However, inferring a genome-scale kinetic model has not been done up to now. In this work, we used kinetic models (more specifically from (Chassagnole et al., 2002) and (Khodayari et al., 2014)) to simulate fluxes and metabolite concentrations. This method uses normalisation steps to get rid of the absolute concentration values in the computations. This implies that to obtain the actual simulated concentrations, it is necessary to have knowledge about the *in vivo* concentrations; however such knowledge is not always available. Most reactions

are also modelled with the Michaelis-Menten law which may not be accurate. *Ensemble modelling* is nevertheless a good initial approach to create larger kinetics models.

Other methods have been developed to study the metabolism globally, that is using all the information known about the capabilities of an organism. Such approaches use the topology of the networks and sometimes the stoichiometry of the reactions. Hence, they do not infer any mechanistic mechanisms about the reaction processes, allowing to treat full networks with fewer information. In this case, the information about the flux and concentration dynamics along time is lost.

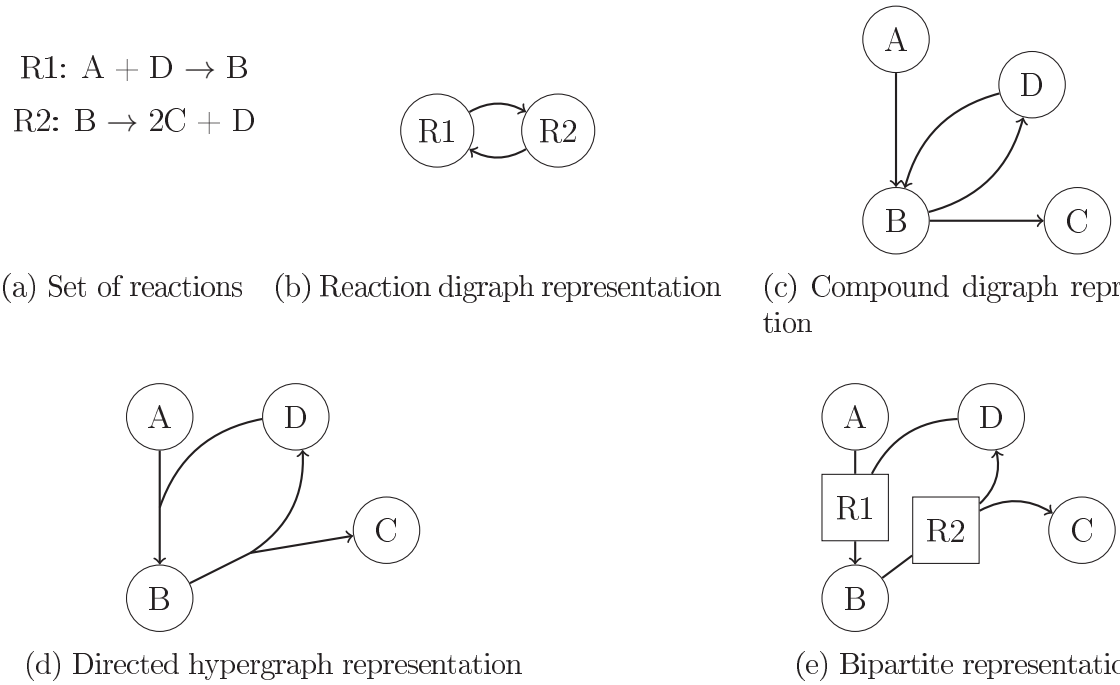
In the next two sections, I present some of the methods that can work at the scale of a whole organism.

1.3.3 Topological models

As presented in Figure 1.4A, it is possible to represent a metabolic network as a directed graph that will then allow to use graph theory for solving the various problems that need to be addressed. In a metabolic network, we can distinguish two entities that are:

1. *the metabolites* which may be seen as chemical species;
2. *the reactions* which represent the enzymatic conversions of some chemical species into other species.

There are several possibilities for representing the relations between metabolites and reactions. These are shown in Figure 1.5 for the set of reactions in Figure 1.5a.



(a) Set of reactions (b) Reaction digraph representation (c) Compound digraph representation

(d) Directed hypergraph representation (e) Bipartite representation

Figure 1.5: Directed graph representations of a set of reactions.

In a *reaction digraph* (Figure 1.5b), the vertices are the reactions and an arc links a reaction that produces a metabolite x to a reaction that uses x as a substrate.

In the case of a *compound digraph*, the vertices are the metabolites and an arc links a substrate to a product of a same reaction. As shown in Figure 1.5c, the information about the joint use of subtrates or products is lost (for example, it is not possible to know that A and D are both needed as substrates for reaction R1 to produce B).

In this PhD, we worked with the *directed hypergraph* representation (Figure 1.5d), where the arcs (hyperarcs) of the hypergraph link the substrates to the products (substrates and products can be multiple). Here, we work exactly with the biological objects considered in a metabolic network, that is, the reactions (hyperarcs in the hypergraph) and the metabolites (vertices in the hypergraph). The structure of the network created thus preserves the functional properties of the metabolism. However, directed hypergraphs have been less studied and taking into account the multiplicity of the vertices used by a hyperarc is more complicated.

The last representation shown in Figure 1.5e is a *bipartite digraph*. Conceptually, it is the same as a hypergraph. It uses simple arcs (that is going only from one vertex to another), but the vertices of the graph are divided into two sets: the metabolites (circles here) and the reactions (squares).

The hypergraph or bipartite representations can be easily translated into the stoichiometric matrix (and vice-versa) by adding the stoichiometry on the hyperarcs or arcs. However, with the stoichiometric matrix it is not possible to represent autocatalytic reactions (where a metabolite is both substrate and product of a reaction).

Once the biological system is converted into a known object in graph theory, it is tempting to apply common graph measures. However, such measures are not necessarily meaningful from a metabolic point of view since they were not initially conceived for studying biological processes (van Helden et al., 2002; Arita, 2004; Croes et al., 2006; Lima-Mendez and van Helden, 2009; Klein et al., 2012). For example, a commonly used measure is the diameter (longest among all the shortest paths between any two vertices). Such measure is influenced by the ubiquitous compounds and co-factors. Those very connected vertices are hubs in the networks but do not really participate in the carbon exchange. Studying the topology of the network allows to use common algorithms from graph theory or propose new ones, that have an interest both in computer science/complexity theory, and in biology.

For example, topological methods have been developed to infer the minimal set of substrates (precursors) to produce target compounds (the biomass here) (Cottret et al., 2008; Acuña et al., 2012b) or to analyse the metabolites exchanged between different organisms (Cottret et al., 2010a). They are also used to perform metabolic pathway enumeration (van Helden et al., 2002) and prediction (Mithani et al., 2009; Carbonell et al., 2012). Finally, some methods try to identify subnetworks used by the organism in certain conditions, or to point out the ones that behave differently between two conditions. Chapter 2 proposes such a topological method. It is an improved model and a new algorithm in relation to the ones proposed by (Milreu et al., 2014; Acuña et al., 2012a) to identify the reactions impacted by a change in conditions.

Currently however, constraint-based models are increasingly more used to explore the capabilities of a metabolism.

1.3.4 Constraint-based models

Kinetic models are useful to predict the dynamics of metabolism in terms of fluxes and also of metabolite concentrations. However, it requires to know (or infer) kinetic laws and parameters. Hence, some methods use the *quasi-steady-state assumption* to predict the flux distribution in a specific condition for genome-scale models (not taking into account the dynamics of such fluxes and concentrations). Such methods only use the stoichiometry of the network and the growth conditions (including medium concentration).

An essential assumption in this case is that the cell is optimising a specific metabolic function. Often in flux balance analysis (FBA), the objective is to maximise the biomass (*i.e.* the growth yield μ). The hidden hypothesis is that evolution is such that the organism that grows best, that is in a fastest way, would win any competition and that the one and only goal is to grow the best possible. Whether this is an appropriate hypothesis is discussed in (Schuster et al., 2008) together with special cases. Various authors proposed new objectives suggesting that the optimal growth yield may not be adapted (Schultz and Qutub, 2015; Schuster et al., 2008). Several methods that use optimisation techniques are currently available, and are discussed in (Zomorrodi et al., 2012; Klamt et al., 2014).

In flux balance analysis, the metabolic network is represented as its stoichiometric matrix $\mathcal{S}$. The steady-state assumption is then represented by the Equation below:

$$\mathcal{S}.v = 0 \tag{1.10}$$

The flux distribution vector v is obtained using an objective function such as maximal growth, or ATP production (see (Orth et al., 2010) for more details). The overall optimisation problem is summarised in Equation (1.11) where c is the vector of weights balancing the reactions to optimise and Z is the resulting objective.

$$\begin{aligned} \max / \min \quad & Z = c^T v \\ s.t \quad & \mathcal{S}.v = 0 \\ & lb_i \leqslant v_i \leqslant ub_i, \forall i \in \{1, ..., n\} \end{aligned} \tag{1.11}$$

The variables lb_i and ub_i are the upper and lower bounds of the flux i (where i goes from 1 to n, n being the number of reactions). Using boundaries for the flux vector allows to reduce the space of possible solutions and to retrieve reasonable flux vectors. In particular, external fluxes can be constrained to a set of measured values, or to values inferred from previous knowledge. One straightforward way to do this, is by constraining the bounds with the experimental measurements of external fluxes (that is substrate uptake and metabolite export) when those are known. However, few measurements exist for the intracellular fluxes since the experiments are costly and prone to errors. FBA analyses have been limited to consider one solution that is solver-dependent. The overall idea of FBA is summarised in Figure 1.6.

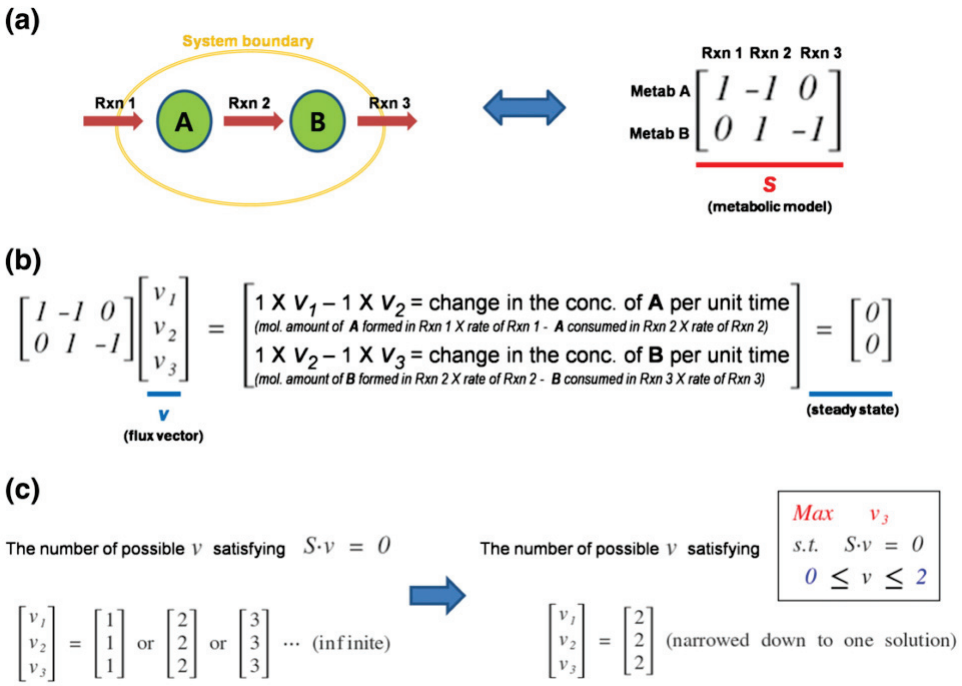


Figure 1.6: From (Kim and Lun, 2014), a simple FBA for a non branching pathway. In **a)**, is presented the network and its stoichiometric matrix. The boundary is defined and allows uptake and export. Here, $s_{1,2}$ = -1 since one mole of A is consumed by the reaction $Rxn2$ to produce one mole of B. In **b)**, the steady-state assumption is added to look for the flux vectors such that the metabolite concentrations do not change along time. In **c)**, we can see that there is an infinite number of flux vectors that are solution (all the vectors represented plus those obtained through multiplication by a scalar). Using the objective function and boundaries for the reaction fluxes, it is possible to obtain one optimal solution (that is, in this case, unique).

Another formulation called Flux Variability Analysis (FVA) is used to understand the variability of the flux distribution vector in a specific condition. It computes the range of values that a certain flux can take for a same value of the objective function (Mahadevan and Schilling, 2003). However, the bounds can remain large and, while it is possible to identify fixed reactions, that is reactions which cannot vary for a given value of the objective function, most of them still have a large interval of variability.

Finally, among the remaining methods using constraint-based techniques for analysing metabolism at the genome-scale level, we can also cite Flux Coupling Analysis which looks for correlations and dependencies between fluxes to determine the coupling of reactions (Burgard et al., 2004).

FBA like approaches propose an optimal solution whereas several exist. Indeed, one flux distribution vector v is computable in FBA, but many flux vectors can respect the steady-state assumption ($S.v = 0$) and flux boundaries, and have a same optimum value Z (which is minimised or maximised). Such vectors form what is called the *flux cone*.

There are often more reactions than metabolites in the network, thus more variables than constraints creating multiple solutions. This multiplicity of solutions in FBA analysis is also caused by structures in the network such as cycles, parallel pathways or branching points that create alternative optimal solutions.

Different approaches attempted to bypass the multiplicity of the FBA solutions. In (Kelk

et al., 2012), by an efficient *pre-processing* of the genome-scale network, the authors managed to characterise the optimal solution space by identifying the topological subnetworks that are creating the combinatorial explosion. They thus offer a structural analysis of the network that can explain FBA variability.

Other approaches are based on sampling the possible solutions to obtain different flux distributions. Several sampling methods have been proposed, for example by (Binns et al., 2015) where the authors propose to penalise at each step the objective function to cover the feasible flux space as much as possible.

Approaches such as elementary fluxes modes (EFMs) or extreme pathways (a subset of elementary flux modes (Schilling et al., 2000)) can define the polyhydron formed by the optimal solutions of the FBA analysis. Obtaining such polyhydron is a difficult enumeration problem. Although different algorithms have been proposed for medium-size networks, it is still computationally expensive to enumerate such objects for genome-scale models.

It is thus possible to model metabolism using the knowledge that we have about the chemical transformations occurring in an organism. However, these transformations depend on the enzymes that may catalyse them, namely on the transcription and translation of the genes that code for such enzymes, and on several regulation mechanisms as is discussed in the next section.

1.4 Regulation and control

Organisms do not function in the same way in different conditions, in particular at the level of their metabolism. Indeed, although an organism is identified by a unique metabolic network, that is a same set of possible reactions, it may not have the same usage of such network in two different conditions. Furthermore, metabolic pathways share metabolites. There is therefore a need to control which pathways are activated for the organism to perform well in certain conditions. For example, in the case of diauxic growth, two substrates are available but they are used sequentially. Moreover, we can compare normal growth with growth in the presence of a toxic compound (*e.g.* yeast exposure to cadmium (Milreu et al., 2014)) or the changes that result from an interference in the expression of a regulatory element (such as *csrA* which is a small RNA binding protein (Wei et al., 2000)).

1.4.1 Mechanisms

(Fell and Cornish-Bowden, 1997b) and others proposed to make the distinction between *regulation* and *control*.

The idea is that an organism *regulates* its metabolism when it maintains some of the latter's variables constant despite external variations. Here, what remains constant is not completely clear and well defined: is it the intracellular metabolite concentrations or the maintenance of growth, or else ATP synthesis? On the contrary, *control* refers to the active change of a system in response to perturbations or to external signals.

These two concepts are not easily distinguishable, and it is difficult to know which are the regulated and which are the controlled variables. For example, is it the intracellular metabolite concentrations that are regulated through the control of the reaction fluxes (by changing the enzymes quantities) or the other way around? Indeed, according to the kinetic laws, reaction

fluxes depend on metabolite concentrations (their substrates and possibly some metabolites that are regulators).

Whether we are dealing with *regulation* or *control* therefore depends on the context and the type of biological objects observed. Notice that in later chapters, we will mostly use the term *regulation* as this is the generalisation that is commonly adopted.

Regulation and control of metabolism rely on several mechanisms. It is possible to distinguish (Metallo and Vander Heiden, 2013):

1. transcriptional regulation: where gene expression can change in-between conditions, allowing a change in enzyme concentration, hence a context-specific response;

2. allosteric regulation: which requires the binding of an allosteric effector to the active site of an enzyme; such effectors can be, for instance, the end-products of pathways;

3. post-translation modification (such as enzyme phosphorylation): those can impact enzyme activity.

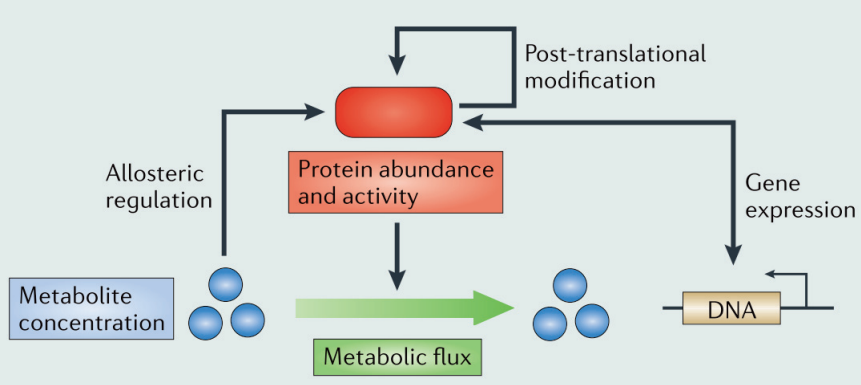


Figure 1.7: From (Chubukov et al., 2014), possible regulations of the metabolism are done by modifying protein abundance (through gene expression) and activity (through post-translational modification or allosteric regulation).

These mechanisms are not acting at the same time-scales. It is common to speak about "coarse" control which is the control of the amount of enzyme in the cell. Transcriptional regulation is usually estimated with a response time in minutes or hours. It is possible to refer to a "fine control", working on a shorter time-scale, that will not act on the concentrations of the enzymes but on their activity using covalent or non-covalent interactions. However, it is known that those different types of regulations can take place together. For example, in the case of multiple substrates availability for *E. coli*, if glucose (which is a preferential substrate) is present, there will be a direct inhibition of the transporter of other substrates through the dephosphorylation of EIIA. Additionally, there is the down-regulation of the other types of carbon-uptake systems. We refer to (Chubukov et al., 2014) for a review on what is known on the regulation and control of microbial metabolism.

1.4.2 Regulating pathways

As previously indicated, it is important to understand how metabolism is controlled to direct the fluxes at branching points in the metabolic networks.

For a long time, the *limiting step* hypothesis was admitted, that is the hypothesis that the rates of the reactions along the pathways depended only on one enzyme/reaction.

However, *metabolic control theory* or *metabolic control analysis* (Fell, 1992; Fell and Cornish-Bowden, 1997b; Sauro, 2009) proposed that the control of the metabolic pathways is, generally, distributed. Furthermore, enzymes are usually considered to be available in abundance. In this case, the main theory is that the regulation of the reaction rates is shared among most of the reactions of a pathway. Indeed, some pathways are regulated by one reaction (thus are in agreement with the limiting step hypothesis), however it is not any more the only possible mechanism for regulation. Shared regulation appears to be more relevant in most cases.

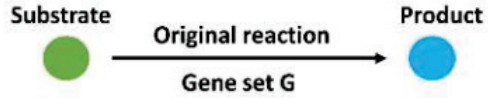
Two roles have been recognised for enzymes (Fell and Cornish-Bowden, 1997b): one that corresponds to a regulatory effect (stabilising pathways), also called *metabolic homeostatis*, and another that corresponds to a *control action* (changing the states of a pathway in response to an external signal).

Metabolic control analysis (MCA) is based on the properties of enzymes and was established by analogy to other mechanistic systems. It is based on a sensitivity analysis. This formalism has been developed to infer responses of a system (here the metabolism of the organism) to perturbations of its current steady-state. It is based on two indicators of the response of a reaction: the control coefficient of the reaction and the control coefficient of the concentration which represents the effect of the change in enzyme concentration on a flux (resp. concentration) (Fell, 1992). Performing MCA requires intensive *in vivo* experimentations to infer those parameters and it is not adequate to genome-scale studies. However, it has been successfully applied in various bioengineering experiments (for a review, see (Costa et al., 2015))

1.4.3 Using regulatory networks and transcriptional data

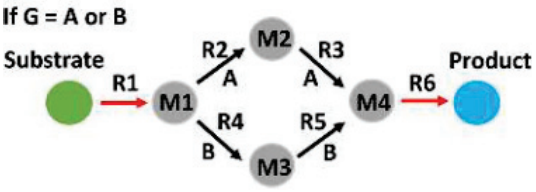
In this section, we briefly discuss some papers linking gene expression to metabolism. An idea in this PhD was to handle both metabolomic and transcriptomic data to try to unravel the metabolic regulation that creates the difference between two steady-states (problems described in Chapters 2 and 3). However, several issues appeared while trying to conceive a model that would encompass both metabolomic and transcriptomic data. Even ignoring possible issues in assembling, quantifying and annotating transcriptomic data, the trancriptomic regulation of metabolism seems to be noisy as pointed out in recent papers.

First, there is not a one-to-one relation between a gene and a reaction (gene-reaction association); several genes can lead to enzymes catalysing a same reaction (called isoenzymes) or several genes can be needed to catalyse a reaction (protein complex). One possible response to such issue was presented by (Zhang et al., 2015) and is called Logical Transformation of genome-scale Models (LTM). The authors transformed a metabolic network using the Gene-Protein-Reaction (GPR) relationships. We present this transformation in Figure 1.8. GPRs are boolean formulae composed of AND and OR operators that indicate the gene requirements for a reaction to be catalysed.

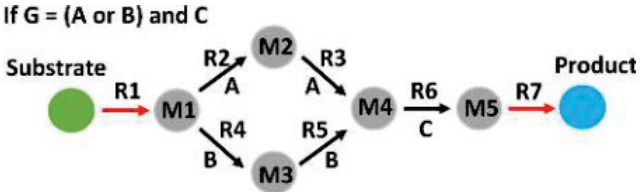


(a) Initial network, one reaction transforming *Substrate* into *Product*.

(b) Case of an AND relation.



(c) Case of an OR relation.



(d) Combined case of an AND and OR relation.

Figure 1.8: From (Zhang et al., 2015), a toy example for the LTM transformation. Here, the transformation is shown in the case of a simple gene set. The metabolic network is transformed by adding new vertices and arcs according to the GPR of the reaction. The arcs in the transformed network do not represent reactions anymore but genes.

However, even if we consider that the network can topologically convey all the information required for the associations of enzymes and genes, various recent publications seem to show that transcription regulation is not specific. Indeed, in various types of organisms such as plants (oilseed (Schwender et al., 2014)), prokaryotes (*Bacillus subtilis* (Chubukov et al., 2013)) and smaller eukaryotes (yeast, *Saccharomyces cerevisiae* (Daran-Lapujade et al., 2004, 2007)), no direct correlation has been identified between the level of expression of the transcripts and the metabolic activities of the organisms. This was the case in (Schwender et al., 2014) even when the experiments were run in two plant oilseed rape accessions. Since it was two different accessions, it could be expected that the difference in fluxes between them would be due to a difference in gene expression (*i.e.* in transcript levels), but no clear correlation was found.

It therefore seems that the main regulations are done by covalent binding and post-transcriptional modifications, or at least that we cannot derive any general regulation rule, since all kinds of regulations appear intertwined.

However, some methods have been developed to integrate transcriptomic data into the metabolic models. The obtained information can be used to restrict flux distributions in constraint-based models for specific growth conditions. Various works adopt a boolean formulation of presence/absence: if transcripts are not detected in replicates or are below a certain threshold, the corresponding enzymes are considered absent.

We refer to (Kim and Lun, 2014) and (Machado and Herrgard, 2014) for more details.

Knowledge on the regulation and function of metabolism has led to gather increasingly more information but it also paved the way for the creation of models and algorithms to help infer the modifications that could be done to metabolism in order to produce specific compounds of interest.

1.5 Synthetic Biology

With the recent progresses in genome editing techniques, synthetic biology has been used to produce compounds of interests with microorganisms. Genome editing technologies such as

CRISPR/Cas9 made it possible to modify genomes either by inserting or deleting DNA sequences (Halweg-Edwards et al., 2015). Up and down regulation of genes can also direct fluxes toward a metabolic goal (Nakamura and Whited, 2003; Jensen and Hammer, 1998; De Ruyter et al., 1996).

Despite increasingly more accessible techniques, there is still a need for computational methods to predict how to modify an organism in order to reach an enhanced productivity. Indeed, a trial-and-error approach requires time, and may lead to side effects (which however are interesting to study in themselves). It is necessary to understand how the system (*i.e.* the organism) will react to insertions/removals for instance, and how globally its metabolism will be modified. A global view of such systems is thus interesting. This can be obtained using computational methods to predict which enzymes to insert or suppress. I will first present the *retrobiosynthesis* approach in synthetic biology and then how *in silico* methods are used to predict an optimum yield.

1.5.1 Retrobiosynthesis

Retrobiosynthesis has been defined by (Hadadi and Hatzimanikatis, 2015) as:

> "a discipline that involves the design, evaluation, and optimization of *de novo* biosynthetic pathways for the production of high-value compounds and drugs from renewable resources and natural or engineered enzymes."

The word *retrobiosynthesis* comes from the fact that the starting point of the associated methods is the target compound, hence there is a need to work backwards to discover the possible pathways and sources.

In Figure 1.9, (Hadadi and Hatzimanikatis, 2015) propose a global pipeline for the implementation (from *in silico* to *in vivo*) of new pathways. The idea here is to first select the target molecule(s) and then to use several methods and databases to screen the possible pathways.

As a first step, enumeration of the possible pathways is important. To do so, it is possible to either select known reactions in several organisms, or to infer chemically possible reactions even if no enzymes or resulting compounds have been annotated. The result is then a set of heterologous pathways composed of enzymes previously annotated, or that are putative. Such approaches have been developed by different groups that use structural chemistry and then score the resulting pathways (see (Hatzimanikatis et al., 2005; Rodrigo et al., 2008; Cho et al., 2010; Carbonell et al., 2011, 2012, 2014) for a non exhaustive list of some related works). This first step (*in silico pathway design*) generates multiple possible pathways. Since in general there is a combinatorial explosion of possible pathways due to an immense space of chemical possibilities (in particular, if *de novo* reactions are considered), it is important to prune non-feasible pathways using known databases, mass balance and thermodynamics analysis. It is then possible to score the remaining pathways that are assumed to be "feasible".

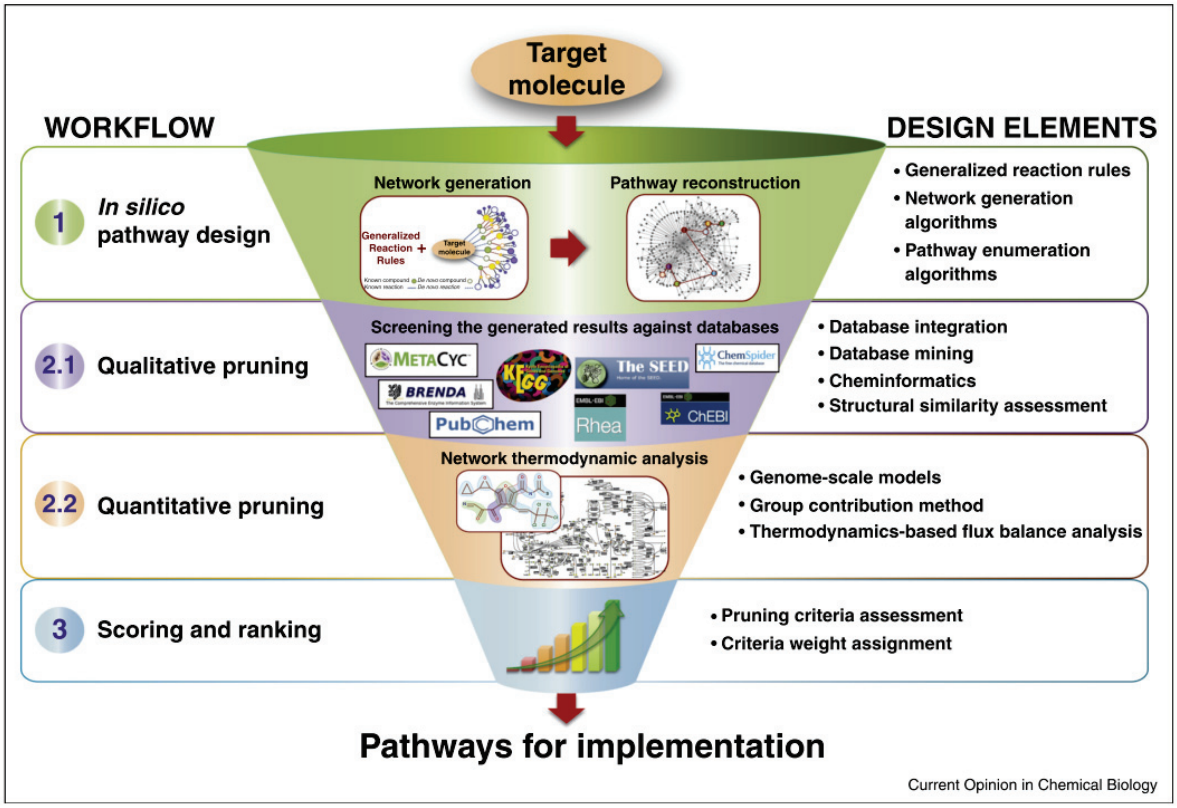


Figure 1.9: From (Hadadi and Hatzimanikatis, 2015), the authors propose a complete pipeline from *in silico* to *in vivo* for the production of target compounds.

Pathway enumeration methods can be roughly divided as in Sections 1.3.3 and 1.3.4: "topological" that correspond to a qualitative research of pathways and "quantitative" which models how to direct the fluxes in order to reach the maximum possible yield. I will not discuss the use of putative reactions (inferred using the chemical structures of compounds) as it was not part of this PhD work.

It is important to know whether an organism is capable of producing a specific product endogenously, that is with its metabolic capabilities. Those are theoretical capabilities in the case of network topology and can be specific of the substrates provided. The goal is to be able to obtain a profitable production. To reach a better gain, it is possible to use low-cost substrates (such as lignocellulosic biomass obtained from plant and wood industries) to produce high-value compounds (such as pharmaceuticals).

If it is not possible to produce endogenously the desired target compounds, one may consider to insert new enzyme-coding genes using the technologies developed in the last 20 years. Several algorithms have been developed to find the metabolic road from substrates to targets. We also propose in Chapter 4 a model of pathway enumeration applied to communities.

Once the production pathway is known, the metabolic network is completed with such a pathway. It is then possible to infer how to reach an optimum production.

1.5.2 Optimising the production yield

The hypothesis commonly encountered is that an organism will always try to optimise its biomass. Unless the desired product is coupled to the growth of the organism, there is therefore the need to force the metabolism to use some of the given resources (substrates in the medium) to produce a compound that might not be useful for its growth.

A commonly used technique is constraint-based modelling that allows to understand which are the gene expression modifications to perform to maximise the yield. Such modifications are usually knock-outs that result in the inactivation of the genes leading to no expression anymore. There have been several constraint-based formalisations proposed, using Integer programming (ILP) or Mixed Integer Programming (MILP). We will not discuss all of them. It is important to observe that the problem formulations are usually really close but some of the hypotheses differ, thus leading to different solutions (Chowdhury et al., 2014). In the case of heterologous pathways, those methods can be applied to the genome-scale model of the chassis organism chosen, modified to include the new reactions. It is therefore necessary to know the production pathway(s).

OptKnock was developed by (Burgard et al., 2003). It is a bilevel framework where the idea is to find a flux distribution vector such that the product yield is maximised (through gene knock-outs) while maximising the biomass by removing some reactions using integer variables (setting the fluxes to zero). The original formulation is presented in Equation (1.12) where y_j is a boolean variable for the reaction j. The value of y_j is 0 if the reaction is knocked out, 1 otherwise. Here the maximum number of knock-outs is set to $\mathcal{K}$. The steady-state is a constraint ($\mathcal{S}.v = 0$). The outer problem is to maximise the product yield v_p for a certain flux vector v. The inner problem is to maximise the biomass μ. As previously, every reaction rate can be bounded by an upper and lower bound (ub_j and lb_j).

$$
\begin{aligned}
&\underset{y_i}{\text{maximise}} \quad v_p \\
&\quad \text{s.t.} \\
&\qquad \begin{cases} \underset{v_j}{\text{maximise}} \quad \mu \\ \quad \text{s.t.} \\ \qquad \mathcal{S}.v = 0 \\ \qquad y_j \cdot lb_j \leqslant v_j \leqslant ub_j \cdot y_j, \forall j \in \mathcal{R} \end{cases} \\
&\qquad y_j = \{0, 1\} \\
&\qquad \sum_{j \in \mathcal{R}} (1 - y_j) \leqslant \mathcal{K}
\end{aligned}
\tag{1.12}
$$

In OptKnock, an optimistic prediction is done. The biomass is optimised for the full network by setting the fluxes of the reactions. The outer problem then tries to maximise the production of the target product. The idea in this case is that the organism after knock-outs of some of its reactions will optimise its biomass, even at the expense of a re-routing of the metabolism and of having to change a big part of its fluxes.

An alternative hypothesis was proposed by (Segrè et al., 2002) with MOMA (Minimal Of Metabolic Adjustement). The formulation of MOMA is such that there is no optimisation of the biomass rate. Instead, the authors minimise the differences between the wild-type flux and the mutant flux distribution. This hypothesis is not unreasonable, hence even if OPTKNOCK gives a best-case scenario (maximised biomass and maximised production), the *in vivo* implementation of the knock-out may lead to no production of the desired product in some cases. There can thus be many solutions and also other effects not taken into account (such as regulatory effects). Furthermore, the wild-type flux distributions must be known (usually by FBA) and must be accurate.

The ROBUSTKNOCK method proposed by (Tepper and Shlomi, 2009) reflects this doubt whether the organism after modifications will optimise its biomass and production. The authors offer a worst-case scenario by maximising the minimum product yield. Hence, there are two layers of optimisation.

The outer problem is a max-min optimisation, where the goal is to find genetic modifications (knock-outs) such that the minimum production rate is maximised. The inner problem is the same as for OPTKNOCK, that is the maximisation of a cellular objective.

Such differences are represented in Figure 1.10.

We can see that the maximum production for A (computed with OPTKNOCK) is higher than for B (computed with ROBUSTKNOCK). Nevertheless, for a same biomass yield, the product yield of A can be smaller than the maximum production predicted as shown by the blue vertical dotted line. The predicted yield in the production of mutant B is lower than the maximum production yield of mutant A. Nevertheless, this yield is guaranteed since it is the minimum production yield for the given growth rate (biomass yield). Mutant A can have a different production yield for the growth rate selected, and it is shown that the worst possible yield (minimum yield of mutant A) can be quite low. The variability in the production yield of A is worrying because this means that the implementation *in vivo* of the proposed knock-outs could lead to a far lower yield than what is predicted.

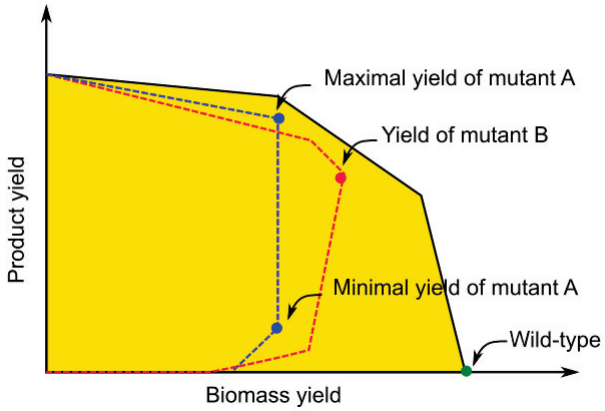


Figure 1.10: From (Klamt et al., 2014), the representation of the solution space for two mutants A (computed with OPTKNOCK, the solution space is in blue-dotted line) and B (computed with ROBUSTKNOCK, the solution space is in red-dotted line). They carry different knock-outs ($y_A \neq y_B$).

For the three methods presented, the solution is always to knock out reactions, that is to

block them completely. However recent methods propose to infer up and down regulation targets. Indeed, totally removing a reaction can be lethal (even if it was not predicted by the metabolic model due to regulation effects). Hence, regulating the expression of the enzymes can help redirect the fluxes without preventing an organism's survival.

RobOKoD (Stanford et al., 2015) does not guarantee a global optimum but instead identifies reactions to either knock out, over or under-express. The claim is that if product formation and growth are not coupled, using FBA and a bilevel approach is not the best solution. Here, there are two assumptions:

1. one is the need to avoid carbon losses through alternative routes and coproduct formation ("peripheral pathways"). To do so, each reaction is attributed a score called MCT (metabolite consumption test). The MCT is an indicator of the importance of a reaction for the producing pathway. It reflects whether a reaction is involved in the biomass rate and/or in the product formation. This score allows to target the knock-out reactions that could lead to carbon loss (*i.e.* not lead to the product formation).

2. the second is the fact that the FVA of a reaction will change depending on its functionality in the biomass and/or product formation.

The authors use Flux Variability Analysis (FVA) to profile reactions in their subnetwork. The procedure reflects the underlying challenge in synthetic biology which is that the production of the compound of interest is rerouting some of the carbon flux that is used for the biomass. The RobOKoD method was applied to the production of butanol in *E. coli*, and showed good results compared to OptKnock and RobustKnock that lead to non viable strains. However, it was not yet applied to large models (in the *E. coli* model used, there were 95 native reactions and in total 125 reactions and 93 metabolites).

Finally, it is worth mentioning OptStrain (Pharkya and Maranas, 2006), which is an optimisation framework to design strains, that will determine the maximum product yield and minimise the number of non native reactions.

Computational tools are therefore of great help, not only to suggest non-trivial pathways but also to propose the most efficient ones. As always, there is here a strong need of collaboration between researchers from different disciplines in order to adapt the models to specific cases by transposing biological knowledge into the computational models and algorithms. Finally, the methods presented propose targets for knock-outs or regulation of the expression levels of enzymes. Nevertheless, some *in vivo* trials can be unsuccessful because of side-effects. Indeed, what is not captured here are regulation effects as well as lethality. Moreover, there is a one-to-one assumption that the change in gene expression will affect the fluxes in proportion. This is not always the case, but such initial models have proven to provide good indications already.

We refer to (Copeland et al., 2012) and (Hadadi and Hatzimanikatis, 2015) for more information, together with (Lewis et al., 2012) for a review on constraint-based reconstruction and analysis methods.

We presented above modelling methods for single organisms, but lately, a lot of effort has been made to study communities, either in their natural habitat, or in artificial ones using the same kind of modelling framework.

1.6 Communities

Communities of organisms are widely present in nature. Several advantages can result from an association with others. Different metabolic capabilities can profit from the division of labour (not having only one species to sequentially perform complex tasks) but also some interactions may alleviate inhibitions or even enable to complement one another's growth. Some syntrophy (meaning cross-feeding) examples are shown in Figure 1.11

(Bernstein and Carlson, 2012) described possible associations of microorganisms, whether these occur in nature or in the laboratory. Some interactions such as cross-feeding can be directly modelled within the small molecule metabolism framework. However, others are more difficult to depict. Indeed, using topological methods, exchanges such as electron transfer are not modelled. However, they can be captured in a quantitative manner using the constraint-based model and the linear programming framework. Another example given for the efficiency of communities is through spatial organisation. In Figure 1.11A), aerobic organisms form a biofilm to protect the anaerobic members of the consortium. (Bernstein and Carlson, 2012) described three different types of consortia. A *natural* consortium is composed of organisms that can be found jointly in nature, whereas *artificial* consortia are formed in the laboratory with organisms that are newly associated. Finally, a *synthetic* consortium is formed with engineered microorganisms.

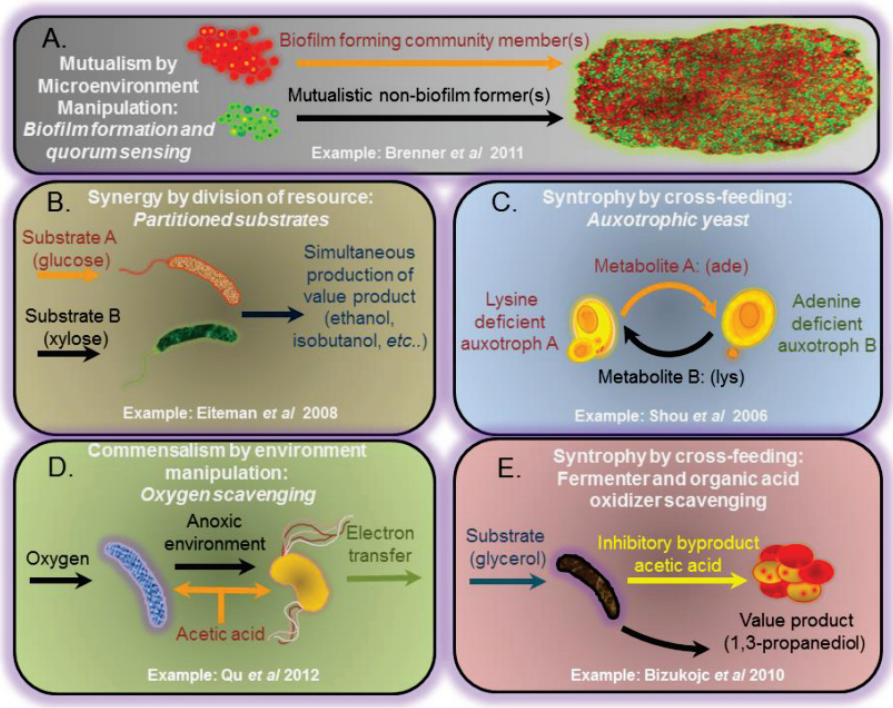


Figure 1.11: From (Bernstein and Carlson, 2012), different types of interactions that can occur in consortia (artificial or natural).
In **A)** is shown a particular type of mutualism where one species is responsible for a an environment that is favourable to the fitness of the other anaerobic member by creating a biofilm that does not let oxygen pass.
In **B)** and **C)**, both organisms require something specific from the environment, in **C)** it is the other strain that can provide the missing amino-acid (adenine or lysine).
In **D)** and **E)**, one organism scavenges an inhibitory by-product, again to create a more profitable environment.

1.6.1 Natural communities

Natural communities are studied in order to understand ecological niches of specific interest, such as the microbiome of human or other animals. Indeed, this system has to be considered as a whole to better understand the mechanisms that can arise from such communities.

Interactions inside communities have been defined previously, depending on the type of benefit or disadvantage the individuals of a consortium retrieve from the interactions.

It is important to observe that modelling methods require to identify the strains present in the environment. Under-represented species remaining unidentified (or undetected) can still play a role of importance in the equilibrium of a consortium. Furthermore, even when the composition of a community is known, in order to apply such methods, the metabolic networks of the members of the community must be available and this issue also restricts the scope of the studies that are possible.

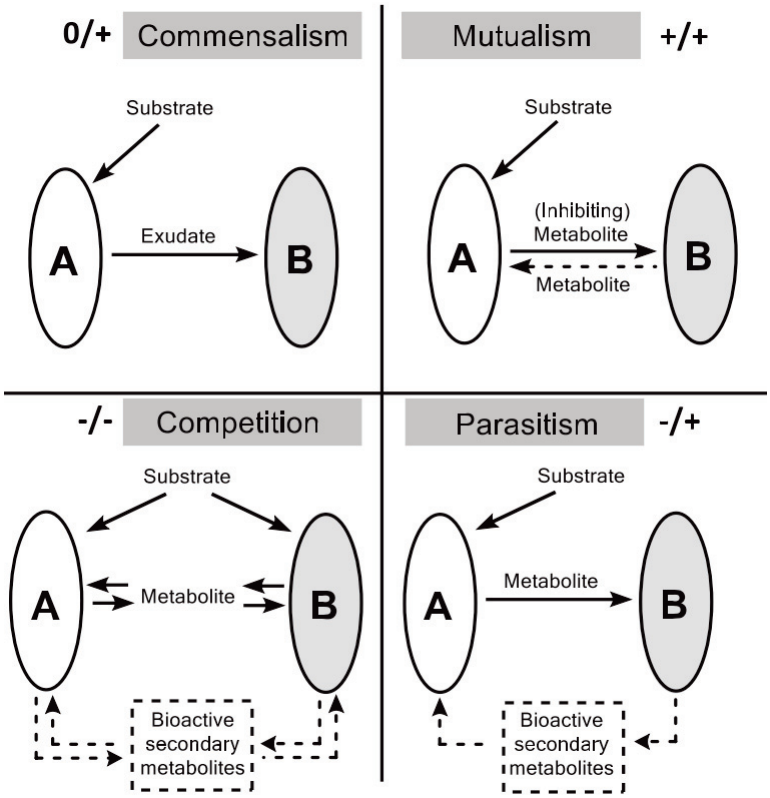


Figure 1.12: Some interactions that can be encountered between two organisms (**A** and **B**), from (Jagmann and Philipp, 2014). In *commensalism*, **A** does not have an advantage to be in contact with **B** but **B** uses a metabolite produced by **A**. In *mutualism*, both organisms exchange metabolites that are useful for one another. In a *competition*, both organisms compete for a substrate or for intermediate metabolites, and this competition can be supported by secondary bioactive metabolites (*e.g.* antibiotics). Finally, in *parasitism*, one organism (here **B**), scavenges on the production of the other (here **A**). **B** can use secondary bioactive metabolites to induce **A** to release the necessary metabolites.

Recently, the same questions as for single organisms arose. It is interesting for example to infer what are the possible flux distributions inside a consortium to identify the division of labour, but also to infer the metabolite exchanged and what are the costs and benefits of the associations. Methods developed for single organisms (presented in Section 1.3) have therefore been extended to communities.

Most of them are based not only on the topology of the network but also on the stoichiometry. The extension is quite straightforward when the organisms have a mutualistic interaction, meaning that the association is mandatory; each organism depends on the other (or more particularly in this case, on a metabolite produced by the other) to grow. Some models have used a compartimentalised formulation, by analogy with the eukaryotic models available. For example (Stolyar et al., 2007) studied the association of *Desulfovibrio vulgaris* and *Methanococcus maripaludis* as a three compartment system (the two organisms and the exchange compartment). In the case of balanced growth, (Khandelwal et al., 2013) also proposed a formulation adapted from FBA called cFBA (community flux balance analysis).

The modelling of interactions other than mutualistic is less direct. Some of the possible interactions are presented in Figure 1.12. OptCom (Zomorrodi and Maranas, 2012) takes advantage of the constraint-based framework using a multi-level and multi-objective formulation. (Zomorrodi and Maranas, 2012) reported that all kinds of interactions within the community can be taken into account. A community objective is maximised as the outer problem, and then the inner problem is composed of the individual objectives of each organism in the community, the whole being modelled as a multi-level problem (shown in Figure 1.13).

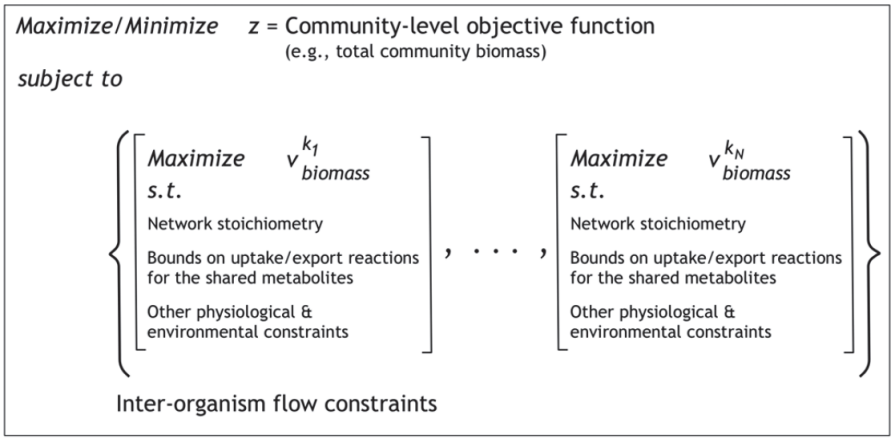


Figure 1.13: From (Zomorrodi and Maranas, 2012), the *OptCom* formulation with k_i corresponding to the organism i that is part of the consortium.

Dynamic flux balance analysis (dFBA) (Hjersted and Henson, 2009) allows to take into account the time dimension by using a dynamic mass balance equation for the concentration of the extracellular metabolites. It has been applied by (Hanly and Henson, 2011) to mono-cultures and then to a co-culture in the case of balanced growth.

Adaptations of flux balance analysis and of other methods are starting to take into account another dimension that appears to be essential: this is the spatial organisation of the communities and of the species involved in it. Indeed this is important in a biofilm for example (see Figure 1.11A)), but also because in nature and in culture, the substrates, metabolites and gases are not always homogeneously distributed, creating some gradients of concentration. Some areas will therefore be more profitable for a certain organism because of the concentrations at its position. The cell localisation is of importance and new methods are also developed using such constraints (see (Cole et al., 2015; Henson, 2015; Chen et al., 2016)).

The positive interactions found in nature can be reproduced in the laboratory, allowing the creation of better growing consortia. Furthermore, this is of importance in synthetic biology and for the production of compounds of interest. Indeed, in this kind of production, there is the possibility of the joint production of an interesting compound together with an inhibitory one for example (see Chapter 4, Section 4.5.2). Using communities is a way of stabilising a production and of reaching higher yields.

1.6.2 Artificial communities

Synthetic biology is a field in expansion that can have a high impact in terms of financial return. Indeed, engineering microorganisms to produce compounds such as 1,3-propanediol (PDO) and 1,4-butanediol, but also for water remediation allows the diminution of the use of non renewable energy such as fuel. The objective can be to build polymers, generate biogas and biodiesel in a sustainable way, or even to do decontamination without an extensive use of chemicals that may be released in the environment. Finally, biopharmaceuticals can be produced using engineered organisms.

Several successes have been reported, using single species fermentors and processes. However, using a single organism can be restrictive in several ways. Indeed, an organism can be genetically modified only to a certain extent. It can be more interesting to divide the insertion of long novel pathways among different organisms or strains.

Furthermore, taking advantage of different metabolic capabilities can allow to use different kinds of substrates (that can be of less cost, such as cellulose, lignin, hemicellulose, etc.) but also, for example, an organism can produce the targeted compounds but this production can be associated to the creation of a co-product that has a concomitant inhibitory effect. Hence, the association with an organism able to utilise and neutralise this co-product is of great interest. Examples of such associations have been described by (Sabra et al., 2010; Bizukojc et al., 2010).

Dividing the work between different microorganisms has therefore represented a novel direction, requiring some new understanding of the organisation of communities.

As discussed by (Jagmann and Philipp, 2014), several parameters must be taken into account to obtain a synthetic community with a sustainable production and growth. It is necessary to create (this does not exist naturally) a beneficial interaction between the organisms of the community (mutualism or syntrophy) to ensure the success of the synthetic community. Mutualism, where both species profit from the association, seems the more adapted interaction. If needed, mutualism can be enforced by genetic engineering, for example by creating auxotrophic strains; this will force a cross-feeding between the organisms, regulating the growth of the species composing the co-culture (Shou et al., 2007; Hosoda et al., 2011).

Several case-studies have been modelled but they are generally restricted to a core-model

and require extensive knowledge of the consortium and the effect of perturbations on it (see (Mahadevan and Henson, 2012) and (Henson and Hanly, 2014) for reviews).

(Eng and Borenstein, 2016) proposed a generic topological method using genome-scale models to produce specific target compounds using preselected sources of carbon. This method uses an integer linear programming (ILP) solver to propose the minimum number of species such that the target can be produced from the sources. It is already a hard problem and work remains to be done to take into account stoichiometry in this procedure. The problem presented in Chapter 4 is related to this one, however we do not attempt to minimise the number of species but instead the weight of the global solutions.

Most of the methods proposed for single organisms can therefore be adapted for multiple organisms. However, as for simple organisms, predictions do not always translate to reality, probably because modellers missed some interactions and side-effects. Nevertheless, modelling is a way of summarising the current state of knowledge, and by an iterative process it allows to know increasingly more about the microorganisms and their possible usage in biology.

Chapter 2

Understanding metabolic shifts: a qualitative analysis

Contents

2.1 Introduction

Metabolomics studies the metabolome (the set of small molecules found in a biological sample) using mass spectrometry approaches and/or NMR (Nuclear Magnetic Resonance) analyses. It is now possible to perform either targeted or global (*i.e.* untargeted) experiments to identify the levels of metabolites present in extracellular and intracellular media.

In targeted metabolic experiments, the concentrations of the selected compounds can be measured in comparison with global approaches where all metabolites (detectable and measurable) are considered.

In global approaches, compounds can be identified *a posteriori* through different statistical analyses, molecular networking, dereplication and/or by structural analysis (structural identification after isolation of pure compounds). One such untargeted approach is called metabolic fingerprinting which provides an image of the metabolome of a cell at a certain time. Possible metabolites have then to be identified with pattern detection tools, while multivariate analysis can be used to observe differences between several experiments. Another untargeted approach is metabolic profiling that provides the level of metabolites present in a sample (Shulaev, 2006).

The results of metabolomic experiments can be used to constrain the solutions of Flux Balance Analysis by computing only those which are thermodynamically feasible (Hoppe et al., 2007). Experiments showed, for instance in *Escherichia coli*, that there is a large variability for the concentrations of metabolites in a certain condition. Several models were elaborated in order to infer the reason for this variability in relation to the quantity of enzymes and the occupancy of their binding sites. Indeed, different studies showed that intracellular metabolite pools are not necessarily small (Cornish-Bowden, 1991; Bennett et al., 2009). A current hypothesis is that there is a compromise on the size of the intracellular metabolite pools. On one hand, high concentrations of metabolites allow to maintain a thermodynamic driving force and improve the enzymatic efficiency (Park et al., 2016). On the other hand, the metabolite pools might be small because of the limited intracellular space (osmotic pressure) and to avoid a cross-talk between pathways (Bennett et al., 2009; Tepper et al., 2013).

We are interested in understanding which part of the metabolism of an organism is impacted, and how, by a change in culture condition (for example a stress) or a mutation. Hence, we focus on the results of metabolomic differential analysis that is the identification of quantitative differences in metabolite concentrations between several culture conditions. Using such differences in metabolite pools, we propose to extract information on the transition between two measured conditions. We call *metabolic shift* the transition from one condition to another.

In formal terms, this translates into a problem of subnetwork (sub-(hyper)graph) extraction. The input is a whole metabolic network and a list of metabolites (vertices of the network) whose concentration changed between two conditions, while the output should provide an explanation that involves such vertices in an ordered manner through reactions (*i.e.* arcs or hyperarcs of the network). Such reactions should correspond to those that played a role in the metabolic shift.

Methods of extraction of subnetworks associated to discriminating metabolites using known pathway databases such as *KEGG* or *MetaCyc* (Leader et al., 2011) are available. Those approaches are interesting and allow for a first identification, however they are biased towards the current biological knowledge of annotated organisms, which can limit the discovery of novel interesting metabolic roads. (Faust et al., 2011) use a bipartite graph and several heuristics for the prediction of metabolic pathways that connect a set of elements of interest.

In our work, we tried to address two questions, which are:

1. What is the subnetwork impacted by a stress or by environmental changes?
2. How did the system (*i.e.* the metabolism of the organism) reorganise itself following such stress or change?

In this chapter, the input is the set of metabolites identified and their observed relative change between two conditions, as well as the genome-scale network of the studied organism.

We first present the dataset used throughout this chapter. We then explore the previous method developed in the team by (Acuña et al., 2012a; Milreu et al., 2014) and its associated software, GOBBOLINO & TOUCHÉ. This first modelling performed well, but was not fully satisfying (see Paragraph 2.3.2). We then proposed to move to a new formulation, in particular as shown in Section 2.4, and present the results obtained for this new method in Section 2.5.

2.2 *Saccharomyces cerevisiae* exposed to cadmium

Yeast (*Saccharomyces cerevisiae*) in the presence of cadmium (Cd^{2+}) experiences an oxydative stress. The metal is partially absorbed and accumulated into the vacuole (Volesky et al., 1993). This biosorption of metals into the cell is possible through classical transport systems that are used to ingest heavy metals with biological function (such as cobalt, iron, etc. (Mendoza-Cózatl et al., 2005)).

Glutathione is present in large quantities in cells creating a reducing environment. It can bind to free heavy metals (through a process called chelation) to keep their intracellular concentrations low. It thus acts as a defence mechanism against cadmium.

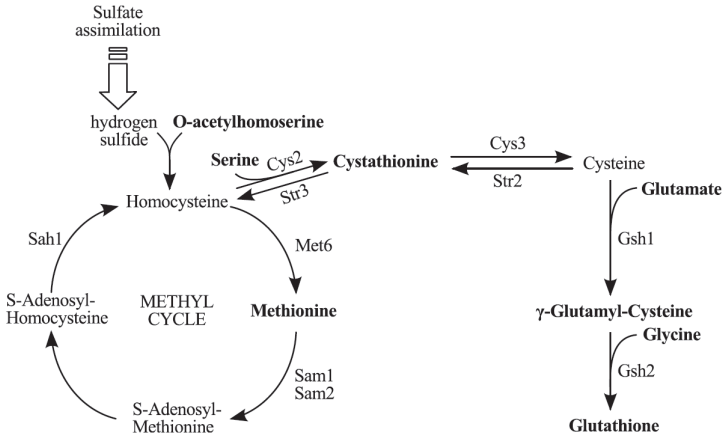


Figure 2.1: From (Milreu et al., 2014), the known pathway for cadmium detoxification. Compounds in bold are the discriminating metabolites measured by (Madalinski et al., 2008) and discussed later on. Glutathione binds to cadmium to reduce the metal intracellular concentration.

Glutathione is the end point of the known pathway of cadmium detoxification (presented in Figure 2.1). There is a redirection of the fluxes from the methyl cycle toward glutathione to capture more cadmium. Hence, in the presence of cadmium, the glutathione synthesis and the

sulfur amino acid biosynthesis pathways are induced. The second one can be seen as a consequence since sulfur is required to obtain glutathione. (Vido et al., 2001) studied the proteome of different *Saccharomyces cerevisiae* strains and noticed such correlation, that is the stimulation of the glutathione and cysteine synthesis in the presence of cadmium. (Wysocki and Tamás, 2010) presented an overall review of the mechanisms of action of *Saccharomyces cerevisiae* in the presence of metals, including sensing and signalling processes that will not be discussed here.

(Madalinski et al., 2008) studied the impact of cadmium on the metabolite concentrations of a Yeast extract (*Saccharomyces cerevisiae* S228c) and we adopted their experimental data. Using both a targeted and a global approach, the authors identified 21 discriminating compounds (actually between 21 and 24 due to identification issues), that is, metabolites whose concentration was detected as different between the cadmium and the standard conditions. These compounds are shown in Table 2.1.

Metabolite ID	Intensity ratio	Present in the pathway
arginine	1.9	no
reduced glutathione	33.9	yes
O-acetylhomoserine (*) and/or 2-aminoadipate (*)	0.5	yes / no
nicotinamide (*) and/or pyridine-3-aldoxime (*)	4.8	no
pyrroline-hydroxy-carboxylate	0.7 [a]	no
methionine	0.3	yes
citrulline (*)	0.7	no
threonine (*) and/or homoserine (*)	0.6	no
glutamine	0.7	no
glutamate	0.8	yes
glutamylcysteine	192.2	yes
5-methylthioadenosine	11	no
serine	0.2	yes
glycine (*)	0.3	yes
cystathionine	50.5	yes
lysine	0.7	no
cysteinylglycine (*)	35.9	no
leucine/isoleucine	1.2	no
tyrosine	2.9	no
histidine	1.2	no
alanine	0.8	no

Table 2.1: Measured metabolites in the cadmium example (adapted from (Milreu, 2012; Madalinski et al., 2008)). Stars (*) indicate metabolites whose identification needs confirmation. [a]: non significant in (Madalinski et al., 2008) analysis. The intensity ratio represents the ratio from the cadmium condition to the control. Metabolites present in the detoxification pathway (Figure 2.1) are indicated by (yes) in the last column.

The Yeast network that is used throughout this chapter was obtained by (Milreu et al., 2014) from *MetExplore* (Cottret et al., 2010b). Reactions involving pairs of cofactors were split, and the cell compartments and ubiquitous compounds removed (the latter are: *water, carbon dioxide, phosphate, diphosphate, ammonia, proton, hydrogen peroxide and oxygen*). Removal of the ubiquitous compounds is a common step in the topological analysis of metabolic networks. Indeed, they are involved in a large number of reactions but do not transfer carbon. Deleting

such metabolites will therefore avoid solutions that are not biologically reasonable. By removing them, we make the implicit hypothesis that those metabolites are present in sufficient quantity, and that they are not important in the metabolic shift in terms of carbon transfer (as their other actions cannot be depicted in this topological framework).

In our case, as in (Milreu et al., 2014), we first tested a small dataset composed of the eight discriminating compounds involved in the known detoxification pathway. We then tested the dataset containing all the identified discriminating metabolites.

In (Milreu et al., 2014), 2 of the compounds reported by (Madalinski et al., 2008) were not used. Indeed pyrroline-hydroxy-carboxylate is not included in the larger dataset of discriminating compounds. It is disconnected in the network (see Figure 2.2) and did not change significantly between the two conditions.

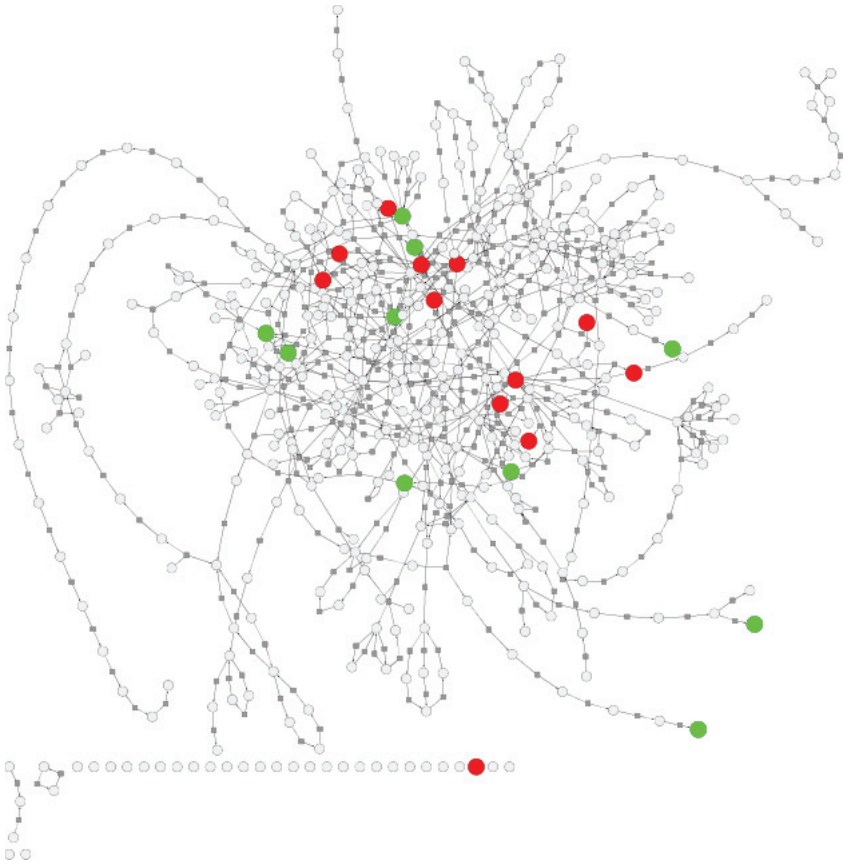


Figure 2.2: Yeast network obtained from *MetExplore* (Cottret et al., 2010b). In red, the metabolites with a decreased concentration in the cadmium exposure condition. In green, the ones with an increased concentration in the cadmium exposure condition. Bottom left: disconnected compounds and reactions, including pyrroline-hydroxy-carboxylate. The squares represent the reactions and the circles represent the metabolites.

Moreover, pyridine-3-aldoxime was not found in the filtered network. It is not annotated in classical Yeast databases (either *YMDB* which is the Yeast metabolome database (Jewison et al., 2012) or the *Saccharomyces genome database* (Issel-Tarver et al., 2002)) nor in the last Yeast genome-scale model available (model version *yeast_ 7.6* (Aung et al., 2013) from July 2016). (Madalinski et al., 2008) reported that in the experiments performed, the identification of pyridine-3-aldoxime was not certain. The authors did not know whether it was both nicotinamide and pyridine-3-aldoxime that were measured or only one of those. There is therefore the

possibility that it was in fact only nicotinamide. This is the hypothesis that we followed here. For this reason, pyridine-3-aldoxime was not included either in the larger dataset of discriminating compounds.

We thus considered two datasets. The first one is composed of 8 metabolites whose concentration changed significantly between the cadmium and the control condition. The second one contained 22 compounds in (Milreu et al., 2014) and as discussed later on, 21 compounds when we applied it to the new method (see Section 2.5).

However, we first present the previous method developed by (Acuña et al., 2012a; Milreu et al., 2014) and discuss the results obtained.

2.3 Previous model and methodology: Gobbolino & Touché

The problem of finding a subnetwork impacted by a change of conditions has been described in (Acuña et al., 2012a; Milreu et al., 2014).

Gobbolino & Touché use a directed graph which represents a compound graph. An arc links a substrate to a product of a given reaction. The software output *stories* (see Definition 5) that are directed acyclic graphs (DAGs) which connect discriminating metabolites (*i.e.* those whose concentration changed between the two conditions). Such stories explain where matter that was lost (in the case of a decreased concentration) went to or from where matter that was gained (in the case of an increased concentration) arrived.

2.3.1 Definitions

In (Milreu et al., 2014), the algorithm uses a compound graph and discriminates two types of compounds, and hence of vertices. One are the vertices called *black* that correspond to the metabolites whose concentration changed significantly and the other to the vertices called *white* which are the metabolites that were not measured or whose concentration did not change.

The authors use the definitions recalled below (Definition 4 and Definition 5).

Definition 4. *Let $\mathcal{G} = (\mathbb{B} \cup \mathbb{W}, \mathcal{A})$ be a directed graph such that $\mathbb{B} \cap \mathbb{W} = \emptyset$. Vertices in $\mathbb{B}$ are said to be black and correspond to the discriminating compounds while those in $\mathbb{W}$ are said to be white vertices.*

Definition 5. ***A metabolic story*** *of $\mathcal{G}$ is a* ***maximal acyclic*** *subgraph $\mathcal{G}' = (\mathbb{B} \cup \mathbb{W}', \mathcal{A}')$ of $\mathcal{G}$ with $\mathbb{W}' \subseteq \mathbb{W}$ and $\mathcal{A}' \subseteq \mathcal{A}$ and such that for each vertex $w \in \mathbb{W}'$, w is not a source nor a target.*

A white vertex that is not a source nor a target means that it has at least one incoming arc ($d^- > 0$) and one outgoing arc ($d^+ > 0$). It thus can only be an intermediate and never the starting point (source) or the end-point (target) of the sub-graph given as solution. Indeed, a white vertex is not a discriminating compound. Not all white compounds are measured but the assumption is that their concentrations are the same in the two conditions. If a white metabolite is a source, it would mean that only outgoing reactions changed in the transient state, but then the metabolite concentration should have changed accordingly. The same reasoning applies for target metabolites.

In this definition, the acyclicity constraint is not based on any biological *a priori*. It is known that some cycles are important in metabolism. This constraint was used to prevent solutions with only regenerating cycles. Indeed, the latter are cycles such that their vertices are consumed and regenerated at the same time (as in Figure 2.3a).

(a) Example of a simple regenerating cycle. (b) The two metabolic stories.

Figure 2.3: Example of a regenerating cycle containing two black vertices (A and C) in Figure 2.3a. Two metabolic stories can be retrieved and are presented in Figure 2.3b. One goes from A to C through B (dotted arcs), the other goes from C to A through D (dashed arc).

Gobbolino & Touché look for an explanation where the flux of matter goes from one metabolite to another, thus there is a need to have a least two metabolites not part of a regenerating cycle, *i.e.* a source and a sink according to Definition 5. This is presented in Figure 2.3. This acyclicity constraint allowed to obtain those two metabolites and also made the problem more tractable. One may however observe that, although it is useful to avoid cycles that do not participate in moving matter, there is no indication that all cycles should be removed. This was part of the motivation for the new approaches that we proposed and that will be presented later in Section 2.4 and in Chapter 3. For now, let us get back to the Gobbolino & Touché model.

The compound graph is obtained from the metabolic network. Reversible reactions are added as separate reactions. Finally, there are no parallel arcs, meaning that if two reactions share a same substrate and a same product, only one arc would link those. Once the compound graph is created, it is *pre-processed*. This is necessary to be able to enumerate the metabolic stories in a genome-scale network. The pre-processing involves a lossless compression composed of four steps that are such that no solution is lost between the uncompressed and the compressed graph. Then a final step uses the lightest paths between the black vertices. The weight of a path is the sum of the degrees of the vertices that are part of it. This approach has also been discussed by (Faust et al., 2011) and is illustrated in Figure 2.4.

The resulting network is composed of the union of all the lightest paths between all pairs of black vertices. The idea of using vertex degree is motivated by the desire to avoid cofactors such as ATP in the paths since they are not involved in the carbon transfer of the reactions. This step performs well to reduce the network size. However, since only the lightest paths are conserved, interesting ones are lost because of intermediates that have a large degree in the network but are not considered as cofactors and can play a key role in the metabolism. A common example of such metabolites is pyruvate which is at the junction of many metabolic pathways.

The network compression can greatly reduce its size. For example in Figure 2.5, the original network of Yeast (*Saccharomyces cerevisiae s288c*) used by (Milreu et al., 2014) contains 600 metabolites and 949 arcs (Fig. 2.5a). The compression steps reduce it to a network containing 10 vertices and 25 arcs, that is, more than 97% of the arcs were removed (Fig. 2.5b).

The complexity of enumerating all metabolic stories of a given input $\mathcal{G} = (\mathbb{B} \cup \mathbb{W}, \mathcal{A})$ remains an open problem. Finding one story is polynomial, but the algorithms that have been proposed

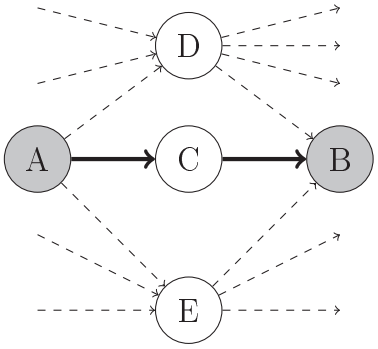


Figure 2.4: Example of the lightest path compression. Here, there are three paths linking the vertices A and B. D has a degree of 7, whereas E has a degree of 6 and C has a total degree of 2. The lightest path linking the vertices A and B is thus the one going through C indicated in bold. After the lightest path compression, there will be only 2 arcs remaining, A $\rightarrow$ C $\rightarrow$ B.

and successfully applied to enumerate all stories are of unknown complexity (Acuña et al., 2012a; Borassi et al., 2013). In (Milreu et al., 2014), the problem formulation was used to infer pathway activations when Yeast (*Saccharomyces cerevisiae*) is exposed to a heavy metal, namely cadmium.

In the next section, I present the solutions obtained and how we proposed to improve the current modelling.

2.3.2 Metabolic stories in Yeast exposed to Cadmium

In (Madalinski et al., 2008), 24 metabolites are measured and identified as changing significantly between the two conditions of growth, during cadmium exposure and the control. Out of those 24 metabolites, eight are known to be involved in the cadmium detoxification pathway. A summary of the measurements obtained can be found in Table 2.1.

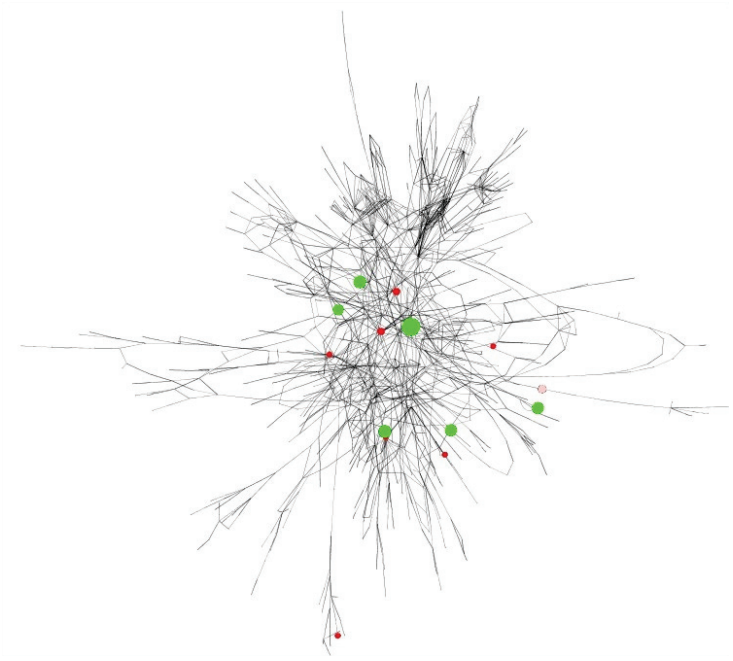
The definition previously proposed led to interesting answers, however it was already computationally expensive. After the preprocessing step in (Milreu et al., 2014), from the yeast network with 8 discriminating metabolites, 222 metabolic stories were enumerated. If the initial 22 discriminating metabolites are taken into account, the number of stories explodes, creating almost 4.10^6 stories.

The main reason for the high number of stories is the acyclicity constraint. Since the solutions must be acyclic but also maximal, there is a combinatorial explosion of their number. Indeed, if we consider as an example a graph containing one cycle, it is necessary to create as many solutions as arcs composing it, since the solutions will be obtained by removing one arc at a time. Thus, in graphs containing more than one cycle, the number of solutions is increasing drastically.

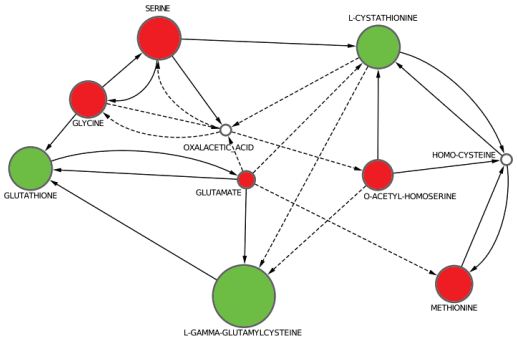
For example, in Figure 2.6b, there are 4 arcs only, but it is possible to extract 4 directed acyclic graphs (see Figure 2.6c).

Furthermore, in the case of reversible reactions, the two directions are inserted, creating even more cycles to break.

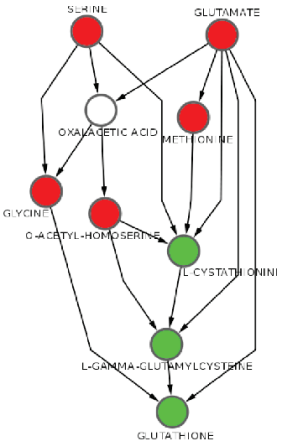
Another reason for the combinatorial explosion is the use of the compound graph representation, which artificially increases the number of arcs. Indeed, if the network is modelled as a compound graph, the joint consumption of substrates or the coproduction of metabolites cannot be modelled as shown in Figure 2.6.



(a) Original network.



(b) Compressed network; it contains 222 stories.

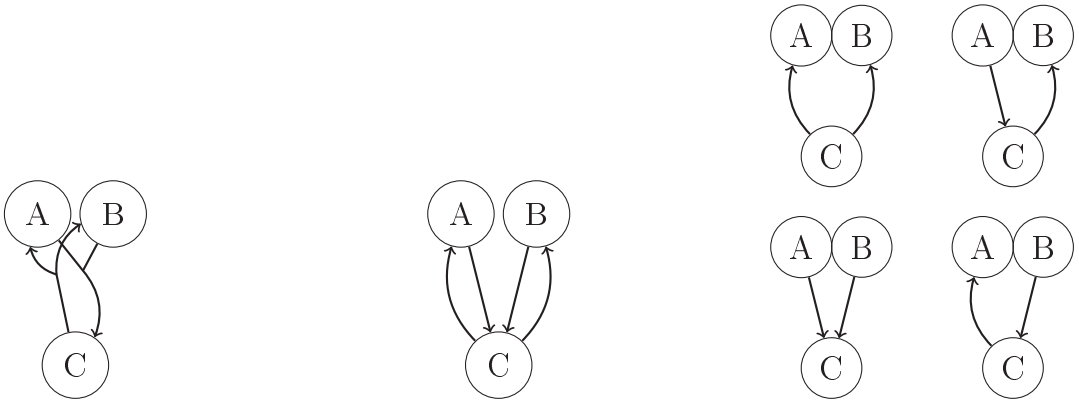


(c) A story

Figure 2.5: Example of a story in the network used in (Milreu et al., 2014). Red vertices are discriminating compounds with a decreased concentration after cadmium exposure, whereas green vertices have an increased concentration.

In Figure 2.6b, the fact that both A and B are needed to produce C is not represented anymore. In the extracted acyclic subgraphs, it is then possible to pass from A to B through C, without acknowledging that such paths are using the reaction $A + B \leftrightarrow C$, as shown in Figure 2.6c.

The notion of reaction is therefore lost in the compound graph representation. We next look more in detail at the solutions obtained by GOBBOLINO & TOUCHÉ to see how to propose a new modelling that might be more realistic in terms of biological interpretation (in particular as concerns the acyclicity constraint) and possibly also more tractable.



(a) Representation as a directed hypergraph.

(b) Representation as a compound graph.

(c) The remaining DAGs.

Figure 2.6: Reaction $A + B \leftrightarrow C$; in the compound graph, this reversible reaction creates 4 arcs, and it is possible to extract 4 directed acyclic graphs (DAGs).

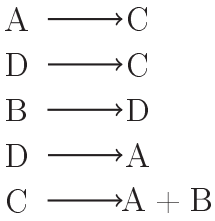
Post-processing the stories

The question here is whether the definition of stories used in (Milreu et al., 2014) provides biologically meaningful solutions. It is thus important to understand the side-effects of the modelling choice (namely, the use of a compound graph to represent the metabolic network and the chosen definition of a metabolic story) as this will enable us to establish whether they need to be bypassed, or they in fact do not represent a problem. A directed hypergraph seems to provide a better representation for a metabolic network. Studying the solutions given by Gobbolino & Touché is also a first step to improve the modelling of the metabolic shifts by placing the obtained solutions in the metabolic network modelled as a directed hypergraph. Moreover, as mentioned previously, the acyclicity constraint is not based on any biological *a priori*. We therefore wanted a new modelling approach that does not forbid cycles, but rather proposes solutions based on a biological interpretation of the discriminating compounds and selected reactions.

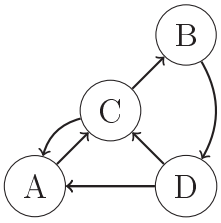
We define a **valid** story as one in which a reversible reaction is not used in both directions (forward and backward). Indeed, having a reaction used in both ways can only happen in cases such as the one presented in Figure 2.6. Since the obtained solutions must be acyclic, using a reaction and its reverse is necessarily the result of a path where we go from one substrate (resp. product) to another substrate (resp. product). It therefore is an artefact of the compound graph representation. Having both directions is also misleading. We indeed recall that we want to extract a subgraph such that the reactions were involved in a metabolic shift. This would mean that at a given point, the flux going through this reaction played a role in the metabolite concentrations. Having one reaction and its reverse is therefore redundant (since the increased flux in one direction can be interpreted as a decreased flux in the other direction).

Furthermore, for every reaction in a solution, we add the possible arcs that were discarded in the extraction of the solution in the compound graph. This means that if an arc of the solution was part of a reaction with several products or several substrates, its co-substrates and co-productsmust be added to the solution with the corresponding arcs. We wanted here to observe a more complete representation of the extracted solutions, in particular whether the joint consumption or production of metabolites are required for the solutions to be topologically valid.

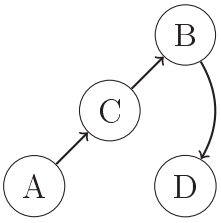
The acyclicity of the solution in this extended compound graph is then tested (see Figure 2.7). In this case, an important number of solutions presented a cycle as shown in Table 2.2.



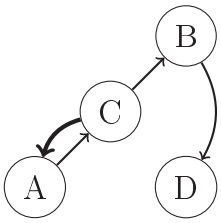
(a) Reactions and metabolites.



(b) The compound graph.



(c) A possible extracted story.



(d) The story is completed by adding the arc from C to A (in bold) as the original reaction of the network was C → A+B.

Figure 2.7: Procedure for verifying the acyclicity of the solution in the compound graph for the network in Figure 2.7a and represented as a compound graph in Figure 2.7b. One of the possible stories is shown in Figure 2.7c. The arc going from C to B is part of a reaction that also produces A. When the new arc C → A is added, the solution is no longer acyclic (Figure 2.7d).

Datasets	# Stories	# Valid Hyperstories	# Acyclic Hyperstories
8 Black vertices	222	12	1
22 Black vertices	3 934 160	43 360	193

Table 2.2: Number of stories found for the two black vertices datasets of (Milreu et al., 2014); one is composed of only black vertices (8) involved in the cadmium detoxification pathway and the other contains the list of metabolites identified by (Madalinski et al., 2008). We remind that a valid hyperstory is a solution not using reversible reactions in both directions and that an acyclic hyperstory is a solution that remains acyclic after retrieving for the reactions all their arcs.

Using the compound graph was necessary in order to have a first idea of how to interpret the differences in metabolite concentrations between two steady-states. The problem of the metabolic stories generates a large number of solutions and requires important computation power when applied to genome-scale networks.

As previously mentioned, it seems important from a modelling point of view to replace the acyclicity constraint by one that is biologically more meaningful. Furthermore, it is essential to use a more faithful network representation such as a directed hypergraph representation. We now discuss the modelling approach developed in the case where we work with a directed hypergraph.

2.4 From metabolic stories to metabolic hyperstories – The TOTORO method

We propose TOTORO (TOpological analysis of Transient metabOlic RespOnse), a new method to decipher metabolic changes between two metabolic equilibria. This method is based on a hypergraph representation of a metabolic network and relies on the direction of variation of the metabolite pools. Indeed, with TOTORO, we discriminate the metabolites with an increased or decreased concentration between two conditions as will be explained later on. We first present the definition of the solutions that we called *metabolic hyperstories*, and then the current implementation of the enumeration method applicable to genome-scale model networks.

2.4.1 Definitions

The definition of a *metabolic hyperstory* relies on the same input data as the previous method, but uses another representation of the metabolic network: a directed hypergraph. Furthermore, as explained later on, we propose a novel interpretation of the results and new constraints for the search of solutions.

We are assuming that we can correctly identify metabolites, measure them, and infer if their concentrations changed significantly between two conditions and in which direction (increased or decreased). As in GOBBOLINO & TOUCHÉ, the idea is to output the reactions that are part of the process that created or followed the metabolic shift, that is those that participated in the change of concentrations and those that are responsible for the transfer of matter because of an increased or a decreased flux. We therefore propose to extract a sub-hypergraph composed of the reactions involved in the transient state, that represents the transition that occurred between the two measured conditions. We call attention to the fact that the reactions that are not selected as being part of a metabolic hyperstory may still have a positive flux in both steady-states (and such flux can even differ between the two states). It is just assumed that these reactions are not "responsible" for the changes in the metabolite pools; they are not the reactions that changed to adapt to the new situation. If such fluxes are of interest, it is possible to compute them using flux balance analysis (FBA). The differences in the metabolite concentrations should be explained by the transfer of atoms between pools that decreased or increased. In itself, it is not the transfer *per se* but more the differences and desynchronisation of the fluxes that led to such pools. Indeed, we work under the hypothesis that the overall cell capacity will remain the same between the two conditions.

A metabolite may play different roles in two metabolic hyperstories. Without any *a priori*

knowledge, one cannot know, for example, if a pool increase is due to the fact that a metabolite is at the end-point of an activated pathway, or is the source of an inhibited pathway, or even is an intermediary. The solution should also include signs for the reactions indicating whether their contribution represents an excess of flux (meaning that an increased matter went through the reaction compared to its neighbours), or the opposite (a decrease of flux). We will use the term activation (represented by $+$) and inhibition (represented by $-$) although we do not infer if a true regulation has occurred or what type of regulation was used (*i.e.* allosteric or transcriptomic).

In GOBBOLINO & TOUCHÉ, there are two types of metabolites. The white vertices ($\mathbb{W}$) that were not measured or did not change, and the black ones ($\mathbb{B}$) that varied between the two conditions. We propose here to divide the black vertices into two categories:

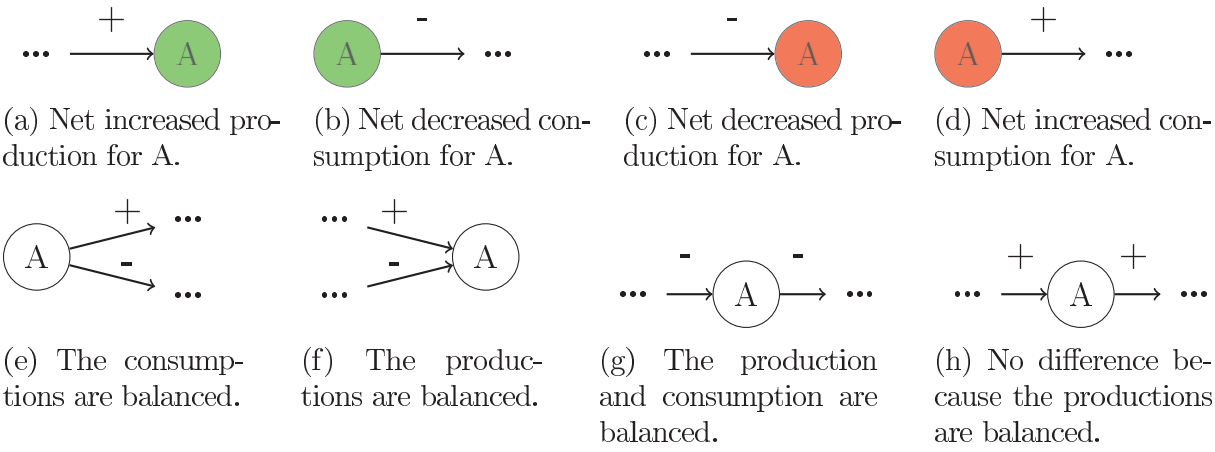
- *Red*: corresponding to a black vertex whose concentration decreased between two conditions;
- *Green*: corresponding to a black vertex whose concentration increased between two conditions.

Observe that $\mathbb{B} = Red \cup Green$ and $Red \cap Green = \emptyset$

A green vertex implies an increased concentration, that is either a lower consumption or a bigger production during the transient state (*i.e.* the time interval between the two measured states). Indeed, the differences of concentration must result from a change in the reaction rates between the two states. Moreover, such differences in consumption or production must have an impact on the other substrates and products. Following a cascade of flux changes that are expected because of the imbalance created, one of the reactions selected as changed must therefore lead to a red vertex with a decreased (resp. increased) production (resp. consumption) during the transient state.

It is possible to infer simple rules and to define what we call *vertex coherence*. This corresponds to the fact that the sign of the hyperarcs around a vertex can explain its colour. Such rules define a minimal set of hyperarcs that need to be part of the solution to make the vertices coherent.

- A white vertex is coherent:
  - if it has two incoming (or outgoing) hyperarcs of opposite sign;
  - *or* if it has one incoming and one outgoing hyperarc of the same sign.
- A green vertex is coherent:
  - if it has an incoming hyperarc of sign $+$;
  - *or* if it has an outgoing hyperarc of sign **-**.
- A red vertex is coherent:
  - if it has an incoming hyperarc of sign **-**;
  - *or* if it has an outgoing hyperarc of sign $+$.



(a) Net increased production for A.

(b) Net decreased consumption for A.

(c) Net decreased production for A.

(d) Net increased consumption for A.

(e) The consumptions are balanced.

(f) The productions are balanced.

(g) The production and consumption are balanced.

(h) No difference because the productions are balanced.

Figure 2.8: Minimal explanation for the colour of the vertices. In (a) and (b), A is green (increased concentration) while in (c) and (d) A is red (decreased concentration). In the remaining examples, A is white, *i.e.* the concentration has not changed.

Such rules are illustrated in Figure 2.8.

We require all our vertices to be *coherent* and linked together by reactions carrying a correct sign. We will use here the definition of hyperpath introduced by (Gallo et al., 1993) (see Chapter 1, Paragraph 1.2.1). This is recalled in Definition 6.

Definition 6. *A hyperpath $\mathcal{P}_{st}$, of length q in the hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{A})$ is a sequence of vertices and hyperarcs $\mathcal{P}_{st} = (v_1 = s, a_{i_1}, v_2, a_{i_2}, ..., a_{i_q}, v_{q+1} = t)$, where, for $j = 2, ..., q$, $s \in tail(a_{i_1}), t \in head(a_{i_q})$, and $v_j \in head(a_{i_{j-1}}) \cap tail(a_{i_j})$.*

By taking into account the fact that a solution would then be the reactions participating in the pool differences with their (+/-) signs, we can propose a new definition of metabolic hyperstories in a directed hypergraph.

Definition 7. *Let $\mathcal{H} = \mathcal{G}(\mathcal{V}, \mathcal{A})$ be a directed hypergraph with $\mathcal{V} = \mathbb{B} \cup \mathbb{W}$ and $\mathbb{B} \cap \mathbb{W} = \emptyset$.*
A metabolic hyperstory *is the extracted sub-hypergraph $\mathcal{H}' = \mathcal{G}'(\mathcal{V}', \mathcal{A}')$ and the sign function associated to $\mathcal{A}'$ such that $sign : \mathcal{A}' \mapsto \{-, +\}$, with $\mathcal{A}' \subseteq \mathcal{A}$, $V' = \mathbb{B} \cup \mathbb{W}'$ for $\mathbb{W}' \subseteq \mathbb{W}$, and all vertices are coherent.*

We can see than this definition implies that:

(a) a black vertex v can be a source (that is, $d^-(v) = 0$) or a target (that is, $d^+(v) = 0$);

(b) a white vertex can be a source (resp. a target) if there are at least two reactions outgoing (resp. incoming) of opposite sign;

(c) there is not always a path $\mathcal{P}_{s,t}$ between $s \in Red$ (resp. $s \in Green$) and $t \in green$ (resp. $t \in Red$);

(d) nevertheless, for every red (resp. green) vertex, there must be at least one green vertex (resp. red) that is part of the same weakly connected component; indeed the matter that accumulated during the transient state in a green vertex should be correlated to the one lost by the red vertices.

The last three points are illustrated in Figure 2.9a.

2.4.2 Computing the metabolic hyperstories

Hypergraph transformation

The rules depicted in Figure 2.8 are symmetrical. Indeed, if the reactions are reversible, taking the forward direction with the sign + (resp. −) is identical in terms of coherence to taking the reverse one with the sign − (resp. +).

The hypergraph is preprocessed such that for every reaction, we insert its reverse.

Definition 8. *Let $\mathcal{H}_t = \mathcal{G}(V, \mathcal{A}_t)$ be the transformed directed hypergraph from $\mathcal{H} = \mathcal{G}(V, \mathcal{A})$ such that $\forall a \in \mathcal{A}$, $a \in \mathcal{A}_t$ and $\bar{a} \in \mathcal{A}_t$ with $\bar{a}$ such that $head(\bar{a}) = tail(a)$ and $tail(\bar{a}) = head(a)$*

We can now define the solutions of our problem using the hyperpath definition (Definition 6) and some additional constraints.

Definition 9. *Let $\mathcal{H}_t = \mathcal{G}(V, \mathcal{A}_t)$ be a transformed directed hypergraph with $V = \mathbb{B} \cup \mathbb{W}$ and $\mathbb{B} \cap \mathbb{W} = \emptyset$*
A hyperstory will be a set $\mathcal{A}' \subseteq \mathcal{A}_t$ with $V' = \mathbb{B} \cup \mathbb{W}'$ for $\mathbb{W}' \subseteq \mathbb{W}$ such that:

(a) it is not possible to take a hyperarc and its reverse in the solution;

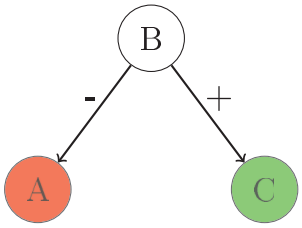
(b) green vertices (in Green) cannot be sources;

(c) red vertices (in Red) cannot be sinks;

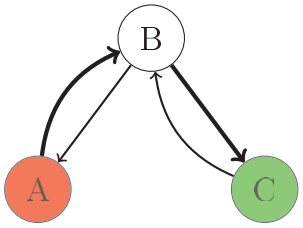
(d) white vertices (in White) cannot be sinks nor sources;

(e) for all $a \in \mathcal{A}'$, we have that:

- *$\forall x \in tail(a)$, $\exists v \in Red$ such that $\exists \mathcal{P}_{v,x}$;*
- *$\forall y \in head(a)$, $\exists v \in Green$ such that $\exists \mathcal{P}_{y,v}$.*



(a) Here, this directed hypergraph meets Definition 7, but there is not a direct hyperpath between the green vertex (C) and the red vertex (A). B is a source but it is *coherent*.



(b) Here, the directed hypergraph of Figure 2.9a was transformed. In bold, the solution meets Definition 9. B is not a source not a sink whereas A is a source and C is a sink.

Figure 2.9: The solutions of Definition 7 in the directed hypergraph can be retrieved from the solutions of Definition 9 that use the transformed directed hypergraph.

As for the black vertices, we propose to divide the white vertices (in $\mathbb{W}$) into two categories:

- *White*: a vertex which is "truly" white, meaning that its concentration did not change between the two conditions;

- *Grey*: a vertex whose concentration was not measured between two conditions.

We then have that $\mathbb{W} = White \cup Grey$ and $White \cap Grey = \emptyset$.

Indeed, we rarely have information on all the metabolites of the genome-scale network. In the case of (Madalinski et al., 2008) for example, there are 24 metabolites reported with a change in concentration for a network of 600 metabolites (and 949 arcs). Only 4% of the metabolites are black whereas the remaining are grey.

It is thus reasonable to relax the constraint (d) in Definition 9 to allow grey vertices to be sources or targets. In this case, grey vertices in the solution do not have to respect the same coherence as white vertices. If they are not coherent, we call them *isolated*.

Definition 10. *A grey vertex v is* isolated *in the hypergraph $\mathcal{H}_t = \mathcal{G}(V, \mathcal{A}_t)$ if it is a source ($d^-(v) = 0$) or a sink ($d^+(v) = 0$)*

Here, we do not have information on the concentrations of the grey vertices. Since we do not have the full information, while staying focused on the vertices we do know, we allow grey vertices as sources or targets if the reaction in which they occur can be part of the solution. Doing so will provide hints on the metabolites that can be measured in the experiments. We will then have the possibility of either considering non measured vertices as "true" white vertices, that can only be intermediary, or to allow grey vertices.

In Definition 9, we gave the conditions that a subnetwork has to verify to be a solution. We will further on consider two cases of optimisation. These are: the search for minimum solutions, or the search for minimal solutions.

The motivation for enumerating minimal solutions is that we infer a local action on the metabolism. Indeed, for an organism, it is less costly in terms of energy to act on a smaller number of reactions than on the complete metabolism. The minimality of the solutions thus implies that we expect the action to be localised but not necessarily on shortest hyperpaths (as it is possible to have one control action that leads to a cascade of effects on the reaction fluxes).

Moreover, this optimisation choice is coherent with a currently used hypothesis (such as adopted in the MOMA formulation) which is that an organism will not fully re-route its metabolism, but instead minimise the changes in its flux distribution compared to the normal growth condition. The minimality requirement results in a large number of solutions (shown in Table 2.3) as we chose to avoid to compress the network using the lightest path compression method since this can lead to lose hyperpaths that do not necessarily involve cofactors.

To overcome this enumeration issue, we propose to directly compute a subset of the minimal solutions, the ones that are also minimum.

Minimum solutions are considered in terms of the number of hyperarcs that are part of the solution. It is possible to restrain even more the set of solutions in the case where grey vertices are taken into account by minimising also the number of isolated grey vertices in the solution.

We can therefore see that the objective is double: inferring the reactions impacted by the change in conditions but also, proposing new experiments in order to measure the uncertain vertices (the isolated grey ones) and thus to narrow down the number of solutions.

Implementation

To enumerate all the solutions, we use *Answer Set Programming* (ASP). *ASP* is a formalism based on disjunctive logic programming that allows to express some problems with a simple

declaration (Bonatti et al., 2010). We therefore can concentrate on formulating the problem in the form of logical rules and constraints. Such formulation uses Prolog, which is a logic programming language. We chose to work with the POSTASCO software and their solver (encompassed in the CLINGO suite (Gebser et al., 2014b; Leuschel and Schrijvers, 2014)).

There is a *preprocessing* step implemented in python to read the SBML file of the metabolic network that will then output a CLINGO compatible format.

SBML is a common exchange format for metabolic networks describing the metabolites and reactions in a XML document (Hucka et al., 2003). After the call to CLINGO, it is possible to post-treat the answers to visualise them using DINGHY (Bulteau et al., 2015), or to retrieve the most common set of reactions encountered in the solutions. This last task can be performed using LCM (Linear time Closed itemset Miner, see (Uno et al., 2005)). Figure 2.10 represents the available workflow.

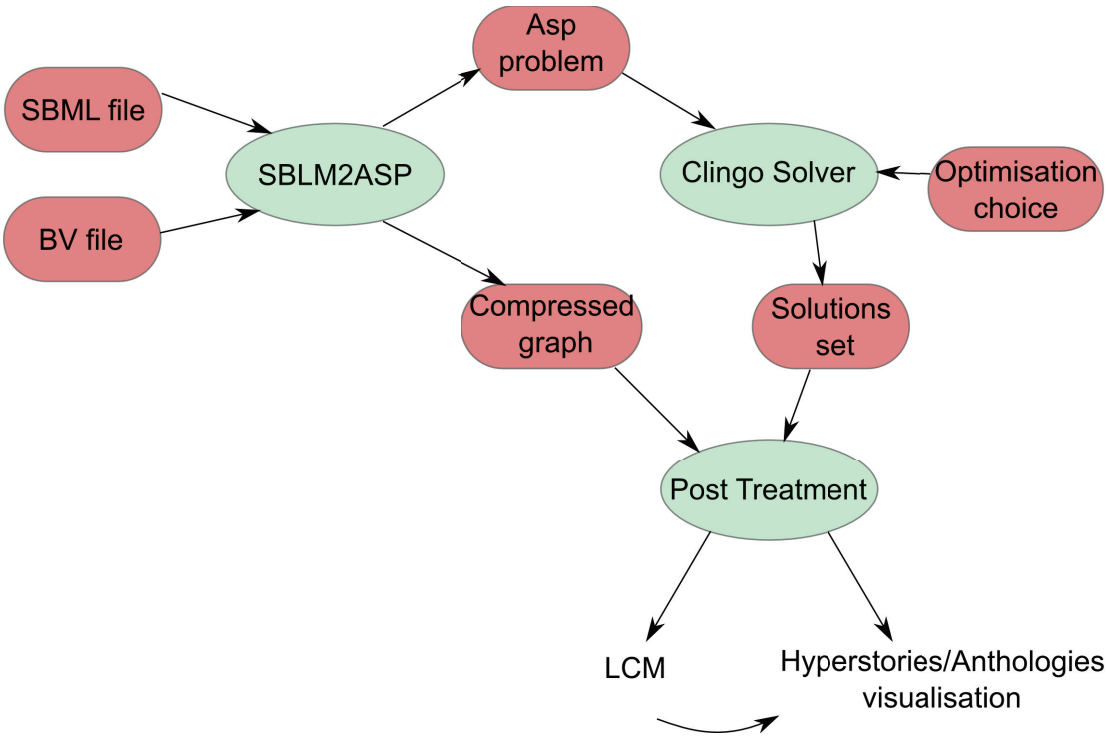


Figure 2.10: Workflow of the method. The preprocessing step (SBML2ASP) implemented in python compresses the network considering the black vertices file (BV file) and outputs the network for the ASP solver CLINGO. This solver requires also as input the optimisation rule(s) that can be applied. The output of CLINGO is then the set of solutions, *i.e.* answer sets than can be post-treated, either for visualisation directly one by one or as groups (creating *anthologies*), or for clustering using Linear time Closed itemset Miner.

Lossless compression. Lossless compression has been used to decrease the size of the input graph but it does not affect the number of solutions. The compression is done iteratively, until no compression is possible anymore.

One important yet trivial compression is the fusion of what we call *twin-hyperarcs*, that are hyperarcs/reactions that share exactly the same substrates and products. They correspond to reactions catalysed by isoenzymes, or they can result from the removal of ubiquitous metabolites and of cofactors if the input network is filtered. Such twin-hyperarcs are merged into one. This compression allows to decrease the number of solutions without losing information since it is

possible to keep the details of the compressed arcs.

A hyperarc $a \in \mathcal{A}$ of the network is removed if one of the following is verified:

- $|tail(a)| = 0$ or $|head(a) = 0|$;
- $\forall v \in head(a), v \in \mathbb{W}$ and $d^+(v) = 0$;
- $\forall v \in tail(a), v \in \mathbb{W}$ and $d^-(v) = 0$.

Disconnected vertices, that is with no incoming nor outgoing hyperarcs, are removed.

These compression steps do not impact the results given as output since the deleted vertices and hyperarcs could not have been part of a solution.

We used the LCM implementation of (Uno et al., 2004) to analyse the results when there is a large number of solutions.

The advantage of LCM is that it is possible to obtain frequent itemsets, that is sets of items (hyperarcs in this case) that are present with at least a certain frequency among the solutions. One can also compute the closed itemsets and the maximal frequent itemsets. An itemset is closed if there is no other itemset that includes it with the same support. A maximal frequent itemset is such that it is not included in another frequent itemset. The space of maximal itemsets is therefore included in the space of the closed itemsets.

Using the definitions and notations of (Uno et al., 2003), we define more formally the *Frequent Closed Itemsets* and the *Frequent Maximal Itemsets*.

The hyperarcs selected in the solutions constitute the set of *items* E from $1, ..., m$. An itemset is a subset X of E. A solution is called a *transaction*, and $\mathcal{T}$ is a set of transactions, *i.e.* of solutions. We denote by $\mathcal{T}(X) = \{t \in \mathcal{T} | X \subseteq t\}$ the set of transactions including X.

Definition 11. *An itemset X is* frequent *if $|\mathcal{T}(X)| \geq \alpha$ for a constant $\alpha \geq 0$*

Definition 12. *A* frequent maximal itemset *is such that it is not included in another frequent itemset.*

Definition 13. *A* closed itemset *is such that $I(\mathcal{T}(X)) = X$ where for a transaction set $S \subseteq \mathcal{T}$, $I(S) = \bigcap_{T \in S} T$*

Identifying frequent itemsets is useful in the case of multiple solutions in order to select the reactions that are jointly required to explain the observed metabolite changes. It allows to select the core reactions that are necessary, or are the most used.

Here, the support, that is the frequency threshold α, is a parameter that has to be determined by the user. We usually set the frequency threshold high (close to 1). The obtained itemsets will then contain reactions that are present in almost all solutions, thus showing the backbone of the organism's response. These will represent mandatory reactions of our solutions, that are required to explain the discriminating metabolites. Furthermore, the itemsets obtained can be helpful to cluster our solutions in order to observe them by groups when an individual analysis is not possible.

Although the LCM implementation performed by (Uno et al., 2004) provides a linear time enumeration of closed patterns, some of our results cannot be treated as we will explain later on.

Visualisation of the solutions. One or several solutions can be visualised afterwards using DINGHY (Dynamic Interactive Navigator for General Hypergraphs in Biology, available at `http:`

//dinghy.gforge.inria.fr/, (Bulteau et al., 2015)). For this, the solutions are retrieved from the output of CLINGO and transformed into a *json* format using a python script.

This post-processing is done in such a way that the colours of the vertices, the direction of the taken hyperarcs and their signs (for vertex coherence) are displayed. We use the information of the original network and of the compression steps.

It is also possible to visualise multiple solutions at once. We re-used the term *anthologies* that had been proposed by (Milreu et al., 2014). In an anthology, multiple solutions are presented. The widths of the hyperarcs representing the reactions are proportional to the frequency of the solutions they appear in.

The developed framework and modelling approach were tested on the same dataset as the one used in (Milreu et al., 2014). The results obtained are presented in the next section.

2.5 Metabolic hyperstories applied to *Saccharomyces cerevisiae* exposed to cadmium

We decided to test the definition of *metabolic hyperstories* on the response of Yeast exposed to cadmium. The measured metabolites were presented previously in Section 2.2, Table 2.1. We first used a subset of the metabolites, the eight compounds known to be involved in the cadmium detoxification pathway. Then, for the large black vertices dataset, some identification issues remained. As aforementioned, for some of the metabolites, (Madalinski et al., 2008) reported that the measurements could be of several metabolites together and that they could not experimentally distinguish them.

This was the case for L-2-aminoadipate. This metabolite is connected to another red metabolite that is Lysine in the network. In the network currently used, there are no reactions such that L-2-aminoadipate can be coherent according to Definition 7. As such, when the latter is included in the dataset, no solution could be found. There can be two explanations here. One is related to an identification issue, meaning that this metabolite was not the one measured and in fact only O-acetylhomoserine was present. The second is that the metabolic network that we use is missing some information. More in particular, it may be missing some reactions necessary to explain the L-2-aminoadipate concentration.

A consequence of this is that we decided to not include L-2-aminoadipate in the datasets.

We therefore considered two datasets of black vertices, one composed of 8 black vertices and the other of 21 black vertices.

The hyperstories that we want to compute are described in Definition 9 in Section 2.4.2.

A complete enumeration would output all the minimal sets of hyperarcs such that the constraints of Definition 9 are respected. Since the metabolic network is highly connected, we obtained a large number of solutions, as shown in Table 2.3. Given this, it is currently not possible to enumerate all of them and even efficient clustering techniques such as LCM cannot treat the amount of solutions obtained.

Grey vertices	Time	# of solutions	Memory requirement
No	168h	88 330 322	80 Gb RAM ; output file $\sim$ 55 Gb
Yes	168h	56 515 391	53 Gb RAM ; output file$\sim$ 25Gb

Table 2.3: Number of minimal solutions for 8 discriminating compounds. In both cases, the computations used 10 threads and were stopped after 1 week. There is a high number of solutions, and due to the high requirement in computation time and memory, we did not run the experiments with the larger dataset (21 Black vertices).

In order to study more in detail the identified subnetworks, we discuss the solutions obtained while minimising the number of hyperarcs, and with or without minimisation of the number of isolated grey vertices.

Table 2.4 presents the results for the minimum constraint. We can see that for the large dataset, it is necessary to allow isolated grey vertices. In this case, at least seven isolated grey vertices are necessary to obtain one solution.

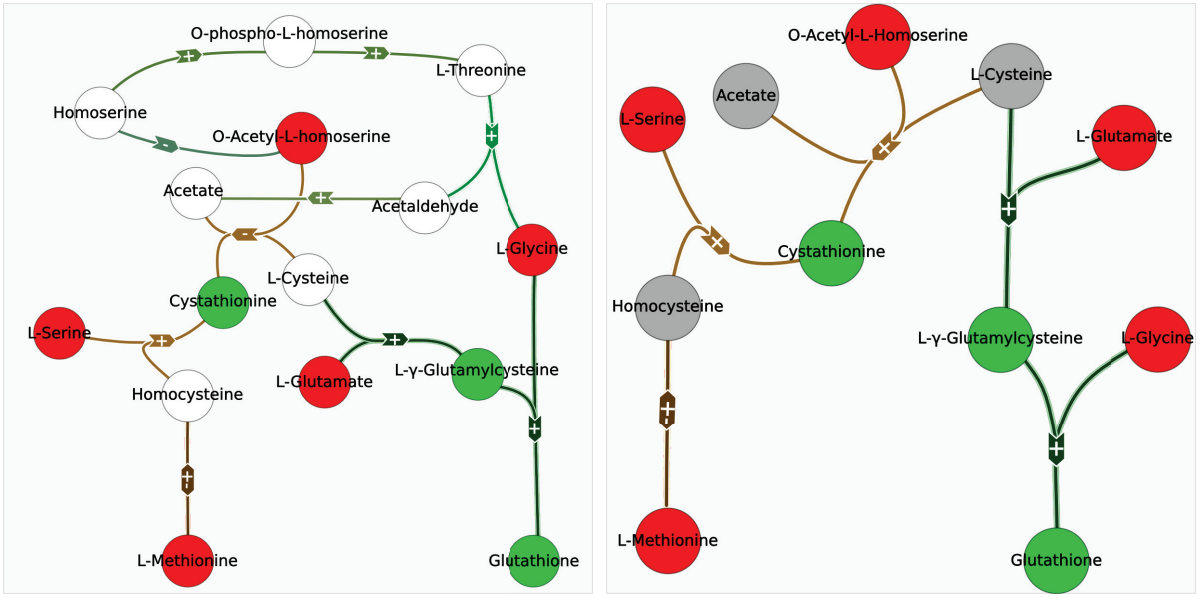
We now discuss more in detail the results of this analysis.

BV file	Grey vertices	Minimisation	# of solutions	Value
8 BV	No	number of hyperarcs	1 solution	10
8 BV	Yes	number of hyperarcs	4 solutions	5
8 BV	Yes	number of hyperarcs and of isolated grey vertices	1 solution	(5 , 2)
21BV	No	number of hyperarcs	No solutions	NA
21 BV	Yes	number of hyperarcs	268	42
21 BV	Yes	number of hyperarcs and isolated grey vertices	1	(42, 13)

Table 2.4: Number of solutions with a minimum number of hyperarcs, and with or without minimising the number of isolated grey vertices. The framework was run for the small dataset (8 black vertices (BV)) and the large one composed of 21 black vertices (BV). The value column indicates the number of hyperarcs in the solution(s). The second value, if one appears, is the number of isolated grey vertices.

2.5.1 Discussion of the results in the case of 8 black vertices

Using the small dataset, there is one solution in the case where we do not consider grey vertices and we minimise the number of hyperarcs that are part of the solutions. Such solution is presented in Figure 2.11a.



(a) Solution when all vertices are considered white (no isolated grey vertices).

(b) Solution with grey vertices. The number of hyperarcs and of isolated grey vertices is minimised.

Figure 2.11: Results for the dataset with eight black vertices. The colours of the reactions represent known pathways annotated in the SBML file, the direction of a hyperarc is the one of the reaction in the original network if it is not reversible, otherwise both directions are included. The sign of a hyperarc represents the sign that explains the observed changes in metabolite concentrations, *i.e.* makes the vertices coherent. A $-$ (resp. $+$) sign indicates that the corresponding reaction should have had a lower flux (resp. higher flux) during the transient state. The colour of the vertices is their colour in our problem -red, green, white, grey-.

We obtain a similar solution when grey vertices are allowed but the number of isolated grey

vertices is minimised. This solution is presented in Figure 2.11b. Here, there are two isolated grey vertices: acetate and L-cysteine.

Those compounds were present in the solution with only white vertices (FIG. 2.11a). Since the isolated grey vertices are not allowed by the problem definition in the solution of Figure 2.11a, there are more hyperarcs connecting them. We remind the reader that in fact, all metabolites that are not green nor red in this dataset were not measured. We cannot really infer that their concentrations did not change between the two conditions. We therefore defined the grey vertices, that can be in some cases *isolated*. When a grey vertex is isolated in a solution, it is a sink or a source in the transformed directed hypergraph (see Definition 8), and is not coherent as presented in Figure 2.8 in the directed hypergraph. We thus expect the concentration of such a grey vertex to change between the two conditions.

As indicated in Table 2.4, in the case where the number of isolated grey vertices is not minimised, four solutions were obtained. The joint representation of those four solutions is presented in Figure 2.12.

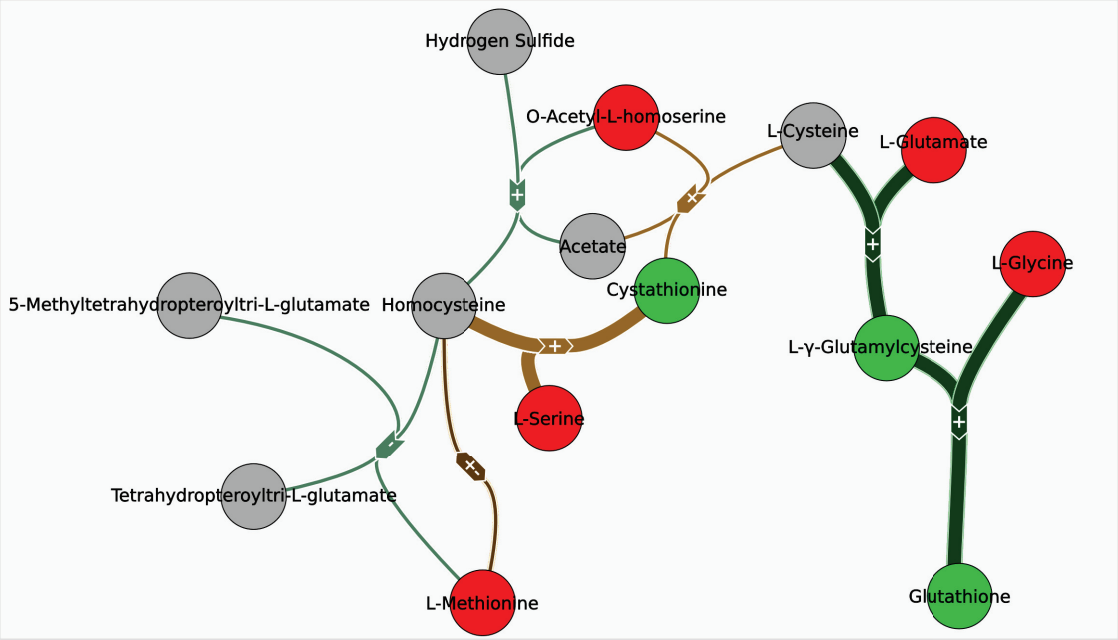


Figure 2.12: Results with eight black vertices and a minimum number of hyperarcs in the solutions. Here I present the 4 optimal ones together. The width of the hyperarcs is proportional to the number of solutions they appear in.

While observing the three cases jointly (Figures 2.11a, 2.11b and 2.12), we can notice three things.

The *Homocysteine S-methyltransferase* reaction is reversible:

$$Homocysteine \Leftrightarrow L\text{-}Methionine.$$

Hence, the post-processing offers two hypotheses, either an inhibition of the methionine synthesis or an activation of the synthesis of homocysteine.

In Figure 2.12, the connexion between cysteine and cystathionine is present only in two out of the four solutions, meaning that two of the solutions are disconnected.

Finally, in Figure 2.11b, the connexion of acetyl-homoserine to the subnetwork is not accurate.

It does not correspond to the one depicted in the detoxification pathway in Figure 2.1. Indeed, there are two reactions that involve acetyl-homoserine and other metabolites of the detoxification pathway. The known reaction is the one using both acetyl-homoserine and hydrogen sulfide to produce homocysteine and acetate. Since hydrogen sulfide was not measured, if this reaction is added to the solution, hydrogen sulfide would be isolated, hence there would be a greater number of isolated grey vertices. In Figure 2.12, the connexion is retrieved since we do not minimise the number of isolated grey vertices.

In these three cases, part of the pathway was obtained and the solutions indicate that it would be interesting to have more precise measurements for the isolated grey vertices that are hydrogen sulfide, L-cysteine, acetate, tetrahydropteroyltri-L-glutamate and 5-methyltetrahydropteroyltri-L-glutamate. Tetrahydropteroyltri-L-glutamate and 5-methyltetrahydropteroyltri-L-glutamate do not seem important for now since there is another reaction connecting L-methionine to homocysteine. However, if the idea is to identify more precisely which of the two possible reactions is active during the metabolic shift, measuring tetrahydropteroyltri-L-glutamate and 5-methyltetrahydropteroyltri-L-glutamate would provide this information.

Furthermore, it was reported in (Lafaye et al., 2005) that there is no increase in the sulfur uptake through sulfate assimilation. However, it is known that the synthesis of glutathione requires sulfur. Sulfur requirement is then met through intracellular metabolites and the protein synthesis drop.

The consensus in terms of the solutions obtained with the small dataset of black vertices is that there has been a flux redirection towards glutathione synthesis. *Glutathione synthase* and *Glutamate–cysteine ligase* (gene GSH1) are of greatest importance at the end of the pathway whereas *Cystathionine β-synthase* (gene CYS4) is also always present in the solutions encountered. GSH1 and CYS4 were reported as strongly induced by the presence of cadmium in (Vido et al., 2001). Those reactions therefore seem to be key points in the metabolic shift observed. They are presented in Equation 2.1.

$$\begin{aligned} &\textit{Glutathione synthase} : \textit{L-}\gamma\textit{-glutamylcysteine} + \textit{L-glycine} \rightarrow \textit{glutathione} \\ &\textit{Glutamate-cysteine ligase} : \textit{L-cysteine} + \textit{L-glutamate} \rightarrow \textit{L-}\gamma\textit{-glutamylcysteine} \\ &\textit{Cystathionine } \beta\textit{-synthase} : \textit{Homocysteine} + \textit{L-serine} \rightarrow \textit{cystathionine} \end{aligned} \tag{2.1}$$

The fact that there are green vertices, that is metabolites with increased concentration, not only at the end point of the pathway but before (*e.g.* L-γ-glutamylcysteine) can appear surprising. However, an increased concentration can result in an increased flux also downstream since there are more substrates available. During the dynamic state, accumulating a metabolite pool can therefore be a strategy to have higher fluxes once the metabolic state is stabilised. Finally, if the enzymes are saturated at one point of the transient state, their substrates must accumulate (Bennett et al., 2009; Tepper et al., 2013).

Saccharomyces cerevisiae adopted a closely related strategy for homocysteine in the presence of cadmium. Indeed (Lafaye et al., 2005) showed that the homocysteine concentration greatly increased in the two first hours of cadmium exposition. However, this concentration then goes back to the one observed before exposure, and then our method cannot capture this precise behaviour as we do not have access to the dynamics in the transient state.

In this work, we retrieved the pathways encountered, without losing possible side-effects due to co-substrates or co-products. This first experiment was performed with a subset of the

discriminating compounds, the ones that are involved in the known detoxification pathway.

We presented here the results obtained while minimising the number of hyperarcs that are part of the solutions. Such hypothesis of minimisation does not seem to be an issue as we obtained the expected pathway. When we search for minimal solutions, we obtained several billions. For those minimal solutions, the number of hyperarcs is not bounded. Thus, we might have some solutions that use really long metabolic roads to connect the green and red vertices. Those answers might not be interesting for now. Indeed, we suggest that the number of reactions impacted by the change in conditions should be small. Observing minimal solutions that have a number of hyperarcs close to the one of the minimum solutions would propose alternative pathways that might be reasonable since not too expensive in terms of reactions.

We now discuss the solutions obtained with the larger black vertices dataset, that is with the 21 discriminating compounds.

2.5.2 Discussion of the results in the case of 21 black vertices

For the set with 21 black vertices, there are no solutions when isolated grey vertices are not allowed, that is when all the non measured vertices are supposed to have a constant concentration across the conditions. Hence, here some metabolite pools have changed when Yeast was exposed to cadmium and were not detected by (Madalinski et al., 2008).

In the case where we minimise the number of hyperarcs and the number of isolated grey vertices, we obtain 1 solution (Figure 2.13).

It is interesting to notice that the greater number of discriminating compounds in the dataset did not increase too much the complexity and the readability of the solution. It is composed of three disconnect components.

The one in the upper right part of Figure 2.13 contains the redirection of the flux from the methyl cycle towards glutathione synthesis using the five reactions previously highlighted and presented in Figure 2.11b for the small dataset. However, there are additional reactions connecting the new discriminating compounds. The unique solution found thus proposes that the cysteinylglycine increase is due to a redirection, as the concentration of reduced glutathione also increases. The hypothesis here is that the *glutathione hydrolase* reaction (that converts glutathione into cysteinylglycine and glutamate) was activated. This is possible since glutathione has an increased concentration: more free substrates could be available for the enzyme. One can notice that, since this reaction is annotated as reversible, another possibility would be that there is less flux in the direction of cyteinylglycine and glutamate towards glutathione. This can be seen as a displaced equilibrium of the reaction with more production than consumption of cyteinylglycine. As this reaction also produces glutamate that is a substrate for L-γ-glutamylcysteine, a precursor of glutathione, glutamate decreases. We can thus see here a cycle that is of interest since it could regenerate glutamate. Furthermore, the other compounds of this component of the solution are also explained. Our Totoro method proposes that the concentration of L-glutamine decreases because of the activation of glutaminase that will convert it into glutamate. Similarly, the decreased concentration of threonine could be a side-effect of an increase in the production of glutathione as it is a precursor of glycine, a substrate of the *glutathione synthase* reaction. Finally, the decrease in lysine can be explained by a redirection of glutamate towards glutathione. The last measured metabolite of this disconnected component is homoserine where the proposed explanation is that there is an increased production of O-acetyl-homoserine which

is then converted into cystathione.

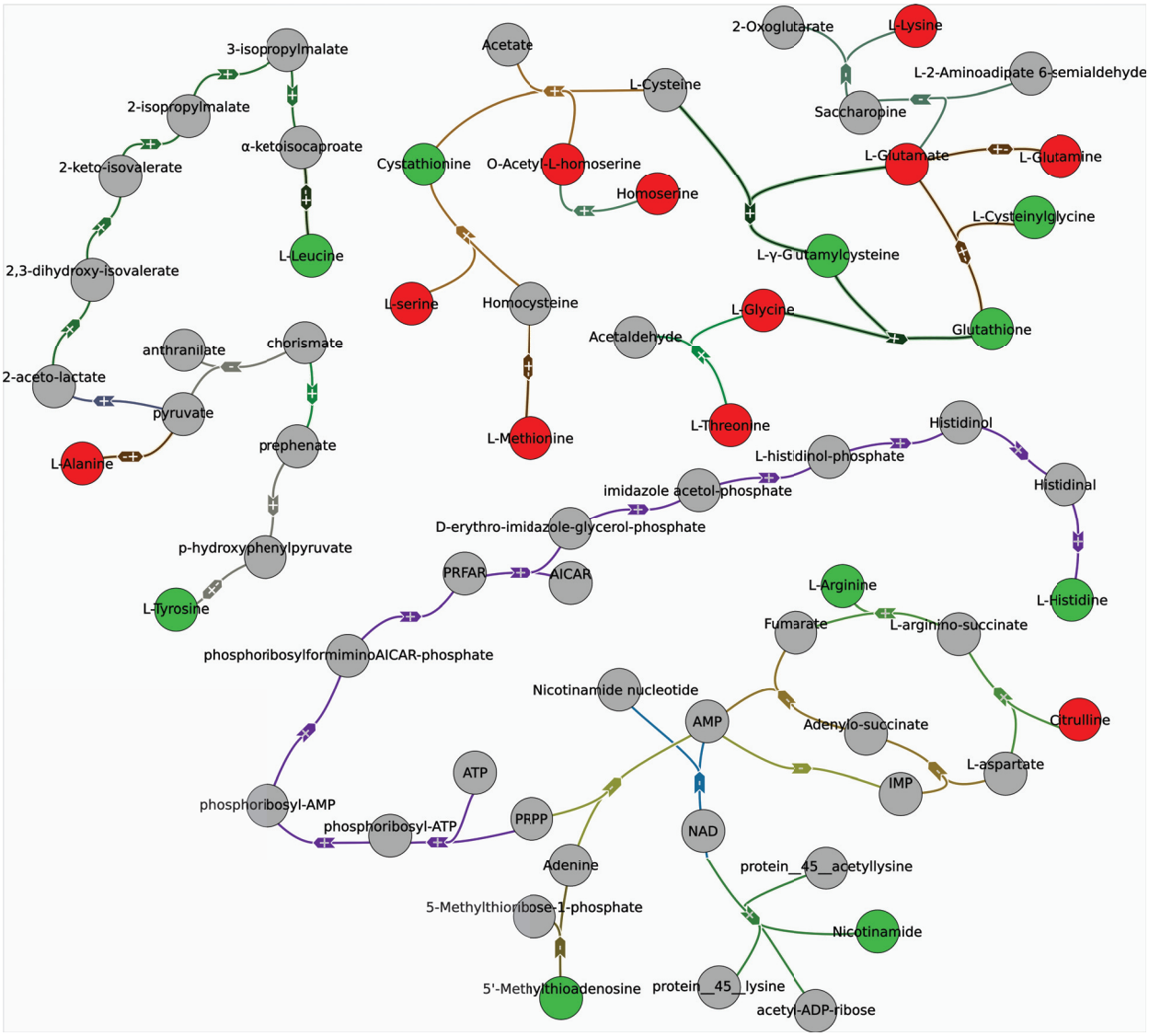


Figure 2.13: Results with 21 black vertices. The number of hyperarcs and of isolated grey vertices is minimised.

The second component of the solution connects L-alanine, L-leucine and L-tyrosine and proposes as an explanation that there is an increase in their synthesis. Here, another explanation for an increased concentration of the amino-acids is that there is less protein synthesis, thus leading to accumulation of amino-acids. However, as our method proposes solutions in the small molecule metabolism, we do not depict the protein synthesis. Indeed, as Yeast is exposed to a stress (namely, cadmium), the protein synthesis diminishes rapidly when cadmium is present in the medium (Lafaye et al., 2005).

However, some of the measured amino acids, such as L-alanine, L-threonine and others were not measured as increased but rather as decreased. This is striking if we do not take into account a possible amino acid degradation. (Link et al., 2015) showed that when protein synthesis is inhibited in *Escherichia coli*, some amino-acids are degraded whereas others accumulate. The authors' main hypothesis was that the non-degraded metabolites are costly to produce. Furthermore, the ones that are degraded are used as energy providers during glucose starvation.

Hence they distinguish two classes of amino acids: Class 1 is such that amino-acids are degraded whereas Class 2 are the amino-acids that are accumulated when protein synthesis is suspended.

If we transpose this observation into *Saccharomyces cerevisiae*, we notice that the green amino-acids (accumulated) indeed correspond to the Class 2 of amino-acids that have a high production cost (hence tend to accumulate) whereas five out of seven of the red amino acids correspond to the identified Class 1 which are amino acids that are degraded. The remaining two are methionine and lysine. As explained previously, methionine decreases because of a redirection of the fluxes from the methyl cycle to glutathione synthesis and lysine. We propose that lysine synthesis is inhibited because of this redirection towards glutathione that consumes a lysine precursor (L-glutamate) through the *Glutamate-cysteine ligase* reaction. We can see here a limitation of the small molecule framework. Since protein synthesis is not modelled, it is not possible to infer that such synthesis diminished rapidly during cadmium exposure.

The last component of the solution contains two amino acids that appear increased (histidine and arginine) and indeed are supposed to accumulate. The reason for the decreased concentration of citrulline together with the increased concentrations of 5-methylthioadenosine and nicotinamide is unclear for now. One could think of a possible relation with oxydo-reduction balance in the organism that is impacted by the increased concentration of reduced glutathione, but here, as for the protein synthesis, such balance is not modelled.

The sub-hypergraph extracted with our framework shows a possible connexion in the metabolic network of such compounds, and the reactions possibly impacted though some of the possible activations/inhibitions are not accurate in the case of amino acids synthesis.

When we do not minimise the number of isolated grey vertices, there are more solutions (268 metabolic hyperstories). To see what are the common reactions, we computed the closed frequent itemsets. By choosing a high threshold (here 200), it is possible to examine the most frequent combination of hyperarcs in our solutions. Such itemsets take into consideration the direction of the reactions used. Hence if in one solution, a hyperarc $a \in \mathcal{A}_t$ is used in one direction, while in another solution its transformed hyperarc $\bar{a} \in \mathcal{A}_t$ is used (in the reverse direction), such reaction will not be part of a frequent itemset containing the common hyperarcs of the two solutions. Here, there are two closed itemsets which are present in at least 200 solutions. One is found in 268 solutions, that is in all of them. These common reactions are shown in Figure 2.14.

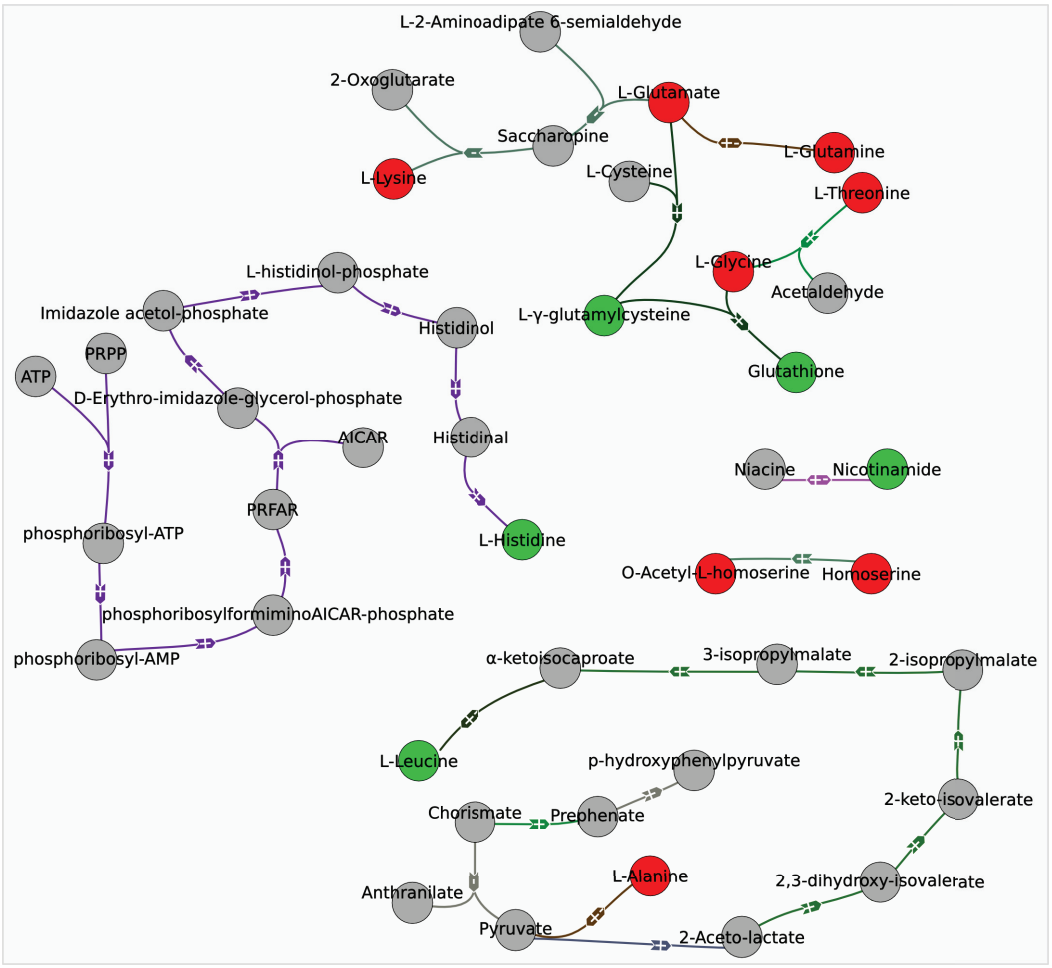


Figure 2.14: Subnetwork obtained only in the closed frequent itemsets with a support of 268.

We can notice that even allowing for isolated grey vertices and working with the large dataset of black vertices, the increased glutathione production is retrieved. However, the methyl cycle is not present. This shows that other explanations are possible as methionine is involved in many reactions (see Figure 2.15 representing the 268 solutions). Also, the closed frequent itemset indicates that the connexion between L-alanine and L-leucine is always present. Homoserine and acetyl-homoserine also always appear connected, but this could be easily explained by the fact that they are direct neighbours (obtained since we minimise the number of reactions) and a cascade effect can easily be supposed as discussed previously.

Looking at the anthology of the 268 solutions in Figure 2.15, it becomes even clearer that there is not a lot of variability. The superpathway of histidine, purine, and pyrimidine biosynthesis (in purple) is shown as activated and starts from ATP and 5-phospho-alpha-D-ribose 1-diphosphate. Lysine synthesis appears decreased and the fluxes redirected toward L-glutamate that is used in the cadmium detoxification pathway.

Thus, Totoro managed to retrieve the known pathway and proposes possible explanations for the metabolites not involved in it.

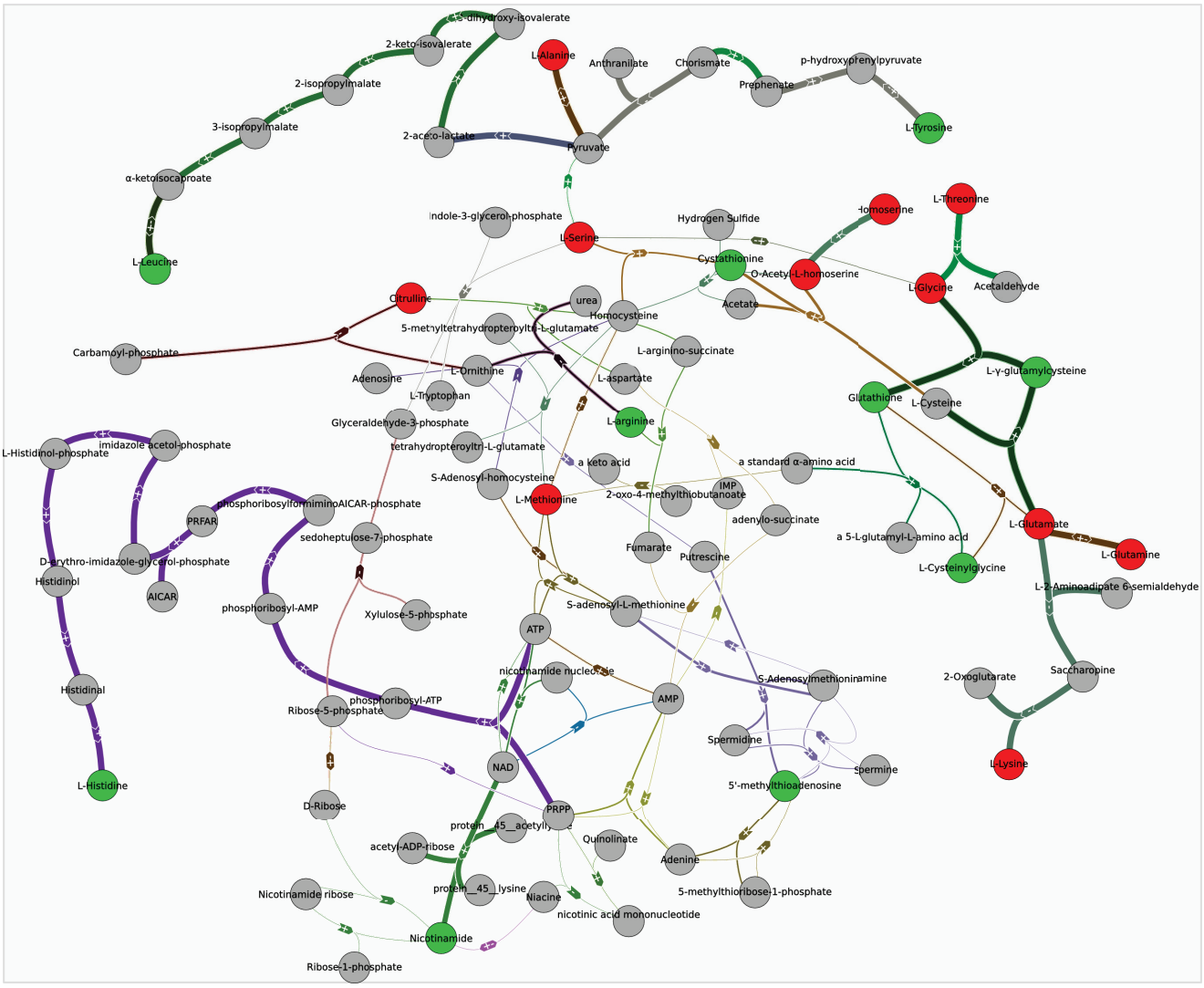


Figure 2.15: Anthology of the 268 solutions when the number of isolated grey vertices is not minimum. *Glutathione synthesis* (right part of the figure, dark green reactions) is present in all solutions as shown by the width of the reactions involved.

2.6 Conclusion

In this chapter, we presented a method, called TOTORO that allows to extract the subnetworks impacted by a change in conditions. Although the identification of impacted subnetworks has already been described in the literature (see (Dittrich et al., 2008) and (Leader et al., 2011) among others), to our knowledge, ours is the first exact enumeration method that uses a directed hypergraph model and infers the potential impacted sub-hypergraphs based on discriminating vertices.

TOTORO was used in the case of a change in culture conditions (*Saccharomyces cerevisiae* exposed to cadmium) in order to infer which reactions were activated or repressed during the transient state. We showed that we are able to retrieve the known pathway and to infer some connexions between the discriminating compounds that are worth exploring. This was possible even in the case of the large dataset of metabolites. These results validate the framework proposed. A more recently obtained yet still unexplored dataset would be of interest to enhance the innovation of this method. We aim to apply it also to the case where there are genetic modi-

fications performed to an organism (in a context of synthetic biology such as will be presented in Chapter 4), where the effects of those modifications on the metabolism are unknown and not directly identifiable at the level of the metabolic network. Indeed, it is possible to alter the regulation of genes that do not code for enzymes but that themselves have a regulatory action. Such genes will create flux changes that will impact the metabolite concentrations at steady-state. We therefore hypothesised that reaching the steady-state in a wild-type or a mutant strain may lead to different metabolite pools due to differences in dynamics, and that such effects might be captured.

We started from a previous method (Milreu et al., 2014) that showed promising results and developed a new model to better understand the metabolic shifts. We then proposed a new definition for the *metabolic hyperstories* that better takes into account the measured variations. Moreover, using a directed hypergraph to model the metabolic network is a progress and we can now infer the direction of the actions on the reactions. We further added two novelties as concerns the metabolites. First, for the discriminating compounds (black vertices in $\mathbb{B}$), we take into account the direction of the variation. Second, in the optimisation process, we proposed to divide the set $\mathbb{W}$ of white vertices into two subsets that reflect the information we have on such vertices (namely measured or not).

This method is dependent on the metabolites identified and measured. Allowing for isolated grey vertices in this new definition relaxes the type of solutions obtained but does not fully address the problem since the solutions obtained remain centred around the discriminating compounds that were measured. Nevertheless, some main regulatory effects can now be retrieved. Furthermore, Totoro can be used to propose new experiments to target the detected isolated grey vertices. It is thus possible to use this framework with an iterative strategy to discover novel modified pathways. One important improvement is that there is no acyclicity constraint anymore, which was one of the limitations of the previous method. Whereas cycles will not always be present (*i.e.* they are not forced), they are no longer prohibited. Finally, the current implementation of minimum solutions is tractable, and thus, there is no need as previously to apply the lightest paths compression. As mentioned, such lightest paths compression would lead to a smaller network, but sometimes at the expense of pathways that would pass through highly connected metabolites. With Totoro, obtaining the minimum solutions of a genome-scale network can be done on a personal computer in a few minutes.

As mentioned previously, the solutions obtained are not necessarily connected. It is entirely possible that two distinct connected components of the network have been influenced separately. This can depict the fact that two separate parts of the small molecule metabolism have been impacted by a common regulator for example. Connectivity therefore really does not have to be a required property of the solutions.

Moreover, we saw in the case of cadmium exposure that the small molecule framework can present limitations. It was not possible to detect the drop in protein synthesis, since this, as well as many other macromolecule synthesis pathways are not modelled. It is not trivial to include such synthesis in our modelling framework. Indeed, a recent development (Link et al., 2015) showed that amino acid accumulation is not uniform due to different degradation behaviours. The question therefore remains open on how to model and interpret such phenomenon. Despite this difficulty, it was possible to see unexpected concentration modifications of the amino acids and to propose a topological explanation even if the sign of the reactions selected was not necessarily accurate.

This shows that more information could be used to depict the complete metabolic process. With the increased usage and availability of omic techniques, the interest of a multiple-layers model that will not only include the small molecule metabolism but also some macromolecules such as proteins including enzymes, as well as gene expression could be useful to provide a clearer picture of the response of an organism to a new condition. Indeed, this is a first approach and without any additional information (such as related to gene expression regulation), it is only possible to draw hypotheses on what are the causes or consequences of the metabolic shifts observed.

A first step to the improvement of the model could be to include protein quantities and enzymatic activities. It could refine the model and possibly restrain the solution space. Moreover, we could score the solutions to get a subset of the minimal ones. Such scoring system should use the stoichiometry of the reactions and the changes in concentrations. Taking into account such information while computing the minimal solutions will help overcome the fact that the minimal solutions are too numerous to be treated correctly. A possible line of research would be to use mixed-integer programming, hence a constraint-based approach, to infer possible flux variations such that inequalities based on the direction of the metabolite concentration changes are respected.

As concerns the minimal solutions, it would also be interesting to analyse a subset only of them, for instance those containing a number of reactions below a certain threshold that can be equal or close to the number of hyperarcs in the minimum solutions. Indeed, it is possible that a slightly longer hyperpath would provide an alternative pathway and be of interest to study. Furthermore, both constraints are based on the hypothesis of a trade-off to conserve energy. Such energy can be for example the number of proteins constituting an enzymatic complex or the Gibbs energy required to perform certain reactions. Hence, one could define a cost for each reaction and use such cost to obtain solutions of minimum weight.

Finally, the developed framework has a larger scope of application. Without the preprocessing of a metabolic network associated to the black vertices, we proposed a solution for the search of hyperpaths (using the definition of (Gallo et al., 1993)) from a set of sources to a set of targets. An application to metabolism would be the search of hyperpaths linking a minimal set of precursors to a target compound. Indeed, (Cottret et al., 2008; Acuña et al., 2012b) proposed algorithms to enumerate the minimal set of metabolites called precursors that are sufficient topologically to produce some target compounds. (Cottret et al., 2010a) then used such minimal precursor sets enumeration to study the symbiotic association of an insect (*Homalodisca coagulata*) with two bacteria (*Baumannia cicadellinicola* and *Sulcia muelleri*) and to infer the possible metabolic exchanges between the two bacteria and their host. In our case, once the minimal precursor sets are known, it would be possible to apply Totoro to infer the metabolic roads taken to reach the target compound(s). This would be done by assigning to red the known precursors and to green the targets.

Moreover, the enumeration of *hyperstories* is an interesting theoretical problem whose complexity remains open (both for finding one hyperstory and for enumerating all minimum or minimal ones). Better algorithms are also required. We used a generic solver, clingo, but an adapted algorithm may lead to more efficient computations.

In the next Chapter 3, we propose a first method to use the metabolite pools measurements when their concentration in both conditions is known. We thus infer in a more quantitative manner the reaction changes during the transient state.

Chapter 3

Quantifying the metabolic responses to perturbations and inferring the transient states

Contents

3.1 Introduction

Various methods were previously proposed to infer which metabolites were possibly involved in regulation and allosteric interactions in condition-specific models. Such regulations were observed around a steady-state. (Reznik et al., 2013) performed a steady-state imbalance analysis derived from the classical flux balance analysis (FBA) framework. The authors showed that *shadow prices*, which are variables of the dual problem of FBA, can be interpreted as the contribution of each metabolite to the objective function, namely optimal growth. (Rohwer and Hofmeyr, 2008) and (Christensen et al., 2015) presented a method to identify regulatory metabolites by varying *in silico* their known concentrations in a measured steady-state using supply-demand analysis. Both methods are therefore based on the response of an organism to a relatively small perturbation, and on the influence of the metabolite concentrations on the reaction rates of the system to return to the original equilibrium. (Link et al., 2013) modelled changes that occur during a transition between two steady-states. The authors proposed novel allosteric interactions using dynamic metabolite data (that is, measurement of the metabolite concentrations along time) and ^{13}C labelling.

Metabolites can regulate reactions and so their variations act on the metabolic fluxes (Wegner et al., 2015). Here, we do not propose to identify regulatory metabolites but rather to infer (in a quantitative way) reactions that may have caused the variations observed in the metabolite pools. Hence we do not work within a metabolic control analysis framework nor do we propose new regulations. In this Chapter, we present KOTOURA (Kantitative analysis Of Transient metabOlic and regUlatory Response And control), that infers the possible rate variations that led to the differences in the measured metabolite concentrations. The idea is to infer quantitative changes of the reactions using information on the metabolites. Such changes occurred during the transition from one steady-state to another. The presented method does not require information on the transient state, that is on the dynamic of the shift from one steady-state to another. The analysis thus cannot infer the exact regulation that took place, but without having any dynamic or kinetic information, it can identify the reactions that led to the measured differences of concentration. We thus focus on identifying and quantifying the differential changes in reaction rates during metabolite state shifts.

As mentioned previously, using a kinetic framework can be limiting as the required information is not always available. Kinetic parameters can be measured *in vitro* but the *in vivo* interactions between metabolites are not captured. Furthermore, the kinetic parameter values that are inferred *in silico* are usually fitted for a specific condition and are often not valid in another one that was not tested, as explained in Chapter 1 and by (Khodayari et al., 2014). Nevertheless, the topology (along with the stoichiometry) already provides insights on the metabolism (Stelling et al., 2002), and we thus wanted to use the known variations in the metabolite pools to infer reactions that were changed during the transient state.

Although this problem is similar to what was previously presented in Chapter 2, it provides a quantitative evaluation of the flux changes without inferring the possible regulators. In this case, we use more information than previously, as we have exact concentration measurements in both conditions (the two steady-states considered). We can quantify exactly the excess or the lack of matter that went through the reactions during the metabolic shift.

As in the remaining of this PhD work, we propose a modelling approach that aims to be general, thus applicable to several types of metabolic shifts, whether in the case of an adaptation

to a new environment, of a response to external signals such as stress, or even of a genetic modification (knock-outs or regulations).

We start by presenting the modelling proposed. Using simulated data, we then describe the possible interpretation of the results obtained, first with a toy model composed of three metabolites and five reactions, then on a larger kinetic model published by (Chassagnole et al., 2002). Finally, we discuss another kinetic model published by (Khodayari et al., 2014), and the current metabolite concentrations inferred. For this last dataset, we are not currently able at the time of writing this manuscript to present the results. We hope to be able to do it during the defence.

This chapter introduces a proof-of-concept, that will then need to be tested on larger biological datasets.

3.2 Using quantitative information

As for the topological method presented in Chapter 2, the idea here is to infer the reactions whose variations led to differences in metabolite concentrations during a switch between two steady-states. It is however possible in this case to use more information which corresponds to the measured differences in concentrations.

3.2.1 Mathematical framework

Using a mathematical approach closely related to Flux Balance Analysis, we would like to infer the reaction variations using information on the metabolite concentration changes.

As presented in Chapter 1, variation of the metabolite concentrations can be written as:

$$\frac{dX}{dt} = \mathcal{S} \cdot v$$

where X represents the concentrations of the metabolites, t the time, $\mathcal{S}$ the stoichiometric matrix, and v the flux distribution vector.

For one metabolite i, it is possible to compute its concentration variation as:

$$\frac{dX_i}{dt} = \mathcal{S}_{i\cdot} \cdot v.$$

Hence the variation in concentration of the metabolite between two time points t_0 and t_f is:

$$\int_{t_0}^{t_f} dX_i = \int_{t_0}^{t_f} \mathcal{S}_{i\cdot} \cdot v(t) \cdot \mathrm{d}t.$$

We thus have that:

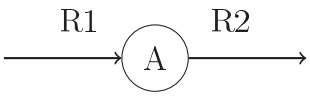
$$
\begin{aligned}
X_{i,t_f} - X_{i,t_0} &= \int_{t_0}^{t_f} \sum_{j=1}^{j=n} s_{ij} \cdot v_j(t) \cdot \mathrm{d}t \\
\Leftrightarrow \Delta X_i &= \sum_{j=1}^{j=n} s_{ij} \cdot \int_{t_0}^{t_f} v_j(t) \cdot \mathrm{d}t \\
\Leftrightarrow \Delta X_i &= \sum_{j=1}^{j=n} s_{ij} \cdot \varphi_j
\end{aligned} \tag{3.1}
$$

where j is the reaction index (which goes from 1 to n), and s_{ij} the stoichiometric coefficient associated to the metabolite i and the reaction j.

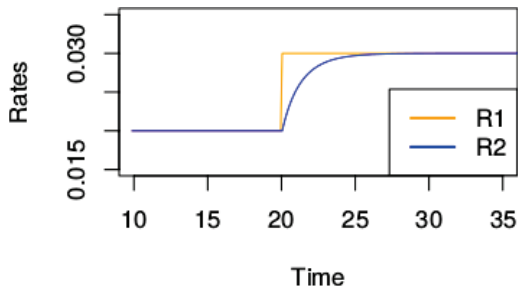
We have that $\varphi_j = \int_{t_0}^{t_f} v_j(t) \cdot dt$ is the overall number of moles that passed through the reaction j during the time interval $[t_0, t_f]$. We may notice that φ is the area under the curve of $v(t)$ for the same time interval.

The graphical idea behind this it that we could interpret the changes of concentration ΔX as an area difference of the reaction fluxes along time (this is illustrated in Figure 3.1 and discussed more in detail in Section 3.3.1) and thus infer φ such that:

$$
\Delta X = \mathcal{S} \cdot \varphi. \tag{3.2}
$$



(a) Small example of two reactions, one producing a metabolite A ($R1$) and one consuming it ($R2$).



(b) Example of reaction rates along time, where A would have increased, with R2 being a bottleneck. The increase of the metabolite A is the area defined between the blue and the orange curves.

Figure 3.1: With the system presented in Figure 3.1a, the obtained vector φ has to be interpreted in comparison to the other reaction. Here, we have that the metabolite variation $\Delta_A = \varphi_{\mathrm{R1}} - \varphi_{\mathrm{R2}}$. Thus, $\varphi_{\mathrm{R1}} > 0$ and $\varphi_{\mathrm{R2}} = 0$ leads to an increased concentration of A. Having $\varphi_{\mathrm{R1}} = 0$ and $\varphi_{\mathrm{R2}} < 0$ would lead to the same result in terms of metabolite concentration. Observe that such solution can be interpreted as R2 having a lower flux than R1 during a certain amount of time of the transient state (such amount of time can be small). There are therefore several explanations: R2 might have an anticipated decreased flux (*i.e.* R2 decreased its flux while R1 was higher than R2), but another option would be that R2 had a delayed increase in flux (being a bottleneck as in Figure 3.1b). Thus, the proposed φ always depends on the remaining of the reactions in the network.

Measurements of concentrations are not always exact. Indeed, there are differences in the measured concentrations across replicates which are due to biological and technical variability.

Moreover, the concentrations obtained are always only as precise as the machines can be. Hence, using exact concentrations in the formulation (that may not be exact in the real world) can lead to infeasible systems. Nevertheless, it is possible to use replicates and the standard deviation between measurements to obtain an interval corresponding to a possible variation of the metabolite concentrations.

We arbitrarily assume that the concentration of a metabolite X_i measured at time t_k will with high probability be in: $\overline{x_{i,t_k}} \pm sd(x_{i,t_k})$, where $\overline{x_{i,t_k}}$ represents the mean of the measured concentrations and $sd(x_{i,t_k})$ the standard deviation. We thus estimate that the possible variation between two conditions stands between:

- $\Delta_{X_i}^{\max} = (\overline{x_{i,t_f}} + sd(x_{i,t_f})) - (\overline{x_{i,t_0}} - sd(x_{i,t_0}))$, the maximum possible variation;
- $\Delta_{X_i}^{\min} = (\overline{x_{i,t_f}} - sd(x_{i,t_f})) - (\overline{x_{i,t_0}} + sd(x_{i,t_0}))$ the minimum possible variation.

The interval of values for the measured metabolites is $\Delta = [\Delta^{\min},\ \Delta^{\max}]$. If we suppose that the non-measured metabolites did not change, they will have a small interval of variation, centred around 0, that is $\Delta = [-\varepsilon,\ \varepsilon]$ with $\varepsilon \in \mathbb{R}^+$.

In this case, the proposition is that the non measured metabolites did not exhibit large changes of concentrations. This can be adapted as it is not always the case.

A small example will be given later on (Section 3.3.1). We make here the same assumption as in the previous Chapter 2, which is that a minimum number of reactions will be changed. We therefore want to select a few reactions that play a major role in the metabolic shift. In this case, we first minimise the number of reactions that are part of a solution, that is, the number of values of the vector φ that are not null. This is equivalent to saying that we suppose the reorganisation of the network in the transient state to be localised, that is mostly based on some reactions only.

Moreover, our second hypothesis is that the differential changes of the reactions, that is their desynchronisation around a metabolite, is also small. The decision is then to minimise $|\varphi|_1$, that is the L_1-norm. Since we minimise $|\varphi|_1$, the solutions obtained will be sparse. This is close to the Lasso (Least Absolute Shrinkage and Selection Operator) method (Tibshirani, 1996). It selects a subset of variables (φ) that are sufficient to explain the observations (the metabolite variations Δ). We will obtain solutions with a large number of null values of the φ vector and avoid solutions that are multiples of a smallest one. Furthermore, $|\varphi|_1$ can be implemented as a linear constraint, thus making the computations easier.

We may notice that the two assumptions presented are similar to the ones made in Moma (Segrè et al., 2002): to reach a new equilibrium, an organism will reorganise its fluxes in a parsimonious way. While this assumption is quite satisfactory in synthetic biology, it can be discussed in the first case described later on, where there is a change in the culture condition and not on the actual genetic content of the organism (such as the knock-out of an enzyme-coding gene). We remind the reader that the Moma approach assumes that modified strains did not undergo evolution and are not (yet) optimal. Their flux distribution is thus not computed as the maximisation of a cell objective but instead as a minimisation of the differences with the flux distribution of the wild-type. In the latter case, since wild-type strains are supposed to have undergone multiple generations, one could propose that their organisation will be totally changed between two culture conditions.

In our case, we nevertheless suppose that a minimal reorganisation will be preferred in terms of energy and possible cascade effects. This is supported by (Schuetz et al., 2012) who showed

that wild-type organisms do not function optimally, but instead in a slightly sub-optimal manner. This hypothesis was tested in *Escherichia coli* and explained as a trade-off. The idea is that even though an organism is not optimal in a culture condition (but remains close to optimal), the changes required by a new condition would be less expensive in terms of flux reorganisation, and thus less energy demanding than if the organism was optimal. As one can imagine, microorganisms do rarely evolve in perfectly stable environments, thus an efficient adaptation is probably indeed a reasonable evolutive assumption.

Since we are performing a quantitative evaluation of the reorganisation of a system, all the solutions obtained will be in agreement with the measured changes of concentration. The solutions will show the reactions whose desynchronisation during the transient state led to the measured changes in concentrations.

3.2.2 Problem formulation

The problem that we want to address can be formulated in the constraint-based framework using a MILP (Mixed-Integer Linear Programming) approach.

The first solution is obtained using the formulation given in Equation 3.3:

$$\begin{aligned} \min_{\varphi} f = & \sum_{j=1}^{n} y_j + \frac{\sum_{j=1}^{n} |\varphi_j|}{K} \\ s.t \quad & \Delta_X^{\min} \leq \mathcal{S}.\varphi \leq \Delta_X^{\max} \\ & lb_j \leqslant \varphi_j \leqslant ub_j, \forall j \in \mathcal{R} \\ & y_j = 0 \leftrightarrow \varphi_j = 0, \forall j \in \mathcal{R} \\ & y_j \in \{0,1\}, \forall j \in \mathcal{R}. \end{aligned} \tag{3.3}$$

In the optimisation problem represented by Equation 3.3, y_j is a boolean variable that relates to reaction j. When a reaction j is involved in the concentration changes during the transient state, y_j is equal to 1. If y_j is null in the proposed solution, the reaction j was not responsible for changes in the metabolite pools between the two conditions. As expressed by the third constraint, y_j is null if and only if φ_j is null. All the values of φ_j are between a lower bound (lb) and an upper bound (up) that can be specified if known. The first constraint expresses the fact that metabolite variations are the result of the fluxes involving the neighbour reactions along time, as explained in Section 3.2.1.

The objective function is a sum of the two assumptions discussed previously in Section 3.2.1.

First, $\sum_{j=1}^{n} y_j$ is the number of reactions with $\varphi \neq 0$. The sum $\sum_{j=1}^{n} y_j$ is an integer (since $y_j \in \{0,1\}$). Here, we minimise the number of reactions involved in the solution.

The second member of the optimisation function is such that we minimise the absolute values of the vector φ: $\sum_{j=1}^{n} |\varphi_j|$. There are no preferences for the sign of φ_j. Indeed, both are valid as explained in Figure 3.1, Section 3.2.1.

The second member of the optimisation function is normalised by K, a parameter which is chosen such that $0 \leq \frac{\sum_{j=1}^{n} |\varphi_j|}{K} < 1$. In the experiments performed in this manuscript, $K = 1000 \cdot n$ where n is the number of reactions.

Hence, first the number of reactions that are part of the solutions is minimised, and then the possible changes corresponding to the overall desynchronisation of the network is assessed.

To obtain more solutions, we iterate this process by adding the constraint given by Equation 3.4:

$$\sum_{j \in I_{y^*}} y_j \leq |I_{y^*}| - 1 \tag{3.4}$$

with (y^*, φ^*) the solution obtained previously and I_y the support of the vector y. Adding the constraint 3.4, we forbid the solution to have the same support vector. The set of obtained reactions (such that $y_j \neq 0$) is then excluded from the next solutions. We can therefore obtain solutions with an increasing number of reactions involved, but for all solutions, the reaction set will be minimal.

This type of approach has already been used to enumerate all solutions in the context of other metabolic-related problems (Lee et al., 2000; de Figueiredo et al., 2009; von Kamp and Klamt, 2014; Andrade et al., 2016) by excluding previous solutions in the solving process.

The problem in Equation 3.3 was implemented in C++ during this PhD work and uses the IBM Cplex optimiser solver.

3.3 Finding the response to a change in culture condition

We now present the results obtained on simulated datasets. Here, we will always consider the vector φ as the solution of the formulation presented in Equation 3.3. The approximation of the integral under the curve for the reaction rates will be named AUC for Area Under the Curve. In the examples given below, we compare the AUC of the reaction rates with the vector φ obtained by our framework. As we use simulated data, we have access to the dynamic of the transient state (and so to the AUC). We can then demonstrate that we do manage to retrieve the reaction variations with the formulation adopted. However, when working with an *in vivo* dataset, the AUC are not available. Their comparison with the solutions of our framework (y^*, φ^*) will thus not be possible.

3.3.1 Small toy example

As a small example of the possible response to a change of growth condition, we simulated in this section a toy system composed of three metabolites and four reactions including an entry (uptake) and an exit (export) of the system. The latter is represented in Figure 3.2. This example is not biologically relevant, however we used the kinetic equations obtained by (Chassagnole et al., 2002) and shown in Figure 3.3 to mimic as closely as possible the behaviour of the reactions. A more complete model will be seen later on (Section 3.4).

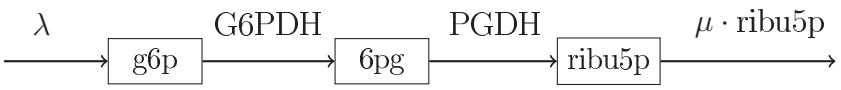


Figure 3.2: Toy example composed of a constant import λ of glucose-6-phosphate (g6p). The latter is then converted into 6-phosphogluconate (6pg) by *glucose-6-phosphate dehydrogenase* (G6PDH). Then *6-phosphogluconate-dehydrogenase* (PGDH) produces ribulose-5-phosphate (ribu5p). We used a mass action law such that matter exits the system in proportion to the ribu5p concentration.

The two intermediate reactions follow a kinetic law called Two-substrate irreversible Michaelis-Menten. The rates are provided in Figure 3.3. These were obtained directly from the literature. We added the uptake and export reactions for explanation purposes and kept the notations of (Chassagnole et al., 2002).

$$
\begin{aligned}
r_{uptake} &= \lambda \\
r_{g6pdh} &= \frac{r_{\text{max}}^{\text{g6pdh}} \cdot \text{g6p} \cdot \text{nadp}}{(\text{g6p} + K_{\text{g6pdh, g6p}}) \cdot (1 + \frac{\text{nadph}}{K_{\text{g6pdh, nadph, g6pin}}}) \cdot (K_{\text{g6pdh, nadp}} \cdot (1 + \frac{\text{nadph}}{K_{\text{g6pdh, nadph, nadphin}}}) + \text{nadp})} \\
r_{pgdh} &= \frac{r_{\text{max}}^{\text{pgdh}} \cdot \text{6pg} \cdot \text{nadp}}{(\text{6pg} + K_{\text{ pgdh, 6pg}}) \cdot (\text{nadp} + K_{\text{pgdh, nadp}} \cdot (1 + \frac{\text{nadph}}{K_{\text{pgdh, nadphin}}}) \cdot (1 + \frac{\text{atp}}{K_{\text{pgdh, atp, inh}}}))} \\
r_{export} &= \mu \cdot \text{ribu5}
\end{aligned}
$$

Figure 3.3: Rates of the reactions implemented as in (Chassagnole et al., 2002) for *6-phosphogluconate dehydrogenase* (PGDH) and for *glucose-6-phosphate dehydrogenase* (G6PDH). The uptake was defined as constant and the export follows a simple mass action kinetic.

Once the reaction rates are defined, one can obtain the concentration variations as shown in Equation 3.5:

$$
\begin{aligned}
\frac{d\ \text{g6p}}{dt} &= r_{\text{uptake}} - r_{\text{g6pdh}} \\
\frac{d\ \text{6pg}}{dt} &= r_{\text{g6pdh}} - r_{\text{pgdh}} \\
\frac{d\ \text{ribu5p}}{dt} &= r_{\text{pgdh}} - r_{\text{export}}
\end{aligned}
\tag{3.5}
$$

In Figure 3.3 and Equation 3.5, g6p, 6pg, and ribu5p refer to the concentrations of glucose-6-phosphate, 6-phosphogluconate and ribulose-5-phosphate respectively. Similarly, nadp, nadph and atp are concentrations. Those compounds are known inhibitors of the reactions and here, their concentrations are considered constant. Constant parameters are provided in Table 3.1 while the initial conditions and the variable λ are given in Table 3.2.

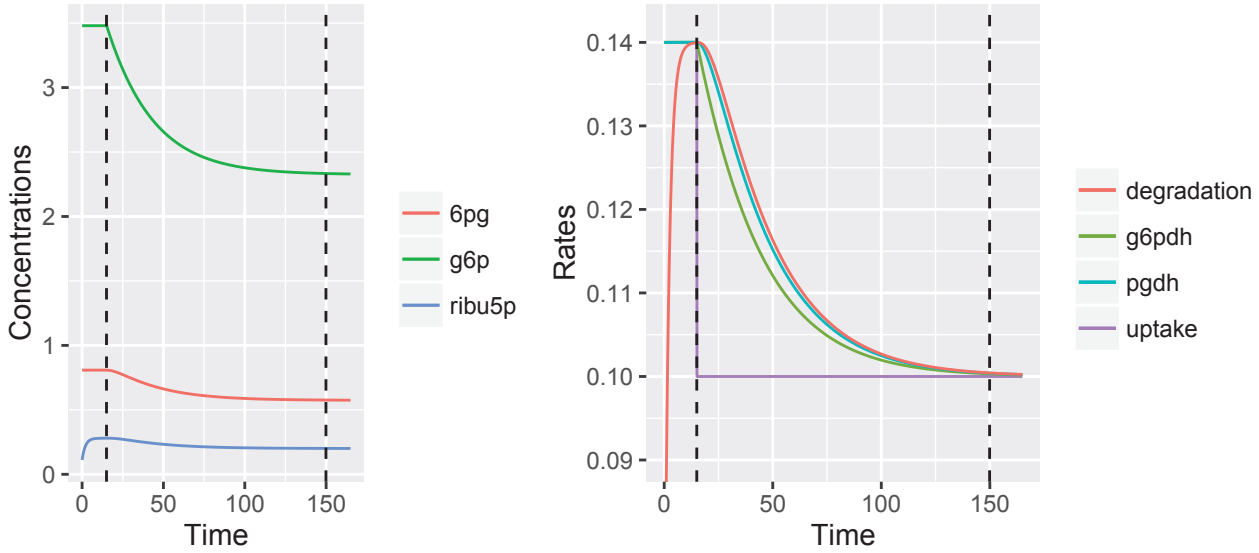
Parameter	Value
$r_{\max}^{\text{g6pdh}}$	1.3802 mmol/(l · s)
$K_{\text{g6pdh, g6p}}$	14.4 mmol/l
$K_{\text{g6pdh, nadph, g6pin}}$	6.43 mmol/l
$K_{\text{g6pdh, nadp}}$	0.0246 mmol/l
$K_{\text{g6pdh, nadph, nadphin}}$	0.01 mmol/l
$r_{\max}^{\text{pgdh}}$	16.234 mmol/(l · s)
$K_{\text{ pgdh, 6pg}}$	37.5 mmol/l
$K_{\text{pgdh, nadp}}$	0.0506 mmol/l
$K_{\text{pgdh, nadphin}}$	0.0138 mmol/l
$K_{\text{pgdh, atp, inh}}$	208 mmol/l
μ	0.215 1/s
nadp	0.196759
nadph	0.062
atp	4.27

Table 3.1: Values of the parameters and of the concentrations considered as constant.

Variable	Condition 1	Starting the transient state
g6p_0	3.48 mmol/l	3.4798301 mmol/l
6pg_0	0.808 mmol/l	0.8081566 mmol/l
ribu5p_0	0.111 mmol/l	0.5882640 mmol/l
λ	0.14 mmol/(l · s)	0.8 mmol/(l · s)

Table 3.2: Values of the variables for the simulation. We start by simulating a first steady-state (called Condition 1). Then, once verified that the concentrations are indeed constants, we switch the uptake of the system λ, and use as new initial conditions for our system of differential equations the concentrations obtained at the end of Condition 1. Here, $(\text{g6p}_0, \text{6pg}_0, \text{ribu5p}_0)$ are the concentrations used as initial conditions for the simulations.

Once the system was implemented, we simulated a first condition during a certain time interval and then changed the uptake of the system. The time-course plots are shown in Figure 3.4. Glucose-6-phosphate which represented the largest pool of the simulation is the most impacted by the decreased uptake as one would expect. As the kinetic laws used for the main reactions are similar, they exhibit the same behaviour, that is a slow decrease of the rates during the transient state. One can observe that this decrease is desynchronised. Indeed, there is a cascade of events from the uptake of the system to the export with a gradual response to the lower glucose input.



(a) Simulated concentrations along time.

(b) Reaction rates along time.

Figure 3.4: Result of the time-course simulations. The dotted vertical lines represent the two time-points for which we computed the variations Δ_X (t_0 and t_f). The time t_0 is also where the uptake switch occurred.

The obtained variations of concentrations (Δ_X) are shown in Table 3.4. As explained in Section 3.2.1, those variations are the result of the overall variation of the reactions (see Equation 3.1). This is represented by the area under the curves of $r(t)$ that can be approximated using the trapezoidal rule. We allowed for an interval of variations for the concentrations. To obtain such intervals, we used a proxy for the standard deviation and computed the interval as:

$$[\Delta^{\min},\ \Delta^{\max}] = [\Delta_X - \gamma \cdot |\Delta_X|,\ \Delta_X + \gamma \cdot |\Delta_X|]. \tag{3.6}$$

We therefore created the interval proportionally to the simulated values with the parameter γ, $0 \leq \gamma \leq 1$. The input and output of the system are of importance. Indeed, the network is not closed, and the uptake and export reactions need to be constrained. It is possible to add what we will call *pseudo-metabolites*. Those pseudo-metabolites can also be referred to as *boundaries* as they delimit our system.

We thus added *sink* and *source* metabolites. Those boundary metabolite variations are defined using the known variations of the reaction rates during the transient state. These are established using the area under the curve (AUC) of the uptake and export reaction rates. Another solution would be to constrain the bounds of the vector φ (ub_j and lb_j in Equation 3.3).

For the sake of simplicity, we chose to add two pseudo-metabolites. A pseudo-substrate of the uptake reaction was added with a difference in concentration $\Delta_{X_{\text{uptake}}}$. The same occurred for the degradation of ribulose-5-phosphate; the pseudo-metabolite difference in concentration is then called $\Delta_{X_{\text{export}}}$. Both values are inferred using the known relations:

$$\Delta_{X_{\text{uptake}}} = -\int_{t_0}^{t_f} r_{\text{uptake}}(t) \cdot dt,$$

$$\Delta_{X_{\text{export}}} = \int_{t_0}^{t_f} r_{\text{export}}(t) \cdot dt.$$

For all the metabolites of the system, we used $\gamma = 0.01$. Once $\Delta_X^{\min}$ and $\Delta_X^{\max}$ were obtained for the five metabolites (the three of the network in Figure 3.2 and the two boundary metabolites), it is possible to compute vector φ using the formulation presented in Section 3.2.2, Equation (3.3). The obtained results are presented in Table 3.3. We compared the vector φ found by our MILP formulation (Equation 3.3) with the approximation of the area under the curves of the reaction rates.

Reactions	Area computed	φ
uptake	13.49382	13.3589
g6pdh	14.64174	14.494
pgdh	14.87379	14.7239
export	14.95281	14.8033

Table 3.3: Computed area under the curves between t_0 and t_f and φ obtained with the formulation in Equation 3.3. We may notice that the values of φ are always smaller. This is due to the fact that we simulated an interval for the concentration changes, hence allowing for smaller values of φ (that are minimised in our formulation).

Finally, we can compare the differences of metabolite concentrations using the information on the reaction rates (either using the vector φ or the computed areas) as shown in Equation 3.5.

Metabolite	Simulated	Inferred with area	Inferred with φ
Δg6p	-1.146613	-1.147920	-1.1351
Δ6pg	-0.2321874	-0.2320457	-0.2299
Δribu5p	-0.07907658	-0.07902390	-0.0794

Table 3.4: Simulated ΔX computed with the concentrations in t_f and t_0 and inferred using the area under the curves representing the rates of the reactions. The small differences can be attributed to numerical error and to the time step selected for the integration of the differential equations.

We may notice that the concentrations inferred from the φ vector are smaller. This is probably due to the fact that we try to minimise $|\varphi|$, hence always reaching for the smallest absolute value of φ_j for all j.

Looking at Table 3.3, it is possible to interpret the φ vector. We can see that $\varphi_{\text{uptake}} < \varphi_{\text{g6phd}} < \varphi_{\text{pgdh}} < \varphi_{\text{export}}$. This indicates that there has been a desynchronisation of the fluxes. Since we have the rates of the reactions in the transient state (Figure 3.4b), we know that they all decreased. The differences of the φ are an indication that overall, the export flux was higher than the other fluxes, that pgdh was higher than the uptake and g6pdh, etc. Hence we know that the decreased rates have been desynchronised in a time slot basically going from the left to the right of the system drawn in Figure 3.2.

If we did not know that all the rates decreased, but instead assumed that they increased, we could infer that the degradation rate was the first to increase, then pgdh, etc.

We can say here for example that the fact that $\varphi_{\text{export}} - \varphi_{\text{pgdh}}$ is equal to 0.0794 should mean that, according to our results, the degradation of ribulose-5-phosphate transferred 0.0794 mmol more than *6-phosphogluconate dehydrogenase* during the metabolic shift.

The vector φ therefore really represents the desynchronisation of the reactions during the metabolic shift, and such desynchronisation leads to the differences of concentrations that were measured. In larger systems, the absolute value of φ_j is thus not really of interest if it is not compared to the remaining of the vector.

We presented this small example as an introduction to the formulation and meaning of φ. We then tested the method on a larger network, using simulated data. We first present the model obtained from the literature, and then the results we got.

3.4 *Escherichia coli* in response to a glucose pulse

As presented in Chapter 1, Section 1.3.2, kinetic models are available to infer the reaction dynamics (outside a steady-state). Such models can be obtained by using metabolite concentrations and reaction rates to estimate the model structures and the value of the kinetic parameters.

We used here this type of model in order to simulate datasets to test the method presented above.

The model presented in (Chassagnole et al., 2002) (shown in Figure 3.5) reproduces the central metabolism of *E. coli*, namely the glycolysis and pentose phosphate pathways and the phosphotransferase system (PTS), as well as the response of those pathways to a glucose pulse. It also takes into account known regulations. The authors confronted the obtained kinetic equations and parameters with dynamic experimental data and showed a good fitting of the predictions with *in vivo* measurements.

3.4.1 Simulating the transient state

We used COPASI (Hoops et al., 2006) to obtain the simulated data and downloaded the model from KYMOSYS (Costa et al., 2014). The motivation was to retrieve the available dynamic fluxes. One drawback of such dynamic models is that we are not always sure of the range of application of the kinetic parameters, as they were often fitted on one or two specific biological conditions. We therefore used the same conditions as in (Chassagnole et al., 2002), following a response to a glucose pulse. We simulate such response when the extracellular glucose concentration is increased and goes from $12mg/L$ to $0.3g/L$.

We simulated the transient state in COPASI and obtained the time-course of the metabolite concentrations and of the reaction rates presented in Figure 3.6.

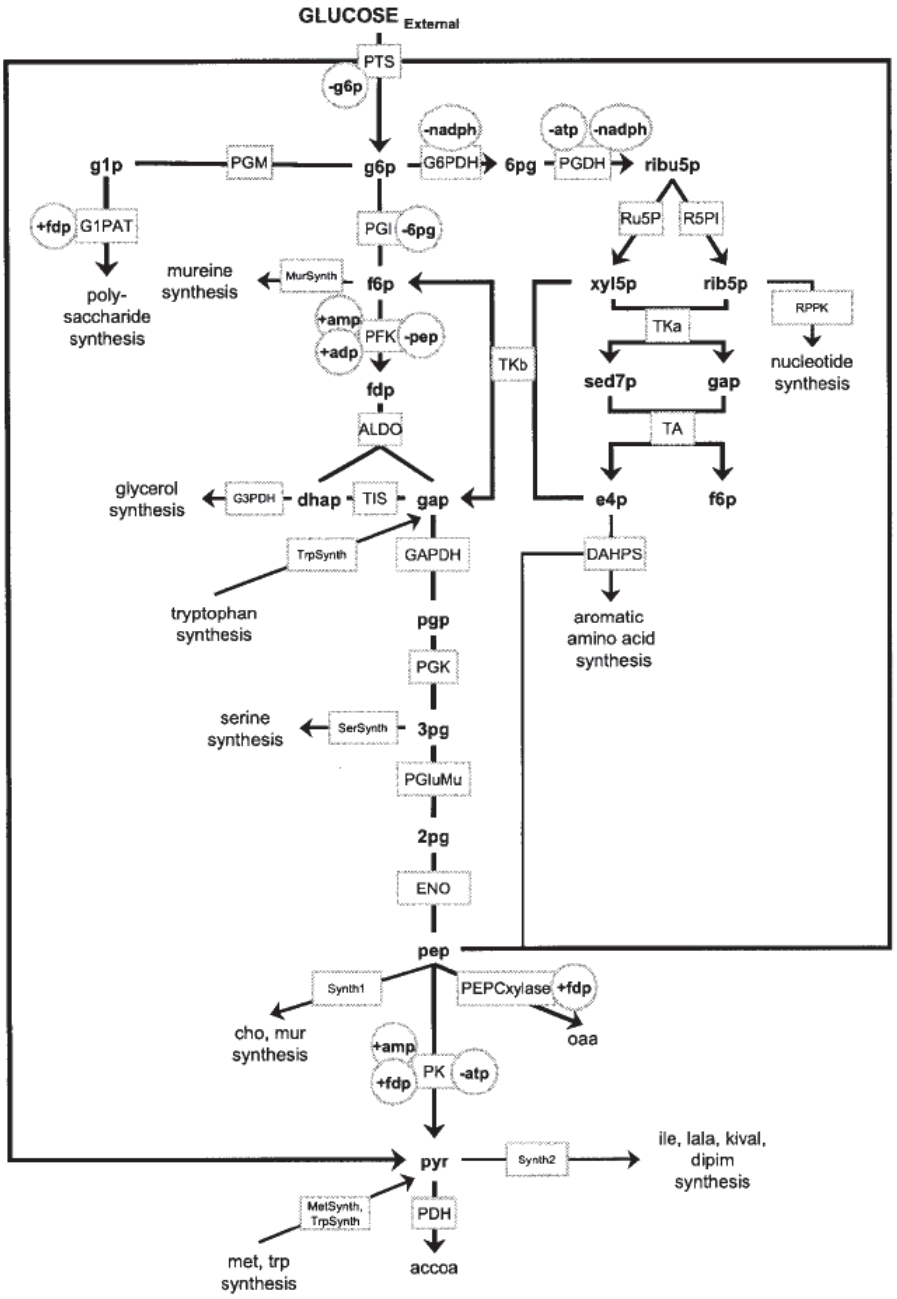
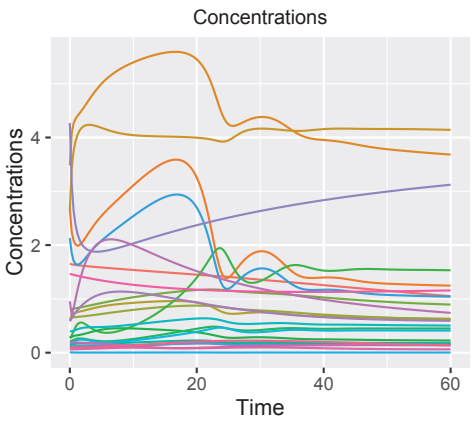


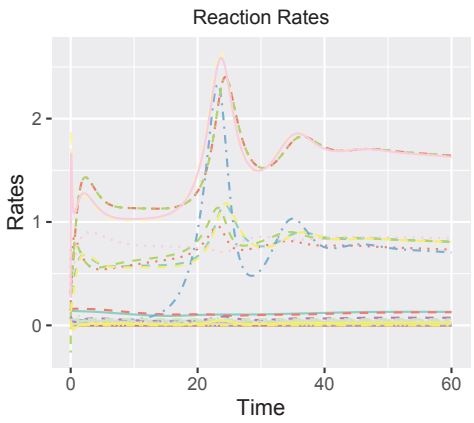
Figure 3.5: Metabolic model from (Chassagnole et al., 2002). The enzymes are represented by squares and the regulatory effects by circles. The complete list of the names of the metabolites and of the reactions is available in Appendix A.1 (Table A.1 and A.2).

3.4.2 Retrieving the data for the model

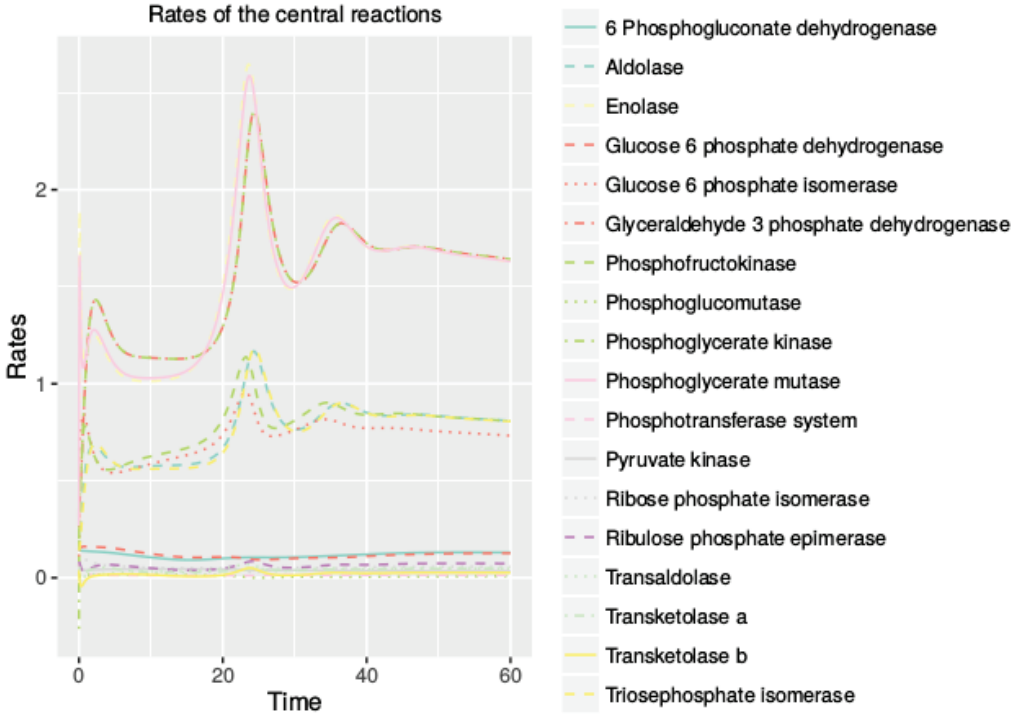
The method proposed requires a metabolic network in SBML format and a text file composed of the metabolite identifiers and their variations $\Delta^{\max}$ and $\Delta^{\min}$. We followed the same procedure as in Section 3.3.1. Here the variations are obtained using the concentrations computed at t_0 ($t_0 = 0$) and t_f (t_f being the end of the time-course simulation).



(a) Simulated concentrations along time.



(b) Reaction rates along time.



(c) Time-course of the 18 reactions that have at least one substrate and one product (neither uptake nor export reactions).

Figure 3.6: Results of the time-course simulations. Detailed representations are given in Appendix A.1.1

As the model contains 30 uptake and export reactions, we added to the original network pseudo-source or sink metabolites for those reactions. We computed the area under the curves (AUCs) of the reaction rates and set the variations of those pseudo-metabolites accordingly. The intervals were simulated with a variation coefficient $\gamma = 0.01$. The resulting variation intervals are shown in Table 3.5.

Once the network and the metabolite variations were obtained, it was possible to run the formulation presented in Section 3.2.2. We discuss in the next section the results obtained.

Metabolite	Δ^{max}	Δ^{min}
cpgp	-0.0049	-0.0050
cpg2	-0.2056	-0.2097
cpg3	-1.0767	-1.0984
cpg	0.0877	0.0860
cdhap	0.2463	0.2414
ce4p	0.0642	0.0630
cglcex	-0.6004	-0.6126
cfdp	1.2731	1.2478
cf6p	0.0316	0.0310
cg1p	-0.0365	-0.0372
cg6p	0.2050	0.2009
gap	0.2345	0.2299
cpep	-1.4084	-1.4368
cpyr	1.4865	1.4571
crib5p	0.1071	0.1050
cribu5p	0.0271	0.0266
csed7p	-0.0499	-0.0509
cxyl5p	0.0401	0.0394

(a) Variation interval for the internal metabolites and external glucose concentrations.

Metabolite	Δ^{max}	Δ^{min}
vDAHPS	1.2388	1.2143
vDHAP	0.0006	0.0006
vE4P	0.0003	0.0003
vEXTER	-0.1810	-0.1847
vf6P	0.0013	0.0013
vfdP	0.0021	0.0021
vG1PAT	0.5814	0.5699
vG3PDH	0.1848	0.1812
vG6P	0.0075	0.0074
vGAP	0.0007	0.0007
vGLP	0.0013	0.0012
vMethSynth	-0.1344	-0.1371
vMURSyNTH	0.0265	0.0260
vPDH	49.3084	48.3320
vPEP	0.0034	0.0033
vpepCxylase	38.8729	38.1032
vPG	0.0017	0.0017
vpg2	0.0005	0.0005
vPG3	0.0028	0.0028
vPGP	0.0000	0.0000
vPPK	0.6579	0.6448
vpyr	0.0069	0.0068
vRIB5P	0.0009	0.0009
vRibu5p	0.0002	0.0002
vSED7P	0.0005	0.0005
vsersynth	0.9430	0.9243
vSynth1	0.7658	0.7506
vSynth2	3.5855	3.5145
vTRPSYNTH	-0.0616	-0.0628
vXYL5P	0.0003	0.0003

(b) Variation interval for the pseudo-metabolites. They are named with the identifier of the uptake or export reaction.

Table 3.5: Variation intervals. All numbers were rounded to the 4^{th} decimal.

3.4.3 Discussion of the results

Using the data presented above, we obtained two solutions (y_1^*, φ_1^*) and (y_2^*, φ_2^*). The raw results are given in Appendix A.1.2, Table A.3. In Table 3.6, we present the computed φ^* and the corresponding AUC for the internal reactions and the glucose export. As mentioned previously, a large number of pseudo-metabolites were added to the network and we assigned their variations using the known ones of the export or uptake reactions. Since those metabolites are connected to the network through only one reaction each, the value of φ_j for the reaction will be directly constrained to the assigned variation of the boundary metabolites and is, as expected, accurately predicted. We thus compare only the values obtained for the internal reactions.

Reaction Id	AUC	φ_1^*	φ_2^*
vPTS	0.7894	0.7814	0.7814
vPGM	0.5401	0.5462	0.5462
vPGI	43.5006	46.6032	48.5447
vPFK	47.1484	47.8206	48.4678
vALDO	45.886	46.5455	47.1926
vTIS	45.4585	46.1137	46.7609
vGAPDH	92.3413	92.4362	93.0834
vPGK	92.3375	92.4411	93.0883
vrpGluMu	92.4947	92.5719	93.2191
vENO	92.6987	92.777	93.4242
vPK	2.3451	2.5117	3.1589
vG6PDH	7.0469	3.4309	1.4894
vPGDH	6.9584	3.3415	1.4
vR5PI	3.2161	2.0198	1.3726
vRu5P	**3.7152**	**1.2943**	**0**
vTKA	2.4578	1.2539	0.6067
vTA	2.5078	1.3033	0.6561
vTKB	**1.2174**	**0**	**-0.6472**

Table 3.6: The φ^* computed and the corresponding simulated values (AUC: Area Under the Curve) for the internal reactions. All numbers were rounded to the 4^{th} decimal. In bold, we show where the reaction supports of the y^* differ between the two solutions.

Such comparison can be graphically visualised (Figure 3.7), but we also propose to perform Spearman coefficient tests to see whether the predicted values do correlate to the computed AUC.

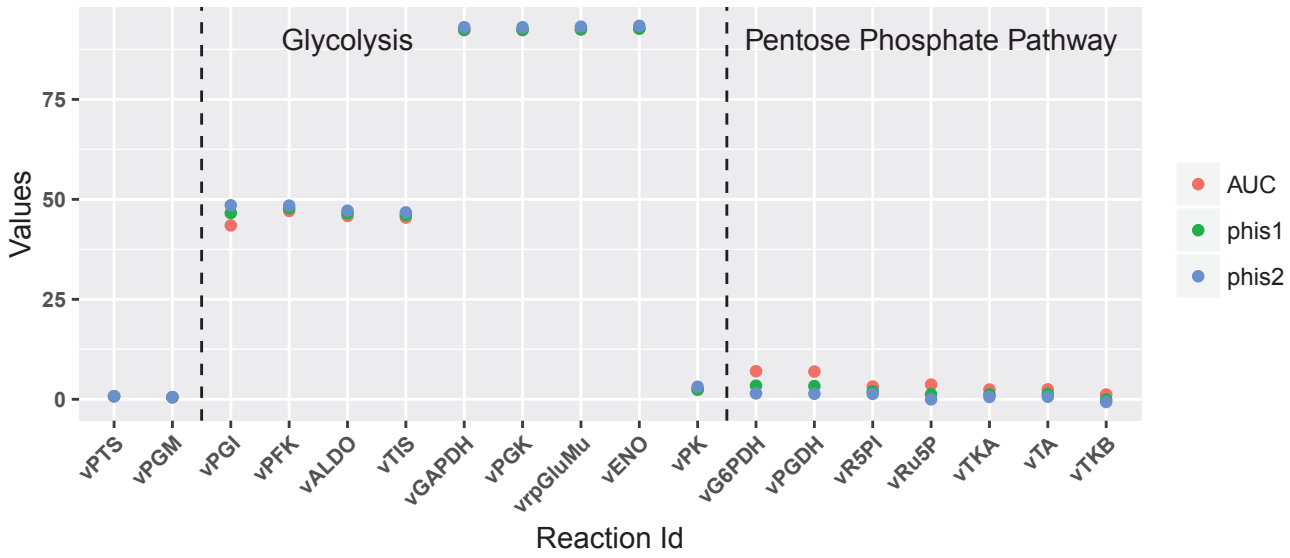


Figure 3.7: Comparison of the values obtained for the two φ^* and for the AUC.

For the solution φ_1^*, we obtain a correlation $\rho_{\varphi_1^*} = 0.9587203$ with a p-value of $5.547 \cdot 10^{-06}$ whereas for φ_2^*, there is a correlation of $\rho_{\varphi_1^*} = 0.8823529$ with a p-value $< 2.2 \cdot 10^{-16}$.

Looking at the results more in particular, we may notice that, on one hand they tend to

overestimate the changes in the glycolysis pathway, while on the other hand the increased fluxes in the pentose phosphate pathway (vG6PDH, vPGDH, vRu5P, vR5PI, vTKa, vTKb, vTA) are underestimated. Hence, part of the redirection towards the pentose phosphate pathway (PPP) was not retrieved. This might be due to an enzyme saturation in the original kinetic model that cannot be captured in our formulation.

Nevertheless, compared to glycolysis, the changes in the PPP are smaller, and to the exception of *Ribulose-Phosphate epimerase* (RU5p) in the second solution, we can consider the results satisfying as the global behaviour of the system was retrieved. Indeed, the increases in PPP and glycolysis are both clear, and the vector φ does correlate with the simulated reaction responses to the pulse of glucose.

Of course, this remains a small example as only part of the metabolism was represented. We thus decided to realise the same type of analysis using a novel kinetic model published by (Khodayari et al., 2014). Such model was fitted using the measurements from different knock-out experiments. In this case, we met with some difficulties to reproduce the data as is discussed in the next section.

3.5 Prediction of knock-out behaviours in *E. coli*

(Khodayari et al., 2014) proposed a larger model of *E. coli* with 91 metabolites, 138 reactions and 60 known regulations. The authors used the ENSEMBLE MODELLING framework proposed by (Tran et al., 2008) which is described in Chapter 1. The overall network is presented in Figure 3.8.

The novelty here is that the kinetic parameters are estimated such that they reproduce the behaviour of seven mutant strains and of the wild-type. The authors proposed the following procedure. Several sets of parameters are estimated such that they fit the fluxes measured by (Ishii et al., 2007) for the wild-type strain. Then, for each mutant strain, the authors simulate the possible fluxes and compare them to the *in vivo* experiments such that the generated sets of parameters converge to the experimental measures. This convergence is obtained using two successive optimisation procedures implemented as genetic algorithms.

This model and the associated procedure appear really promising as the model is larger than the ones previously published and it can reproduce the behaviour of several strains. We thus decided to test our framework on this larger instance.

As will be explained later on, some reproducibility problems appeared while trying to simulate the data. After discussion with the authors, we simulated one dataset, that corresponds to the knock-out of the GND (6-phosphogluconate dehydrogenase) reaction. We are currently running the simulations, but the results are not yet available. We present for now the procedure followed to simulate the data and discuss the discrepancies between the published paper and our simulations.

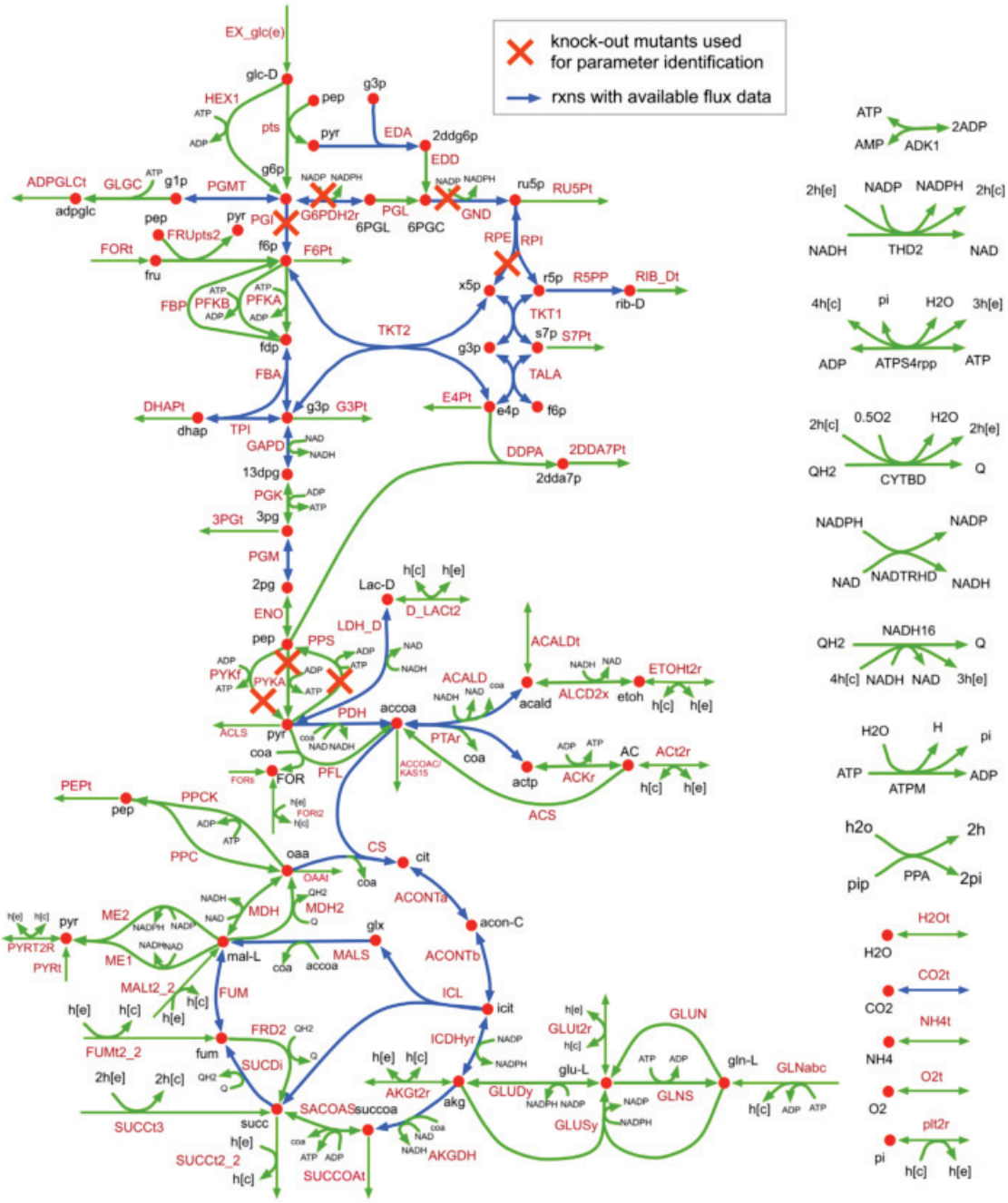


Fig. 1. (Colour online) The constructed kinetic model of *E. coli* core metabolism.

Figure 3.8: From (Khodayari et al., 2014), the authors reported that 93 metabolites were simulated. As there were no experimental conditions with fructose as a substrate, and as glucose was directly imported through the known phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS), those two metabolites were excluded in the final model. The simulated metabolic network contains 91 metabolites. Metabolite names and formulas can be found in Appendix A, Table A.6 along with information on the reactions (Table A.7).

3.5.1 Model reproducibility

The kinetic model presented by (Khodayari et al., 2014) is available along with implementation of the ordinary differential equations (ODEs) simulating the dynamic evolutions from the wild-type to an enzyme knock-out. This dynamics is not based on biological experiments. However, it

simulates the possible flux reorganisation from a wild-type strain (that will be our first condition) to a mutant strain (that will be the second condition).

Starting from $t_0 = 0$, the Matlab code solves the ODEs and thus simulates the transient state. Once a new steady-state is reached (where the metabolite concentrations do not vary anymore), the framework stops and we can thus consider that it reached the final state, that is t_f.

In the ENSEMBLE MODELLING framework, metabolite concentrations are normalised by the wild-type (WT) concentrations. The model thus reports fold-change concentrations that can be transformed into absolute concentrations if the wild-type ones are known. This is the case for 23 metabolites (taken from (Ishii et al., 2007) and (Bennett et al., 2009)). When no measurements were available, (Khodayari et al., 2014) used the range proposed by (Feist et al., 2007), namely $[0.01, 20]$mM.

Using the Matlab code available, we could simulate the pseudo-transient states that the organism would supposedly undergo. For now, we tested the GND mutant predictions. The gene GND encodes the *6-phosphogluconate dehydrogenase* enzyme that converts 6-phospho-D-gluconate into D-ribulose 5-phosphate. This reaction uses as coenzyme Nicotinamide adenine dinucleotide phosphate (NADP+) that is converted into reduced nicotinamide adenine dinucleotide phosphate (NADPH). The two other coproducts are carbon dioxyde (CO2) and hydron (H+), and the reaction is annotated as reversible.

We retrieved the new steady-state fluxes and metabolite pools.

The procedure suggested by the authors of the model is to use the mean of the concentration intervals to obtain an estimation of the mutant metabolite concentrations. As mentioned previously, we take into account the variation in the concentrations in a same condition using intervals. In this case, we consider that the concentration intervals for the mutant are obtained using the fold-change of the simulation and the wild-type concentration intervals as shown in Equation 3.7:

$$[FC_i \cdot WT_{\min,i},\ FC_i \cdot WT_{\max,i}] \tag{3.7}$$

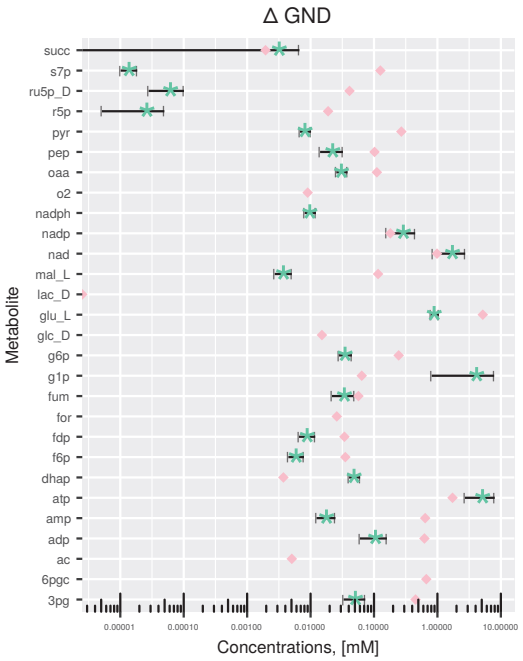
where for a metabolite i, FC_i is the computed fold-change, $WT_{\min,i}$ is the lower limit of the wild-type metabolite concentration, and $WT_{\max,i}$ is the upper limit of the wild-type metabolite concentration. FC_i is obtained from our simulations, and the WT concentrations were obtained from the authors.

We present the results of our simulation in Table 3.7.

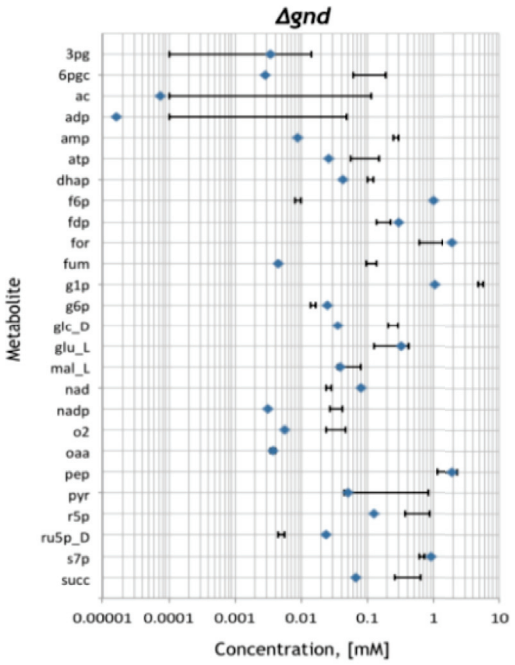
Name	WT concentrations (mM)	Fold change	Mutant GND concentrations (mM)
3pg	[0.317, 0.696]	0.1047	[0.0332, 0.0728]
adp	[0.307, 0.818]	0.1952	[0.0599, 0.1597]
amp	[0.382, 0.76]	0.0325	[0.0124, 0.0247]
atp	[0.623, 1.84]	4.3525	[2.7116, 8.0086]
dhap	[0.101, 0.151]	0.4004	[0.0404, 0.0605]
f6p	[0.0318, 0.056]	0.1405	[0.0045, 0.0079]
fdp	[0.0185, 0.0334]	0.3545	[0.0066, 0.0118]
fum	[0.035, 0.0793]	0.6209	[0.0217, 0.0492]
g1p	[0.0119, 0.115]	68.1507	[0.811, 7.8373]
g6p	[0.139, 0.223]	0.2013	[0.028, 0.0449]
glu_L	[2.97, 3.97]	0.2661	[0.7902, 1.0563]
mal_L	[0.048, 0.0893]	0.0567	[0.0027, 0.0051]
nad	[0.343, 1.1]	2.4883	[0.8535, 2.7371]
nadp	[0.046, 0.129]	3.4543	[0.1589, 0.4456]
nadph	[0.13, 0.196]	0.062	[0.0081, 0.0122]
oaa	[0.024, 0.0361]	1.061	[0.0255, 0.0383]
pep	[0.0729, 0.167]	0.1939	[0.0141, 0.0324]
pyr	[0.103, 0.154]	0.0661	[0.0068, 0.0102]
r5p	[0.0068, 0.0657]	$8 \cdot 10^{-4}$	[0, 0]
ru5p_D	[0.0555, 0.2]	$5 \cdot 10^{-4}$	$[0, 1 \cdot 10^{-4}]$
s7p	[0.157, 0.286]	$1 \cdot 10^{-4}$	[0, 0]
succ	[0, 0.022]	0.3033	[0, 0.0067]

Table 3.7: Known metabolite concentrations for the 23 metabolites of the model. As explained in the main text, the WT concentration intervals are extracted from the literature. The fold change was obtained from the simulation and represents the value $\frac{GND_i}{WT_i}$, where WT_i is the concentration of the metabolite i for the wild-type and GND_i is the concentration of the metabolite i for the mutant once the new steady-state is reached. The mutant concentration intervals are thus obtained by multiplying the fold change by the WT concentrations. The remaining of the metabolites are considered to have a concentration in the interval proposed by (Feist et al., 2007) for the wild-type, that is $[0.01, 20]mM$. All values are rounded to the 4th digits.

Currently, the simulated concentrations do not correspond to the ones published (see Figure 3.9). The authors informed us that some of their figures were mislabelled and following our request, they sent the results obtained for the publication. When comparing our results, the published figures and the published data, we may notice that they are not coherent. Here, we only focused on the metabolites actually measured since for the others, the interval proposed by (Feist et al., 2007) is wide ($[0.01, 20]mM$). As we directly used the kinetic model (already parameterised), there are no heuristics that could lead to such divergent results. Moreover, the observed differences cannot be explained by rounding errors during integration of the ODEs as they are quite large. For now then, such differences are not expected nor explained. This raised doubts on the published model. However, the authors confirmed that the results presented are the ones that they currently obtain with the model.



(a) Results obtained using the model provided in comparison to the results sent by the authors that were obtained at the time of their publication.



(b) From (Khodayari et al., 2014), concentrations provided in the publication. Image obtained from the Supplementary Material.

Figure 3.9: Results obtained using the model provided and results obtained by the authors. In Figure 3.9a, the intervals and blue stars are our simulation results whereas the pink diamonds were simulated by the authors. As mentioned, we only computed the concentrations for the metabolites with WT measurements. We can notice that the intervals rarely superpose with the diamond points, and thus that the concentrations provided by the authors do not correspond to the ones simulated. In Figure 3.9b, the figure presented in (Khodayari et al., 2014). The intervals here are the measured concentrations *in vivo* by (Ishii et al., 2007) whereas the blue diamonds are the computed mean concentrations. After discussion with the authors, they informed us that the labelling of the figure was wrongly indicated in reverse order. For example, the metabolite concentration of 3-phosphoglycerate (which labels the first row of the figure) is in fact presented in the last row of the figure (which is labelled as succinate), while, inversely, the metabolite concentration of succinate is presented in the first row.

Not managing to reproduce the results of the paper is troublesome. This is also worrying as the authors do not manage either to reproduce their published figures. One may wonder then what would be the biological meaning of the results we could obtain. However, as previously, we will here compare the results we get with the transient state simulation, that is the integration of the ODEs providing certain rates and concentrations along time. The coupling between concentrations and reaction rates should thus be maintained at least from a theoretical point of view. We observe however that later on, we would need a more biologically convincing model for the knock-out mutants. Such datasets are difficult to obtain. Indeed, to our knowledge, few *in vivo* measurements of metabolite pools both in a wild-type and a mutant cultivated in similar condition exist. Moreover, the use of kinetic models is not always straightforward as shown here.

3.5.2 Preparing the dataset

Once the simulation performed, the concentrations and transient fluxes can be obtained.

In this case, as previously for the smallest example we used (Section 3.4), we can define the variations $\Delta^{\max}$ and $\Delta^{\min}$ for the metabolites of the model, and add boundary metabolites for the uptake and export reactions.

We have here precise intervals for 22 out of the 91 metabolites. For the ones where no *in vivo* measurements are available, we define the possible variation using an artificially large interval. If the reported fold-change is higher than 1, then $\Delta_X^{\min} = \alpha$ and $\Delta_X^{\max} = \beta$. Otherwise $\Delta_X^{\min} = -\beta$ and $\Delta_X^{\max} = -\alpha$. As previously, 34 boundary metabolites were defined, together with their possible intervals using α and β. This is summarised in Equation 3.8:

$$\begin{aligned} \Delta_{X_i}^{\min} &= \begin{cases} \alpha, & \text{if } \mathrm{FC}_i \geq 1 \text{ or } (s_{ij} \cdot \mathrm{AUC}_j) \geq 0 \\ -\beta, & \text{otherwise} \end{cases} \\ \text{and } \Delta_{X_i}^{\max} &= \begin{cases} -\beta, & \text{if } \mathrm{FC}_i < 1 \text{ or } (s_{ij} \cdot \mathrm{AUC}_j) < 0 \\ -\alpha, & \text{otherwise} \end{cases} \end{aligned} \tag{3.8}$$

where FC_i stands for the Fold-Change of the metabolite i, AUC_j for the Area Under the Curve of the rate of reaction j and s_{ij} is the stoichiometric coefficient associated to the metabolite i and the reaction j.

Fold change is thus used for the internal metabolites while AUC is used for the boundary metabolites and is obtained by approximating the integral of the reaction rates that were used as a sink or a source. The complete table of metabolites is provided in the Appendix A, Tables A.4 and A.5.

Here, we used $\alpha = 10^{-4}$ and $\beta = 10^{5}$. The intervals of variation for the measured metabolites are presented in Table 3.8.

The SBML network was established using the data obtained from the model in Matlab. In (Khodayari et al., 2014), isoenzymes were implemented as several reactions since fluxomics data were available. As mentioned previously, φ indicates if an excess or decrease of matter occurred between two sets of metabolites during the transient state. Thus, if two reactions in the kinetic model share the same substrates and the same products, they will create two solutions that will just offer a choice on which reaction was regulated. Similarly, if a reaction has for products (resp. substrates) the substrates (resp. products) of another, this will also create two solutions with opposite φ_j from those two reactions. The computed φ will thus be symmetrical. To reduce the solution space, and since we cannot distinguish those reactions, we decided to group them in order to treat them afterwards. We thus reduce the following pairs of reactions into one:

- *6-phosphofructokinase 1* (pfkA) and *6-phosphofructokinase 2* (pfkB);
- 2 reactions encoding for *6-phosphogluconolactonase* (pgl and pgl_spoon);
- *pyruvate kinase I* (pykA) and *pyruvate kinase II* (pykF).

Name	Δ^{min}	Δ^{max}
3pg	-0.6628	-0.2442
adp	-0.7581	-0.1473
amp	-0.7476	-0.3573
atp	0.8716	7.3856
dhap	-0.1106	-0.0405
f6p	-0.0515	-0.0239
fdp	-0.0268	-0.0067
fum	-0.0576	0.0142
g1p	0.696	7.8254
g6p	-0.195	-0.0941
glu_L	-3.1798	-1.9137
mal_L	-0.0866	-0.0429
nad	-0.2465	2.3941
nadp	0.0299	0.3996
nadph	-0.1879	-0.1178
oaa	-0.0106	0.0143
pep	-0.1529	-0.0405
pyr	-0.1472	-0.0928
r5p	-0.0657	-0.0068
ru5p_D	-0.2	-0.0554
s7p	-0.286	-0.157
succ	-0.022	0.0067

Table 3.8: Computed variation intervals for the subset of metabolites with measurements from (Ishii et al., 2007) and (Bennett et al., 2009).

Our final network contains 134 reactions and 126 metabolites including 32 boundary ones.

As mentioned in the introduction to this chapter, we are currently expecting the results of the simulations. Some doubt however remains here, more particularly because of the discrepancies between the published results and the ones obtained from the model provided by the authors. Several metabolites exhibit a fold change larger than 100 and were not reported in (Khodayari et al., 2014). One may wonder whether they are reliable results, or are due to a robustness issue.

Finally, in this case we defined large intervals of variation using the parameters α and β. First, one can see how semi-quantitative methods could be applicable here, as we just take into account for certain metabolites the known direction of change. Second, this parameterisation led to define large intervals and slowed the computation. Narrower intervals when possible should thus always be preferred.

3.6 Discussion

The method proposed in this chapter is not a replacement of a kinetic modelling of metabolism. Indeed, our desire here is to infer some important modifications from a global point of view using a small set of information. This global point of view is therefore not driven towards a specific pathway, apart from the fact that the measured metabolites do circumvent the solutions obtained to a certain subnetwork. Furthermore, some behaviours cannot be depicted, such as enzyme saturation that would lead to the usage of a different reaction as shown in the second example (Section 3.4). Nevertheless, if such saturation resulted in new metabolite pools, then this method could propose points of bifurcation, *i.e.* possible alternative pathways.

We therefore showed that without any kinetic information, we could retrieve the possible overall variations of the reaction rates during the transient state. Here, we compared with the simulations of the fluxes obtained. However, *in vivo* measurements of the reactions during a transient state are rarely available. The ambition in this case was to be able to point towards putatively modulated reactions that are responsible for the observed changes in metabolite pools. However, we do not infer the possible regulators, allosteric or transcriptomic.

As in Chapter 2, incoherence in the model can suggest to perform a new measurement of the intracellular metabolites. Indeed, if the system is not satisfiable, *i.e.* if the solver cannot find a solution, it is possible to see which constraint is not respected. Such constraint would indicate a metabolite whose variation interval is not coherent with the remaining of the measurements. It is possible to allow for some unknown but suspected metabolite variations by constraining their upper and lower bounds. This is required particularly for reactions that are at the limit of the system. Those can be export and uptake reactions, but also boundaries of the small molecule metabolism that model other cell functions. Indeed, as we saw in Chapter 2, such reactions are of importance when there is for example an increase in glycogen storage, a change in the protein synthesis rate, or a change in the growth rate of an organism.

Finally, as a perspective, one can constrain the known reactions using thermodynamic rules since the metabolite concentrations are known in both states. Indeed, the Gibbs free energy of a reaction ΔG (mentioned in Chapter 1, Section 1.2.2) depends on the concentration of its substrates and products. According to the second law of thermodynamics, the direction of a reaction is such that its Gibbs energy is negative. Such thermodynamic laws have been used previously to constrain the solutions of flux balance analysis (Hoppe et al., 2007). In our current work, some reversible reactions might be proposed in one direction only. This would not restrain the number of solutions, but instead the interpretation of a solution, more particularly of the vector φ.

We also met here with the difficulty of evaluating such theory. We presented the method using simulated data from (Chassagnole et al., 2002). This model uses different kinetic laws and takes into account known regulations, thus providing an interesting non-linear behaviour to them. However, this model remains small to test the method proposed. Furthermore, we are limited by the boundaries of the system as mentioned previously. Moreover, since we worked with simulated data, all the metabolite variations were known. Thus a larger model could help evaluate the impact of missing measurements and unknown behaviour of the boundaries.

A more recent kinetic model was available from (Khodayari et al., 2014). However, even after an extensive usage and several exchanges with the authors, the results we obtained did not reproduce those from the published paper accompanying the model. This is a problematic issue

that I suppose will continue to be present in research. The importance of the reproducibility of published results and of the availability of the developed kinetic models must be emphasised.

Whereas this method presents a simple way of inferring a transient response to an environmental or to a genetic change, obtaining data of sufficient quality remains an issue, and a continuous dialogue with experimentalists is important to test *in vivo* the proposed hypotheses.

Moreover, the objective function we adopted should be discussed. Indeed, minimising the sum of the absolute values of the vector φ rests on the hypothesis of a parsimonious response meaning that most of the metabolism would be synchronised in its variations. One may think here that we are using the limiting step hypothesis, namely that one reaction only of a pathway is regulated. It is not the case. Since we rely on measured metabolite concentrations, a perfectly synchronised regulation of pathways cannot be captured as the metabolite concentrations would not change. However, if several concentrations changed, the method will provide information on the asynchrony of the reaction rates. As for the steady-state, synchronisation is relative to the time-scale observed. Metabolite shifts in the case of growth limiting conditions have been reported to be quite rapid. Also here, we infer a parsimonious regulation but as presented in Chapter 1, metabolic regulation can be done at several scales: genomic, transcriptomic, or allosteric. A parsimonious regulation could also be done at the genomic scale for example, by acting on a regulator that itself will act on several enzymes, meaning that several reactions would be impacted. This is not taken into account in the model for the moment. It would therefore be interesting to obtain a multi-level method that takes into account genomic, proteomic, fluxomic and/or enzymatic data. Initial attempts in that direction have been published, for example by (Ryll et al., 2014) who link signalling and gene expression data with metabolism through the use of ordinary differential equations. Such approaches remain restricted to small biological examples as many parameters are required. Moreover, the authors emphasised the fact that there is a lack of a standardised exchange format and notations that prevent the creation of an automated data integration framework. Providing such framework to interpret possible metabolic shifts at those different layers would be a huge step forward. Finally, in Chapter 2, we presented a method that addresses the same type of problem, which is to identify the flux reorganisation when an organism is confronted to a change of conditions. One could therefore propose a mixed approach using both quantitative and qualitative measurements. A first such approach can be performed by using a large interval of variations as presented in Section 3.5.2 for the non measured metabolites.

Chapter 4

Communities and synthetic biology

Contents

4.1 Introduction

During this PhD, I also focused on the use of communities for the production of compounds of interest. As interactions between microorganisms are ubiquitous and span over several levels of associations (see Chapter 1), recent research has focused on such communities, whether from an ecological point of view with the objective to understand the evolution of niches, or from a medical point of view through the study of the microbiome. As for single organisms, novel ways of engineering communities are now possible besides only understanding them. In this chapter, we proposed to infer these communities, whether the association between the organisms involved exists already in nature or is artificially established, as well as the possible metabolic roads to produce one or several target compounds from specific substrates. As we took into account the possibility of genetic modifications, this work falls within the scope of synthetic biology. The results presented in this chapter were published in (Julien-Laferrière et al., 2016).

Synthetic biology has been defined by the European Commission as "the application of science, technology, and engineering to facilitate and accelerate the design, manufacture, and/or modification of genetic materials in living organisms to alter living or nonliving materials". It is a field that has boomed since the early 2000s when in particular Jay Keasling showed that it was possible to efficiently synthetise a compound – artemisinic acid – which after a few more tricks then leads to an effective anti-malaria drug, artemisinin (Ro et al., 2006). Such chemical compounds were naturally produced only in the plant *Artemisia annua*, a type of wormwood, in quantities too small to enable a cheap production of the drug. To address this problem, a living organism, *Saccharomyces cerevisiae*, was used for such rapid, and therefore much more effective synthetic production. Since the work by J. Keasling, many other species, in particular bacteria, have also been manipulated with a similar objective of more efficiently producing some compounds of interest for health, environmental or industrial purposes.

However, engineering a single microorganism to optimise one or a collection of metabolic tasks may often lead to considerable difficulties in terms either of getting an efficient production system, or of avoiding toxic effects for the recruited microorganism (Bernstein and Carlson, 2012). The idea of using a microbial consortium has thus started being developed in the last decade (Bernstein and Carlson, 2012; Brenner et al., 2008; Momeni et al., 2011; Sabra et al., 2010). This can indeed allow to perform more complex tasks, for example by splitting the work between the members of the consortium, by alleviating an inhibition due to toxic compounds as we show later, or even by obtaining a culture more resistant to environmental changes. Microorganisms may thus be more efficient synthetic factory workers as a group than as individual species, as already shown for problems related to remediation or energy (Masset et al., 2012; Bourdakos et al., 2014; Mnif et al., 2015). However, difficulties may arise, limiting or preventing the success of such community approaches (Timan, 1982; Oliveira et al., 2014). Finally, selecting the members of the consortium to produce one or several compounds remains a challenge (Bernstein and Carlson, 2012).

Two types of consortia are studied. The first is a synthetic consortium of strains carrying genetic and/or regulatory modifications. This follows the same spirit as in the work of Jay Keasling for the production of artemisinic acid. In our case, the goal is the synthetic production of two bioactive compounds with antibacterial properties: penicillin and cephalosporin C. Four microorganisms were considered for such production. Notice that already an important question is whether the best option is to use all four in the consortium, or instead only a subset thereof,

and of course, which subset is then most efficient. In this first case study, the compounds of interest are exogenous to the consortium.

In the second case study addressed, the two microorganisms form an artificial consortium in the sense that the species involved in it do not naturally interact, and both organisms are able to endogenously produce the target compounds. One of these is 1,3-propanediol (PDO), a building block of polymers. Associating microorganisms in a consortium can lead to a better yield of production as already demonstrated by (Bizukojc et al., 2010). This however is not the only consortium that may be considered.

In both cases, it is necessary to infer the transfer of metabolites from one organism to another and, if the compounds are exogenous to the selected organisms, which reactions need to be inserted in the consortium. For such problems, computer models are crucial in providing hints on how to best divide a given metabolic production line among different organisms that are then made to interact with one another.

Various methods exist that enable to better understand the metabolic capabilities and the interactions observed in natural communities (Zomorrodi et al., 2014; Zomorrodi and Maranas, 2012), but they do not take into consideration the production of specific products from selected substrates. This issue was addressed more recently by (Eng and Borenstein, 2016) while minimising the number of species in the community. In this PhD work, we present a different model to solve both biological cases considered above that attempts to strike a balance between the exchanges that would be required among the species involved in the consortium and the genetic modifications that would be needed. To this purpose, we use a weighted network, thus assigning a priority of use to some reactions over others. This enables on one hand to either favour or on the contrary, disfavour a transport reaction, and on the other hand to reflect the difficulties associated with inserting exogenous genes. Indeed, the problem of obtaining an optimal consortium includes at least the following two parallel objectives: one is to have a small number of reactions exogenous to the consortium that need to be added to it, the second is to have a small number of compounds that need to be transported across different species of the consortium. Both are indeed costly and should thus be avoided whenever possible. Other aspects would also need to be taken into consideration, such as the efficiency of the consortium in terms of both survival and growth of each species composing it, as well as of production of the compounds of interest. Here, we address only the first two objectives of minimising the number of insertions of exogenous reactions and of transitions. Our approach is purely combinatorial and topological. We do not take into account stoichiometry for the moment. This approach however represents a first step that, as we show, leads already to a hard problem. We start by some preliminaries that present the basic notations and definitions used, the model adopted, and a formal description of the problem addressed. Following the idea initially introduced by (Fellows and Rosamond, 2007; Fellows et al., 2009), we then explore how different parameters of the problem and combinations thereof influence its complexity. We propose an initial algorithm, MultiPus (MULTIple species for the synthetic Production of Useful biochemical Substances), for addressing this problem. Because of an increasing running time on genome-scale metabolic models (GEMs) for the first version based on dynamic programming, MultiPus is also available using an *Answer Set Programming* (ASP) solver (Gebser et al., 2014a) which is more efficient in general. Finally, we present the two production cases explored with MultiPus.

4.2 Preliminaries

4.2.1 Notations and basic definitions

We work with a directed hypergraph representation of a metabolic network, using genome-scale metabolic models (GEMs). Let then $\mathcal{H}$ be a directed hypergraph defined on a set of vertices, denoted by $\mathcal{V}$, that corresponds to the compounds, and a set of directed hyperedges, that is of *hyperarcs*, denoted by $\mathcal{A}$, that corresponds to the reactions. Given a hyperarc a, we denote by $\mathsf{src}(a)$ and $\mathsf{tgt}(a)$ the sets of source and target vertices of a, respectively, that is the set of substrates and of products. In the problem described below, the main issue comes from the hyperarcs with multiple source vertices. The possible multiplicity of the target vertices of a hyperarc does not affect the complexity of the problem. Moreover, we can, without loss of information, decompose such hyperarcs into ones that each have the same set of source vertices but only one of the target vertices of the original hyperarc (as explained in Section 4.2.2). We therefore make this assumption from now on.

For a subset of hyperarcs $\mathcal{A}' \subseteq \mathcal{A}$, $V(\mathcal{A}')$ denotes the set of vertices that are involved in at least one hyperarc of $\mathcal{A}'$, that is the set of compounds that participate in at least one of the reactions represented by $\mathcal{A}'$. By abuse of notation, given a set of hyperarcs $\mathcal{A}'$, we often refer to the hypergraph $(V(\mathcal{A}'), \mathcal{A}')$ simply as $\mathcal{A}'$.

Since a reaction needs all its substrates to be activated, we consider that the multiple source vertices of a hyperarc correspond to a multiplicity of tentacles (often used for grasping), each associated to one substrate. A hyperarc is therefore like an octopus, only with a number of tentacles that may be different from eight. The greater the number of tentacles, the more tentacular is the hyperarc considered to be.

We formally introduce the notion of a *tentacular* hyperarc as follows.

Definition 14. *A hyperarc a is called* tentacular *with* number of tentacles*, or* spreadness *for short, b if $b = |\mathsf{src}(a)| > 1$.*

Finally, we define the notion of the *total number of tentacles*, *total spreadness* for short, of a directed hypergraph.

Definition 15. *Given a directed hypergraph $\mathcal{H}(\mathcal{V}, \mathcal{A})$, its total spreadness is the sum of the number of sources of the tentacular hyperacs in $\mathcal{H}$.*

For the sake of simplicity, we will use the term *arc* to refer to non tentacular hyperarcs. It will later become clear why we need to consider the total spreadness of the input.

4.2.2 Model adopted

We recall that the problem we want to address concerns the production by a consortium of organisms, microbes for instance, of a set of compounds denoted by T. The compounds of interest may not be produced naturally by the members of the consortium. Indeed, they could instead be produced by other organisms (in the example given in the introduction, this is a plant). We denote these two sets by, respectively, O_w (the workers to be used to synthetically produce the compounds in X) and O_o (those other organisms, used as reference, where the compounds in T are naturally produced). As indicated, we may have $|O_o| = 0$ meaning here that the workers are naturally able to produce the compounds.

Let $N_1, \ldots, N_k$ be the genome-scale metabolic models (GEMs) for the organisms in O_w, and let $V_1, \ldots, V_k$ respectively correspond to the sets of vertices in these networks. Actually, this is a superset of the consortium that may really be required for the production of T and that will be a solution of the problem as defined below. The hyperarcs in N_i have weight w_{worker}, independently of i.

Typically w_{worker} will be set equal to zero, or to a value that is close to zero for reasons that will be explained later, in Section 4.5. The set of hyperarcs in the metabolic models for O_o is denoted by A_o.

The directed hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{A})$ that is the input to our problem is constructed in the following way.

First, we perform the disjoint union of the networks $N_1, \ldots, N_k$. Let $\mathcal{H}$ be such that $\mathcal{H} = N_1 \cup \ldots \cup N_k$. Thus for now $\mathcal{V} = \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_k$ and $\mathcal{A} = \mathcal{A}_1 \cup \ldots \cup \mathcal{A}_k$. Then, for each network N_i, and for each hyperarc $a \in \mathcal{A}_o$ that corresponds to a reaction not already in N_i, we create a copy of it in N_i, and thus in $\mathcal{H}$. We add the hyperarc a labelled as a_i to $\mathcal{A}_i$. We further add to V_i, and thus to V any vertex corresponding to a compound not already in N_i if such exists. The added hyperarc has weight w_{other}. Typically, $w_{other} > w_{worker}$: introducing a reaction in the metabolism of an organism that does not contain the corresponding enzyme(s) is indeed costly. Finally, for each pair of vertices $v_i \in V_i$ and $v_j \in V_j$ with $i, j \leq k$ and $i \neq j$ such that the corresponding compound is the same, we create a hyperarc that has v_i for single source and v_j for single target (it therefore is an arc) and has weight $w_{transition}$. Typically, we will have that $w_{transition} > w_{worker}$: making a transition from one organism of the synthetic consortium to another, which implies transporting a compound, is also costly.

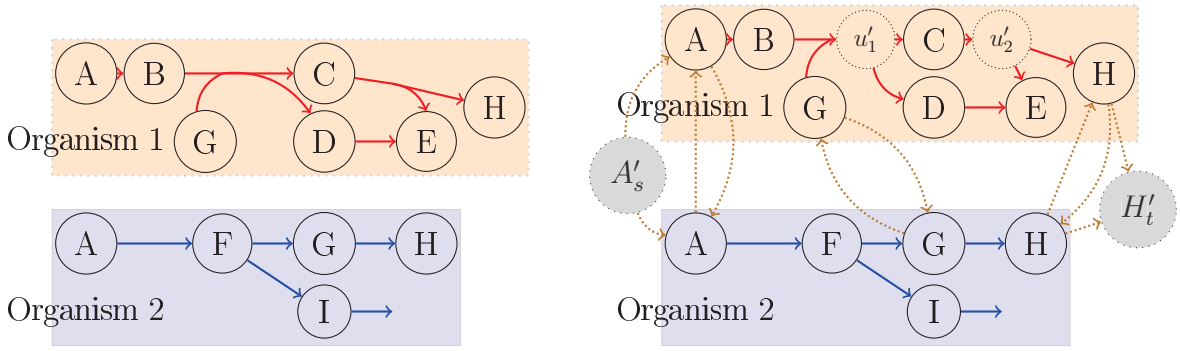
The production of the target compounds (in T) is not mandatory in all workers. Indeed, the objective here is to have at least one organism producing it. The same reasoning applies for the source set. We thus create pseudo-sources and pseudo-targets that will be linked to the remaining of the network.

Every target $t_{i,j}$ corresponding to the metabolite i of an organism/network j is connected to a pseudo-target t'_i by an arc $((t_{i,j}, t'_i)$ with a weight that is negligible compared to the other (regular) weights used. Reaching t'_i guarantees that at least one $t_{i,j}$ is reached. The same procedure is applied for the sources.

Finally, we only want to take into account the hyperarcs of spreadness > 0 (*i.e.* such that $|\mathsf{src}(a)| > 1$), then all the hyperarcs of spreadness 0 are replaced by simple arcs in the network. Hence, for all $a \in \mathcal{A}$ such that $|\mathsf{src}(a)| = 1$ and $|\mathsf{tgt}(a)| > 1$, a vertex that will play the role of a pseudo-metabolite u_a is added to the network. The hyperarc selected is then removed and some arcs are added, one going from the tail vertex (substrate) to the pseudo-metabolite and the others from the pseudo-metabolite to the head vertices (products). To summarise, one hyperarc (the original one) is removed and $1 + |tgt(a)|$ simple arcs are added: the arc $(\mathsf{src}(a), u_a)$ and for each $v \in \mathsf{tgt}(a)$, an arc (u_a, v).

An example of the steps described above is depicted in Figure **??**.

It is worth calling attention to the fact that we are considering here that adding a reaction from O_o to an organism from the consortium O_w (when such operation is required) implies a cost that does not depend on the reaction. Similarly, we are considering that any transition from one organism in O_w to another is equally costly. These assumptions may however be refined by making such costs, and thus the weights of the added hyperarcs (tentacular hyperarcs or arcs) depend on the reaction or on the transition (see later for a further discussion on this).



(a) Original networks, the two coloured squares represent two different organisms.

(b) Network after addition of the pseudo-source and pseudo target for A and H and of the transitions.

Figure 4.1: Toy example for the transition reactions and addition of pseudo-sources and pseudo-targets. In Figure 4.1b, the pseudo-source for A and pseudo target for H were added. Also six transitions are present since both networks (organisms) share the vertices A, G and H. Moreover, the hyperarcs $C \to E + H$ and $B + G \to C + D$ were decomposed.

4.2.3 Problem definition

We first introduce the notion of a directed rooted hypergraph.

Definition 16. *A directed hypergraph $\mathcal{H}' = (\mathcal{V}', \mathcal{A}')$ is rooted at $S \subseteq \mathcal{V}'$ if there exists an ordering of its hyperarcs $(a_1, ..., a_m)$ such that for all $i \leq m$,* $\mathsf{src}(a_i) \subseteq S \cup \mathsf{tgt}(\{a_1, \ldots, a_{i-1}\})$.

The problem that we address here is then defined as follows:

Directed Steiner Hypertree (DSH) problem
Input: A weighted directed hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{A}, w)$ where w is the set of weights associated to the hyperarcs in $\mathcal{A}$, a set of sources S and a set of targets T.
Output: A directed hypergraph $\mathcal{H}' = (\mathcal{V}', \mathcal{A}')$ rooted at S, with $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{A}' \subseteq \mathcal{A}$, of minimum weight such that $T \subseteq \mathsf{tgt}(\mathcal{A}')$.

Notice that the term Directed Steiner Hypertree abuses language in the sense that there may be more than one root. In the case of directed graphs (digraphs), it would correspond to a set of trees, hence to a forest.

4.2.4 Relation to known problems

If the directed hypergraph is a digraph, then it is a minimal directed hypergraph rooted at a node s if and only if it is an arborescence rooted at s (*i.e.* a directed tree with an orientation from the root s to the leaves). If there is more than one source, then it is a set of arborescences. In the case of digraphs, the DSH problem coincides with the well-studied Directed Steiner Tree (DST) problem defined as follows:

Directed Steiner Tree (DST) problem
Input: A weighted directed graph $G = (V, A, w)$ where w is the set of weights associated to the arcs in A, a source s and a set of targets T.
Output: A subset A' of A of minimum weight such that $T \subseteq \mathsf{closure}_{A'}(s)$.

The closure operation is defined as follows: Given a directed hypergraph $\mathcal{H} = (V, A)$, a set of vertices X such that $X \subseteq V$ and a set of hyperarcs A' such that $A' \subseteq A$, $\mathsf{closure}_{A'}(X)$ is the smallest set $C \subseteq V$ such that $X \subseteq C$ and for each $a \in A'$, if $\mathsf{src}(a) \subseteq C$, then $\mathsf{tgt}(a) \subseteq C$.

Intuitively, $\mathsf{closure}_{A'}(X)$ is the set of vertices that can be reached from X following the hyperarcs in A'. In the context of metabolic networks, it is the set of compounds that the reactions from A' can produce using only the compounds of X as sources.

4.2.5 Complexity of the problem

We start by investigating the complexity of the problem. We first observe that the Directed Steiner Tree problem is NP-hard (Garey and Johnson, 1979). The Directed Steiner Hypertree problem is also NP-hard, even on graphs, indicating that it is highly unlikely that there exists an efficient (polynomial time) delay algorithm for its solution. However, if the number of targets is considered a constant, then there exists an algorithm with polynomial running time. DST is said to be Fixed Parameter Tractable (FPT) with the number of targets as parameter. This implies that DSH also admits an FPT algorithm for a constant number of targets in the case where the input is a directed graph.

In the general setting however, Proposition 1 indicates that the problem is doomed to be intractable when using only parameters related to the solution size. The proofs of the propositions are available in Appendix B and in Figures B.1 and B.2.

Proposition 1. *The problem is W[1]-hard when parameterised by any combination of:* $|A'|$, $\mathsf{weight}(A')$, $|T|$, $|S|$, *total number of tentacles of the hyperarcs in* A'.

Part of the difficulty indeed comes from the choice of tentacular hyperarcs that must belong to the solution. However, taking into account only the number of tentacular hyperarcs in the instance is not sufficient to obtain tractability.

Proposition 2. *The problem is NP-hard even when* $|T| = 1$ *and* A *contains only one tentacular hyperarc.*

Overall, the problem remains intractable when either of these constraints applies to the input: there are few targets, or the total number of tentacles of the hyperarcs is bounded. However, there remains the stronger case when both quantities (number of targets and total number of tentacles of the hyperarcs) are bounded. We present a fixed-parameter tractable algorithm for this case in the next section.

4.3 Algorithm

We now present our main algorithm that exactly solves the Directed Steiner Hypertree problem provided that the number of targets and the total number of tentacles of the hyperarcs remain small. Intuitively speaking, the algorithm identifies the best combinations of tentacular hyperarcs by trying all those in parallel, and for each such combination, it outputs the solutions (if any exists) having minimum weight. More precisely the algorithm enumerates all possible combinations of tentacular hyperarcs that will be used in a solution, where a combination is a subset of the tentacular hyperarcs ordered according to the topological order of the solution (with k tentacular hyperarcs, there are $2^k k!$ such combinations to consider). For each combination, it remains to compute the optimal way of linking these tentacular hyperarcs with regular arcs. This problem is solved by extending the FPT algorithm for the Directed Steiner Tree problem which requires the number of targets as a parameter. In our case, we need the number of targets plus the total number of tentacles of a solution. For a given directed weighted hypergraph $\mathcal{H} = (V, A)$, we denote by $G(\mathcal{H})$ the graph obtained from $\mathcal{H}$ by removing all tentacular hyperarcs. Let $\mathrm{ST}(x, X)$ be the best directed Steiner tree of $G(\mathcal{H})$ rooted in x that has X as set of leaves.

Given an ordered subset $M := (a_1, ..., a_k)$ of the tentacular hyperacs of $\mathcal{H}$, we describe a dynamic programming algorithm to find the best Directed Steiner Hypertree with hyperarc set A' that uses exactly the tentacular hyperarcs of M following their ordering.

The following definitions are illustrated in Figure 4.2. Since all tentacular hyperarcs of M must be used, we have that, for all $i \leq k$, $\mathsf{src}(a_i) \subseteq \mathsf{tgt}(A') \cup S$, and so the set $\mathsf{src}(M)$ can be seen as an additional set of targets. We establish $T' := T \cup \mathsf{src}(M)$ to be the new set of targets, and for $t \in T'$, we define $Layer_T(t) := \min\{i \leq k : t \in \mathsf{src}(a_i)\}$. If $t \in T \setminus \mathsf{src}(M)$, we define $Layer_T(t)$ as $k+1$, and for a subset $X \subseteq T'$, we define $Layer_T(X) := \min\{Layer_T(t) : t \in X\}$. Similarly, since all tentacular hyperarcs of M must be used, intuitively $\mathsf{tgt}(M)$ can be seen as an additional set of sources. We write $S' := S \cup \mathsf{tgt}(M)$ and $Layer_S(s) := \min\{i \leq k : s \in \mathsf{tgt}(a_i)\}$ if $s \in \mathsf{tgt}(M) \setminus S$, and $Layer_S(s) := 0$ if $s \in S$. To respect the ordering of M, the target of a tentacular hyperarc $a_i \in M$ can be used to "reach" only the sources of the tentacular hyperarcs that come after a_i in M. For all $Y \subseteq T'$, we define $S_Y := \{s \in S' \mid Layer_S(s) < Layer_T(Y)\}$.

Observe that for any minimal Directed Steiner Hypertree A', the vertices in $G(A')$ must have in-degree one if they are not in S', and, by minimality, out-degree at least one if they are not in T'.

Given a (directed) forest F, we denote by $V(F)$ and $\mathsf{leaves}(F)$ respectively the vertices and the leaves of all the trees of F. For any vertex t, we denote by $\mathsf{root}(F, t)$ the root of the tree in F containing t when $t \in V(F)$ (the root is the farthest vertex we can reach starting from t by following only branches of F), or $\mathsf{root}(F, t) = t$ otherwise (t is an isolated node).

For a set of targets $Y \subseteq T'$, we say that a forest F of $G(\mathcal{H})$ *covers* Y if $\mathsf{leaves}(F) \subseteq Y$ and $\mathsf{root}(F, t) \in S_t$ for all $t \in Y$.

Lemma 1. *For any optimal solution $\mathcal{A}'$ of the Directed Steiner Hypertree problem given $(\mathcal{H}, S, T)$ as input, if $\mathcal{A}'$ uses exactly the tentacular hyperarcs of an ordered subset M, then $G(\mathcal{A}')$ is a forest covering T'.*

Proof. Consider a Directed Steiner Hypertree $\mathcal{A}'$. First notice that by minimality, $G(\mathcal{A}')$ is a forest. Indeed, if some vertex x has two incoming arcs in A', denoted by a and a', a appearing before a' in $\mathcal{A}'$, then removing arc a' yields a strictly better solution to the Directed Steiner

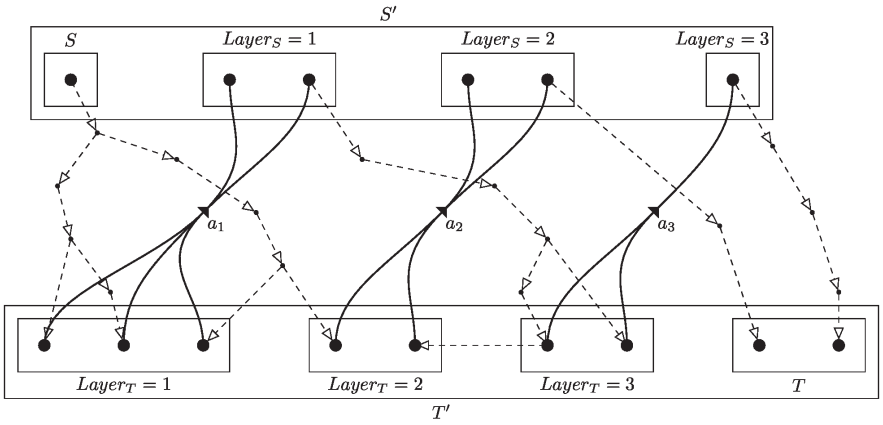


Figure 4.2: Illustration of the notion of layers: Given $M = (a_1, a_2, a_3)$ (thick tentacular hyperarcs), and a solution $\mathcal{A}'$ containing M, $G(\mathcal{A}')$ (dashed arcs) consists of a forest covering all T', *i.e.* each vertex in $t \in T'$ is part of a tree whose source is in a lower "layer" than t.

Hypertree problem. Furthermore, if any $x \notin T'$ is a leaf of $G(\mathcal{A}')$, with incoming arc a, then x is not the head of any arc nor is it part of T'. In this case, a can be deleted and all leaves are in T'.

Consider now any $t \in T'$. Let $s = \mathsf{root}(F, t)$. Consider the path from s to t: by minimality, the arcs of the path must appear in the same order as in $\mathcal{A}'$ (otherwise some arcs must be deleted), and s must appear in the targets of a tentacular hyperarc ordered before any hyperarc of which t is a source (or $s \in S$). This implies that $s \in S_t$. □

Lemma 2. *Given an ordered subset of tentacular hyperarcs M and any forest F covering T', there exists a solution $\mathcal{A}'$ of the Directed Steiner Hypertree problem with $(\mathcal{H}, S, T)$ as input, where $\mathcal{A}'$ uses exactly the tentacular hyperarcs of M in this order, and such that $G(\mathcal{A}') = F$.*

Proof. We build $\mathcal{A}'$ as follows. We first insert the tentacular hyperarcs $(a_1, \ldots, a_k)$ of M, in this order. We then insert the arcs of F between the tentacular hyperarcs, according to the layer of the root of the tree to which they belong. Formally, let $D_1, \ldots, D_p$ be the directed trees in F, and $s_1, \ldots, s_p$ their respective roots. Observe that since all leaves are in T', then each s_i can be written as $\mathsf{root}(F, t)$ for some $t \in T'$, and thus $s_i \in S'$ and $Layer_S(s_i)$ is well-defined and can be computed. For each j, $1 \leq j < k$ (respectively, $j = 0$ or $j = k$), we insert between a_j and a_{j+1} (resp. before a_1 or after a_k), all arcs of all trees D_i such that $Layer_S(s_i) = j$. Within each tree, the arcs are inserted in topological order. There remains to prove that this ordering has the required properties.

We first verify that for any $t \in T$, t is reached by some hyperarc (tentacular or not) of $\mathcal{A}'$. Two cases are possible:

1. If $t = \mathsf{root}(F, t)$ (*i.e.*, either t is the root of some tree of F or $t \notin V(F)$), then $t \in \mathsf{tgt}(a_i)$ for some $a_i \in M \subseteq \mathcal{A}'$, thus $t \in \mathsf{tgt}(\mathcal{A}')$.

2. Otherwise, $t \in \mathsf{tgt}(a)$ for some arc a in F, so $a \in \mathcal{A}'$ and $t \in \mathsf{tgt}(\mathcal{A}')$.

For any vertex $x \in \mathsf{src}(a)$ with $a \in \mathcal{A}'$, we now need to verify that $x \in S$ or x is the target of some hyperarc selected before a. Three cases apply:

1. If $a \in F$ and x is not the root of any tree D_i, then it has an incoming arc appearing in $\mathcal{A}'$ before a (since we kept the topological order of each tree).

2. If $a \in F$ and x is the root of some tree D_i, then $x = s_i$. If $x \notin S$, then $Layer_S(x) > 0$, and x is produced by the tentacular hyperarc $a_{Layer_S(x)}$ which appears before a.

3. If a is not an arc of F, then it is a tentacular hyperarc, $x \in T'$, and $a = a_j$ for some $j > Layer_T(x)$. Let $s_i = \mathsf{root}(F, x)$, then $Layer_S(s_i) + 1 \leq Layer_T(x)$, and the arc producing x is placed before $a_{Layer_S(s_i)+1}$, which in turn is before (or equal to) the arcs $a_{Layer_T(x)}$ and $a = a_j$.

Overall, we indeed have a Directed Steiner Hypertree for $(\mathcal{H}, S, T)$ using M, where, by construction, $G(A') = F$. □

Lemma 3. *For any optimal solution A' of Directed Steiner Hypertree of $(\mathcal{H}, S, T)$, if $\mathcal{A}'$ uses exactly the tentacular hyperarcs of an ordered subset of tentacular hyperarcs M in this order, then $G(\mathcal{A}')$ is a forest covering T' of minimum weight.*

Proof. By Lemma 1, $F = G(\mathcal{A}')$ is already a forest, and it has a total weight of $weight(F) = weight(A') - weight(M)$. Consider any forest F' of weight w' covering T'. By Lemma 2, there exists a solution with weight $weight(F') + weight(M)$, which must be larger than $weight(A')$, hence $w' \geq weight(F)$, *i.e.* F has minimal weight. □

For a set of targets $Y \subseteq T'$, let $\mathrm{SH}_M(Y)$ be the minimum weight of a forest F covering Y under the ordering M. By Lemma 3, the weight of an optimal solution A' of the Directed Steiner Hypertree problem given $(\mathcal{H}, S, T)$ as input is $weight(M) + \mathrm{SH}_M(T')$ where M is the ordered set of tentacular hyperarcs used by $\mathcal{A}'$.

Theorem 1. *The optimal value of an instance of $(\mathcal{H}, S, T)$ of the Directed Steiner Hypertree problem has value $\mathrm{SH}_M(T') + weight(M)$ for some ordering M. Furthermore, SH_M can be computed recursively as follows. For any $Y \subseteq T'$,*

$$\mathrm{SH}_M(Y) = \min(\min\{\mathrm{ST}(s, Y), s \in S_Y\}, \min_{Y' \subset Y}\{\mathrm{SH}_M(Y') + \mathrm{SH}_M(Y \setminus Y')\})$$

Proof. Assume first that the optimal forest F covering Y is a tree and let $s \in S_Y$ be its root. Then $\mathrm{SH}_M(Y) = \mathrm{ST}(s, Y) = \min\{\mathrm{ST}(s, Y), s \in S_Y\}$.

Assume now that F has at least two trees. Let $Y' := \mathsf{leaves}(F_1)$ where F_1 is a tree of F. Notice that since F_1 and the other trees of F do not intersect, we have $weight(F) = weight(F_1) + weight(F \setminus F_1)$. Furthermore, F_1 is an optimal forest covering Y' and $F \setminus F_1$ is an optimal forest covering $Y \setminus Y'$ since otherwise, the union of two better solutions would lead to a better forest covering Y. We then have that $\mathrm{SH}_M(Y) = \mathrm{SH}_M(Y') + \mathrm{SH}_M(Y \setminus Y')$ and $\mathrm{SH}_M(Y) \geq \min_{Y' \subset Y}\{\mathrm{SH}_M(Y') + \mathrm{SH}_M(Y \setminus Y')\})$. Finally, assume that there exists $Y' \subseteq Y$ such that $\mathrm{SH}_M(Y) > \mathrm{SH}_M(Y') + \mathrm{SH}_M(Y \setminus Y')$ and let F' (resp. F'') be an optimal forest covering Y' (resp. $Y \setminus Y'$). Then $F' \cup F''$ would be forest covering Y of weight $weight(F' \cup F'') \leq \mathrm{SH}_M(Y') + \mathrm{SH}_M(Y \setminus Y') < F$, contradicting the optimality of F. Thus $\mathrm{SH}_M(Y) = \min_{Y' \subset Y}\{\mathrm{SH}_M(Y') + \mathrm{SH}_M(Y \setminus Y')\})$.

□

Theorem 2. *The Directed Steiner hypertree problem is Fixed-Parameter Tractable for the parameters $|T|$ and total number of tentacles of the hypergraph.*

Proof. The algorithm computes $\mathrm{SH}_M(T')$ for each ordered subset M of tentacular hyperarcs. Since the number of tentacular hyperarcs is bounded by the total number of tentacles k of the hypergraph, there are at most $2^k k!$ ordered subsets of tentacular hyperarcs. For a given M, we now compute $\mathrm{SH}_M(T')$ using a dynamic programming algorithm induced by the recursion of Theorem 1. We need to store the value of $\mathrm{SH}_M(Y')$ for every subset Y' of T'. Since the size of T' is bounded by $k + |T|$, we have at most $2^{k+|T|}$ such subsets. Finally, since for every vertex s and every $Y' \subseteq T'$, the computation of $\mathrm{ST}(s, Y')$ is FPT in $|Y'| \leq |T'| \leq k + |T|$, the total running time of the algorithm is FPT in $k + |T'|$.

□

4.4 MultiPus framework

MultiPus is composed of three steps:

1. Network merging and pre-processing;
2. Computing the best Directed Steiner Hypertree(s) from the set of sources to the set of targets;
3. Post-processing the solutions and visualising them using Dinghy.

We now describe the pre-processing and post-processing steps and discuss the central part of the framework.

Once the merged network is obtained as explained in Section 4.2.2, a lossless filtering is performed. This filtering is such that no solutions are lost between the merged network and the filtered merged network.

4.4.1 Directed hypergraph filtering

We call a reaction $a \in \mathcal{A}$ a sink reaction if it has no product, namely is such that $|\mathsf{src}(a)| = 0$. Similarly, an uptake reaction $a \in \mathcal{A}$ has no defined substrate ($|\mathsf{tgt}(a)| = 0$).

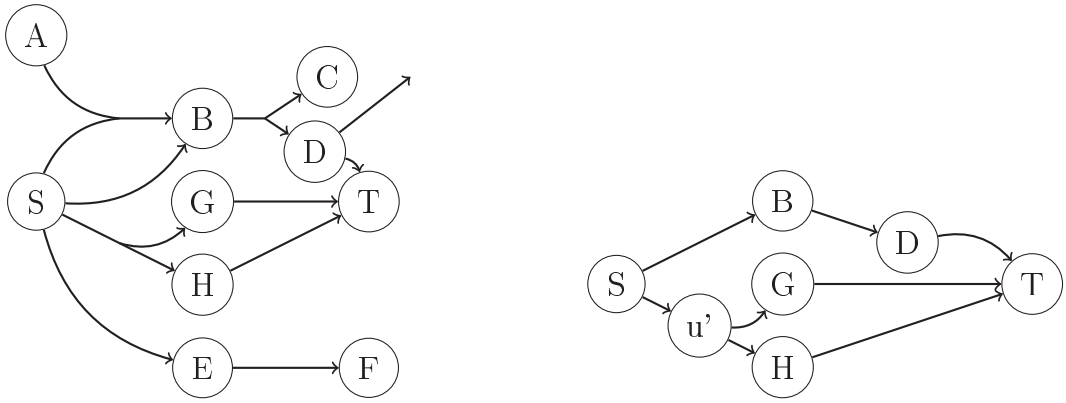
The filtering rules for the merged network are twofold:

1. Removal of the sink and import reactions: removes any arc (reaction) $a \in \mathcal{A}$ such that $|\mathsf{tgt}(a)| = 0$ or $|\mathsf{src}(a)| = 0$;
2. Removal of the source and target metabolites that are not part of S and T: removes any vertex (metabolite) $v \in V$ such that $d^-(v) = 0$ or $d^+(v) = 0$. If $d^-(v) = 0$, then all the outgoing reactions are removed. If $d^+(v) = 0$, then the entering reactions $a \in \mathcal{A}$ such that $v \in \mathsf{tgt}(a)$ (that is, which have v as product) are removed if $|src(a)| = |\mathsf{tgt}(a)| = 1$. Otherwise the reaction is simplified by removing v from its products ($\mathsf{tgt}(a) = \mathsf{tgt}(a) \setminus v$).

Both steps are repeated until the network is stable (no vertex and no arc fit the requirements of the filter).

In Figure 4.3a, the hyperarc $S \to G + H$ has a spreadness of 0. It is thus divided into 3 arcs with the introduction of a node u'. The three created arcs are $(S, u'), (u', G), (u', H')$. The filtering step is applied twice. The first time, using rule 2, the vertices A, C, and F are removed, as are the reactions $A + S \to B$ and $E \to F$. The reaction $B \to C + D$ is simplified into $B \to D$. Using rule 1, the sink reaction of D ($D \to$) will be removed. The second time, according to rule 2, vertex E is deleted as is the reaction $S \to E$. The resulting network can be seen in Figure 4.3b.

In Figure **??**, the vertex I would be removed along with the two reactions $I \to$ and $F \to I$.



(a) Original network

(b) Network after filtering.

Figure 4.3: Toy example, before and after filtering. Here only one network is considered. All arcs have a weight of w_{worker} in 4.3a. In 4.3b, all arcs kept their weight of w_{worker} except for $u' \to G$ and $u' \to H$ that have each a weight negligible w_{pseudo}

Once the preprocessing and the filtering are finished, once can apply the central part of MULTIPUS, that is the research for the Directed Steiner Hypertrees of lightest weight, such that all targets are reached starting from one or several sources of the source set.

4.4.2 Obtaining the best weighted Directed Steiner Hypertree(s)

We proposed here two possibilities, one that is the implementation of the algorithm described in Section 4.3 and the other that solves the problem using an *ASP* formulation by calling the solver CLINGO.

The main reason for the second possibility is that, due to the larger number of tentacular hyperarc combinations, our *in-house* algorithm can only take into account a bounded number of tentacular hyperarcs. We thus have to establish a parameter k that is the maximum number of tentacular hyperarcs in the solutions.

Furthermore, while this implementation can enumerate solutions for such bounded number of tentacular hyperarcs, it does not provide for a same set M of tentacular hyperarcs all the directed forests of lightest weight but only one.

Finally, while this algorithm is interesting from a theoretical point of view, it remains relatively slow. CLINGO outperforms it in terms of speed but also because it does not have to work with a bounded number of tentacular hyperarcs and can provide all the forests if required.

One further subtlety is that the Directed Hypersteiner Hypertree Algorithm cannot work with a weight of 0. We thus have $w_{pseudo} = 0$ when working with CLINGO and $w_{pseudo} = 10^{-6}$ when we use the Directed Hypersteiner Forest Algorithm.

Both methods were applied to the case studies in Section 4.5 and will be compared later on in Section 4.5.3.

4.4.3 Visualising the obtained solutions

Once the solutions using either MultiPus or clingo are obtained, it is possible to observe them using Dinghy (Bulteau et al., 2015). In order to get the solutions in a *json* format, we post-process the solutions using an in-house python program.

4.5 Application

The main objective of microbial consortia engineering is to highlight their capacity to reach enhanced productivity, stability or metabolic functionality (Brenner et al., 2008). More in particular in this work, we explore the possibility of such consortia to produce compounds of interest using low cost substrates (such as, for instance, the waste of other industries).

We initially focused attention on the production of two bioactive compounds: penicillin and cephalosporin C, useful to the pharmacology industry for their antibiotic properties. For this production, a synthetic consortium defined as a system of metabolically engineered microbes which are modified by genetic manipulations and/or regulatory processes (Bernstein and Carlson, 2012) has been tested, using distant species as will be explained in the first example. The goal in this case was to take advantage of the different metabolic capabilities of the organisms composing the consortium for the *de novo* synthesis of bioactive metabolites and to show that the model is able to select the Directed Steiner Hypertree of least cost to produce one or a set of metabolites of interest.

We then considered the case of an artificial consortium. This corresponds to a system composed of wild-type populations that do not naturally interact (Bernstein and Carlson, 2012). We tested the association of a natural 1,3-propanediol (PDO) producer *Klebsiella pneumoniae* with an acetogenic Archae *Methanosarcina mazei*. The goal is to obtain a higher yield of 1,3-propanediol. Indeed, production of this compound in a pure culture of *K. pneumoniae* is associated with production of acetate. The latter has an inhibiting effect on bacterial growth, and ultimately also on the production of PDO. Hence associating *K. pneumoniae* with a methanogen has been proposed to reduce such effect (Sabra et al., 2010; Bizukojc et al., 2010).

All the genome-scale models (GEMs) used were extracted from Kegg (Kanehisa and Goto, 2000) using MetExplore (Cottret et al., 2010b). In both examples, cofactors and co-enzymes obtained from a list available in Kegg (Kanehisa and Goto, 2000) were removed. This list is available in Appendix (B B.2.1) The networks, constructed as explained previously, were filtered using a lossless compression step (see Section 4.4.1, Figure 4.3).

The resulting networks have a high number of tentacular hyperarcs. In the first case, the directed hypergraph contains 10087 arcs and 285 tentacular hyperarcs (that is, arcs with at least two substrates). The total number of tentacles of the graph is 575. In the case of improved PDO production, the network contains 1606 arcs and 71 tentacular hyperarcs for a total number of tentacles of 142.

In the absence of any prior knowledge, the weights were set uniformly using as *a priori* the fact that endogenous reactions should be easier to use than transport ones (no need to export or

to uptake compounds) and than insertions (since this implies introducing one or several genes and over-expressing them).

Therefore, the following weights were first applied: $w_{worker} = 1$, $w_{other} = 100$, $w_{transition} = 100$. Notice that the weight of the (hyper)arcs that are present in the organisms forming the consortium is not zero, but instead equal to a value above zero that remains however small in relation to the weights of an insertion or of a transition. The motivation for this is to favour solutions which, while minimising the number of insertion or transition hyperarcs that are used, also minimise the number of hyperarcs corresponding to reactions that are internal to the microorganisms in the consortium.

In the second application, two sets of transport weights were adopted, one a refinement of the first, as will be explained later on.

4.5.1 Antibiotics production

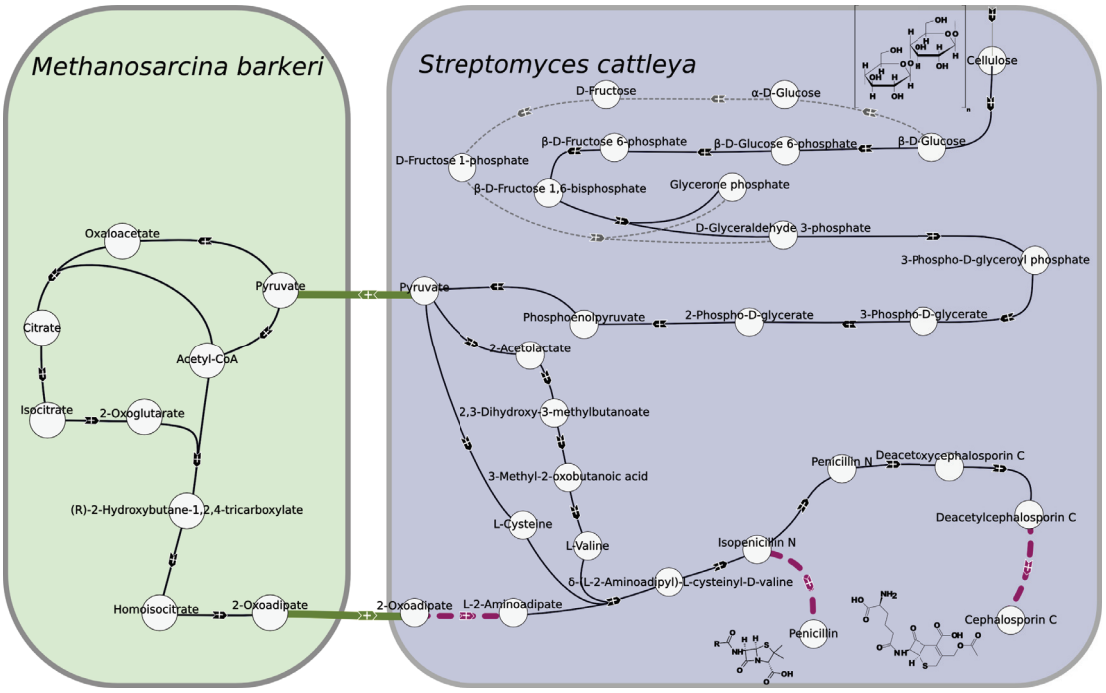
In this first application, a synthetic consortium of three Actinobacteria (*Streptomyces cattleya*, *Rhodococcus jostii* RAH_1, *Rhodococcus erythropolis* BG43) and one methanogenic Archaea (*Methanosarcina barkeri*) was tested to determine which microbial consortium could produce a set of metabolites of interest. In this case, two well-known beta-lactam antibiotics (penicillin and cephalosporin C) were selected. Both active compounds belong to the cephalosporin/penicillin pathway and share several metabolic reactions. They also have a common precursor, namely isopenicillin N, are commonly used for their antibacterial properties and are naturally produced by fungi belonging to the *Aspergillus* and *Cephalosporium* species (*Aspergillus chrysogenum* and *Cephalosporium acremonium* respectively) (Katz and Baltz, 2016). In this case, cellulose was used as carbon source. Indeed, life on earth depends on photosynthesis, which results in the production of plant biomass having cellulose as major component, and cellulosic materials are particularly attractive in this context because of their relatively low cost and plentiful supply (Lynd et al., 2002).

The microorganisms were chosen because of the availability of Actinobacteria to produce bioactive compounds (representing about 45% of all the microbial bioactive products discovered (Jose et al., 2013)). Furthermore, the phylogenetic distance between Actinobacteria and Archaea suggests variability in their metabolisms. The presence of reactions that are specific to each organism means that there might be a gain in the overall metabolic capabilities from making the two bacteria work together. Using a consortium could thus be more efficient to produce one or several of the metabolites of interest. In addition, two other organisms (henceforth called *reference organisms*) were used for reaction insertion: *Aspergillus nidulans* and *Streptomyces rapamycinicus*. The first is a fungus known to produce penicillin while the second possesses reactions in the penicillin/cephalosporin C pathway, and in particular those needed to produce cephalosporin C. All the reactions present in the reference organisms were added to the four prokaryotes forming the consortium (as described in Section 4.2.2).

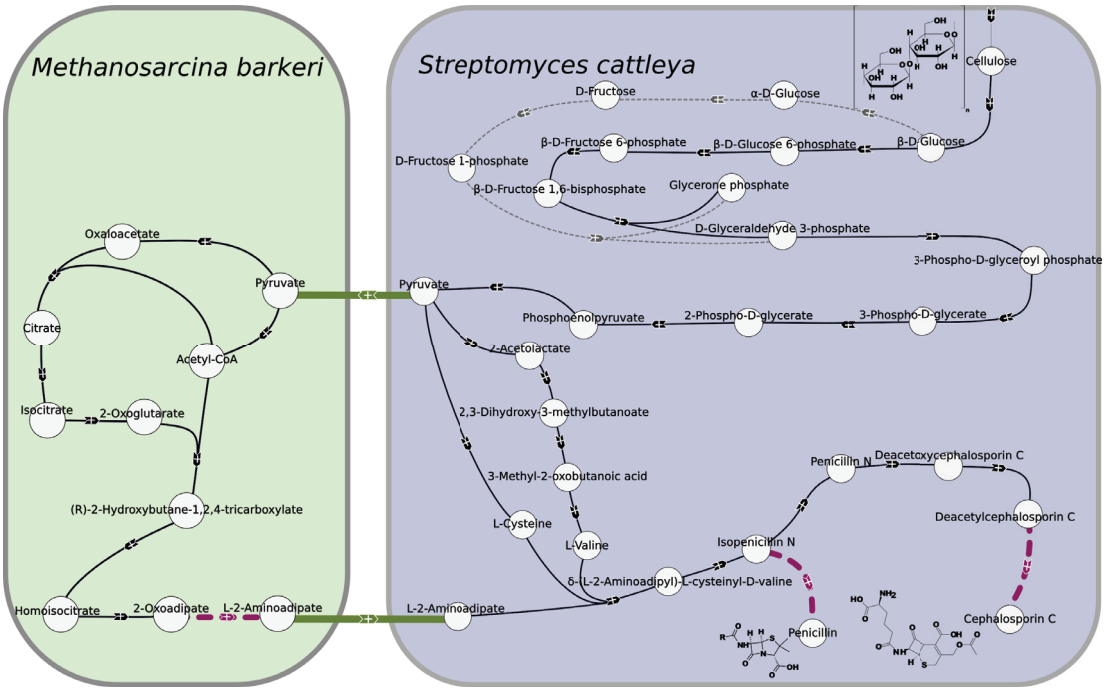
Four solutions with a minimum cost of 528 (2 transports, 3 insertions, and 28 endogenous reactions) are found. All of them are composed of *Streptomyces cattleya* and *Methanosarcina barkeri* showing that topologically, there is no need to use the other two Actinobacteria to produce both beta-lactam antibiotics. Two of them are presented in Figure 4.4a. In this case, the consortia exchanges pyruvate and 2-oxoadipate.

The other two use another metabolite transport (*i.e.* L-2-aminoadipate) and are illustrated

in the Figure 4.4b. In this case, the insertion of the reaction transforming 2-oxoadipate into L-2-aminoadipate is proposed in *M. barkeri* and L-2-aminodipate is transported into *S. cattleya*.



(a) Two solutions of minimum weights. Here, pyruvate and 2-oxoadipate are exchanged.



(b) Two solutions of minimum weights. Here, pyruvate and L-2-aminoadipate are exchanged.

Figure 4.4: Representation of the four solutions of minimum weight. The circles are compounds. Black hyperarcs are endogenous reactions, that is reactions already present in the organisms forming the consortium, while purple-dashed hyperarcs are the reactions that were inserted. The widths of the arcs are proportional to the assigned weights. Grey-dashed arcs represent an alternative path of endogenous reactions in the upper part of glycolysis. Green arcs represent the transport of pyruvate *Streptomyces cattleya* to *Methanosarcina barkeri* and of 2-oxoadipate (Figure 4.4a) or L-2-aminoadipate (Figure 4.4b).

Three tentacular hyperacs are used in this case. One of the reactions is *N-(5-amino-5-carboxypentanoyl)-L-cysteinyl-D-valine synthase* that converts L-2-aminoadipate, L-valine and L-cysteine into δ-(L-2-aminoadipyl)-L-cysteinyl-D-valine, which is the starting point for the production of penicillin and cephalosporin C. All metabolites previously mentioned can be produced from pyruvate. The requirements to produce the three substrates of N-(5-amino-5-carboxypentanoyl)-L-cysteinyl-D-valine synthase using a solution of minimum weight therefore force to go back into the bacterium producing both amino-acids (L-valine and L-cysteine), in this example *S. cattleya*. The two other tentacular hyperarcs correspond to the reactions for *citrate synthase* (converts acetylCoA, H_20 and oxaloacetate into citrate and CoA) and *acetylCoA:2-oxoglutarate C-acetyltransferase* (transforms 2-oxoglutarate and acetylCoA into homocitrate ((R)-2-hydroxybutane-1,2,4 tricarboxylate).

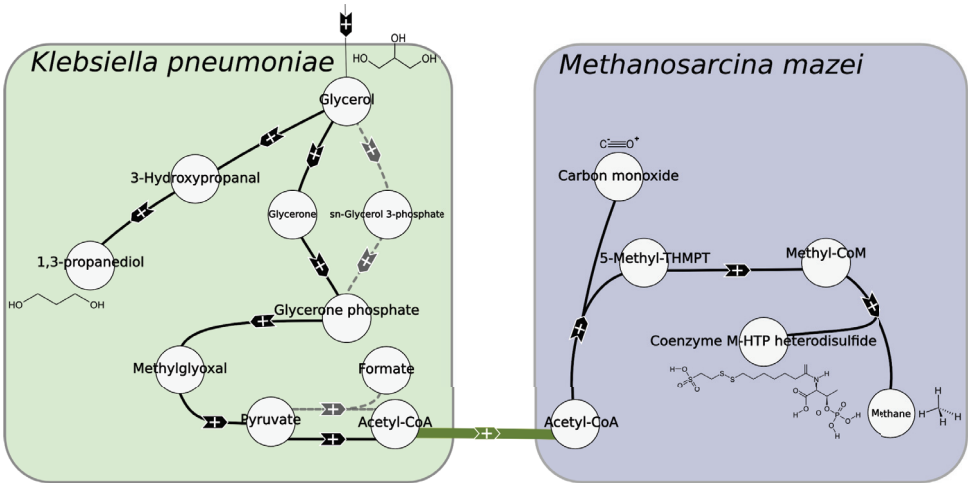
4.5.2 Production of 1,3-propanediol and methane

The compound 1,3-propanediol (PDO) is of high interest in biotechnology since it is used as a building block in polymers (Saxena et al., 2009). Bizukojc *et al* (Bizukojc et al., 2010) reported that the co-culture of the 1,3-propanediol producer *Clostridium butyricum* with a methanogenic Archaea, namely *Methanosarcina mazei*, could lead to a better yield of production. Indeed, in *C. butyricum*, production of PDO leads to the production of acetate as well as of a side-compound, the latter then participating in the production in *M. mazei* of methane, which is the main molecule in the composition of biogas.

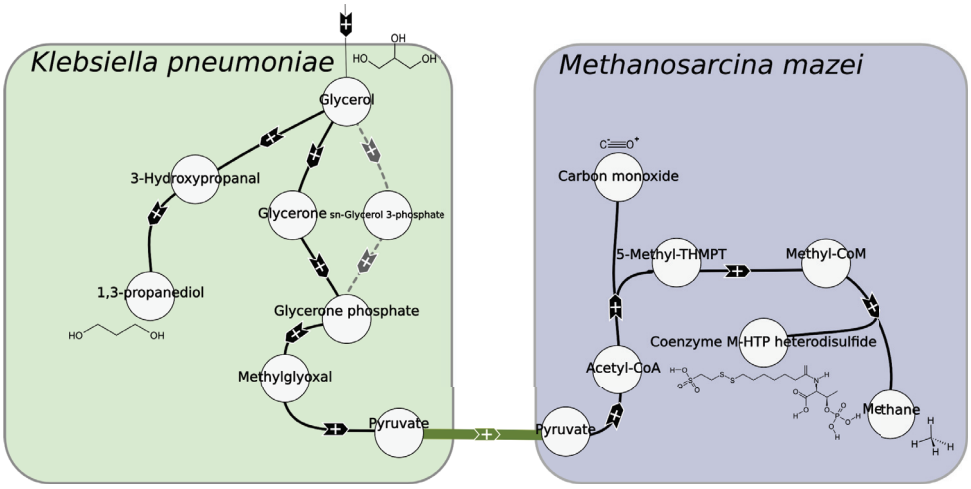
In this example, another PDO producer and Enterobacteria glycerol scavenger, namely *Klebsiella pneumoniae*, is associated with *Methanosarcina mazei* to produce 1,3-propanediol and methane. Both organisms have the capacity to produce the target compounds. Hence, no reference organisms were used. The weights were first set as in the previous section (*i.e.* $w_{worker} = 1$, $w_{other} = 100$, $w_{transition} = 100$). The only authorised source was glycerol. Indeed, glycerol is a by-product of biodiesel biodiesel. It therefore is a substrate of choice for biotechnological processes (da Silva et al., 2009). In this case, we have two targets: 1,3-propanediol and methane.

We obtain six solutions with the same weight of 110 (1 transition and 10 endogenous reactions) represented in Figure 4.5.

In *K. pneumoniae*, there are two ways of reaching glycerone phosphate from glycerol. Moreover, two different reactions are possible to transform pyruvate into acetyl-CoA, one of them forming also formate. Finally, in the solutions obtained, there is also the possibility to exchange pyruvate instead of Acetyl-CoA. This therefore leads to six solutions.



(a) Four of the six solutions. Here, the consortium exchanges acetyl-CoA.



(b) The two others solutions. Here, the consortium exchanges pyruvate.

Figure 4.5: Solutions obtained with an uniform weight for the production of 1,3-propanediol from glycerol in *K. pneumoniae* and *M. mazei*. Black hyperarcs are endogenous reactions and green arcs represent transports. Grey dashed hyperarcs represent alternative paths.

In this case, the community does not exchange acetate but acetyl-CoA or pyruvate. In eukaryotes, transporters of acetyl-CoA are known in several pluricellular organisms and also in yeast. However, no transporter of acetyl-CoA has been detected in organisms close to the ones used in our case. Moreover, the pool of acetyl-CoA is essential to *K. pneumoniae*. Indeed, Jung *et al.* (Jung et al., 2014) reported that a mutant with a reduced pool of acetyl-CoA showed growth retardation and redox imbalance. Therefore, it is not clear whether *K. pneumoniae* has an advantage in sharing acetyl-CoA or pyruvate (which is a substrate for the reactions producing acetyl-CoA). However, as stated previously, the production of 1,3-propanediol is associated with the synthesis of acetate and formate. Those by-products are inhibiting for *K. pneumoniae* and can reduce both its growth and the production of 1,3-propanediol (Jung et al., 2014; Cheng et al., 2005). Finally, *K. pneumoniae* possesses a citrate/acetate exchanger (Kästner et al., 2002) which is CitW, and *Methanosarcina* spp. can grow on acetate although other substrates might be preferred. This indicates the possibility of an exchange of acetate between the two organisms since transport is possible in both species. We therefore decided to diminish the weight of the transport of those organic acids to $w_{transition} = 50$.

Two minimum solutions were obtained with a weight of 61 (the acetate transport with $w_{transition} = 50$ and 11 endogenous reactions). They are presented in Figure 4.6.

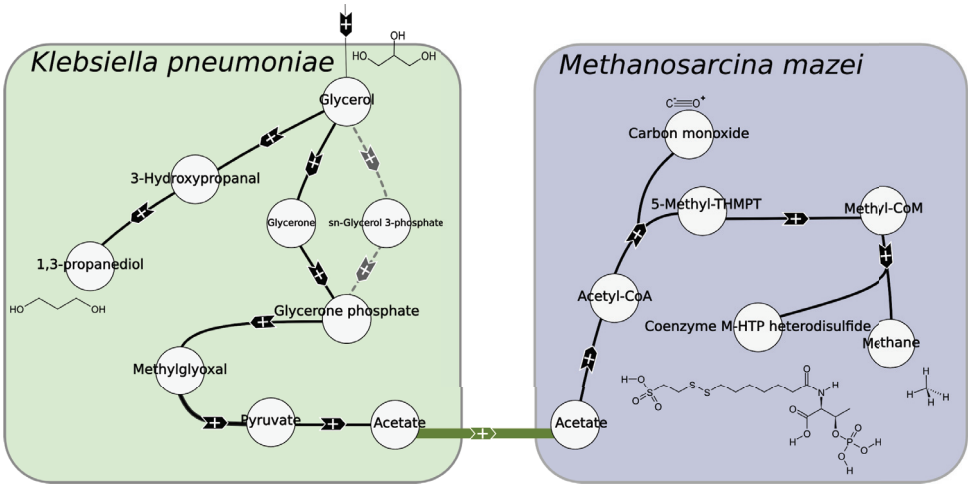


Figure 4.6: Solutions with reduced weight for acetate and formate to produce 1,3-propanediol from glycerol in *K. pneumoniae* and *M. mazei*. Black hyperarcs are endogenous reactions and green arcs represent transports, here acetate from *K. pneumoniae* to *M. mazei*. The grey dashed arcs represent the alternative solution.

We can observe that this solution is really close to the previous one. Here, pyruvate is used to produce acetate (pyruvate:ubiquinone oxidoreductase) which is then exchanged from *K. pneumoniae* to *M. mazei*. The resulting pathway is in agreement with the one described by (Sabra et al., 2010).

4.5.3 Comparison between the Directed Steiner Hypertree algorithm and the *ASP* formulation

As mentioned previously, the experiments were ran for the two current versions of MultiPus. One uses the implementation of the algorithm presented in Section 4.3 whereas the other uses an ASP formulation and calls the solver clingo to obtain the solutions of lightest weight. We compared the running time and number of solutions of the two methods. The results are presented in Table 4.1.

First, we can observe that the running time greatly decreases with the *ASP* formulation. Clearly, it is necessary to improve the in-house algorithm to match those results. The current computation time of clingo for this problem is not an issue, and it can run on a personal computer without problem.

One may notice that the number of solutions obtained with the *ASP* formulation is higher than the number of solutions with our *in-house* implementation of the problem. This is because our algorithm only outputs one solution (*i.e.* lightest forest) per set of tentacular hyperarcs whereas clingo can enumerate all Directed Steiner Hypertrees. We thus chose to present the results obtained by clingo in the previous section.

Both versions of MultiPus are available at the address: `http://multipus.gforge.inria.fr`.

Method	Dataset used	Running Time	# of threads	# of solutions
1	Antibiotics Production	286 hours	20	1
2	Antibiotics Production	< 7 minutes	4	2
1	PDO production, uniform weight	9 hours	20	1
2	PDO production, uniform weight	< 1 minute	4	6
1	PDO production, $w_{acetate} = 0.5$	9 hours	20	1
2	PDO production, $w_{acetate} = 0.5$	< 1 minute	4	2

Table 4.1: For the in-house algorithm, the value of the parameter k, that is the maximum number of tentacular hyperarcs in the solutions, was set to 3. Column method: 1 stands for the in-house algorithm and 2 stands for the ASP solver.

4.6 Discussion

The method introduced here allows to infer topological subnetworks to produce target compounds using one or several microorganisms forming a consortium. Ensuring that a component will be produced as much as it will be consumed according to stoichiometric coefficients leads to a more complex problem. Since we do not use such coefficients, a conservative hypothesis was adopted. This induces the exclusion of some cycles where a substrate used in a reaction is immediately formed again (such phenomenon appears for example in the phosphotransferase system in *E. coli*). Without stoichiometric coefficients, we cannot guarantee that the intermediate substrates of the cycles will be all regenerated by a solution. Prohibiting those cycles allows us to ensure that all solutions are feasible by themselves, meaning that all intermediates are at least as much produced as they are consumed (regardless of the remaining of the network).

Once a solution is obtained several points must be verified.

In the first example, only two of the four bacteria were selected to produce the two compounds of interest, showing the ability of our algorithm MultiPus to not only identify the less costly solution, but also to select the best consortium among a larger set of microorganisms given as input.

In this synthetic bacterial consortium defined by *Streptomyces cattleya* and *Methanosarcina barkeri*, pyruvate and either 2-oxoadipate or L-2-aminoadipate are exchanged between the two prokaryotes. The organisms therefore need to be able to export and uptake the three compounds. It was shown that *Methanosarcina barkeri* – the model species of the genus *Methanosarcina* whose properties are shared by most of the others (Balch et al., 1979) – grows on pyruvate, the uptake being done by passive diffusion (Bock et al., 1994).

Moreover, *Streptomyces coelicolor* is able to transport monocarboxylates such as pyruvate by secondary carriers and active transporters (Getsin et al., 2013). Although pyruvate transporters have not yet been shown to exist in *S. cattleya*, it is probable that the transport of pyruvate is nevertheless possible since it happens in a closely related organism (*i.e. S. coelicolor*) (Getsin et al., 2013).

As concerns the second exchange, mitochondrial transporters for oxodicarboxylic acids (oxodicarboxylate carrier proteins (ODCs)) such as 2-oxoadipate and 2-oxoglutarate were reported in yeast (*Saccharomyces cerevisiae*) and in human (Palmieri et al., 2001; Fiermonte et al., 2001). Both human and yeast ODCs catalyse the transport of 2-oxoadipate and 2-oxoglutarate by a counter-exchange mechanism. Moreover, L-2-aminoadipate is also transported by the human ODC (Fiermonte et al., 2001). However, no homologous genes were found in Archaea and Acti-

nobacteria (using a BLAST analysis), neither did we find any information about the presence of such transporters in *Methanosarcina* or *Streptomyces*. Further experiments will therefore be needed to determine whether the two species constituting the microbial consortium do possess the ability to uptake and export 2-oxoadipate. Moreover, if it is confirmed that these two bacterial strains indeed lack this ability, an insertion of ODCs might still be possible, similarly to what was performed in *Escherichia coli* using human ODCs (Fiermonte et al., 2001).

Although the production of two beta-lactam antibiotics destroys the walls of positive Gram bacteria, *Streptomyces* is well-known for possessing a gene cluster which orchestrates antibiotic biosynthesis. Such cluster consists of resistance, transport and regulatory genes physically linked and coordinately regulated with genes encoding biosynthetic enzymes (Chater and Bibb, 1997). Among such species, *Streptomyces clavuligerus* produces several beta-lactam compounds, such as cephamycin C, clavulanic acid (an inhibitor of several beta-lactamases able to inactivate penicillins (Katz and Baltz, 2016)) and other structurally related clavams (Alexander and Jensen, 1998). Moreover, thienamycin, a carbapenem compound belonging to a class of beta-lactam antibiotics, is produced by *S. cattleya*. This metabolite employs a similar mode of action as penicillins through disrupting the cell wall synthesis (peptidoglycan biosynthesis) of various Gram-positive and Gram-negative bacteria. It further presents a resistance to bacterial beta-lactamases enzymes (Katz and Baltz, 2016; Nuñez et al., 2003). Therefore, *S. cattleya* could produce the two beta-lactam antibiotics without affecting its bacterial growth.

One must however call attention here to the fact that cultivating an aerobiose Actinobacteria and an anaerobiose Archaea in a same culture may be difficult. On one hand, several anaerobic-aerobic co-cultures have already been reported (Field et al., 1995). Indeed, because of the low solubility and diffusibility of oxygen in water, anaerobic micro-niches can be created and maintained in an aerobic environment (Field et al., 1995). On the other hand, we have here two mesophilic species: *Streptomyces* sp. (with a temperature growth interval between 25°C and 35°C) and *Methanosarcina* sp. (with an optimum of growth around 37°C)(Gunnigle et al., 2013). In this context, the synthetic bacterial consortium will be able to grow together without major difficulties.

At their bacterial growth temperature (between 25°C and 37°C), we exclude a possible temperature-dependent biosynthetic pathway of antibiotic compounds as already reported for actinorhodin (Chen and Qin, 2011). Indeed, the expression of the actinorhodin gene cluster was showed to be impossible at high temperatures (45°C) and instead realised at 30°C and at 37°C, suggesting that it could thus depend on the temperature (Chen and Qin, 2011). Under such conditions, the penicillin and cephalosporin C gene cluster should therefore be heterologously expressed by the consortium which should be able to produce the two well-known beta-lactam antibiotics.

In the second example, we retrieved a possible network for the joint production of 1,3-propanediol and methane. In (Jung et al., 2014), attempts to reduce the production of by-products such as acetate through gene deletion led to a growth defect in *K. pneumoniae*. In those experiments, the yield of 2,3-butanediol (BDO) is improved by deletion of *pflB*, possibly because of the accumulation of pyruvate, a precursor of BDO. Indeed, *pflB* with *ldhA* encodes the *pyruvate formate-lyase* enzyme. Nevertheless in our case, pyruvate is not a precursor of PDO, hence the deletion of the same gene (*pflB*) would have a negative impact since the growth of the cells would be impaired by the redox imbalance created. Hence the possibility of the association with an acetogenic Archaea is of great interest to regulate acetate production.

In (Bizukojc et al., 2010), an *in silico* simulation of the co-culture of another propanediol producer, namely *Clostridium butyricum*, with *M. mazei* showed an improvement in the growth of *C. butyricum* due to the consumption of acetate by *M. mazei*. Such consumption alleviates the inhibition of acetate. A similar effect should be expected for *Klebsiella pneumoniae*. The lighter weight assigned to the exchange of acetate allowed us to retrieve a feasible solution. Although acetate can be utilised almost completely by *M. mazei* for its growth, it is necessary to have methanol (present in raw glycerol obtained from biodiesel plant) in the medium to produce methane. However, even if the production of methane is low, the association of the two organisms will decrease the concentration of extracellular acetate, which is toxic for *K. pneumoniae*, hence increasing the yield of PDO. Co-cultures of *Clostridium* sp. associated to methanogenes such as *Methanosarcina sp.* CHTI55 have been described in the literature, showing acetate utilisation by methanogene organisms (Koesnandar et al., 1990). The use of an Enterobacteria, *Klebsiella pneumoniae*, as the propanediol producer in co-culture with methanogenes has been less described. Hence, more extensive tests on the feasibility using classical optimisation techniques are needed, even though the process and apparatus for such associations have been patented (Friedmann and Zeng, 2013).

As shown in this second application, we can assign a non uniform weight to the exchange of compounds between organisms, the insertion of exogenous reactions or the use of internal reactions. Using a biological *a priori* to tune the weights assigned to each reaction is helpful to obtain a realistic solution. Indeed, the weight of an inserted reaction can be set more precisely by taking into account, for example, gene-reaction associations. Reactions catalysed by protein complexes require the insertion of several genes, hence may be harder to handle than those associated to single genes. Using the AND/OR relations available in the SBML models, insertion weights may thus be adapted to reflect those difficulties. Moreover, if information about the inserted organisms is available, more complex weights can be computed, taking into account enzyme promiscuity, catalytic performance, gene compatibility (Carbonell et al., 2011), but also for example the toxicity of side-products or even a known difficulty of enzyme incorporation. The exchange weights are harder to evaluate, however information about transporters (active or passive) for export and uptake may be taken into account to tune the exchange reactions. For example, a passive transporter is costless, molecules move across the membrane without energy input; on the contrary, an active transporter such as an ATP-powered pump will be costly since it requires the hydrolysis of ATP into ADP. Attributing a relative weight inside each category as briefly described above may be straightforward. What may be more difficult is to decide on how to balance such weights across the three categories. This may require some trial and error, and be dependent on the *in silico* experiment that is considered.

4.7 Conclusion

We proposed a new topological method, called MultiPus, to select possible microbial consortia for the production of compounds of interest.

With MultiPus, any situation of both exogenous and endogenous compounds might be considered, as well as larger initial consortia whose final composition in terms of species is then optimised by the method. Finally, by setting the sources required, one can test the possibility of using low-cost substrates for the production of high value chemicals.

As a *post-processing* step, classical methods of flux balance analysis (using the inferred topological network) can be employed to predict product yield (Chowdhury et al., 2014; Orth et al., 2010; Segrè et al., 2002). Gene over-expression and knock-out can moreover be explored in order to guarantee both growth and production of the compound(s) of interest, but also interaction among the species present in the consortium (Pharkya and Maranas, 2006; Tepper and Shlomi, 2009).

Indeed, the species that are part of the consortium may not have the same growth rate, hence may not reach an equilibrium in terms of composition when all organisms are present.

Stable growth and equilibrium in biomass of the community which is being considered is of importance, and stoichiometric models could be used to predict such equilibrium (Koch et al., 2016; Zomorrodi et al., 2014). Current constraint-based techniques are, to our knowledge, multi-level (usually bi-level), or use a linear combination of objectives. The formulation does not take into account multiple objectives independently. It appears clear however that several objectives must be considered while evaluating a consortium. Indeed, whereas the final goal is an interesting product yield, growth of the organisms that are part of the consortium is necessary. We are thus currently working on a multi-objective framework that would be able to propose several optimal values for multiple objectives.

If balance cannot be reached, it is necessary to create a beneficial interaction among the organisms involved (mutualism or syntrophy) to guarantee the success of the synthetic community (Jagmann and Philipp, 2014). If needed, mutualism can be enforced by genetic engineering, for example by creating auxotrophic strains; this will force a cross-feeding between organisms, regulating the growth of the species composing the co-culture (Shou et al., 2007; Hosoda et al., 2011). Furthermore, (Zampieri and Sauer, 2016) proposed a framework to select culture media in order to favour cooperation.

This first model allows to infer topologically possible insertions for heterologous expression and the usage of a mixed culture for the production of exogenous and/or endogenous target compounds. Moreover, MultiPus may thus enable to establish which co-cultures could be interesting to use in order to avoid the inhibition of co-products (*e.g.* 1,3-propanediol). It is a good starting point, that should be associated in the future with more quantitative methods in order to guarantee maintenance and growth of the organisms in communities (for instance, taking into account electron transport and/or red/ox balance).

Conclusion and Perspectives

During these three years, I focused on modelling metabolism. The idea was to unravel some metabolic behaviours using simple *hypothesis-driven* models applicable to several organisms or contexts.

Such models were developed along two different questions. The first concerns how organisms respond to changes and switch from one steady-state to another, and the second how they may work together (as a community) towards a bioprocess goal. In this work, such goal was the production of specific compounds but the issue is more general and could relate to other biological processes.

We started by presenting in Chapter 1 some current knowledge on metabolism, its regulation and the already existing modelling approaches.

In Chapter 2, we presented a topological analysis of metabolic shifts and proposed a method that we called Totoro (TOpological analysis of Transient metabOlic RespOnse). Totoro enumerates sub-hypergraphs, or more particularly sets of hyperarcs (the reactions) based on some coloured vertices (the metabolites). We started from a previous method developed in the team but proposed a new definition allowing to infer the direction of the reactions (*i.e.* inhibition or activation), without constraining the solutions to be acyclic which was one of the drawbacks of the previous method. Furthermore, this new modelling is done using a directed hypergraph representation which allows to maintain the coupling of substrates or products and does not require anymore to use the extensive lightest path compression. We propose that the regulations act upon some of the reactions selected, by allosteric or transcriptomic responses, and that part of the impacted metabolism could be unravelled. Totoro can also help design new metabolomic experiments as it suggests possible metabolites to measure in order to refine the solutions. As an illustration, the method was applied to a published dataset of Yeast exposed to cadmium. We could retrieve the cadmium detoxification pathways and propose links between the metabolites that were impacted. The assumption here is that there is a parsimonious reorganisation of the metabolism. The solutions we provide are minimal or minimum. Whereas the minimal ones are in general not treatable in practice, the minimum solutions obtained did retrieve the expected effect of the cadmium presence. Some minimal solutions could however also be of interest. Using a parsimonious response hypothesis, we could propose to discard the solutions that are large, and focus only on those with a number of reactions close to the minimum.

Several more improvements can be performed. For a theoretical point of view, obtaining an efficient algorithm and proving the complexity of the enumeration problem is thus an important perspective. One option to handle the large number of solutions, in addition to the clustering by frequent itemsets, could be to use big data analysis techniques, such as weakly supervised learning, to group possible common reactions. This is an option that is currently being considered. Furthermore, in the new definition, we did not consider the stoichiometry whereas it is

an important feature of the metabolic network that is directly linked to the quantities of matter transformed by a reaction. Including stoichiometry in the current definition of *metabolic hyperstories* will thus provide a more precise information, as well as more complete solutions possibly in a smaller number. Taking into account the stoichiometry can be done in a topological way, and a definition for what would thus be a *stoichiometric metabolic hyperstory* remains to be found. Such new definition can be computed using a combinatorial algorithm on weighted directed hypergraphs but we also proposed a method based on constraint-based programming in Chapter 3.

Indeed, in Chapter 3, we presented KOTOURA (Kantitative analysis Of Transient metabOlic and regUlatory Response And control), a method using stoichiometry and a quantitative measurement of the concentrations. This new framework uses constraint-based modelling to formulate more quantitatively the changes that occurred during the transient state, using the stoichiometric matrix and precise concentration measurements. Such quantitative analysis is performed without requiring to know the exact kinetics of the model. We show that the overall variations of the reaction rates during the transient state can be estimated. In this method, the same hypothesis of parsimonious changes of the reactions was made. We used a combined objective to enumerate the possible reactions selected, and implemented the method using Mixed-Integer Linear Programming (MILP) and a CPLEX solver. The framework developed was presented using small simulated datasets. We then discussed the simulations obtained with a larger and more recently published kinetic model. The analyses in this case are still on-going, but we aim to combine quantitative measurements of concentrations with qualitative knowledge of increasing or decreasing concentrations to create a mixed method that would use the two types of data presented in Chapters 2 and 3, *i.e.* both quantitative and qualitative measurements of the concentration variations between the two conditions.

In both cases, we want to unravel the mechanisms in action when there is the necessity to respond to new environmental conditions but it could also apply when an organism is impacted by mutations. Such mutations could affect enzyme-coding genes but also genes that do not appear directly related to metabolism but which, by unknown direct or indirect regulation(s), affect the metabolism of the organism.

Both of the above chapters were thus concerned with deciphering an organism's response to new conditions. The methods could be extended to communities. However, the work we propose on communities is related to another objective, which is the inference of consortia for the production of target compounds. Those two biological problems are nevertheless related. Indeed, it is through a better understanding of metabolism that current biotechnological processes are now possible at a larger scale.

In Chapter 4, MULTIPUS (MULTIple species for the synthetic Production of Useful biochemical Substances), proposes a topological analysis of the production pathways. The modelling approach adopted uses a weighted directed hypergraph representation. The weights of the hyperarcs, that is of the reactions, are based on the presumed costs of catalysis, transport or expression of heterologous genes. MULTIPUS selects a consortium (which might be composed of a subset of the proposed species) together with the reactions needed to produce the metabolites of interest. We applied it to two cases and proposed new consortia for each one of them. However, in the future, more appropriate weight functions should be considered. Since we can propose modifications of the metabolic network, classical methods for optimisation of product yield can be applied (see Chapter 1), but positive interactions among the organisms in the consortium

must then be enforced.

MultiPus was developed with a bioprocess design in mind. The problem of enumerating Directed Steiner Hypertree(s) of minimum/minimal weight from a set of root vertices (the substrate(s)) towards some specific leaves could however have wider applications. In particular, if the weights are set to zero, then looking for all Directed Steiner Hypertrees with MultiPus would enable to infer the metabolite exchanges in natural communities.

In this PhD, we thus presented three methods that are not organism-specific but instead are generic. Despite such characteristic, they provided interesting insights on the biological cases studied.

In those works, we remained at the level of the small molecule metabolism (SMM). However, we saw that one ambition would be to start to push back the frontier of the SMM by explicitly modelling processes such as storage of macromolecules (*e.g.* protein synthesis), oxydo-reduction balance and so on. For example, in the case where Yeast is exposed to cadmium, while looking for *metabolic hyperstories* (in Chapter 2), it would be possible to model protein synthesis as a metabolite and to link such protein pool to the amino-acids.

Moreover, as mentioned, we work with a common hypothesis which is that the metabolic changes during a transient state should be restricted to small variations on a limited quantity of reactions. For now, we focused on the reactions. However, regulation as well as other processes, for example at the genomic and the transcriptomic levels, should also be taken into account. Creating a multi-layer model that would be able to do this will require an extensive reflection on how to combine regulation and signalling mechanisms. It would also depend on accessibility to datasets to test such model. This nevertheless seems a necessary step if we want to go much further.

Finally, we addressed the problem of designing communities. However, becoming able to model natural ones together with their behaviour in response to perturbation remains a challenge. It it is known that some reactions which are not possible in pure culture, could occur when different microorganisms are associated in a community. This is the case when one species catalyses an endergonic reaction whose product is immediately consumed by the exergonic reaction of another partner in the community, thus creating a driving force as the product concentration will always remain low (Wintermute and Silver, 2010).

Introducing known thermodynamics in our models (not only in MultiPus but also in Totoro) is an important perspective of this work. This can be done explicitly in the case of known concentrations while working with a constraint-based framework. In the case of a topological analysis, we propose to add weights to the hyperarcs such that known endergonic reactions would be disfavoured. For the method presented in Chapter 3, the Gibbs energy of the reactions computed with the metabolite concentrations can be used to interpret the results of the method.

The two topological methods presented (Totoro in Chapter 2 and MultiPus in Chapter 4) did not take into account the stoichiometry. In the case of MultiPus, including stoichiometry could be done in such a way that for every compound of the solution, its production would be higher than its consumption (this would translate into having, for each compound, the sum of the stoichiometric coefficients greater than one). Using the stoichiometry in MultiPus would allow the inclusion of cycles in the solutions. Moreover, this would alleviate the need to *a priori* filter out the cofactors and coenzymes of the network. However, in the case of Totoro, the definition of a stoichiometrically valid solution remains not fully clear.

The biological problems addressed during this PhD led to different mathematical formulations

and different theoretical and computational problems. There is always the matter of getting an accurate (even if simplistic) definition that is also tractable. Starting from a biological question, the proposed model may not always be easy to compute, thus obtaining one solution, or enumerating several is in itself an important problem. We used different approaches based on graph theory or constraint-based modelling to obtain biologically relevant solutions. The modelling approaches required some simplifications, but nevertheless led to interesting results.

The approaches where we obtain generic models are thus useful, and now it is possible in all methods to work by adding up layers to our models to include known regulation or signalling, but but also known biochemistry such as stoichiometry or thermodynamics. In all cases, to obtain tractable problems, we aim to work in collaboration with computer scientists, biologists and bioinformaticians.

Bibliography

Acuña, V., Birmelé, E., Cottret, L., Crescenzi, P., Jourdan, F., Lacroix, V., Marchetti-Spaccamela, A., Marino, A., Milreu, P. V., Sagot, M.-F., and Stougie, L. (2012a). Telling stories: Enumerating maximal directed acyclic graphs with a constrained set of sources and targets. *Theoretical Computer Science*, 457:1–9.

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

Alexander, D. C. and Jensen, S. E. (1998). Investigation of the *Streptomyces clavuligerus* Cephamycin C Gene Cluster and Its Regulation by the CcaR Protein. *Journal of Bacteriology*, 180(16):4068–4079.

Andrade, R., Wannagat, M., Klein, C. C., Milreu, P. V., Stougie, L., and Sagot, M.-f. (2016). Enumeration of minimal stoichiometric precursor sets in metabolic networks. *Algorithms for Molecular Biology*, pages 1–23.

Arita, M. (2004). The metabolic world of Escherichia coli is not small. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1543–1547.

Aung, H. W., Henry, S. A., and Walker, L. P. (2013). Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Industrial Biotechnology*, 9(4):215–228.

Ausiello, G., Franciosa, P. G., and Frigioni, D. (2001). Directed hypergraphs: Problems, algorithmic results, and a novel decremental approach. In *Italian Conference on Theoretical Computer Science*, pages 312–328. Springer.

Balch, W. E., Fox, G. E., Magrum, L. J., Woese, C. R., and Wolfe, R. S. (1979). Methanogens: reevaluation of a unique biological group. *Microbiological Reviews*, 43(2):260–296.

Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., and Rabinowitz, J. D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nature Chemical Biology*, 5(8):593–599.

Bernstein, H. C. and Carlson, R. P. (2012). Microbial Consortia Engineering for Cellular Factories: in vitro to in silico systems. *Computational and Structural Biotechnology Journal*, 3(4):e201210017.

Binns, M., de Atauri, P., Vlysidis, A., Cascante, M., and Theodoropoulos, C. (2015). Sampling with poling-based flux balance analysis: optimal versus sub-optimal flux space analysis of *Actinobacillus succinogenes*. *BMC Bioinformatics*, 16(1):49.

Bizukojc, M., Dietz, D., Sun, J., and Zeng, A. P. (2010). Metabolic modelling of syntrophic-like growth of a 1,3-propanediol producer, *Clostridium butyricum*, and a methanogenic archeon, *Methanosarcina mazei*, under anaerobic conditions. *Bioprocess and Biosystems Engineering*, 33(4):507–523.

Bock, A.-K., Prieger-Kraft, A., and Schönheit, P. (1994). Pyruvate—a novel substrate for growth and methane formation in methanosarcina barkeri. *Archives of Microbiology*, 161(1):33–46.

Bonatti, P., Calimeri, F., Leone, N., and Ricca, F. (2010). Answer set programming. In *A 25-year perspective on logic programming*, pages 159–182. Springer-Verlag.

Borassi, M., Crescenzi, P., Lacroix, V., Marino, A., Sagot, M.-F., and Milreu, P. V. (2013). Telling Stories Fast Via Linear-Time Delay Pitch Enumeration. In Bonifaci, V., Demetrescu, V., and Marchetti-Spaccamela, A., editors, *Experimental Algorithms*, pages 200–2011, Rome. Springer.

Bourdakos, N., Marsili, E., and Mahadevan, R. (2014). A defined co-culture of *Geobacter sulfurreducens* and *Escherichia coli* in a membrane-less microbial fuel cell. *Biotechnology and Bioengineering*, 111:709–781.

Brenner, K., You, L., and Arnold, F. H. (2008). Engineering microbial consortia: a new frontier in synthetic biology. *Trends in Biotechnology*, 26(9):483–489.

Bulteau, L., Julien-Laferrière, A., Lacroix, V., Parrot, D., and Sagot, M.-F. (2015). DINGHY: Dynamic Interactive Navigator for General Hypergraphs in Biology. In *Jobim*, Clermont-Ferrand.

Burgard, A. P., Nikolaev, E. V., Schilling, C. H., and Maranas, C. D. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research*, 14(2):301–12.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6):647–57.

Campbell, N. A., Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B. (2015). *Biology: a global approach*, chapter 5: Biological Macromolecules and Lipids. Pearson.

Carbonell, P., Fichera, D., Pandit, S. B., and Faulon, J.-L. (2012). Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Systems Biology*, 6(1):10.

Carbonell, P., Parutto, P., Herisson, J., Pandit, S. B., and Faulon, J.-L. (2014). XTMS: Pathway design in an eXTended metabolic space. *Nucleic Acids Research*, 42(W1):389–394.

Carbonell, P., Planson, A. G., Fichera, D., and Faulon, J.-L. (2011). A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Systems Biology*, 5(1):122.

Carneiro, S., Ferreira, E. C., and Rocha, I. (2013). Metabolic responses to recombinant bioprocesses in *Escherichia coli*. *Journal of Biotechnology*, 164(3):396–408.

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K., and Reuss, M. (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnology and Bioengineering*, 79(1):53–73.

Chater, K. and Bibb, M. (1997). Chapter 2. regulation of bacterial antibiotic production. In *Biotechnology vol. 6: Products of Secondary Metabolism*, pages 57–105. VCH, Weinheim.

Chen, J., Gomez, J. A., Höffner, K., Phalak, P., Barton, P. I., and Henson, M. A. (2016). Spatiotemporal modeling of microbial metabolism. *BMC Systems Biology*, 10(1):21.

Chen, W. and Qin, Z. (2011). Development of a gene cloning system in a fast-growing and moderately thermophilic *Streptomyces* species and heterologous expression of *Streptomyces* antibiotic biosynthetic gene clusters. *BMC Microbiology*, 11(1):243.

Cheng, K.-k. K., Liu, H.-j. J., and Liu, D.-h. H. (2005). Multiple growth inhibition of *Klebsiella pneumoniae* in 1,3-propanediol fermentation. *Cell*, 27(1):19–22.

Cho, A., Yun, H., Park, J. H., Lee, S. Y., and Park, S. (2010). Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Systems Biology*, 4(1):35.

Chowdhury, A., Zomorrodi, A. R., and Maranas, C. D. (2014). Bilevel optimization techniques in computational strain design. *Computers and Chemical Engineering*, 72:363–372.

Christensen, C. D., Hofmeyr, J. S., and Rohwer, J. M. (2015). Tracing regulatory routes in metabolism using generalised supply-demand analysis. *BMC Systems Biology*, pages 1–18.

Chubukov, V., Gerosa, L., Kochanowski, K., and Sauer, U. (2014). Coordination of microbial metabolism. *Nature Reviews Microbiology*, 12(5):327–340.

Chubukov, V., Uhr, M., Le Chat, L., Kleijn, R. J., Jules, M., Link, H., Aymerich, S., Stelling, J., and Sauer, U. (2013). Transcriptional regulation is insufficient to explain substrate-induced flux changes in *Bacillus subtilis*. *Molecular Systems Biology*, 9(709):709.

Cole, J. A., Kohler, L., Hedhli, J., and Luthey-Schulten, Z. (2015). Spatially-resolved metabolic cooperativity within dense bacterial colonies. *BMC Systems Biology*, 9:15.

Copeland, W. B., Bartley, B. a., Chandran, D., Galdzicki, M., Kim, K. H., Sleight, S. C., Maranas, C. D., and Sauro, H. M. (2012). Computational tools for metabolic engineering. *Metabolic Engineering*, 14(3):270–280.

Cornish-Bowden, A. (1991). Failure of channelling to maintain low concentrations of metabolic intermediates. *European Journal of Biochemistry*, 195(1):103–108.

Costa, R. S., Hartmann, A., and Vinga, S. (2015). Kinetic modeling of cell metabolism for microbial production. *Journal of Biotechnology*, 219:126–141.

Costa, R. S., Veríssimo, A., and Vinga, S. (2014). KiMoSys: a web-based repository of experimental data for KInetic MOdels of biological SYStems. *BMC Systems Biology*, 8(1):85.

Cottret, L., Milreu, P., Acuña, V., Marchetti-Spaccamela, A., Viduani Martinez, F., Sagot, M.-F., and Stougie, L. (2008). Enumerating sets of precursors of target metabolites in a metabolic network. In Crandall, K. and Lagergren, J., editors, *Lecture Notes in Computer Science*, pages 233–244. Springer.

Cottret, L., Milreu, P. V., Acuña, V., Marchetti-Spaccamela, A., Stougie, L., Charles, H., and Sagot, M.-F. (2010a). Graph-based analysis of the metabolic exchanges between two co-resident intracellular symbionts, *Baumannia cicadellinicola* and *Sulcia muelleri*, with their insect host, *Homalodisca coagulata*. *PLoS Computational Biology*, 6(9).

Cottret, L., Wildridge, D., Vinson, F., Barrett, M. P., Charles, H., Sagot, M.-F., and Jourdan, F. (2010b). MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Research*, 38(Web Server):W132–W137.

Croes, D., Couche, F., Wodak, S. J., and Van Helden, J. (2006). Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356(1):222–236.

da Silva, G. P., Mack, M., and Contiero, J. (2009). Glycerol: A promising and abundant carbon source for industrial microbiology. *Biotechnology Advances*, 27(1):30–39.

Daran-Lapujade, P., Jansen, M. L. A., Daran, J. M., Van Gulik, W., De Winde, J. H., and Pronk, J. T. (2004). Role of Transcriptional Regulation in Controlling Fluxes in Central Carbon Metabolism of *Saccharomyces cerevisiae*: A chemostat culture study. *Journal of Biological Chemistry*, 279(10):9125–9138.

Daran-Lapujade, P., Rossell, S., van Gulik, W. M., Luttik, M. A. H., de Groot, M. J. L., Slijper, M., Heck, A. J. R., Daran, J.-M., de Winde, J. H., Westerhoff, H. V., Pronk, J. T., and Bakker, B. M. (2007). The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proceedings of the National Academy of Sciences of the United States of America*, 104(40):15753–15758.

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

De Ruyter, P., Kuipers, O. P., and De Vos, W. M. (1996). Controlled gene expression systems for lactococcus lactis with the food-grade inducer nisin. *Applied and Environmental Microbiology*, 62(10):3662–3667.

Dersch, L. M., Beckers, V., and Wittmann, C. (2015). Green pathways: Metabolic network analysis of plant systems. *Metabolic Engineering*, 34:1–24.

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Müller, T. (2008). Identifying functional modules in protein-protein interaction networks: An integrated exact approach. *Bioinformatics*, 24(13):223–231.

Donaldson, D. S. and Mabbott, N. A. (2016). The influence of the commensal and pathogenic gut microbiota on prion disease pathogenesis. *Journal of General Virology*, 97(97):1725–1738.

Eng, A. and Borenstein, E. (2016). An algorithm for designing minimal microbial communities with desired metabolic capacities. *Bioinformatics*.

Faust, K., Croes, D., and van Helden, J. (2011). Prediction of metabolic pathways from genome-scale metabolic networks. *BioSystems*, 105(2):109–21.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. O. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, 3(121):121.

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

Fell, D. and Cornish-Bowden, A. (1997a). *Understanding the control of metabolism*, volume 2, chapter 3 Enzyme activity: the molecular basis for its regulation. Portland press London.

Fell, D. and Cornish-Bowden, A. (1997b). *Understanding the control of metabolism*, volume 2, chapter 1 Introduction: regulation and control. Portland press London.

Fell, D. A. (1992). Metabolic Control Analysis : a survey of its theoretical and experimental development. *Biochemical Journal*, 330(Pt 2):313–330.

Fellows, M. and Rosamond, F. (2007). The complexity ecology of parameters: an illustration using bounded max leaf number. In *Computation and Logic in the Real World*, pages 268–277. Springer.

Fellows, M. R., Lokshtanov, D., Misra, N., Mnich, M., Rosamond, F. A., and Saurabh, S. (2009). The complexity ecology of parameters: An illustration using bounded max leaf number. *Theory of Computing Systems*, 45(4):822–848.

Field, J. a., Stams, A. J. M., Kato, M., and Schraa, G. (1995). Enhanced biodegradation of aromatic pollutants in cocultures of anaerobic and aerobic bacterial consortia. *Antonie van Leeuwenhoek*, 67(1):47–77.

Fiermonte, G., Dolce, V., Palmieri, L., Ventura, M., Runswick, M. J., Palmieri, F., and Walker, J. E. (2001). Identification of the human mitochondrial oxodicarboxylate carrier bacterial expression, reconstitution, functional characterization, tissue distribution, and chromosomal location. *Journal of Biological Chemistry*, 276(11):8225–8230.

Franssens, V., Bynens, T., Van Den Brande, J., Vandermeeren, K., Verduyckt, M., and Winderickx, J. (2013). The benefits of humanized yeast models to study Parkinson's disease. *Oxidative Medicine and Cellular Longevity*, 2013.

Friedmann, H. and Zeng, A.-P. (2013). Process and apparatus for the microbial production of a specific product and methane. US Patent 8,426,162.

Gallo, G., Longo, G., Pallottino, S., and Nguyen, S. (1993). Directed hypergraphs and applications. *Discrete applied mathematics*, 42(2):177–201.

Garey, M. R. and Johnson, D. S. (1979). Computers and intractability: a guide to the theory of np-completeness. 1979. *San Francisco, LA: Freeman*.

Gebser, M., Kaminski, R., Kaufmann, B., and Schaub, T. (2014a). Clingo = ASP + Control: Preliminary Report. *Technical Communications of the Thirtieth International Conference on Logic Programming (ICLP'14)*, pages 1–9.

Gebser, M., Kaminski, R., Kaufmann, B., and Schaub, T. (2014b). *Clingo* = ASP + control: Preliminary report. In Leuschel and Schrijvers (2014). Theory and Practice of Logic Programming, Online Supplement.

Getsin, I., Nalbandian, G. H., Yee, D. C., Vastermark, A., Paparoditis, P. C., Reddy, V. S., and Saier, M. H. (2013). Comparative genomics of transport proteins in developmental bacteria: *Myxococcus xanthus* and *Streptomyces coelicolor*. *BMC Microbiology*, 13(1):279.

Gunnigle, E., McCay, P., Fuszard, M., Botting, C. H., Abram, F., and O'Flaherty, V. (2013). A functional approach to uncover the low-temperature adaptation strategies of the archaeon *Methanosarcina barkeri*. *Applied and Environmental Microbiology*, 79(14):4210–4219.

Hadadi, N. and Hatzimanikatis, V. (2015). Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Current Opinion in Chemical Biology*, 28:99–104.

Halweg-Edwards, A. L., Grau, W. C., Winkler, J. D., Garst, A. D., and Gill, R. T. (2015). The emergence of commodity-scale genetic manipulation. *Current Opinion in Chemical Biology*, 28:150–155.

Hanly, T. J. and Henson, M. A. (2011). Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures. *Biotechnology and Bioengineering*, 108(2):376–385.

Hatzimanikatis, V., Li, C., Ionita, J. a., Henry, C. S., Jankowski, M. D., and Broadbelt, L. J. (2005). Exploring the diversity of complex metabolic networks. *Bioinformatics (Oxford, England)*, 21(8):1603–9.

Henriques-Normark, B. and Normark, S. (2010). Commensal pathogens, with a focus on *Streptococcus pneumoniae*, and interactions with the human host. *Experimental Cell Research*, 316(8):1408–1414.

Henson, M. A. (2015). Genome-scale modelling of microbial metabolism with temporal and spatial resolution. *Biochemical Society Transactions*, 43(6):1164–1171.

Henson, M. A. and Hanly, T. J. (2014). Dynamic flux balance analysis for synthetic microbial communities. *IET Systems Biology*, 8(5):214–29.

Hjersted, J. L. and Henson, M. a. (2009). Steady-state and dynamic flux balance analysis of ethanol production by *Saccharomyces cerevisiae*. *IET Systems Biology*, 3(August 2008):167–179.

Hollinshead, W., He, L., and Tang, Y. J. (2014). Biofuel production: An odyssey from metabolic engineering to fermentation scale-up. *Frontiers in Microbiology*, 5(JULY):1–8.

Hoops, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). COPASI - A COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074.

Hoppe, A., Hoffmann, S., and Holzhütter, H.-G. (2007). Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Systems Biology*, 1:23.

Hosoda, K., Suzuki, S., Yamauchi, Y., Shiroguchi, Y., Kashiwagi, A., Ono, N., Mori, K., and Yomo, T. (2011). Cooperative Adaptation to Establishment of a Synthetic Bacterial Mutualism. *PLoS ONE*, 6(2):e17105.

Huang, C. J., Lin, H., and Yang, X. M. (2012). Industrial production of recombinant therapeutics in *Escherichia coli* and its recent advancements. *Journal of Industrial Microbiology & Biotechnology*, 39(3):383–399.

Hube, B. (2004). From commensal to pathogen: Stage- and tissue-specific gene expression of *Candida albicans*. *Current Opinion in Microbiology*, 7(4):336–341.

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

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., Hirasawa, T., Naba, M., Hirai, K., Hoque, A., Ho, P. Y., Kakazu, Y., Sugawara, K., Igarashi, S., Harada, S., Masuda, T., Sugiyama, N., Togashi, T., Hasegawa, M., Takai, Y., Yugi, K., Arakawa, K., Iwata, N., Toya, Y., Nakayama, Y., Nishioka, T., Shimizu, K., Mori, H., Masaru, T., and Tomita, M. (2007). Multiple High-Throughput Analyses Monitor the Response of *E. coli* to Perturbations. *Science*, 316(April):593–598.

Issel-Tarver, L., Christie, K. R., Dolinski, K., Andrada, R., Balakrishnan, R., Ball, C. A., Binkley, G., Dong, S., Dwight, S. S., Fisk, D. G., et al. (2002). *Saccharomyces* genome database. *Methods in Enzymology*, 350:329–346.

Jagmann, N. and Philipp, B. (2014). Design of synthetic microbial communities for biotechnological production processes. *Journal of Biotechnology*, 184:209–218.

Jensen, P. R. and Hammer, K. (1998). Artificial promoters for metabolic optimization. *Biotechnology and Bioengineering*, 58(2-3):191–195.

Jewison, T., Knox, C., Neveu, V., Djoumbou, Y., Guo, A. C., Lee, J., Liu, P., Mandal, R., Krishnamurthy, R., Sinelnikov, I., Wilson, M., and Wishart, D. S. (2012). YMDB: The yeast metabolome database. *Nucleic Acids Research*, 40(D1):815–820.

Jose, P. A., Robinson, S., and Jebakumar, D. (2013). Non-streptomycete actinomycetes nourish the current microbial antibiotic drug discovery. *Frontiers in Microbiology*, 4(August):2008–2010.

Julien-Laferrière, A., Bulteau, L., Parrot, D., Marchetti-Spaccamela, A., Stougie, L., Vinga, S., Mary, A., and Sagot, M.-F. (2016). A Combinatorial Algorithm for Microbial Consortia Synthetic Design. *Scientific Reports*, 6(July):29182.

Jung, M.-Y., Mazumdar, S., Shin, S. H., Yang, K.-S., Lee, J., and Oh, M.-K. (2014). Improvement of 2,3-butanediol yield in *Klebsiella pneumoniae* by deletion of the pyruvate formate-lyase gene. *Applied and Environmental Microbiology*, 80(19):6195–203.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

Kästner, C. N., Schneider, K., Dimroth, P., and Pos, K. M. (2002). Characterization of the citrate/acetate antiporter CitW of *Klebsiella pneumoniae*. *Archives of Microbiology*, 177(6):500–506.

Katz, L. and Baltz, R. H. (2016). Natural product discovery: past, present, and future. *Journal of Industrial Microbiology & Biotechnology*.

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

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

Khodayari, A., Zomorrodi, A. R., Liao, J. C., and Maranas, C. D. (2014). A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. *Metabolic Engineering*, 25:50–62.

Kim, M. K. and Lun, D. S. (2014). Methods for integration of transcriptomic data in genome-scale metabolic models. *Computational and Structural Biotechnology Journal*, 11(18):59–65.

Klamt, S., Hädicke, O., and von Kamp, A. (2014). Stoichiometric and constraint-based analysis of biochemical reaction networks. In *Large-Scale Networks in Engineering and Life Sciences*, pages 263–316. Springer.

Klein, C. C., Marino, A., Sagot, M.-F., Vieira Milreu, P., and Brilli, M. (2012). Structural and dynamical analysis of biological networks. *Briefings in Functional Genomics*, 11(6):420–33.

Koch, S., Benndorf, D., Fronk, K., Reichl, U., and Klamt, S. (2016). Predicting compositions of microbial communities from stoichiometric models with applications for the biogas process. *Biotechnology for Biofuels*, 9(1):17.

Koesnandar, Nishio, N., Kuroda, K., and Nagai, S. (1990). Methanogenesis of glucose by defined thermophilic coculture of *Clostridium thermoaceticum* and *Methanosarcina* sp. *Journal of Fermentation and Bioengineering*, 70(6):398–403.

Lafaye, A., Junot, C., Pereira, Y., Lagniel, G., Tabet, J. C., Ezan, E., and Labarre, J. (2005). Combined proteome and metabolite-profiling analyses reveal surprising insights into yeast sulfur metabolism. *Journal of Biological Chemistry*, 280(26):24723–24730.

Leader, D. P., Burgess, K., Creek, D., and Barrett, M. P. (2011). Pathos: A web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry. *Rapid Communications in Mass Spectrometry*, 25(22):3422–3426.

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

Leuschel, M. and Schrijvers, T., editors (2014). *Technical Communications of the Thirtieth International Conference on Logic Programming (ICLP'14)*, volume 14(4-5). Theory and Practice of Logic Programming, Online Supplement.

Lewis, N. E., Nagarajan, H., and Palsson, B. Ø. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 81(4):291–305.

Lima-Mendez, G. and van Helden, J. (2009). The powerful law of the power law and other myths in network biology. *Molecular bioSystems*, 5(12):1482–1493.

Link, H., Fuhrer, T., Gerosa, L., Zamboni, N., and Sauer, U. (2015). Real-time metabolome profiling of the metabolic switch between starvation and growth. *Nature Methods*, 12(11):1091–1097.

Link, H., Kochanowski, K., and Sauer, U. (2013). Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo. *Nature Biotechnology*, 31(4):357–61.

Lynd, L. R., Weimer, P. J., VanZyl, W. H., and Pretorius, I. S. (2002). Microbial cellulose utilization: Fundamentals and Biotechnology. *Microbiology and Molecular Biology Reviews*, 66(3):506–577.

Machado, D., Costa, R. S., Rocha, M., Ferreira, E. C., Tidor, B., and Rocha, I. (2011). Modeling formalisms in Systems Biology. *AMB Express*, 1(1):45.

Machado, D. and Herrgard, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Computational Biology*, 10(4):e1003580.

Madalinski, G., Godat, E., Alves, S., Lesage, D., Genin, E., Levi, P., Labarre, J., Tabet, J.-C., Ezan, E., and Junot, C. (2008). Direct introduction of biological samples into a LTQ-Orbitrap hybrid mass spectrometer as a tool for fast metabolome analysis. *Analytical Chemistry*, 80(9):3291–303.

Mahadevan, R. and Henson, M. A. (2012). Genome-based Modeling and Design of Metabolic Interactions in Microbial Communities. *Computational and Structural Biotechnology Journal*, 3(October):e201210008.

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

Masset, J., Calusinska, M., Hamilton, C., Hiligsmann, S., Joris, B., Wilmotte, A., and Thonart, P. (2012). Fermentative hydrogen production from glucose and starch using pure strains and artificial co-cultures of clostridium spp. *Biotechnology for Biofuels*, 5(1):1.

Mazumdar, V., Snitkin, E. S., Amar, S., and Segrè, D. (2009). Metabolic network model of a human oral pathogen. *Journal of Bacteriology*, 91(1):74–90.

Mendoza-Cózatl, D., Loza-Tavera, H., Hernández-Navarro, A., and Moreno-Sánchez, R. (2005). Sulfur assimilation and glutathione metabolism under cadmium stress in yeast, protists and plants. *FEMS Microbiology Reviews*, 29(4):653–671.

Metallo, C. M. and Vander Heiden, M. G. (2013). Understanding Metabolic Regulation and Its Influence on Cell Physiology. *Molecular Cell*, 49(3):388–398.

Milreu, P. V. (2012). *Enumerating Functional Substructures of Genome-Scale Metabolic Networks: Stories, Precursors and Organisations*. PhD thesis, Université Lyon 1.

Milreu, P. V., Klein, C. C., Cottret, L., Acuña, V., Birmelé, E., Borassi, M., Junot, C., Marchetti-Spaccamela, A., Marino, A., Stougie, L., Jourdan, F., Crescenzi, P., Lacroix, V., and Sagot, M.-F. F. (2014). Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure. *Bioinformatics (Oxford, England)*, 30(1):61–70.

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

Mnif, I., Mnif, S., Sahnoun, R., Maktouf, S., Ayedi, Y., Ellouze-Chaabouni, S., and Ghribi, D. (2015). Biodegradation of diesel oil by a novel microbial consortium: Comparison between co-inoculation with biosurfactant-producing strain and exogenously added biosurfactants. *Environmental Science and Pollution Research*, 22:14852–14861.

Momeni, B., Chen, C.-C., Hillesland, K. L., Waite, A., and Shou, W. (2011). Using artificial systems to explore the ecology and evolution of symbioses. *Cellular and Molecular Life Sciences*, 68(8):1353–1368.

Nakamura, C. E. and Whited, G. M. (2003). Metabolic engineering for the microbial production of 1,3-propanediol. *Current Opinion in Biotechnology*, 14(5):454–459.

Nuñez, L., Méndez, C., Brãna, A., Blanco, G., and Salas, J. A. (2003). The Biosynthetic Gene Cluster for the Beta-Lactam Carbapenem Thienamycin in *Streptomyces cattleya*. *Chemistry & biology*, 10:301–311.

Oliveira, N. M., Niehus, R., and Foster, K. R. (2014). Evolutionary limits to cooperation in microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 111(50):201412673.

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

Palmieri, L., Agrimi, G., Runswick, M. J., Fearnley, I. M., Palmieri, F., Walker, J. E., L, P., G., A., MJ, R., IM, F., F, P., and JE., W. (2001). Identification in *Saccharomyces cerevisiae* of two isoforms of a novel mitochondrial transporter for 2-oxoadipate and 2-oxoglutarate. *Journal of Biological Chemistry*, 276(3):1916–1922.

Park, J. O., Rubin, S. A., Xu, Y.-F., Amador-Noguez, D., Fan, J., Shlomi, T., and Rabinowitz, J. D. (2016). Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nature Chemical Biology*, advance on(7):482–489.

Pharkya, P. and Maranas, C. D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic Engineering*, 8(1):1–13.

Reznik, E., Mehta, P., and Segrè, D. (2013). Flux Imbalance Analysis and the Sensitivity of Cellular Growth to Changes in Metabolite Pools. *PLoS Computational Biology*, 9(8).

Ro, D.-K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. a., Eachus, R. a., Ham, T. S., Kirby, J., Chang, M. C. Y., Withers, S. T., Shiba, Y., Sarpong, R., and Keasling, J. D. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086):940–943.

Rodrigo, G., Carrera, J., Prather, K. J., and Jaramillo, A. (2008). DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics (Oxford, England)*, 24(21):2554–6.

Rohwer, J. M. and Hofmeyr, J.-H. S. (2008). Identifying and characterising regulatory metabolites with generalised supply-demand analysis. *Journal of Theoretical Biology*, 252(3):546–54.

Rollié, S., Mangold, M., and Sundmacher, K. (2012). Designing biological systems: Systems Engineering meets Synthetic Biology. *Chemical Engineering Science*, 69(1):1–29.

Ryll, A., Bucher, J., Bonin, A., Bongard, S., Gonçalves, E., Saez-Rodriguez, J., Niklas, J., and Klamt, S. (2014). A model integration approach linking signalling and gene-regulatory logic with kinetic metabolic models. *Biosystems*, 124:26–38.

Sabra, W., Dietz, D., Tjahjasari, D., and Zeng, A.-P. (2010). Biosystems analysis and engineering of microbial consortia for industrial biotechnology. *Engineering in Life Sciences*, 10(5):407–421.

Sauro, H. M. (2009). Network Dynamics. In Ireton, R., Montgomery, K., Bumgarner, R., Samudrala, R., and McDermott, J., editors, *Computational Systems Biology*, Methods in Molecular Biology, pages 269—-309. Springer, Totowa, NJ.

Saxena, R., Anand, P., Saran, S., and Isar, J. (2009). Microbial production of 1,3-propanediol: Recent developments and emerging opportunities. *Biotechnology Advances*, 27(6):895–913.

Schilling, C. H., Letscher, D., and Palsson, B. Ø. (2000). Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective. *Journal of Theoretical Biology*, 203(3):229–248.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional Optimality of Microbial Metabolism. *Science*, 336:601–604.

Schultz, A. and Qutub, A. a. (2015). Predicting internal cell fluxes at sub-optimal growth. *BMC Systems Biology*, 9(1):18.

Schuster, S., Pfeiffer, T., and Fell, D. A. (2008). Is maximization of molar yield in metabolic networks favoured by evolution? *Journal of Theoretical Biology*, 252(3):497–504.

Schwender, J., König, C., Klapperstück, M., Heinzel, N., Munz, E., Hebbelmann, I., Hay, J. O., Denolf, P., De Bodt, S., Redestig, H., Caestecker, E., Jakob, P. M., Borisjuk, L., and Rolletschek, H. (2014). Transcript abundance on its own cannot be used to infer fluxes in central metabolism. *Frontiers in Plant Science*, 5(November):668.

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

Shetty, P. H. and Jespersen, L. (2006). *Saccharomyces cerevisiae* and lactic acid bacteria as potential mycotoxin decontaminating agents. *Trends in Food Science and Technology*, 17(2):48–55.

Shivlata, L. and Satyanarayana, T. (2015). Thermophilic and alkaliphilic Actinobacteria: Biology and potential applications. *Frontiers in Microbiology*, 6(SEP):1–29.

Shou, W., Ram, S., and Vilar, J. M. G. (2007). Synthetic cooperation in engineered yeast populations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1877–82.

Shulaev, V. (2006). Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, 7(2):128–139.

Stanford, N. J., Millard, P., and Swainston, N. (2015). RobOKoD: microbial strain design for (over)production of target compounds. *Frontiers in Cell and Developmental Biology*, 3(March):1–12.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–3.

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. a., and Stahl, D. a. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology*, 3(92):92.

Storey, K. B. (2005). *Functional metabolism: regulation and adaptation*, chapter 2 Enzymes, the basis of catalysis. John Wiley & Sons.

Tepper, N., Noor, E., Amador-Noguez, D., Haraldsdóttir, H. S., Milo, R., Rabinowitz, J., Liebermeister, W., and Shlomi, T. (2013). Steady-State Metabolite Concentrations Reflect a Balance between Maximizing Enzyme Efficiency and Minimizing Total Metabolite Load. *PLoS ONE*, 8(9):1–13.

Tepper, N. and Shlomi, T. (2009). Predicting metabolic engineering knockout strategies for chemical production: Accounting for competing pathways. *Bioinformatics*, 26(4):536–543.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Timan, D. (1982). *Resource Competition and Community Structure*. Princeton Press.

Tran, L. M., Rizk, M. L., and Liao, J. C. (2008). Ensemble modeling of metabolic networks. *Biophysical Journal*, 95(12):5606–5617.

Uno, T., Asai, T., Uchida, Y., and Arimura, H. (2003). LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets. *Fimi*, 90.

Uno, T., Asai, T., Yuzo, U., and Arimura, H. (2004). An efficient algorithm for enumerating closed patterns in transaction databases. *Discovery Science*, 3245:16–31.

Uno, T., Kiyomi, M., and Arimura, H. (2005). LCM ver . 3 : Collaboration of Array , Bitmap and Prefix Tree for Frequent Itemset Mining. In ACM, editor, *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pages 77–86.

van Helden, J., Wernisch, L., Gilbert, D., and Wodak, S. J. (2002). Graph-based analysis of metabolic networks. *Ernst Schering Research Foundation workshop*, (38):245–274.

Vido, K., Spector, D., Lagniel, G., Lopez, S., Toledano, M. B., and Labarre, J. (2001). A Proteome Analysis of the Cadmium Response in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, 276(11):8469–8474.

Volesky, B., May, H., and Holan, Z. R. (1993). Cadmium biosorption by *Saccharomyces cerevisiae*. *Biotechnology and Bioengineering*, 41(8):826–829.

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

Wegner, A., Meiser, J., Weindl, D., and Hiller, K. (2015). How metabolites modulate metabolic flux. *Current Opinion in Biotechnology*, 34:16–22.

Wei, B., Shin, S., Laporte, D., Wolfe, A. J., and Romeo, T. (2000). Global regulatory mutations in csrA and rpoS cause severe central carbon stress in *Escherichia coli* in the presence of acetate. *Journal of Bacteriology*, 182(6):1632–1640.

Wintermute, E. H. and Silver, P. A. (2010). Dynamics in the mixed microbial concourse. *Genes & Development*, 24(23):2603–2614.

Wu, A. and Ross, D. (2016). Evolutionary Game between Commensal and Pathogenic Microbes in Intestinal Microbiota. *Games*.

Wysocki, R. and Tamás, M. J. (2010). How *Saccharomyces cerevisiae* copes with toxic metals and metalloids. *FEMS Microbiology Reviews*, 34(6):925–951.

Zamboni, N., Fendt, S.-M., Rühl, M., and Sauer, U. (2009). 13C-Based Metabolic Flux Analysis. *Nature Protocols*, 4(6):878–892.

Zampieri, M. and Sauer, U. (2016). Model-based media selection to minimize the cost of metabolic cooperation in microbial ecosystems. *Bioinformatics*, 32(February).

Zhang, C., Ji, B., Mardinoglu, A., Nielsen, J., and Hua, Q. (2015). Logical transformation of genome-scale metabolic models for gene level applications and analysis. *Bioinformatics*, 31(14):2324–2331.

Zomorrodi, A. R., Islam, M. M., and Maranas, C. D. (2014). D-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synthetic Biology*, 3(4):247–257.

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

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

Appendix A

Quantifying the transient states

A.1 Model of *Escherichia coli* in response to glucose pulse

A.1.1 Results of the time-course simulations

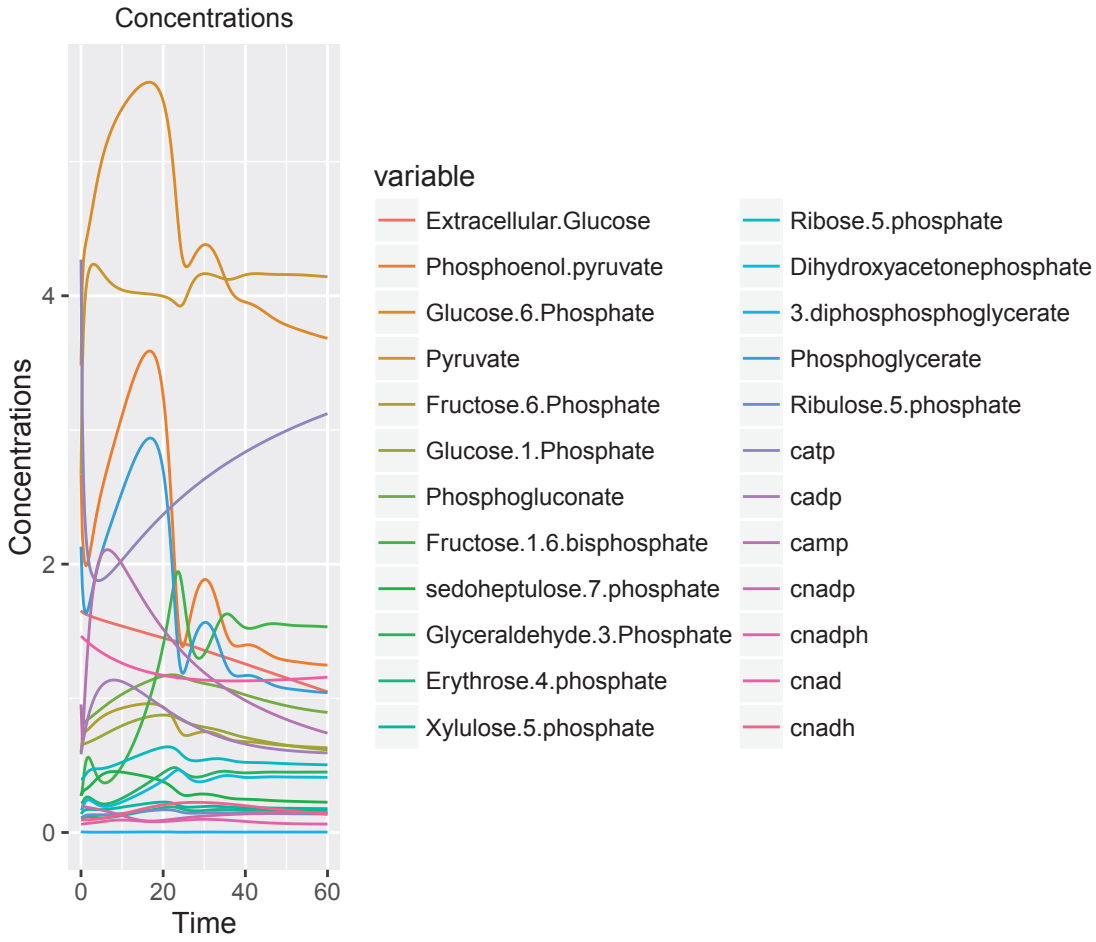


Figure A.1: Simulated concentrations along time.

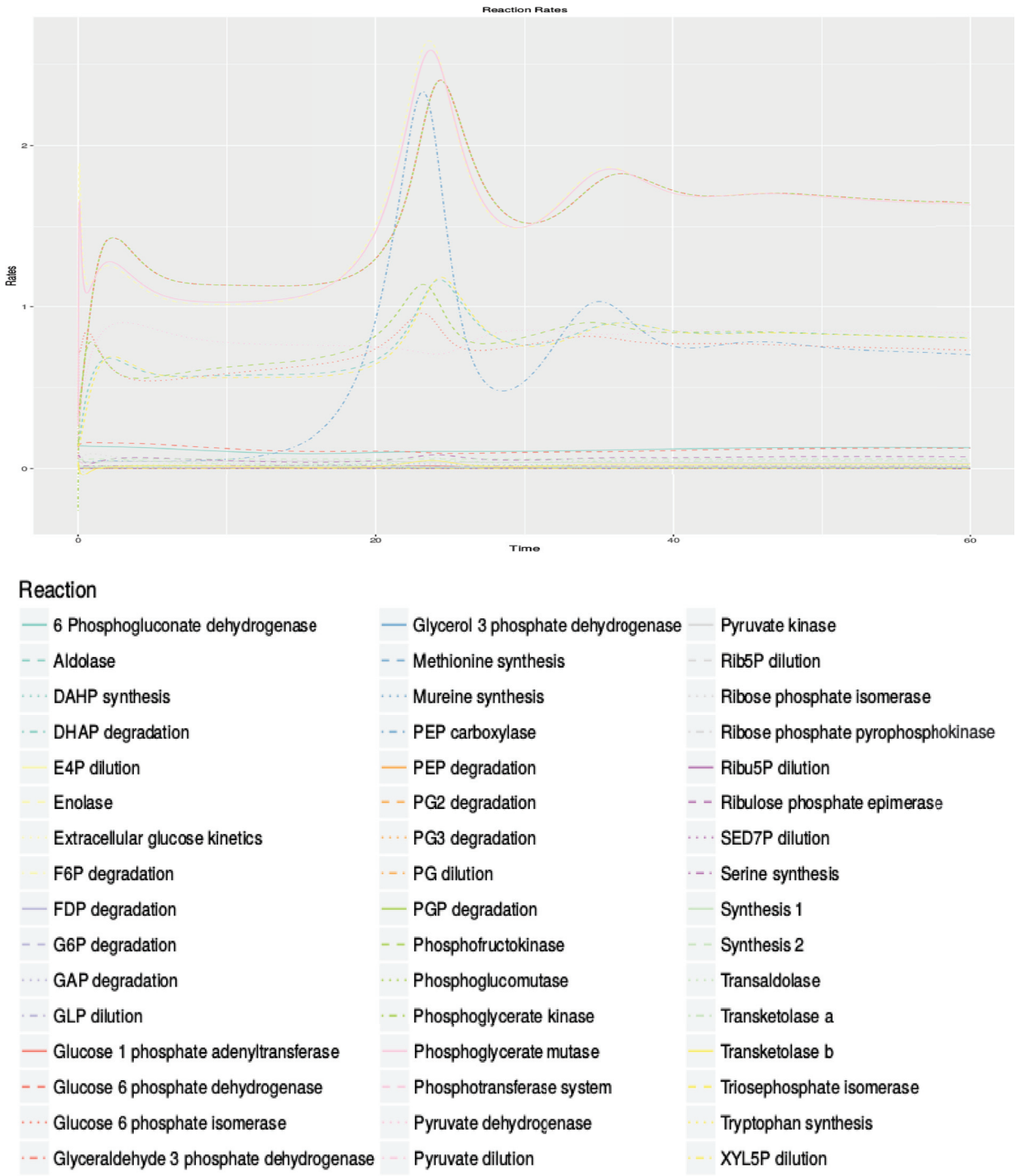


Figure A.2: Simulated reactions rates along time.

A.1.2 The model variables

Metabolite identifier	Metabolite name
cpep	Phosphoenol pyruvate
cglcex	Extracellular Glucose
cg6p	Glucose-6-Phosphate
cpyr	Pyruvate
cf6p	Fructose-6-Phosphate
cg1p	Glucose-1-Phosphate
cpg	6-Phosphogluconate
cfdp	Fructose-1,6-bisphosphate
csed7p	sedoheptulose-7-phosphate
cgap	Glyceraldehyde-3-Phosphate
ce4p	Erythrose-4-phosphate
cxyl5p	Xylulose-5-phosphate
crib5p	Ribose-5-phosphate
cdhap	Dihydroxyacetonephosphate
cpgp	1,3-diphosphosphoglycerate
cpg3	3-Phosphoglycerate
cpg2	2-Phosphoglycerate
cribu5p	Ribulose-5-phosphate

Table A.1: Identifiers and names of the metabolites in the model described by (Chassagnole et al., 2002)

A.1.3 KOTOURA results

Reaction identifier	Reaction name
vPTS	Phosphotransferase system
vPGI	Glucose-6-phosphate isomerase
vPGM	Phosphoglucomutase
vG6PDH	Glucose-6-phosphate dehydrogenase
vPFK	Phosphofructokinase
vTA	Transaldolase
vTKA	Transketolase a
vTKB	Transketolase b
vMURSyNTH	Mureine synthesis
vALDO	Aldolase
vGAPDH	Glyceraldehyde-3-phosphate dehydrogenase
vTIS	Triosephosphate isomerase
vTRPSYNTH	Tryptophan synthesis
vG3PDH	Glycerol-3-phosphate dehydrogenase
vPGK	Phosphoglycerate kinase
vsersynth	Serine synthesis
vrpGluMu	Phosphoglycerate mutase
vENO	Enolase
vPK	Pyruvate kinase
vpepCxylase	PEP carboxylase
vSynth1	Synthesis 1
vSynth2	Synthesis 2
vDAHPS	DAHP synthesis
vPDH	Pyruvate dehydrogenase
vMethSynth	Methionine synthesis
vPGDH	6-Phosphogluconate dehydrogenase
vR5PI	Ribose-phosphate isomerase
vRu5P	Ribulose-phosphate epimerase
vPPK	Ribose phosphate pyrophosphokinase
vG1PAT	Glucose-1-phosphate adenyltransferase
vG6P	G6P degradation
vf6P	F6P degradation
vfdP	FDP degradation
vGAP	GAP degradation
vDHAP	DHAP degradation
vPGP	PGP degradation
vPG3	PG3 degradation
vpg2	PG2 degradation
vPEP	PEP degradation
vRibu5p	Ribu5P dilution
vRIB5P	Rib5P dilution
vXYL5P	XYL5P dilution
vSED7P	SED7P dilution
vpyr	Pyruvate dilution
vPG	PG dilution
vE4P	E4P dilution
vGLP	GLP dilution
vEXTER	Extracellular glucose kinetics

Table A.2: Identifier and names of the reactions in the model described by (Chassagnole et al., 2002)

Reaction	AUC	φ_1^*	φ_2^*
vALDO	45.886	46.5455	47.1926
vDAHPS	1.2265	1.2388	1.2388
vDHAP	$6 \cdot 10^{-4}$	$6 \cdot 10^{-4}$	$6 \cdot 10^{-4}$
vE4P	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$
vENO	92.6987	92.777	93.4242
vEXTER	0.1828	0.181	0.181
vf6P	0.0013	0.0013	0.0013
vfdP	0.0021	0.0021	0.0021
vG1PAT	0.5757	0.5814	0.5814
vG3PDH	0.183	0.1848	0.1848
vG6P	0.0074	0.0075	0.0075
vG6PDH	7.0469	3.4309	1.4894
vGAP	$7 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$7 \cdot 10^{-4}$
vGAPDH	92.3413	92.4362	93.0834
vGLP	0.0012	0.0013	0.0013
vMethSynth	0.1358	0.1344	0.1344
vMURSyNTH	0.0262	0.0265	0.0265
vPDH	48.8202	48.4216	49.1399
vPEP	0.0034	0.0034	0.0034
vpepCxylase	38.488	38.8729	38.8729
vPFK	47.1484	47.8206	48.4678
vPG	0.0017	0.0017	0.0017
vpg2	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
vPG3	0.0028	0.0028	0.0028
vPGDH	6.9584	3.3415	1.4
vPGI	43.5006	46.6032	48.5447
vPGK	92.3375	92.4411	93.0883
vPGM	0.5401	0.5462	0.5462
vPGP	0	0	0
vPK	2.3451	2.5117	3.1589
vPPK	0.6514	0.6579	0.6579
vPTS	0.7894	0.7814	0.7814
vpyr	0.0068	0.0069	0.0068
vR5PI	3.2161	2.0198	1.3726
vRIB5P	$9 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	$9 \cdot 10^{-4}$
vRibu5p	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-4}$
vrpGluMu	92.4947	92.5719	93.2191
vRu5P	**3.7152**	**1.2943**	**0**
vSED7P	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
vsersynth	0.9337	0.943	0.943
vSynth1	0.7582	0.7658	0.7658
vSynth2	3.55	3.5855	3.5145
vTA	2.5078	1.3033	0.6561
vTIS	45.4585	46.1137	46.7609
vTKA	2.4578	1.2539	0.6067
vTKB	**1.2174**	**0**	**-0.6472**
vTRPSYNTH	0.0622	0.0616	0.0616
vXYL5P	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$

Table A.3: Values for the φ^* computed and the corresponding simulated value (AUC: Area Under the Curve) for all reactions. All numbers were rounded to the 4^{th} decimal. In bold, we show where the reaction supports of the y^* differ between the two solutions.

name	Δ^{min}	Δ^{max}
3pg	-0.6628	-0.2442
adp	-0.7581	-0.1473
amp	-0.7476	-0.3573
atp	0.8716	7.3856
dhap	-0.1106	-0.0405
f6p	-0.0515	-0.0239
fdp	-0.0268	-0.0067
fum	-0.0576	0.0142
g1p	0.696	7.8254
g6p	-0.195	-0.0941
glu_L	-3.1798	-1.9137
mal_L	-0.0866	-0.0429
nad	-0.2465	2.3941
nadp	0.0299	0.3996
nadph	-0.1879	-0.1178
oaa	-0.0106	0.0143
pep	-0.1529	-0.0405
pyr	-0.1472	-0.0928
r5p	-0.0657	-0.0068
ru5p_D	-0.2	-0.0554
s7p	-0.286	-0.157
succ	-0.022	0.0067
13dpg	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
2dda7p	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
2ddg6p	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
2pg	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
6pgc	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
6pgl	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
ac	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
acald	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
accoa	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
acon_C	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
actp	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
adpglc	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
akg	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
Mit	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
Mo2	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
e4p	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
etoh	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
for	0	0
g3p	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
glc_D	0	0
gln_L	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
glx	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
h	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
h2o	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
icit	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
lac_D	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$

name	Δ^{min}	Δ^{max}
nadh	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
nh4	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
o2	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
pi	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
q8	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
q8h2	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
rib_D	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
succoa	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
xu5p_D	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
ppi	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
ac_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
acald_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
akg_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
Mo2_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
etoh_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
for_e	0	0
fru_e	0	0
fum_e	0	0
glc_D_e	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
gln_L_e	0	0
glu_L_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
h_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
h2o_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
lac_D_e	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
mal_L_e	0	0
nh4_e	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
o2_e	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
pi_e	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
pyr_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
succ_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
e4p_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
s7p_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
ru5p_D_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
oaa_e	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
succoa_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
accoa_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
dhap_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
3pg_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
f6p_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
g3p_e	$1 \cdot 10^{-4}$	$1 \cdot 10^{4}$
pep_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
2dda7p_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$
adpglc_e	0	0
rib_D_e	$-1 \cdot 10^{4}$	$-1 \cdot 10^{-4}$

Table A.4: Metabolite variations used for the simulation

name	Δ^{min}	Δ^{max}
R_EX_3PG_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_OAA_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_mal_L_e_	0	0
R_EX_succ_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_E4P_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_ACCOA_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_pi_e_	$-1 \cdot 10^{5}$	$-1 \cdot 10^{-4}$
R_EX_fru_e_	0	0
R_EX_nh4_e_	$-1 \cdot 10^{5}$	$-1 \cdot 10^{-4}$
R_EX_RIB_D_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_pyr_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_etoh_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_S7P_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_glc_e_	$-1 \cdot 10^{5}$	$-1 \cdot 10^{-4}$
R_EX_h2o_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_DHAP_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_ac_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_ADPGLC_e_	0	0
R_EX_G3P_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_2DDA7P_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_RU5P_e_	$-1 \cdot 10^{5}$	$-1 \cdot 10^{-4}$
R_EX_for_e_	0	0
R_EX_akg_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_h_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_PEP_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_fum_e_	0	0
R_EX_gln_L_e_	0	0
R_EX_F6P_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_lac_D_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_glu_L_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_o2_e_	$-1 \cdot 10^{5}$	$-1 \cdot 10^{-4}$
R_EXo2_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_acald_e_	$1 \cdot 10^{-4}$	$1 \cdot 10^{5}$
R_EX_SUCCOA_e_	$-1 \cdot 10^{5}$	$-1 \cdot 10^{-4}$

Table A.5: Boundary metabolite variations used for the simulation

A.2 Predicting mutant knock-outs

Abbreviation	Name	Formula
2dda7p	2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate	C7H10O10P
2dda7p(e)	2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate	C7H10O10P
2ddg6p	2-Dehydro-3-deoxy-D-gluconate 6-phosphate	C6H8O9P
akg	2-Oxoglutarate	C5H4O5
akg(e)	2-Oxoglutarate	C5H4O5
3pg	3-Phospho-D-glycerate	C3H4O7P
3pg(e)	3-Phospho-D-glycerate	C3H4O7P
13dpg	3-Phospho-D-glyceroyl phosphate	C3H4O10P2
6pgc	6-Phospho-D-gluconate	C6H10O10P
6pgl	6-phospho-D-glucono-1,5-lactone	C6H9O9P
acald	Acetaldehyde	C2H4O
acald(e)	Acetaldehyde	C2H4O
ac	Acetate	C2H3O2
ac(e)	Acetate	C2H3O2
actp	Acetyl phosphate	C2H3O5P
accoa	Acetyl-CoA	C23H34N7O17P3S
acetylco(e)	Acetyl-CoA	C23H34N7O17P3S
adp	ADP	C10H12N5O10P2
adpglc	ADPglucose	C16H23N5O15P2
adpglc(e)	ADPglucose	C16H23N5O15P2
r5p	alpha-D-Ribose 5-phosphate	C5H9O8P
nh4	Ammonium	H4N
nh4(e)	Ammonium	H4N
amp	AMP	C10H12N5O7P
atp	ATP	C10H12N5O13P3
co2(e)	C02	C102
acon-C	cis-Aconitate	C6H3O6
cit	Citrate	C6H5O7
co2	CO2	CO2
e4p	D-Erythrose 4-phosphate	C4H7O7P
e4p(e)	D-Erythrose 4-phosphate	C4H7O7P
fru	D-Fructose	C6H12O6
fru(e)	D-Fructose	C6H12O6
fdp	D-Fructose 1,6-bisphosphate	C6H10O12P2
f6p	D-Fructose 6-phosphate	C6H11O9P
f6p(e)	D-Fructose 6-phosphate	C6H11O9P
glc-D	D-Glucose	C6H12O6
glc_D(e)	D-Glucose	C6H12O6
g1p	D-Glucose 1-phosphate	C6H11O9P
g6p	D-Glucose 6-phosphate	C6H11O9P
2pg	D-Glycerate 2-phosphate	C3H4O7P
lac-D	D-Lactate	C3H5O3
lac_D(e)	D-Lactate	C3H5O3
rib-D	D-Ribose	C5H10O5
rib_D(e)	D-Ribose	C5H10O5
ru5p-D	D-Ribulose 5-phosphate	C5H9O8P
ru5p_D(e)	D-Ribulose 5-phosphate	C5H9O8P
xu5p-D	D-Xylulose 5-phosphate	C5H9O8P
dhap	Dihydroxyacetone phosphate	C3H5O6P
dhap(e)	Dihydroxyacetone phosphate	C3H5O6P
ppi	Diphosphate	HO7P2
etoh	Ethanol	C2H6O
ethanol(e)	ethanol	C2H6O
for	Formate	CH1O2
for(e)	Formate	CH1O2
fum	Fumarate	C4H2O4
fum(e)	Fumarate	C4H2O4
g3p	Glyceraldehyde 3-phosphate	C3H5O6P

g3p(e)	Glyceraldehyde 3-phosphate	C3H5O6P
glx	Glyoxylate	C2H1O3
h(e)	h	H
h	H +	H
h2o	H2O	H2O
icit	Isocitrate	C6H5O7
glu-L	L-Glutamate	C5H8NO4
glu_L(e)	L-Glutamate	C5H8NO4
gln-L	L-Glutamine	C5H10N2O3
gln_L(e)	L-Glutamine	C5H10N2O3
mal-L	L-Malate	C4H4O5
mal_L(e)	L-Malate	C4H4O5
nad	Nicotinamide adenine dinucleotide	C21H26N7O14P2
nadh	Nicotinamide adenine dinucleotide - reduced	C21H27N7O14P2
nadp	Nicotinamide adenine dinucleotide phosphate	C21H25N7O17P3
nadph	Nicotinamide adenine dinucleotide phosphate - reduced	C21H26N7O17P3
o2	O2	O2
o2(e)	O2	O2
oaa	Oxaloacetate	C4H2O5
oaa(e)	Oxaloacetate	C4H2O5
pi	Phosphate	HO4P
pi(e)	Phosphate	HO4P
pep	Phosphoenolpyruvate	C3H2O6P
pep(e)	Phosphoenolpyruvate	C3H2O6P
pyr	Pyruvate	C3H3O3
pyr(e)	Pyruvate	C3H3O3
s7p	Sedoheptulose 7-phosphate	C7H13O10P
s7p(e)	Sedoheptulose 7-phosphate	C7H13O10P
succ	Succinate	C4H4O4
succ(e)	Succinate	C4H4O4
succoa	Succinyl-CoA	C25H35N7O19P3S
succo(e)	Succinyl-CoA	C25H35N7O19P3S
q8h2	Ubiquinol-8	C49H76O4
q8	Ubiquinone-8	C49H74O4
h2o(e)	Water	H2O

Table A.6: Metabolite names from (Khodayari et al., 2014) model

Abbreviation	Name	Reaction
EDA	2-dehydro-3-deoxy-phosphogluconate aldolase	[c] : 2ddg6p → g3p + pyr
EX_akg(e)	2-Oxoglutarate exchange	(1)akg[e] ⟷
AKGDH	2-Oxogluterate dehydrogenase	(1)akg[c] + (1)coa[c] + (1)nad[c] → (1)co2[c] + (1)nadh[c] + (1)succoa[c]
DDPA	3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase	[c] : e4p + h2o + pep → 2dda7p + pi
EDD	6-phosphogluconate dehydratase	[c] : 6pgc → 2ddg6p + h2o
PGL	6-phosphogluconolactonase	(1)6pgl[c] + (1)h2o[c] → (1)6pgc[c] + (1)h[c]
PGL_spon	6-phosphogluconolactonase	(1)6pgl[c] + (1)h2o[c] → (1)6pgc[c] + (1)h[c]
ACALD	acetaldehyde dehydrogenase (acetylating)	(1)acald[c] + (1)coa[c] + (1)nad[c] ⟷ (1)accoa[c] + (1)h[c] + (1)nadh[c]
EX_acald(e)	Acetaldehyde exchange	(1)acald[e] ⟷
EX_ac(e)	Acetate exchange	(1)ac[e] ⟷
ACKr	acetate kinase	(1)ac[c] + (1)atp[c] ⟷ (1)actp[c] + (1)adp[c]
ACS	acetyl-CoA synthetase	[c] : ac + atp + coa → accoa + amp + ppi
ACONTa	aconitase (half-reaction A, Citrate hydro-lyase)	(1)cit[c] ⟷ (1)acon-C[c] + (1)h2o[c]
ACONTb	aconitase (half-reaction B, Isocitrate hydro-lyase)	(1)acon-C[c] + (1)h2o[c] ⟷ (1)icit[c]
ADK1	adenylate kinase	(1)amp[c] + (1)atp[c] ⟷ (2)adp[c]
ALCD2x	alcohol dehydrogenase (ethanol)	(1)etoh[c] + (1)nad[c] ⟷ (1)acald[c] + (1)h[c] + (1)nadh[c]
EX_nh4(e)	Ammonia exchange	(1)nh4[e] ⟷
ATPM	ATP maintenance requirement	(1)atp[c] + (1)h2o[c] → (1)adp[c] + (1)h[c] + (1)pi[c]
ATPS4rpp	ATP synthase (four protons for one ATP) (periplasm)	(1)adp[c] + (4)h[e] + (1)pi[c] ⟷ (1)atp[c] + (1)h2o[c] + (3)h[c]
CS	citrate synthase	(1)accoa[c] + (1)h2o[c] + (1)oaa[c] → (1)cit[c] + (1)coa[c] + (1)h[c]
EX_co2(e)	CO2 exchange	(1)co2[e] ⟷
AKGt2rpp_ex	composed of 2-oxoglutarate reversible transport via symport (periplasm) and alpha-ketoglutarate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)akg[e] + (1)h[e] ⟷ (1)akg[c] + (1)h[c]
ACALDtpp_ex	composed of acetaldehyde transport via diffusion (extracellular to periplasm) and acetaldehyde reversible transport (periplasm)	(1)acald[e] ⟷ (1)acald[c]
ACt2rpp_ex_H	composed of acetate reversible transport via proton symport (periplasm) and Acetate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)ac[e] + (1)h[e] ⟷ (1)ac[c] + (1)h[c]

NH4tpp_ex	composed of ammonia transport via diffusion (extracellular to periplasm) and ammonia reversible transport (periplasm)	(1)nh4[e] ⟷ (1)nh4[c]
CO2tpp_ex	composed of CO2 transport via diffusion (periplasm and extracellular to periplasm)	(1)co2[e] ⟷ (1)co2[c]
CYTBD2pp_H	composed of cytochrome oxidase bd (menaquinol-8: 2 protons) (periplasm) and proton transport via diffusion (extracellular to periplasm)	(2)h[c] + (0.5)o2[c] + (1)q8h2[c] → (1)h2o[c] + (2)h[e] + (1)q8[c]
D-LACt2pp_ex_H	composed of D-lactate transport via diffusion (periplasm and extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)h[e] + (1)lac-D[e] ⟷ (1)h[c] + (1)lac-D[c]
ETOHt2rpp_ex_H	composed of ethanol reversible transport via proton symport (periplasm) and ethanol transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)etoh[e] + (1)h[e] ⟷ (1)etoh[c] + (1)h[c]
FORtppi_ex	composed of formate transport via diffusion (extracellular to periplasm) and formate transport via diffusion (cytoplasm to periplasm)	(1)for[c] → (1)for[e]
FORt2pp_ex_H	composed of formate transport via proton symport (uptake only, periplasm) and formate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)for[e] + (1)h[e] → (1)for[c] + (1)h[c]
FRUpts2pp_ex	composed of Fructose transport via PEP:Pyr PTS (f6p generating) (periplasm) and D-fructose transport via diffusion (extracellular to periplasm)	(1)fru[e] + (1)pep[c] → (1)f6p[c] + (1)pyr[c]
FUMt2_2pp_ex_H	composed of Fumarate transport via diffusion (extracellular to periplasm) and Fumarate transport via proton symport (2 H) (periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)fum[e] + (2)h[e] → (1)fum[c] + (2)h[c]
GLCptspp_ex_exi	composed of glucose transport via diffusion (extracellular to periplasm) and D-glucoseMaltotriose transport via diffusion (extracellular to periplasm) irreversible and D-glucose transport via PEP:Pyr PTS (periplasm)	(1)glc-D[e] + (1)pep[c] → (1)g6p[c] + (1)pyr[c]
H2Otpp_ex	composed of H2O transport via diffusion (extracellular to periplasm) and H2O transport via diffusion (periplasm)	(1)h2o[e] ⟷ (1)h2o[c]
GLUt2rpp_ex_H	composed of L-glutamate transport via proton symport, reversible (periplasm) and L-glutamate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)glu-L[e] + (1)h[e] ⟷ (1)glu-L[c] + (1)h[c]

GLNabcpp_ex	composed of L-glutamine transport via ABC system (periplasm) and L-glutamine transport via diffusion (extracellular to periplasm)	(1)atp[c] + (1)gln-L[e] + (1)h2o[c] → (1)adp[c] + (1)gln-L[c] + (1)h[c] + (1)pi[c]
MALt2_2pp_ex_H	composed of Malate transport via proton symport (2 H) (periplasm) and Malate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(2)h[e] + (1)mal-L[e] → (2)h[c] + (1)mal-L[c]
THD2pp_H	composed of NAD(P) transhydrogenase (periplasm) and proton transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(2)h[e] + (1)nadh[c] + (1)nadp[c] → (2)h[c] + (1)nad[c] + (1)nadph[c]
NADH16pp_H	composed of NADH dehydrogenase (ubiquinone-8 & 3 protons) (periplasm) and proton transport via diffusion (extracellular to periplasm)	(4)h[c] + (1)nadh[c] + (1)q8[c] → (3)h[e] + (1)nad[c] + (1)q8h2[c]
O2tpp_ex	composed of oxygen transport via diffusion (extracellular to periplasm) and o2 transport via diffusion (periplasm)	(1)o2[e] ⟷ (1)o2[c]
PIt2rpp_ex_H	composed of phosphate reversible transport via symport (periplasm) and phosphate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)h[e] + (1)pi[e] ⟷ (1)h[c] + (1)pi[c]
PYRt2rpp_ex_H	composed of pyruvate reversible transport via proton symport (periplasm) and pyruvate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)h[e] + (1)pyr[e] ⟷ (1)h[c] + (1)pyr[c]
SUCCt3pp_ex_H	composed of succinate transport out via proton antiport (periplasm) and succinate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(1)h[e] + (1)succ[c] → (1)h[c] + (1)succ[e]
SUCCt2_2pp_ex_H	composed of succinate transport via proton symport (2 H) (periplasm) and succinate transport via diffusion (extracellular to periplasm) and proton transport via diffusion (extracellular to periplasm)	(2)h[e] + (1)succ[e] → (2)h[c] + (1)succ[c]
EX_fru(e)	D-Fructose exchange	(1)fru[e] ⟷
EX_glc(e)	D-Glucose exchange	(1)glc-D[e] ⟷
LDH_D	D-lactate dehydrogenase	(1)lac-D[c] + (1)nad[c] ⟷ (1)h[c] + (1)nadh[c] + (1)pyr[c]
EX_lac_D(e)	D-lactate exchange	(1)lac-D[e] ⟷
ENO	enolase	(1)2pg[c] ⟷ (1)h2o[c] + (1)pep[c]
EX_etoh(e)	Ethanol exchange	(1)etoh[e] ⟷
EX_for(e)	Formate exchange	(1)for[e] ⟷
FBP	fructose-bisphosphatase	(1)fdp[c] + (1)h2o[c] → (1)f6p[c] + (1)pi[c]

FBP_Glpx	fructose-bisphosphatase	(1)fdp[c] + (1)h2o[c] → (1)f6p[c] + (1)pi[c]
FBA	fructose-bisphosphate aldolase	(1)fdp[c] ⟷ (1)dhap[c] + (1)g3p[c]
FUM	fumarase	(1)fum[c] + (1)h2o[c] ⟷ (1)mal-L[c]
EX_fum(e)	Fumarate exchange	(1)fum[e] ⟷
FRD2	fumarate reductase	(1)fum[c] + (1)q8h2[c] → (1)q8[c] + (1)succ[c]
G6PDH2r	glucose 6-phosphate dehydrogenase	(1)g6p[c] + (1)nadp[c] ⟷ (1)6pgl[c] + (1)h[c] + (1)nadph[c]
GLGC	glucose-1-phosphate adenylyltransferase	[c] : atp + g1p + h → adpglc + ppi
PGI	glucose-6-phosphate isomerase	(1)g6p[c] ⟷ (1)f6p[c]
GLUDy	glutamate dehydrogenase (NADP)	(1)glu-L[c] + (1)h2o[c] + (1)nadp[c] ⟷ (1)akg[c] + (1)h[c] + (1)nadph[c] + (1)nh4[c]
GLUSy	glutamate synthase (NADPH)	(1)akg[c] + (1)gln-L[c] + (1)h[c] + (1)nadph[c] → (2)glu-L[c] + (1)nadp[c]
GLUN	glutaminase	(1)gln-L[c] + (1)h2o[c] → (1)glu-L[c] + (1)nh4[c]
GLNS	glutamine synthetase	(1)atp[c] + (1)glu-L[c] + (1)nh4[c] → (1)adp[c] + (1)gln-L[c] + (1)h[c] + (1)pi[c]
GAPD	glyceraldehyde-3-phosphate dehydrogenase	(1)g3p[c] + (1)nad[c] + (1)pi[c] ⟷ (1)13dpg[c] + (1)h[c] + (1)nadh[c]
GLCS1	glycogen synthase (ADPGlc)	(1)adpglc_c ↔ (1)adp_c + (1)glycogen + (1)h_c
EX_h(e)	H2 exchange	(1)h[e] ⟷
EX_h2o(e)	H2O exchange	(1)h2o[e] ⟷
HEX1	hexokinase (D-glucose:ATP)	[c] : atp + glc-D → adp + g6p + h
PPA	inorganic diphosphatase	(1)h2o_c + (1)ppi_c ↔ (2)h_c + (2)pi_c
ICDHyr	isocitrate dehydrogenase (NADP)	(1)icit[c] + (1)nadp[c] ⟷ (1)akg[c] + (1)co2[c] + (1)nadph[c]
ICL	Isocitrate lyase	(1)icit[c] → (1)glx[c] + (1)succ[c]
EX_glu_L(e)	L-Glutamate exchange	(1)glu-L[e] ⟷
EX_gln_L(e)	L-Glutamine exchange	(1)gln-L[e] ⟷
EX_mal_L(e)	L-Malate exchange	(1)mal-L[e] ⟷
MDH2	Malate dehydrogenase (ubiquinone 8 as acceptor)	[c] : mal-L + q8 → oaa + q8h2
MALS	malate synthase	(1)accoa[c] + (1)glx[c] + (1)h2o[c] → (1)coa[c] + (1)h[c] + (1)mal-L[c]
MDH	malate synthase	(1)mal-L[c] + (1)nad[c] ⟷ (1)h[c] + (1)nadh[c] + (1)oaa[c]
ME1	malic enzyme (NAD)	(1)mal-L[c] + (1)nad[c] → (1)co2[c] + (1)nadh[c] + (1)pyr[c]
ME2	malic enzyme (NADP)	(1)mal-L[c] + (1)nadp[c] → (1)co2[c] + (1)nadph[c] + (1)pyr[c]
NADTRHD	NAD transhydrogenase	(1)nad[c] + (1)nadph[c] → (1)nadh[c] + (1)nadp[c]

EX_o2(e)	O2 exchange	(1)o2[e] ⟷
EX_pi(e)	Phosphate exchange	(1)pi[e] ⟷
PPCK	phosphoenolpyruvate carboxykinase	(1)atp[c] + (1)oaa[c] → (1)adp[c] + (1)co2[c] + (1)pep[c]
PPC	phosphoenolpyruvate carboxylase	(1)co2[c] + (1)h2o[c] + (1)pep[c] → (1)h[c] + (1)oaa[c] + (1)pi[c]
PPS	phosphoenolpyruvate synthase	(1)atp[c] + (1)h2o[c] + (1)pyr[c] → (1)amp[c] + (2)h[c] + (1)pep[c] + (1)pi[c]
PFKA	phosphofructokinase	(1)atp[c] + (1)f6p[c] → (1)adp[c] + (1)fdp[c] + (1)h[c]
PFKB	phosphofructokinase	[c] : atp + f6p → adp + fdp + h
PGMT	phosphoglucomutase	[c] : g1p ↔ g6p
GND	phosphogluconate dehydrogenase	(1)6pgc[c] + (1)nadp[c] → (1)co2[c] + (1)nadph[c] + (1)ru5p-D[c]
PGK	phosphoglycerate kinase	(1)3pg[c] + (1)atp[c] ⟷ (1)13dpg[c] + (1)adp[c]
PGM	phosphoglycerate mutase	(1)2pg[c] ⟷ (1)3pg[c]
PTAr	phosphotransacetylase	(1)accoa[c] + (1)pi[c] ⟷ (1)actp[c] + (1)coa[c]
PDH	pyruvate dehydrogenase	(1)coa[c] + (1)nad[c] + (1)pyr[c] → (1)accoa[c] + (1)co2[c] + (1)nadh[c]
EX_pyr(e)	Pyruvate exchange	(1)pyr[e] ⟷
PFL	pyruvate formate lyase	(1)coa[c] + (1)pyr[c] → (1)accoa[c] + (1)for[c]
PYKF	pyruvate kinase	(1)adp[c] + (1)h[c] + (1)pep[c] → (1)atp[c] + (1)pyr[c]
PYKA	pyruvate kinase	[c] : adp + h + pep → atp + pyr
R5PP	ribose 5-phosphate phosphatase	[c] : h2o + r5p → pi + rib-D
RPI	ribose-5-phosphate isomerase	(1)r5p[c] ⟷ (1)ru5p-D[c]
RPE	ribulose 5-phosphate 3-epimerase	(1)ru5p-D[c] ⟷ (1)xu5p-D[c]
SUCDi	succinate dehydrogenase (irreversible)	(1)q8[c] + (1)succ[c] → (1)fum[c] + (1)q8h2[c]
EX_succ(e)	Succinate exchange	(1)succ[e] ⟷
SUCOAS	succinyl-CoA synthetase (ADP-forming)	(1)atp[c] + (1)coa[c] + (1)succ[c] ⟷ (1)adp[c] + (1)pi[c] + (1)succoa[c]
TALA	transaldolase	(1)g3p[c] + (1)s7p[c] ⟷ (1)e4p[c] + (1)f6p[c]
TKT1	transketolase	(1)r5p[c] + (1)xu5p-D[c] ⟷ (1)g3p[c] + (1)s7p[c]
TKT2	transketolase	(1)e4p[c] + (1)xu5p-D[c] ⟷ (1)f6p[c] + (1)g3p[c]
TPI	triose-phosphate isomerase	(1)dhap[c] ⟷ (1)g3p[c]
EX_E4P(e)		e4p [c] →
EX_S7P(e)		s7p [c] →
EX_RU5P(e)		ru5p [c] →
EX_OAA(e)		oaa [c] →

EX_SUCCOA(e)		succoa [c] →
EX_ACCOA(e)		accoa [c] →
EX_DHAP(e)		dhap [c] →
EX_3PG(e)		3pg [c] →
EX_F6P(e)		f6p [c] →
EX_G3P(e)		→ gap [c]
EX_PEP(e)		pep [c] →
EX_2DDA7P(e)		(1)2dda7p_e ↔
EX_ADPGLC(e)		(1)adpglc_e ↔
ADPGLCtpp_ex		(1)adpglc_e ↔ (1)adpglc_c
2DDA7Ptpp_ex		(1)2dda7p_e ↔ (1)2dda7p_c
E4Ptpp_ex		(1)e4p_e ↔ (1)e4p_c
S7Ptpp_ex		(1)s7p_e ↔ (1)s7p_c
RU5Ptpp_ex		(1)ru5p_e ↔ (1)ru5p_c
OAAtpp_ex		(1)oaa_e ↔ (1)oaa_c
SUCCOAtpp_ex		(1)succoa_e ↔ (1)succoa_c
ACCOAtpp_ex		(1)accoa_e ↔ (1)accoa_c
DHAPtpp_ex		(1)dhap_e ↔ (1)dhap_c
3PGtpp_ex		(1)3pg_e ↔ (1)3pg_c
F6Ptpp_ex		(1)f6p_e ↔ (1)f6p_c
G3Ptpp_ex		(1)g3p_e ↔ (1)g3p_c
PYRpp_ex		(1)pyr_e ↔ (1)pyr_c
PEPtpp_ex		(1)pep_e ↔ (1)pep_c
RIB_Dtpp_ex		(1)rib_d_e ↔ (1)rib_d_c
EX_RIB_D(e)		(1)rib_d_e →

Table A.7: Reactions from the model presented by (Khodayari et al., 2014).

Appendix B

Communities and synthetic biology

B.1 Complexity proofs

In Section 4.2.5 in Chapter 4, we presented two complexity results whose proofs are given below.

Proposition 3. *The problem is W[1]-hard when parameterised by any combination of:* $|A'|$, $\mathsf{weight}(A')$, $|T|$, $|S|$, *total number of tentacles of hyperarcs in* A'.

Proof. The reduction is from the Multi-Colour Clique (MCC) problem, defined as follows:

Input: A k-partite graph $G = (V, E)$ where $V = V_1 \cup V_2 \cup \ldots \cup V_k$.
Output: A k-clique of G, *i.e.* a set $V' = \{v_1, \ldots, v_k\}$, with $v_i \in V_i$ for all i, such that $\{v_i, v_j\} \in E$ for every pair $\{i, j\}$, $i \neq j$.

The MCC problem is W[1]-hard when parameterised by k.

Each set V_i is called a *(vertex-)class*. The set of edges can be partitioned into $\binom{k}{2}$ *edge-classes* $E_{i,j}$, according to the class of their endpoints (formally, $E_{i,j} = E \cap (V_i \times V_j)$).

Given an instance of MCC $G = (V, E)$, we build an instance of the Directed Steiner Hypertree problem as follows (see Supplementary Figure B.1).

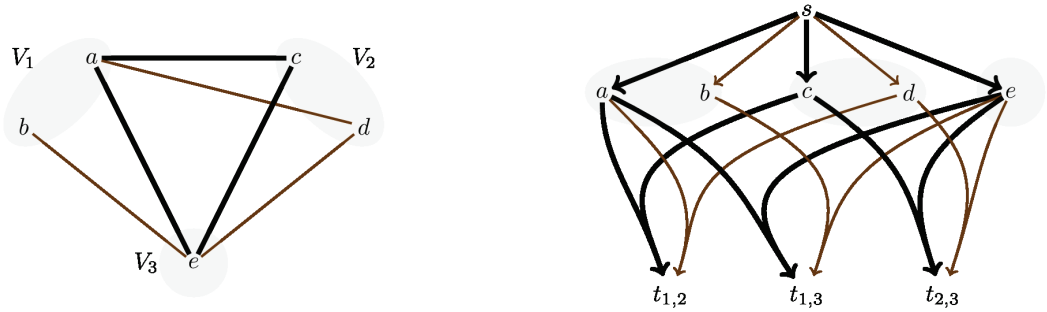


Figure B.1: Illustration of the parameterised reduction from Multi-Colour Clique (left) to Directed Steiner Hypertree (right) with 5 vertices and $k = 3$. A clique of the input graph and the corresponding optimal directed rooted hypergraph are depicted with bold edges.

Define a set of vertices W:

$$\begin{aligned} W = &\{s\} \\ &\cup \{x_v \mid v \in V\} \\ &\cup \{t_{i,j} \mid 1 \leq i < j \leq k\} \end{aligned}$$

Intuitively, this vertex set contains one source, one vertex for each original vertex of G, and one target for each edge-class. We write $S = \{s\}$ and $T = \{t_{i,j} \mid 1 \leq i < j \leq k\}$.

Define now a set of hyperarcs A, all having weight 1:

$$\begin{aligned} A = &\{a_v = (\{s\}, \{x_v\}) \mid v \in V\} \\ &\cup \{a_e = (\{x_u, x_v\}, \{t_{i,j}\}) \mid e \in E, e = \{u, v\} \in E_{i,j}\} \end{aligned}$$

This set contains two parts: first a set of simple arcs a_v allowing to reach any vertex x_v, then for each edge a hyperarc starting from both endpoints of the edge and reaching the target of the corresponding colour.

We now prove that $(\mathcal{H} = (W, A), S, T)$ admits a solution of weight at most $k + \binom{k}{2}$ if, and only if, G admits a k-clique. Note that this equivalence completes the parameterised reduction, since $|S|$, $|T|$, the solution size and its total number of tentacles are all bounded by a function of k.

If. Consider a k-clique K of G. Write E' for the set of $\binom{k}{2}$ edges used in K. It is easy to verify that $A' = \{a_v \mid v \in K\} \cup \{a_e \mid e \in E'\}$ is a valid solution: starting from the source s, each x_v, $v \in K$ can be reached using the corresponding arc $a_v \in A'$. Then for each $1 \leq i < j \leq k$, pick the edge $e = \{u, v\}$ having class $\{i, j\}$ in E': both vertices x_u and x_v have already been reached, hence the hyperarc a_e allows us to reach $t_{i,j}$. Overall, all targets have been reached with exactly $k + \binom{k}{2}$ hyperarcs.

Only if. Consider a solution A' of DSH for $(\mathcal{H}, S, T)$. Write $K = \{u \in V \mid a_u \in A'\}$ and $E' = \{e \in E \mid a_e \in A'\}$, that is, the set of vertices and edges of the original graph for which the corresponding hyperarc of $\mathcal{H}$ is used in A'. We will show that (K, E') forms a clique. The first observation is that $|K| + |E'| \leq k + \binom{k}{2}$, since this is the maximum total weight of the solution.

For every $1 \leq i < j \leq k$, since $t_{i,j} \in T$, A' contains a hyperarc ending in $t_{i,j}$, so E' must contain some edge e having class $\{i, j\}$. This already shows that $|E'| \geq \binom{k}{2}$, which in turn yields $|K| \leq k$.

Write $\{u, v\}$ for the endpoints of any $e \in E'$. Since x_u and x_v have in-degree 1 in $\mathcal{H}$, the arcs a_u and a_v must also belong to A', and $u, v \in K$. Hence all the endpoints of edges in E' are in K.

To sum up: E' is a set of at least $\binom{k}{2}$ edges with a total of only k endpoints: they are the edges of a clique of G.

□

Proposition 4. *The problem is NP-hard even when $|T| = 1$ and A contains only one tentacular hyperarc.*

Proof. The reduction is from the Directed Steiner Tree problem. Consider an instance of this problem, *i.e.* a directed graph $G = (V, A)$, a source $s \in V$, and a set of targets $T \subseteq V$. Create a directed hypergraph $\mathcal{H}$ from G by adding a vertex t and a hyperarc h from T to $\{t\}$. Then $(\mathcal{H}, \{s\}, \{t\})$ is an instance of the Directed Steiner Hypertree problem with a single target and a single tentacular hyperarc (see Supplementary Figure B.2). Also, any solution necessarily takes the hyperarc h, and thus must cover its whole head, *i.e.* T. Hence the solutions of the Directed Steiner Hypertree problem exactly correspond to the solutions of the Directed Steiner Tree problem by adding h.

□

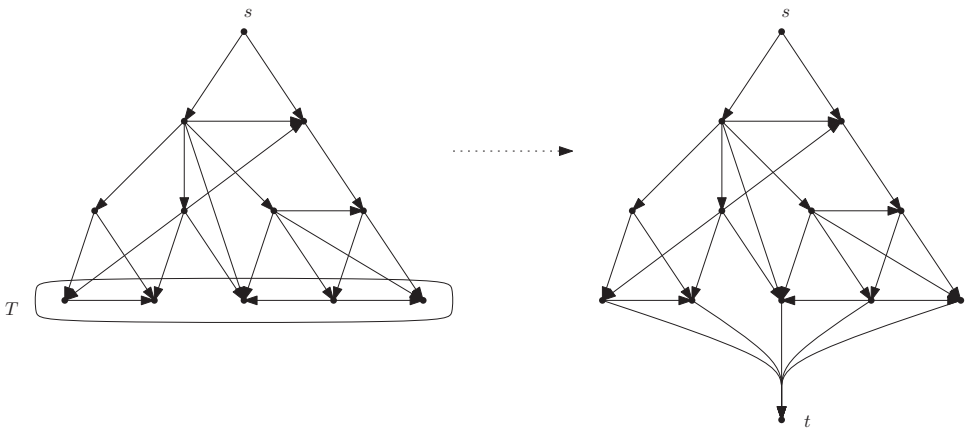


Figure B.2: Illustration of the reduction from Directed Steiner Tree (left) to Directed Steiner Hypertree (right), with a single tentacular hyperarc and a single target.

B.2 Directed hypergraph construction

B.2.1 Metabolites removal

When creating the hypergraphs from the reconstructed metabolic networks, common cofactors and co-enzymes were removed. They were identified using the BRITE functional hierarchies of Kegg. The list of all filtered metabolites is given below:

- Adenosine 5'-triphosphate
- Nicotinamide adenine dinucleotide
- Nicotinamide adenine dinucleotide phosphate
- Coenzyme A
- Flavin adenine dinucleotide
- Pyridoxal phosphate
- S-Adenosyl-L-methionine
- UPD-glucose
- Heme
- Glutathione
- 3'-Phosphoadenylyl sulfate
- Riboflavin-5-phosphate
- Cytidine 5'-triphosphate
- Thiamin diphosphate
- Tetrahydrofolate
- Pyrroloquinoline quinone
- Coenzyme R
- Cobamide coenzyme
- 5-Dehydro-D-fructose
- Lipoate
- Methanofuran

- 5,6,7,8-Tetrahydromethanopterin
- 2-Mercaptoethanesulfonate (Coenyzme M)
- N-(7-Mercaptoheptanoyl)threonine O3-phosphate (Coenzyme B)
- Coenzyme F430
- Ubiquinone-10
- Heme A
- Heme O
- Molybdenum cofactor

This list has been made for general applications. Hence some of the metabolites may not be present in the networks used in the Application of the method in this manuscript.

Modéliser le métabolisme: expliciter les réponses aux perturbations et composer des consortia microbiens.

Résumé en français

Lors de cette thèse, je me suis intéressée à la modélisation du métabolisme des micro-organismes. Nous nous sommes focalisé sur le métabolisme des petites molécules qui ne prend pas en compte les réactions associées aux macromolécules, telle que la synthèse des protéines.

Nous avons ainsi utilisé différents formalismes de modélisation.

Tout d'abord, nous avons développé TOTORO où les réseaux métaboliques sont représentés par des hypergraphes dirigés et qui permet d'identifier les réactions ayant participé à une transition métabolique. TOTORO a été utilisé sur un jeu de données sur la levure en présence de cadmium. Nous avons pu montrer que nous retrouvons les mécanismes connus de désintoxication.

Ensuite, en utilisant une méthode de modélisation par contraintes, nous discutons d'un développement en cours, KOTOURA, qui propose d'utiliser les connaissances actuelles de concentrations de métabolites entre différentes conditions pour inférer de manière quantitative les possibles asynchronies des réactions lors du passage d'un état stable à un autre. Nous avons testé son implémentation sur des données simulées.

Enfin, nous proposons MULTIPUS, une méthode d'extraction d'(hyper)-arbres de Steiner dirigés qui permet de sélectionner les voies métaboliques pour la production de composés au sein d'une communauté bactérienne. Les réseaux métaboliques sont modélisés en utilisant des hypergraphes dirigés et pondérés. Nous proposons un algorithme de programmation dynamique paramétré ainsi qu'une formulation utilisant la programmation par ensemble réponse. Ces deux propositions sont ensuite comparées dans deux cas d'applications.

MOTS-CLEFS en français

Métabolisme des petites molécules; modélisation des réseaux métaboliques; communautés microbiennes; consortium synthétique; hypergraphes dirigés; programmation sous contraintes; hyper-histoires métaboliques; transition métabolique.

Models and algorithms applied to metabolism: From revealing the responses to perturbations towards the design of microbial consortia

Abstract in english

In this PhD work, we proposed to model metabolism. Our focus was to develop generic models, that are not specific to one organism or condition, but are instead based on general assumptions that we tried to validate using data from the literature. We first present TOTORO that uses a qualitative measurement of concentrations in two steady-states to infer the reaction changes that lead to differences in metabolite pools in both conditions.

TOTORO enumerates all sub-(hyper)graphs that represent a sufficient explanation for the observed differences in concentrations. We exploit a dataset of Yeast (*Saccharomyces cerevisiae*) exposed to cadmium and show that we manage to retrieve the known pathways used by the organisms.

We then address the same issue, but using a constraint-based programming framework, called KOTOURA, that allows to infer more quantitatively the reaction changes during the perturbed state. We use in this case exact concentration measurements and the stoichiometric matrix, and show on simulated datasets that the overall variations of reaction fluxes can be captured by our formulation.

Finally, we propose MULTIPUS, a method to infer microbial communities and metabolic roads to produce specific target compounds from a set of defined substrates. We use in this case a weighted directed hypergraph. We apply MULTIPUS to the production of antibiotics using a consortium composed of an archae and an actinobacteria and show hat their metabolic capacities are complementary. We then infer for another community the excretion of an inhibitory product (acetate) by a 1,3-propanediol (PDO) producer and its consumption by a methanogene archae.

Keywords in english

Small molecule metabolism; metabolic network modelling; bacterial communities; synthetic consortium; directed hypergraphs; constraint-based programming; metabolic hyperstories; metabolic shifts.