Dynamique d'un réseau métabolique avec un modèle à base de contraintes : approche par échantillonnage des trajectoires solutions

Thomas Duigou

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

I2BC

Institut de Biologie Intégrative de la Cellule

Université Paris-Sud

École Doctorale :	Agriculture, Alimentation, Biologie, Environnement, Santé
Laboratoire :	Institut de Biologie Intégrative de la Cellule
Discipline :	Sciences agronomiques, biotechnologies agro-alimentaires
Scpécialité :	Bio-informatique

Thèse de doctorat

Soutenue le 13 mai 2015 par

Thomas Duigou

Dynamique d'un réseau métabolique avec un modèle à base de contraintes : approche par échantillonnage des trajectoires solutions

Directeur de thèse :	Armel Guyonvarch	PU (Univ. Paris-Sud)
Co-encadrant :	Bruno Bost	MCU (Univ. Paris-Sud)

Composition du jury :

Rapporteurs :	Gilles Bernot	PU (Univ. Nice)
	Hidde de Jong	DR (INRIA Grenoble)
Examinateurs :	Christine Dillmann	PU (Univ. Paris-Sud)
	Guillaume Letellier	Docteur entreprise

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

I2BC

Institut de Biologie Intégrative de la Cellule

Université Paris-Sud

École Doctorale :	Agriculture, Alimentation, Biologie, Environnement, Santé
Laboratoire :	Institut de Biologie Intégrative de la Cellule
Discipline :	Sciences agronomiques, biotechnologies agro-alimentaires
Scpécialité :	Bio-informatique

Thèse de doctorat

Soutenue le 13 mai 2015 par

Thomas Duigou

Dynamique d'un réseau métabolique avec un modèle à base de contraintes : approche par échantillonnage des trajectoires solutions

Directeur de thèse :	Armel Guyonvarch	PU (Univ. Paris-Sud)
Co-encadrant :	Bruno Bost	MCU (Univ. Paris-Sud)

Composition du jury :

Rapporteurs :	Gilles Bernot	PU (Univ. Nice)
	Hidde de Jong	DR (INRIA Grenoble)
Examinateurs :	Christine Dillmann	PU (Univ. Paris-Sud)
	Guillaume Letellier	Docteur entreprise

Remerciements

Une thèse, c'est des kilogrammes de café, des kilomètres de documents parcourus à la force d'une vaillante molette de souris, ou encore des milliers de pressions simultanées sur les touches [ctrl] et [s].

Mais, une thèse c'est aussi, et surtout, des personnes.

Je remercie Gilles Bernot et Hidde de Jong d'avoir accepté d'être rapporteurs de mon travail de thèse. Je les remercie également avec Christine Dillmann et Guillaume Letellier d'avoir bien voulu faire partie de mon jury de thèse.

Mes remerciements vont bien sûr à Armel Guyonvarch et Bruno Bost qui m'ont encadré tout au long de ces trois ans et demi, et m'ont accordé leur confiance pour mener à bien mon projet de thèse.

Je tiens également à remercier les membres de mon ancienne équipe, l'équipe physiologie et métabolisme des corynébactéries ; ainsi que les membres de ma nouvelle équipe, l'équipe bio-informatique moléculaire, qui m'ont accueilli pendant les derniers mois de ma thèse.

Enfin, je remercie particulièrement ma famille, ma belle-famille, mes amis, et ma compagne, Mylène, pour son soutien sans faille.

Table des matières

Abréviations

ADN	acide désoxyribonucléique
ARN	acide ribonucléique
ATCC	*American type culture collection*
ATPm	ATP de maintenance
C. glutamicum	*Corynebacterium glutamicum*
DO	densité optique
EC number	*enzyme commission number*
E. coli	*Escherichia coli*
flux E/S	flux d'entrées/sorties
RM	réseau métabolique
RRG	réseau de régulation génétique
TCA	*tricarboxylic acid cycle*, cycle des acides tricarboxyliques

Abbréviations de méthode

`cda`	*coordinates directions algorithm*
DFBA	*dynamic flux balance analysis*
DMFA	*dynamic metabolic flux analysis*
DOA	*dynamic optimization approcah* (DFBA)
EDLP	équations différentielles linéaires par partie
EDO	équations différentielles ordinaires
FBA	*flux balance analysis*
FSA	*flux spectrum approach*
FVA	*flux variability analysis*
MBC	modèle à base de containtes
MCMC	*Monte Carlo markov chain*
MFA	*metabolic flux analysis*
ML	*modèle logique*
`rda`	*random directions algorithm*
RdP	réseau de Petri
SOA	*static optimization approcah* (DFBA)

Préambule

L'émergence de la biologie des systèmes

Aux origines de l'étude des êtres vivants, « l'histoire naturelle » portait principalement sur le recensement et la description des caractères observables : organisation générale des organismes, taille, durée de vie... Au fur et à mesure des développements techniques, de nouveaux dispositifs ont permis de « zoomer » progressivement dans les êtres vivants, et d'y découvrir des entités biologiques de plus en plus petites. La biologie cellulaire, l'étude des fonctions et de la structure des cellules, s'est ainsi développée avec les premiers microscopes. Vint ensuite l'avènement de la génétique [1], de la biochimie [2] et de la biologie moléculaire [3], domaines de la biologie dédiés à l'étude des entités qui constituent les cellules.

Après plusieurs décennies d'expérimentation, nous disposons aujourd'hui d'une grande masse d'informations biologiques. Récemment, l'augmentation de la quantité d'information a connu un nouveau bond avec la généralisation de la biologie « -omique » (génomique, transcriptomique, protéomique ou encore métabolomique), qui consiste en l'étude simultanée d'un grand nombre d'entités biologiques (par exemple l'ensemble des gènes, des protéines ou des métabolites d'une cellule).

Les données disponibles à l'heure actuelle sont de nature variée et fournissent des informations détaillées sur de nombreuses entités qui existent au sein des êtres vivants. Cependant, connaître chaque entité n'est pas suffisant pour comprendre le fonctionnement des organismes vivants.

En effet, la ou les fonctions d'une entité biologique résulte(nt) de ses relations avec d'autres entités au sein de l'organisme. L'étude de ces relations n'est pas chose facile, car la quantité et l'hétérogénéité des informations

1. étude de l'organisation des génomes et des gènes
2. étude des processus biologiques
3. approche visant l'étude des mécanismes moléculaires au sein des cellules

à considérer rendent difficile leur interprétation sans l'utilisation d'outils adaptés.

Cela a conduit à l'émergence de la biologie des systèmes, aussi appelée *biologie systémique* ou encore *biologie intégrative*. La biologie des systèmes s'intéresse aux interactions rencontrées dans les systèmes biologiques et à la manière dont ces interactions donnent naissance à des fonctions biologiques. Pour faire face à la complexité des systèmes vivants, l'étude des interactions s'appuie sur la construction de modèles mathématiques, statistiques et informatiques. Ces modèles, qui décrivent de façon plus ou moins précise la réalité biologique, sont ensuite utilisés pour analyser, tester et prédire le fonctionnement des processus biologiques. À terme, l'idéal à atteindre pour la biologie des systèmes est la compréhension de l'ensemble des phénomènes biologiques.

Ce but ultime amène des enjeux importants, notamment dans les domaines de la santé et des biotechnologies. La biologie des systèmes est ainsi déjà utilisée pour identifier de nouvelles cibles thérapeutiques et pour produire des composés d'intérêts industriels en utilisant des organismes modifiés.

Les challenges de la biologie des systèmes

Un premier palier à atteindre dans le but de comprendre l'ensemble des phénomènes biologiques est la modélisation de chaque processus considéré séparément des autres. Par exemple, à l'échelle d'un réseau métabolique, il s'agit de recenser les entités du réseau (métabolites, enzymes), les relations entre ces entités (réactions biochimiques), ainsi que les conditions nécessaires pour que ces relations existent (concentrations en métabolites et en enzymes). Bien que simple, comparé à l'objectif final, ce palier représente le premier challenge de la biologie des systèmes : la traduction des données biologiques disponibles en informations utilisables dans un modèle.

Un deuxième palier est la prise en compte de la dynamique des relations entre entités au cours du temps. Dans le contexte d'un réseau métabolique, cela se traduit généralement par la modélisation du changement des vitesses de réaction au fil du temps. Cette prise en compte de la dynamique constitue un deuxième challenge pour la biologie des systèmes.

Le palier suivant est d'étudier ensemble plusieurs processus biologiques. Cela consiste par exemple à modéliser un réseau métabolique et un réseau

de régulation de l'expression des gènes — aussi appelé réseau de régulation génétique — et à tenir compte des influences réciproques que chaque réseau peut avoir sur l'autre. Cette intégration de plusieurs processus biologiques représente un troisième challenge.

Karr et *coll.* ont récemment présenté un travail de modélisation « cellule complète » [Karr 2012]. En intégrant des connaissances issues de plus 900 expériences, publications et bases de données, l'ensemble des processus biologiques connus chez *Mycoplama genitalium* a été décrit dans des modèles. En couplant ces modèles, le comportement de la cellule a pu être prédit au cours du temps.

Ces travaux illustrent à la fois les progrès importants réalisés dans le domaine de la biologie des systèmes, mais aussi la quantité colossale de connaissances nécessaires pour parvenir à un tel résultat.

Dans de nombreuses situations, la quantité de connaissances et les moyens disponibles pour les amasser sont moins favorables. Outre le besoin de données biologiques, des apports méthodologiques sont encore nécessaires à tous les « paliers » de la biologie des systèmes pour proposer des méthodes de modélisation capables de décrire le comportement des systèmes à partir d'une quantité raisonnable de données.

Les contributions de mon travail de doctorat

Partie 1. Mon projet était initialement dédié au développement d'approches pour modéliser ensemble les réseaux métaboliques et les réseaux de régulation génétique. Dans une première partie « Des réseaux biologiques aux modèles » (p. 8), je présente les entités et les phénomènes qui sont communément rencontrés lorsque l'on s'intéresse à ce type de réseaux biologiques. J'y présenterai aussi les notions et concepts de base inhérents à l'usage de la modélisation en biologie.

Partie 2. Pour comprendre le fonctionnement des cellules, en particulier *comment* les cellules adaptent leur comportement selon l'environnement, il est important de considérer *ensemble* le métabolisme et la régulation génétique. En effet, la régulation génétique a un impact sur le métabolisme en modifiant la concentration des enzymes. En retour, certains métabolites interviennent dans la régulation génétique en (dés-)activant les régulateurs

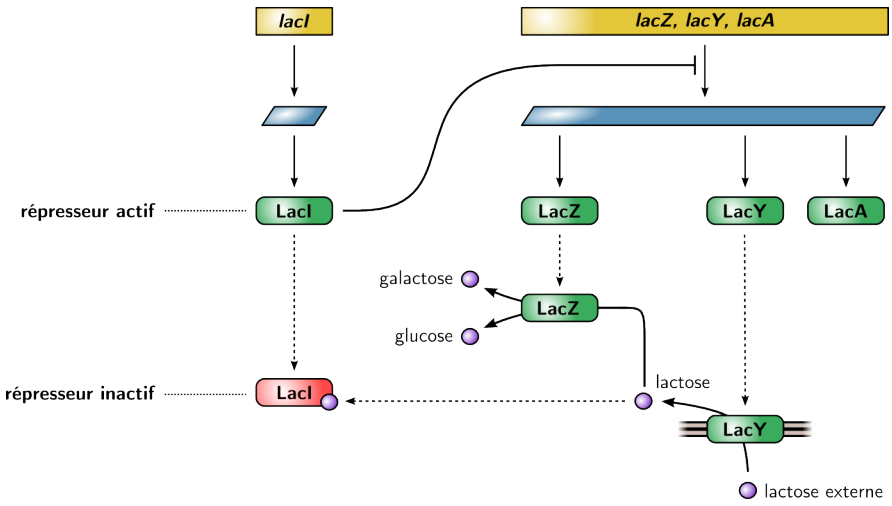


Figure 0.1. – Couplage entre régulation génétique et métabolisme : exemple de l'opéron lactose chez *Escherichia coli*. Le gène *lacZ* produit l'enzyme β-galactosidase qui transforme le lactose en glucose et en galactose. Le régulateur LacI réprime l'expression des gènes *lacZ*, *lacY* et *lacA*. La présence de lactose entraine la formation d'un complexe *lactose–LacI*, ce qui désactive la capacité de répression de LacI.

de l'expression des gènes. Un exemple bien connu de couplage entre régulation génétique et métabolisme est celui de l'opéron lactose chez *Escherichia coli* : le lactose forme un complexe *lactose–LacI*, levant la répression de l'opéron, et induisant notamment la synthèse de l'enzyme LacZ et de la perméase LacY (figure 0.1). Notons que le couplage entre l'opéron lactose et le métabolisme fait aussi intervenir d'autres acteurs, comme l'AMPc et le régulateur Crp, qui participent à la régulation des niveaux d'expression des gènes de l'opéron.

Afin de prendre en compte ce couplage, il est important de considérer la dynamique des interactions entre les deux réseaux. Ainsi, dans l'exemple de l'opéron lactose, LacI est continuellement exprimé et il réprime en l'absence de lactose l'expression de l'opéron *lacZ, lacY, lacA*. Dans cette situation, les trois gènes de l'opéron ne sont que très faiblement transcrits. Lorsque du lactose apparait dans l'environnement extracellulaire, il traverse — *dans un premier temps* — faiblement la membrane *via* les quelques perméases LacY présentes, et forme un complexe avec LacI, levant ainsi la répression de l'opéron. Cela entraine la synthèse plus importante de la perméase LacY, et des enzymes LacZ et LacA, ce qui permet — *dans un second temps* —

l'entrée accélérée et l'utilisation du lactose. *Dans un troisième temps*, une fois le lactose épuisé, le répresseur LacI retrouve sa capacité de régulation et réprime de nouveau l'expression de l'opéron.

Un des premiers objectifs de mon travail de thèse a été d'identifier les méthodes de modélisation qui permettent de modéliser ensemble les réseaux métaboliques et de régulation génétique, et leurs interactions. Axé sur les procaryotes, ce travail de revue est présenté dans la seconde partie « Modélisation simultanée des réseaux métaboliques et de régulation génétique, revue » (p. 50).

Partie 3. Au cours de mon travail de revue sur les méthodes, la modélisation à base de contraintes est apparue comme très intéressante. Les modèles à base de contraintes, ou MBC, permettent de prédire la répartition des flux au sein d'un réseau métabolique. Comparés à d'autres méthodes, les MBC nécessitent peu d'informations pour décrire le comportement d'un réseau métabolique.

Cependant, la dimension dynamique ne peut être directement prise en compte dans les méthodes MBC : ces méthodes sont basées sur l'hypothèse de l'état stationnaire du système, autrement dit sur l'hypothèse que l'état du réseau métabolique ne change pas pendant la période de temps modélisée. Afin d'élargir le champ des applications des MBC, des approches ont déjà été proposées pour intégrer la dimension dynamique.

Lorsque la quantité de données est faible, la répartition des flux dans le réseau ne peut pas être estimée de manière unique : il existe différentes répartitions de flux qui sont toutes en accord avec les données à disposition. Une manière de choisir parmi toutes les répartitions possibles est alors de poser une hypothèse sur un phénomène biologique qui serait optimisé au sein du métabolisme étudié. Typiquement, l'hypothèse sera que le système optimise la multiplication cellulaire ou la production d'énergie. Cependant, le « but » vers lequel tend le métabolisme n'est connu que dans certaines situations particulières, ce qui limite les possibilités d'utilisation de cette méthode.

Compte tenu des améliorations qu'il était possible d'apporter aux méthodes MBC, mon sujet de thèse a évolué vers la recherche d'une approche qui utilise les méthodes MBC pour modéliser la dynamique du métabolisme, considéré indépendamment de la régulation, et ce lorsque la quantité de données est faible et sans poser d'hypothèse sur l'optimum du système métabolique.

En associant le formalisme des MBC avec l'échantillonnage de l'ensemble des répartitions de flux possibles, j'ai mis au point une nouvelle approche pour modéliser la dynamique du métabolisme. Ce travail est présenté au fil de la troisième partie « Dynamique d'un réseau métabolique avec un modèle à base de contraintes : approche par échantillonnage des trajectoires solutions » (p. 96).

Dans le premier chapitre de cette dernière partie, je propose tout d'abord un tour d'horizon des méthodes qui permettent (*i*) de gérer la multitude de répartition de flux possibles ou (*ii*) d'introduire une dimension dynamique dans les MBC (chapitre 6, p. 99).

Je présente ensuite l'approche développée, sous la forme d'un article, dans le chapitre 7 (p. 129). Afin d'illustrer les capacités de cette méthode, je l'ai appliquée pour modéliser le comportement du métabolisme de la bactérie *Corynebacterium glutamicum* lors de la croissance en limitation en biotine. Les résultats obtenus sont aussi présentés au fil de l'article.

Dans le chapitre 8 (p. 161), j'approfondis plusieurs points brièvement abordés dans l'article, notamment, le choix de la méthode d'échantillonnage, et le choix des paramètres de l'approche. Le chapitre 9 (p. 187) est ensuite l'occasion de discuter d'une part, de l'impossibilité de modéliser l'une des périodes de la culture de *C. glutamicum*, et d'autre part de montrer en quoi les prédictions peuvent être utilisées pour améliorer le modèle. Enfin, le chapitre 10 (p. 223) conclut ce travail de thèse, et propose quelques éléments de réflexions pour la conception d'autres approches.

Première partie

Des réseaux biologiques aux modèles

Une première étape dans l'exercice de modélisation est d'une part de formaliser les entités biologiques qui vont être modélisées, et d'autre part de définir les relations qui existent entre ces entités. *Quelles sont les entités mises en jeu dans les réseaux métaboliques ? Dans les réseaux de régulation génétique ? Qu'entend-on par réseau métabolique ? Par réseau de régulation génétique ?* Ce travail de formalisation est présenté dans le chapitre 1 (p. 9).

Le modèle est le principal outil utilisé pour l'étude des interactions entre entités biologiques. De manière plus générale, l'outil *modèle* est largement rencontré en biologie. *Qu'est-ce qu'un modèle ? Quel est son usage en biologie ?* Dans le chapitre 2 (p. 39), je propose une ouverture sur la notion de modèle, ses usages en biologie, et plus particulièrement dans le champ disciplinaire de la bio-informatique.

Chapitre 1

Entités et réseaux biologiques

D'un point de vue purement chimique, une cellule n'est composée que d'une collection de molécules. La « vie », cette capacité à s'adapter et à se reproduire, nait à partir de l'organisation complexe de ces molécules.

Ce chapitre présente les molécules, leurs organisations, et les phénomènes biochimiques qui sont impliqués dans le métabolisme et la régulation génétique chez les procaryotes. Afin d'en faciliter la compréhension par les lecteurs non biologistes, certaines portions de ce chapitre abordent délibérément des connaissances de base en biologie.

La figure 1.1 (p. 11) propose une représentation de l'organisation des entités et phénomènes biologiques impliqués dans le métabolisme et la régulation génétique. Cette organisation est inspirée, en partie, de l'ontologie utilisée dans les bases de données BioCyc [Karp 2000]. Dans cette figure, je distingue 4 grands types d'entités :

— Les molécules qui contiennent l'information génétique. Dans le contexte du projet, ces molécules sont considérées comme statiques, l'information génétique et les concentrations de ces molécules n'évoluent pas.

— Les molécules qui sont les acteurs du réseau. Ces acteurs sont des entités physiques, dont la concentration peut varier selon l'état du réseau.

— Les fonctions biochimiques, que possèdent certains acteurs du réseau. Une fonction biochimique représente la capacité d'une entité à intervenir dans un processus biologique. Le concept de fonction biochimique permet de relier chaque acteur au(x) processus dans le(s)quel(s) il est impliqué. Si un acteur participe à trois processus, par exemple à trois réactions chimiques différentes, alors cet acteur possède trois fonctions biochimiques.

— Les processus biologiques, qui sont le résultat de la réalisation des fonctions biochimiques.

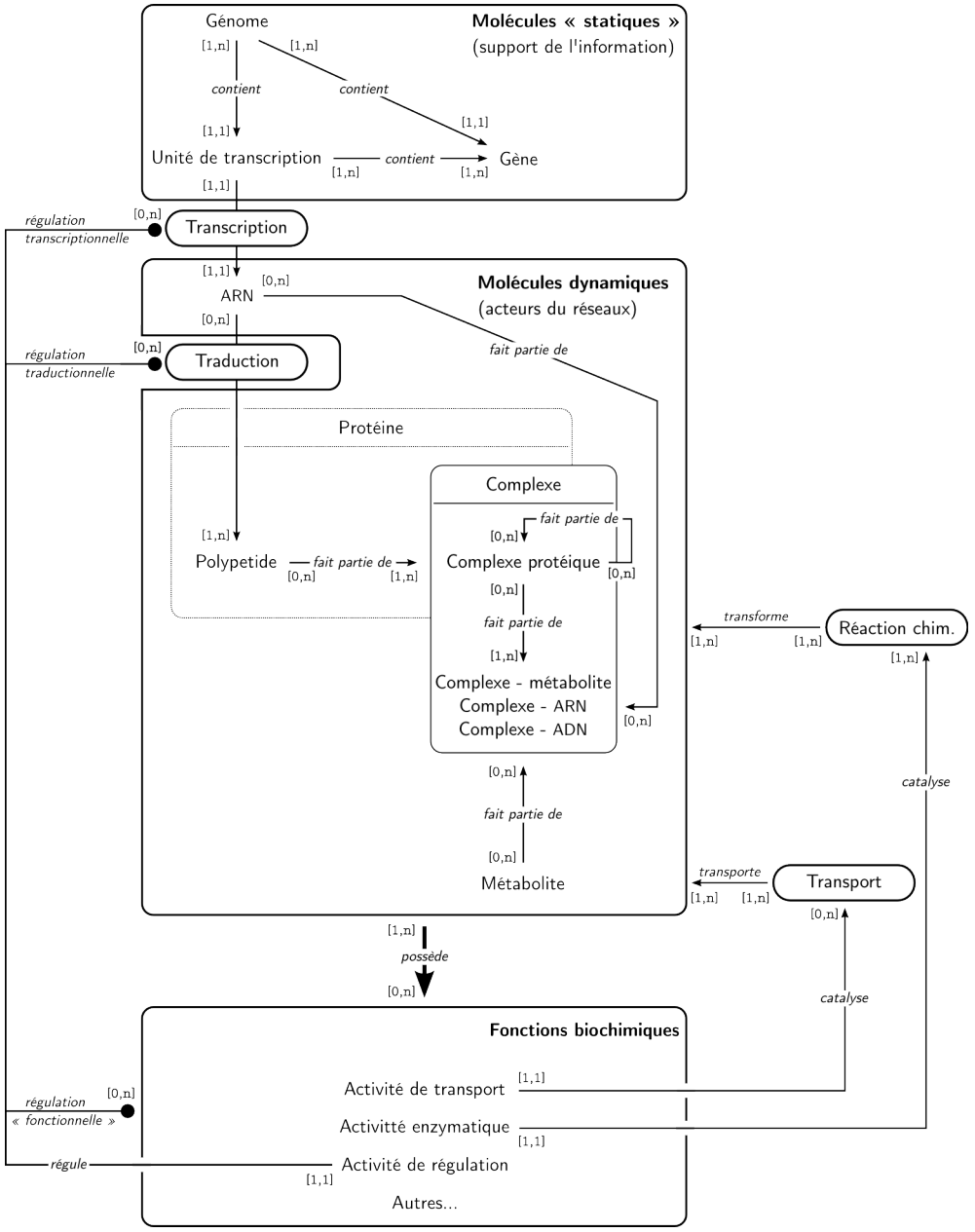


Figure 1.1. – Organisation simplifiée des entités et des phénomènes biologiques impliqués dans le métabolisme et la régulation génétique. Les relations entre entités sont indiquées par des flèches. Les valeurs entre crochets indiquent les cardinalités, c'est-à-dire le nombre minimum et maximum de relations possibles entre deux entités. Par exemple : une unité de transcription peut contenir un à plusieurs gènes, et un gène peut être contenu par une à plusieurs unités de transcription.

1.1. Support de l'information génétique

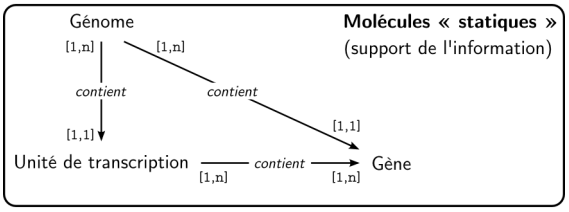


Figure 1.2. – Organisation des molécules qui détiennent l'information génétique.

La définition stricte du terme *information génétique* correspond à l'ensemble des informations qui peuvent être transmises à une descendance par l'intermédiaire des gènes. Le terme d'*expression* de l'information génétique est couramment employé pour désigner la « fabrication », en suivant les informations génétiques, de molécules qui joueront un rôle actif dans la cellule.

Ces informations génétiques sont portées par le matériel génétique. Chez les procaryotes, et de manière plus générale, chez la plupart des organismes à l'exception de quelques virus, ce matériel génétique est constitué de molécules d'ADN (acide désoxyribonucléique).

1.1.1. ADN, chromosome et génome

Les informations génétiques sont stockées dans l'ADN sous la forme d'un code constitué de 4 nucléotides : l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). La séquence dans laquelle ces nucléotides sont agencés détermine les informations à disposition de la cellule.

Les molécules d'ADN sont le plus souvent rencontrées sous la forme d'une double hélice. Il s'agit de deux molécules d'ADN enroulées l'une autour de l'autre. L'appariement des deux molécules — des deux brins — est réalisé grâce à la formation de liaisons faibles entre certains couples de nucléotides : entre l'adénine et la thymine, et entre la cytosine et la guanine. La figure 1.3 illustre la structure de la double hélice d'ADN.

L'ADN, sous sa forme double brin, est le principal constituant des *chromosomes*. Le chromosome est une structure dans laquelle l'ADN, guidé par

Figure 1.3. – Nucléotides, double hélice d'ADN, sur-enroulement et chromosome.

des protéines spécifiques, est compacté en formant des sur-enroulements (figure 1.3). Les modalités de cette compression sont différentes selon que l'on considère les procaryotes ou les eucaryotes. Cependant, dans les deux situations, ce phénomène est impliqué dans la régulation de l'expression de l'information génétique, en modifiant l'accessibilité des séquences d'ADN [Willenbrock 2004].

La plupart des procaryotes ne possèdent qu'un seul chromosome de forme circulaire. Dans cette situation, cet unique chromosome contient la majorité de l'information génétique. Le terme *majorité* est ici employé, car les procaryotes peuvent aussi posséder un ou plusieurs plasmides. Les plasmides, qui ne seront pas abordés plus en détail, sont des molécules d'ADN physiquement distinctes des chromosomes, et capables de se répliquer indépendamment du chromosome.

Le concept de *génome* regroupe, pour un organisme donné, l'ensemble des supports de l'information génétique. Pour les procaryotes, ce génome sera constitué le plus souvent d'un chromosome, et parfois d'un ou plusieurs plasmides additionnels.

1.1.2. Gène et unité de transcription

Si l'on considère plus en détail l'organisation des séquences génomiques, certaines régions jouent un rôle particulier dans l'expression de l'information génétique. Il s'agit des gènes.

Un *gène* est une portion d'ADN qui contient les informations nécessaires pour produire une séquence d'ARN (acide ribonucléique). Selon le gène

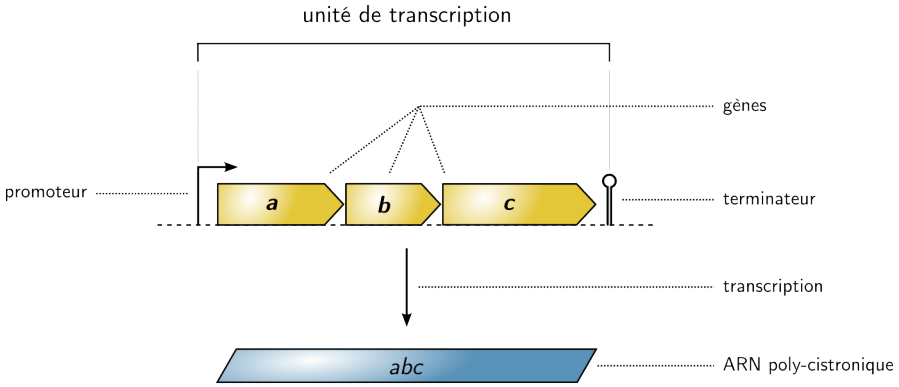


Figure 1.4. – Unité de transcription et ARN polycistronique.

considéré, la séquence d'ARN produite est soit un ARN messager qui code un polypeptide, soit un ARN non-codant tel qu'un ARN ribosomal ou un ARN de transfert. Un gène peut être localisé sur l'un ou l'autre des brins de la double hélice d'ADN.

La production d'une molécule d'ARN à partir du gène correspondant est effectuée lors d'un processus cellulaire nommé transcription. Ce processus est réalisé par un ensemble de molécules dédiées, qui reconnaissent le début du gène à transcrire, synthétisent la séquence d'ARN correspondante, et stoppent la transcription lorsque la fin du gène est atteinte.

Pour guider ces opérations, un ensemble de séquences entourent généralement les gènes. Par exemple, la séquence promotrice — ou promoteur — participe à la phase d'initiation de la transcription, tandis que la séquence terminatrice — ou terminateur — induit la fin de la transcription.

Chez les procaryotes, certains gènes contigus partagent un unique promoteur et un unique terminateur. Le promoteur se situe alors en amont du premier gène, et la séquence terminatrice en aval du dernier gène. La figure 1.4 donne un exemple d'une telle organisation génétique.

Dans cette situation, la transcription commence au début du premier gène et se termine à la fin du dernier gène. En conséquence, l'information génétique portée par ce groupe de gène est codée dans un unique ARN, qui est alors qualifié d'ARN polycistronique. Cet ARN sera ensuite traité par la machinerie cellulaire et conduira à la formation de plusieurs produits de gènes.

La diversité des organisations possibles des séquences qui permettent la production des ARN a conduit à l'émergence du concept d'unité de transcription.

Une *unité de transcription* est définie comme une séquence d'ADN, qui code un unique ARN mono- ou polycistronique, et qui contient les séquences nécessaires à sa transcription. Ces séquences incluent un promoteur, la ou les séquences des gènes transcrits, et un terminateur.

1.1.3. Support de l'information : point de vue statique

L'ADN et les chromosomes sont des molécules qui portent l'information génétique.

Cette information génétique peut évoluer au fil de temps, notamment par le biais de mutations ponctuelles ou de grands remaniements chromosomiques. Ces phénomènes peuvent aussi modifier l'organisation et le nombre des unités de transcription.

Le génome et son organisation possèdent donc une dynamique, mais qui se mesure sur une échelle de temps bien plus longue que le contexte dans lequel se situe mon projet de recherche.

En conséquence, pour la suite de ce manuscrit, je vais considérer que la quantité de ces molécules et les informations qui y sont stockées sont stables et constantes à l'échelle des réactions métaboliques et des régulations génétiques.

1.2. Molécules effectrices

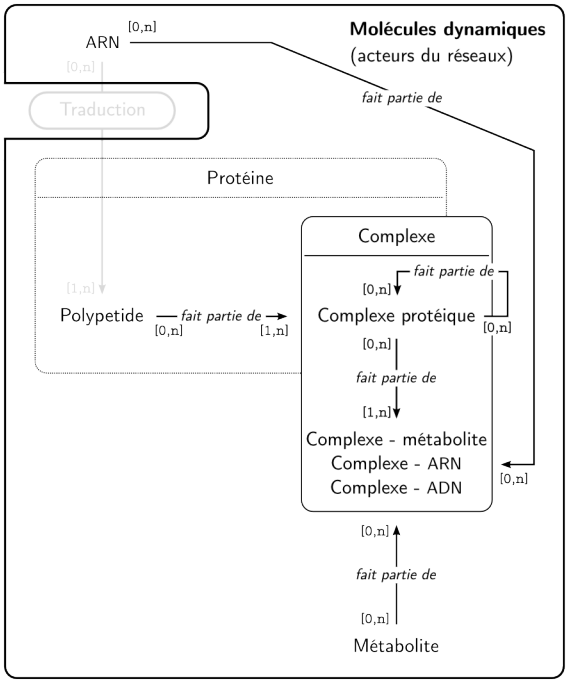


Figure 1.5. – Molécules effectrices.

Lors de la section précédente, nous avons vu que l'information génétique était exprimée sous la forme d'ARN. Les ARN peuvent ensuite être traduits en protéines. Les ARN et les protéines forment, avec les métabolites, les principaux acteurs des processus cellulaires.

1.2.1. Métabolites

Les métabolites sont des composés chimiques, le plus souvent de petite taille, qui peuvent être consommés ou produits par des réactions chimiques au sein des cellules. Les métabolites peuvent aussi être importés dans la cellule, ou être exportés en dehors.

Les métabolites jouent différents rôles au sein des cellules. Les métabolites permettent de produire de l'énergie, de fournir le pouvoir réducteur, de

synthétiser de nouvelles molécules, et ils tiennent aussi un rôle dans des phénomènes de régulation et de signalisation cellulaire.

1.2.2. ARN

Les ARN, aussi appelés transcrits, sont issus de l'expression des unités de transcription. On peut distinguer deux grandes classes d'ARN : les ARN messagers et les ARN non-codants.

Les ARN messagers sont traduits en polypeptides. Ce type d'ARN peut être considéré comme un support temporaire de l'information génétique.

Les ARN non-codants ne codent pas un polypeptide. Ces ARN peuvent directement contribuer à l'apparition d'activités biochimiques et de processus cellulaires. C'est le cas, par exemple, des ARN ribosomaux : en se liant avec d'autres ARN et des polypeptides, ils forment le ribosome, complexe moléculaire essentiel dans le processus de traduction. Un autre exemple d'ARN fonctionnel est celui des petits ARN (*sRNA ; small RNA* en anglais), qui sont capable de réguler l'expression de gènes en inhibant leur transcription.

À l'origine sous-estimée, l'importance des ARN non-codants dans les processus biologiques et leurs régulations est aujourd'hui reconnue.

1.2.3. Polypeptides et protéines

Les protéines sont des molécules constituées d'une ou plusieurs séquences d'acides aminés. Ces séquences d'acides aminés, aussi appelée polypeptides, sont synthétisées au cours du processus de traduction des ARN messagers.

Lorsqu'une protéine n'est constituée que d'une seule séquence, les termes *polypeptide* et *protéine* peuvent être considérés comme équivalent. Lorsque la protéine est constituée de plusieurs polypeptides, le terme de *complexe protéique* est préféré.

Chaque acide aminé possède des caractéristiques physico-chimiques particulières. L'ordre des acides aminés détermine ainsi la forme tridimensionnelle des protéines. La figure 1.6 (page suivante) illustre les différents niveaux de structure qui sont communément considérés pour étudier les protéines.

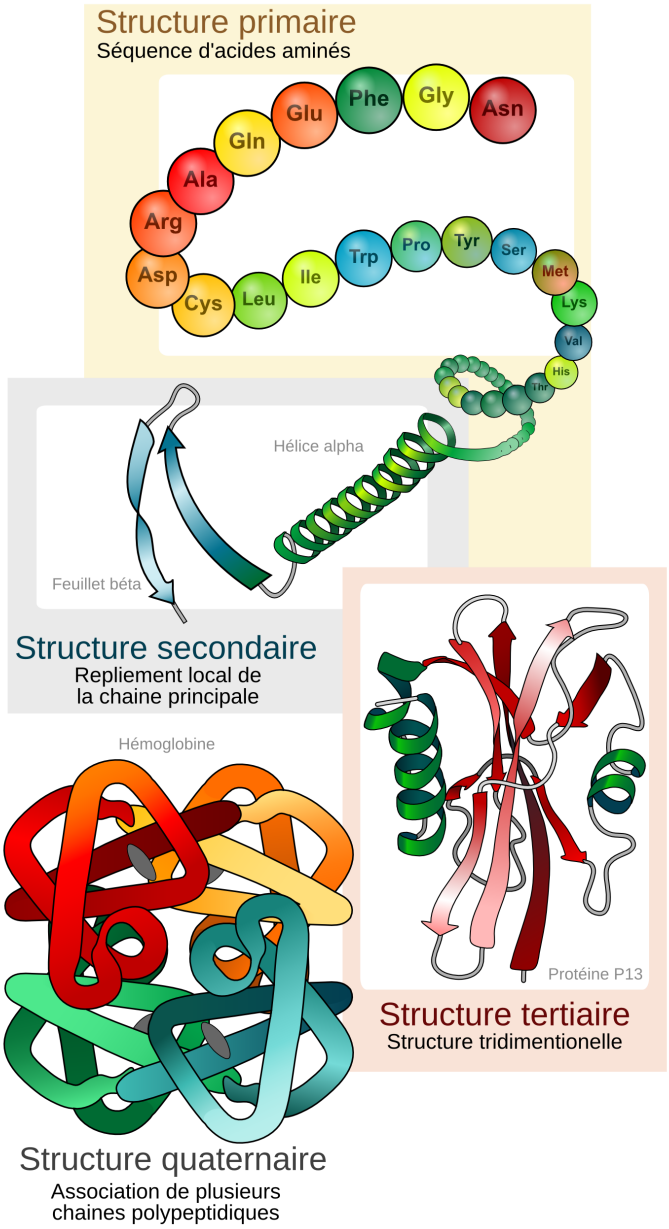


Figure 1.6. – Niveaux de structure communément considérés chez les protéines. La structure quaternaire correspond à la notion de complexe protéique.

De par la diversité de leurs structures, les protéines remplissent de nombreux rôles au sein des cellules et elles sont impliquées dans la plupart des processus cellulaires.

1.2.4. Complexes

Dans le cadre de cette revue, un complexe est un assemblage d'entités biologiques. On peut définir différents types de complexes selon les entités qui le constituent. À noter qu'un complexe peut lui-même faire partie d'un autre complexe.

Un *complexe protéique* correspond à l'assemblage de plusieurs polypeptides. Chaque polypeptide est alors considéré comme une sous-unité. Selon que les sous-unités sont identiques ou non, le complexe est qualifié d'*homo-* ou d'*hétéro*multimère. La figure 1.7 illustre trois structures quaternaires de complexes protéiques.

De nombreux autres types de complexes peuvent être formés au sein des cellules, de manière stable ou transitoire. Citons notamment : les complexes protéine-métabolite, où une protéine est liée à un métabolite ; les complexes protéine-ARN, parfois appelés complexes ribonucléiques ; et les complexes protéine-ADN.

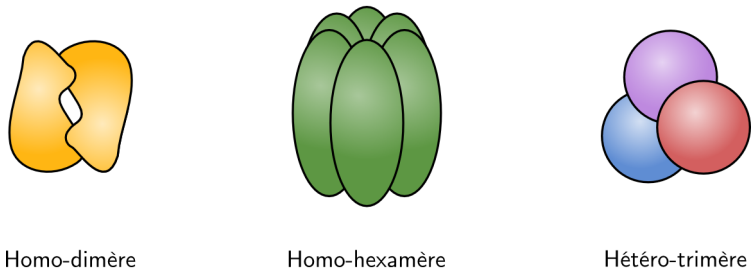


Figure 1.7. – Exemples de complexes protéiques.

1.3. Processus biologiques

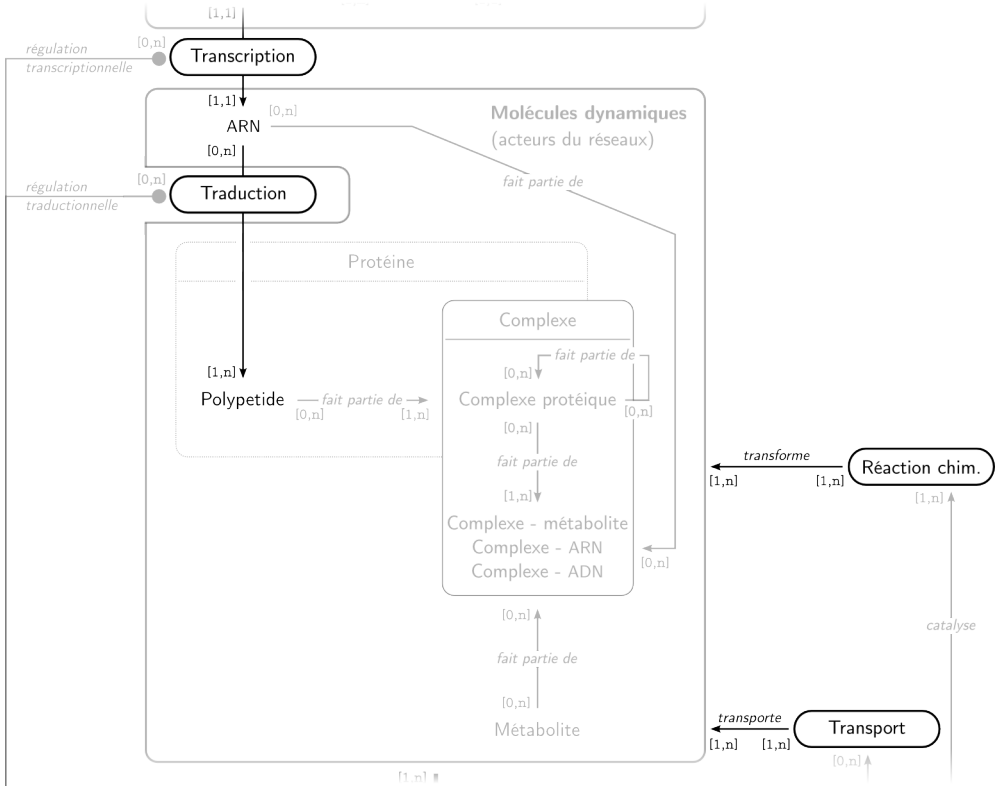


Figure 1.8. – Processus biologiques.

Les processus biologiques résultent des interactions coordonnées entre les différentes entités qui constituent les cellules. Dans cette section, j'aborde 4 processus qui sont impliqués dans le métabolisme et la régulation génétique : la transcription, la traduction, le transport de molécules et les réactions chimiques.

1.3.1. Transcription

La transcription permet la transmission de l'information portée par les unités de transcription vers des molécules d'ARN.

Ce processus fait intervenir un complexe protéique, nommé ARN polymérase, qui est responsable de la « lecture » de l'ADN et de la synthèse de l'ARN correspondant. Un certain nombre de protéines, appelées facteurs de transcription, permettent de reconnaître les unités de transcription qui seront transcrites par l'ARN polymérase.

La transcription d'une unité de transcription se fait en trois étapes :

1. Étape d'initiation. L'ARN polymérase (sous sa forme holoenzyme) associée au facteur sigma reconnaît la séquence promotrice de l'unité de transcription.
2. Étape d'élongation. L'ARN polymérase (sous sa forme *core* enzyme) parcourt la séquence de nucléotides du (ou des) gène(s) et synthétise au fur et à mesure la séquence d'ARN complémentaire du brin d'ADN en cours de lecture.
3. Étape de terminaison. L'ARN polymérase, l'ADN, et l'ARN se séparent. Cette séparation est due à la présence d'une séquence terminatrice au niveau de l'unité de transcription. La séquence terminatrice peut éventuellement faire intervenir des protéines appelées facteurs de terminaison (protéines Rho).

Outre la capacité plus ou moins importante du promoteur à induire l'étape d'initiation de la transcription (on parle de force du promoteur), la fréquence et la vitesse à laquelle les unités de transcription sont transcrites sont contrôlées par divers mécanismes cellulaires. Cette régulation de la transcription passe notamment par l'action d'entités biologiques qui vont favoriser l'étape d'initiation de la transcription, on parle alors d'activateur transcriptionnel, ou au contraire limiter l'étape d'initiation, on parlera de répresseur transcriptionnel.

1.3.2. Traduction

Le processus de traduction correspond au décodage de l'information contenue dans les ARN messagers et mène à la formation des polypeptides.

Ce décodage fait intervenir le ribosome, complexe protéique qui va parcourir l'ARN et polymériser le polypeptide, ainsi que des ARN de transfert qui vont fournir les acides aminés qui seront assemblés. La correspondance entre la séquence des nucléotides qui constituent l'ARN et la séquence des acides aminés est déterminée par les règles du code génétique. Chaque triplet de nucléotides code un acide aminé, sauf trois codons qui signalent la fin de la séquence (codons *stop*).

Tout comme la transcription, la traduction est réalisée en trois étapes :

1. Étape d'initiation. Chez les procaryotes, le ribosome reconnaît le premier codon, aussi appelé codon *start*, qui indique où débute la séquence à traduire. L'acide aminé correspondant est alors incorporé. La reconnaissance du codon *start* est généralement facilitée par la présence d'une courte séquence, la séquence de Shine-Dalgarno, qui est située quelques nucléotides en amont du codon *start* (entre 9 et 11 nucléotides).
2. Étape d'élongation. Le ribosome lit la séquence codon par codon, en se déplaçant le long de l'ARN. À chaque codon lu, l'acide aminé correspondant est fourni par un ARN de transfert et est ajouté au polypeptide en cours de formation.
3. Étape de terminaison. Lorsque le ribosome atteint un codon *stop*, aucun nouvel acide aminé n'est ajouté, le ribosome se sépare de l'ARN en cours de lecture, et le polypeptide est relâché.

La traduction peut être régulée par différentes entités. Généralement, cette régulation portera sur l'étape d'initiation de la traduction. Lorsqu'un gène code une protéine, la traduction est une étape indispensable pour la bonne expression de ce gène. En conséquence, la régulation de la traduction peut être considérée comme faisant partie de la régulation de l'expression génique.

1.3.3. Réaction chimique

Une réaction chimique est un processus au cours duquel une ou plusieurs molécules sont transformées en d'autres molécules. Cette transformation peut se faire soit par un transfert d'atomes entre plusieurs molécules, soit par la modification de l'arrangement des atomes au sein d'une molécule. Les molécules avant transformation sont appelées des substrats ou des réactifs, tandis que celles issues de la réaction sont nommées produits.

Dans le contexte des cellules, les substrats et les produits des réactions sont le plus souvent des métabolites. Mais d'autres types d'entités peuvent aussi être transformés. C'est le cas, par exemple, des protéines qui sont phosphorylées, et des ARN qui subissent des méthylations après avoir été transcrit.

L'un des principes fondamentaux des réactions chimiques est la conservation du nombre d'atomes de chaque espèce chimique. Ainsi, dans la réaction de combustion du méthane,

$$\mathbf{1}\ \mathsf{CH_4} + \mathbf{2}\ \mathsf{O_2} \rightarrow \mathbf{1}\ \mathsf{CO_2} + \mathbf{2}\ \mathsf{H_2O}$$

les nombres d'atomes de carbone (C), d'oxygène (O) et d'hydrogène (H) sont identiques avant et après la transformation. Le nombre qui précède chaque molécule représente la stœchiométrie de la réaction, c'est-à-dire la proportion nécessaire de chaque entité pour que la quantité de matière soit maintenue.

Un autre principe fondamental des réactions chimiques est la conservation de l'énergie. Ainsi, lors de la combustion du méthane,

$$1\ \mathsf{CH_4} + 2\ \mathsf{O_2} \rightarrow 1\ \mathsf{CO_2} + 2\ \mathsf{H_2O} + \textbf{énergie}$$

l'énergie observée sous forme de chaleur est le résultat d'un transfert d'une partie de l'énergie potentielle contenue dans les réactifs vers l'environnement.

L'énergie potentielle d'une molécule est déterminée, en partie, par les liaisons entre les atomes qui la constitue. De manière générale, sans entrer dans des considérations thermodynamiques, les réactions chimiques ne sont réalisables qu'à la condition que l'énergie potentielle détenue par les substrats soit supérieure à celle des produits. Lors de la transformation, une partie de l'énergie des substrats est libérée dans le milieu environnant, le plus souvent sous la forme de chaleur.

L'énergie libre d'activation est l'énergie nécessaire pour qu'une réaction chimique se réalise. Cette énergie est apportée par le milieu environnant. Lors de la combustion du méthane, la réaction est ainsi déclenchée par une augmentation de la chaleur, par exemple une étincelle.

$$1\ \mathsf{CH_4} + 2\ \mathsf{O_2} \xrightarrow{\text{🗲}} 1\ \mathsf{CO_2} + 2\ \mathsf{H_2O} + \text{énergie}$$

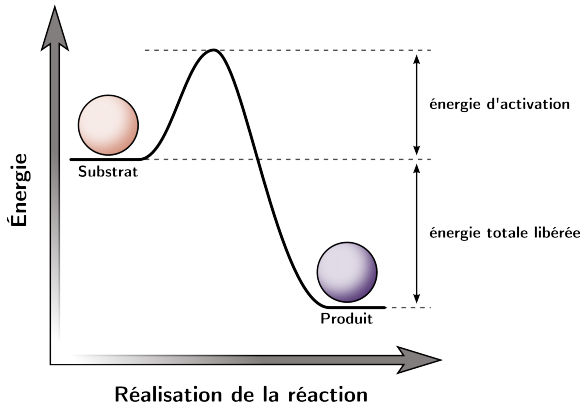


Figure 1.9. – Énergie d'activation d'une réaction chimique.

La « barrière » d'activation joue un rôle indispensable pour la stabilité des molécules. En effet, le besoin en énergie pour amorcer les réactions chimiques limite grandement la fréquence des transformations. La figure 1.9 illustre les notions énergétiques mises en jeu lors des réactions chimiques.

Les réactions chimiques sont indispensables à la vie des cellules. Des réactions de dégradation permettent aux cellules d'obtenir des molécules de structure simple (par ex. : les acides aminés) et des molécules énergétiques (par ex. : l'adénosine triphosphate, ATP). Ces produits sont ensuite utilisés pour constituer des structures complexes telles que l'ADN ou les protéines.

Ces ensembles de transformations nécessitent un contrôle précis de l'enchainement des réactions chimiques. De plus, la fréquence des réactions chimiques doit être suffisamment importante pour être compatible avec la vie. Par exemple, si elle n'est pas accélérée, la transformation d'une molécule de glucose en glucose-6-phosphate a — en théorie — une fréquence de réalisation d'une réaction tous les 300 000 ans [Koshland 1956]. En comparaison, en conditions favorables, une bactérie telle qu'*Escherichia coli* a un temps de division cellulaire de 20 minutes, durée pendant laquelle de très nombreuses réactions chimiques ont lieu de manière coordonnée.

Ces deux caractéristiques indispensables, contrôle et vitesse élevée des transformations, sont assurées par certaines entités biologiques qui possèdent soit une activité enzymatique soit une activité de régulation. Ces fonctions biochimiques seront développées dans la section suivante (p. 27).

1.3.4. Transport

Le transport est un processus au cours duquel une ou plusieurs molécules passent de l'intérieur de la cellule au milieu extracellulaire (ou inversement), ou bien d'un compartiment cellulaire à un autre.

Chez les procaryotes, de manière simplifiée, on peut considérer qu'il n'existe qu'un seul compartiment cellulaire, le cytoplasme. La membrane plasmique sépare l'intérieur de la cellule de l'environnement extérieur. Cela permet notamment à la cellule de maintenir des concentrations élevées de molécules dans le cytoplasme. Cependant, la membrane plasmique n'est pas totalement imperméable : plusieurs mécanismes de transport permettent l'entrée et la sortie de molécules spécifiques. Les mécanismes les plus communs sont présentés par la figure 1.10.

La diffusion directe et la diffusion facilitée sont deux mécanismes de transport qui ne nécessitent pas l'apport direct d'énergie. Ils sont qualifiés de transports passifs. Au cours de ces transports, le mouvement des molécules se fait toujours depuis le compartiment où les concentrations sont les plus élevées vers un autre compartiment. On parle de gradient de concentration ou encore de gradient électrochimique.

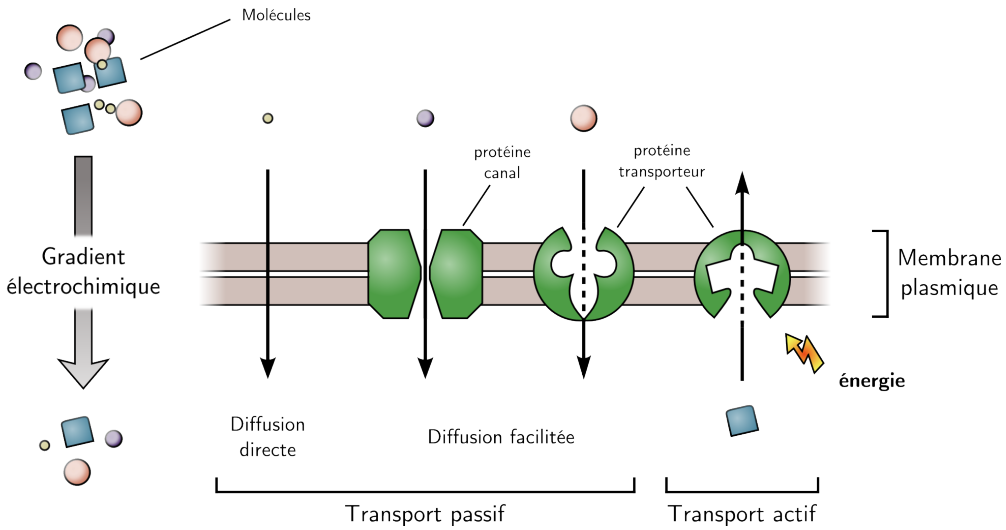


Figure 1.10. – Les différents types de transport cellulaire.

Le mécanisme de transport le plus simple est la diffusion directe à travers la membrane. La membrane plasmique étant en grande partie constituée de lipides, ce type de transport concerne essentiellement les molécules qui sont lipophiles. Certaines molécules de très petite taille et non lipophiles peuvent aussi traverser directement la membrane.

La diffusion facilitée exploite aussi le gradient électrochimique pour transporter des molécules, mais ce transport est cette fois facilité par des protéines qui traversent la membrane. Je caractérise ces protéines comme possédant une fonction biochimique de type *activité de transport*. Chacune de ces protéines permet le transport spécifique d'une ou de quelques molécules. Deux grandes familles de protéines interviennent dans la diffusion facilitée : les protéines *canaux* et les protéines *transporteurs*. Une protéine *canal*, aussi appelée pore, forme un canal qui traverse la membrane plasmique et laisse passer des ions et des molécules de petite taille. Une protéine *transporteuse*, aussi appelée perméase, est capable de transférer des molécules d'un côté à l'autre de la membrane en changeant de forme.

Le dernier mécanisme communément rencontré est le transport actif. Le terme *actif* est ici employé, car ce mécanisme nécessite l'apport d'énergie, ce qui permet un mouvement des molécules contre le gradient électrochimique. L'énergie est fournie, de manière directe ou indirecte, en couplant des protéines de transports avec une réaction chimique.

De la même manière que le contrôle des réactions chimiques est crucial pour diriger les flux de molécules à l'intérieur des cellules, la coordination des entrées et des sorties de molécules est indispensable au bon fonctionnement de la machinerie cellulaire. Pour cela, des entités qui possèdent une activité de régulation permettent de contrôler les protéines impliquées dans les transports facilités et les transports actifs.

1.4. Fonctions biochimiques

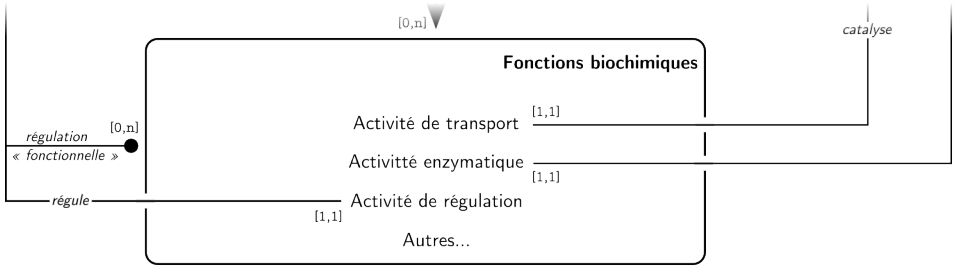


Figure 1.11. – Fonctions biochimiques

Une fonction biochimique représente la capacité d'une entité biologique à agir sur d'autres entités. Les métabolites, les ARN, les protéines et les complexes peuvent posséder des fonctions biochimiques.

Dans le cadre des réseaux métaboliques et de régulation génétique, trois grandes fonctions biochimiques peuvent être identifiées. Il s'agit des activités enzymatiques, de transport, et de régulation.

Les activités biochimiques qui réalisent les processus de transcription et de traduction ne seront pas abordées dans cette partie : ces activités font intervenir des ensembles complexes de phénomènes qui sont en dehors du cadre de cette thèse. La régulation de ces processus sera toutefois abordée dans la partie « activité de régulation ».

1.4.1. Activité enzymatique

Une activité enzymatique est la faculté que possèdent certaines entités à catalyser une réaction chimique. Au sein des cellules, ce type de fonction biochimique est principalement rencontrée chez les protéines et les complexes, mais aussi chez certains ARN. Ces entités biologiques, capables de catalyser des réactions chimiques, sont appelées *enzymes*, ou *ribozyme* lorsque l'entité est un ARN.

Le concept d'activité enzymatique permet de créer des paires *entité biologique/réaction chimique*, en reliant chaque enzyme à chacune des réactions qu'elle catalyse.

La catalyse est un phénomène dans lequel la présence d'une substance, appelée catalyseur, va augmenter la fréquence de réalisation d'un processus. La catalyse d'une réaction chimique va ainsi induire une augmentation de la vitesse de la réaction, autrement dit, du nombre de transformations réalisées par unité de temps.

Une caractéristique importante des catalyseurs est qu'ils ne sont pas modifiés par les réactions : après avoir facilité une première transformation, une enzyme peut directement intervenir dans une deuxième transformation. Ce caractère « réutilisable » des enzymes leur permet d'avoir un impact important tout en étant présentes à de très faibles concentrations.

Les enzymes facilitent les réactions en abaissant l'énergie d'activation nécessaire pour déclencher la transformation (figure 1.12). Cette diminution de l'énergie d'activation permet aux enzymes d'accélérer de manière très importante les réactions chimiques. À titre d'exemple, la réaction de phosphorylation du glucose en glucose-6-phosphate est théoriquement 13 milliards de fois plus fréquente lorsqu'elle est catalysée par l'enzyme hexokinase [Koshland 1956].

La diminution de l'énergie d'activation nécessite une interaction entre l'enzyme et les substrats de la réaction. Cette interaction se fait au niveau d'une région particulière de l'enzyme, appelée site actif, et conduit à la formation d'un complexe enzyme-substrat (figure 1.13). Le site actif, formé par le repliement en trois dimensions de l'enzyme, peut être divisé en deux parties : le site de reconnaissance et le site catalytique. Le site de reconnaissance est capable de fixer spécifiquement un ou quelques substrats. Le site catalytique va ensuite faciliter la transformation du ou des substrats en produit(s). Une fois la transformation réalisée, enzyme et produit(s) se séparent, et l'enzyme peut être réutilisée pour une nouvelle itération de réaction catalysée.

Du fait du haut niveau de spécificité d'un site actif — spécificité nécessaire à la fois pour fixer correctement les substrats et pour faciliter une réaction chimique particulière — une enzyme n'est généralement capable de catalyser qu'une seule réaction chimique.

Chaque activité enzymatique est associée à un code numérique appelée numéro EC (*Enzyme Commission number* en anglais). Ce numéro EC permet d'indiquer de façon précise la réaction chimique qui est catalysée. Un numéro EC est constitué de 4 parties séparées par des points. De gauche à droite, chaque partie apporte de plus en plus d'informations.

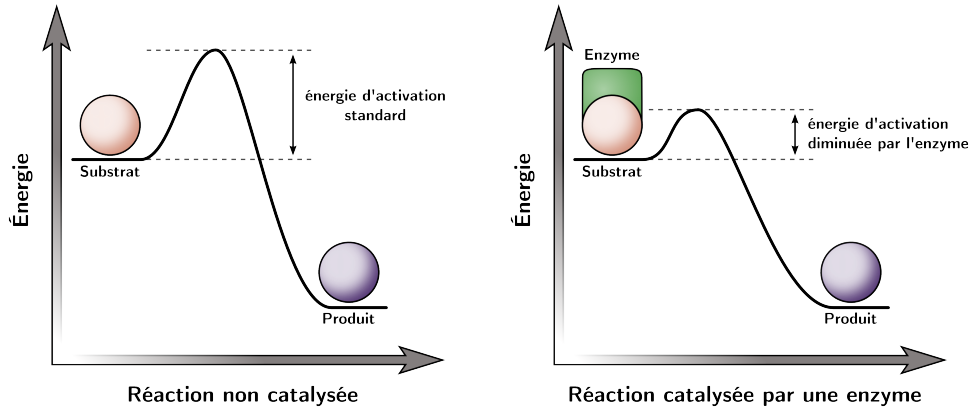


Figure 1.12. – Diminution de l'énergie d'activation par une enzyme.

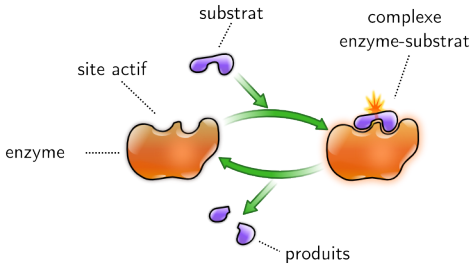


Figure 1.13. – Site actif, complexe enzyme-substrat, et réutilisation de l'enzyme.

Par exemple, le numéro « EC 1.1.1.27 » désigne :

1.-.-.-	une oxydoréductase...
1.1.-.-	... qui agit sur un groupement CH-OH d'un substrat...
1.1.1.-	... et qui utilise le NAD^+ ou le $NADP^+$ comme accepteur...
1.1.1.27	... et dont le substrat est le L-lactate.

Pour une concentration d'enzyme fixée, plus la concentration en substrat augmente, plus la proportion d'enzyme complexée avec un substrat augmente, et plus la vitesse globale de la réaction est importante. Cependant, l'augmentation de la quantité de substrat va progressivement conduire à un phénomène de saturation de l'enzyme : la proportion d'enzyme disponible pour catalyser de nouvelles transformations diminue, et donc l'augmentation de la vitesse de la réaction est moins importante (figure 1.14). En conséquence, la vitesse d'une réaction chimique est dépendante de la quantité de substrats *et* de la quantité d'enzyme disponibles.

Divers mécanismes cellulaires font varier la quantité des enzymes disponibles. Ces mécanismes seront abordés dans le point « activité de régulation ».

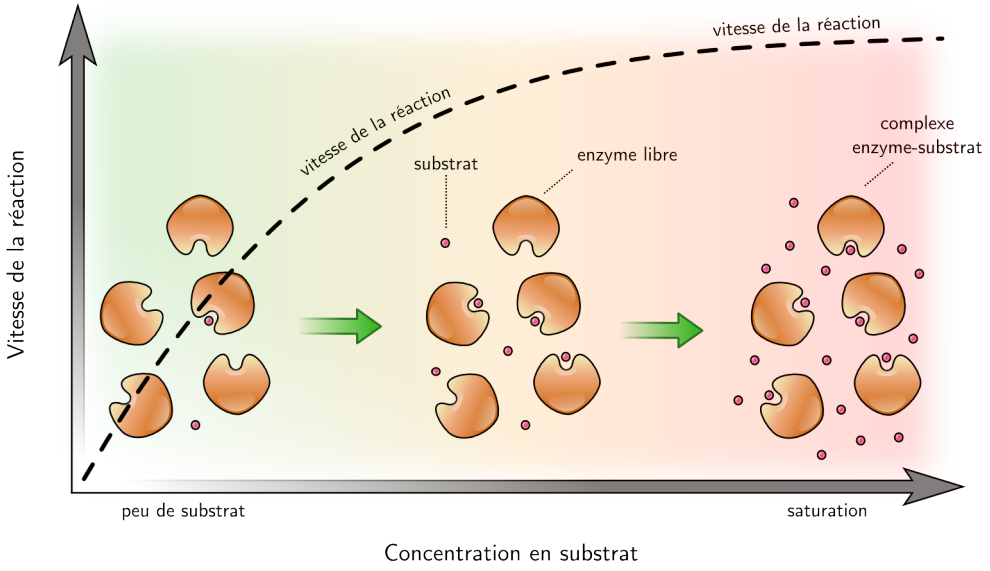


Figure 1.14. – Concentration en substrat, saturation de l'enzyme et vitesse de la réaction catalysée.

1.4.2. Activité de transport

Une activité de transport est la faculté que possèdent certains acteurs du réseau à catalyser le transport de molécules d'un compartiment à un autre.

Les entités qui possèdent cette fonction biochimique sont principalement des protéines, généralement assemblées sous la forme de complexe protéique. Les phénomènes de transport catalysés sont la diffusion facilitée et le transport actif.

De la même manière que les enzymes reconnaissent de façon spécifique les substrats, une entité de transport donnée ne permet le passage que d'une seule ou de quelques molécules spécifiques. Par ailleurs, la concentration en transporteur a une influence sur la vitesse des transports de molécules.

1.4.3. Activité de régulation

Une activité de régulation représente la capacité d'une entité biologique à contrôler la réalisation ou non des processus biologiques, ainsi que la vitesse à laquelle ces processus ont lieu.

Pour cela, une activité de régulation peut agir de deux manières différentes : soit modifier la quantité d'entités biologiques, soit changer la fonction biochimique des entités déjà produites.

Modification de la quantité d'entités. La variation de la quantité d'ARN et de protéines dépend des vitesses de synthèse et de dégradation de ces entités.

La régulation de la transcription, aussi appelée *régulation transcriptionnelle*, module la fréquence à laquelle un début de transcription a lieu (pour une unité de transcription donnée). Cette modulation est généralement due à une protéine régulatrice, parfois sous la forme d'un complexe protéine-métabolite, qui se fixe sur une portion du chromosome proche de l'unité de transcription régulée. Selon le type de régulateur, activateur ou répresseur, l'expression de l'unité de transcription sera augmentée ou diminuée. En conséquence la vitesse de production de l'ARN correspondant sera modifiée.

Lorsque ces ARN sont des ARN messagers, la traduction est une étape nécessaire pour que le produit de gène soit fonctionnel. La régulation de la traduction, ou *régulation traductionnelle*, contrôle le nombre de polypeptides synthétisés à partir d'un ARN messager.

À l'inverse, la dégradation diminue le nombre de molécules d'ARN et de protéines. La dégradation de l'ARN peut contribuer à diminuer le nombre d'ARN non-codants fonctionnels, mais aussi à diminuer le nombre d'ARN messagers, et en conséquence à diminuer l'expression d'un gène codant un polypeptide. La régulation de la dégradation des ARN et des protéines passe généralement par une étape de « marquage » de la molécule à dégrader, les molécules « marquées » sont ensuite prises en charge par des enzymes spécifiques. Le marquage d'un ARN pour qu'il soit dégradé peut être considéré comme une régulation qualifiée de *post-transcriptionnelle*, c'est-à-dire une régulation qui a lieu après la transcription, sur une séquence d'ARN, et avant l'éventuelle traduction.

Modification de la fonction des entités. La fonction biochimique des ARN, des protéines et des complexes dépend de leur structure en trois dimensions. En agissant sur cette structure, les phénomènes de régulation peuvent amplifier, atténuer ou supprimer une fonction biochimique.

Un premier mécanisme est la transformation d'une entité au moyen d'une réaction chimique pour lui ajouter ou lui retirer des groupements d'atomes, comme un groupement phosphate. Ce type de modification correspond soit à une régulation *post-transcriptionnelle* lorsque l'entité modifiée est un ARN fonctionnel, soit à une régulation *post-traductionnelle* si l'entité modifiée est une protéine. La phosphorylation peut, par exemple, faire passer un répresseur transcriptionnel d'un état actif (répression) à inactif (pas de répression), ou activer une enzyme en rendant son site catalytique fonctionnel.

Un second mécanisme est la modification de la structure par la fixation d'une autre entité. Ce type de régulation est rencontré chez certains régulateurs transcriptionnels : l'activité de régulation est (dés-)activée par la fixation d'un métabolite, comme la fixation du lactose sur le répresseur LacI dans le cas de l'opéron lactose. La régulation par modification de la structure est aussi communément rencontrée chez les enzymes, où la fixation d'un métabolite module l'activité enzymatique. Le métabolite sera considéré comme un inhibiteur s'il diminue l'activité enzymatique, ou comme un activateur s'il l'augmente.

Si l'on se concentre sur la régulation des enzymes, on peut distinguer trois types de régulation : les inhibitions (*i*) compétitive, (*ii*) non compétitive et (*iii*) incompétitive. L'inhibition compétitive est un phénomène dans lequel la fixation de l'inhibiteur sur l'enzyme empêche la fixation du substrat. Il s'agit ici d'une régulation par interaction physique, mais sans modification de la structure de l'enzyme : les fixations du substrat et de l'inhibiteur sont mutuellement exclusives, et les deux molécules se retrouvent en *compétition* pour se fixer à l'enzyme. Dans l'inhibition non compétitive, la fixation de l'inhibiteur n'a pas d'effet sur celle du substrat (et réciproquement). Par contre, la fixation de l'inhibiteur empêche la transformation du substrat par le site catalytique de l'enzyme (mécanisme de modification de la structure). Enfin, dans l'inhibition incompétitive, ou inhibition par blocage du complexe intermédiaire, l'inhibiteur favorise la formation du complexe enzyme-substrat. Puis, l'inhibiteur se fixe à ce complexe et empêche la transformation du substrat (mécanisme de modification de la structure).

Remarque sur les régulations globales des processus de transcription et de traduction. Les entités biologiques qui réalisent les processus de transcription — les ARN polymérases — sont présentes en quantités limitées au sein d'une cellule. En conséquence, les gènes sont en « compétition » pour être transcrits [Nyström 2004, De Vos 2011]. Par ailleurs, l'activité des ARN polymérases et des ribosomes peut être régulée, ce qui modifie de manière globale la fréquence de transcription des gènes, et de traduction des ARN messagers.

Bien que ces phénomènes « globaux » participent à la régulation de l'expression de l'information génétique, ils ne seront pas considérés dans la suite de cette thèse. Je ne m'attacherai qu'aux phénomènes de régulation « ciblés », n'agissant que sur une ou quelques entités, qui ont été décrits plus en amont dans cette partie.

1.5. Notion de réseaux biologiques

Au sens commun du terme, un réseau est un ensemble d'éléments qui sont reliés les uns aux autres.

Cette notion d'interconnexion entre éléments est aussi valable en biologie : un réseau biologique représente un ensemble d'entités biologiques qui interagissent entre elles. La nature des entités et des interaction va dépendre du type de réseau biologique considéré. Dans cette section, je présente les notions de réseau métabolique et de réseau de régulation génétique. Pour donner une vision plus globale de la notion de réseau biologique, j'aborderai rapidement deux autres types de réseaux fréquemment étudiés : les réseaux de signalisation et les réseaux d'interaction protéines-protéines.

1.5.1. Réseau métabolique

Le métabolisme est l'ensemble des réactions chimiques qui se déroulent au sein d'un être vivant. On parle de *métabolisme + substantif* lorsque l'on fait référence à une portion particulière du métabolisme, par exemple le *métabolisme des lipides* qui se réfère aux réactions impliquées dans la synthèse des lipides.

Un *réseau métabolique* regroupe l'ensemble des réactions chimiques impliquées dans un métabolisme, ainsi que les entités biologiques mises en jeu lors de ces réactions (figure 1.15).

On peut distinguer deux catégories d'entités, en fonction du rôle qu'elles jouent dans les réactions chimiques : les catalyseurs de réactions, et les molécules consommées et produites par les réactions. Les catalyseurs sont les

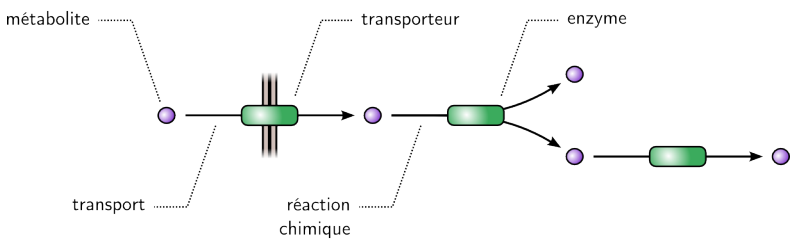


Figure 1.15. – Représentation d'un réseau métabolique.

entités biologiques qui possèdent une activité enzymatique, les enzymes. Les enzymes sont majoritairement des protéines, mais d'autres entités telles que des ARN non-codants et des complexes biologiques peuvent aussi catalyser des réactions. Les molécules consommées et produites sont majoritairement des métabolites, même si d'autres entités comme les protéines ou les ARN peuvent aussi être transformées par des réactions chimiques.

Les métabolites transportés par les processus de transport sont souvent les substrats et les produits des réactions chimiques. En conséquence, bien qu'un phénomène de transport ne soit pas à proprement parler une réaction chimique, les processus de transport et les entités qui y sont mobilisées sont souvent pris en compte dans les réseaux métaboliques.

La notion « étendue » d'un réseau métabolique correspond donc à : (*i*) un ensemble de processus de transport et de réactions chimiques, (*ii*) l'ensemble des métabolites et autres entités transportées et transformées par ces processus, et (*iii*) l'ensemble des entités qui permettent la réalisation de ces processus.

1.5.2. Réseau de régulation génétique

Mot à mot, un réseau de régulation est un ensemble d'entités qui sont liées entre elles par des relations de régulation. Comme nous l'avons vu, les phénomènes de régulations sont omniprésents dans la cellule : ils peuvent concerner presque tous les types d'entités, et ils utilisent une grande diversité de mécanismes.

Un réseau de régulation génétique, aussi appelé réseau de régulation transcriptionnelle, est un réseau de régulation qui se limite aux phénomènes de régulations de l'expression des gènes, et aux entités qui sont impliquées dans ces régulations (figure 1.16, page suivante).

La régulation de l'expression des gènes peut se faire à trois niveaux : au niveau de la transcription (régulations transcriptionnelles), au niveau de l'ARN transcrit (régulations post-transcriptionnelles), et au niveau de la traduction (régulation traductionnelle). En général, les régulations post-traductionnelles ne sont pas considérées dans les réseaux de régulation génétique.

Les entités considérées dans les réseaux de régulation génétique sont les gènes (ou les unités de transcription), les ARN et les protéines. Les ARN et les

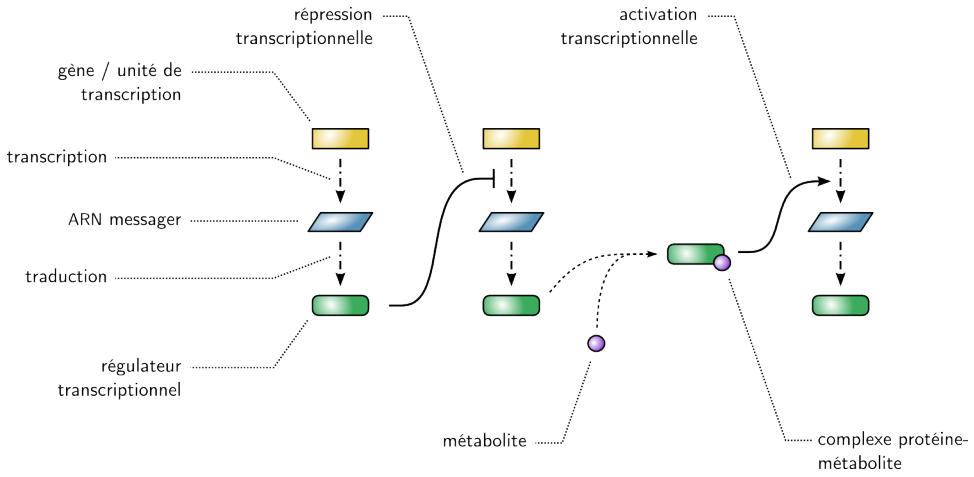


Figure 1.16. – Représentation d'un réseau de régulation génétique.

protéines sont à la fois les acteurs et les cibles (indirectes) des régulations. Par exemple, si l'on considère un gène x qui code une protéine X, et un régulateur Y capable de réprimer le gène x : la présence du répresseur Y va réprimer l'expression du gène x, et la diminution de l'expression du gène x va entrainer une baisse de la concentration en protéine X. On peut donc considérer que le régulateur Y — l'acteur — régule indirectement la concentration de la protéine X — la cible.

La régulation de l'expression des gènes peut aussi faire intervenir des métabolites. La formation d'un complexe entre le régulateur Y et un métabolite peut, par exemple, rendre inactif le régulateur Y et l'empêcher de réprimer l'expression du gène x. Le fait que certains métabolites peuvent avoir un impact sur les régulations génétiques sera abordé plus en détail dans la seconde partie « Modélisation simultanée des réseaux métaboliques et de régulation génétique, revue » (p. 50).

1.5.3. Autres réseaux biologiques

Mis à part les réseaux métaboliques et les réseaux de régulation génétique, deux autres types de réseaux biologiques sont souvent étudiés : les réseaux d'interaction protéines-protéines et les réseaux de signalisation.

Un réseau d'interaction protéines-protéines représente un ensemble de protéines qui interagissent entre elles par des liens physiques. Ce réseau peut être représenté par un graphe où les nœuds sont les protéines, et les arrêtes indiquent l'existence des liens physiques (figure 1.17, partie a).

Un réseau de signalisation correspond à un ensemble d'entités biologiques qui se transmettent un signal informatif. La signalisation cellulaire fait généralement intervenir des protéines qui sont transformées *en cascade* par des réactions chimiques telles que des phosphorylations. Dans un graphe d'un réseau de signalisation, les nœuds sont les entités biologiques transformées, et les arrêtes représentent la transmission de l'information (figure 1.17, partie b).

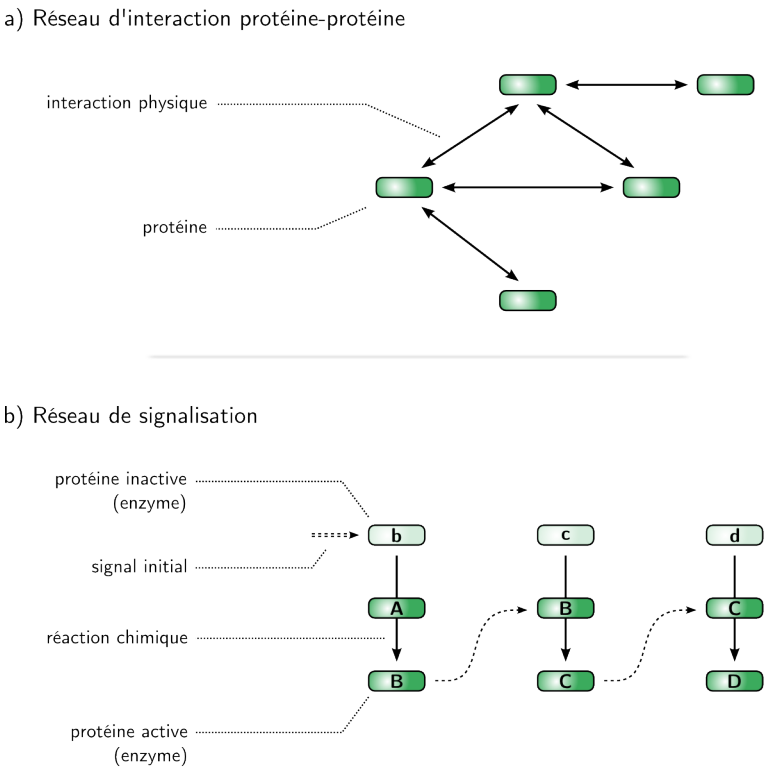


Figure 1.17. – **a)** Représentation d'un réseau d'interaction protéine-protéine et **b)** d'un réseau de signalisation.

Chapitre 2

Concept de modèle, usage en biologie

Dans le chapitre précédent, nous avons vu que de nombreuses molécules participent au fonctionnement des cellules. Au fur et à mesure des travaux scientifiques, les connaissances ont été accumulées sur les rôles et les mécanismes d'action de ces molécules. Aujourd'hui, la quantité de connaissances est telle que l'emploi de modèles formels — modèles basés sur les mathématiques et l'informatique — est nécessaire pour appréhender la complexité des cellules et de leur fonctionnement.

Dans ce chapitre, je vais en premier lieu parler de la notion de modèle formel, et de son usage en biologie. Le travail de modélisation nécessite une étape préalable de questionnement afin de définir les objectifs du modèle. Je discuterai de cette étape préliminaire et de ses enjeux dans une deuxième section.

2.1. Modèle formel, notion et usage en biologie

Un modèle est une description d'une chose réelle. Lorsque la *chose* modélisée est un système constitué de plusieurs éléments, le modèle de ce système consiste en une description des éléments et de leurs relations (figure 2.1, p. 41). Plus on a de connaissances sur la *chose* modélisée, et plus le modèle peut être détaillé. Ainsi, la représentation (le modèle) d'un ordinateur sera différente selon que l'on est un simple utilisateur ou un expert en circuit imprimé.

2.1.1. Notion de modèle

À l'heure actuelle, dans le domaine de la biologie, les termes de modèle et de modélisation font immédiatement penser aux modèles basés sur les mathématiques et l'informatique. Mais il existe d'autres types de modèles, bien plus répandus et plus utilisés. Ces modèles sont les schémas, les diagrammes, et d'autres représentations structurées des connaissances.

Ces représentations — comme la figure 2.1, partie b — sont des modèles, dans le sens où ils décrivent des choses réelles (un vieil ordinateur pour l'exemple).

Les schémas, et tous les modèles de manière globale partagent une caractéristique commune : la simplification et la généralisation de la réalité. Les schémas sont simplifiés pour une raison pratique qui est le manque de place sur le support pour faire figurer toutes les connaissances. Mais les modèles sont surtout simplifiés afin de rester lisibles et compréhensibles.

Les « modèles illustratifs » restent utiles : ils permettent de rassembler les connaissances et de faciliter leur interprétation. Cependant, ils ont une limite majeure : l'interprétation du fonctionnement de la *chose* décrite reste à la charge de l'observateur.

Par exemple, lorsque la *chose* décrite est un système biologique constitué de plusieurs centaines de protéines interagissant entre elles (figure 2.2), il est difficile, voire impossible, de comprendre à partir d'une simple représentation graphique comment telle ou telle protéine a un impact sur l'ensemble du système.

2.1.2. Modèles formels

Les modèles formels repoussent cette limite du passage à grande échelle. Le terme *formel* est employé, car ces modèles décrivent les relations entre les éléments d'un système au moyen d'expressions mathématiques, de relations logiques, de lois statistiques, ou encore d'instructions informatiques. Pour la suite de cette thèse, les termes *modèles formels* et *modèles* seront utilisés comme des synonymes.

Les modèles formels ont plusieurs intérêts : ils décrivent de manière précise et non ambiguë les constituants d'un système, ils permettent de simuler

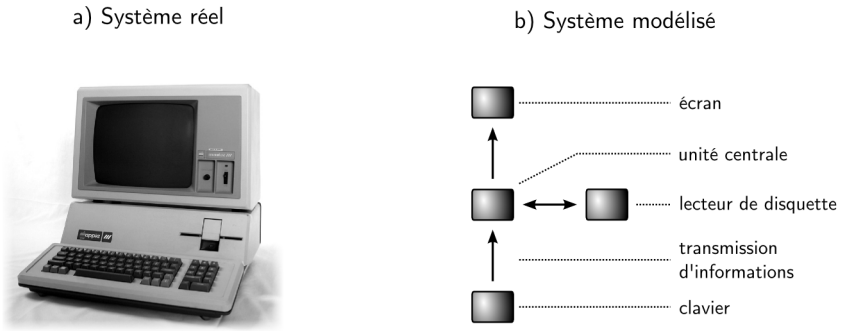


Figure 2.1. – Système réel et système modélisé.

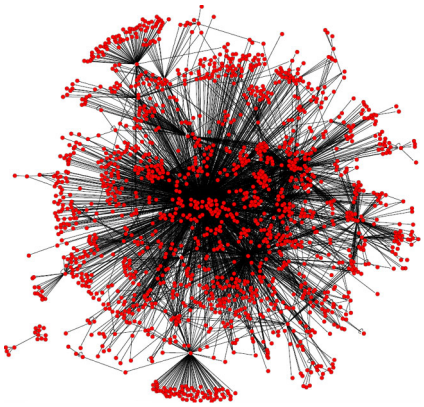


Figure 2.2. – Exemple du réseau de régulation génétique chez *Escherichia coli*. Chaque point rouge représente une protéine, chaque trait noir représente une relation de régulation. Source : © Nature Education.

et de prédire le fonctionnement global du système, et ainsi ils permettent d'étudier l'importance de chacun des constituants.

Chacun de ces points représente potentiellement une aide précieuse pour la biologie. La description formelle permet de lever les ambiguïtés que l'on rencontre dans des descriptions textuelles ou graphiques. Les simulations et les prédictions génèrent des résultats qui seraient difficilement obtenus par des moyens expérimentaux. Lorsque des résultats expérimentaux sont déjà disponibles, la comparaison entre les résultats prédits et ceux observés permet de préciser le modèle et d'améliorer les connaissances biologiques. Enfin, la possibilité d'étudier *in silico* les entités biologiques représente un gain de temps précieux. Par exemple, pour identifier l'effet de la suppression d'un gène, une approche expérimentale nécessite généralement plusieurs semaines de travail pour (espérer) observer un changement dans le phénotype de l'organisme. En comparaison, supprimer un gène dans un modèle ne prend que quelques minutes.

Les modèles sont utilisés pour étudier un vaste ensemble de phénomènes biologiques. Par exemple, les modèles formels sont des outils de choix pour prédire le repliement en 3 dimensions des protéines, ou pour comprendre le fonctionnement des réseaux de régulation. La figure 2.3 illustre des systèmes communément modélisés en fonction du niveau d'organisation biologique.

Selon le système modélisé, la nature des objets décrits change. Par exemple, la modélisation des structures protéiques s'intéresse aux propriétés physico-chimiques des acides aminés, tandis qu'un modèle de réseau de régulation génétique va prendre en compte les régulations entre des gènes, des ARN et des protéines.

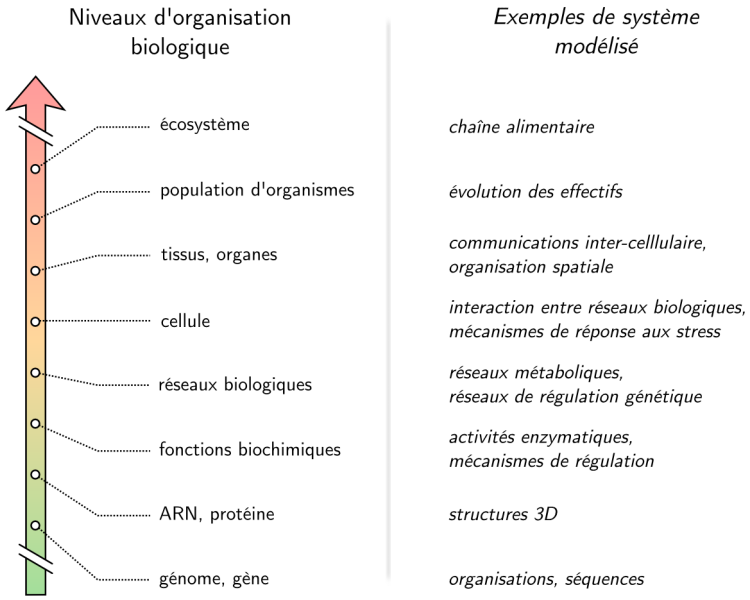


Figure 2.3. – Niveaux d'organisation cellulaire et exemples de systèmes biologiques modélisés.

2.2. Objectifs, simplifications et hypothèses d'un modèle

Les modèles sont des outils qui aident à mieux comprendre le fonctionnement des êtres vivants. En tant qu'outil, un modèle se doit d'être adapté à la problématique, à la question biologique posée.

La construction d'un modèle nécessite une première étape de questionnement : quel est l'objectif, quels sont les résultats que doit produire le modèle pour répondre à la problématique biologique ? Compte tenu de cet objectif, quelles connaissances doivent être intégrées pour modéliser le système ? Ces connaissances sont-elles disponibles ? Quel(s) formalisme(s) utiliser pour décrire le système ?

Pour cette partie, je vais prendre en exemple un système simple, celui de la réaction de transformation du glucose en glucose-6-phosphate catalysée par l'enzyme hexokinase. Comme le montre la figure 2.4, ce système comprend plusieurs métabolites, et une protéine possédant une activité enzymatique.

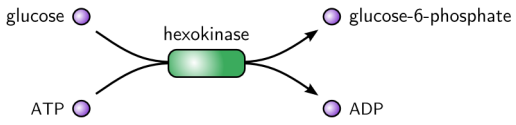


Figure 2.4. – Exemple d'un système biologique simple.

2.2.1. Problématique biologique

À partir d'un système donné, on peut étudier différents aspects biologiques. Pour la réaction catalysée, on peut par exemple s'intéresser au réarrangement des atomes lors de la transformation des substrats en produits, à la structure du site actif de l'enzyme, ou encore à la vitesse de la réaction. Selon la problématique, l'objectif du modèle change, et les résultats produits seront différents.

2.2.2. Simplifications et hypothèses

Afin de restreindre la complexité des modèles, c'est-à-dire le nombre d'informations prises en compte, seules les informations considérées comme essentielles sont décrites dans un modèle (table 2.1). En d'autres termes, les modèles font abstraction d'une partie de la réalité. Cette abstraction de la réalité porte à la fois sur la description des constituants du système, et sur la description des relations entre ces constituants. Par exemple, la structure de la protéine n'est pas considérée pour étudier la vitesse de la réaction, car elle n'y est pas indispensable. Il s'agit ici d'une simplification de la description du constituant « enzyme ».

L'abstraction de la réalité est basée sur des hypothèses simplificatrices. Par exemple, la vitesse d'une réaction dépend de nombreux facteurs tels que l'affinité de l'enzyme pour les substrats, le pH, la température, et les concentrations en métabolites et en enzyme. Pour simplifier le calcul de la vitesse, on pose généralement l'hypothèse que le pH et la température sont à des valeurs « standards » et que ces valeurs n'évoluent pas au cours de la dynamique.

Table 2.1. – Problématiques biologiques, modèles permettant d'y répondre, et connaissances nécessaires pour ces modèles.

Aspect biologique	Objectif du modèle	Connaissances essentielles
• Réarrangement des atomes des substrats	• simuler le devenir des atomes au cours de la transformation	• mécanisme de la transformation chimique • structure électronique des atomes
• Site actif de l'enzyme	• prédiction de la structure 3D de la protéine	• séquence d'acides aminés de la protéine • propriétés physico-chimiques des acides aminés • conformation 3D des métabolites
• Vitesse de la réaction	• simuler la dynamique de la réaction	• concentration en enzyme et en métabolites • paramètres cinétiques

La définition des hypothèses qui seront utilisées par le modèle représente un enjeu important. Idéalement, ces hypothèses simplifient grandement la réalité biologique, tout en permettant d'obtenir des prédictions suffisamment précises et pertinentes. Bien souvent, cette situation idéale ne peut pas être atteinte, et il s'agit alors de trouver un compromis entre simplification et précision.

2.2.3. Choix du formalisme

Nous avons vu qu'un système biologique réel peut être considéré selon différents aspects. Une fois que l'on a déterminé quel aspect biologique sera étudié au travers du modèle, viennent ensuite la question de l'objectif du modèle et des connaissances qui y seront mises en jeu. Un autre critère à considérer avant de commencer toute modélisation est le choix du formalisme qui sera utilisé pour décrire le système.

En modélisation, un formalisme est une manière de décrire les entités et leurs relations au sein d'un modèle. De nombreux formalismes différents existent pour décrire les systèmes biologiques (par ex. : équation différentielle, syntaxe logique).

Le choix d'un formalisme implique certaines simplifications et hypothèses qui sont inhérentes aux possibilités même du formalisme utilisé. Par exemple, un modèle booléen, modèle qui est décrit par un formalisme basé sur des variables booléennes, ne pourra pas prendre en compte précisément les concentrations des molécules : une molécule sera par exemple considérée comme absente (0) ou présente (1).

2.2.4. Confrontation des résultats avec la réalité

L'utilité des modèles repose sur la confrontation des résultats prédits avec des faits réels. Cette confrontation peut conduire à l'amélioration des connaissances biologiques (les prédictions sont cohérentes avec la réalité, la description du système semble correcte) ou à une révision du modèle (les prédictions ne sont pas cohérentes, une cause possible est la mauvaise description du système).

En conséquence, un autre point important à considérer est la possibilité de confronter les prédictions avec la réalité. En effet, si les moyens techniques et les connaissances ne permettent pas d'évaluer les prédictions d'un modèle, l'exploitation dudit modèle peut devenir difficile. Une alternative à l'absence (ou à l'insuffisance) de données expérimentales est la construction de modèles pour un usage théorique. Dans cette situation, la cohérence des résultats est évaluée sans confrontation avec les données existantes.

Le simple exercice d'identifier les connaissances nécessaires à la réalisation d'un modèle fournit des retours intéressants au modélisateur. Notamment, cela permet de définir le périmètre de la problématique étudiée, ainsi que de planifier des expériences en vue d'obtenir des connaissances manquantes ou de vérifier les prédictions du modèle.

Pour une même problématique biologique, il existe donc différentes manières de modéliser un système. Les constituants modélisés, les simplifications utilisées, les hypothèses posées, et les résultats générés par le modèle doivent être adaptés au contexte de l'étude.

Deuxième partie

Modélisation simultanée des réseaux métaboliques et de régulation génétique, revue

Dans cette partie, je reviens sur la problématique de mon sujet initial : le développement d'une approche pour modéliser le fonctionnement d'un système biologique constitué de (1), un réseau de régulation génétique (RRG) et de (2), un réseau métabolique (RM) et ce (3), en tenant compte des effets réciproques entre les deux réseaux.

Le terme d'*approche* est employé, car il ne s'agissait pas de construire un modèle chez un organisme en particulier, mais au contraire de mettre au point une méthode « générale » qui puisse être appliquée à différentes situations, chez différents organismes.

Quels sont les phénomènes biologiques qui lient les réseaux métabolique et de régulation génétique ? Quelles importances ont ces phénomènes dans le fonctionnement d'une cellule ? Quels sont les enjeux autour de la modélisation simultanée de ces deux réseaux ? J'aborde ces questions dans le chapitre 3 (p. 51).

Dans le contexte de la modélisation simultanée d'un RM et d'un RRG, *quelles sont les caractéristiques qu'il convient de considérer dans le choix de formalismes de modélisation ?* Je présente les caractéristiques qui m'ont semblé être importantes au fil du chapitre 4 (p. 59).

Un vaste choix de formalismes existe pour modéliser les relations entre entités biologiques. *Quelles sont ces méthodes ? Leurs caractéristiques ? Leurs avantages ? Leurs inconvénients ?* Dans le chapitre 5 (p. 65), je présente quelques-unes des méthodes de modélisation qui m'ont semblé être intéressantes pour la modélisation simultanée d'un RRG et d'un RM.

Chapitre 3

Biologie du couplage et enjeux de la modélisation simultanée

3.1. Couplage entre réseau métabolique et réseau de régulation génétique

Au sein des cellules, le métabolisme et la régulation génétique sont intimement liés. La régulation génétique a un impact sur le métabolisme en modifiant la quantité des enzymes et des transporteurs. En retour, le métabolisme a une influence sur les régulations génétiques *via* certains métabolites qui interviennent dans la (dés-)activation de régulateurs transcriptionnels.

Ce *couplage* entre réseau métabolique et réseau de régulation génétique a un rôle central au sein des cellules, à la fois parce qu'il fait intervenir un nombre considérable d'entités biologique, et parce qu'il participe à la coordination du fonctionnement des cellules.

3.1.1. Entités et relations mises en jeu

Le *couplage* entre réseau métabolique et réseau de régulation génétique est illustré par la figure 3.1. Le réseau de régulation modifie l'état du réseau métabolique *via* les processus de transcription et de traduction, et *via* les activités de régulation qui agissent sur ces processus. Le réseau métabolique modifie l'état du réseau de régulation *via* les concentrations de certains métabolites, qui possèdent une activité de régulation, qui agissent sur la capacité de régulation d'acteurs du réseau de régulation.

Au final, les entités et les relations mises en jeu dans le couplage sont toutes celles que j'ai présentées jusqu'alors. Les gènes sont transcrits en ARN. Les ARN messagers sont traduits en protéines. Les ARN non-codants, les protéines et les complexes peuvent catalyser des réactions chimiques, transporter des métabolites, ou réguler les processus biologiques. Enfin, les métabolites sont transformés et transportés par le métabolisme, et ils peuvent intervenir dans les régulations génétiques.

3.1.2. Importance du couplage, par le nombre d'entités

Les interactions entre les deux réseaux sont nombreuses au sein des cellules. La figure 3.2 (p. 55) donne, pour la bactérie *Escherichia coli*, quelques chiffres qui reflètent l'importance de ce couplage. Ces données sont issues de

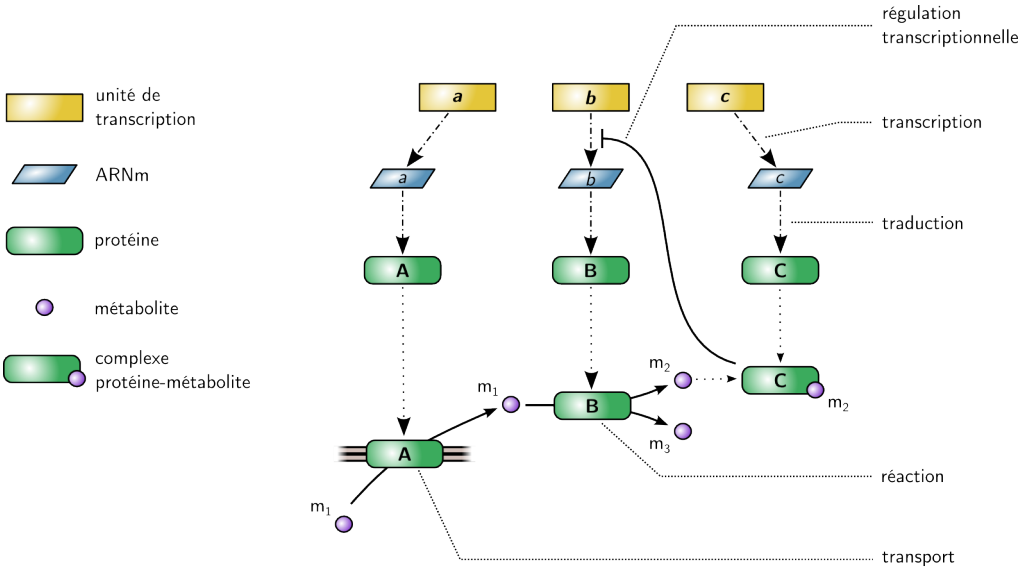


Figure 3.1. – Entités et relations mises en jeu dans le couplage entre réseau de régulation génétique et réseau métabolique. L'unité de transcription *a* code une protéine transporteur A qui permet l'entrée du métabolite m_1 dans la cellule. L'unité de transcription *b* code une enzyme B qui catalyse la transformation du métabolite m_1 en deux autres métabolites m_2 et m_3. L'unité de transcription *c* code un régulateur transcriptionnel C qui, lorsqu'il forme un complexe avec le métabolite m_2, réprime l'expression de l'unité de transcription *b*.

la base de données EcoCyc, [Keseler 2013], au mois de novembre 2014. Les requêtes utilisées pour obtenir ces chiffres sont disponibles dans l'annexe A (p. 243).

On recense 4 489 gènes qui codent une entité de type ARN ou protéine. Si l'on considère uniquement les régulations transcriptionnelles, 86 gènes codent un régulateur transcriptionnel pouvant former un complexe avec un métabolite (figure 3.2, partie c). Ces régulateurs transcriptionnels « pouvant former un complexe protéine-métabolite » régulent l'expression de 1 209 gènes, soit plus de 25% des gènes codants. Parmi ces gènes régulés, 742 codent une enzyme ou un transporteur, ce qui représente 41% du nombre total de gènes codant des acteurs du métabolisme.

Il est intéressant de noter que l'effet du métabolisme sur le réseau de régulation génétique se fait *via* un faible nombre de métabolites. Ainsi, chez *E. coli*, 1 413 métabolites sont impliqués dans des réactions chimiques ou des transports (catalysés par des enzymes et des transporteurs connus). Parmi ces métabolites, seuls 85 interviennent dans des phénomènes de régulation génétique en formant des complexes avec des régulateurs transcriptionnels. À l'inverse, le réseau de régulation génétique agit sur le métabolisme *via* la modification de l'expression d'un grand nombre de gènes. Par exemple, toujours chez *E. Coli*, 1 781 gènes codent des entités impliquées dans le métabolisme. L'expression de 1 017 de ces gènes peut être modifiée par des régulations transcriptionnelles.

3.1.3. Importance du couplage, par son rôle

Dans une cellule, toutes les régulations génétiques ne sont pas « effectives » en même temps. Les évènements de régulation, les gènes régulés, et l'intensité de ces régulations dépendent du milieu environnant et de l'état interne de la cellule. De la même manière, la vitesse des réactions chimiques et des transports de molécules évolue selon les conditions.

Dans ce contexte, le métabolisme joue un double rôle : il permet d'une part de produire l'énergie et les molécules nécessaires au fonctionnement des cellules, et d'autre part il contribue à « rapporter » l'état des environnements extra- et intracellulaires *via* l'activité régulatrice de métabolites. En réponse, la régulation de l'expression des gènes ajuste le fonctionnement de la cellule, notamment *via* la modification des concentrations en enzymes et transporteurs, afin de s'adapter au mieux aux nouvelles conditions de

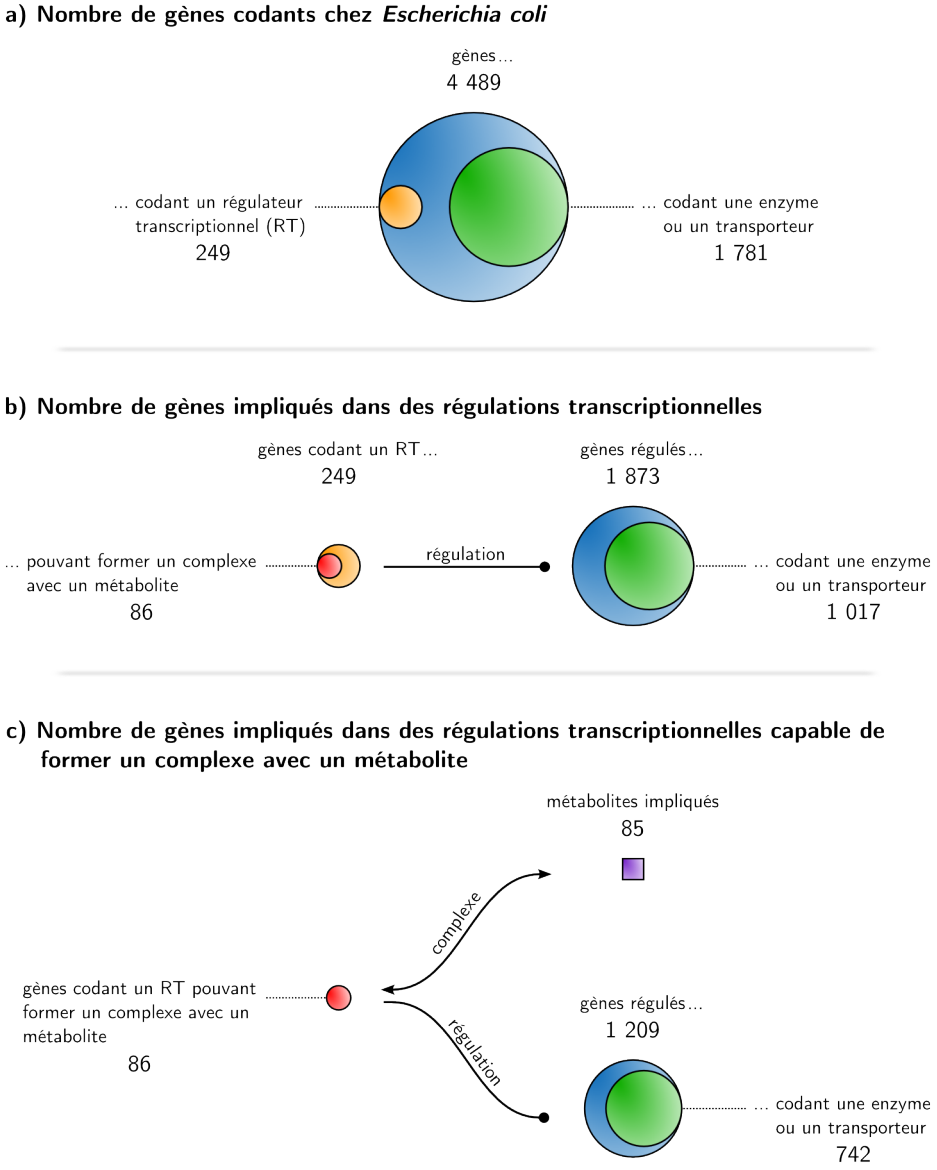


Figure 3.2. – Quelques chiffres qui reflètent l'importance du couplage entre réseau métabolique et réseau de régulation génétique chez *E. coli*. **a)** : 4 489 gènes codent une entité de type ARN ou protéine, dont 249 codent un régulateur transcriptionnel (RT) et 1 781 codent une enzyme ou un transporteur. **b)** : les RT régulent l'expression de 1 873 gènes, dont 1 017 codent une enzyme ou un transporteur. 86 gènes codent un RT qui peut former un complexe avec un métabolite. **c)** : ces 86 RT interagissent avec 85 métabolites, et régulent l'expression de 1 209 gènes, dont 742 codent une enzyme ou une protéine.

vie. Un exemple d'un tel circuit d'ajustements est donné par les travaux de Jacob et Monod sur l'opéron lactose [Jacob 1961]. L'article de revue récemment écrit par Chubukov et coll. présente d'autres exemples de couplage entre métabolisme et régulation génétique, et aussi entre le métabolisme et le réseau de signalisation [Chubukov 2014].

La compréhension de ces couplages — non plus à l'échelle de quelques réactions, mais à celle du métabolisme entier — est l'un des enjeux majeurs de la biologie des systèmes. Cependant, la complexité des phénomènes — le nombre d'acteurs qui interviennent dans ces réseaux — est telle que l'utilisation d'outils adaptés est nécessaire.

3.2. Enjeux autour de la modélisation simultanée

L'enjeu du développement d'approches de modélisation pour décrire le fonctionnement combiné d'un RRG et d'un RM est à la fois biologique et méthodologique.

3.2.1. Enjeu biologique

Il y a un enjeu biologique, car la modélisation intégrée de plusieurs réseaux biologiques permet d'avoir une vision plus globale des phénomènes qui régissent le comportement d'une cellule.

Si l'on ne considère qu'un seul réseau biologique, un modèle de ce réseau ne tiendra compte que d'un nombre limité d'entités biologiques et de relations. Par exemple, la modélisation du métabolisme prendra en compte les enzymes et les métabolites en tant qu'entités, et les réactions et les transports en tant que relations. Du fait du périmètre limité d'un tel modèle, il ne sera pas possible d'étudier la régulation de l'expression des gènes codant les enzymes, ou encore les mécanismes qui permettent à la cellule d'adapter son métabolisme selon les conditions environnementales.

Une vision intégrée d'un système vivant ouvre de nombreuses perspectives. Par exemple, il devient possible d'étudier l'impact de la modification d'une entité biologique sur le reste du système, à la fois au niveau du métabolisme et du réseau de régulation génétique. Par ailleurs, la modélisation du

couplage RM ↔ RRG permet d'identifier des boucles de régulation entre les deux réseaux.

Le besoin d'une vision systémique du fonctionnement des cellules est aujourd'hui accru par l'arrivée de nombreuses données « -omiques », produites par les techniques de transcriptomique, de protéomique ou encore de métabolomique. Ces données fournissent, à l'échelle de la cellule entière, une grande quantité d'informations. Par exemple, la métabolomique permet d'obtenir simultanément la concentration intracellulaire de nombreux métabolites [Patti 2012].

Mais les données « -omiques » ne fournissent pas d'information directe sur les mécanismes cellulaires, sur les causes et sur les conséquences des valeurs mesurées. En conséquence, l'interprétation de ces données nécessite l'utilisation de diverses connaissances préalablement acquises. Les modèles, en particulier les modèles multi-réseaux, représentent une aide précieuse pour l'exploitation des données « -omiques ». Par exemple, Colijn et coll. utilisent un modèle métabolique pour comprendre l'impact d'un profil d'expression du génome (données de transcriptomique) sur le fonctionnement du métabolisme [Colijn 2009].

De manière générale, le fait de prendre en compte l'état des deux réseaux permet aux modèles de générer des résultats plus pertinents. Des publications récentes montrent par ailleurs que d'autres phénomènes, notamment les régulations post-traductionnelles et le réseau de signalisation, jouent aussi un rôle important [Chubukov 2013].

3.2.2. Enjeu méthodologique

Le développement d'approches pour la modélisation simultanée de (i) l'effet des régulations génétiques sur le fonctionnement du métabolisme et de (ii) l'effet retour du métabolisme représente aussi des enjeux méthodologiques.

Les RRG et les RM peuvent être modélisés en utilisant différents formalismes [Machado 2011]. Certaines familles de formalismes peuvent servir pour modéliser l'un et l'autre des réseaux, comme les équations différentielles (modèles à bases d'équations différentielles). D'autres formalismes sont plus adaptés pour modéliser l'un des réseaux. C'est le cas du formalisme logique, adapté à la modélisation des réseaux de régulation (modèles

logiques), ou des matrices stœchiométriques pour la modélisation des réseaux métaboliques (modèles à base de contraintes).

Compte tenu des formalismes aujourd'hui disponibles, la modélisation simultanée des deux réseaux peut se faire selon deux approches. Une première approche est de construire un modèle *intégré*, qui utilise un même formalisme pour modéliser les deux réseaux. L'utilisation d'un même formalisme facilite la description des interactions entre les deux réseaux. La deuxième approche est de construire un modèle *couplé*, qui utilise un formalisme différent pour chaque réseau. Les formalismes étant différents, il est nécessaire de définir comment les deux modèles interagissent entre eux.

À l'heure actuelle, dans la majorité des travaux scientifiques, les modèles ne décrivent qu'un seul réseau biologique au moyen d'un unique formalisme.

Parmi les travaux qui mettent en jeux les deux réseaux, les méthodes développées portent principalement sur un couplage « unidirectionnel », où seul l'effet de l'un des réseaux sur l'autre est pris en compte : action du modèle métabolique sur le modèle de régulation OU action du modèle de régulation sur le modèle métabolique. Par exemple, Covert et coll. ont développé la méthode rFBA (*regulatory Flux Balance Analysis*), où chaque réseau est décrit au moyen d'un formalisme différent (modèle logique pour le réseau de régulation, modèle à base de contraintes pour le réseau métabolique [Covert 2003]). Grâce à cette méthode, les prédictions issues du modèle de réseau de régulation sont utilisées pour améliorer les prédictions sur le fonctionnement du métabolisme d'*E. coli*.

Enfin, certaines méthodes permettent une modélisation « bidirectionnelle » des deux réseaux (prenant en compte les effets réciproques). Simão et coll. ont ainsi développé une méthode basée sur un unique formalisme, les réseaux de Petri [Simão 2005]. Un autre exemple est celui de la méthode idFBA (*integrated dynamic Flux Balance Analysis*), développée par Lee et coll. [Lee 2008].

Dans le contexte de la biologie des systèmes, des méthodes permettant d'intégrer le métabolisme et la régulation génétique sont aujourd'hui souvent citées parmi les outils les plus en vue afin d'améliorer notre compréhension du fonctionnement des cellules. Un autre enjeu méthodologique important est l'intégration, la prise en compte des résultats expérimentaux dans des modèles multi-réseaux.

Chapitre 4

Caractéristiques pour le choix du formalisme

La difficulté de la modélisation simultanée de plusieurs réseaux biologiques est liée, d'une part, à la complexité de la réalité biologique, et d'autre part, aux contraintes imposées par chaque formalisme.

Les relations entre les entités changent selon le réseau biologique considéré. Un RRG véhicule principalement des flux d'information (les gènes contiennent l'information génétique, les régulateurs modulent l'expression de cette information), tandis qu'un RM fait intervenir des flux de matière (transport et transformation de métabolites).

D'autre part, comme l'expriment si bien Machado et coll. [Machado 2011],

> *Many of the proposed formalisms, such as Petri nets or process algebras, were originally created by the computational community for the specification of software systems, where the final system has to comply to the model. The biological community faces the opposite problem, where the model has to mimic the system's behavior, and where most components cannot even be measured directly.*

les formalismes utilisés n'ont pas été créés pour la biologie, mais sont issus d'autres domaines scientifiques. En conséquence, l'utilisation de ces formalismes doit s'adapter à la complexité et l'incertitude inhérentes à la biologie.

Dans cette section, je présente les caractéristiques — biologiques et méthodologiques — qu'il convient à mon sens de considérer lors de la conception d'une méthode de modélisation multi-réseaux.

4.0.1. Inférence *vs* simulation

On peut, grossièrement, différencier deux types d'utilisation des modèles : d'un côté les modèles inférentiels, de l'autre les modèles pour la simulation. Les modèles inférentiels utilisent des données expérimentales afin de déterminer, d'inférer, les relations entre les entités d'un système. En d'autres termes, l'objectif est de trouver une structure/topologie du système qui soit cohérente avec les données expérimentales [Segal 2003, Yeang 2006]. À l'inverse, dans les modèles pour la simulation, les relations entre les entités sont déjà définies. L'objectif est alors de prédire comment le système va évoluer étant donnée cette description du réseau. Dans ce type de modèle, les données sont utilisées pour construire le modèle et pour définir l'état initial du système au début d'une simulation. Par la suite, je ne m'intéresserai qu'aux modèles pour la simulation.

4.0.2. Constituants considérés dans le modèle

Idéalement, les entités et les relations considérées dans le modèle sont toutes celles présentées dans le premier chapitre de cette thèse. La figure 4.1 propose un résumé de ces informations. Le réseau de régulation génétique a un impact sur le métabolisme *via* la modification des quantités d'ARN non-codant et de protéines qui interviennent comme catalyseurs ou transporteurs dans le métabolisme. En retour, le réseau métabolique a un impact sur les régulations génétiques *via* la modification des concentrations des métabolites qui interviennent dans certaines régulations.

Nous verrons dans le chapitre suivant que, selon le formalisme utilisé, les entités et leurs caractéristiques ne sont pas considérées de la même manière. Un exemple, peut-être extrême, est la manière différente dont on considère une protéine :

— modèles à base d'équations différentielles : la concentration est considérée comme une variable continue ;

— modèles logiques : la concentration est considérée comme une variable discrète ;

— modèles à base de contraintes : la concentration n'est pas considérée (elle peut éventuellement être considérée de manière indirecte).

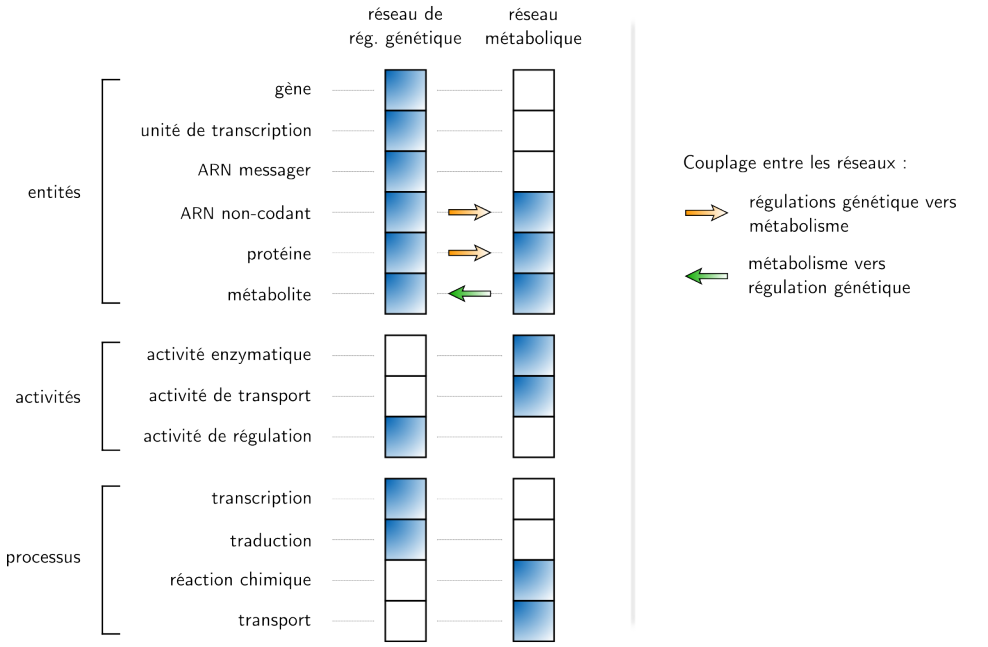


Figure 4.1. – Entités et relations idéalement prises en compte dans un modèle décrivant simultanément un réseau de régulation génétique et un réseau métabolique, et les interactions réciproques entre les deux réseaux. Une case bleutée indique que l'élément de la ligne intervient dans le réseau de la colonne. Les activités et les processus représentent les relations entre les entités. Par exemple, une réaction chimique relie une protéine qui possède une activité enzymatique (enzyme), et les métabolites qui sont consommés et produits par cette réaction.

4.0.3. Tailles des réseaux, connaissances disponibles

La taille des réseaux modélisés, en termes de nombre d'entités et de relations, est aussi un critère important dans le choix de l'approche : plus les réseaux sont grands, et plus l'interprétation des résultats générés par le modèle sera potentiellement complexe.

Il faut aussi tenir compte du fait que, de manière générale, plus le système considéré est grand, moins on dispose d'informations sur chaque constituant. Ceci est dû au fait que certaines portions des réseaux sont beaucoup plus étudiés que d'autres. En conséquence, dès lors que l'on souhaite intégrer des portions de réseaux « faiblement » étudiés, la quantité d'informations disponible est fortement réduite. Par exemple, si l'on considère le métabolisme global d'une cellule, la glycolyse est relativement bien caractérisée grâce à des décennies de travaux sur le sujet, alors que d'autres parties du métabolisme restent mal connues.

Par ailleurs, les connaissances sont souvent obtenues dans des conditions expérimentales standards (culture bactérienne sur substrat carboné unique, conditions optimales de croissance...). Si l'on souhaite étudier le fonctionnement d'un système dans des conditions différentes, certaines connaissances ne seront pas utilisables.

4.0.4. Paramétrage du modèle

Les paramètres sont des valeurs, le plus souvent numériques, qui peuvent être nécessaires pour décrire les relations entre les différents constituants d'un modèle.

Le nombre de paramètres nécessaire varie selon le formalisme utilisé. Ainsi, les modèles à base de contraintes (MBC) ne requièrent aucun paramètre pour décrire une réaction chimique, tandis qu'un modèle à bases d'équations différentielles nécessite au minimum un paramètre (si la cinétique suit la loi d'action de masse), et souvent plus.

Le paramétrage, autrement dit le fait de déterminer les valeurs des paramètres, est une phase importante dans la conception d'un modèle, car les paramètres décrivent souvent l'intensité des relations entre les éléments, ou les conditions nécessaires pour qu'une relation ait lieu. La dynamique du système dépendra en grande partie de la valeur de ces paramètres.

Selon le formalisme, la taille des réseaux modélisés, et les connaissances disponibles, le paramétrage d'un modèle est plus ou moins délicat. Il y a plusieurs raisons à cela. Tout d'abord, les paramètres et leurs valeurs ne représentent pas forcément une entité ou un phénomène biologique qui peut être mesurée directement expérimentalement. Par exemple, dans la fonction de Hill — $F = [x]^n/(K + [x]^n)$ — le nombre de Hill n, qui exprime le degré de coopérativité du ligand x, est difficile à déterminer. Deuxièmement, les valeurs des paramètres sont souvent estimées à partir de données expérimentales : la précision des valeurs estimées dépend donc de la quantité et de la qualité des mesures effectuées. Enfin, les mécanismes biologiques modélisés sont plus ou moins bien connus.

En conséquence, il est important de tenir compte du nombre et de la nature des paramètres requis par chaque formalisme. Par ailleurs, plus les réseaux modélisés seront grands (en termes de nombre d'entités et de relations), plus le nombre de paramètres sera potentiellement important.

4.0.5. Temps discret ou continu

Les cellules sont des systèmes dynamiques dont le fonctionnement est ajusté en permanence. En conséquence, la prise en compte du temps est un élément clé pour une meilleure compréhension du vivant. Selon le formalisme utilisé, le temps est considéré soit de façon discrète, soit de façon continue.

Lorsque le temps est considéré de façon discrète, comme dans les modèles logiques, l'écoulement du temps est représenté par la succession des interactions entre éléments qui provoquent un changement dans le système. L'utilisation d'un temps discret permet d'étudier l'ordre dans lequel les évènements (notamment les phénomènes de régulation) ont lieu. Cette conception du temps a l'avantage de ne pas nécessiter d'information sur la vitesse à laquelle ces évènements ont lieu (paramètres cinétiques). Elle est fréquemment utilisée pour étudier les réseaux de régulation.

Lorsque le temps est considéré de façon continue, comme dans les modèles à base d'équations différentielles, l'écoulement du temps est considéré de manière « conventionnelle » et est exprimé par exemple en secondes. Le temps continu permet de considérer la vitesse des évènements, comme la vitesse d'une réaction chimique. Le temps continu est en particulier utilisé pour étudier les réseaux métaboliques.

4.0.6. Résultats qualitatifs ou quantitatifs

Les modèles de simulation permettent de capturer la dynamique d'un système en rendant compte des effets des interactions entre les constituants, et de leur intensité.

Selon le formalisme, les effets des interactions et leur intensité peuvent être considérés soit de manière qualitative (par ex. : « production d'une *faible* quantité de protéines »), soit de manière quantitative (par ex. : « vitesse de production de 12 mmol/h de protéine »). Notons que des résultats quantitatifs peuvent parfois être analysés de manière qualitative (comme c'est le cas pour les équations différentielles linéaires par partie).

À taille de système modélisé égale, des modèles qui produisent des résultats quantitatifs permettent d'étudier plus précisément la dynamique du système, mais ces modèles nécessitent aussi plus d'informations.

4.0.7. Comportement déterministe ou stochastique

La simulation du fonctionnement d'un système implique de calculer, à partir d'un état donné, l'évolution du système au cours du temps.

Les modèles déterministes sont caractérisés par le fait que les conséquences des interactions entre les constituants sont totalement déterminées par la topologie et le paramétrage du modèle. Dans ce type de modèle, à partir d'un état donné du système, la simulation prédira toujours la même dynamique.

En opposition, les modèles stochastiques introduisent une part d'aléatoire dans la description des constituants et de leurs interactions. Cette dimension stochastique permet de prendre en compte les incertitudes liées aux connaissances et aux mesures expérimentales disponibles. Par ailleurs, notons que les phénomènes biologiques dans les cellules sont en partie soumis au hasard : par exemple, la vitesse d'une réaction chimique dépend de la probabilité de rencontre entre une enzyme et ses substrats.

Il est possible d'introduire une dimension stochastique dans la plupart des méthodes déterministes. Pour la suite de cette partie, je vais me concentrer sur l'aspect déterministe des méthodes de modélisation.

Chapitre 5

Formalismes pour la modélisation simultanée

Dans ce chapitre, je présente plusieurs formalismes qui sont communément utilisés pour modéliser des réseaux biologiques. Il s'agit des modèles logiques, des modèles à base d'équations différentielles ordinaires ou linéaires par partie, des réseaux de Petri, et des modèles à base de contraintes. La table 5.1 se propose de situer chaque méthode par rapport aux diverses caractéristiques présentées dans le chapitre précédent.

Il existe de (nombreux) autres formalismes qui ne seront pas abordés dans cette section. Divers articles de revue peuvent compléter les propos de cette section. Citons notamment Machado et coll. pour une vision générale sur l'usage des formalismes selon le type de réseau étudié [Machado 2011], ainsi que l'article écrit par H. de Jong et celui écrit par Karlebach et coll. pour

Table 5.1. – Caractéristiques de méthodes de modélisation. ML : modèle logique ; EDO : modèle à base d'équations différentielles ordinaires ; EDLP : modèles à base d'équations différentielles linéaires par partie ; MBC : modèles à base de contraintes. Les réseaux de Petri ne sont pas représentés, car les caractéristiques de ce formalisme dépendent de l'extension considérée.

	Formalisme			
Caractéristiques	ML	EDO	EDLP	MBC
Type de réseau	RRG	RRG et RM	RRG	RM[†]
Taille de réseau	grand	petit	petit	grand
Paramétrage	faible	élevé	élevé	très faible
Temps	discret	continu	continu	spécial[††]
Résultats générés	qualitatif	quantitatif	quantitatif	quantitatif
Comportement du modèle	déterministe	déterministe	déterministe	déterministe

[†] quelques modèles à base de contraintes ont aussi été utilisés pour étudier des RRG.
[††] le temps est continu, mais le système est considéré à l'état stationnaire, cf. sous-section dédiée.

des revues dédiées à la modélisation des réseaux de régulation génétique [de Jong 2002, Karlebach 2008].

5.1. Modèles logiques

Les modèles à base de variables logiques, ou modèles logiques ont été introduits par Kauffman [Kauffman 1969]. Dans ces modèles, les entités biologiques sont représentées par des variables qui ne peuvent prendre que deux valeurs : 0 (l'entité est absente/inactive), ou 1 (l'entité est présente/active). Le principe de cette méthode a ensuite été étendu par R. Thomas [Thomas 1991]. Dans cette nouvelle méthode, appelée méthode logique généralisée, chaque variable peut avoir plus de deux valeurs.

5.1.1. Caractéristiques et réseaux modélisés

La méthode logique généralisée est une méthode de modélisation déterministe, qualitative, et qui prend en compte le temps de façon discrète.

Cette méthode est principalement utilisée pour modéliser les réseaux de régulation génétique, et permet de prendre en compte les ARN non-codants, les protéines et les métabolites, ainsi que les relations de régulation entre ces entités.

5.1.2. Principe

Un RRG décrit par un modèle logique peut être représenté par un graphe orienté et pondéré (figure 5.1). Les nœuds représentent les entités biologiques, et les arcs représentent les actions de régulation entre les entités. Bien que les gènes ne soient pas directement considérés, la régulation de leur expression est modélisée en faisant varier le niveau de concentration des entités (ARN, protéines) codées par ces gènes.

Les concentrations réelles des entités sont discrétisées en valeurs logiques (figure 5.2). Cette simplification forte est basée sur le fait qu'un régulateur est souvent inefficace en dessous d'une concentration « seuil » et que son effet régulateur plafonne ensuite rapidement pour des concentrations

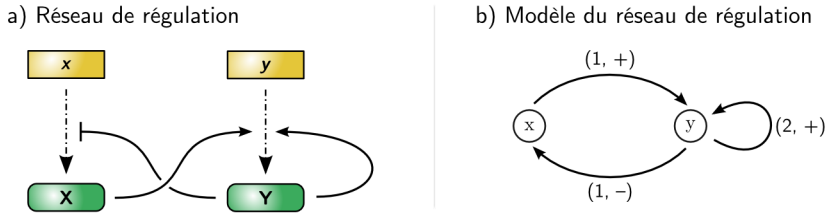


Figure 5.1. – Exemple d'un réseau de régulation génétique et de sa représentation sous la forme d'un graphe. Dans **a)**, les processus de transcription et de traduction des gènes sont combinés sous la forme d'une unique flèche en pointillé, et les relations de régulation sont représentées par des flèches en trait plein. Dans **b)**, les nœuds représentent les entités effectrices, ici les protéines régulatrices, et les arcs représentent les relations de régulation. Le chiffre associé à chaque arc indique la valeur seuil à partir de laquelle une entité est capable d'effectuer la régulation, et le signe (+ ou −) spécifie l'effet de cette régulation.

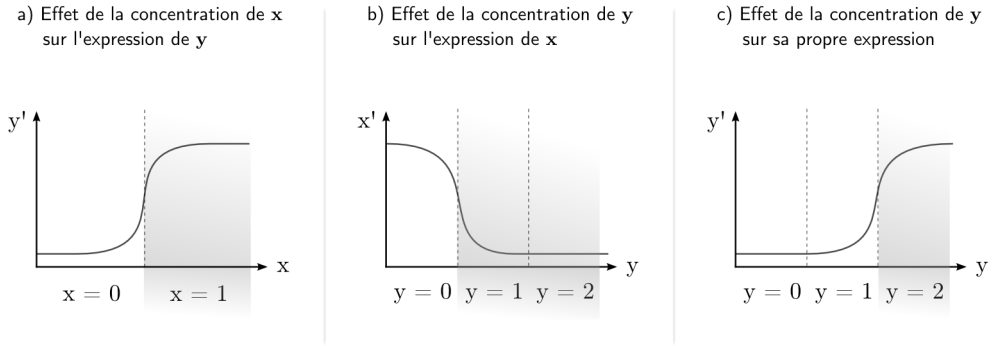


Figure 5.2. – Discrétisation de la concentration des entités effectrices de la figure précédente. **a)** La protéine x à deux valeurs logiques (0 et 1), car elle régule uniquement le niveau de concentration de la protéine y : si x a une concentration de « niveau 1 », alors x active l'expression de y. **b)** et **c)** La protéine y a trois valeurs logiques (0, 1 et 2), car elle régule le niveau d'expression de x et son propre niveau d'expression : l'expression de x est activée lorsque $y \geq 1$ et sa propre expression est activée lorsque $y = 2$.

supérieures. Chaque entité modélisée est ainsi décrite par une variable logique. Le nombre n de valeurs que peut prendre une variable est égal au nombre m d'actions de régulation de l'entité associée, auquel on ajoute 1 ($n = m + 1$).

La dynamique du modèle est exprimée par les transitions entre les états du système. Chaque état est caractérisé par un vecteur de taille v qui contient la valeur logique des v entités du système. La transition d'un état à un autre est réalisée par la modification d'une variable logique.

Les transitions sont calculées à l'aide de fonctions logiques (figure 5.3, parties a et b). Chaque variable est associée à une fonction logique. Pour une entité donnée, représentée par sa variable z, la fonction logique f_z détermine la valeur vers laquelle devra tendre z compte tenu de l'état des autres variables (de la concentration des autres entités) qui peuvent agir sur z. Les intensités des régulations entre les entités du réseau peuvent être spécifiées à l'aide de paramètres (paramètres logiques, introduits par E. Snoussi, [Snoussi 1989] ; figure 5.3, partie b). Finalement, en appliquant les fonctions logiques, on peut établir la table des états qui décrit pour chaque état possible vers quel nouvel état va tendre le système (figure 5.3, partie c).

Les évènements de régulation sont considérés de manière asynchrone. Autrement dit, à un instant donné on considère qu'un seul évènement de régulation peut être effectué. En conséquence, une seule variable logique est modifiée lors de la transition d'un état à un autre. La justification biologique de cette hypothèse se retrouve dans le fait qu'il est peu probable que deux entités passent leur seuil d'action exactement en même temps.

En exploitant l'hypothèse « une seule modification de variable entre chaque transition », et en attribuant des valeurs logiques aux paramètres, on peut déterminer l'ensemble de la dynamique du système modélisé (représentée sous la forme d'un graphe des états, figure 5.4).

a) Fonctions logiques

$X = f_x(y)$
$Y = f_y(x, y)$

b) Valeurs retournées et paramètres logiques

$f_x(y)$	y
$k_{x.y}$	0
0	1
0	2

$f_y(x,y)$	x	y
0	0	0
0	0	1
$k_{y.y}$	0	2
$k_{y.x}$	1	0
$k_{y.x}$	1	1
$k_{y.xy}$	1	2

c) Table des états

x	y	X	Y
0	0	$k_{x.y}$	0
0	1	0	0
0	2	0	$k_{y.y}$
1	0	$k_{x.y}$	$k_{y.x}$
1	1	0	$k_{y.x}$
1	2	0	$k_{y.xy}$

Figure 5.3. – Fonctions logiques et table des états d'un modèle logique **a)** Fonctions logiques associées à chaque entité du système. **b)** Valeurs retournées par les fonctions selon les valeurs actuelles des variables. Par exemple, si $y = 1$, le niveau de concentration de y est suffisant pour réprimer l'expression de x. En conséquence, le niveau de concentration x va tendre vers 0. Les valeurs retournées peuvent être spécifiées à l'aide de paramètres logiques (notés $k_{...}$). **c)** La table des états décrit, pour chaque état possible (colonne de gauche), vers quel nouvel état va tendre le système (colonne de droite). Par exemple, lorsque $x = 0$ et $y = 1$, le système va tendre vers un nouvel état où x reste à 0 (car l'expression de x est réprimée par la présence de y) et y passe à 0 (puisque x est absent, l'expression de y n'est pas activée).

a) Exemple de valeurs pour les paramètres

Param.	Valeurs
$k_{x.y}$	1
$k_{y.x}$	2
$k_{y.y}$	2
$k_{y.xy}$	2

b) Table des états correspondants

x	y	X	Y
$\overset{+}{0}$	0	1	0
0	$\overset{-}{1}$	0	0
$[0$	$2]$	0	2
1	$\overset{+}{0}$	1	2
$\overset{-}{1}$	$\overset{+}{1}$	0	2
$\overset{-}{1}$	2	0	2

c) Graphe des états correspondants

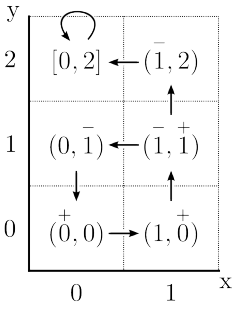


Figure 5.4. – Dynamique d'un modèle logique. **a)** Exemple de valeurs de paramètres. **b)** Table des états correspondants. Les variables dont la valeur doit changer sont indiquées par des signes + ou − (la concentration de l'entité doit augmenter ou diminuer, respectivement). L'état entre crochets représente un état stable, autrement dit, un état dont les évènements de régulation conduisent le système à rester tel qu'il est. **c)** Dynamique du système représenté sous la forme d'un graphe des états. Les variables logiques, qui correspondent aux concentrations discrétisées des entités, sont représentées en abscisse (entité x) et en ordonnée (entité y). Chaque combinaison de variables (x, y) représente un état du système. Les transitions entre les états sont représentées par des flèches. Du fait de l'hypothèse de la non-synchronicité des évènements de régulation, l'état (1,1) a deux transitions possibles.

5.1.3. Résultats générés

Les modèles logiques permettent d'étudier la dynamique des RRG. Cette dynamique est ici vue de manière discrète, sous la forme de la séquence des transitions d'états. À partir d'un état initial donné, on peut suivre l'évolution du RRG. En modifiant certains paramètres des fonctions logiques (ce qui correspond à modifier l'intensité des régulations), il est aussi possible de tester l'importance de chaque entité dans le réseau de régulation.

5.1.4. Avantages et inconvénients

L'aspect qualitatif est particulièrement bien adapté au réseau de régulation génétique, d'une part parce que les phénomènes de régulation présentent souvent des courbes de réponse en forme de sigmoïde, et d'autre part parce que les connaissances sur les effets des régulations sont souvent exprimées de façon binaire (« un gène s'exprime ou ne s'exprime pas »). Par ailleurs, la construction et le paramétrage de ces modèles nécessitent relativement peu de données (en comparaison avec d'autres méthodes de modélisation). Notons aussi que la prise en compte du temps discret ajoute un autre niveau d'abstraction, ce qui facilite l'utilisation de méthodes d'analyse formelle pour étudier les résultats issus d'un modèle logique (*Model-checking* et *Computation tree logic*, [Bernot 2004]).

Cependant, ces simplifications présentent aussi des inconvénients. En particulier, il est difficile d'utiliser les modèles logiques pour étudier les réseaux métaboliques. Une information indispensable lors de l'étude des réseaux métaboliques est la vitesse des réactions. Le fait de prendre en compte le temps de manière discrète empêche de calculer ces vitesses (une vitesse étant exprimée par unité de temps). De plus, la variation des vitesses des réactions n'est que rarement sujette à des effets de seuil, ce qui rend délicate l'utilisation de variables qualitatives pour décrire ces phénomènes.

5.1.5. Considérations sur la modélisation simultanée

L'utilisation du formalisme des modèles logiques pour modéliser à la fois un RRG et un RM (modélisation *intégrée*) semble peu adaptée, du fait de la difficulté à décrire les réseaux métaboliques.

L'autre alternative est de coupler un modèle logique d'un RRG avec une autre méthode décrivant le RM (modélisation *couplée*). Dans cette situation, il est nécessaire de déterminer comment les deux modèles vont interagir. La plupart des méthodes de modélisation du RM sont quantitatives et considèrent le temps de manière continue. En conséquence, la problématique du couplage entre un modèle logique et un modèle métabolique est double. D'une part, il s'agit de « traduire » les informations qualitatives en informations quantitatives, et inversement. D'autre part, il faut être en mesure de faire correspondre l'échelle de temps continue du modèle métabolique avec l'échelle discrète des modèles logiques.

5.2. Modèles à base d'équations différentielles

Les modèles à base d'équations différentielles ordinaires (EDO) sont probablement les modèles les plus utilisés pour la modélisation de systèmes dynamiques. Cela est notamment dû au fait que les équations différentielles offrent une grande liberté dans la manière de décrire les phénomènes biologiques.

5.2.1. Caractéristiques et réseaux modélisés

Les modèles EDO décrivent de manière quantitative les entités et les relations entre ces entités. Le temps y est pris en compte de manière continue. Ce formalisme permet de modéliser aussi bien les réseaux de régulation génétique que les réseaux métaboliques.

Dans cette partie, je parlerai aussi des équations différentielles linéaires par partie (EDLP), extension des équations différentielles ordinaires. Les EDLP exploitent certaines hypothèses supplémentaires qui limitent leur usage aux réseaux de régulation génétique.

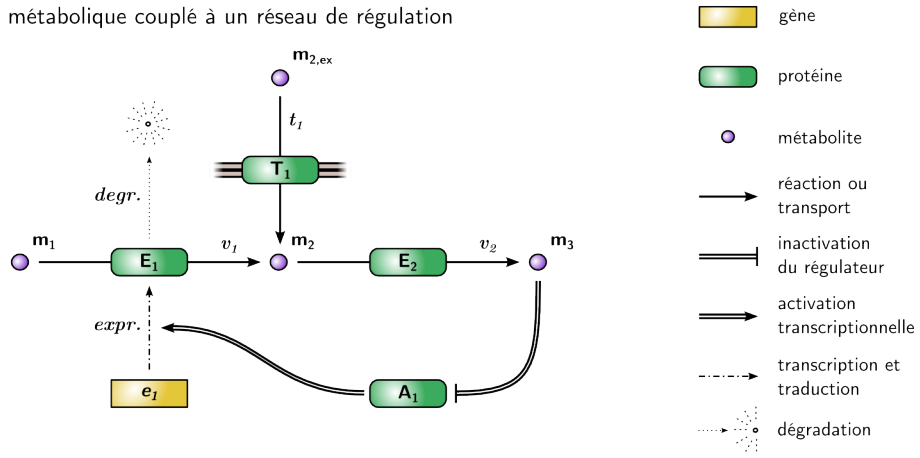
5.2.2. Principe

Les modèles EDO décrivent un réseau biologique à l'aide d'un ensemble d'équations. Chaque équation spécifie comment évolue la concentration d'une entité du réseau en fonction du temps, et en fonction de la concentration des autres entités du système.

La figure 5.5 présente un système biologique constitué d'un réseau métabolique couplé à un réseau de régulation (partie a), et le système d'équations associé (partie b). Chaque équation est composée d'une ou plusieurs fonctions, qui décrivent la vitesse des processus qui contribuent à l'augmentation ou à la diminution de la concentration des entités.

Dans les modèles EDO, la dynamique du système est représentée par les trajectoires de concentration de chaque entité au cours du temps. Pour cela, on pose l'hypothèse que, pour des intervalles de temps courts, la concentration des entités ne change pas. Grâce à cette hypothèse, il est possible

a) Réseau métabolique couplé à un réseau de régulation



b) Système d'équations différentielles

réaction v_1

$$\frac{dm_1}{dt} = -f_1(E_1, m_1)$$

transport t_1

$$\frac{dm_2}{dt} = +f_1(E_1, m_1) + g_1(T_1, m_{2,ex}) - f_2(E_2, m_2)$$

réaction v_2

$$\frac{dm_3}{dt} = +f_2(E_2, m_2)$$

transcription et traduction

dégradation

$$\frac{dE_1}{dt} = expr._{min} + expr._{max} \cdot h_1(A_1) \cdot h_2(m_3) - d_1(E_1)$$

Figure 5.5. – **a)** Exemple de système biologique constitué d'un RM couplé à un RRG. **b)** Système d'équations différentielles qui décrit le comportement du système. Les annotations indiquent la correspondance entre les termes des équations et les processus biologiques.

de calculer la variation des concentrations pour chaque intervalle de temps, d'actualiser la concentration « finale » de chaque entité, puis de continuer la simulation pour un nouvel intervalle.

L'usage des mathématiques offre une grande souplesse pour décrire les processus. Une fonction bien connue pour décrire les vitesses des réactions chimiques est la formule de Michaelis-Menten ($\frac{V_m S}{K_m + S}$). Elle met en relation la concentration de l'enzyme, celle du substrat, et la capacité de l'enzyme à catalyser la réaction *via* son activité enzymatique (figure 5.6, partie a). La formule de Hill ($\frac{R^n}{R^n + k_d}$; figure 5.6, partie c) est quant à elle couramment employée pour décrire les processus de régulation génétique. Elle permet d'obtenir une courbe de réponse en forme de sigmoïde, forme caractéristique des interactions génétiques.

Il existe de nombreuses autres fonctions. Il s'agit souvent de trouver un compromis entre le nombre de paramètres requis, la possibilité d'estimer la valeur de ces paramètres, et le réalisme de la fonction vis-à-vis du phénomène décrit.

5.2.3. Cas des équations linéaires par partie

Les équations différentielles linéaires par partie (EDLP) sont un type particulier d'équations. Elles sont utilisées pour modéliser les réseaux de régulation génétique [Gebert 2007].

Dans ces équations, toutes les fonctions d'échelon renvoient une valeur binaire, soit 0, soit 1 (figure 5.7, p. 77). Cette simplification permet de réduire le nombre de paramètres, tout en conservant l'essentiel de la courbe de réponse d'une régulation.

Par ailleurs, l'usage exclusif des fonctions d'échelon rend les trajectoires simulées linéaires pour chaque combinaison de valeurs retournées par les fonctions. Cela permet d'étudier la dynamique du système de manière analytique (étude de l'espace des phases [de Jong 2004]).

5.2.4. Avantages et inconvénients

La description mécanistique des équations différentielles — la variation d'une entité est fonction de la concentration d'autres entités — suppose peu de simplifications en comparaison avec d'autres méthodes. Par ailleurs,

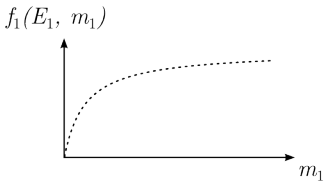
a) Fonction de Michaelis-Menten, réaction v_1

1. Formule

$$f_1(E_1, m_1) = \frac{k_{cat,1}.E_1.m_1}{k_{m,1} + m_1}$$

activité enzymatique de la protéine E_1

2. Courbe de réponse



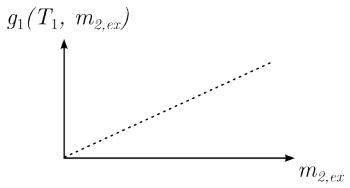
b) Fonction d'action de masse, transport t_1

1. Formule

$$g_1(T_1, m_{2,ex}) = k_{t,1}.T_1.m_{2,ex}$$

activité de transport de la protéine T_1

2. Courbe de réponse



c) Fonction de Hill, régulation de l'expression de E_1 par A_1

1. Formule

$$h_1(A_1) = \frac{A_1^i}{A_1^i + \theta_{A,1}}$$

activité de régulation de la protéine A_1

2. Courbe de réponse

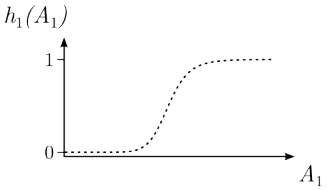


Figure 5.6. – Exemple de fonctions et de courbes de réponse qui décrivent les processus biologiques, selon la concentration en métabolite. **a)** Fonction de Michaelis-Menten (réaction v_1). **b)** Fonction d'action de masse (transport t_1). **c)** Fonction de Hill (régulation de l'expression de E_1 par A_1). Les annotations indiquent la correspondance entre les termes des équations et les activités biochimiques.

les équations différentielles offrent une grande souplesse dans la manière de décrire les phénomènes biologiques. Un autre point fort de la méthode est le fait de prendre en compte la concentration des entités (enzymes, régulateurs, métabolites...), chose qui n'est pas possible dans les modèles à base de contraintes (méthode abordée plus loin).

Cependant, les modèles EDO ont aussi des inconvénients. Premièrement, le besoin de déterminer la valeur de nombreux paramètres. Même s'il existe aujourd'hui de nombreux outils qui aident dans cette tâche, le paramétrage des modèles EDO reste souvent l'étape limitante de ce type de modèle [Berthoumieux 2013]. Deuxièmement, il est difficile de comparer l'ensemble des trajectoires simulées des concentrations avec des résultats expérimentaux. En effet, les modèles EDO génèrent des résultats très détaillés (calcul de la concentration des entités toutes les secondes, parfois plus fréquemment), alors que les techniques expérimentales ne permettent pas encore une telle fréquence d'échantillonnage. La construction d'un modèle EDO demande donc beaucoup d'efforts, tandis qu'on ne peut exploiter qu'une partie des résultats générés par le modèle.

5.2.5. Considérations sur la modélisation simultanée

La possibilité de décrire à la fois les phénomènes métaboliques et les régulations génétiques permet de modéliser avec le même formalisme un RM et un RRG [Usuda 2010]. Cependant, la quantité de données nécessaires pour paramétrer les modèles EDO limite son usage à des systèmes biologiques de petite taille et très bien étudiés.

Une manière de réduire la quantité de paramètres nécessaires est d'utiliser des fonctions simples pour modéliser le réseau de régulation génétique, telles que les fonctions d'échelon. À noter que dès lors que ces fonctions sont utilisées en conjonction de fonctions continues (comme l'équation de Michaelis-Menten), le système ne peut plus être considéré comme linéaire par partie.

a) Équation linéaire par partie

$$\frac{dE_1}{dt} = expr._{min} + expr._{max} \cdot h_x(A_1, \theta_{A,1}) \cdot (1 - h_x(m_3, \theta_{m,3})) - \gamma_1 . E_1$$

b) Fonction d'échelon : régulation de l'expression de E_1 par A_1

1. Formule

$$h_x(A_1, \theta_{A,1}) = \begin{cases} 1, & A_1 \geq \theta_{A,1} \\ 0, & A_1 < \theta_{A,1} \end{cases}$$

2. Courbe de réponse

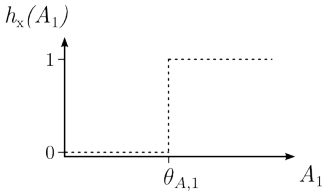


Figure 5.7. – Équation linéaire par partie et fonction d'échelon. **a)** Équation linéaire par partie décrivant le réseau de régulation de la figure 5.5. **b)** Fonction d'échelon, formule et courbe de réponse.

5.3. Réseaux de Petri

Le formalisme des réseaux de Petri (RdP) a été inventé par C. Petri pendant sa thèse [Petri 1962]. À l'origine dédiés à l'étude des flux d'informations dans un système à évènements discrets, les RdP représentent aujourd'hui une large famille de méthodes.

5.3.1. Caractéristiques et réseaux modélisés

Les RdP et leurs extensions permettent de considérer des variables qualitatives et quantitatives (RdP hybrides), avec une prise en compte du temps discrète ou continue (RdP temporels), et un comportement déterministe ou stochastique (RdP stochastique). Cette polyvalence permet aux RdP de modéliser aussi bien les RRG que les RM.

Dans cette partie, je ne présenterai que les principes essentiels des RdP qualitatifs (variables et temps discrets). Je n'aborderai pas les nombreuses extensions existantes (pour une revue dédiée aux RdP, voir [Chaouiya 2007]).

5.3.2. Principe

Un réseau de Petri, ou réseau place/transition, peut être représenté sous la forme d'un graphe bipartite orienté. Les quatre éléments principaux des réseaux Petri sont les places, les transitions, les arcs et les jetons (figure 5.8).

Les *places*, premier type de nœud, représentent des conditions. Les places peuvent contenir des *jetons*. Les jetons représentent les variables qui seront évaluées dans les conditions. Ils sont consommés et produits par les transitions. On parle de marquage d'un réseau de Petri pour désigner la répartition des jetons au sein du réseau. Les *transitions*, deuxième type de nœud, représentent les évènements du réseau. La réalisation (le tir) d'une transition nécessite que toutes les conditions associées soient vérifiées, c'est-à-dire que le nombre de jetons dans chaque place (situées en amont de la transition) est suffisant. Le nombre de jetons suffisant est spécifié par la valeur des *arcs*. Les arcs sont orientés et valués. Ils vont soit d'une place à une transition, soit d'une transition vers une place. La valeur de chaque arc

indique le nombre de jetons consommés (place → transition) ou produits (transition → place) lors de la réalisation d'une transition.

Dans un contexte biologique, les places correspondent souvent aux entités du système (métabolites, protéines...), et les jetons expriment la concentration ou l'état de ces entités. La combinaison des arcs et des transitions représente les relations entre ces entités (processus biologiques, et activités de régulation).

La dynamique d'un réseau de Petri est exprimée par la succession des transitions. À chaque réalisation d'une transition, des jetons sont produits et consommés, et le marquage du réseau change (figure 5.9, p. 81).

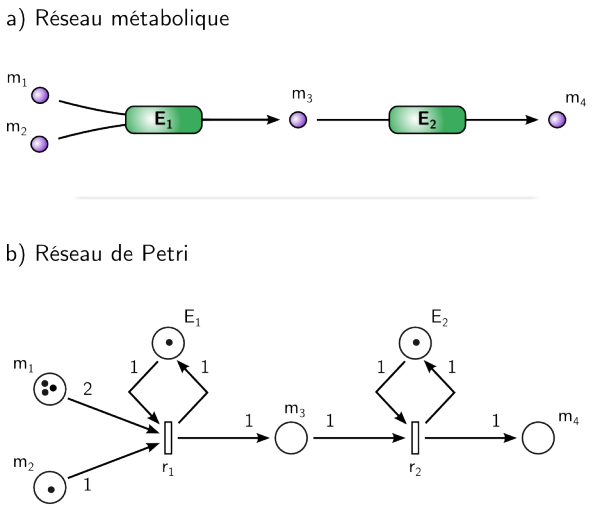


Figure 5.8. – **a)** Exemple d'un réseau métabolique et **b)** de sa représentation par un réseau de Petri. Les places et les transitions sont respectivement représentées par des cercles et des rectangles. Un exemple de marquage du réseau est donné, où les jetons sont représentés par des points noirs. La valeur associée à chaque arc indique le nombre de jetons consommés ou produits par la réalisation des transitions.

5.3.3. Résultats générés

Les RdP permettent d'analyser la dynamique du réseau de manière qualitative, en analysant l'ensemble des séquences de transition possibles. Notons ici que la dynamique d'un RdP est asynchrone : lorsque, à partir d'un marquage donné, plusieurs transitions sont possibles, les transitions sont effectuées séquentiellement.

Le type de résultat généré dépend grandement des extensions de RdP utilisées. L'emploi de variables quantitatives pour les places et de fonctions pour les transitions (vitesse de transition en fonction du contenu des places ; RdP hybride fonctionnel) permet par ailleurs d'étudier l'évolution de l'état du réseau en s'intéressant aux concentrations réelles des entités.

5.3.4. Considérations sur la modélisation simultanée

Le formalisme des RdP permet de décrire tous les types de réseaux biologiques. Cette polyvalence fait de ce formalisme un bon candidat pour modéliser simultanément un RM et RRG, voire aussi un réseau de signalisation.

Une approche possible est par exemple d'utiliser un RdP à variables discrètes pour modéliser le réseau de régulation génétique, et de le combiner avec un RdP fonctionnel pour modéliser le réseau métabolique. Cette dernière extension utilise des fonctions mathématiques pour décrire les transitions. Une limite potentielle de cette approche est le paramétrage des fonctions lorsque le réseau modélisé est grand (de façon similaire aux modèles EDO).

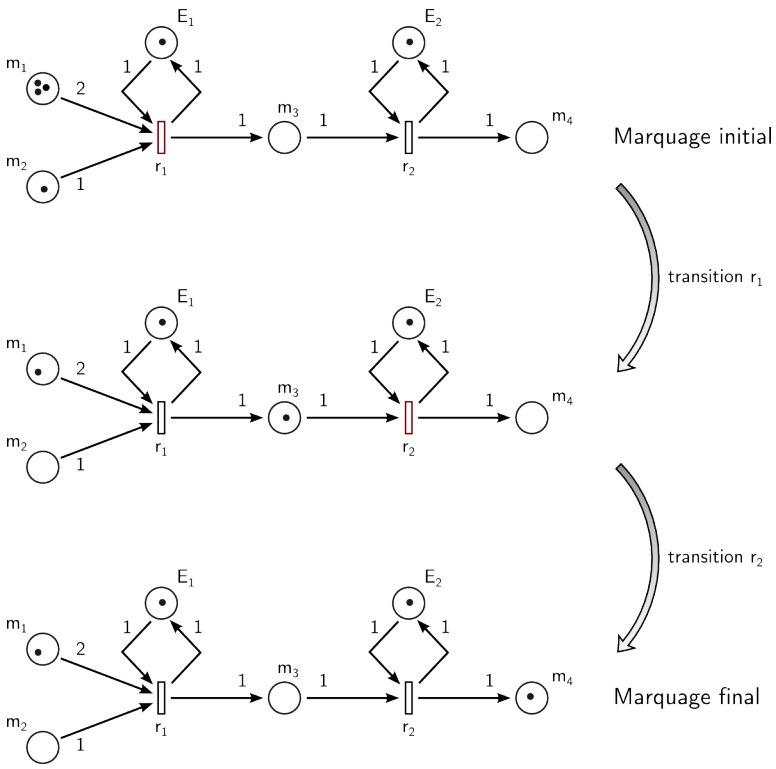


Figure 5.9. – Dynamique d'un réseau de Petri, exprimée par la séquence des transitions.

5.4. Modèles à base de contraintes

Dans un modèle à base de contraintes (MBC), les relations entre les entités sont décrites par une matrice stœchiométrique. Dans ce type de modèle, on s'intéresse généralement aux flux de matière qui traversent le réseau modélisé.

5.4.1. Caractéristiques et réseaux modélisés

La modélisation à base de contraintes est une méthode qui ne nécessite pas de connaître les paramètres cinétiques des réactions pour décrire les relations entre les entités du système.

Les MBC sont largement utilisés pour modéliser les réseaux métaboliques. L'absence de paramètre permet de modéliser des réseaux de grande taille, allant jusqu'à une échelle « cellule complète » (description de l'ensemble des processus métaboliques qui peuvent avoir lieu au sein d'une cellule).

Dans ces modèles, les flux de métabolites sont représentés par la vitesse des réactions chimiques et des transports. Ces vitesses sont exprimées à l'aide de variables quantitatives. Quelques MBC ont aussi été réalisés pour décrire des réseaux de régulation génétique (par exemple [Gianchandani 2009]), mais cette utilisation reste rare. Je ne vais considérer ici que la modélisation des réseaux métaboliques.

Contrairement aux autres méthodes de modélisation présentées jusqu'ici, les MBC ne permettent pas de simuler la dynamique du système, car ils supposent que le système modélisé est à l'état stationnaire. L'intégration de la dynamique du système dans un MBC fait partie du cœur de mon projet de thèse. Les méthodes actuellement développées sur le sujet seront présentées plus loin dans une partie dédiée (chapitre 6, p. 99).

5.4.2. Principe

La figure 5.10 présente la manière dont on considère un réseau métabolique dans un MBC. Dans cette représentation, le système métabolique est constitué d'un ensemble de métabolites, reliés entre eux par des flux qui correspondent aux réactions chimiques. Les MBC introduisent la notion de

limite du système, qui sépare le système modélisé — constitué des métabolites et des flux internes — de l'environnement extérieur. Le système est relié à l'extérieur par des flux d'échange qui correspondent classiquement à des processus de transport.

Les MBC supposent que le réseau métabolique est à l'état stationnaire, autrement dit que les concentrations des métabolites internes ne varient pas au cours du temps. Cette hypothèse est basée sur le fait que, à l'échelle d'une cellule entière, la dynamique des réactions intracellulaires est beaucoup plus rapide que l'évolution des conditions environnementales. En conséquence, les MBC ne tiennent pas compte des variations transitoires des concentrations en métabolites, et considèrent que ces concentrations atteignent rapidement une valeur stable.

Soulignons ici que l'hypothèse d'état stationnaire n'implique pas des vitesses de flux nulles, puisque le modèle tient compte des flux d'échange avec l'*extérieur* du système. Comme l'illustre la figure 5.11 (page suivante), l'état stationnaire représente en fait un équilibre dynamique : pour chaque métabolite interne, la somme des flux qui consomment ce métabolite est

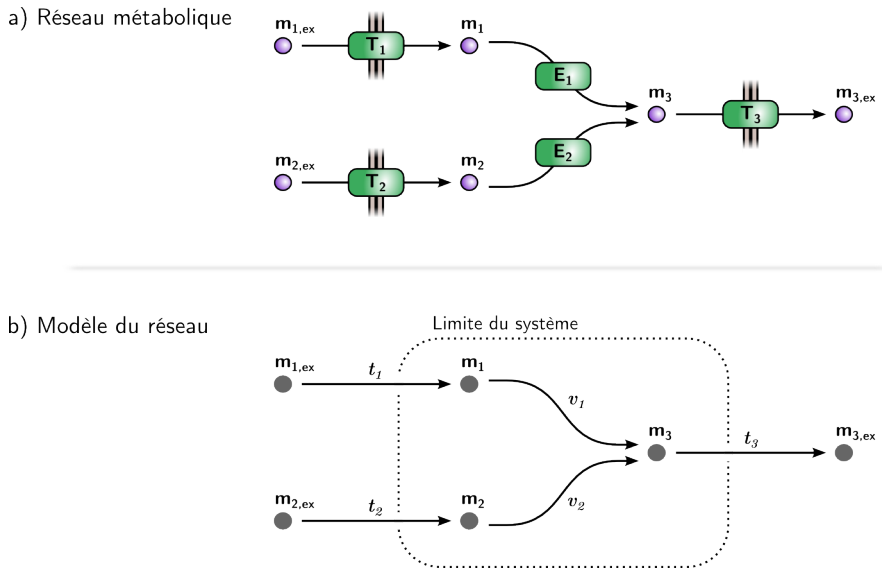


Figure 5.10. – Représentation d'un réseau métabolique avec un modèle à base de contraintes. La limite du système est représentée par des pointillés. Les métabolites m_1, m_2 et m_3 sont des métabolites internes, tandis que $m_{1,ex}$, $m_{2,ex}$ et $m_{3,ex}$ sont considéré comme des métabolites externes au système. Les flux v_1 et v_2 sont des flux internes. t_1, t_2 et t_3 sont des flux d'échange entre le système et l'environnement extérieur.

égale à la somme des flux qui le produisent.

Sous l'hypothèse de l'état stationnaire, les relations entre flux peuvent être décrites par un système d'équations linéaires (figure 5.12). Ces équations sont classiquement réécrites sous forme matricielle :

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

où la matrice stœchiométrique S donne pour chaque flux la stœchiométrie des métabolites consommés et produits, v représente la vitesse de chaque flux, et l'égalité à 0 impose l'équilibre entre les flux produisant et ceux consommant chaque métabolite interne.

En l'état, il existe une infinité de vecteurs v solutions de l'équation 5.1 (infinité de valeurs possibles de flux). Ces solutions sont considérées dans un espace des solutions, où chaque point représente un vecteur v qui satisfait l'équation (figure 5.13). On peut ici relier l'existence de plusieurs solutions à la réalité biologique. En effet, le métabolisme des cellules peut avoir des comportements différents (des régimes de fonctionnement différents) en fonction de l'environnement, par exemple en fonction de la disponibilité des substrats.

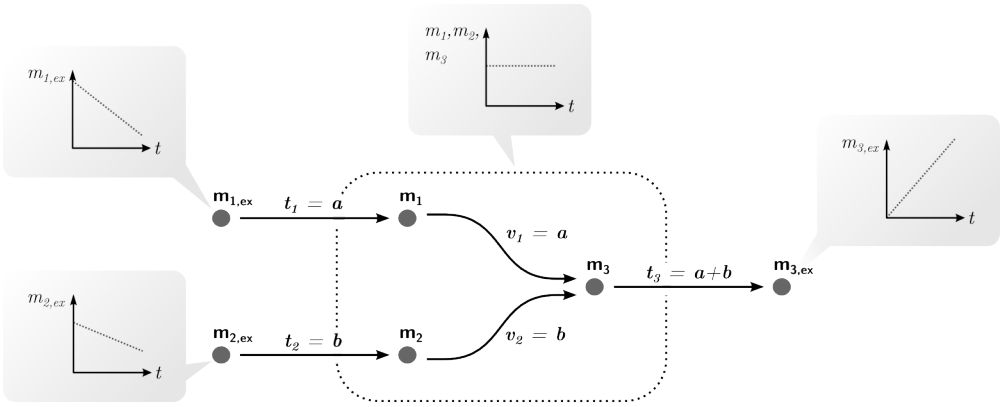


Figure 5.11. – Exemple d'équilibre dynamique à l'état stationnaire. Les métabolites $m_{1,ex}$ et $m_{2,ex}$ sont respectivement consommés *via* les flux d'échange t_1 (à une vitesse a) et t_2 (à une vitesse b). Cet apport en m_1 et m_2 est équilibré par les flux v_1 et v_2 qui produisent m_3 : on a les vitesses $t_1 = v_1 = a$ et $t_2 = v_2 = b$, les concentrations de m_1 et m_2 restent donc stables. L'apport en m_3 est équilibré par la sortie de $m_{3,ex}$ à l'extérieur du système : on a $t_3 = v_1 + v_2 = a + b$, la concentration de m_3 reste stable.

a) Système d'équations linéaires

$$\begin{aligned}\frac{dm_1}{dt} &= +t_1 - v_1 &&= 0\\ \frac{dm_2}{dt} &= +t_2 - v_2 &&= 0\\ \frac{dm_3}{dt} &= +v_1 + v_2 - t_3 &&= 0\end{aligned}$$

b) Système d'équations sous forme matricielle

S : matrice de stœchiométrie — v : vecteur des valeurs de flux

$$\begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \end{array} \begin{array}{c} \begin{array}{ccccc} v_1 & v_2 & t_1 & t_2 & t_3 \end{array} \\ \begin{bmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & -1 \end{bmatrix} \end{array} \cdot \begin{bmatrix} v_1 \\ v_2 \\ t_1 \\ t_2 \\ t_3 \end{bmatrix} = 0$$

Figure 5.12. – Description mathématique du modèle.

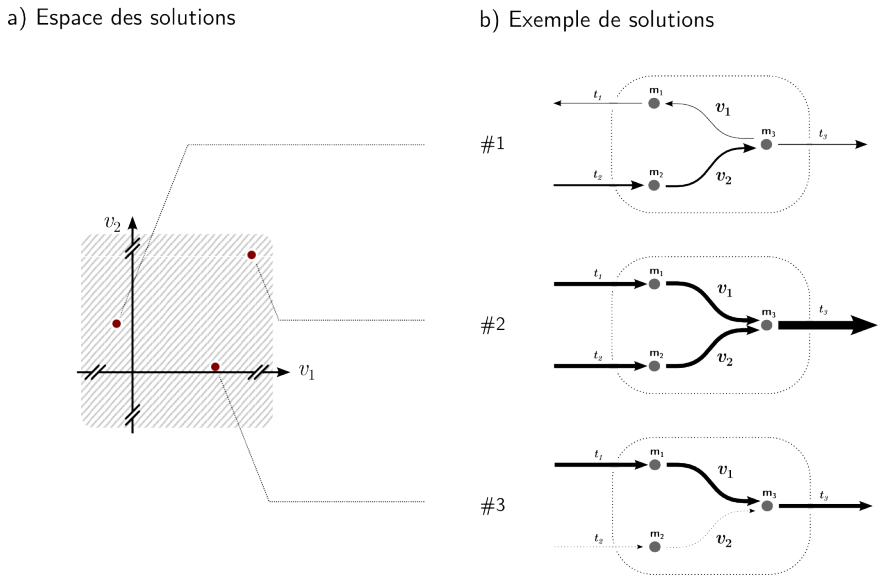


Figure 5.13. – **a)** Espace des solutions d'un modèle à base de contraintes. Chaque point dans l'espace représente une combinaison de valeur des flux v_1 et v_2 solution du problème. Sans imposition de contraintes, l'espace est infini. **b)** Exemples de solutions. L'épaisseur des flèches est proportionnelle à la vitesse des flux.

L'utilisation de contraintes sur la vitesse des flux permet de réduire la taille de l'espace des solutions, voire de déterminer un unique vecteur v solution. Deux types de contraintes peuvent être utilisés :

— Contraintes d'inégalité.

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

Un flux peut être contraint par une limite inférieure et/ou une limite supérieure, qui définissent les valeurs minimale et maximale possibles du flux. Ces contraintes peuvent être utilisées lorsque des connaissances sont disponibles sur la capacité du système. Par exemple, le fait qu'une réaction soit irréversible peut être décrit en fixant à 0 la valeur minimale du flux associé.

— Contraintes d'égalité.

$$v_{min} = v_{const}$$

Un flux peut être contraint à une unique valeur. Ce type de contrainte est typiquement utilisé lorsque certaines réactions du système sont connues comme étant inactives dans les conditions de l'étude ($v_{const} = 0$).

L'application de ces contraintes permet de réduire l'espace des solutions initialement défini par l'équation 5.1 (figure 5.14). L'espace des solutions contraint peut être formulé de la manière suivante :

$$S \cdot v = 0$$
$$\text{sujet à} \quad \begin{cases} v_{min} \leq v \leq v_{max} \\ v = v_{const} \end{cases}$$

5.4.3. Réaction de production de biomasse

La croissance bactérienne peut être modélisée en introduisant dans un MBC une *pseudo-réaction* qui décrit les métabolites nécessaires (considérés comme des substrats de la réaction) pour produire une nouvelle cellule (produit de la réaction). Cette pseudo-rédaction de production de biomasse est considéré comme un flux d'échange qui va de l'intérieur vers l'extérieur du système.

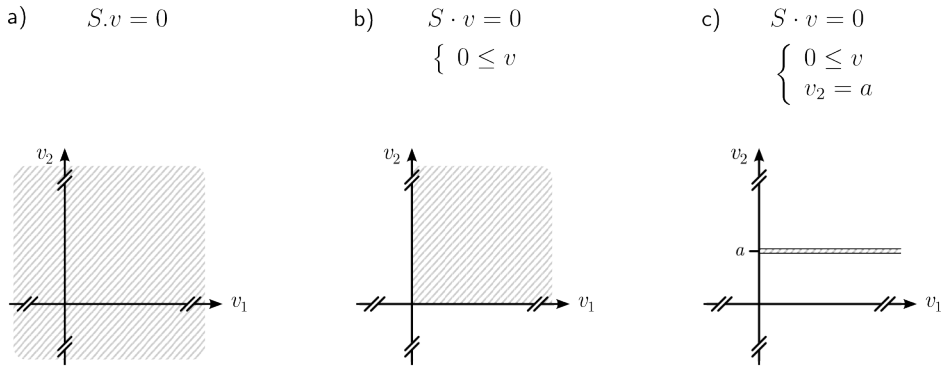


Figure 5.14. – Espace des solutions **a)** sans contraintes, **b)** avec des contraintes d'inégalité, et **c)** avec des contraintes d'inégalité et une contrainte d'égalité.

On peut déterminer expérimentalement la composition en macromolécules d'une cellule (les protéines, les lipides...). Si l'on connait suffisamment le métabolisme de la bactérie, il est aussi possible de déterminer les métabolites précurseurs nécessaires pour produire chacune de ces macromolécules. La table 5.2 donne un exemple d'une telle « conversion » pour les acides aminés qui constituent les protéines d'*E. coli*. En tenant compte des proportions respectives de chaque constituant, on peut alors exprimer la production de biomasse en termes de métabolites précurseurs et d'énergie nécessaires pour produire un gramme de biomasse. La table 5.3 donne un exemple pour *E. coli*. Ces informations sont ensuite réécrites sous la forme d'une pesudo-réaction qui est intégrée au MBC :

$$205 \times \text{glucose 6-P} + 70,9 \times \text{fructose 6-P} + \cdots \Rightarrow 1 \times \text{biomasse} + \ldots$$

5.4.4. Résultats générés

Les modèles à base de contraintes permettent d'obtenir une ou plusieurs cartes de flux qui décrivent la répartition et la vitesse des réactions (et des transports) au sein d'un réseau métabolique.

Ces cartes de flux peuvent être considérées comme des « clichés » qui révèlent l'état du métabolisme à un *moment* donné. La période de temps pendant laquelle une carte de flux est pertinente va dépendre de la durée pendant laquelle le système est à l'état stationnaire. Cette durée peut varier de quelques minutes jusqu'à plusieurs jours, par exemple lors de cultures bactériennes en chemostat (avec renouvellement en continu du milieu de culture).

Si l'on dispose des connaissances suffisantes pour contraindre totalement l'espace des solutions, il est possible d'obtenir une seule carte de flux. Lorsque le réseau est sous-déterminé (le nombre de flux est supérieur au nombre de métabolites) ou que les connaissances ne sont pas suffisantes, l'espace des solutions ne peut pas être pas totalement contraint. En conséquence, plusieurs cartes de flux sont possibles.

Différentes extensions existent afin de gérer cette multiplicité de solution, dont certaines seront présentées plus loin dans cette thèse. De manière simple, ces méthodes permettent soit de ne choisir qu'une seule carte parmi toutes celles possibles, soit d'analyser la variabilité des solutions possibles.

Table 5.2. – Composition en acides aminés des cellules d'*E. coli* et besoins en métabolites précurseurs. Souche *B/r*, données issues de [Neidhardt 1990].

Acide aminé	Quantité présente (µmol/g de cellule)	Besoins en précurseurs pour produire chaque acide aminé (µmol/µmol)			
		Métabolites[†]	ATP	NADH[††]	NADPH
Alanine	488	1 PYR	0	0	1
Arginine	281	1 AKG	7	−1	4
Asparagine	229	1 OA	3	0	1
Aspartate	229	1 OA	0	0	1
Cystéine	87	1 3PG	4	−1	5
Glutamate	250	1 AKG	0	0	1
Glutamine	250	1 AKG	1	0	1
Glycine	582	1 3PG	0	−1	1
Histidine	90	1 R5P	6	−3	1
Isoleucine	276	1 OA, 1 PYR	2	0	5
Leucine	428	2 PYR, 1 ACCOA	0	−1	2
Lysine	326	1 OA, 1 PYR	2	0	4
Méthionine	146	1 OA	7	0	8
Phénylalanine	176	1 E4P, 2 PEP	1	0	2
Proline	210	1 AKG	1	0	3
Sérine	205	1 3PG	0	−1	1
Thréonine	241	1 OA	2	0	3
Tryptophane	54	1 R5P, 1 E4P, 1 PEP	5	−2	3
Tyrosine	131	1 E4P, 2 PEP	1	−1	2
Valine	402	2 PYR	0	0	2

[†] 3PG, 3-phosphoglycérate ; ACCOA, acétyl-CoA ; AKG, α-kétoglutarate ; E4P, érythrose 4-phosphate OA, oxaloacétate ; PEP, phospho*énol*pyruvate ; PYR, pyruvate ; R5P, ribose 5-phosphate ;
[††] Une valeur négative indique que la synthèse de l'acide conduit à la production de NADH.

Table 5.3. – Métabolites précurseurs et énergie nécessaires pour produire un gramme de cellule d'*E. coli*. Souche *B/r*, données issues de [Neidhardt 1990].

Métabolite précurseur	Quantité (µmol/g de cellule)	Métabolite précurseur	Quantité (µmol/g de cellule)
Glucose 6-P	205	Pyruvate	2 832,8
Fructose 6-P	70,9	Acétyl-CoA	3 747,8
Ribose 5-P	897,7	α-kétoglutarate	1 078,9
Érythrose 4-P	361	Oxaloacétate	1 786,7
Glycéraldéhyde 3-P	129	ATP	18 485
3-phosphoglycérate	1 496	NADPH	18 225
Phospho*énol*pyruvate	519.1	NADH	−3 547

La modélisation à base de contraintes a donné lieu a de nombreux travaux. Parmi ceux-ci, citons la modélisation de métabolismes « complets » (« *genome-scale* ») [Kim 2012]. Par exemple Weaver et coll. ont construit un modèle chez *E. coli* décrivant 2 286 réactions et 1 453 métabolites [Weaver 2014]. D'autres modèles complets ont aussi été construits pour des organismes pluricellulaires, tels que les plantes [de Oliveira Dal'Molin 2013].

Les MBC ont de nombreuses applications ([Price 2004a, Bordbar 2014]). Ils permettent, par exemple, d'évaluer la réorganisation des flux selon les conditions environnementales [Almaas 2004], de comparer les répartitions de flux chez des souches bactériennes différentes [Coze 2013], ou encore d'étudier la compétition pour les ressources entre deux populations bactériennes [Zhuang 2011].

5.4.5. Avantages et inconvénients

La modélisation à base de contraintes ne nécessite pas de connaitre les paramètres cinétiques des réactions, cela permet de modéliser des réseaux de grandes tailles.

Par ailleurs, les vitesses de consommation ou de production de métabolites, couramment mesurées lors de l'étude du métabolisme, sont facilement intégrables : ces informations peuvent être directement traduites en contraintes sur les flux d'échange du système.

Un autre type de donnée classiquement mesurée lors de l'étude des procaryotes est la production de biomasse. Comme abordé plus en amont, cette production de biomasse peut être décrite dans un MBC comme un flux de sortie du système.

La modélisation à base de contraintes a aussi des désavantages. Par construction, un modèle à base de contraintes ne peut pas considérer la dynamique du système, puisque l'on suppose que ce système est à l'état stationnaire. Cela pose notamment des limites si l'on souhaite étudier le comportement d'un métabolisme dans un environnement changeant. Une façon de contourner ce problème est de considérer la dynamique du système comme étant une succession d'états stationnaires (hypothèse d'état quasi-stationnaire). Sur cette base, on peut ainsi générer une ou plusieurs cartes de flux représentatives de chaque état. Ce type d'approche a par exemple été utilisé par Covert et Palsson pour modéliser la dynamique du métabolisme chez *E. coli* à l'aide d'un MBC [Covert 2002].

Un second inconvénient est le fait qu'un modèle peut déboucher sur plusieurs cartes de flux solutions. Cela peut notamment poser des difficultés lorsque l'on souhaite modéliser la dynamique du système en calculant une succession de cartes de flux.

Un autre limite est l'impossibilité de considérer directement les concentrations des entités du système — métabolites, enzymes et transporteurs — que ce soit lors de la construction du modèle, ou lors de l'analyse des cartes de flux solutions, puisque les MBC considèrent que la concentration des métabolites internes n'évolue pas ($dm/dt = 0$).

5.4.6. Considérations sur la modélisation simultanée

Des approches qui utilisent les modèles à base de contraintes ont déjà été développées pour modéliser simultanément un réseau de régulation et un réseau métabolique. Lee et coll. ont ainsi présenté un travail où les réseaux sont décrit à l'aide d'un MBC [Lee 2008] [1] (modélisation *intégrée*, utilisant les matrices stœchiométriques). Covert et coll. ont quant à eux utilisé un modèle logique pour décrire le réseau de régulation, et une combinaison d'un modèle à base de contraintes et d'équations différentielles pour décrire le métabolisme [Covert 2008] (modélisation *couplée*). Ces deux travaux ont porté sur des réseaux de taille « moyenne » (de l'ordre de 50 à 100 réactions, environ 10 régulateurs transcriptionnels). Plus récemment, Karr et coll. ont développé une approche « cellule complète » dans laquelle la description du métabolisme utilise la modélisation à base de contraintes [Karr 2012].

Les régulations génétiques se font en réponse à des stimuli, et ces réponses entrainent *ensuite* des modification du fonctionnement du métabolisme. Cette dimension dynamique pose une première difficulté dans le cadre de la modélisation à base de contraintes : puisque les MBC supposent l'état stationnaire du système, la vitesse des flux — dans un MBC du métabolisme — ou le niveau d'expression des gènes — dans un MBC du réseau de régulation génétique — sont constants dans chaque carte solution.

En conséquence, la modélisation simultanée d'un réseau de régulation génétique et d'un réseau métabolique passe nécessairement par la génération d'une succession de cartes solutions, où chaque carte représente un état transitoire du système. Il faut alors définir comment l'état du système à

1. Pour être exact, les auteurs modélisent aussi un réseau de signalisation

un instant t détermine les contraintes du système pour un instant suivant $t + 1$.

Le réseau de régulation agit sur le métabolisme en régulant la concentration d'une enzyme, concentration qui peut ensuite avoir une influence sur la vitesse de la réaction catalysée. Si le métabolisme est modélisé avec un MBC, l'effet de la modification de la concentration d'une enzyme peut être modélisé en modifiant les contraintes sur le flux associé à cette enzyme. Ce type de couplage a été utilisé par Covert et coll. [Covert 2001]. Dans ce travail, un modèle logique décrit le réseau de régulation génétique, et un MBC décrit le métabolisme. Chaque protéine du RRG ne peut avoir que deux états : présente ou absente. Lorsqu'une enzyme est absente, la vitesse du flux associé est contrainte à 0 dans le MBC. Une manière plus fine de réguler les flux est d'utiliser des contraintes d'inégalité, afin de définir la vitesse maximale d'un flux en fonction du niveau d'expression de l'enzyme. Cependant, il faut noter que le niveau d'expression d'une enzyme n'est pas toujours corrélé à la vitesse d'une réaction dans un contexte *in vivo* : la concentration en enzyme active peut être régulée par des phénomènes de régulation post-traductionnelle, comme c'est le cas pour l'enzyme isocitrate dehydrogenase chez *E. coli* [Cozzone 2005].

L'action du métabolisme sur le réseau de régulation génétique dépend de la concentration des métabolites. Or, le résultat d'un MBC du métabolisme est une carte de flux, qui ne donne pas d'information sur le niveau de concentration des métabolites. Le couplage avec un modèle du RRG nécessite donc un travail de « traduction » des vitesses de flux vers des concentrations en métabolites (ces concentrations pouvant être exprimées de manière qualitative ou quantitative, selon la méthode utilisée pour modéliser le RRG). Une piste éventuellement intéressante est l'utilisation des propriétés thermodynamiques des réactions, afin d'estimer des jeux de concentrations thermodynamiquement compatibles avec le sens et la vitesse des flux [Beard 2005]. Une autre piste est de modéliser le métabolisme en utilisant de manière combinée la modélisation à base de contraintes et les équations différentielles [Smallbone 2010].

Un dernier point important dans le cadre du couplage RRG $\leftrightarrow$ RM est la différence de vitesse des processus. Au niveau du métabolisme la vitesse des réactions chimiques change rapidement lorsque la concentration des substrats ou des produits évolue (changement de l'ordre de quelques secondes, [Mashego 2006]). Au niveau du réseau de régulation génétique, l'activation d'un gène nécessite un certain délai avant que le produit soit

exprimé et fonctionnel (de l'ordre de quelques minutes, [Zubay 1973, McAdams 1998]).

Troisième partie

Dynamique d'un réseau métabolique avec un modèle à base de contraintes : approche par échantillonnage des trajectoires solutions

Suite à mon travail de revue sur les formalismes de modélisation, la modélisation à base de contraintes est apparue comme très puissante pour étudier les réseaux métaboliques : le fait que les MBC ne nécessitent pas de paramétrage permet l'étude de ces réseaux même lorsque les connaissances sont incomplètes.

Revers de la médaille, la dimension dynamique ne peut être prise en compte directement dans les MBC, puisque l'on pose l'hypothèse que le système étudié est à l'état stationnaire. De plus, lorsque l'espace des solutions n'est pas suffisamment contraint, la répartition des flux ne peut être estimée de manière unique — le système est dit sous-déterminé.

Afin d'élargir le champ des applications des MBC, de nombreuses approches ont été proposées pour (*i*) gérer la multiplicité des cartes solutions, ou (*ii*) produire une dynamique à l'aide du formalisme des MBC, voire (*iii*) combiner ces deux caractéristiques. Je présente quelques-unes de ces approches dans le chapitre 6 (p. 99).

La dynamique d'un réseau métabolique peut être intégrée dans un MBC en considérant que le système passe par une succession d'états stationnaires. Dans cette approche, le temps est discrétisé en périodes, et le « potentiel » de fonctionnement du système est décrit pour chaque période par un espace des solutions contraint par des valeurs spécifiques (les valeurs des contraintes dépendent de la période de temps).

Les MBC dynamiques sont notamment utilisés pour modéliser le fonctionnement du métabolisme des bactéries au cours d'une culture. Dans ce cadre, les flux d'échange entre le système et l'environnement extérieur sont typiquement les flux de consommation de substrats et de production de biomasse. Ces flux d'entrées / sorties (flux E/S) sont généralement déterminés à partir de mesures expérimentales, réalisées à différents moments de la culture, et servent à contraindre l'espace des solutions de manière temps dépendant.

La variabilité est une composante fondamentale de l'étude des systèmes biologiques. Cette variabilité résulte d'une part de différences de fonctionnement entre les cellules d'une même population, et d'autre part de la chaîne des manipulations nécessaires pour effectuer des mesures sur les entités et processus biologiques (technicité de l'expérimentateur, qualité de l'équipement, précision des dispositifs...). La variabilité mesurée expérimentalement peut être « traduite » dans les MBC en utilisant des contraintes d'inégalité sur les flux (plutôt que des contraintes d'égalité). Ainsi, un flux

d'entrée v_e dans le système peut être exprimé par une contrainte d'égalité, par exemple $v_{e,glu} = 10$ (mmol/h). Toujours pour l'exemple, si l'on sait que ce flux varie entre 8 et 12 (mmol/h), on peut reformuler la contrainte d'égalité par deux contraintes d'inégalité : $8 \leq v_{e,glu} \leq 12$.

Lorsque la quantité de données est faible et que de la variabilité est introduite sur les contraintes, la répartition des flux dans le réseau ne peut pas être estimée de manière unique : il existe différentes répartitions de flux qui sont toutes en accord avec les données à disposition. Une manière de choisir parmi toutes les répartitions possibles est alors de poser une hypothèse sur un phénomène biologique qui serait optimisé au sein du métabolisme étudié. Typiquement, l'hypothèse sera que le système optimise la production de biomasse ou d'énergie. Cependant, le « but » vers lequel tend le métabolisme n'est réellement connu que dans certaines situations.

Plutôt que de choisir une carte particulière, une autre manière de gérer la multiplicité des cartes solutions est d'évaluer l'espace des solutions en entier. En sélectionnant aléatoirement des cartes de flux solutions, l'échantillonnage de l'espace des solutions permet ainsi d'estimer la variabilité et la distribution de probabilité de chacun des flux du système.

Bien que conceptuellement intéressante, la prise en compte de la variabilité dans les MBC dynamiques est encore relativement peu étudiée. Compte tenu des améliorations qu'il était possible d'apporter aux méthodes MBC, mon sujet de thèse a évolué vers la recherche d'une approche qui utilise les méthodes MBC pour modéliser la dynamique du métabolisme, considéré indépendamment de la régulation, et ce lorsque la quantité de données est faible et sans poser d'hypothèse sur l'optimum du système métabolique. L'enjeu majeur autour du développement d'une telle méthode est de permettre d'estimer de manière *fine* la variabilité des flux dans les modèles MBC dynamiques.

En associant le formalisme des MBC avec l'échantillonnage de l'espace des solutions, j'ai mis au point une nouvelle approche qui permet d'estimer au cours du temps les densités de distribution des valeurs de flux. La dynamique du système y est évaluée en calculant une population de *trajectoires solutions*, où chaque trajectoire est constituée d'une succession de cartes solutions. En introduisant une contrainte de faisabilité entre les cartes de chaque trajectoire, la méthode permet de faire des prédictions plus réalistes, et de réduire la variabilité des flux prédits.

Cette méthode est présentée sous la forme d'un article dans le chapitre 7 (p. 129). Afin d'illustrer les capacités de la méthode, je l'ai appliquée pour modéliser le comportement du métabolisme de la bactérie *Corynebacterium glutamicum* lors de la croissance en limitation en biotine. Les résultats obtenus sont aussi abordés au fil de l'article.

Dans le chapitre 8 (p. 161), j'approfondis plusieurs points brièvement abordés dans l'article, notamment, le choix de la méthode d'échantillonnage, et le choix des paramètres de l'approche. Le chapitre 9 (p. 187) est ensuite l'occasion de discuter d'une part, de l'impossibilité de modéliser l'une des périodes de la culture de *C. glutamicum*, et d'autre part de montrer en quoi les prédictions peuvent être utilisées pour améliorer le modèle. Enfin, le chapitre 10 (p. 223) conclut ce travail de thèse, et propose quelques éléments de réflexions pour la conception d'autres approches.

Chapitre 6

Gestion de la multitude de solutions et de la dynamique dans les MBC

Depuis plus de 30 ans, les MBC sont utilisés pour modéliser les réseaux métaboliques. Au fil des années, de nombreux développements sont venus améliorer le principe de base de ce formalisme. Les récents articles de revue de Bordbar et coll. [Bordbar 2014], et de Lewis et coll. [Lewis 2012] illustrent bien la quantité et la diversité des approches basées sur les MBC. Dans le premier, les auteurs proposent de revenir sur des résultats récemment obtenus grâce aux MBC. L'article est accompagné d'une base de données qui référence plus 600 travaux scientifiques publiés entre 1986 et 2013 [1] (figure 6.1).

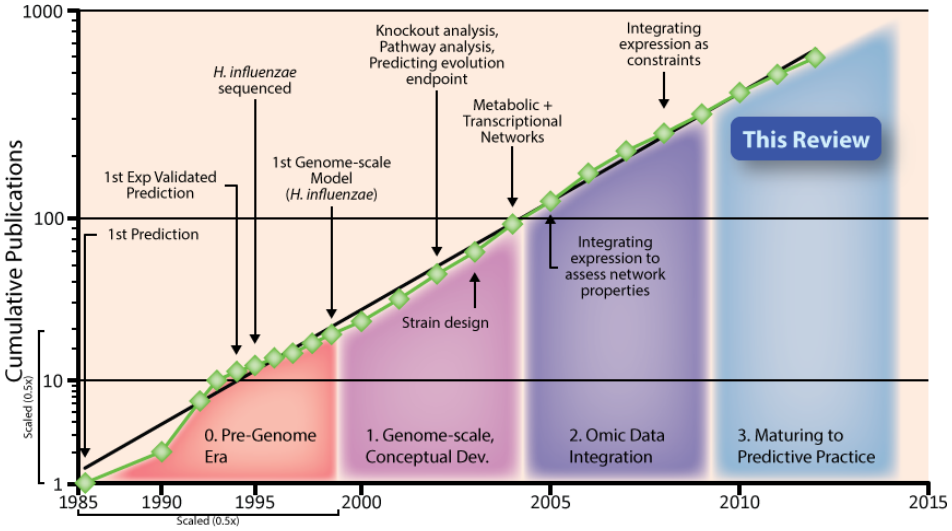


Figure 6.1. – Nombre d'articles publiés sur les MBC de 1985 à 2013. Figure associée à [Bordbar 2014], issue de `http://sbrg.ucsd.edu/cobra-predictions`.

1. En ligne, `http://sbrg.ucsd.edu/cobra-predictions`, accédé le 15 janvier 2015

Dans le second article, les auteurs proposent une « phylogénie » des approches existantes, regroupées selon quelques caractéristiques (figure 6.2). Une liste de méthodes, ainsi qu'une liste de *packages* logiciels dédiés aux MBC, a été mise en ligne par les auteurs [2]. Un troisième article de revue particulièrement intéressant, qui présente entre autre les différents types d'analyses réalisables à partir d'un modèle MBC, est celui de Durot et coll. [Durot 2009].

Dans ce chapitre, je vais me concentrer sur deux caractéristiques : la gestion de la multitude de solutions lorsque le réseau est sous-déterminé, et la production d'une dynamique.

2. En ligne, http://cobramethods.wikidot.com, accédé le 15 janvier 2015

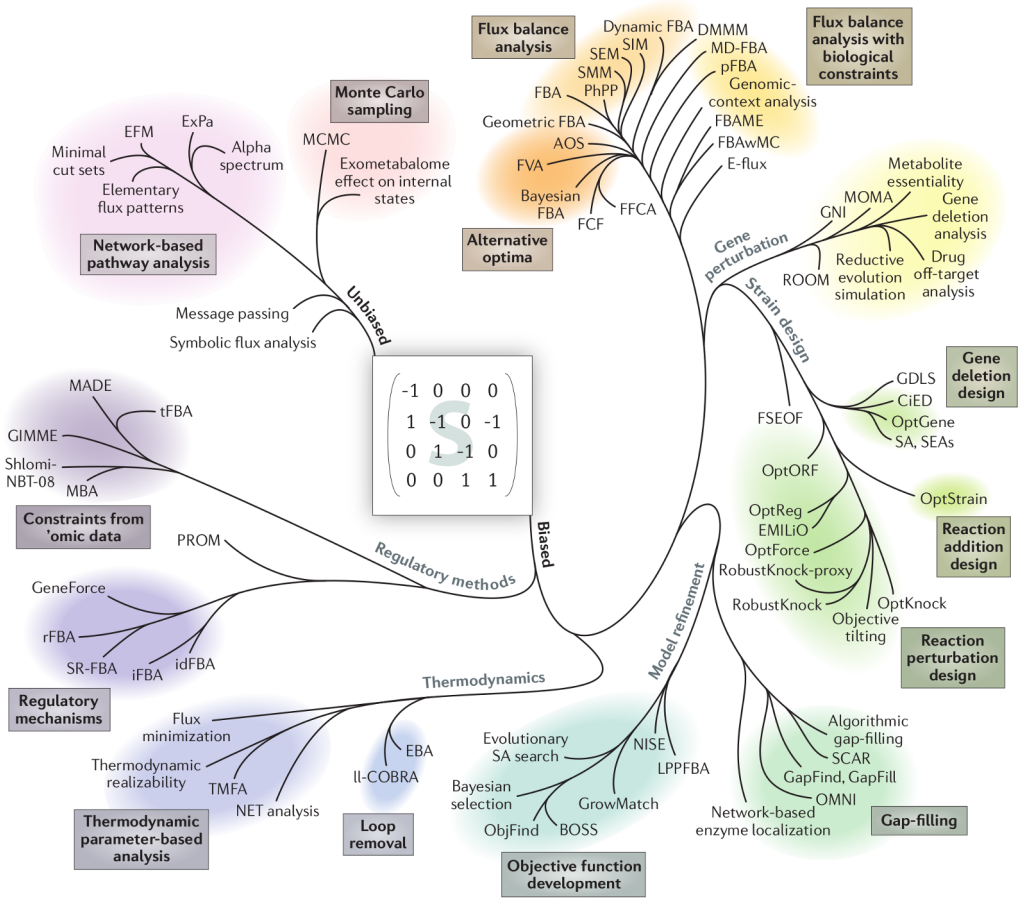


Figure 6.2. – Représentation « phylogénique » des approches MBC. Figure issue de [Lewis 2012].

6.1. Gestion de la multitude de solutions

6.1.1. Notion de sous-détermination d'un réseau

Nous avons vu dans le chapitre précédent que la topologie d'un réseau métabolique peut être représentée sous la forme d'une matrice stœchiométrique S, où chaque ligne représente un métabolite interne, où chaque colonne représente une réaction, et où les valeurs de la matrice représentent la stœchiométrie de chaque réaction. Dans les MBC, la dynamique du système n'est pas considérée, et l'on pose l'hypothèse de l'état stationnaire des métabolites internes. En conséquence, le modèle peut être décrit par un système d'équations linéaires : chaque ligne/métabolite de la matrice correspond à une équation linéaire, et le flux de chaque colonne/réaction est une variable inconnue (figure 6.3).

Lorsque le nombre r d'équations indépendantes est inférieur au nombre n de variables inconnues, le système d'équations est dit *sous-déterminé* : il peut exister plusieurs solutions (combinaisons de valeurs de flux) qui satisfont le système d'équations. Le nombre de *degrés de liberté* d d'un tel système est $d = n - r$. L'ensemble de ces solutions réside dans un espace vectoriel de dimension d, l'espace des solutions, qui représente le « potentiel » du réseau métabolique, en terme des répartitions possibles des flux au sein du réseau métabolique étudié.

Il est intéressant de noter ici que les contraintes d'égalité et d'inégalité ont un impact différent quant à la manière dont elles réduisent le volume de l'espace des solutions (on parlera d'*hyper volume* lorsque $d > 3$) :

— L'utilisation des contraintes d'égalité, par exemple $v_2 = a$, diminue le nombre de variables inconnues du système, et donc le nombre de degrés de liberté. En conséquence, le volume de l'espace des solutions est diminué *via* la réduction du nombre de dimensions.

— L'utilisation de contraintes d'inégalité, par exemple $v_1 \geq 0$, diminue l'intervalle des valeurs solutions pour une variable inconnue, mais sans toutefois diminuer le nombre de variables. Par conséquent, le volume de l'espace des solutions est diminué uniquement *via* la réduction de l'étendue des valeurs possibles sur l'un des axes de l'espace, sans réduction du nombre de dimensions.

a) Système d'équations linéaires

$$\begin{aligned}
\frac{dm_1}{dt} &= +t_1 - v_1 &&= 0\\
\frac{dm_2}{dt} &= +t_2 - v_2 &&= 0\\
\frac{dm_3}{dt} &= +v_1 + v_2 - t_3 &&= 0
\end{aligned}$$

b) Système d'équations sous forme matricielle

$$\begin{array}{c} \\ m_1 \\ m_2 \\ m_3 \end{array}
\overbrace{\begin{array}{c} \begin{array}{ccccc} v_1 & v_2 & t_1 & t_2 & t_3 \end{array} \\ \begin{bmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & -1 \end{bmatrix} \end{array}}^{S\text{ : matrice de stœchiométrie}}
\cdot
\overbrace{\begin{bmatrix} v_1 \\ v_2 \\ t_1 \\ t_2 \\ t_3 \end{bmatrix}}^{v\text{ : vecteur des valeurs de flux}}
= 0$$

Figure 6.3. – Système d'équations linéaires d'un MBC.

6.1.2. Choix d'une solution parmi toutes : méthode *flux balance analysis* (FBA)

Une manière de choisir parmi toutes les répartitions possibles est de poser une hypothèse sur un phénomène biologique qui serait optimisé au sein du métabolisme étudié. Cette démarche peut être qualifiée de *biaisée,* car elle se base sur des hypothèses supplémentaires afin d'identifier une unique solution. Les nombreuses approches qui exploitent cette démarche sont représentées dans la branche *biased* de la figure 6.2 (p. 101). Je vais présenter ici l'approche *flux balance analysis* (FBA) qui est l'une des plus utilisées [Kauffman 2003, Palsson 2006, Raman 2009, Orth 2010].

La méthode FBA utilise l'optimisation pour identifier une répartition de flux qui optimise un comportement prédéfini du système modélisé. Le comportement à optimiser est défini à l'aide d'une fonction de coût, appelée *fonction objectif*, que l'on peut écrire sous la forme

$$Z = p \cdot v$$

où :

— v est le vecteur des flux (variables inconnues du système d'équations) ;

— p est un vecteur de poids, de même taille que v, qui pondère la contribution de chaque flux au score de la fonction *objectif*.

Il s'agit ensuite d'utiliser une procédure d'optimisation afin d'identifier la répartition de flux qui maximise (ou minimise, selon le comportement recherché) la fonction *objectif*, tout en prenant en compte la topologie du réseau et les contraintes d'égalité et d'inégalité :

$$\begin{array}{ll} \text{Optimiser} & Z = p \cdot v \\ \text{tel que} & S \cdot v = 0 \\ \text{et} & \left\{ \begin{array}{l} v_{min} \leq v \leq v_{max} \\ v = v_{contr} \end{array} \right. \end{array}$$

Pour illustrer simplement le fonctionnement de la méthode FBA, on peut reprendre le réseau métabolique fictif donné dans le chapitre précédent à la page 83 (réseau redonné par la figure 6.4), auquel on applique les contraintes listées par la table 6.1. Sur ce modèle, on peut par exemple s'intéresser à la répartition des flux qui maximise la consommation du métabolite m_2, ce qui revient à maximiser la fonction de score $Z = t_2$. La figure 6.5 représente la position de la carte de flux solution au sein de l'espace des solutions.

Les fonctions *objectif* permettent de formuler un large panel de comportements d'intérêt. De nombreuses fonctions *objectif* ont été utilisées pour étudier des réseaux métaboliques. La maximisation de la production de biomasse est probablement la plus rencontrée dans les études sur les systèmes bactériens. Elle est par exemple utilisée pour prédire le taux de croissance maximal d'*E. coli* en fonction des conditions expérimentales [Varma 1994, Edwards 2001]. Parmi les autres fonctions *objectif*, citons la maximisation de la production d'ATP, utilisée pour étudier le fonctionnement de la mitochondrie [Ramakrishna 2001] ; la minimisation de la consommation de substrat [Famili 2003], ou encore la maximisation de la production de métabolites secondaires [Varma 1993]. Notons qu'il est aussi possible de formuler des fonctions *objectif* qui ne soient pas des combinaisons linéaires de flux. On peut par exemple rechercher la solution qui optimise la production de biomasse, tout en minimisant les ajustements des vitesses de flux vis-à-vis d'une carte de référence [Segrè 2002, Shlomi 2005].

L'usage d'une fonction *objectif* pour prédire le comportement d'un système métabolique nécessite de poser une hypothèse quant à l'optimalité du réseau métabolique étudié : on suppose que l'organisme étudié s'est adapté, au fil de l'évolution, de manière optimale aux conditions environnementales. L'une des difficultés de la méthode FBA réside dans l'identification du comportement, du « but », qui serait optimisé par l'organisme.

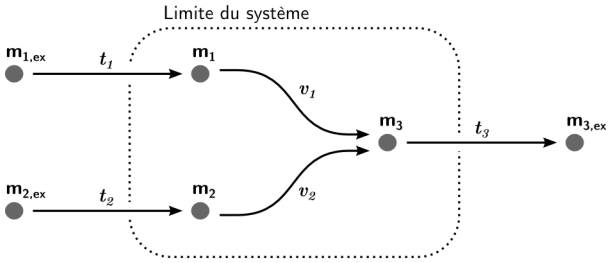


Figure 6.4. – Exemple d'un réseau métabolique.

Table 6.1. – Exemples de contraintes appliquées sur le réseau.

Contraintes	Explications
$v \geq 0$	toutes les réactions sont considérées comme irréversibles.
$v_1 \leq 5$	Le flux v_1 ne peut excéder une vitesse de 5.
$t_3 \leq 10$	Le flux d'échange t_3 ne peut excéder 10. Compte tenu de la topologie, cela implique que $v_1 + v_2 \leq 10$.

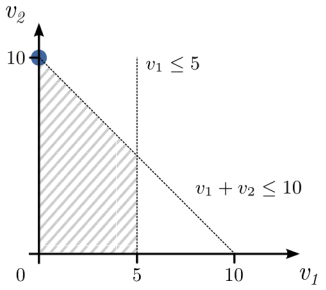


Figure 6.5. – Espace des solutions et solution FBA. L'espace des solutions du réseau contraint est représenté en hachures sur le plan v_1-v_2. Les contraintes sur les valeurs maximales des flux sont représentées par des droites en pointillé. La carte de flux qui optimise la fonction $Z = t_2$ est représentée par un point bleu.

Dans le cadre de l'étude du métabolisme de bactéries, la production de biomasse est souvent utilisée comme fonction *objectif*. L'hypothèse forte qui est posée est alors que le fonctionnement des bactéries a évolué afin de convertir le plus efficacement possible les substrats disponibles en biomasse (production de nouvelles cellules par division cellulaire). L'utilisation de cette fonction *objectif* « production de biomasse » a permis d'obtenir des prédictions cohérentes avec les résultats expérimentaux dans de nombreuses publications ([Edwards 2001, Schilling 2002, Ibarra 2002, Shinfuku 2009], par exemple).

Toutefois, l'hypothèse que le métabolisme a été optimisé pour produire de la biomasse n'est pas valide dans toutes les situations [Schuster 2008]. Le comportement optimisé par le métabolisme semble en fait dépendre de l'organisme considéré [Schuster 2008], et des conditions environnementales [Schuetz 2007]. Ces dernières années, l'identification des principes qui gouverneraient le fonctionnement du métabolisme a donné lieu à de nombreux travaux. Schuetz et coll. proposent par exemple que le métabolisme bactérien ait évolué selon un compromis entre (i) l'optimalité dans une condition environnementale spécifique et (ii) une minimisation des ajustements pour s'adapter à d'autres conditions [Schuetz 2012]. Zarecki et coll. constatent quant à eux que la maximisation des flux d'échange entre l'environnement et le système prédit des taux de croissance plus pertinents que la maximisation de la production de biomasse [Zarecki 2014].

Notons que la connaissance de « l'objectif » qui serait optimisé par le système n'est pas suffisante pour générer des prédictions pertinentes : le réseau métabolique étudiée doit aussi être suffisamment bien décrit et contraint pour que l'optimisation de la fonction *objectif* conduise à des résultats conformes avec une réalité.

Outre l'importance du choix de la fonction *objectif*, la méthode FBA présente deux limites majeures. Premièrement, plusieurs solutions optimales peuvent avoir le même score (score de la fonction *objectif* optimisée), tout en présentant des répartitions de flux différentes. Dans le réseau de la figure 6.4, ce serait par exemple le cas si l'on cherchait à optimiser la fonction *objectif* $Z = t_3$: il y a une infinité de solutions optimales, situées sur la droite $v_1 + v_2 = 10$ de l'espace des solutions (figure 6.5). Deuxièmement, comme la méthode FBA identifie une seule répartition de flux, on ne peut pas évaluer la variabilité potentielle de chacun des flux du réseau.

6.1.3. Estimation de l'intervalle des valeurs possibles pour chaque flux : méthode *flux variability analysis* (FVA)

La méthode *flux variability analysis* (FVA) a été initialement développée afin d'évaluer la variabilité des solutions optimales alternatives [Mahadevan 2003]. L'idée est ici d'évaluer, parmi la « famille des répartitions optimales de flux », l'intervalle des valeurs possibles de chacun des flux du système.

Pour cela, il s'agit dans un premier temps de réaliser une analyse FBA classique afin de déterminer la valeur optimale z atteignable par la fonction *objectif*. Cette valeur est ensuite utilisée comme contrainte supplémentaire sur le MBC :

$$\begin{aligned} & S \cdot v = 0 \\ \text{sachant} \quad & \begin{cases} v_{min} \leq v \leq v_{max} \\ v = v_{contr} \end{cases} \\ \textbf{et} \quad & p.v = z \end{aligned} \tag{6.1}$$

De cette manière, l'espace des solutions est restreint et ne contient plus que l'ensemble des solutions optimales. Dans un second temps, la méthode FVA consiste à minimiser et maximiser, par optimisation linéaire et de manière itérative, la valeur de chacun des flux : pour chaque flux v_i du système, calculer $Min(v_i)$ et $Max(v_i)$ en respectant l'équation (6.1).

Par exemple, dans le système décrit par la figure 6.4 et la table 6.1 (p. 105), on peut s'intéresser à la valeur optimale de la fonction $Z = t_3$: on obtient $z = 10$ (du fait de la contrainte $t_3 \leq 10$) [3]. L'espace des solutions défini par l'équation (6.1) correspond alors à la droite $v_1 + v_2 = 10$ représentée sur la figure 6.6 (p. 109). On détermine ensuite les intervalles des valeurs possibles par la méthode FVA, ce qui donne $0 \leq v_1 \leq 5$ et $5 \leq v_2 \leq 10$ pour les flux internes du réseau.

La méthode FVA peut aussi être utilisée afin d'évaluer les solutions *sous*-optimales [Mahadevan 2003, Reed 2004]. Dans cette situation, un pourcentage de la valeur optimale z est utilisé comme contrainte d'inégalité.

3. Notons que, dans une situation réelle, l'optimisation d'une variable dont la valeur est contrainte est peu judicieuse, mais que, pour l'exemple, cette situation a le mérite d'être simple.

Par exemple, l'espace des solutions qui contient l'ensemble des solutions capables d'atteindre au moins 95% d'une valeur optimale est contraint par :

$$\begin{aligned} & S \cdot v = 0 \\ \text{sachant} \quad & \begin{cases} v_{min} \leq v \leq v_{max} \\ v = v_{contr} \end{cases} \\ \textbf{et} \quad & p.v \geq 0.95 \times z \end{aligned}$$

La méthode FVA, tel qu'originalement publiée, peut être qualifiée de méthode biaisée puisque l'on s'intéresse uniquement a un sous-ensemble particulier des solutions. À ce titre, elle est représentée dans la branche *biased* de la classification donnée par la figure 6.2 (p. 101). Mais, l'approche de la méthode FVA peut aussi être utilisée de manière non biaisée : pour cela, il suffit de ne pas contraindre le modèle avec une valeur z. On peut alors déterminer les intervalles des valeurs possibles pour chacun des flux sans poser d'hypothèse sur l'optimalité du système. Llaneras et coll. ont appelé cette variation *flux spectrum approach* (FSA) [Llaneras 2007a].

La méthode FVA présente toutefois deux limites majeures. Tout d'abord, le calcul des valeurs extrêmes des flux est réalisé itérativement, de manière indépendante. En conséquence, certaines combinaisons des valeurs de flux (valeurs comprises dans les intervalles calculés) ne sont pas solutions du système. C'est par exemple le cas dans la figure 6.6, où la combinaison $v_1 = 5$ et $v_2 = 10$ n'est pas une solution respectant les contraintes imposées au système. La méthode FVA fournit donc une surestimation de l'espace des solutions. Par ailleurs, la méthode ne permet pas d'estimer la probabilité d'occurrence de chaque valeur de flux. Ainsi, dans l'espace des solutions de la figure 6.5 (p. 105), le calcul des intervalles ne permet pas de voir qu'il existe « plus » de combinaisons de valeurs de flux lorsque $v_2 = 5$ que lorsque $v_2 = 10$, c'est-à-dire que l'intervalle de valeurs possibles pour v_1 est plus grand quand $v_2 = 5$ que quand $v_2 = 10$.

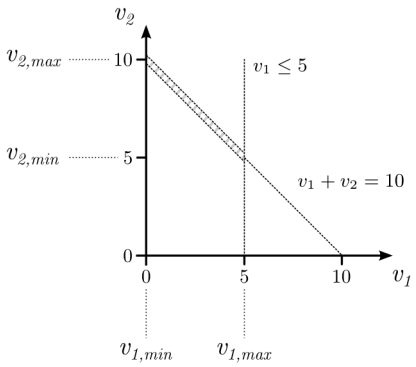


Figure 6.6. – Espace des solutions et méthode FVA. L'espace des solutions est représenté en hachure sur le plan v_1-v_2. L'égalité $z = 10 = t_3$ contraint fortement l'espace des solutions, et l'ensemble des solutions optimales est présent sur la droite $v_1+v_2 = 10$. La minimisation et la maximisation successive des flux permettent ensuite de déterminer les intervalles de valeurs possibles pour chaque flux.

6.1.4. Échantillonnage de l'espace des solutions : méthode *Monte Carlo markov chain* (MCMC)

L'échantillonnage de l'espace des solutions consiste à choisir au hasard un ensemble de points de l'espace des solutions. Le sous-ensemble — l'échantillon — de solutions ainsi obtenu permet ensuite d'analyser différentes caractéristiques du système [Palsson 2006, Schellenberger 2009]. Cette approche permet d'étudier un MBC de manière non biaisée, puisqu'aucune hypothèse sur le fonctionnement du réseau n'est nécessaire.

Les algorithmes les plus utilisés afin de réaliser l'échantillonnage d'un espace des solutions sont ceux basés sur la méthode de Monte-Carlo par chaînes de Markov (*Monte Carlo markov chain* ; MCMC). De manière générale, l'exécution de l'algorithme consiste à déterminer un point initial solution du système, puis à déplacer itérativement ce point au sein de l'espace des solutions. À chaque itération, la direction et la distance parcourue par le point sont déterminées selon des règles probabilistes.

Dans l'algorithme « mirror », développé par Van den Meersche et coll. [Van den Meersche 2009], à chaque itération, la distance parcourue sur chaque axe de l'espace des solutions est tirée au sein d'une loi normale de moyenne 0 et d'écart-type fixé. Si le mouvement du point amène celui-ci en dehors de l'espace des solutions, autrement dit si le mouvement du point implique d'enfreindre l'une des contraintes d'inégalité, alors la contrainte est utilisée comme un plan de réflexion (à la façon d'un « miroir ») : le mouvement du point est réfléchi par le plan, et le trajet est poursuivi à l'intérieur de l'espace des solutions (figure 6.7). De cette manière, chaque point généré est une solution valable du MBC.

Le fait de disposer d'une population de cartes de flux « solutions » donne accès à de nombreuses analyses. On peut notamment s'intéresser à la distribution statistique des valeurs de flux. Contrairement à la méthode FVA où seules les valeurs minimales et maximales de chaque flux sont connues, l'échantillonnage de l'espace des solutions fournit une population de valeurs. Cela permet alors d'estimer la fonction de densité de probabilité pour chaque flux du système (figure 6.8). Les fonctions de densité de probabilité ont ainsi été utilisées pour estimer la taille et la forme de l'espace des solutions [Wiback 2004] ou pour étudier l'impact de différentes contraintes génétiques sur les capacités du système [Price 2004b].

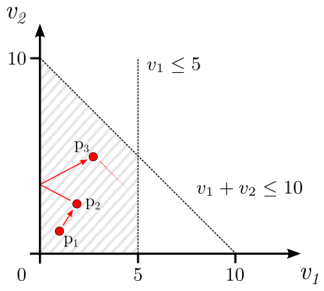


Figure 6.7. – Principe de l'algorithme « mirror » pour l'échantillonnage de l'espace des solutions. Les points de couleur rouge représentent les solutions échantillonnées, les flèches indiquent le mouvement appliqué entre chaque point.

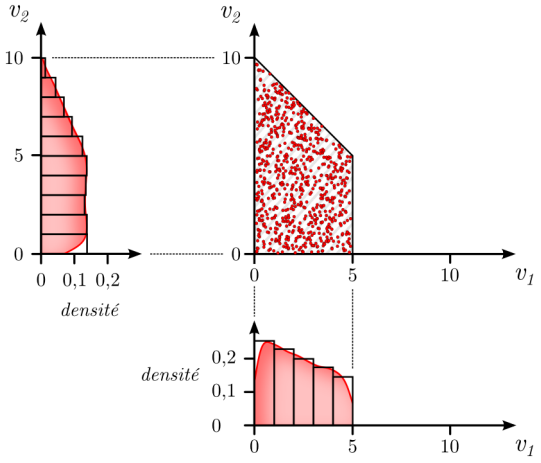


Figure 6.8. – Fonctions de densité des probabilités estimées à partir de l'échantillonnage de l'espace des solutions. Dans le graphique central, les points de couleur rouge représentent les solutions échantillonnées. Cette population permet d'estimer la répartition des valeurs de chaque flux. Le graphique de gauche présente la distribution des valeurs pour le flux v_2 sous la forme d'un histogramme (en noir) et d'une fonction de densité de probabilités (en rouge). Le graphique du bas présente les mêmes informations pour le flux v_1.

On peut aussi exploiter les fonctions de densité de probabilité afin de choisir, pour chaque flux, une valeur qui sera considérée comme représentative de l'état du système.

Rappelons ici que, pour un flux donné v_i, la fonction de densité de probabilité est estimée à partir de la distribution des solutions le long de l'axe v_i de l'espace des solutions. Sous l'hypothèse que l'échantillonnage est homogène, la *probabilité* associée à $v_i = a$ reflète alors la *proportion* de l'espace des solutions qui est située aux coordonnées a sur l'axe v_i de l'espace des solutions. Par exemple, dans la figure 6.8, la valeur $v_2 = 5$ est plus probable que $v_2 = 10$ parce que le « sous-espace » des solutions est plus grand quand $v_2 = 5$ (toutes les valeurs de v_1 comprise entre 0 et 5 sont solutions) que quand $v_2 = 10$ (la seule valeur solution de v_1 est 0). D'un point de vue biologique, l'espace des solutions représente l'ensemble des comportements possibles du réseau. Si ces comportements sont tous équiprobables (en référence à l'échantillonnage supposé homogène), alors la probabilité associée à une valeur de flux représente la chance que l'on aurait d'observer cette vitesse par hasard.

Le mode, la médiane et la moyenne sont trois descripteurs qui peuvent être utilisés comme valeur représentative d'un flux. Le *mode* d'une distribution de valeur v_i représentera la valeur du flux la plus probable. Dans [Thiele 2005], les modes des distributions sont comparés à des mesures *in vivo*, et ils semblent être de bons estimateurs pour certaines conditions expérimentales. Cependant, le mode d'une distribution est relativement sensible aux fluctuations des valeurs. Plus robuste, la *médiane* d'un flux représentera la valeur qui coupera l'espace des solutions, selon l'axe v_i, en deux parts égales. Notons ici qu'une carte de flux « reconstruite » à partir des modes ou des médianes ne respecte pas nécessairement l'hypothèse d'état stationnaire. Pour cette raison, dans [D'Huys 2012], le mode des distributions est utilisé uniquement comme critère pour choisir une fonction *objectif*, puis les auteurs étudient le fonctionnement du réseau à partir de la solution FBA. van Oevelen et coll. préfèrent utiliser la valeur moyenne comme estimateur : du fait de ses propriétés algébriques, une carte de flux reconstruite à partir des valeurs moyennes garantit l'état stationnaire du système [van Oevelen 2010]. Cependant, des trois estimateurs abordés, la moyenne est celui le plus sensible aux fluctuations des données. Ajoutons que le fait de disposer de distributions de valeurs permet aisément de calculer des intervalles de confiance autour de la valeur « représentative » choisie.

Une autre manière d'exploiter une population de solutions est de calculer la corrélation entre les valeurs de chaque paire de flux [Price 2004b] afin d'identifier des ensembles de réactions corrélées (*correleted reaction sets* ; « Co-Sets »). Le calcul des Co-Sets permet, par exemple, de déterminer des « modules métaboliques fonctionnels » [Thiele 2005]. Les Co-Sets parfaitement corrélés ($R^2 = 1$) peuvent être identifiés par une analyse de la topologie du réseau, tandis l'échantillonnage permet en plus d'identifier des Co-Sets partiellement corrélés [Xi 2011].

Finissons en abordant quelques limites de l'échantillonnage de l'espace des solutions. Tout d'abord, il est nécessaire d'échantillonner un grand nombre de points afin d'obtenir un échantillon représentatif. En conséquence, le temps de calcul nécessaire à l'échantillonnage peut être très important pour un réseau métabolique de grande taille, notamment pour les modèles « cellules complètes ». Une manière d'évaluer la qualité de l'échantillonnage est de comparer l'intervalle des valeurs de chaque flux avec les valeurs calculées avec la méthode FVA [D'Huys 2012]. Une autre approche consiste à réaliser plusieurs échantillonnages indépendants, puis de comparer la similitude entre les fonctions de densité de probabilité estimées [Thiele 2005].

6.1.5. Analyse d'un système surdéterminé : méthode *metabolic flux analysis* (MFA)

Même si la méthode *metabolic flux analysis* (MFA) est une approche peu adaptée à l'analyse des systèmes sous-déterminés, elle est très utilisée pour les systèmes (sur-)déterminés [Wiechert 2001, Antoniewicz 2015], c'est pourquoi il est intéressant de la positionner par rapport aux autres approches présentées jusqu'ici.

Dans la méthode MFA traditionnelle, seules les contraintes d'égalité v_{contr} sont utilisées. Parmi l'ensemble des flux v du système, je note v_c les flux contraints.

Ces contraintes correspondent généralement à des flux mesurés expérimentalement. Le système est dit déterminé lorsque le nombre d'équations linéaires indépendantes est égal au nombre de variables (flux) inconnues du système. Dans cette situation, il existe au plus une unique solution. Lorsque le nombre d'équations est supérieur au nombre de flux inconnus, il existe des contraintes redondantes, et le système est dit surdéterminé. En général, un tel système d'équations n'a pas de solution, car certaines contraintes sont incompatibles entre elles (figure 6.9).

Dans cette situation, la méthode MFA recherche la répartition des flux solutions v qui minimise — au niveau des flux contraints v_c — les écarts entre les valeurs v_{contr} initialement fixées et les valeurs prédites v_{pred} qui permettent de trouver une solution. D'un point de vue mathématique, ceci est réalisé par une minimisation de la somme du carré des écarts pondérés :

$$\begin{aligned} &\text{Minimiser} \quad SCE = \sum \left(v_{contr} - v_{pred}\right)^2 / \sigma^2_{v_{contr}} \\ &\text{tel que} \quad\quad S \cdot v = 0 \end{aligned}$$

Il existe de nombreuses extensions de la méthode MFA. Parmi celles-ci, citons notamment la méthode ^{13}C-MFA, qui exploite les mesures expérimentales d'isotopes stables (comme le carbone C^{13}) afin de déterminer la répartition des flux au niveau des points de branchement du réseau. De manière simplifiée, la connaissance des ratios de flux permet d'ajouter des équations linéaires dans le système, et ainsi d'augmenter le « niveau de détermination du système ». Par exemple, dans le réseau de la figure 6.4 (p. 105), l'ajout de l'équation $v_1 = v_2$ contraint le système à des solutions où la répartition des flux entre v_1 et v_2 est $v_1/v_2 = 1$. Cette approche a

par exemple permis à Coze et coll. de passer d'un réseau sous-déterminé à un degré de liberté à un réseau surdéterminé [Coze 2013]. Citons aussi les travaux de Antoniewicz et coll. qui ont présenté une méthode exploitant les mesures d'isotopes stables pour estimer des intervalles de confiance autour des flux prédits, et ce sans perdre l'aspect surdéterminé du réseau [Antoniewicz 2006]. L'estimation de l'incertitude sur les flux prédits est aussi introduite par la méthode *Possibilistic MFA* développée par Llaneras et coll. [Llaneras 2009]. Remarquons que cette méthode peut être utilisée pour étudier un système sous-déterminé, mais que les résultats générés sont alors proches des résultats générés par la méthode FVA.

La méthode MFA permet de prédire une unique répartition de flux, mais contrairement à la méthode FBA, cette résolution du système ne nécessite pas de poser d'hypothèse quant à l'optimalité du fonctionnement du métabolisme. Cependant, cette approche ne peut être appliquée que dans le cas où le système est (sur-)déterminé. La surdétermination d'un système peut être atteinte pour des réseaux tailles raisonnables (dans [Coze 2013], le système contient 46 *pseudo*-réactions qui représentent 265 réactions enzymatiques réelles), mais cela nécessite d'accompagner la construction du modèle avec des expérimentations relativement sophistiquées. Par ailleurs, les MFA ne peuvent gérer certains motifs métaboliques, comme les voies métaboliques parallèles ou les réactions réversibles [Wiechert 2001].

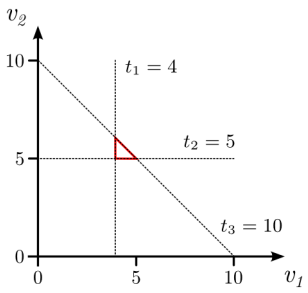


Figure 6.9. – Espace des solutions d'un système surdéterminé. Le système contraint par les égalités $t_1 = 4$, $t_2 = 5$ et $t_3 = 10$ n'a pas de solution. Le triangle rouge représente la zone où les contraintes sont en contradiction.

6.2. Gestion de la dynamique

Du fait de l'hypothèse de l'état stationnaire, la dynamique du système ne peut être directement prise en compte dans un MBC. Dans cette section, je présente deux familles d'approche qui permettent de prendre en compte la dimension dynamique. Une classification différente des approches dynamiques pour les MBC, plutôt centrée sur les modèles surdéterminés, est donnée dans Antoniewicz et coll. [Antoniewicz 2013].

6.2.1. Notion de succession d'états stationnaires

La dynamique est intégrée dans les MBC en considérant que le système passe par une succession d'états stationnaires. Cette démarche implique (i) de « discrétiser » l'écoulement du temps en périodes de temps, et (ii) de considérer que les flux restent stables pendant chaque période. En conséquence, les changements de vitesse ont lieu au moment de la transition entre une période p_i et une période suivante p_{i+1} (figure 6.10).

Il n'y a pas de consensus quant à la durée, c'est-à-dire l'intervalle de temps, pendant lequel un unique état stationnaire est représentatif de l'état réel du système. D'un point de vue biologique, la stabilité du métabolisme dépend grandement des conditions expérimentales et environnementales. Par exemple, une culture bactérienne cultivée dans un chemostat avec renouvellement du milieu peut maintenir « indéfiniment » une phase de croissance exponentielle, phase pendant laquelle la répartition et la vitesse des flux restent stables. À l'inverse, une culture en milieu non renouvelé induira diverses phases de croissance, et potentiellement différents états métaboliques, dont les durées dépendront par exemple des concentrations initiales en substrats et du déclenchement de phénomènes de régulation. D'un point de vue technique, l'espace des solutions est souvent contraint au niveau des flux d'entrées/sorties du système. Ces contraintes étant déterminées à l'aide de mesures expérimentales des concentrations extracellulaires, la durée de chaque période p_i d'un MBC dynamique peut dépendre de la fréquence à laquelle ces mesures ont été réalisées.

Bien que cette démarche « succession d'états stationnaires » simplifie fortement la réalité biologique, elle permet d'obtenir des résultats pertinents dans de nombreuses études, dont certaines seront citées au fil de cette section.

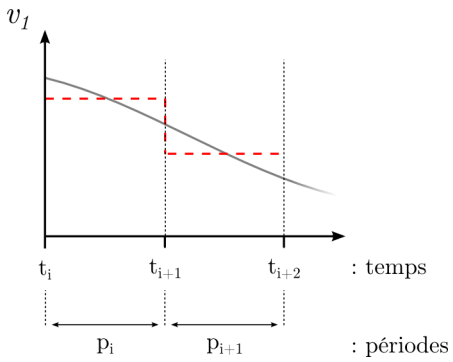


Figure 6.10. – Discrétisation du temps et dynamique des flux dans un MBC. Le graphique représente la vitesse d'un flux v_1 au fil du temps. L'écoulement du temps est discrétisé en périodes : l'intervalle de temps entre t_i et t_{i+1} est par exemple modélisé par la période p_i. La courbe grise représente la vitesse réelle du flux. Les droites de couleur rouge représentent la vitesse modélisée du flux v_1 pendant chaque période.

6.2.2. Dynamique par succession de périodes indépendantes

L'utilisation de contraintes sur les flux d'entrées/sorties (flux E/S) est un moyen simple de contraindre le potentiel métabolique d'un MBC. Dans le cadre d'un MBC dynamique, ces contraintes sont le plus souvent déduites à partir des concentrations en métabolites extracellulaires et en biomasse, mesurées expérimentalement à différents temps de la culture.

La modélisation de la dynamique d'un MBC par succession de cartes indépendantes se fait en deux grandes étapes :

1. Estimation des contraintes sur les flux E/S à différentes périodes de temps. Pour chaque période de la culture, il s'agit d'estimer les vitesses de consommation (ou de production) des métabolites extracellulaires et de la biomasse. Ces vitesses permettent de déduire des jeux de contraintes sur les flux E/S qui seront spécifiques de chaque période.
2. Analyse indépendante de chaque période. Pour chaque période : le jeu de contraintes approprié est appliqué sur le MBC et l'état du système est étudié par des techniques classiques (par exemple en utilisant la méthode FBA).

Estimation des contraintes sur les flux E/S. Afin de calculer les contraintes à chaque période de temps, différentes méthodes sont possibles (figure 6.11).

Une première méthode consiste à diviser l'ensemble de la culture en quelques phases distinctes. Pour chaque phase de culture, les flux E/S moyens sont estimés (figure 6.11, partie a). On obtient ainsi un jeu de contraintes spécifiques pour chacune des phases de culture.

Dans une deuxième méthode, la culture n'est pas divisée en phases. Au contraire, on souhaite ici être capable de calculer les flux E/S à tout instant de la culture. Pour cela, l'évolution continue de la concentration de chaque métabolite et de la biomasse est approximée à partir des points de mesure (figure 6.11, partie b). Les approximations les plus rencontrées sont les interpolations linéaires (segment entre deux points successifs), les régressions polynomiales (fonction polynomiale unique, sur l'ensemble des points), ou les régressions par courbes *splines* (fonction polynomiale par morceaux). À partir de ces courbes de concentrations en métabolites extracellulaires et

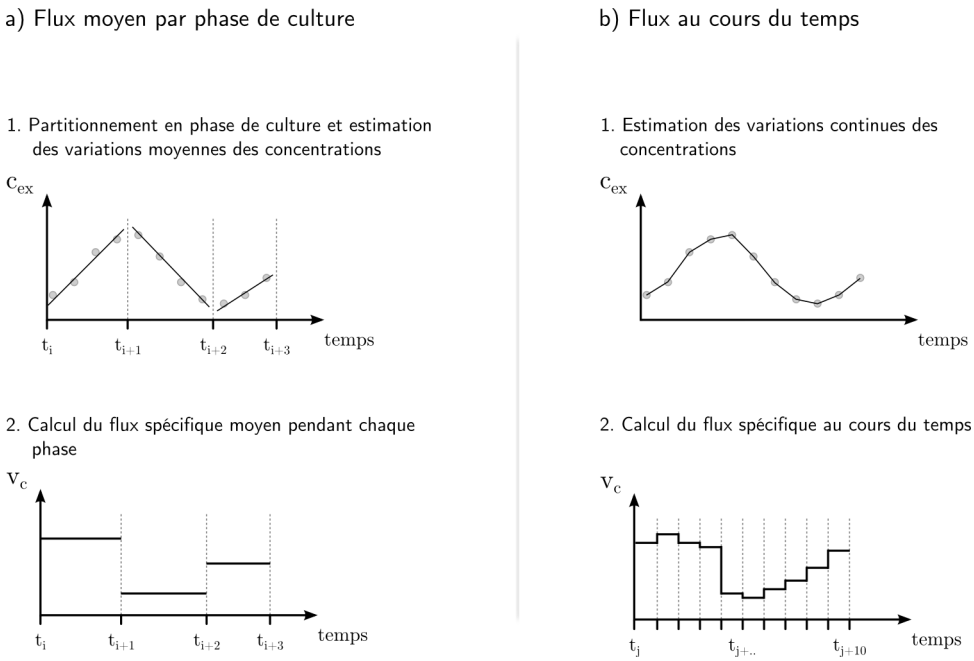


Figure 6.11. – Méthodes pour estimer les contraintes sur les flux E/S à partir des concentrations en métabolites extracellulaires mesurées expérimentalement. **a)** Calcul des flux spécifiques moyens par phase de culture. (1) Le temps de culture est partitionné en quelques phases (délimitées par des pointillées). Pour chaque phase, les concentrations du métabolite extracellulaire c_{ex} (représentées par des points) sont utilisées pour estimer une variation moyenne de concentration (droites). (2) Pour chaque période, la variation de concentration est utilisée pour calculer la vitesse du flux spécifique v_c (droites horizontales). **b)** Calcul des flux au cours du temps. (1) Les mesures de concentrations du métabolite extracellulaire c_{ex} (représentées par des points) sont utilisées pour estimer la variation continue du métabolite au cours du temps (segments de droite). (2) Le temps de culture est discrétisé en périodes (délimité par des pointillés), et le flux spécifique est estimé pour chaque période (segments de droites).

en biomasse, les dérivées des concentrations sont utilisées pour calculer les flux E/S pour différents intervalles de temps de la culture.

Ces deux méthodes ont par exemple été utilisées par Niklas et coll. pour étudier la culture d'une lignée cellulaire humaine [Niklas 2011]. Avec la première méthode, les auteurs divisent les 230 heures de culture en trois phases caractéristiques. Avec la deuxième méthode, ils estiment l'évolution continue des concentrations extracellulaires à l'aide de courbes *spline*, puis ils calculent les flux E/S pour des intervalles de temps d'environ 2,5h. Du fait du nombre de périodes de temps plus important, la seconde méthode permet d'obtenir une modélisation de la dynamique plus complète. De plus, elle ne nécessite pas de diviser *a priori* la culture en phases.

Analyse de chaque période. Quelle que soit la méthode utilisée pour calculer les contraintes, à chaque période de culture correspond un espace des solutions qui est contraint de manière « période spécifique ».

Je qualifie cette approche de « succession de périodes indépendantes » car le résultat d'une analyse (la ou les répartitions de cartes calculées pour une période de temps donnée) pour une période donnée n'a pas d'influence sur les contraintes appliquées aux autres périodes.

En conséquence, chaque espace des solutions peut être analysé indépendamment par des techniques classiques comme la méthode FBA [Calik 2011], la méthode FSA [Llaneras 2007b], ou encore par la méthode MFA lorsque le système est (sur-)déterminé [Niklas 2011].

6.2.3. Succession de périodes indépendantes et prise en compte de la variabilité

Afin de prendre en compte la variabilité des mesures expérimentales — ou l'incertitude liée aux dispositifs de mesure utilisés — dans un modèle dynamique, les contraintes sur les flux E/S associées à chaque période de temps peuvent être exprimées par des contraintes d'inégalité. Par exemple, les contraintes minimales et maximales du flux v_c qui consomme le composé extracellulaire c pendant la période p_i peuvent être définies par (figure 6.12, p. 123) :

$$v_{c,min,p_i} \leq v_{c,p_i} \leq v_{c,max,p_i}$$

$$\text{avec} \quad \begin{cases} v_{c,min,p_i} = \dfrac{c_{max,t_{i+1}} - c_{min,t_i}}{t_{i+1} - t_i} \\ v_{c,max,p_i} = \dfrac{c_{min,t_{i+1}} - c_{max,t_i}}{t_{i+1} - t_i} \end{cases}$$

où :

- v_{c,min,p_i} et v_{c,max,p_i} représentent respectivement les contraintes minimales et maximales du flux v pendant la période p_i ;
- c_{min,t_i} et c_{max,t_i} représentent respectivement les bornes minimales et maximales de l'intervalle de confiance des concentrations du composé extracellulaire c au temps t_i ;
- $c_{min,t_{i+1}}$ et $c_{max,t_{i+1}}$ représentent respectivement les bornes minimales et maximales de l'intervalle de confiance des concentrations du composé extracellulaire c au temps t_{i+1} ;

Cependant, cette manière de prendre en compte la variabilité conduit mathématiquement à estimer des cartes qui génèrent potentiellement des résultats incohérents avec les concentrations mesurées. Ainsi, lors de l'analyse de la période p_i, si la répartition des flux retenue est une carte avec une valeur de flux $v_{c,p_i} = v_{c,min,p_i}$, alors cette carte est uniquement valable dans la situation où la concentration du composé c au temps t_i est $c_{t_i} = c_{min,t_i}$. Pour toutes les autres concentrations possibles de c_{t_i} (comprises dans l'intervalle $c_{min,t_i} < c_{t_i} \leq c_{max,t_i}$), la consommation de c à la vitesse v_{c,min,p_i} conduit le système à prédire une concentration de c au temps t_{i+1} qui est en dehors de l'intervalle de confiance (figure 6.13, p. 123, partie a).

Si l'on se place maintenant dans le contexte de deux périodes successives, les contraintes d'inégalité sur les flux E/S ne permettent pas de garantir que la succession des deux cartes solutions (une pour chaque période) soit réalisable, et ce même si chaque carte considérée séparément génère des résultats cohérents. Une telle situation est illustrée par la figure 6.13 (partie b).

J'ai ici pris pour exemple la situation où une seule carte est utilisée afin d'analyser l'espace des solutions (carte qui serait par exemple obtenue par la méthode FBA). Cependant, ces limites liées à la prise en compte de la variabilité sont aussi valables pour les résultats générés par d'autres méthodes, comme les cartes échantillonnées par une approche MCMC.

Par ailleurs, il ne faut pas oublier que les systèmes métaboliques sont étudiés dans la grande majorité des cas en relativisant les flux par rapport à la concentration de biomasse (flux spécifique, exprimé en mmol/h/g. de biomasse). La biomasse étant considérée comme proportionnelle à la quantité de cellules, les flux spécifiques expriment la vitesse des réactions par nombre moyen de cellules. Ils permettent ainsi de comparer les flux à différentes périodes en s'affranchissant du fait que la concentration de biomasse peut être différente selon la période. Les contraintes minimale et maximale sur la vitesse du flux spécifique v_{c,p_i} pendant la période p_i peuvent en fait être calculées selon :

$$v_{c,min,p_i} = \frac{c_{max,t_{i+1}} - c_{min,t_i}}{t_{i+1} - t_i} / \boldsymbol{bio_{t_i}}$$
$$v_{c,max,p_i} = \frac{c_{min,t_{i+1}} - c_{max,t_i}}{t_{i+1} - t_i} / \boldsymbol{bio_{t_i}}$$

où :

— bio_{t_i} représente la concentration de biomasse au début de la période p_i définie sur l'intervalle de temps t_i à t_{i+1}. On considère ici la concentration initiale de biomasse comme représentative de la concentration de biomasse pendant toute la période p_i.

Dans cette situation, la prise en compte de la variabilité sur les mesures de biomasse apporte un degré supplémentaire de complexité, puisque la valeur de chacun des flux du système modélisé dépendra de la concentration de biomasse.

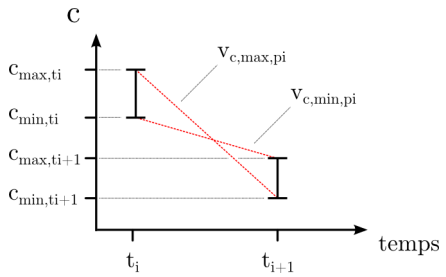


Figure 6.12. – Prise en compte de la variabilité expérimentale dans le calcul des contraintes sur les flux E/S. Les intervalles de concentration du métabolite c aux temps t_i et t_{i+1} sont représentés en noir. Les vitesses de consommation qui correspondent aux contraintes minimales et maximales sur le flux v_c sont représentées par des droites en pointillés de couleur rouge.

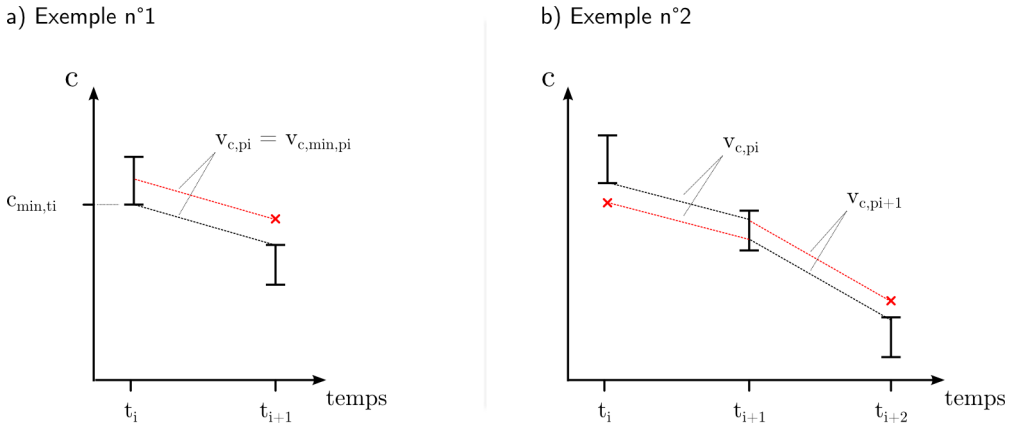


Figure 6.13. – Exemple de flux prédits qui conduisent potentiellement à des concentrations incohérentes avec les mesures expérimentales. **a)** Le flux prédit v_{c,p_i} n'est cohérent que si l'on considère que la concentration initiale du métabolite c_{t_i} est égale à c_{min,t_i} (situation figurée en noir). Dans tous les autres cas, le flux prédit conduit à une concentration $c_{t_{i+1}}$ qui est supérieure à l'intervalle de confiance établi au temps t_{i+1}. **b)** Pris séparément, les flux v_{c,p_i} et $v_{c,p_{i+1}}$ correspondent chacun à une évolution de la concentration du métabolite c qui est cohérente avec les intervalles de confiance (situation présentée en noir). Cependant, la succession de ces deux flux ne peut pas conduire à des concentrations cohérentes avec les mesures expérimentales : le flux v_{c,p_i} n'est cohérent que si l'on considère que $c_{t_{i+1}}$ est *élevée*, tandis que le flux $v_{c,p_{i+1}}$ n'est cohérent que si $c_{t_{i+1}}$ est *faible*.

6.2.4. Dynamique par succession de périodes dépendantes

Dans cette sous-section, je vais aborder l'approche *dynamic flux balance analysis* (DFBA) [Varma 1994, Mahadevan 2002]. À ma connaissance, cette approche est la seule qui permette à la fois (*i*) d'instaurer un lien de dépendance entre les périodes de la culture modélisée et (*ii*) de gérer le fait que le système soit sous-déterminé.

Principe de l'approche DFBA. L'approche DFBA est utilisée pour simuler l'évolution d'un système métabolique au cours du temps. Pour cela, on considère que le système métabolique évolue dans un environnement clos (figure 6.14), et les ressources initialement disponibles dans l'environnement sont prédéfinies (concentrations en métabolites externes et en biomasse). Le principe de la méthode s'apparente ensuite à ce qui peut se passer dans un erlenmeyer lors d'une culture bactérienne : au fil du temps, le système consomme les métabolites extracellulaires, forme de la biomasse, et produit éventuellement des métabolites secondaires.

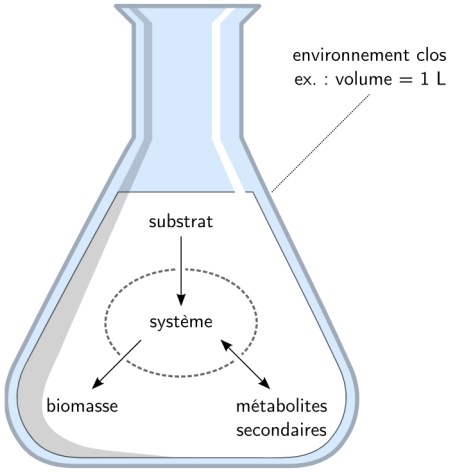


Figure 6.14. – Manière dont le système et l'environnement sont considérés en DFBA. Le système (cercle en pointillé) est considéré comme évoluant dans un environnement clos (figuré par l'erlenmeyer). Les métabolites présents à l'intérieur du système sont considérés comme étant à l'état stationnaire, tandis que les métabolites (et la biomasse) présents dans l'environnement peuvent varier.

Pour simuler cette dynamique, l'algorithme de la méthode consiste à générer une succession de cartes de flux, où chaque carte représente une période de temps, et est générée en utilisant la méthode FBA. À chaque de fin de période, la carte générée détermine les composés de l'environnement extérieur qui ont été consommées (métabolites) et produits (biomasse et métabolites) par le système. Les quantités de ressources présentes dans le milieu extérieur sont ensuite actualisées, et une nouvelle période est simulée. Ce faisant, une dépendance entre les périodes est instaurée par le fait que les ressources disponibles au début d'une période p_i dépendent de la carte de flux de la période précédente. Le critère d'arrêt de la simulation est, par exemple, la pénurie de substrat, ou le fait que la simulation a atteint une certaine durée.

Pour une période donnée p_i, définie sur l'intervalle de temps Δt allant de t_i à t_{i+1}, l'algorithme est le suivant :

1. Calcul des contraintes « période spécifique ». La concentration du substrat au début de la période, notée c_{t_i}, et la concentration en biomasse, notée bio_{t_i}, sont utilisées pour déterminer la vitesse maximale v_{c,max,p_i} du flux spécifique de consommation pendant la période p_i (à volume constant) :

$$v_{c,max,p_i} = c_{t_i}/bio_{t_i}/\Delta t$$

2. Calcul d'une carte de flux avec la méthode FBA. Pour chaque période, deux types de contraintes sont appliquées : des contraintes constantes qui sont valables, quelle que soit la période, et des contraintes « période spécifique » calculées à l'étape précédente. La fonction *objectif* optimisée est typiquement la production de biomasse, que l'on veut maximiser.

3. Actualisation des concentrations en métabolites externes et en biomasse dans le milieu extérieur. La carte de flux obtenue à l'étape précédente permet de connaitre les valeurs prédites du flux de consommation du substrat v_{c,p_i} et du flux de production de biomasse v_{bio,p_i}. Ces valeurs de flux sont utilisées pour calculer les concentrations à la fin de la période (à volume constant) :

$$\begin{aligned} c_{t_{i+1}} &= c_{t_i} - v_{c,p_i} \cdot bio_{t_i} \cdot \Delta t \\ bio_{t_{i+1}} &= bio_{t_i} + v_{bio,p_i} \cdot bio_{t_i} \cdot \Delta t \end{aligned}$$

Extensions et applications. Cette approche a été initialement développée par Varma et Palsson afin de simuler la croissance d'une culture d'*E. coli* dans différentes conditions environnementales [Varma 1994]. Le principe de la méthode a ensuite été repris par Mahadevan et coll. [Mahadevan 2002]. Dans ce second travail, les auteurs implémentent deux approches différentes du DFBA : l'approche par optimisation statique (*static optimization approach*, SOA), et l'approche par optimisation dynamique (*dynamic optmization approach*, DOA).

La SOA correspond à l'algorithme que je viens de présenter : la production de biomasse est optimisée de manière locale, à chaque période de temps. Cet algorithme s'apparente à un algorithme « glouton » où, à chaque étape, le maximum local est recherché, sans certitude que la quantité de biomasse produite à la fin de la simulation soit l'optimum global. La DOA optimise quant à elle la production de biomasse par programmation dynamique : l'optimisation de la production de biomasse est reformulée sous la forme d'un problème non linéaire. La solution est alors évaluée en une fois en considérant l'ensemble du temps de simulation. Cette seconde approche implique un nombre de variables bien plus important que la méthode SOA, ce qui limite l'usage de la DOA à des réseaux de petite taille. À titre d'exemple, le métabolisme central d'*E. coli* est décrit par 4 réactions dans [Mahadevan 2002], ce qui nécessite l'évaluation de 204 variables. Remarquons ici que, d'après les simulations réalisées dans [Mahadevan 2002], une optimisation instantanée (approche « gloutonne ») de la production de biomasse génère des résultats plus proches des mesures expérimentales.

Le principe de l'approche DFBA à donné lieu à de nombreuses applications. Citons notamment : [Sainz 2003] chez la levure (méthode SOA-DFBA), [Luo 2006] où une extension de l'approche DOA-DFBA est utilisée pour étudier le métabolisme énergétique du myocarde, [Lee 2008] où les auteurs modélisent simultanément un métabolisme et un réseau de régulation chez *E. coli*, [Zhuang 2011] pour modéliser la compétition entre deux populations bactériennes. Notons aussi le développement de la méthode *Possibilistic-FBA* permettant d'estimer sous la forme d'intervalles l'étendue des valeurs de flux qui conduisent à des solutions sous-optimales [Llaneras 2012].

Considération sur l'utilisation de l'approche DFBA. La méthode DFBA permet de simuler la dynamique d'un système, et cette dynamique est ensuite comparée avec des résultats expérimentaux. Cet usage est différent des approches présentées jusqu'ici, où les résultats expérimentaux étaient

utilisés afin de contraindre le système et prédire la ou les répartitions de flux les plus cohérentes.

La DFBA est rendue possible grâce à l'utilisation d'une fonction *objectif* qui permet de choisir une répartition de flux parmi tous les comportements potentiellement réalisables par le système. Cependant, comme discuté plus tôt dans ce chapitre, le « comportement » qui serait optimisé par le système métabolique n'est pas toujours connu, en particulier lorsque l'organisme modélisé est cultivé dans des conditions non optimales de croissance.

Chapitre 7

Dynamique d'un modèle à base de contraintes : approche par échantillonnage des trajectoires solutions

Dans le chapitre précédent, nous avons vu que la dynamique du métabolisme était intégrée dans un MBC en (*i*) discrétisant le temps en périodes et (*ii*) en considérant que chaque période représente un intervalle de temps pendant lequel le système est stable (sous-section 6.2.1, p. 116). À partir de cette base commune, on rencontre ensuite deux types d'approche.

La premier type d'approche, que je qualifie de « dynamique par succession de périodes indépendantes », consiste à analyser de manière indépendante chaque période de temps (sous-section 6.2.2, p. 118). Cette approche a pour avantages d'être relativement simple à mettre en place et de permettre d'utiliser les techniques d'analyse classiquement rencontrées pour étudier les MBC (section 6.1, p. 102). Cependant, lorsque de la variabilité est introduite sur les flux d'E/S, ce type d'approche conduit à surestimer les possibilités du système (sous-section 6.2.3, p. 121).

Le second type d'approche, que je qualifie de « dynamique par succession de périodes dépendantes », est représenté par la méthode DFBA dans le cadre des MBC sous-déterminés. Elle propose de simuler la dynamique au cours d'une succession de périodes, où le « potentiel métabolique » du système pour une période donnée dépend de ce qui s'est passé pendant les périodes précédentes (sous-section 6.2.4, p. 124). Cependant, cette méthode ne peut être utilisée que lorsque le « but » optimisé par le réseau métabolique est connu.

Dans ce chapitre, je présente sous la forme d'un article l'approche que j'ai développée pendant ma thèse. La motivation principale à l'origine de cette approche est de pouvoir (*i*) considérer la variabilité des flux d'E/S du système tout en étant (*ii*) capable de prédire une dynamique « réaliste » —

autrement dit de prédire une succession d'états du système qui soit cohérente avec les concentrations mesurées expérimentalement — et ce (*iii*) sans nécessiter d'hypothèse sur l'optimalité du système.

Afin d'y parvenir, l'approche développée propose de générer une population de « trajectoires solutions », où chaque trajectoire représente une dynamique — une succession de cartes solutions — qui est cohérente avec les mesures expérimentales des concentrations. Pour chaque trajectoire et chaque période, « l'histoire » du système pendant les périodes précédentes est prise en compte pour contraindre l'espace des solutions, et une carte est choisie parmi toutes celles possibles en utilisant une approche d'échantillonnage de l'espace des solutions. L'étude de la population des trajectoires générées permet ensuite d'évaluer les répartitions et les valeurs de flux les plus fréquemment prédites pour chaque période de temps modélisée.

La méthode a été utilisée pour étudier le fonctionnement du métabolisme central de la bactérie *Corynebacterium glutamicum* lors d'une croissance en condition de limitation en biotine. Remarquons ici que la limitation en biotine entraîne une courbe de croissance arithmétique (le nombre total de bactéries augmente de façon linéaire au cours du temps), et que cette courbe de croissance ne peut pas être reproduite par une approche DFBA avec l'hypothèse de l'optimisation de biomasse.

L'article a été soumis au journal BMC System Biology (section méthodologie), sous le titre

« *A constraint-based model for time course studies of under-determined metabolic networks : a Monte Carlo sampling approach* »

avec pour auteurs moi-même, Armel Guyonvarch et Bruno Bost (par ordre d'apparition). Les annexes de l'article sont données dans l'annexe B (p. 253). L'annexe C donne, en complément, les schémas qui correspondent au réseau métabolique étudié (p. 261).

chapitre x, section x.y, p. zzz

Les parties de l'article qui sont développées dans les chapitres suivants sont signalées par une note en marge, comme celle ci-contre.

7.1. Background

Biological networks involve numerous interactions between biological entities. As knowledge of complex biological systems rapidely increases, dedicated methods are requiered to help our understanding and prediction of the behavior of such systems.

Constraint-Based Models (CBMs) are widely used to study metabolic networks (see [Bordbar 2014] for a detailed review). Compared to other methods, these models require few data to describe a given metabolic network. This allows use of CBMs for modeling large metabolic networks, up to the level of the "full genome".

This is rendered possible, since CBMs are based on the assumption that the modeled metabolic networks are in a steady-state, i.e. the concentration of metabolites within the network are assumed to be constant. As a consequence, the kinetic parameters of reactions are not required to describe the behavior of the network.

CBMs allow the calculation of flux distributions or flux maps that describe the partitioning of matter fluxes between the different pathways of the network. Each flux map represents a theoretically possible steady-state of the network, taking into account the network topology and some constraints on reaction rates.

Information that is commonly used as a constraint in CBMs, is the overall knowledge of reactions involved in the network and the production or consumption rates of external metabolites. In most cases, these external metabolites are extracellular compounds that are produced or consumed by the cells, and the corresponding fluxes are easily measured experimentaly.

The method we present in this paper is, to our knowledge, the first that combines a CBM method and the sampling of the solution spaces, in order to predict the dynamics of metabolic networks. Due to this combination, our method allows a non-biased prediction of the flux distribution dynamics in an under-determined metabolic network, while taking into account the incertitude coming from experimental results.

Taking experimental variability into account

Variability is inherent to biological experiments: the amount of a biological entity that is measured several times will always present some variability. This variability is due to accuracy thresholds of measurement devices. Moreover, when measurements are made on independent replicates, the observed variability is also due to variability of the individual behavior of cells that compose the biological system.

For a given under-determined metabolic network, if the constraints are not sufficient to determine a unique solution flux map, the solution space contains several different flux maps. Moreover, the experimental variability is expressed in the model as inequality constraints on unknown fluxes, and thus this cannot permit us to reduce the under-determination of the network.

A method related to CBMs, the Flux Balance Analysis method (FBA), uses an objective function to calculate a unique flux map that optimizes a specified biological function [Kauffman 2003]. Besides the fact that the choice of the objective function may be sometimes difficult, this method does not give access to the full range of possible flux maps for the network studied under some specific conditions, and does not allow quantification of the incertitude around the values calculated in the optimal flux map.

Some other methods can give insights on the variability of fluxes. The Flux Variability Analysis method (FVA) evaluates the range of possible values for each flux in the network [Mahadevan 2003]. Nevertheless, this method does not allow the estimation of the most probable value for each flux, but only the minimal and maximal values. In addition, dependencies between fluxes are not considered during the calculation of the flux ranges. Probabilistic informations can be introduced as described by Antoniewicz *et al.* [Antoniewicz 2006]. This allows the estimation of a confidence interval for each flux. However, their algorithm works only when sufficient relevant information is available, so that all the fluxes in the system are unambiguously determined (the system is fully determined). This prevents the use of uncertainty for constraints. Llaneras *et al.* developed the Possibilistic Metabolic Flux Analysis method (Possibilistic MFA), which allows, for each flux in a fully determined system, calculation of the distribution of "possible" values [Llaneras 2009]. Although this method can be used for under-determined systems, the distribution of flux values that are gen-

erated are uniform distributions, thus one could not distinguish the most probable flux values amongst all "possible" flux values.

To our knowledge, only methods using sampling of possible flux maps (sampling of the solution space, [Wiback 2004, Price 2004b, D'Huys 2012]) are helpful to estimate probabilities of occurence for flux values in under-determined metabolic systems.

Taking time into account

Another important phenomenon to consider are the dynamics of the metabolic system. The study of transient biological phenomena in the context of CBMs requires the time-course dynamics of the flux maps to be predicted.

During recent years, several methods exploiting CBMs have been proposed for predicting the dynamics of a metabolic system when taking into account uncertainty due to experimental measurements. These methods assumed that the system was in a pseudo-steady-state. Leighty *et al.* thereby developed the Dynamic MFA method, which allows the prediction of dynamic flux maps in fully determined systems [Leighty 2011]. Llaneras *et al.* proposed two methods for modeling the dynamics of an under-determined system: the flux spectrum approach [Llaneras 2007b] and an extension of the Possibilistic MFA method [Llaneras 2012]. These two methods also allowed the investigator to take into account uncertainties. However these two methods, when applied to under-determined systems, only predicted uncertainty in the form of an interval of possible values for each flux, without giving an estimation of the most probable value in each interval.

The algorithm of the method we propose is exposed in the first two parts of the "results and discussion" section. In a third part, we present the features of the method through its application in a study of the core metabolism of *Corynebacterium glutamicum* during biotin-limited growth. The effect of experimental variability is discussed, as well as the effect of taking into account dependencies between successive time periods that are modeled.

7.2. Results and discussion

7.2.1. General considerations

Constraint-Based Models

Metabolic networks can be represented with a stoichiometric matrix S. The rows of S correspond to the c internal compounds and the columns to the r reactions. The flux value for each reaction can be represented as a vector v of length r. The mass conservation of the system can then be expressed as follows:

$$S \cdot v = \frac{dc}{dt}$$

where c is the vector of the concentration of the internal compounds.

Constraint-based models (CBM) assume that internal compounds in the system are at the metabolic steady-state. A metabolic steady-state corresponds to a state in which all the concentrations of internal compounds of the system remain constant. Under the metabolic steady-state assumption, the system can then be formulated as follows:

$$S \cdot v = 0 \tag{7.1}$$

When the number of linearly independent columns (reactions) is greater than the number of rows (compounds), the system is said to be underdetermined. In such situations, there is an infinite number of flux vectors v possible in order to satisfy equation (7.1). These valid vectors live in a so-called "solution space". A solution space is defined in a hyper-plane of dimension r, where each dimension corresponds to a reaction flux. A point of this solution space corresponds to a unique set of flux values, giving one flux vector v that solves equation (7.1).

A typical method that reduces the size of a solution space is the use of constraints on reachable flux values. Several types of constraint can be used:

— *Flux range — Inequality constraints.*

$$v_{min} \leq v \leq v_{max} \tag{7.2}$$

Each flux has an upper and a lower bound, each of which define the maximal and minimal possible values of the flux, respectively. These constraints can be used when some biological knowledge is available to describe the system's capabilities. For instance, irreversibility of a given reaction can be set as v_{min} equals 0.

— *Flux value — Equality constraints.*

$$v = v_{const} \tag{7.3}$$

Fluxes can be set to a unique value. This constraint is commonly used when some reactions are known to be inactive under specific circumstances (for instance: flux is set to 0).

Biological knowledge allows us to define inequality (7.2) and equality constraints (7.3). Applying these constraints reduces the size of the solution space that was initially defined by equation (7.1). The constrained solution space can then be formulated as follows:

$$\begin{aligned} & S \cdot v = 0 \\ \text{subject to } & \begin{cases} v_{min} \leq v \leq v_{max} \\ v = v_{const} \end{cases} \end{aligned}$$

Sampling the solution space

A way to characterize the solution space is to explore it with sampling methods that randomly choose solution flux maps. Furthermore, sampling a high number of randomly solution flux maps is a way to estimate the marginal distributions of values for each flux.

chapitre 8, section 8.1, p. 164

Sampling of the solution space is carried out with the "mirror" algorithm developed by Van den Meersche *et al.* [Van den Meersche 2009]. This method is implemented in the "xsample" function from the "limSolve" package [Soetaert 2009] that runs under the R software [R Core Team 2014].

The "mirror" algorithm is iterative. The choice of a starting solution is mandatory to initiate the exploration of the solution space. We used a "central solution" as a starting point. For a given network and a given set of constraints, the central solution is calculated as follows:

1. For each flux k, the solutions v_{kmin} and v_{kmax} are calculated. These minimize and maximize the flux k, respectively.
2. For each flux j, the mean value of the flux F_j is calculated such as:

$$F_j = \frac{1}{r} \sum_{k=1}^{r} \left(\frac{v_{kmin,j} + v_{kmax,j}}{2} \right)$$

where $v_{kmin,j}$ et $v_{kmax,j}$ represent the values of flux j obtained by minimization and maximization of flux k, respectively.

Discretization of an experiment into time periods

Studying the functionning of a metabolic system over time is commonly done through time-course measurements of extracellular metabolites and biomass concentrations.

Below and throughout this paper, the period between 2 successive times at which extracellular concentrations have been measured during the experiment is called a "time period".

If we consider that the system is at steady-state during each time period, the global dynamics of the system can be studied as a sequence of steady-states. Under this assumption, each time period can be described by a CBM. Note that the topological description of the system remains the same for all the time periods (equation 7.1), whereas the set of constraints could be different for each time period.

Such assumptions have been used with success in previous studies. Varma and Palsson built-up the "dynamic FBA" in order to predict cellular growth [Varma 1994]. Lequeux *et al.* developed the "dynamic MFA" method to study the metabolic shift between two substrate limitations in *Escherichia coli* [Lequeux 2010].

However, to our knowledge, CBMs and temporal discretization of experiment have never been combined with sampling of the solution space to study the dynamics of a metabolic system.

7.2.2. Algorithm of the method

By combining temporal discretization of the experiment, CBM and sampling of the solution space, our method gives access to feasable sequences of solution flux maps. A sequence of solution flux maps is a sorted set of flux maps, with a solution map for each time period. Feasability is introduced by simply keeping sets of maps that predict concentrations of extracellular compounds that are coherent with experimental measurements.

In order to identify feasable sequences of flux maps, our method works iteratively:

1. At the beginning of each time period p_i, defined as the time interval from time t_i to time t_{i+1}, the constraints on input and output fluxes of the system are calculated by using the experimental measurements (Figure 7.1, panels a and b).
2. The constrained solution space is sampled (Figure 7.1, panel c).
3. One flux map is randomly chosen from amongst the sample of the solution space (Figure 7.1, panel d).
4. From (i) concentrations of extracellular compounds at time t_i and (ii) the chosen flux map, the concentrations of extracellular compounds at time t_{i+1} are calculated (Figure 7.1, panel e).
5. Back to step number 1: the calculated concentrations at time t_{i+1} are used to calculate specific constraints for the next time period p_{i+1}, defined as the interval from time t_{i+1} to time t_{i+2} (Figure 7.1, panel f).

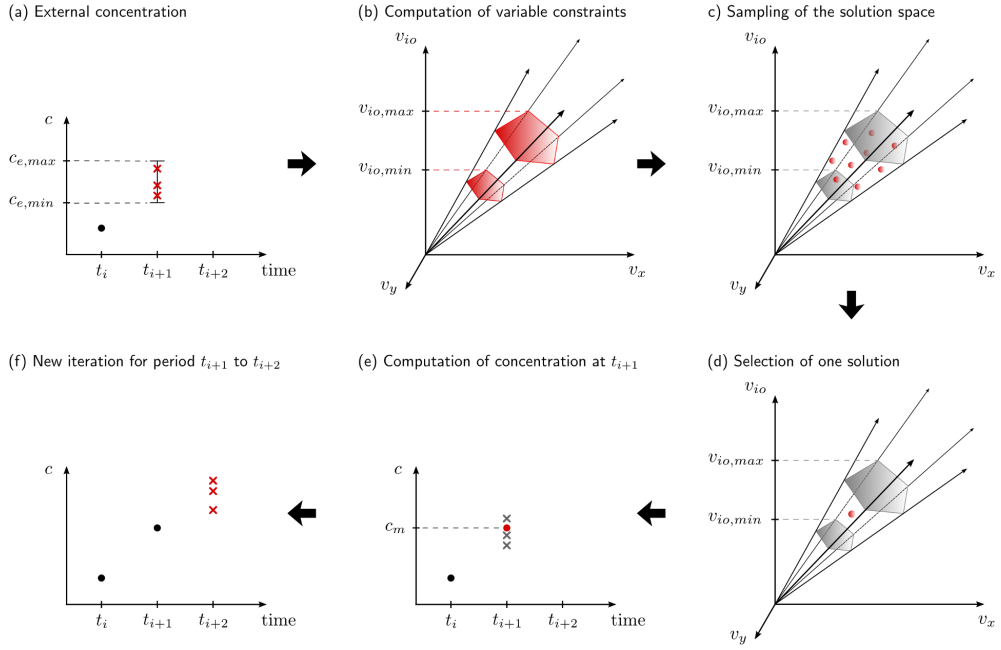


Figure 7.1. – Steps for modeling a time period p_i, defined as the interval form the time to time. **(a)** The extracellular concentrations c_{m,t_i} (black filled circle, •) predicted by the model for the time t_i, and the extracellular concentrations $c_{e,t_{i+1}}$ (red cross, ×) obtained for the experimental measurements at time t_{i+1}, are known. **(b)** The "variable" constraints on the flux v_{io,p_i} that describe production or consumption of the extracellular compound c_e are calculated from the extracellular concentrations c_{m,t_i}, $c_{e,t_{i+1}}$ and the biomass concentration bio_{m,t_i}. The experimental variability allows to define the inequality constraints (v_{io,min,p_i}, and v_{io,max,p_i}, see equation 7.5). These contraints (red lines, - -) are applied to the solution space. **(c)** Solution flux maps are randomly selected by sampling the constrained solution space. **(d)** One of the sampled flux maps is kept for building the metabolic trajectory over time periods. **(e)** Knowing the concentration of the extracellular compound c_{m,t_i} and of the biomass concentration bio_{m,t_i} at time t_i, and the specific flux v_{io,p_i} for the time period p_i, the concentration $c_{m,t_{i+1}}$ is calculated (red filled circle, •). **(f)** Knowing the concentration $c_{m,t_{i+1}}$ predicted by the model for time t_i, and the concentrations $c_{e,t_{i+2}}$ experimentaly measured, the next time period p_{i+1} can be modeled.

Calculation of the constraints on input and output fluxes of the system

For a time period p_i, defined as the interval between times t_i and t_{i+1}, the mean $\bar{c}_{e,t_{i+1}}$ and the standard deviation $\sigma_{c_{e,t_{i+1}}}$ of the measurements of the extracellular concentration of the compound c_e at time t_{i+1} are estimated. From these statistics, an interval of "target" concentrations to be reached by the model at time t_{i+1} is calculated as follow:

$$\begin{aligned} c_{e,min,t_{i+1}} &= \bar{c}_{e,t_{i+1}} - d \cdot \sigma_{c_{e,t_{i+1}}} \\ c_{e,max,t_{i+1}} &= \bar{c}_{e,t_{i+1}} + d \cdot \sigma_{c_{e,t_{i+1}}} \end{aligned} \tag{7.4}$$

where:

— $c_{e,min,t_{i+1}}$ and $c_{e,max,t_{i+1}}$ represent the minimal and maximal values of the interval of target concentrations for the extracellular compound c_e at time t_{i+1}, respectively (namely $c_{e,min}$ and $c_{e,max}$ in Figure 7.1, panel b);

— d is a parameter used to modulate how to take the experimental variability into account.

This interval of target concentrations is converted into an interval of allowed values for the flux v_{io,p_i} that describes production or consumption of the extracellular compound c_e during the time period p_i from t_i to t_{i+1}:

$$\begin{aligned} & v_{io,min,p_i} \leq v_{io,p_i} \leq v_{io,max,p_i} \\ \text{with } & \begin{cases} v_{io,min,p_i} = \dfrac{c_{e,min,t_{i+1}} - c_{m,t_i}}{t_{i+1} - t_i} / bio_{m,t_i} \\ v_{io,max,p_i} = \dfrac{c_{e,max,t_{i+1}} - c_{m,t_i}}{t_{i+1} - t_i} / bio_{m,t_i} \end{cases} \end{aligned} \tag{7.5}$$

where:

— v_{io,min,p_i} and v_{io,max,p_i} represent the minimal and maximal constraints applied on flux v_{io} during time period p_i (v_{io,p_i}), respectively (see Figure 7.1, panel b);

— c_{m,t_i} and bio_{m,t_i} represent the concentrations of extracellular compound c and of biomass bio predicted by the model at the previous time t_i, respectively.

For each time period, inequality constraints on input and output fluxes of the system lead to a "time specific" constraint on the solution space. As some of these constraints can vary when considering various time periods, we call them "variable constraints" throughout this article. In contrast, the constraints applied on the system that remain constant for all the time periods are called "constitutive constraints" of the system.

Sampling the solution space for one time period

From a mathematical point of view, the sampling for a time period p_i is achieved with a solution space defined as:

$$\begin{aligned} & S \cdot v = 0 \\ \text{constrained with} \quad & \begin{cases} v_{const,min} \leq v \leq v_{const,max} \\ v = v_{const} \end{cases} \\ \text{and with} \quad & v_{io,min,p_i} \leq v \leq v_{io,max,p_i} \end{aligned}$$

where:

— S represents the topology of the system, given by its stoichiometric matrix;

— v_{const}, $v_{const,min}$ and $v_{const,max}$ represent the constraints of equality, lower inequality and upper inequality, respectively, that are constitutive of the system;

— v_{io,min,p_i} et v_{io,max,p_i} represent the constraints of lower and upper inequality applied on exchange fluxes, respectively ; these constraints may vary according to time periods.

Calculation of the concentration of extracellular compounds at time t_{i+1}

Knowing at time t_i the concentration of an extracellular compound c_{m,t_i}, the biomass bio_{m,t_i} and the value of the flux v_{io,p_i} that produces or consumes the compound c during the time period p_i, the concentration c_m at time t_{i+1} is calculated by integration as follows:

$$c_{m,t_{i+1}} = c_{m,t_i} + v_{io,p_i}.bio_{m,t_i}$$

First iteration of the method: time period p_1

A special treatment is performed for the first time period p_1. During this period, the concentrations of extracellular compounds at time t_1 cannot be predicted from a flux map predicted for the previous time period. To take into account the experimental variability in this case, the concentration at time t_1 for each extracellular compound and for biomass is randomly chosen in a fixed distribution. The choice of the shape of these starting distributions for each compound is considered as a parameter of the method.

Going back to previous periods when there is no solution for the period p_i

chapitre 8, section 8.3, p. 180

In some cases, the predicted extracellular concentrations at time t_{i+1} do not allow solution flux maps to be obtained for the time period p_{i+1}, defined as the interval from time t_{i+1} to time t_{i+2}. In these cases, another map is chosen from amongst those sampled for the time period p_i, and the predicted concentrations at time t_{i+1} are recalculated (see Figure 7.2).

Finally, a sequence of flux maps representing a possible "scenario", namely a trajectory that connects successive metabolic states that take place in the cells during the experiment, is obtained. Each flux map represents a combination of flux values that is not only a solution for the CBM at the corresponding time period p_i but also allows solution flux maps for all the subsequent time periods to be found. Each flux map is the result of a sampling of the solution space. The solution space is constrained by inequalities that are specific to the time period. These variable constraints are inferred from the concentrations predicted at time t_i and also from the experimental "target" concentrations at time t_{i+1}.

As a consequence, for a same metabolic system, several different metabolic scenarii can be predicted, when taking the experimental variability into consideration. Differences between these scenarii can been observed:

1. for the predicted concentrations of extracellular compounds that will have an effect on the solution flux maps for the subsequent time periods;
2. for the flux maps selected amongst the sampling set that will have an effect on the final extracellular concentrations during the modeled time period.

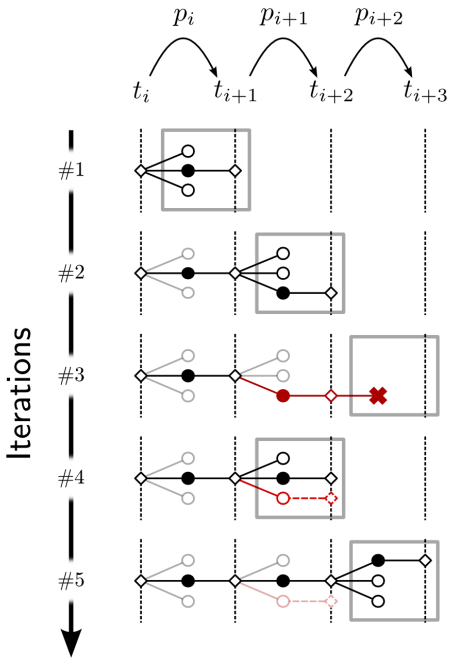


Figure 7.2. – Iterations and steps back during the building of a metabolic trajectory. The represented trajectory contains 3 time periods p_i (from ti to t_{i+1}), p_{i+1} (from t_{i+1} to t_{i+2}) and p_{i+2} (from t_{i+2} to t_{i+3}). (#**1**) The first time period p_i is built at the first iteration of the method: from a set of concentrations for an extracellular compound measured at time t_i (diamond, $\diamond$), the variable constraints are applied and solution flux maps (circles, $\circ$) are sampled, amongst which one is selected (black filled circle, $\bullet$) in order to extend the trajectory. The concentrations at time t_{i+1} are calculated taking into account the fluxes of the selected map for the time period p_i. (#**2**) The time period p_{i+1} is then produced during the second iteration of the method and a set of concentrations at time t_{i+2} is calculated. (#**3**) At the third iteration, no solution exists given the variable constraints calculated from the set of concentrations at time t_{i+2} (red cross, $\times$). As a consequence, the flux map selected for the time period p_{i+1} is labeled as not being able to give a solution (red filled circle, $\bullet$). (#**4**) At the fourth iteration, the method goes back to the previous time period p_{i+1} and selects another flux map among those sampled from the solution space at iteration #2. A new set of concentrations for the time t_{i+2} is calculated from the selected flux map (black filled circle, $\bullet$). (#**5**) The new set of concentrations at time t_{i+2} allows to find flux maps that are solution for the time period p_{i+2}: the trajectory is extended until the time t_{i+3}.

The production of a random set of metabolic trajectories give n flux maps for each time period, and therefore n values for each flux at each time period. With our method, the flux maps that constitute a metabolic scenario are linked by the existence of a solution for each time period until the last one. Thus, flux predicted distributions allows the analysis of the dynamics of the solution space more precisely and more realistically than with a method where the solution spaces associated to each time period are sampled independently.

7.2.3. Application to the core metabolism of *C. glutamicum*

To illustrate the usefulness of our method, we applied it to model the core metabolism of *C. glutamicum* during biotin-limited growth. Variable constraints on input/output fluxes of glucose, acetate, lactate, pyruvate as well as biomass were calculated from the extracellular concentrations measured in the study of Dietrich *et. al.* [Dietrich 2009]. Figure 7.3 presents the extracellular concentrations measured in this study. Description of the model and the parameters used for this application are detailed in the Methods section of this paper. Figure 7.4 (p. 146) shows a simplified topology of the *C. glutamicum* network we studied.

Figure 7.5 (p. 147) presents, for a subset of six "reporting" fluxes, the dynamics of specific fluxes (mmol/g cell dry weight/h) predicted by our method, as well as the total carbon uptake of the system. In order to compare the states of the system between different time periods, we normalized the specific fluxes with respect to the total carbon uptake ($flux_{carb}$). For each flux map, the value of each normalized flux $flux_{k,norm}$ was computed according to the formula:

$$\begin{aligned} flux_{k,norm} &= flux_{k,spe} \cdot NC_k \cdot 1/flux_{carb} \\ \text{with} \quad flux_{carb} &= \sum_{io=1}^{n} \left[max \begin{pmatrix} v_{io} \\ 0 \end{pmatrix} \cdot NC_{io} \right] \end{aligned} \tag{7.6}$$

where:

— NC_k represents the number of carbon atoms that are involved in the flux v_k;

— v_{io} represents a input/output flux of the system;

— NC_{io} represents the number of carbon atoms involved in the reaction corresponding to the flux v_{io}.

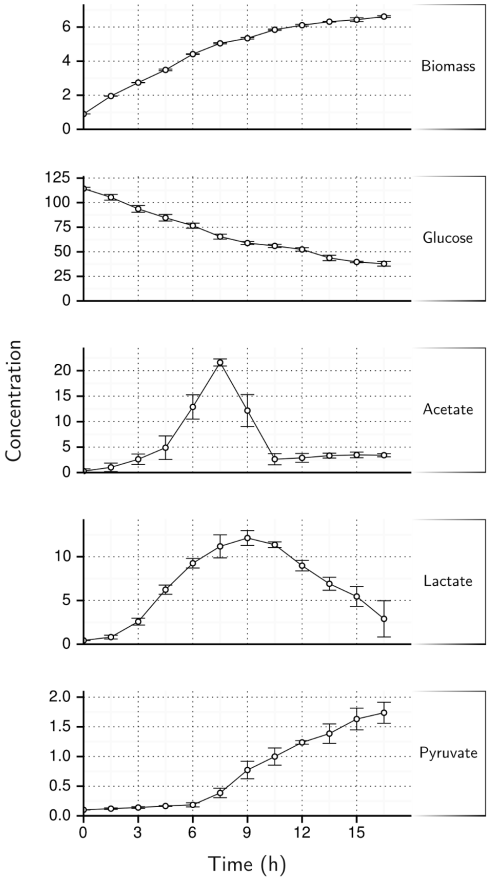


Figure 7.3. – Concentrations of biomass and extracellular compounds measured during growth of *C. glutamicum* under biotin limitation, using glucose as growth substrate. Biomass values are given in g cell dry weight/L, and compound concentrations are given in mmol/L. Each point represents the mean value calculated from three independent measurements. Each error bar represents the standard deviation around the mean.

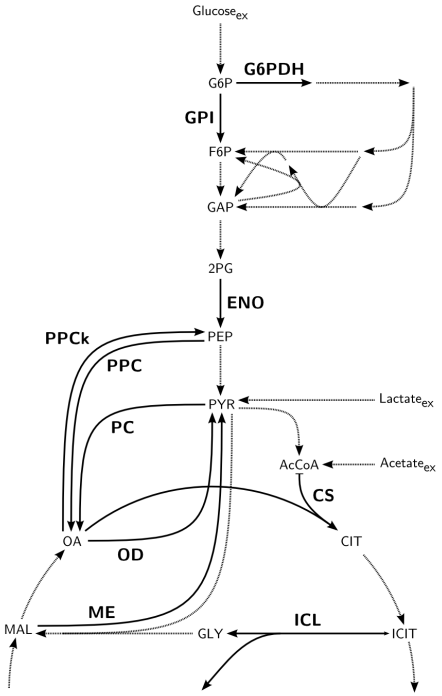


Figure 7.4. – Simplified representation of the studied metabolic network. The reactions that are discussed in the "Application to the core metabolism of *C. glutamicum*" section are represented with solid lines, with the abbreviated name of the enzyme that catalyzes the reaction displayed in bold. The extracellular metabolites are identified by the $_{ex}$ suffix. The complete names of enzymes and metabolites are given at the end of the article, in the "list of abbreviations" section.

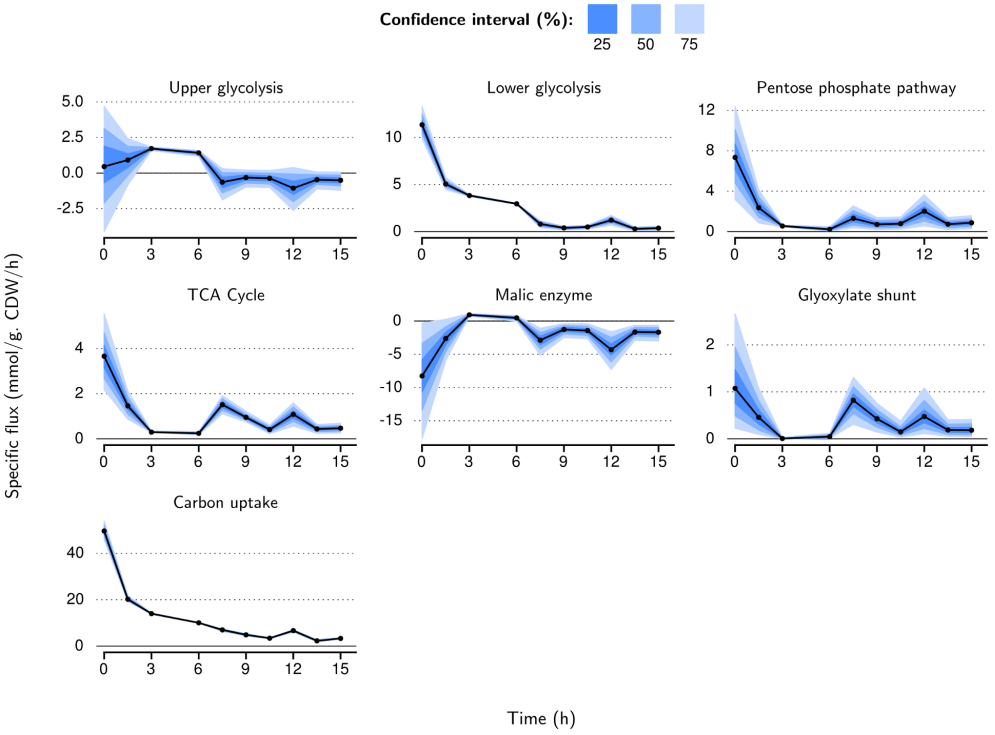


Figure 7.5. – Dynamics and variability of the predicted specific fluxes. The median value of each flux is given by the black solid line with points. The ordinate of each point is the flux value for a modeled time period, and its abscissa corresponds to the time at the beginning of the period (for instance, the 0h → 1.5h period is placed at 0). Confidence intervals at 25, 50 and 75% are displayed with three shades of blue (see the colour code on the figure). The fluxes shown on the figure are "reporter" fluxes for different parts of the network: upper glycolysis (GPI), lower glycolysis (ENO), pentose phosphate pathway (G6PDH), TCA cycle (CS), malic enzyme (ME) and glyoxylate shunt (ICL). The total of carbon uptake ($flux_{carb}$) is calculated following the equation (7.6, p. 144).

As seen in Figure 7.6, the increased variability of normalized fluxes observed during growth of the culture is almost exclusively due to the decrease of the carbon uptake flux. Following equation (7.6, p. 144), this leads to an increase of the term $1/flux_{carb}$, and then to an amplification of the variability of specific fluxes $flux_{spe}$.

From the analysis of the input/output fluxes of extracellular compounds, we can define three phases in the culture, as described by Dietrich *et. al* [Dietrich 2009]:

1. first phase, from 0 to 7.5 h, glucose is the sole carbon source, and part of the consumed carbon is excreted as organic acids (predominantly acetate and lactate).
2. second phase, from 7.5 to 10.5 h, glucose and acetate are co-metabolized. Almost all the acetate producted during the first phase is consumed during this second phase.
3. third phase, from 10.5 to 16.5 h, glucose and lactate are co-metabolized.

The results produced by our method allowed us to identify the specific behavior of each phase of the culture. These behaviors are presented in the following paragraphs.

Median of distributions of flux values: glycolysis, pentose phosphate pathway and tricarboxylic acid cycle

A way to analyze the results produced by our method is to use the median of the distributions of normalized values for each flux. These median values were used here to characterize the behavior of glycolysis, of the pentose phosphate pathway (PP pathway) and the tricarboxylic acid cycle (TCA cycle) throughout the cultivation time.

During the first phase of growth, our predictions show that the use of glucose-6-phosphate (G6P) by the PP pathway (G6PDH flux) gradually decreases, with a simultaneous increase of G6P use by the upper part of glycolysis (GPI flux). The flux through the PP pathway varies from a median value of 92% of the total carbon uptake during the period 0h $\rightarrow$ 1.5h to 13% for the period 6h $\rightarrow$ 7.5h. This phenomenon is accompanied by a gradual increase of the flux through the lower part of glycolysis (ENO flux). In parallel, the TCA cycle is weakly activated, with median values varying between 15% and 45% for the CS flux.

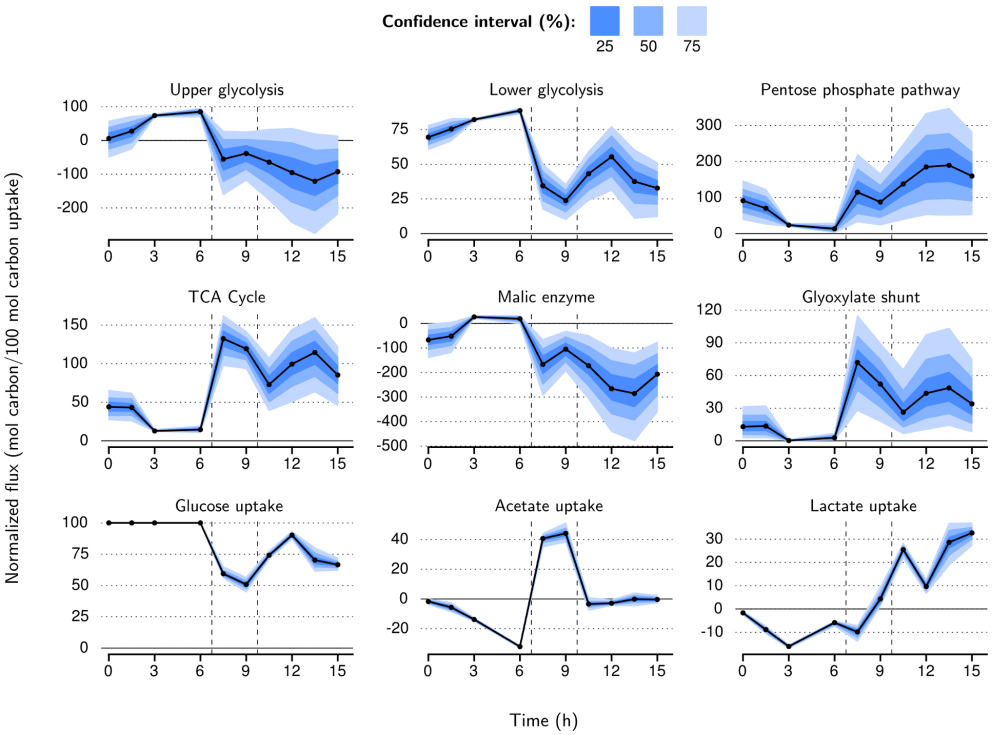


Figure 7.6. – Dynamics and variability of the predicted normalised fluxes. The median value of each flux is given by the black solid line with points. The ordinate of each point is the flux value for a modeled time period, and its abscissa corresponds to the time at the beginning of the period (for instance, the 0h → 1.5h period is placed at 0). Confidence intervals at 25, 50 and 75% are displayed with three shades of blue (see the colour code on the figure). The flux shown on the figure are "reporter" fluxes for different parts of the network: upper glycolysis (GPI), lower glycolysis (ENO), pentose phosphate pathway (G6PDH), TCA cycle (CS), malic enzyme (ME) and glyoxylate shunt (ICL), and for glucose, acetate and lactate uptake. The normalised fluxes are computed following the equation (7.6, p. 144). The boundaries of the three phases of the culture are marked by vertical dotted lines.

During the second phase, the flux distribution totally changes. Our predictions show a hight level of activity of the PP pathway (median value of 100% for the G6PDH flux). The negative values for the GPI flux indicate that gluconeogenesis occurs during this period (median value of 46% in the direction of G6P production). Unlike the upper part, the lower part of glycolysis always presents positive flux values, thus in the direction of pyruvate (PYR) production. Finally, the TCA cycle activity is greatly increased, with a median value of 125% for the CS flux during all of the second phase.

Flux distributions during the third phase are relatively similar to those existing during the second phase. According to our model, high level activity of the PP pathway and gluconeogenesis are still observed (median values of respectively 173% and 99% for G6PDH and GPI fluxes). When compared to the second phase, the ENO flux is higher (median value of 41%). The TCA cycle has a relatively high activity, with a median value of 92% for the CS flux.

Calculation of the most probable flux values: glyoxylate cycle

Another way to utilize the distributions of flux values is to calculate flux ratios at branching points of the network. This method was applied here to estimate the activity of the glyoxylate shunt during growth phases.

In our model, the glyoxylate shunt consists of ICL and MS fluxes. As we have applied no specific constraints on these fluxes, the utilization of the glyoxylate cycle is theoretically possible during all the time periods. In this context, we studied the dynamics of the flux through the glyoxylate shunt (ICL flux) and compared it to the flux entering the TCA cycle (CS flux).

Figure 7.7 presents the estimated probability density of flux ratio ICL/CS for three time periods representative of the three phases of the culture. The most probable flux ratios are 35% and 81% for the time periods 0h $\rightarrow$ 1.5h and 7.5h $\rightarrow$ 9h, respectively. The estimated probability density for the time period 13.5h $\rightarrow$ 15h is not significantly different from a uniform distribution.

Analysis of correlations between fluxes: anaplerotic routes

In this section, we used the distribution of flux values to identify sets of correlated reactions, applying the method proposed by Price *et al.* [Price 2004b] ("correlated subsets").

In addition to the glyoxylate shunt, five other anaplerotic routes are known in *C. glutamicum* (see Figure 7.4, p. 146): pyruvate carboxylase (PC), oxaloacetate decarboxylase (OD), phospho*enol*pyruvate carboxylase (PPC), phospho*enol*pyruvate carboxykinase (PPCk) and malic enzyme (ME). In our model, the PC flux was constrained to a value of 0, as a consequence of biotin limitation. Despite this constraint, the network was clearly under-determined around the pyruvate (PYR), phospho*enol*pyruvate (PEP), malate (MAL) and oxaloacetate (OA) nodes.

The direct consequence of this under-determination was a large variability of the predicted anaplerotic fluxes (see Figure 7.6, p. 149, ME flux). Looking at the correlations between anaplerotic fluxes, we identified two internal loops in this part of the network:

1. a long loop through the following metabolites: PEP → PYR → MAL → OA → PEP;
2. a short loop through the following metabolites: PYR → MAL → OA → PYR.

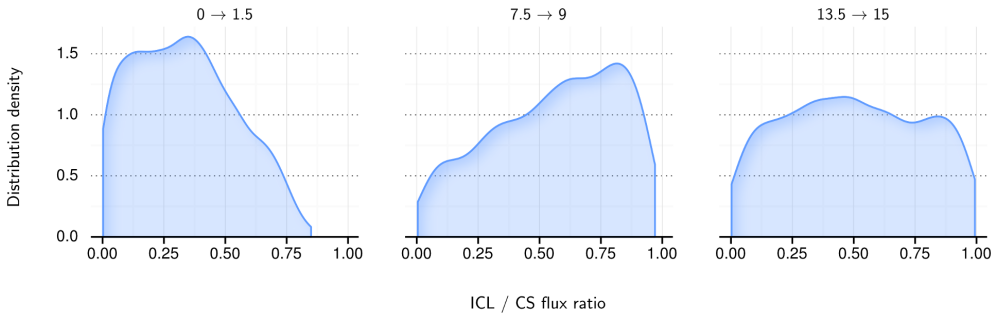


Figure 7.7. – Kernel density estimation of the distribution of ratio between isocitrate lyase (ICL) and citrate synthase (CS) fluxes. The time periods for which the distribution of ICL / CS flux ratio has been computed are mentioned above each graph.

Discussion of results

Under biotin-limited growth conditions, the core metabolism of *C. glutamicum* shows phenomena that have been scarcely studied, such as a linear growth together with co-metabolization of glucose and acetate first and then of glucose and lactate.

Using our method to analyze the behavior of the core metabolism of *C. glutamicum*, we have identified several phenomena that are coherent with previously published papers. Thereby, the flexibility of the G6P node we predicted during the first phase of growth is in concordance with the fact that the values of the flux through the PP pathway may vary widely. Depending on the study, this flux varyied from 89% [Wendisch 2000] to 25% [Klapa 2003] during growth on glucose as the sole carbon source. The glyconeogenesis and the high level of activity of the glyoxylate shunt we predicted during the second phase of the growth is consistent with the study of Wendisch *et al.* on the co-metabolization of glucose and acetate [Wendisch 2000].

Our predictions pointed out a weak activity of the glyoxylate shunt during the first phase of growth. This result could be surprising when referring to *E. coli* metabolism. In *E. coli*, the glyoxylate shunt is repressed by catabolic repression during growth on glucose [Walsh 1984]. In *C. glutamicum*, such catabolic repression is absent. Moreover, Rolin *et. al.* reported a low activity of the glyoxylate shunt during growth on glucose as the sole carbon source [Rollin 1995]. Under the experimental conditions we have analysed, the utilization of the glyoxylate shunt during the first phase of growth could be explained by the fact that there was already co-metabolism of glucose and acetate during this phase, due to an increase of the intracellular concentration of acetate.

We have shown the existence of two futile cycles involving the metabolites PEP, PYR, MAL and OA. Interestingly, Petersen *et. al.* have experimentaly shown a futile cycle between the metabolites PEP $\rightarrow$ PYR $\rightarrow$ OA $\rightarrow$ PEP [Petersen 2000]. To take into account the experimental conditions in our model, the PYR $\rightarrow$ OA flux, which is catalyzed by PC, was set to 0 because of biotin limitation. Therefore, the combined utilization of PYR $\rightarrow$ MAL and MAL $\rightarrow$ OA fluxes that we have predicted could be a means to compensate the absence of an active PC.

Despite the overall validity of our method, we have not been able to find a trajectory solution including the measurements at time 4.5h. The behavior of the system was modeled in the single period 3h to 6h. The carbon balance calculated from experimental measurements indicated a net loss of carbon for 2 of the 3 replicates during the 4.5h → 6h period: the input carbon flux from glucose was lower than the sum of the output carbon fluxes for biomass, acetate, lactate and pyruvate syntheses. This may be first explained by the fact that the equation for biomass composition we used was determined under optimal growth conditions. During biotin-limited growth, the cell envelope of *C. glutamicum* shows a decrease in the phospholipid content [Kimura 2005]. The carbon deficit that we observed could be explained by a lower carbon flux into biomass. This carbon deficit could also be the result of the production of an unknown metabolite, not included in the model, that could be reconsumed during the 4.5h → 6h period, thus compensating the apparent carbon deficit.

chapitre 9, section 9.1, p. 188

7.2.4. Effect of the dependence between time periods on the variability of fluxes

A solution trajectory is a succession of one flux map per period, where each map is (*i*) a solution for the considered time period and (*ii*) allows a solution for all the other time periods to be found. To show that calculation of solution trajectories improves the prediction of flux distributions, we have compared the distribution of flux values predicted with our "complete trajectory" method to the distribution of flux values that can be obtained with a more simple method, leading to "independent flux maps".

We used the "complete trajectories" presented in the "Application to the core metabolism of *C. glutamicum*" section. We have generated the "independent flux maps" with the same information (experimental measurements, network, parameters), but each map was built independently: for each time period and for each map of this period, a set of initial concentrations for extracellular metabolites was randomly sampled, according to the method we have defined for the first time period to study "complete trajectories".

Figure 7.8 (p. 155) compares the distribution of flux values for a set of reactions of the network. As can be seen, the flux variability diminishes when the dependence between successive time periods is taken into account.

This decrease is mainly due to the fact that some flux maps for the time period $t \rightarrow t+1$ lead to extracellular concentrations at time $t+1$ that do not allow any solution to be found for at least one of the subsequent time periods. As a consequence, the variability of the possible solutions on the entire culture is drastically reduced.

Furthermore, as seen for the time period 1.5h $\rightarrow$ 3h for all the fluxes shown in Figure 7.8, the median values are different depending on the method. The reason of these shifts of the median values is the modification of the solution spaces due to the additional constraint introduced by the dependency between time periods.

From a biological point of view, linking the successive time periods is more realistic than studying each period independently. Thus, the predictions produced by our method are likely to be more relevant to the understanding of the dynamics of a metabolic system than methods that do not take this temporal dependency into account.

7.3. Conclusion

In this paper, we present a new method to estimate the dynamics of metabolic fluxes over time, with integration of the experimental measurement variability. Our method combines an interative algorithm for sampling the solution space together with the addition of a temporal dependency between the successive time periods of the culture. The set of "solution trajectories" generated by this method allows the estimation of the most probable dynamics for the system. The temporal dependency that we introduced leads to a significant reduction of the variability of the predicted fluxes.

The sampling of the solution space gives distributions of flux values that are more informative than simple intervals of values, such as those that are calculated with the FVA method. With our distributions of flux values, we are able to study the variability of flux ratios at branching points of the network, and to look for sets of correlated fluxes.

We have applied our method to study the core metabolism of *C. glutamicum* during biotin-limited growth on glucose as the initial carbon source. Our results show that, within the limits of available knowledge on the functioning of *C. glutamicum* metabolism under these unusual conditions, the

predictions given by our method are consistent with previously published studies on this organism.

Several improvements can be considered. At the level of the method, integration of other biological information, such as transcriptomic data, could lead to an increase of the constraints exerted on the solution space. This integration will require the translation of expression levels into constraints on metabolic fluxes. Considering the metabolism of *C. glutamicum*, the highly constrained solution spaces observed at the end of the first phase of the growth could denote the fact that some other metabolites are produced and consumed by the organism. Additional experiments are planned in order to test this hypothesis, especially a more complete metabolomic study and the monitoring of gas exchanges during cultivation.

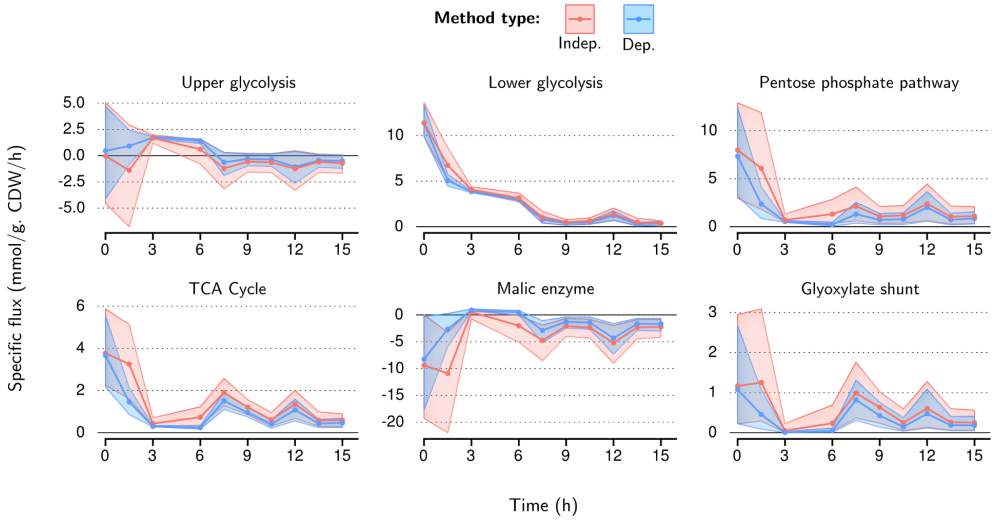


Figure 7.8. – Variability of specific fluxes depending on the calculation method. The median value of each flux is given by the solid line with points. The ordinate of each point is the flux value for a modeled time period, and its abscissa corresponds to the time at the beginning of the period (for instance, the 0h → 1.5h period is placed at 0). The coloured areas represent the confidence interval at 75%. The variability calculated without taking into account the dependence between successive time periods is in red, and the variability calculated by taking into account the temporal dependence (our method) is in blue. The fluxes shown on the figure are "reporter" fluxes for different parts of the network: upper glycolysis (GPI), lower glycolysis (ENO), pentose phosphate pathway (G6PDH), TCA cycle (CS), malic enzyme (ME) and glyoxylate shunt (ICL).

7.4. Methods

7.4.1. Construction of the *C. glutamicum* network model

In order to test our method, we constructed a constraint based model of the core metabolism of *C. glutamicum*. The section below presents the scope and the settings of this model.

The constructed model was focussed on the core metabolism of *C. glutamicum*. The modelled metabolic pathways were glycolysis, the pentose phosphate pathway, the TCA cycle, anaplerotic reactions and the respiratory chain. In order to model inputs and outputs of acetate, lactate and pyruvate, respective metabolic pathways and transport reactions were also included. Table 7.1 gives details on the most important features of each pathway.

Enzymatic reactions were taken from the literature and from a pathway/genome database we built for *C. glutamicum* metabolism, using Pathway Tools software [Karp 2010]. Reversibility of reactions were assessed on the basis of both literature mining and using the eQuilibrator thermodynamic analysis tool [Noor 2012]. Needs in intracellular compounds for the production of biomass were taken from earlier research ([Cocaign-Bousquet 1996] ; see Table 7.2).

The non-growth-associated maintenance requirements were incorporated into the model as an ATP to ADP reaction. In *C. glutamicum*, the mATP value for cell maintenance increases linearly as a function of osmolarity, from 1.8 to 9.2 mmol/g CDW/h between 300 and 1800 mosmol/kg ([Varela 2004]). Under our culture conditions, the calculated osmolarity was in the range 1300 to 1400 mosmol/kg. Thus, the mATP value in our model was constrained to a minimum value of 5 mmol/g CDW/h with no upper constraint.

The modeled respiratory chain includes the cytochrome *bd* and the cytochrome bc_1-aa_3 branches and the F_0-F_1 ATPase. The proton translocation efficiency was set to 2 protons per menaquinone for the cytochrome *bd* branch and 6 for the cytochrome bc_1-aa_3 branch [Bott 2003]. The number of protons required for the formation of one ATP was set to 3 [Bott 2003].

Table 7.1. – Functional characteristics of the metabolic pathways which are parts of the model of *C. glutamicum* core metabolism.

Metabolic Pathways	Important features
Glycolysis	Glucose input Production of biomass precursors Co-factors involved in the respiratory chain
Pentose phosphate	Production of biomass precursors NADPH generation
TCA cycle	Production of biomass precursors Co-factors involved in the respiratory chain NADPH generation
Anaplerotic reactions	Biotin limitation effect
Respiratory chain	ATP production
Acetate	Acetate input and output Co-factors involved in the respiratory chain
Lactate	Lactate input and output Co-factors involved in the respiratory chain
Pyruvate	Pyruvate input and output

Table 7.2. – Metabolites and energy needs for 1 g of *C. glutamicum* biomass.

Compounds	Conc. (mmol/g)
Glucose 6-phosphate	0.318
Fructose 6-phosphate	0.126
Ribose 5-phosphate	0.684
Erythrose 4-phosphate	0.198
Triose phosphate	0.062
3-Phosphoglycerate	0.872
Phospho*enol*pyruvate	0.458
Pyruvate	2.809
Acetyl coenzyme A	3.122
α-Ketoglutarate	1.389
Oxaloacetate	1.168
ATP	29.204
$NADPH_2$	11.509
$NADH_2$	0.388

Our model used a P/O ratio of 2.0, the commonly accepted value for aerobic bacteria ([Cocaign-Bousquet 1996, Varela 2004, Shimizu 2003, Varela 2003, Shirai 2005]) with a predominant use of the bc_1-aa_3 branch.

Glucose uptake by *C. glutamicum* is achieved by 2 different systems: the glucose phosphotransferase system (PTS) [Parche 2001, Ikeda 2012] and a PTS-independent glucose uptake system [Cocaign-Bousquet 1996, Lindner 2011]. Glucose uptake through the PTS was constrained to be at least 90% of the total uptake of glucose [Cocaign-Bousquet 1996].

Biotin is the cofactor of carboxylating enzymes, especially pyruvate carboxylase [Peters-Wendisch 1998]. In *C. glutamicum*, pyruvate carboxylase is the major anaplerotic enzyme, supplying oxaloacetate to the TCA cycle. Under our experimental conditions, the amount of biotin was limited. This limitation led to a strong decrease of this enzyme's activity. To model biotin limitation, the flux associated to pyruvate carboxylase was forced to a value of 0.

All other cofactors involved in metabolic reactions were balanced in the model. Reversible fluxes were set within the value range -100 to 100. Irreversible fluxes were set within the range 0 to 100. These constraints are only reachable by fluxes forming internal loops, therefore constraining the solution space without interfering with variability of informative fluxes.

To summarize, the following assumptions were made to build the model. Needs of intracellular compounds to produce 1g of dry cells were assumed to remain constant over the cultivation time. Respiratory chain efficiency was estimated to be constant, with a P/O ratio of 2. Biotin-limited growth was expected to suppress the pyruvate carboxylase activity. ATP maintenance was assumed to be at least equal to 5 mmol/g CDW/h.

Finally, the model contained 64 reactions (including 11 transport reactions) and 60 compounds (including 10 external compounds). Note that the 64 lumped metabolic reactions correspond to about 250 enzymatic reactions [Neidhardt 1990]. A detailed list of compounds and reactions included in our model is given in the annex part B (p. 253).

7.4.2. Biological data

Experimental data were taken from a previous study, during which *C. glutamicum* cells were analyzed under biotin-limited growth conditions [Dietrich 2009].

The data used in the present work were the extracellular concentrations of glucose, acetate, lactate, pyruvate and biomass. These concentrations were assessed every 90 min during 25 h. At each time and for each compound, 3 independent measurements were performed (by using 3 independent cultures). These data are shown in Figure 7.3 (p. 145) and available in the annex part B (p. 253).

Details about the strain, growth conditions and metabolite analyses are described in [Dietrich 2009].

No smoothing methods were applied to the measured data.

7.4.3. Setting of the parameters of the method

Several parameter values have to be fixed before using the method:

chapitre 8, section 8.2, p. 174

1. the number n of trajectories produced;
2. the number of flux maps sampled from the solution space for each time period;
3. the parameters of the distribution of initial extracellular concentrations (compound and biomass) during the first time period p_i;
4. the coefficient d applied to the experimental variability for the calculation of constraints on input/output fluxes (see 7.4, p. 140).

We have chosen to produce $n = 1,000$ trajectories. This allowed the estimation of the flux distributions at each time period from 1,000 values. According to our tests, increasing the number of trajectories up to $n = 10,000$ led to only a very small improvement in the solution space sampling (data not shown), whereas it multiplied the running time of the method by a factor of 10.

For each trajectory, the number of flux maps sampled from the solution space for each time period was 10. When building our method, we identified a bias in the set of flux maps sampled from the solution space, due to the "mirror" method: the first 10 flux maps were not randomly distributed in

the solution space (data not shown). In order to avoid this bias, we used the "mirror" method for sampling 20 flux maps, in which we only kept the 10 last maps, corresponding to iterations 11 to 20 of the algorithm. Amongst these 10 selected maps, only one was randomly chosen (the 10 maps are equiprobables) to build the trajectory. Even if the number of sampled flux maps seems to be quite low, we did not observe any significant improvement of the solution space sampling by increasing this number (data not shown).

The initial extracellular concentrations for the first time period p_1 were randomly selected from a Gaussian distribution with the location parameter m and scale parameter s, which are equal to the mean and the standard deviation of the experimental measurements of the concentrations, respectively. The Gaussian distribution is truncated at bounds $m - s$ and $m + s$, avoiding negative values.

The variable constraints on input/output fluxes of the system were calculated by applying a coefficient $d = 1$ to the experimental variability (see equation 7.4, p. 140).

Chapitre 8

Approfondissements

L'approche développée propose de générer un ensemble de trajectoires « solutions ». Une trajectoire est constituée d'une succession de cartes de flux, avec une carte solution pour chaque période de temps modélisée. La dépendance entre les différentes périodes est instaurée en intégrant une contrainte de « faisabilité » : pour une trajectoire donnée, la succession des cartes de flux doit conduire à une évolution des concentrations extracellulaires qui soit cohérente avec les mesures expérimentales. Une concentration prédite est considérée comme cohérente si elle est comprise dans un intervalle « cible » de valeurs qui sont déterminées à partir des mesures expérimentales (et de leur variabilité).

Pour une période p_i donnée, les concentrations initiales au temps t_i peuvent varier selon « l'histoire » de la trajectoire — autrement dit, selon la succession des cartes prédites pour les périodes précédentes. Comme les contraintes variables sur les flux E/S sont calculées en fonction des concentrations au temps t_i, ces contraintes peuvent changer d'une trajectoire à l'autre. Un jeu de contraintes variables définit donc un espace des solutions qui représente les répartitions de flux pertinentes, étant donnés les concentrations de départ (à t_i) et les intervalles de concentrations cibles (à t_{i+1}).

Sur cette base, l'exploration des répartitions de flux possibles pendant chaque période ne correspond donc pas à l'étude d'un unique espace des solutions. Elle consiste en fait à évaluer un « méta » espace des solutions, constitué d'un ensemble d'espaces qui sont relativement similaires, sans toutefois être identiques.

Afin d'évaluer ces « méta » espaces des solutions, j'ai utilisé l'échantillonnage de Monte-Carlo par chaîne de Markov (MCMC) d'une manière différente de ce que l'on rencontre habituellement dans les études sur les MBC [Schellenberger 2009]. Dans ces études, la notion de « méta » espace des

solutions n'est pas présente, et un nombre réduit d'espaces des solutions est échantillonné par une ou quelques grandes chaînes de Markov.

Dans l'approche que je propose, un espace des solutions est échantillonné pour chaque période de chaque trajectoire. Pour une trajectoire $traj_j$, et pour une période p_i de cette trajectoire, une unique solution est choisie parmi toutes celles échantillonnées, et cette solution est utilisée pour représenter l'état du système pendant la période p_i dans la trajectoire $traj_j$.

En générant un nombre j de trajectoires, le « méta » espace des solutions de chaque période peut alors être évalué à partir de j valeurs indépendantes. De plus, les seules solutions retenues pour évaluer un « méta » espace sont celles qui participent à la prédiction d'une évolution des concentrations cohérente avec les mesures expérimentales. En conséquence, chaque « méta » espace des solutions généré représente un espace de solutions qui est enrichi en solutions cohérentes avec la dynamique de la culture. Les j solutions échantillonnées pour chaque période permettent ensuite d'estimer la distribution des valeurs de chaque flux du système pour chaque période et de caractériser le comportement du système au cours du temps, comme cela a été présenté dans le chapitre précédent.

Dans ce chapitre, je vais approfondir certains points de l'approche proposée. Dans une première section, je présente l'étude réalisée afin d'identifier la méthode d'échantillonnage la plus performante vis-à-vis du système étudié (section 8.1, p. 164). Dans une seconde section, je présente les analyses réalisées afin de déterminer (i) les indices des itérations des chaînes de Markov adéquats pour réaliser une bonne exploration de chaque espace des solutions, et (ii) le nombre de trajectoires nécessaires pour estimer correctement les « méta » espaces des solutions (section 8.2, p. 174). Dans un troisième section, je reviens sur la fonctionnalité de « retour en arrière » qui a été ajoutée à l'algorithme initial de la méthode (section 8.3, p. 180).

Dans l'approche proposée, le paramètre d permet de moduler la quantité de variabilité expérimentale prise en compte dans le calcul des intervalles des concentrations cibles. D'après les analyses préliminaires réalisées, il n'y a pas de schéma simple qui se dégage quant à l'impact de ce paramètre sur les prédictions. D'autres analyses sont nécessaires, mais, faute de temps, elles n'ont pu être effectuées, c'est pourquoi je ne détaillerai pas les analyses préliminaires.

8.1. Choix de la méthode d'échantillonnage

Il s'agit ici de déterminer la méthode d'échantillonnage la plus performante parmi trois algorithmes implémentés dans le package R « limSolve » [Soetaert 2009], et exécutables au moyen de la fonction « xsample » :

— Méthode *random directions algorithm* (`rda`). La méthode `rda` [Smith 1984] fonctionne en deux étapes (figure 8.1, partie a) :

1. À partir d'un point solution p_i de l'espace des solutions, une direction est choisie au hasard. Le point et la direction choisie définissent un segment qui traverse l'espace des solutions de part en part.
2. Un nouveau point p_{i+1} est tiré au hasard, selon une loi uniforme, sur le segment défini à la première étape. Ce nouveau point est utilisé comme nouveau point de départ pour l'étape n° 1.

— Méthode *coordinates directions algorithm* (`cda`). L'algorithme de la méthode `cda` [Smith 1984] est identique à celui de la méthode `rda`, à l'exception que la direction est choisie parmi un ensemble fini de directions possibles, où chaque direction correspond à l'un des axes de l'espace des solutions.

— Méthode *mirror*. L'algorithme de la méthode `mirror` [Van den Meersche 2009] fonctionne en deux étapes (figure 8.1, partie b) :

1. À partir d'un point p_i de l'espace des solutions, un *mouvement* à réaliser pour obtenir un nouveau point p_{i+1} est calculé. Pour cela, pour chaque axe j de l'espace des solutions, une distance $dist_j$ — distance à parcourir le long de l'axe j — est tirée au hasard au sein d'une distribution normale de moyenne 0 et d'écart-type jmp_j paramétrable. L'ensemble des distances $dist_j$ représente ainsi le mouvement à appliquer depuis le point p_i pour obtenir le point suivant p_{i+1}.
2. Le mouvement est appliqué depuis le point p_i. Si la réalisation de ce mouvement implique d'enfreindre une ou plusieurs contraintes d'inégalité, ces contraintes sont utilisées comme plans de réflexion et le mouvement à réaliser est actualisé. À l'issue du mouvement, le point suivant p_{i+1} est défini, et ce point est utilisé comme nouveau point de départ pour l'étape n° 1.

Au final, la réalisation de chacun de ces algorithmes produit une chaîne de Markov constituée de n points (générés en n itérations). Dans notre situation, chaque point représente une carte de flux solution, et donne donc une valeur pour chaque flux du système. À supposer que le nombre de points échantillonnés soit suffisant, les points qui constituent une chaîne de Markov peuvent ensuite être utilisés pour estimer la distribution des valeurs de chaque flux au sein de l'espace des solutions échantillonné.

La notion de « nombre de points suffisant » est liée à la loi des grands nombres et à la notion de convergence d'une chaîne de Markov. Lorsque le nombre de points échantillonnés tend vers l'infini, les distributions empiriques des valeurs de chaque variable convergent vers les distributions réelles (caractéristiques de la population, autrement dit de l'espace des solutions). Le nombre d'itérations n suffisant pour obtenir une bonne estimation des distributions réelles peut être évalué en déterminant la valeur de n à partir de laquelle la réalisation de nouvelles itérations ne modifie pas les paramètres des distributions estimées.

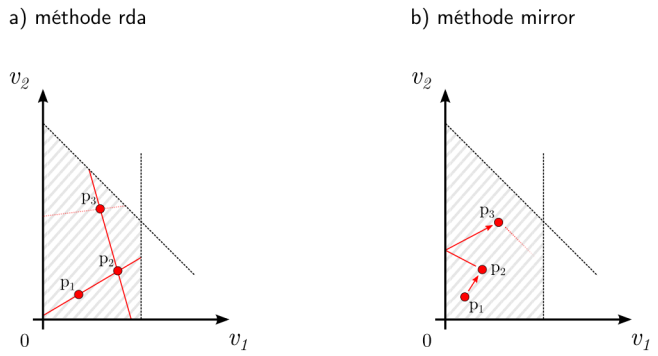


Figure 8.1. – Algorithmes utilisés par les méthodes `rda` (a) et `mirror` (b) pour échantillonner un espace des solutions. Pour chaque graphique, la partie hachurée représente un espace des solutions vu sur un plan v_1 - v_2. Chaque point de couleur rouge représente une solution p_i échantillonnée. Les explications du fonctionnement de chaque algorithme sont données dans le texte.

8.1.1. Espace des solutions pour les tests

Afin de réaliser les comparaisons en utilisant des conditions « expérimentales » communes, un espace des solutions de test est défini de la manière suivante :

— Système métabolique. Le système métabolique utilisé est le métabolisme central de *C. glutamicum*, présenté dans le chapitre précédent, et dont la description est donnée en annexe (annexe B, p. 253). Une fois les contraintes constantes appliquées, le réseau correspond à un système sous-déterminé avec 14 degrés de liberté.

— Période de temps p_i. La période est définie sur l'intervalle $t_i = 0$ à $t_{i+1} = 1,5\text{h}$.

— Contraintes variables. Pour le calcul des contraintes variables appliquées sur les flux E/S v_{io,p_i}, les concentrations initiales en métabolite c_{m,t_i} et en biomasse bio_{m,t_i} sont calculées en prenant les moyennes des mesures expérimentales ($c_{m,t_i} = \bar{c}_{e,t_i}$ et $bio_{m,t_i} = \bar{bio}_{e,t_i}$, respectivement). Les contraintes variables sont ensuite calculées comme décrit dans le chapitre précédent (p. 140). Le paramètre d est fixé à 1.

8.1.2. Critères de performance

Afin d'identifier la méthode d'échantillonnage la plus performante, je me suis basé sur deux critères :

— le nombre d'itérations n nécessaires pour que les paramètres des distributions estimées convergent vers une valeur stable ;

— le temps d'exécution nécessaire pour obtenir le nombre n d'itérations.

8.1.3. Comparaison sur la base du nombre d'itérations

Pour cette première comparaison, l'espace des solutions de test est échantillonné en générant une chaîne de Markov avec chacune des trois méthodes. Pour chaque échantillonnage, un même point de départ correspondant à la solution centrale de l'espace des solutions est utilisé (cf. p. 136), et le nombre d'itérations n est fixé à 5 000. Les valeurs du paramètre jmp de la méthode

`mirror` n'ont pas été spécifiés pour l'exploration : dans cette situation, de « bonnes » valeurs de *jmp* sont calculées par la fonction xsample, sur la base du domaine de variation possible pour chaque axe de l'espace des solutions.

Afin d'évaluer si les solutions échantillonnées par les méthodes couvrent l'ensemble de l'espace des solutions, les valeurs minimales et maximales atteignables par chaque flux sont calculées en utilisant la méthode FVA : la valeur de chaque flux est successivement utilisée comme fonction *objectif*, et cette valeur est minimisée ou maximisée afin d'obtenir l'intervalle des valeurs possibles.

La figure 8.2 illustre les distributions de valeurs obtenues pour trois des flux du système. Comme on peut le voir, les étendues des valeurs échantillonnées sont clairement plus importantes avec la méthode `mirror`. Pour chaque flux, les valeurs échantillonnées couvrent quasiment l'ensemble de la plage des valeurs possibles (plage bornée par les valeurs minimales et maximales calculées avec la méthode FVA). En comparaison, les méthodes `rda` et `cda` n'ont exploré qu'une faible portion des valeurs possibles, centrée autour du point de départ de la chaîne. Ce constat est valable pour l'ensemble des flux du système, et la génération de nouvelles chaînes conduit à des

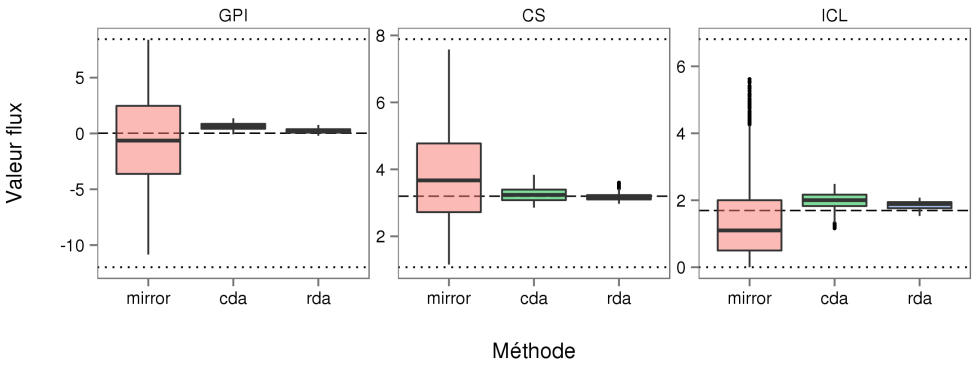


Figure 8.2. – Comparaison des distributions de valeurs obtenues à partir des différentes méthodes d'échantillonnage lorsque le nombre d'itérations n est fixé à 5 000. Les distributions de valeurs sont représentées pour trois flux, de gauche à droite : GPI, CS et ICL. La valeur des flux est exprimée en mmol/g biom./h. Pour un flux donné, chaque boîte à moustaches correspond à l'une des méthodes d'échantillonnage. Les droites horizontales en pointillé, en haut et en bas de chaque graphique, correspondent respectivement aux valeurs maximales et minimales atteignables. La droite horizontale en tiret, dans chaque graphique, représente la valeur issue de la solution centrale.

résultats reproductibles. En conséquence, à nombre d'itérations équivalent, la méthode `mirror` semble plus performante pour explorer l'espace des solutions de test.

Toutefois, il convient ici de remarquer ici que le temps d'exécution nécessaire pour générer les 5 000 itérations est beaucoup plus long avec la méthode `mirror` : 160 secondes, contre environ 3 secondes pour les méthodes `rda` et `cda`. Le temps d'exécution plus long de la méthode `mirror` est notamment dû à des valeurs élevées du paramètre jmp (non spécifiées, calculées par la fonction). Le fait de parcourir une grande distance entre chaque itération nécessite de réévaluer fréquemment la direction dans laquelle l'espace des solutions est exploré. En conséquence, une manière plus « juste » de confronter les méthodes est de comparer la qualité des échantillonnages lorsque le temps d'exécution est similaire pour les trois méthodes.

8.1.4. Comparaison sur la base du temps d'exécution

Pour cette seconde comparaison, le même protocole de test est utilisé, mais le nombre d'itérations est choisi de façon à ce que le temps d'exécution de chaque méthode soit similaire :

— 5 000 itérations pour la méthode `mirror` (160 secondes) ;
— 1 500 000 itérations pour les méthodes `rda` et `cda` (environ 162 secondes).

La figure 8.3 montre que, à temps d'exécution égal, les différences entre la méthode `mirror` et les deux autres méthodes sont moins marquées en termes de plages de valeurs échantillonnées. Cependant, les plages de valeurs explorées par les méthodes `rda` et `cda` restent toujours moins importantes en comparaison de celles explorées par la méthode `mirror`.

Un point important visible sur cette figure est le fait que la médiane estimée pour chaque flux est différente selon la méthode considérée. La figure 8.4 montre que cette différence de médiane est due au fait que les chaînes de Markov générées par les méthodes `rda` et `cda` n'ont pas convergé après les 1 500 000 itérations. Cela indique que le nombre d'itérations pour ces deux méthodes n'est pas suffisant pour estimer les distributions des valeurs de flux.

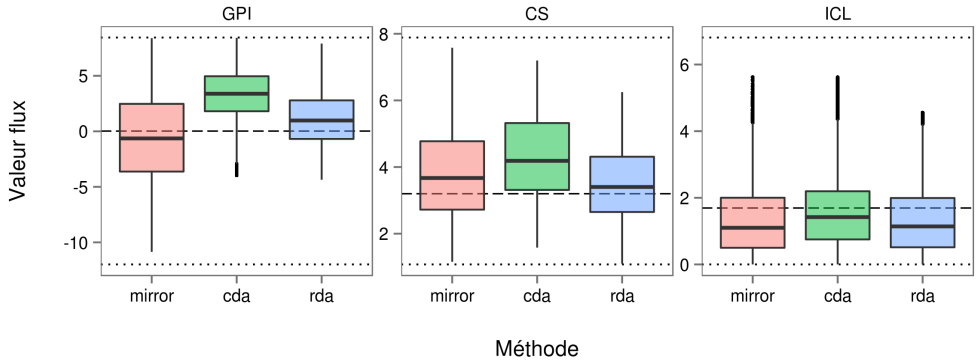


Figure 8.3. – Comparaison des distributions de valeurs obtenues à partir des différentes méthodes d'échantillonnage lorsque le temps d'exécution est d'environ 160 secondes. Les distributions des valeurs sont calculées à partir de 5 000 points pour la méthode `mirror`, et de 1 500 000 pour les méthodes `rda` et `cda`. La valeur des flux est exprimée en mmol/g biom./h. Pour chaque flux, les droites horizontales en pointillé, en haut et en bas de chaque graphique, correspondent respectivement aux valeurs maximales et minimales atteignables. La droite horizontale en tiret, dans chaque graphique, représente la valeur issue de la solution centrale.

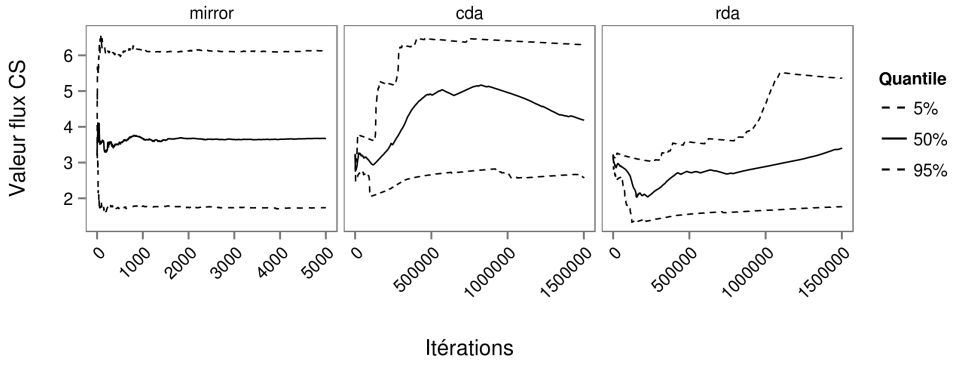


Figure 8.4. – Évolution des quantiles de la distribution du flux CS en fonction du nombre d'itérations, selon la méthode. La valeur du flux est exprimée en mmol/g biom./h. Le trait plein représente l'évolution de la médiane en fonction du nombre n d'itérations prises en compte (à compter de la première itération). Les traits pointillés inférieur et supérieur représentent respectivement la valeur des quantiles à 5% et 95%.

Dans l'analyse précédente (comparaison à nombre d'itérations égal), les valeurs élevées du paramètre *jmp* de la méthode `mirror` représentaient potentiellement un désavantage, puisque la méthode nécessite un temps de calcul beaucoup plus long que les deux autres pour générer les 5 000 points. Dans les faits, le parcours d'une grande distance entre chaque point permet à la méthode `mirror` d'échantillonner des valeurs (chaque point représente une solution, chaque solution représente une valeur pour chaque flux) qui sont peu corrélées d'une itération à la suivante. Autrement dit, la position dans l'espace des solutions du point échantillonné à une itération i n'a que peu d'influence sur la position du point échantillonné à l'itération $i+1$. La dépendance entre les itérations peut être estimée en calculant l'autocorrélation des valeurs de chaque variable (c'est-à-dire de chaque flux), selon la formule

$$c_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

où

- x_i et $x_i + k$ représentent respectivement les valeurs de la variable aux itérations i et $i + k$;
- n est la longueur totale de la série de valeur ;
- et c_k donne la corrélation entre les valeurs distantes de k itérations.

La figure 8.5 montre l'autocorrélation du flux CS, calculée avec k variant de 1 à 1 000, pour la chaîne générée par chaque méthode. On peut facilement voir que la corrélation entre les itérations devient rapidement proche de 0 pour la méthode `mirror`, tandis que les méthodes `rda` et `cda` présentent toujours une très forte corrélation entre les valeurs échantillonnées à 1 000 itérations de distance. Cette forte corrélation explique la lente convergence de ces deux méthodes : la chaîne de Markov visite lentement (en terme d'itérations) l'espace des solutions. Ce phénomène est illustré par des « traces » qui fluctuent peu en comparaison de celles générées par la méthode `mirror` (figure 8.6).

La taille effective d'un échantillon (*effective size sample*) est une autre mesure de la qualité de l'échantillonnage. De manière simplifiée, cette mesure est obtenue en relativisant le nombre de valeurs échantillonnées par rapport à leur variabilité. Elle exprime le nombre de points échantillonnés que l'on peut considérer comme indépendants, et elle est donc liée au niveau de corrélation entre les itérations.

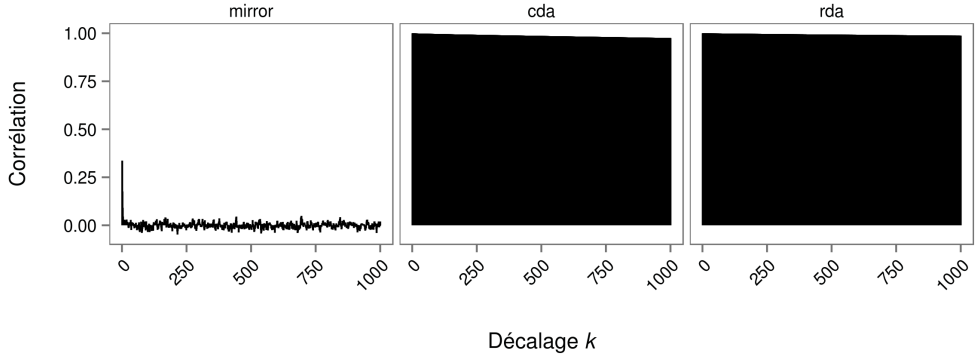


Figure 8.5. – Autocorrélation des valeurs échantillonnées pour le flux CS, selon la méthode d'échantillonnage. La corrélation est donnée en ordonnée, le décalage k est donné en abscisse.

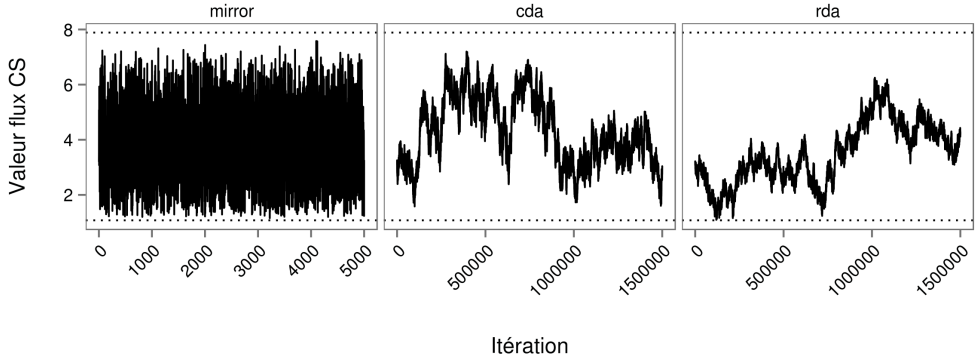


Figure 8.6. – Trace des valeurs échantillonnées pour le flux CS au fil des itérations et selon la méthode d'échantillonnage. L'axe des ordonnées indique la valeur du flux (en mmol/g biom./h), l'axe des abscisses indique le numéro de l'itération dans la chaîne de Markov. Les droites horizontales en pointillé en haut et en bas de chaque graphique correspondent respectivement aux valeurs maximales et minimales atteignables.

La taille effective obtenue pour le flux CS est drastiquement différente entre la méthode `mirror` (2 152 valeurs indépendantes) et les méthodes `rda` et `cda` (27 et 12, respectivement). Les tailles effectives ont été calculées avec le package R "coda" [Plummer 2006].

Si l'on revient à la figure 8.4 (p. 169), la chaîne de Markov générée par la méthode `mirror` semble avoir convergé. Ceci a été confirmé en générant 5 couples de chaînes de Markov avec la méthode `mirror`, puis en testant deux à deux la similarité des distributions des valeurs échantillonnées de chaque flux avec un test de Kolmogorov et Smirnov (test non paramétrique). La grande majorité des flux ne présente pas de différence significative (*p-value* > 0.05). Les seuls flux qui présentent une différence très significative (*p-value* < 0.01) sont les couples de flux PFKA - FBP, qui forment une boucle interne au niveau de la glycolyse, et les flux PPC - PPCk, qui forment une boucle interne au niveau des voies « plérotiques » [1]. Comme ces flux forment des boucles internes — le produit de l'un des flux est le substrat de l'autre et *vice-versa* — ils ne sont pas bornés par les contraintes variables exercées sur les flux E/S. De fait, ces flux sont peu informatifs pour caractériser la dynamique du métabolisme. L'augmentation du nombre d'itérations n'a pas d'impact sur les paramètres de distribution des autres flux : cela permet « uniquement » d'obtenir une résolution plus précise des distributions.

Finissons en remarquant que, en dehors du fait que les méthodes `rda` et `cda` soient peu performantes en comparaison de la méthode `mirror`, le nombre important de points générés représente un frein potentiel quant à leurs utilisations. En effet, les solutions générées par ces deux méthodes représentent chacune une matrice de 1 500 000 lignes (nombre d'itérations) et de 60 colonnes (nombre de variables), ce qui nécessite pour chaque matrice une allocation mémoire de 687 méga-octets dans le logiciel R.

8.1.5. Choix de la méthode d'échantillonnage : bilan

L'objectif de cette analyse était d'identifier la méthode d'exploration la plus performante. Sur la base des analyses réalisées sur l'espace des solutions de test, la méthode `mirror` ressort clairement comme la meilleure méthode : en 160 secondes et 5 000 itérations, la méthode explore l'espace

1. une réaction *ana*plérotique, comme la réaction PPC, est une réaction chimique qui produit un métabolite intermédiaire d'une voie métabolique, en l'occurrence du cycle de Krebs ; à l'inverse une réaction *cata*plérotique, comme la réaction PPCk, consomme un métabolite intermédiaire

des solutions de manière satisfaisante et converge vers des distributions de valeurs stables. La même analyse a été réalisée pour les autres périodes de temps, et les résultats obtenus conduisent à la même conclusion. La méthode d'échantillonnage `mirror` est donc la méthode retenue pour étudier la dynamique du métabolisme.

8.2. Paramétrage de la méthode

Dans l'approche que je propose, les « méta » espaces des solutions sont échantillonnés au moyen de la sélection d'une multitude de solutions, où chaque solution est issue d'une chaîne de Markov différente.

Compte tenu de cette approche, le bon échantillonnage d'un « méta » espace des solutions dépend de deux critères. Le premier critère est la manière dont est sélectionnée la solution issue de chaque échantillonnage. En particulier, il s'agit de déterminer à partir de quelle itération d'une chaîne de Markov la solution échantillonnée est indépendante de la solution choisie pour la première itération. Le deuxième critère est le nombre total de solutions nécessaires pour obtenir une bonne caractérisation d'un « méta » espace des solutions. Rappelons ici que le nombre total de solutions échantillonnées pour chaque « méta » espace correspond au nombre de trajectoires solutions générées.

8.2.1. Nombre minimum d'itérations pour obtenir une solution indépendante du point de départ

On peut distinguer deux parties dans une chaîne de Markov : une partie de « préchauffage » (*warmup*), qui correspond aux itérations pour lesquelles les valeurs échantillonnées sont encore dépendantes du point de départ, et une partie stable dans laquelle les valeurs échantillonnées ne dépendent plus du point de départ.

Afin d'analyser le biais lié au point de départ de l'échantillonnage, l'espace des solutions de test (présenté dans la section précédente) a été échantillonné par 1 000 chaînes de Markov en utilisant la méthode `mirror`. Pour chaque échantillonnage, le nombre d'itérations est fixé à 30, et le point de départ utilisé correspond à la solution centrale de l'espace des solutions.

La figure 8.7 montre, pour trois flux, la corrélation entre les 1 000 valeurs échantillonnées à la deuxième itération et les 1 000 valeurs échantillonnées à la k^{e} itération suivante, avec k variant de 1 (comparaison entre la deuxième et la troisième itération) à 28 (comparaison entre la deuxième et la trentième itération). D'après ces résultats, il semble que la dépendance vis-à-vis du point est relativement faible dès la seconde itération (corrélation d'environ 0,25) et disparait totalement après la 7^{e} itération. Ces résultats sont

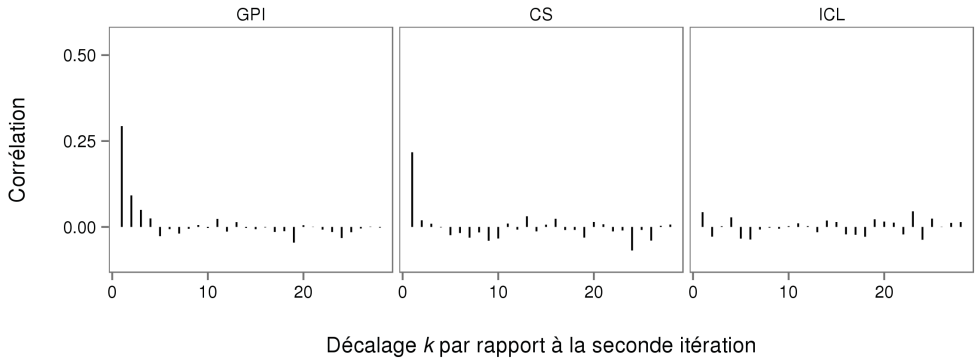


Figure 8.7. – Corrélation entre la deuxième itération et la k^{e} itération suivante, pour les flux GPI, CS et ICL. La corrélation est donnée en ordonnée, le décalage k est donné en abscisse. Chaque corrélation a été calculée à partir de 1 000 valeurs.

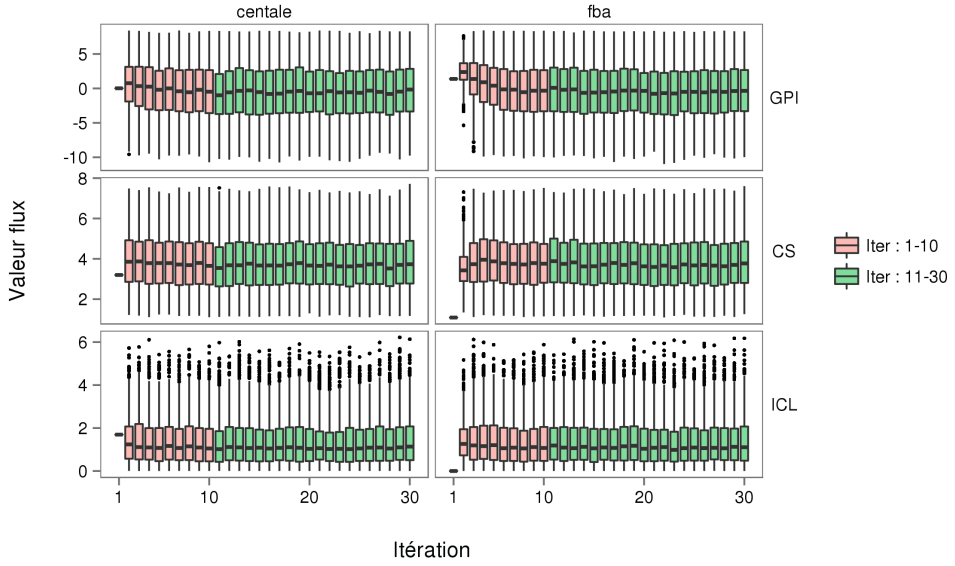


Figure 8.8. – Distribution des valeurs entre les itérations 1 et 30, selon le point de départ utilisé, pour les flux GPI, CS et ICL (de haut en bas). La valeur des flux est donnée en ordonnée (en mmol/g biom./h), et les itérations des échantillonnages sont données en abscisse. Les graphiques de gauche représentent les distributions obtenues lorsque la solution « centrale » est utilisée comme point de départ, tandis que les graphiques de droite correspondent au point de départ « maximisation de la biomasse » (« fba »). Les couleurs différencient la partie de préchauffage (en rouge), de celle considérée comme stable (en vert).

cohérents avec ceux obtenus par la méthode de diagnostic de Raftery et Lewis (disponible dans le package R "coda" [Plummer 2006]) : utilisée pour analyser 10 chaînes de Markov de 5 000 itérations, le diagnostic indique que le nombre d'itérations de préchauffage est d'environ 2,5 itérations.

Afin de confirmer ces résultats, j'ai utilisé le test d'égalité des médianes de Kruskal-Wallis (test non paramétrique) pour tester l'existence d'un effet « itération » sur la médiane des valeurs échantillonnées entre les itérations 11 à 20. Réalisé sur l'ensemble des flux, ce test indique que les médianes ne sont pas significativement différentes dans 90% des cas avec une *p-value* seuil à 0,05, et dans 95% des cas lorsque la *p-value* seuil est fixée à 0,01. Ce test a aussi été réalisé sur les distributions de valeurs issues de 1 000 échantillonnages, générés comme décrit plus haut, mais en utilisant cette fois un point de départ correspondant à la solution qui maximise la production de biomasse. Pour ce jeu de valeurs, les tests indiquent des médianes ne sont pas significativement différentes dans 95% et 97% des cas, avec respectivement une *p-value* seuil à 0,05 et 0,01. La figure 8.8 (page précédente) montre les distributions de valeurs comparées pour trois flux du système.

En conclusion, tous les résultats indiquent que la 11^{e} itération et les suivantes sont dans la partie stable des chaînes de Markov : les solutions échantillonnées à ces itérations ne dépendent pas du point de départ et peuvent donc être situées à n'importe quel endroit de l'espace des solutions. Ces analyses ont été réalisées pour les autres périodes de temps et ont conduit aux mêmes conclusions. Bien que le nombre d'itérations nécessaires paraisse faible, il faut se rappeler que la méthode `mirror` produit une série de solutions où chaque solution échantillonnée à une itération i est relativement indépendante de la solution échantillonnée lors l'itération précédente $i-1$. Cela permet à la méthode d'atteindre rapidement (en terme de nombre d'itérations) les différentes régions d'un espace des solutions, sans biais dû au point de départ de l'échantillonnage.

8.2.2. Nombre de trajectoires solutions

Afin de déterminer le nombre de trajectoires suffisant pour estimer la distribution des flux au sein de chaque « méta » espace des solutions, 20 jeux de 1 000 trajectoires solutions ont été générés.

L'algorithme utilisé pour générer chaque trajectoire est décrit dans le chapitre précédent (p. 138). Le système étudié est le métabolisme central de

C. glutamicum. Chaque trajectoire prédit l'évolution possible du système pendant le temps de culture allant de $t = 0$ à $t = 16,5$h. Chaque trajectoire contient une succession de 10 cartes de flux, et chaque carte correspond à une période de 1,5h de la culture, à l'exception de la période p_3 qui modélise le système pendant un intervalle de temps allant de $t = 3$ à $t = 6$h. Pour chaque période de chaque trajectoire, l'espace des solutions est échantillonné avec la méthode `mirror` en générant une chaîne de Markov de 20 itérations. La carte solution retenue pour décrire une période de la trajectoire est ensuite choisie au hasard parmi les solutions échantillonnées entre les itérations 11 et 20. La valeur du paramètre d, qui permet de calculer les intervalles de concentrations cibles en pondérant la variabilité expérimentale, est fixé à $d = 1$.

Afin d'évaluer si les distributions de valeurs de flux sont reproductibles d'un jeu de trajectoires à l'autre, les distributions de valeurs issues de chaque jeu de trajectoires sont comparées avec celles d'un autre jeu (jeux n° 1 *vs.* n° 11, jeux n° 2 *vs.* n° 12... jeux n° 10 *vs.* n° 20). Pour chaque comparaison, la reproductibilité de la distribution de chaque flux à chaque période de temps est évaluée en utilisant trois critères :

— La similarité générale entre la distribution de valeurs issue du premier jeu *versus* la distribution de valeurs issue du second jeu. Ce critère est évalué en utilisant le test de Kolmogorv et Smirnov (test de conformité entre deux distributions, non paramétrique).

— L'égalité des médianes entre les deux distributions de valeurs, évaluée en utilisant le test de Kruskal-Wallis (non paramétrique).

— L'égalité des variances entre les deux distributions de valeurs, évaluée en utilisant le test de Fligner-Killeen (non paramétrique).

Afin de supprimer d'éventuels artefacts, les flux dont la vitesse est totalement dépendante de celle d'un autre flux (phénomène lié aux contraintes topologiques du réseau, rencontré dans les suites linéaires de réactions), et les flux contraints par une égalité (par exemple le flux PC, dont la vitesse est contrainte à 0) ont été supprimés. Ce faisant, le nombre des flux comparés est réduit à 42. Au final, chaque comparaison de jeux de trajectoires consiste à réaliser 1 260 tests statistiques : 10 périodes de temps $\times$ 42 flux $\times$ 3 tests statistiques. Pour chaque test, une absence de différence significative contribue à montrer que 1 000 trajectoires sont suffisantes pour estimer les distributions de flux.

Afin de se placer dans des conditions défavorables vis-à-vis de la conclusion « satisfaisante », aucune correction des *p-values* n'est réalisée.

La partie gauche de la figure 8.9 présente les proportions de *p-value* non-significatives obtenues à l'issue des comparaisons entre les 10 couples de jeux de 1 000 trajectoires. Les allures, les médianes, et les variances des distributions des valeurs de flux ne sont pas significativement différentes dans respectivement 94,7%, 95,8% et 93,3% des cas. Les proportions obtenues pour chaque période de temps présentent des résultats similaires aux résultats synthétiques présentés ici. Ces résultats montrent donc qu'un nombre de trajectoires de 1 000 permet d'estimer de manière reproductible la distribution des flux pour le système modélisé.

La même procédure de test a été réalisée en comparant 10 jeux de 1 000 trajectoires avec un jeu de 10 000 trajectoires. L'idée est ici de vérifier que le fait d'augmenter le nombre de trajectoires ne modifie pas les distributions de valeurs. La partie droite de la figure 8.9 montre les proportions de *p-value* non significatives obtenues à l'issue de cette comparaison. Les allures, les médianes, et les variances ne sont pas significativement différentes dans respectivement 96,1%, 96,4% et 93,8% des cas. Ces résultats indiquent que l'augmentation du nombre de trajectoires ne modifie pas les distributions de valeurs de flux estimées.

En conclusion, un jeu de 1 000 trajectoires solutions est suffisant pour obtenir une estimation reproductible des distributions de flux au sein de chaque « méta » espace des solutions. Le fait d'augmenter le nombre de trajectoires conduit a une meilleure « résolution » des distributions (densité de points plus importante), sans toutefois modifier les paramètres de ces distributions.

8.2.3. Paramétrage de la méthode : bilan

L'objectif de cette étude était d'une part d'identifier à partir de quelle itération les solutions échantillonnées étaient indépendantes du point de départ, et d'autre part de déterminer le nombre de trajectoires solutions nécessaire pour obtenir une estimation reproductible des distributions des valeurs de flux.

Les analyses ont montré que la 11^{e} itération des chaînes de Markov correspondaient déjà à la partie stable (indépendante du point de départ) de la procédure d'échantillonnage. L'espace des solutions de chaque période p_i de

chaque trajectoire sera donc échantillonné avec 20 itérations de la méthode `mirror`, et la solution utilisée pour représenter l'état du système pendant la période p_i sera choisie au hasard parmi celles correspondant aux itérations 11 à 20. Les comparaisons des distributions issues de différents jeux de trajectoires montrent quant à elles qu'un jeu de 1 000 trajectoires est suffisant pour obtenir une estimation reproductible des distributions des valeurs de flux au sein de chacun des « méta » espaces des solutions.

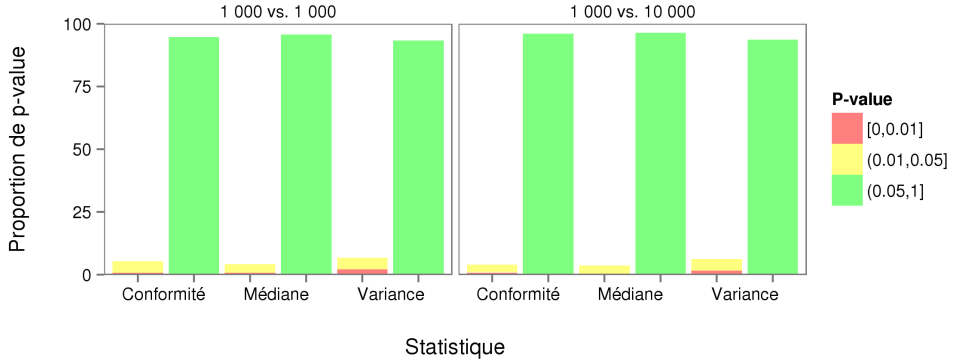


Figure 8.9. – Proportions de *p-value*, selon le niveau significativité, issues des comparaisons entre plusieurs jeux de trajectoires solutions. Le graphique de gauche représente les proportions obtenues à partir des comparaisons entre 20 jeux de 1 000 trajectoires, celui de droite représente les proportions obtenues pour les comparaisons entre 10 jeux de 1 000 trajectoires et 1 jeu de 10 000 trajectoires. La proportion de *p-value*, exprimée en pourcentage du nombre total de *p-value* obtenues pour chaque test, est représentée en ordonnée. Pour chaque graphique, l'axe des abscisses indique le type de test réalisé : test de similarité des distributions (« Conformité »), test d'égalité des médianes (« Médiane »), et test d'égalité des variances (« Variance »). Pour chaque statistique de test, les valeurs de *p-values* sont séparées en deux barres : à gauche, les *p-values* significatives (*p-value* $\leq 0{,}05$), à droite les *p-values* non significatives (*p-value* $> 0{,}05$). Le code couleur donne le niveau de significativité.

8.3. Optimisation du temps d'exécution par « retour arrière »

Dans cette section, je présente la fonctionnalité de « retour arrière » que j'ai ajoutée à l'algorithme initial de l'approche. Cette fonctionnalité a permis de fortement diminuer le temps d'exécution nécessaire pour calculer une trajectoire solution (trajectoire modélisant l'ensemble des périodes de la dynamique).

8.3.1. Algorithme initial

Afin d'identifier des trajectoires solutions qui permettent de modéliser l'ensemble d'une dynamique, l'approche développée construit chaque trajectoire solution de manière itérative :

1. Calculer les contraintes sur les flux E/S de la période p_i, définie sur l'intervalle de temps $t_i \rightarrow t_{i+1}$. Pour cela les concentrations prédites au temps t_i (issue de la période précédente) et les mesures expérimentales à t_{i+1} sont utilisées.
2. Choisir au hasard une solution pour la période p_i.
3. Prédire les concentrations en composés extracellulaires à t_{i+1}, en prenant en compte les valeurs de flux correspondant à la solution obtenue à l'étape n° 2, et les concentrations initiales au temps t_i. Revenir ensuite à l'étape n° 1 pour traiter la période suivante p_{i+1}, définie sur $t_{i+1} \rightarrow t_{i+2}$.

Initialement, lorsque cet algorithme rencontrait une période p_i pour laquelle aucune solution n'existait (compte tenu des concentrations initiales prédites à t_i et des concentrations cibles à t_{i+1}), alors la construction de la trajectoire en cours était « abandonnée », et la construction d'une nouvelle trajectoire était recommencée depuis la période initiale p_0.

Dans la section 9.2 (p. 201), je montrerai que certaines périodes de la culture de *C. glutamicum* sont difficiles à modéliser, notamment la période 3h $\rightarrow$ 6h où des solutions n'existent que pour certaines combinaisons de concentrations en composés à $t = 3$h. Conséquence de ces périodes difficilement modélisables, le temps d'exécution pour trouver une *bonne* trajectoire est particulièrement long : sans la fonctionnalité de « retour en arrière », le

temps d'exécution est d'environ 70 minutes pour trouver une *unique* trajectoire solution.

8.3.2. Optimisation de l'algorithme

En plus de l'identification des raisons pour lesquelles des périodes étaient difficilement modélisables, je me suis intéressé à l'amélioration de l'efficacité de l'algorithme.

L'approche construit une trajectoire en traitant les périodes les unes après les autres, et cette construction reprend depuis le début si une période s'avère être sans solution. Afin de diminuer le temps moyen nécessaire pour générer une trajectoire, une manière de procéder est de ne pas recommencer entièrement la construction lorsqu'une période sans solution est rencontrée. Pour cela, j'ai ajouté à l'algorithme une fonctionnalité de « retour en arrière » que je vais discuter ici.

Les figures 8.10 (entrées et sorties de la fonction) et 8.11 (algorithme de la fonction, p. 183) donnent une description synthétique de l'algorithme de la fonction principale « `CalculerSegmentTraj` » finalement utilisée pour construire les trajectoires.

Fonction : CalculerSegmentTraj($traj$, t_i, t_{i+1}, $datExp$)

Entrées :
- $traj$: objet qui représente la trajectoire jusqu'au temps t_i, contient notamment :
  - $desc$, description du réseau
  - $const$, contraintes constantes
  - $conc$, concentrations prédites jusqu'au temps t_i ($conc_{t_0}$, …, $conc_{t_i}$)
  - sol, solutions retenues aux périodes précédentes (sol_{p_0}, …, $sol_{p_{i-1}}$)
- t_i : temps de début de la période
- t_{i+1} : temps de fin de la période
- $datExp$: mesures expérimentales à tous les temps

Sortie :
- $traj$: objet qui représente la trajectoire jusqu'au temps t_{i+1}

Figure 8.10. – Fonction principale utilisée pour générer une trajectoire, entrées et sortie de la fonction.

Notons que cette fonction est récursive : une instance de la fonction (instance *mère*) peut exécuter une autre instance de la fonction (instance *fille*).

Chaque instance de la fonction recherche une solution pour une période p_i, définie sur l'intervalle de temps allant de t_i à t_{i+1}. Lorsqu'une solution est trouvée pour p_i, la fonction en cours d'exécution appelle une instance *fille* qui recherchera une solution pour la période suivante p_{i+1}. La construction d'une trajectoire se termine lorsque le temps de fin d'une période t_{i+1} correspond au temps de fin de la dynamique modélisée. Dans cette situation, si des solutions ont pu être échantillonnées, l'une de ces solutions est choisie au hasard et la fonction *fille* qui correspond à la dernière période renvoie la description de la trajectoire complètement construite (sortie « A »). Le retour de cette fonction *fille* est alors évalué par la fonction *mère*, et la description de la trajectoire complète est retournée jusqu'à la première instance de la fonction (sortie « C »).

La fonctionnalité de « retour en arrière » se situe au niveau des sorties « B » et « D » de la fonction `CalculerSegmentTraj`. Au cours de la construction d'une trajectoire, si la solution choisie à la période p_{i-1} (au niveau de l'instance *mère* de la fonction) conduit à l'absence de solution pour la période p_i (instance *fille*), l'instance fille renvoie la valeur 0 (sortie « D »). Le processus de construction revient alors à la période p_{i-1} (instance *mère*) et une autre solution est choisie pour prolonger la trajectoire vers la période p_i. Par ailleurs, si après avoir testé 10 solutions pour la période p_{i-1}, aucune ne s'avère être une « bonne » solution pour la période p_i, un autre « retour en arrière » est effectué vers la période p_{i-2} (sortie « B » de l'instance *mère*).

Pour chaque période (autrement dit, pour chaque instance de la fonction), les 10 solutions testées avant d'effectuer un « retour en arrière » correspondent aux solutions échantillonnées aux itérations 11 à 20 des chaînes de Markov (solutions indépendantes du point de départ). Ce faisant, il n'est pas nécessaire de réaliser un nouvel échantillonnage à chaque « retour en arrière », ce qui contribue aussi à réduire le temps de calcul.

8.3.3. Impact sur le temps d'exécution

Sans la fonctionnalité de « retour en arrière », le temps d'exécution moyen nécessaire pour obtenir une unique trajectoire solution est d'environ 70 minutes. Lorsque la fonctionnalité de « retour en arrière » est utilisée, ce

Début

Obtenir concentrations prédites à t_i (depuis $traj$)
Obtenir concentrations expérimentales à t_{i+1} (depuis $datExp$)
Calculer contraintes variables sur flux E/S
sys ← Appliquer contraintes constantes et variables pour définir le système

Si il existe au moins une solution pour sys **alors** :
| Échantillonner espace des solutions de sys
| $candidates$ ← Retenir solutions aux itérations 11 à 20
| $test$ ← 0
|
| **Tant que**
| (1) il reste au moins une solution dans $candidates$ **et**
| (2) $test$ == 0
| **faire :**
| | sol ← Choisir une solution parmi $candidates$
| | $candidates$ ← Retirer sol de $candidates$
| | $conc_{ti+1}$ ← Prédire conc. à t_{i+1} selon la solution choisie
| | Retenir $conc_{ti+1}$ et sol dans $traj$
| |
| | **Si** t_{i+1} est le dernier temps de la dynamique **alors** :
| | | **Retourner** $traj$ *Sortie A*
| | **Sinon** :
| | | $test$ ← CalculerSegmentTraj($traj$, t_{i+1}, t_{i+2}, $datExp$)
| | **Fin si**
| |
| **Fin tant que**
|
| **Si** $test$ == 0 **alors** :
| | **Retourner** 0 *Sortie B*
| **Sinon** :
| | $traj$ ← $test$
| | **Retourner** $traj$ *Sortie C*
| **Fin si**
|
Sinon :
| **Retourner** 0 *Sortie D*
Fin si

Fin

Figure 8.11. – Fonction principale utilisée pour générer une trajectoire, algorithme synthétique. Les différentes sorties possibles de la fonction sont indiquées en marge par un code couleur (sorties A, B, C et D).

temps d'exécution moyen passe à environ 7 minutes et 45 secondes (plus ou moins 2 minutes à un écart-type). Les temps moyens ont été estimés à partir de la génération de 1 000 trajectoires pour l'algorithme sans « retour arrière », et de 9 000 trajectoires pour l'algorithme avec « retour arrière ». Chaque trajectoire modélise la dynamique de la culture de *C. glutamicum* allant de $t = 0$ à $t = 16,5$h, sans prendre en compte les mesures à $t = 4,5$h. Les trajectoires ont été générées en utilisant des processeurs Intel Xeon E5-4607, où chaque cœur est cadencé à 2,2 gigahertz.

Lors de l'utilisation de la fonctionnalité de « retour en arrière », l'essentiel du temps de calcul correspond à la recherche d'une « bonne » solution pour la période de temps allant de $t = 3$ à $t = 6$h. La figure 8.12 montre le nombre moyen de chaque type de sortie de la fonction `CalculSegmentTraj` lors de la construction d'une trajectoire, selon la période de la dynamique traitée. Comme on peut le voir, les « retours en arrière » sont effectués en très grande majorité lors du traitement de la période p_3 (période allant de $t = 3$ à $t = 6$h), où chaque sortie de type D correspond à une situation où les concentrations prédites à $t = 3$h impliquent une absence totale de solution pour cette période. Le fait que des sorties de type B soient utilisées pour la période $p_{1,5}$ indique par ailleurs que l'obtention d'une bonne solution pour la période p_3 peut aussi nécessiter l'essai de plusieurs solutions pour la période $p_{1,5}$.

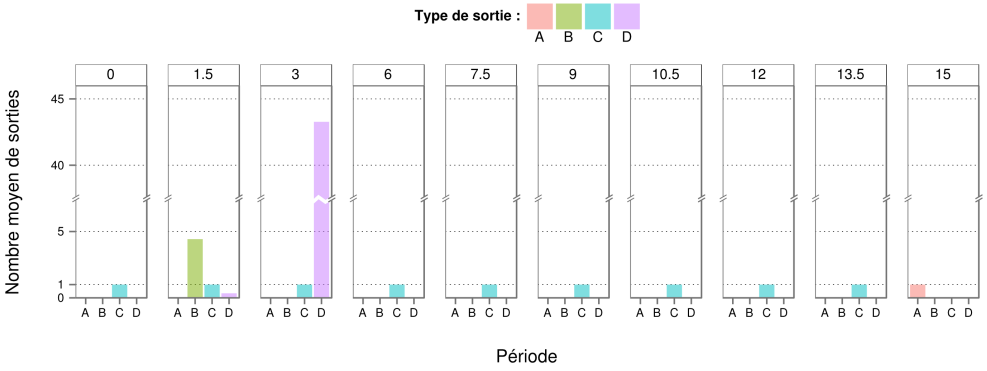


Figure 8.12. – Nombre moyen de chaque type de sortie de la fonction `CalculSegmentTraj` selon la période de la dynamique traitée. Chaque graphique correspond à une période, dont le temps de début est donné au-dessus. Le nombre de sorties est donné en ordonnée. Chaque graphique correspond à l'une des périodes traitées. Le code couleur indique le type de sortie.

8.3.4. Optimisation du temps d'exécution par « retour en arrière » : bilan

Dans cette section, il s'agit de présenter l'optimisation réalisée sur l'algorithme de génération des trajectoires solutions. L'ajout de la fonctionnalité de « retour en arrière » m'a permis de passer d'un temps moyen de 70 minutes à environ 7 minutes et 45 secondes pour générer une trajectoire. Cela représente une accélération de la vitesse de génération par un facteur voisin de 9.

Remarquons que, comme (i) le calcul de chaque trajectoire est indépendant et (ii) l'ensemble de l'approche est implémenté dans le langage R, la génération des trajectoires peut facilement être parallélisée sur plusieurs cœurs, processeurs, ou machines différentes.

Ainsi, en utilisant 12 processus simultanés, le temps nécessaire pour générer un jeu de 1 000 trajectoires était à l'origine d'environ 4 jours. Dans le même laps de temps, 9 000 trajectoires peuvent maintenant être construites. Cela permet potentiellement d'obtenir une meilleure résolution des distributions de flux, et de tester plus facilement différentes hypothèses sur le fonctionnement du métabolisme étudié.

Le nombre de 10 essais avant d'effectuer un « retour en arrière » d'une période p_{i-1} vers une période p_{i-2} (sortie « B ») peut être discuté : il s'agit avant tout de trouver un compromis entre (i) rejeter suffisamment vite la solution choisie pour p_{i-2}, car les concentrations prédites à partir de cette solution peuvent conduire indirectement à une « impasse » (les concentrations prédites à t_{i-1} permettent encore de trouver des solutions, mais quelles que soient ces solutions, les concentrations qui seront prédites à t_i ne pourront plus déboucher sur des solutions pour la période $t_i \rightarrow t_{i+1}$), et (ii) ne pas rejeter trop vite la solution choisie pour p_{i-2}, car elle n'est pas nécessairement en cause dans le fait qu'aucune des solutions échantillonnées à p_{i-1} ne permet de poursuivre la trajectoire à p_i (la solution p_{i-1} est peut être la cause). Remarquons que le fait que les 10 solutions testées correspondent à des itérations successives d'une même chaîne de Markov ne diminue pas la qualité de l'échantillonnage des « méta » espaces des solutions, puisque, au mieux, seule une des solutions sera retenue pour poursuivre la construction de la trajectoire.

Chapitre 9

Discussions

Dans une première section (p. 188), j'aborde le problème du temps 4,5h qui n'a pas pu être modélisé lors de l'étude de la dynamique du métabolisme de *C. glutamicum* (cf. chapitre 7, p. 144). Cette première discussion sera l'occasion d'émettre différentes hypothèses sur les origines possibles du problème.

Dans la seconde section (p. 201), je m'intéresse ensuite aux distributions de concentrations prédites à l'issue des trajectoires solutions. Au fil de cette seconde discussion, je montre notamment que les distributions de concentrations prédites à l'issue des trajectoires solutions peuvent être utilisées afin d'améliorer le modèle.

9.1. Absence de trajectoire solution incluant le temps 4,5h

Lors de l'étude de la dynamique du métabolisme central de *C. glutamicum*, aucune trajectoire solution n'a pu être construite sur l'ensemble de la culture lorsque le temps 4,5h est pris en compte. Plus exactement, des trajectoires peuvent être générées jusqu'au temps 4,5h, mais aucune solution ne peut ensuite être trouvée pour la période suivante allant de $t = 4,5$ à $t = 6$h.

Ce désaccord entre l'approche employée pour modéliser la dynamique et les données expérimentales peut avoir plusieurs origines :

— un problème au niveau des mesures expérimentales, par exemple des valeurs erronées ;

— un problème lié aux connaissances biologiques intégrées dans le modèle, comme une description inexacte du réseau métabolique ;

— un problème lié à l'approche que j'ai développée, comme un biais dans les solutions échantillonnées ;

— une combinaison des trois situations.

Dans cette section, j'aborde ces différents aspects à travers plusieurs analyses.

9.1.1. Problème lié aux mesures de concentration de glucose

Dans le cadre de mon étude du métabolisme de *C. glutamicum*, les données expérimentales correspondent à des concentrations de métabolites extracellulaires et de biomasse mesurées expérimentalement à différents temps de la culture. Ces données ont été obtenues à partir de trois cultures indépendantes (« répétitions biologiques »).

L'absence de solution au temps 4,5h pourrait être due à la cinétique des concentrations de glucose mesurées pour la répétition n° 2 (figure 9.1). Comparée aux autres répétitions, la répétition n° 2 présente une consommation de glucose plus importante pendant la période $1,5\text{h} \rightarrow 3\text{h}$, suivi d'une consommation plus faible pendant la période $4,5\text{h} \rightarrow 6\text{h}$.

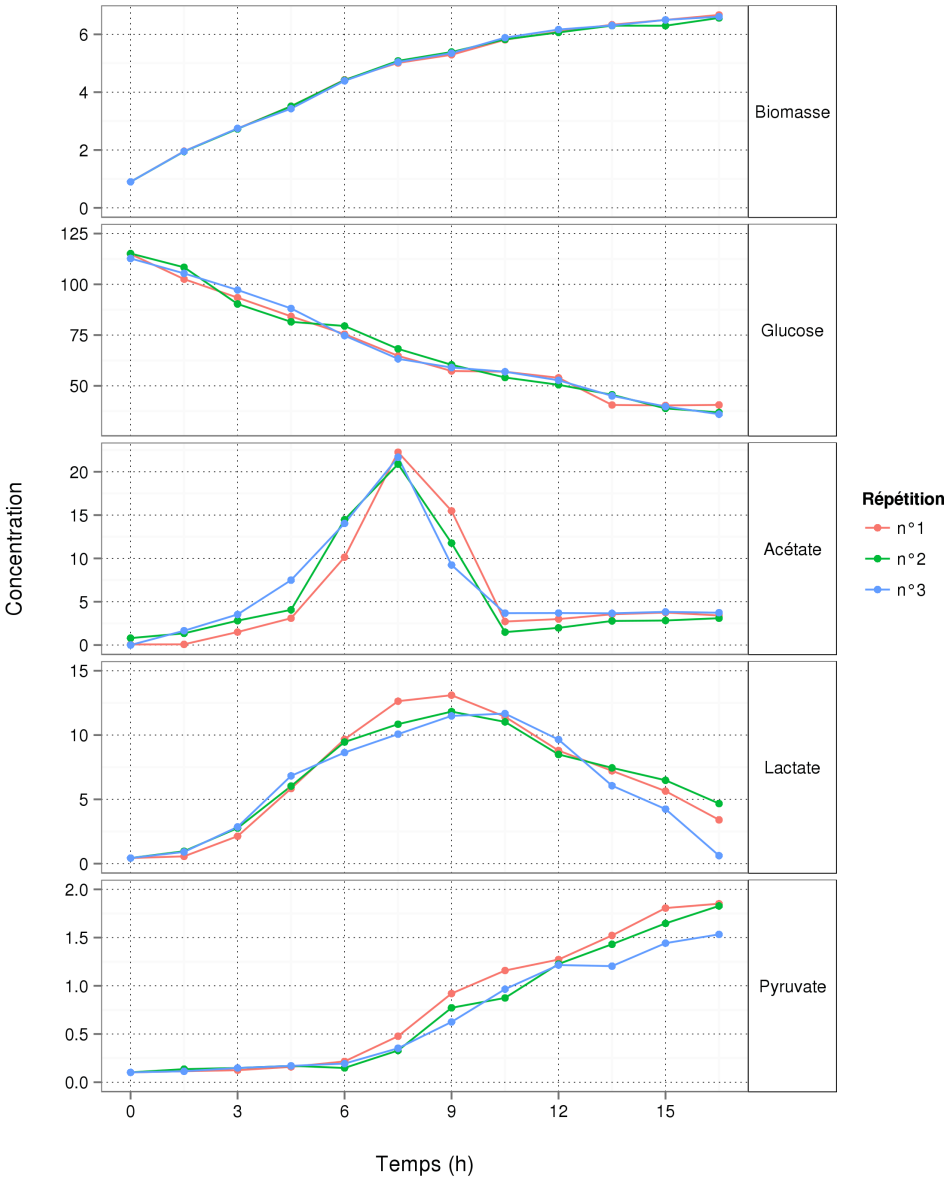


Figure 9.1. – Concentration des métabolites et de la biomasse au cours du temps, pour chaque répétition. La concentration de la biomasse est exprimée en g/L, les concentrations des métabolites sont exprimées en mmol/L. Le code couleur différencie les répétitions.

Afin d'évaluer les conséquences de cette cinétique, j'ai calculé, pour chaque répétition, le bilan carbone (bilan des flux de carbone entrant et sortant des cellules) à chaque période de temps. Compte tenu des données expérimentales disponibles, le calcul du bilan carbone nécessite de poser les hypothèses suivantes :

1. le système est à l'état stationnaire pendant chaque période (il n'y a pas de variation des concentrations de métabolites dans les cellules) ;
2. l'acétate, la biomasse, le glucose, le lactate et le pyruvate sont les seuls composés carbonés « échangés » entre les cellules et le milieu de culture (en plus du CO_2 qui n'a pas été mesuré) ;
3. la quantité de carbone représenté par la biomasse peut être déterminée en utilisant l'équation de biomasse du modèle ;
4. la différence entre le carbone total consommé et celui produit peut être attribuée au CO_2 (non mesuré).

La figure 9.2 montre, pour la répétition n° 2, la contribution de chaque composé dans le flux total de carbone consommé et produit par les cellules. Pendant la période $4,5\text{h} \rightarrow 6\text{h}$, on voit que le glucose consommé ne représente que 19% du carbone total entrant (100% du carbone total entrant correspond à la quantité nécessaire pour équilibrer le bilan, étant donné le carbone total sortant). Afin de maintenir un bilan carbone équilibré, le reste du carbone consommé doit être attribué au CO_2. Cette situation est aussi rencontrée pour la répétition n° 1, mais de manière beaucoup moins marquée puisque la consommation de glucose représente alors 88% du flux de carbone total entrant. Compte tenu des mesures, on peut envisager au moins deux explications simples : soit les cellules sont effectivement capables d'utiliser le CO_2 comme source de carbone, soit les mesures réalisées aux temps 3 et 4,5h de la répétition n° 2 sont erronées.

Afin d'évaluer la première explication, j'ai testé s'il était possible, d'après le modèle, de prédire des répartitions de flux qui conduisent à une consommation nette de CO_2. Pour cela, j'ai utilisé la méthode FBA avec pour fonction *objectif* la valeur du flux d'échange de CO_2 (flux CO2x), et pour optimum la maximisation de l'entrée de CO_2. Lorsque les seules contraintes appliquées sur le système sont les contraintes constantes, il est possible d'obtenir une légère consommation de CO_2 (3,17 mmol/g biom./h, valeur donnée à titre indicatif). Cependant, cette consommation est accompagnée par un flux très élevé d'excrétion de pyruvate (81,15 mmol/g biom./h). Dès lors que les flux E/S sont contraints de manière cohérente avec les concentrations en métabolites et en biomasse mesurées expérimentalement, il n'existe plus

aucune solution permettant la consommation de CO_2, et ce quelle que soit la période de la culture considérée. Notons ici qu'il existe, au sein des procaryotes, des mécanismes qui permettent la co-assimilation de CO_2 avec un autre métabolite carboné [Fuchs 2011]. Cependant, un tel phénomène n'a pas été décrit chez *C. glutamicum*. En conclusion, l'hypothèse d'un flux net d'entrée de CO_2 dans les cellules est improbable.

La première possibilité étant rejetée, on peut se poser la question du niveau de fiabilité à accorder aux deux mesures de concentrations de glucose réalisées pour la répétition n° 2 aux temps 3 et 4,5h. Afin de tester si l'absence de trajectoires solutions est uniquement due à ces « mauvaises » mesures de glucose, j'ai corrigé les concentrations de glucose de la répétition n° 2 aux temps 3 et 4,5h, de telle sorte que les flux de consommation de glucose pendant les périodes $1,5\text{h} \rightarrow 3\text{h}$ et $4,5\text{h} \rightarrow 6\text{h}$ soient similaires à ceux des deux autres répétitions. Ce faisant, le bilan carbone de la répétition n° 2 ne présente plus qu'un léger déficit au niveau de la période $4,5\text{h} \rightarrow 6\text{h}$ (le glucose représente 88% du carbone total sortant).

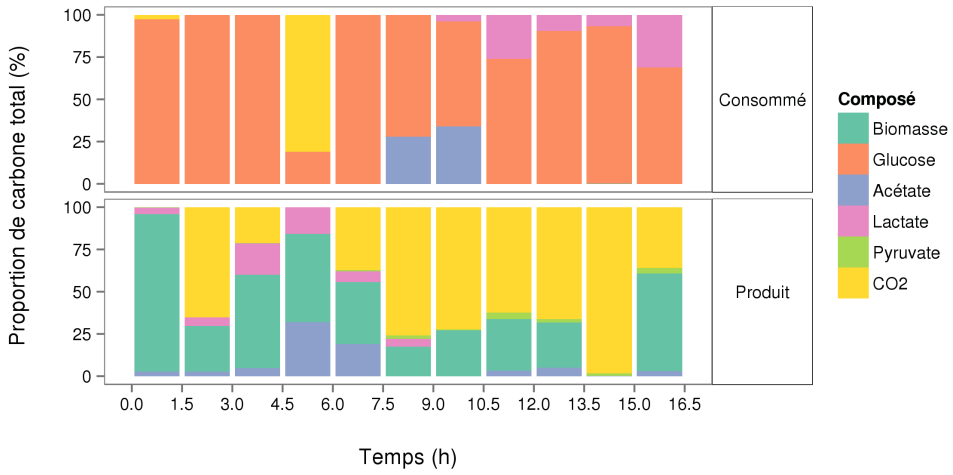


Figure 9.2. – Bilans carbone estimés à partir des mesures expérimentales de la répétition n° 2, pour chaque période de temps. Le graphique du haut montre la proportion que représente chaque composé dans le flux de carbone total consommé, le graphique du bas donne le même type d'information pour le flux de carbone total produit. La contribution du CO_2 (non mesuré) est estimée de manière à équilibrer le bilan entre carbone consommé et carbone produit. Les différentes périodes sont données en abscisse, le code couleur différencie les composés.

En utilisant l'approche par trajectoires solutions, il devient possible de générer des trajectoires lorsque les mesures expérimentales incluant les concentrations de glucose « modifiées » sont utilisées. L'existence de trajectoires solutions alors que le bilan de carbone de la répétition n° 2 reste en déficit s'explique par le fait que l'approche développée intègre la variabilité des mesures expérimentales, ce qui permet de trouver des solutions au bilan carbone neutre ou positif pendant la période $4,5\text{h} \rightarrow 6\text{h}$.

Si l'on revient aux valeurs expérimentales réelles, on peut donc proposer une première explication quant à l'origine de l'incohérence entre le modèle et les données biologiques : les concentrations de glucose mesurées aux temps 3 et 4,5h sont erronées. Toutefois, les deux concentrations *a priori* aberrantes ont été mesurées au sein d'une même culture, correspondent à deux temps consécutifs, et leurs valeurs sont toutes les deux inférieures à celles attendues. On peut se demander si ces constats ne sont vraiment que pure coïncidence, ou si les concentrations de glucose rapportées sont les concentrations réellement présentes dans le milieu de culture. Si les concentrations sont réellement celles présentes, d'autres causes sont alors à envisager.

9.1.2. Problèmes liés aux connaissances biologiques

Dans la sous-section précédente, j'ai calculé les bilans carbone en posant deux hypothèses importantes : (*i*) tous les composés carbonés, à l'exception du CO_2, ont été mesurés, et (*ii*) l'équation de biomasse utilisée est valable dans les conditions expérimentales de l'étude. Si l'on revient sur ces hypothèses, d'autres causes peuvent être envisagées pour expliquer les incohérences entre le modèle et les données.

Production d'un métabolite non mesuré

On peut ainsi revenir sur la première hypothèse et envisager qu'au moins un autre composé carboné extracellulaire n'a pas été mesuré. La cinétique de consommation/production de ce métabolite X pourrait, par exemple, correspondre à :

— Une production du métabolite X par les bactéries en début de culture, notamment pendant la période $1,5\text{h} \rightarrow 3\text{h}$. La production de X représenterait ainsi une part du carbone sortant attribué au CO_2 dans la figure 9.2 (p. 191).

— Une reconsommation du métabolite X par les bactéries pendant la période $4,5\text{h} \rightarrow 6\text{h}$. La consommation de X représenterait alors une partie ou la totalité du carbone entrant que j'ai attribué au CO_2 dans la figure 9.2 (p. 191).

Notons ici que l'espèce *C. glutamicum* a été identifiée dans les années 1950 par Kinoshita et coll. lors de la recherche d'organismes capables de produire naturellement du glutamate [Kinoshita 1957] [1]. La croissance en limitation en biotine de souches de *C. glutamicum* a par la suite été utilisée comme procédé industriel pour provoquer une excrétion importante de glutamate par les cellules. Par ailleurs, les acides aminés nécessitant une dérivation pour être détectés en ultra-violet (par HPLC), ils n'ont pu être détectés lors des expérimentations qui ont conduit aux données que j'ai utilisé pour les analyses. Par conséquent, sous l'hypothèse qu'une production suivie d'une reconsommation de métabolite(s) X a effectivement eu lieu au cours de la culture étudiée, le (ou les) métabolite(s) impliqué(s) dans ces phénomènes pourrai(en)t potentiellement être des acides aminés.

Équation de biomasse différente dans les conditions expérimentales étudiées

Pour ce qui est de la deuxième hypothèse, l'équation de biomasse utilisée dans le modèle a été déterminée dans des conditions de croissance exponentielle [Cocaign-Bousquet 1996]. Or, les cultures bactériennes sont ici différentes, puisque réalisées dans des conditions de limitation en biotine, ce qui induit une courbe de croissance arithmétique. La biotine est un cofacteur de la pyruvate carboxylase [Peters-Wendisch 1998] qui contribue à l'apport de métabolites à quatre carbones vers le cycle de Krebs, et de l'acetyl-CoA carboxylase [Jäger 1996] qui catalyse la première étape de synthèse des acides gras. Le déficit en biotine induit un fonctionnement différent du métabolisme. Cela est par exemple matérialisé par des phénomènes d'excrétion, suivis de reconsommations d'acides organiques — qui

1. l'espèce aujourd'hui nommée *C. glutamicum* avait à l'époque été rattachée au genre *Micrococcus*

ne sont pas observés lors de la croissance exponentielle sur glucose — ou encore par la plus faible teneur en phospholipides de l'enveloppe de *C. glutamicum* [Kimura 2005]. Ce fonctionnement — cette répartition des flux différente — implique probablement des besoins en précurseurs nécessaires à la production de biomasse et des besoins énergétiques qui sont différents de ceux déterminés dans des conditions de croissance exponentielle. Sous l'hypothèse que l'intégration du carbone dans la biomasse soit moins importante en condition de limitation de biotine, le flux de sortie de carbone à attribuer à la biomasse deviendrait moins important. Cela pourrait expliquer, tout du moins en partie, le déficit de carbone entrant pendant la période $4,5\text{h} \rightarrow 6\text{h}$.

9.1.3. Caractérisation de l'absence de solution du point de vue de l'approche « trajectoires solutions »

Aucune trajectoire solution n'a pu être générée lorsque les mesures du temps 4,5h sont prises en compte. Cependant, lors des analyses réalisées pour déterminer la méthode d'échantillonnage la plus adaptée (p. 164), des solutions ont pu être obtenues pour la période $4,5\text{h} \rightarrow 6\text{h}$ lorsque cette période est échantillonnée indépendamment.

En comparaison d'une analyse de la dynamique où chaque période est considérée de manière indépendante, l'approche par trajectoires solutions ajoute une « contrainte » de faisabilité sur l'ensemble du temps de culture : chaque trajectoire doit permettre de prédire des cinétiques de concentrations qui soient cohérentes avec les mesures expérimentales. Des concentrations cohérentes sont matérialisées par des intervalles de concentrations cibles pour chaque temps de la dynamique. Pour chaque trajectoire, la contrainte de faisabilité est introduite en prenant en compte le jeu de concentrations prédites au temps t_i — concentration de biomasse, de glucose, d'acétate, de lactate et de pyruvate — dans le calcul des contraintes d'inégalité sur les flux E/S de la période suivante $t_i \rightarrow t_{i+1}$.

Positionnement de la contrainte de faisabilité vis-à-vis de l'absence de solution

Afin de tester si la « contrainte » de faisabilité qui conduit à l'absence de solution est exercée par une période suivante, j'ai généré un jeu de 1 000

trajectoires solutions avec un temps de fin de la dynamique fixé à $t = 6$h. De cette manière, la période $4,5\text{h} \rightarrow 6\text{h}$ devient la dernière période. Dans ces conditions, des solutions sont identifiées jusqu'au temps $t = 4,5$h, mais aucune solution n'est trouvée pour la période $4,5\text{h} \rightarrow 6\text{h}$. Cela signifie donc que la « contrainte » de faisabilité est exercée par des prédictions situées en amont de cette période.

La figure 9.3 montre la distribution des concentrations prédites à l'issue des trajectoires solutions allant jusqu'au temps 4,5h. On remarque que les concentrations en glucose prédites au temps 4,5h sont clairement dans la partie basse de l'intervalle des concentrations cibles. Le positionnement des concentrations prédites sur l'ensemble de la culture sera discuté dans la section 9.2 (p. 201).

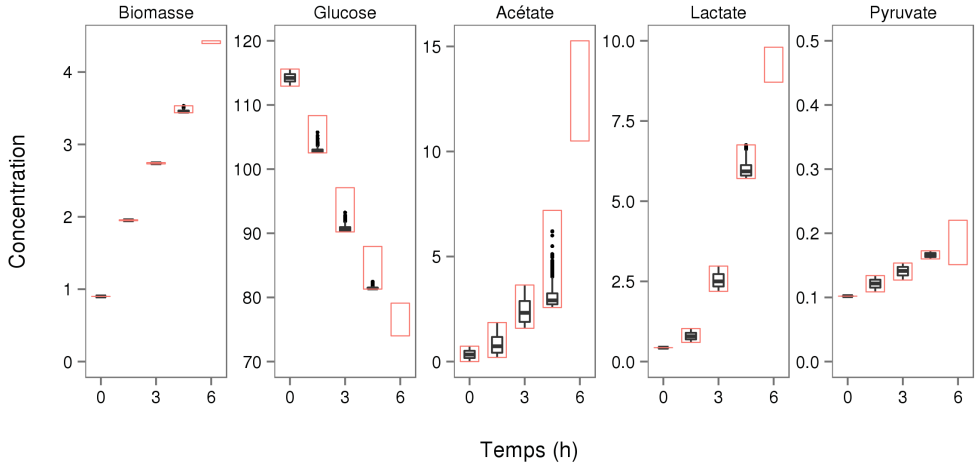


Figure 9.3. – Distributions des concentrations en composés prédites lors de la génération de trajectoires solutions allant de $t = 0$ à $t = 4,5$h. Chaque graphique correspond à un composé. La concentration, en ordonnée, est exprimée en g/L pour la biomasse, et en mmol/L pour les autres composés. Le temps est indiqué en abscisse. Les boîtes à moustaches représentent la distribution des valeurs prédites par les trajectoires solutions. Les rectangles de couleur rouge représentent les intervalles de concentrations cibles utilisés pour calculer les contraintes sur les flux E/S. Les trajectoires solutions ne peuvent être prolongées jusqu'au temps 6h, aucune concentration ne peut être prédite pour ce temps.

Caractérisation des combinaisons de concentrations solutions pour la période $4,5\text{h} \rightarrow 6\text{h}$

Il s'agit ici de comprendre en quoi le positionnement des concentrations prédites au temps 4,5h empêche de prolonger les trajectoires solutions jusqu'au temps 6h. Pour cela, j'ai utilisé une approche par « grille » pour caractériser l'existence ou non d'une solution selon la combinaison des concentrations initiales au temps 4,5h. Une combinaison de concentrations correspond à une valeur de concentration pour chaque composé (glucose, biomasse, acétate, lactate et pyruvate).

Pour chaque composé, j'ai calculé un intervalle de concentrations initiales ($t = 4,5\text{h}$), borné par la moyenne des mesures expérimentales plus ou moins un écart-type. Les intervalles correspondent ainsi aux intervalles « cibles » qui sont calculés dans l'approche à base de trajectoire solution (avec le paramètre d fixé à 1). Chaque intervalle est discrétisé en 20 valeurs de concentrations équidistantes, allant de la valeur minimale à la valeur maximale de l'intervalle. Les concentrations en pyruvate n'ayant qu'un impact négligeable sur l'existence ou non d'une solution (fait constaté dans une analyse préliminaire), l'intervalle associé au pyruvate n'est pas discrétisé et est remplacé par une unique valeur correspondant à la moyenne des mesures expérimentales (centre de l'intervalle). Ce faisant, on dispose de 20 concentrations initiales pour la biomasse, le glucose, l'acétate et le lactate, et d'une unique valeur pour le pyruvate. L'ensemble de ces concentrations représente 160 000 combinaisons de concentrations possibles (4^{20}), où chaque combinaison contient une valeur de concentration pour chaque composé.

J'ai ensuite utilisé la méthode FBA pour tester si chaque combinaison de concentrations au temps $t = 4,5\text{h}$ permet d'obtenir au moins une solution pour la période $4,5\text{h} \rightarrow 6\text{h}$. Les contraintes variables sur les flux E/S sont calculées de façon à ce que les concentrations prédites à l'issue d'une carte de flux soient comprises dans des intervalles de concentrations cibles au temps $t = 6\text{h}$ (intervalle borné par la moyenne des mesures expérimentales plus ou moins un écart-type).

La figure 9.4 présente, pour chaque composé, la proportion de combinaisons pour lesquelles une solution existe en fonction de la concentration utilisée au temps 4,5h. Pour un composé donné, une proportion de 100% indique que la concentration initiale du composé permet toujours d'obtenir une solution, quelles que soient les concentrations des autres composés. Ces résultats montrent qu'il existe des combinaisons de concentrations qui

permettent d'obtenir une solution pour la période $4,5\text{h} \rightarrow 6\text{h}$. Cela était attendu, puisque l'échantillonnage de cette période est possible lorsqu'elle est considérée de manière indépendante.

Toutefois, toutes les combinaisons de concentrations ne permettent pas de déboucher sur une solution. Le composé qui a clairement la plus forte influence sur l'existence ou non d'une solution est le glucose : il existe toujours une solution lorsque la concentration en glucose est « élevée » à $t = 4,5\text{h}$ (relativement à l'intervalle des concentrations autorisées), et il n'existe aucune solution lorsque la concentration de glucose est « faible », et ce peu importe la concentration des autres composés. Ceci est à relier avec les analyses présentées précédemment. Comme le glucose est l'unique composé carboné qui est consommé pendant la période $4,5\text{h} \rightarrow 6\text{h}$, et que du CO_2 ne peut pas être co-assimilé, la quantité de glucose consommé doit être suffisante pour fournir l'ensemble du carbone nécessaire à la production de biomasse, d'acétate et de lactate. Compte tenu des concentrations cibles à $t = 6\text{h}$, une concentration en glucose trop « faible » à $t = 4,5\text{h}$ ne permet pas d'obtenir un flux d'entrée de glucose suffisant pour « répondre » à l'ensemble de ces besoins.

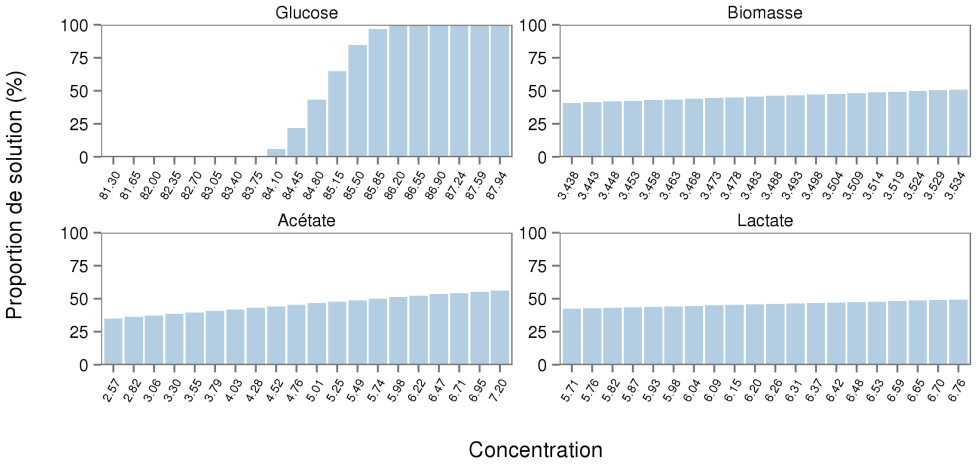


Figure 9.4. – Proportion de combinaisons de concentrations pour lesquelles une solution existe pour la période $4,5\text{h} \rightarrow 6\text{h}$, en fonction de la concentration utilisée au temps 4,5h, vue pour chaque composé. La proportion est indiquée en ordonnée. La concentration, en abscisse, est exprimée en g/L pour la biomasse, et en mmol/L pour les autres composés. Pour un composé et une concentration donnés, la proportion est calculée à partir de 8 000 tests, soit 8 000 combinaisons de concentrations avec les autres composés.

Dans l'hypothèse où le glucose est le seul substrat consommé, il doit par ailleurs fournir l'énergie nécessaire à la survie de la cellule. Cette énergie, qui est associée à la maintenance de l'intégrité des cellules (énergie non associée à la division cellulaire), est modélisée dans le système par un flux de dissipation d'ATP (flux ATPm ; ATP → ADP + Pi ; flux discuté dans la section suivante, p. 209), dont la vitesse minimale est contrainte à 5 mmol/g biom./h. Notons ici que le fait de relâcher cette contrainte à 2 mmol/g biom./h n'a pas d'effet notable sur les proportions de solutions obtenues. Cela peut s'expliquer par le fait que la concentration en biomasse au temps 4,5h est relativement faible. En conséquence, la portion de glucose consommé qui « doit » être dédiée à la maintenance cellulaire est faible.

L'influence des autres composés sur l'existence ou non d'une solution est minime, et n'est visible que pour des concentrations de glucose « intermédiaires ». Dans cette situation, plus la concentration de biomasse, de lactate et d'acétate est « faible » au départ, moins il y a de combinaisons de concentrations permettant d'obtenir une solution.

D'après le modèle, l'existence d'une solution pour la période $4,5\text{h} \rightarrow 6\text{h}$ nécessite donc une concentration en glucose « moyenne » à « élevée » à $t = 4,5\text{h}$ (relativement à l'intervalle des concentrations de départ autorisées), et la concentration des autres composés n'a que peu d'influence. Ces résultats sont à confronter avec les distributions des concentrations prédites lors de la génération de trajectoires solutions (figure 9.3, p. 195). D'après cette figure, la majorité des concentrations de glucose prédites au temps $4,5\text{h}$ se situent dans la partie basse de l'intervalle des concentrations autorisées, ce qui implique une absence de solution pour la période $4,5\text{h} \rightarrow 6\text{h}$. Par ailleurs, quelques trajectoires solutions conduisent à une concentration en glucose « moyenne » (relativement à l'intervalle des concentrations), mais ces trajectoires prédisent par ailleurs des concentrations en lactate et en acétate « faibles » (toujours relativement aux intervalles des concentrations autorisées).

Compte tenu de ces résultats, une question essentielle qui n'a pas été évoquée est « *pourquoi les concentrations prédites par les trajectoires solutions se situent-elles dans la partie basse des intervalles de concentrations cibles ?* ». J'aborderai cette question dans la section suivante (p. 201).

9.1.4. Absence de trajectoire solution incluant le temps 4,5h : bilan

Dans cette section, je me suis intéressé aux causes possibles de l'absence de trajectoire solution incluant les mesures expérimentales du temps 4,5h. Une première origine peut être les mesures expérimentales : (*i*) les concentrations de glucose mesurées au temps 3 et 4,5h semblent anormalement faibles pour la répétition n^o 2, et (*ii*) au moins un autre composé, qui n'aurait pas été mesuré, pourrait être produit et reconsommé par les cellules. Une deuxième origine touche aux connaissances biologiques intégrées dans le modèle : (*iii*) dans les conditions de croissance en limitation en biotine, les besoins en précurseurs et en énergie pour la croissance bactérienne sont probablement différents de ceux utilisés dans le modèle. Enfin (*iv*), une autre origine possible pourrait être liée à l'approche par « trajectoires solutions » employée.

Afin de valider ou d'invalider les hypothèses émises, plusieurs expériences devraient être réalisées. Tout d'abord, la réalisation d'au moins une autre culture dans des conditions similaires permettrait de comparer les répétitions *a priori* aberrantes (hypothèse *i*) avec de nouvelles mesures. Idéalement, cette expérience serait accompagnée de la mesure des gaz échangés pendant la croissance (O_2 et CO_2). La connaissance des flux de CO_2 permettrait ainsi d'affiner le bilan carbone, et de consolider l'hypothèse de la production d'un autre composé (hypothèse *ii*). Ces mesures pourront par ailleurs être intégrées comme contraintes sur les flux E/S du modèle.

Les concentrations de métabolites que j'ai utilisées ont été déterminées en 2007 par des dosages enzymatiques et une analyse par chromatographie liquide. Les appareils d'analyse haut débit, comme les spectromètres de masse ou de résonnance magnétique nucléaire, sont aujourd'hui plus faciles d'accès. Ce type d'appareillage pourrait être utilisé afin de détecter (et de quantifier) de manière exhaustive les éventuels autres métabolites extracellulaires produits et consommés au cours de la culture (hypothèse *ii*).

Enfin, une analyse ICPMS (*inductively coupled plasma mass spectrometry*) pourrait permettre de déterminer la formule élémentaire de la biomasse produite lors de la croissance en condition de limitation en biotine. À la manière de la formule chimique d'une molécule, la formule élémentaire de la biomasse donne les proportions relatives des atomes de carbone, d'hydrogène, d'oxygène, d'azote et de phosphate qui constituent un gramme de

biomasse. Par exemple, Dominguez et coll. ont déterminé que la formule élémentaire des cellules de *C. glutamicum* en condition de croissance exponentielle était $C_{4,5}H_8O_{2,2}N$ [Dominguez 1998]. Des proportions en atomes différentes, lorsque les cellules sont cultivées en condition de limitation en biotine, seraient un bon indicateur d'une évolution des besoins en précurseurs et en énergie nécessaires à la croissance bactérienne (hypothèse *iii*).

9.2. Analyse des distributions de concentrations prédites

Les concentrations en composés prédites par les trajectoires solutions, en particulier les concentrations en glucose, se situent majoritairement dans la partie basse des intervalles de concentrations cibles. Ceci est illustré par la figure 9.5 qui montre les concentrations prédites à l'issue d'un jeu de 1 000 trajectoires solutions, allant du temps $t = 0$ au temps $t = 16,5$h, sans prendre en compte les mesures au temps $t = 4,5$h.

Dans cette section, je vais discuter des phénomènes qui sont à l'origine de ces distributions excentrées : d'une part l'absence de solution pour certaines combinaisons de concentrations en composés extracellulaires, et d'autre part un phénomène de dissipation énergétique excessive du à quelques flux internes du système.

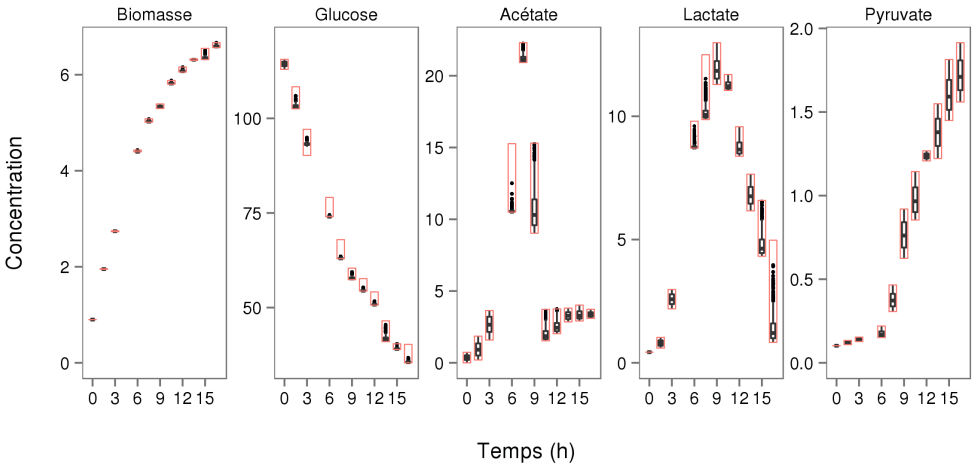


Figure 9.5. – Distributions des concentrations en composés prédites lors de la génération de trajectoires solutions allant de $t = 0$ à $t = 16,5$h, sans prise en compte des mesures au temps $t = 4,5h$. Chaque graphique correspond à un composé. La concentration, en ordonnée, est exprimée en g/L pour la biomasse, et en mmol/L pour les autres composés. Le temps est indiqué en abscisse. Les boîtes à moustaches représentent la distribution des valeurs prédites par les trajectoires solutions. Les rectangles de couleur rouge représentent les intervalles de concentrations cibles (moyenne des mesures plus ou moins un écart-type) qui ont été utilisés pour calculer les contraintes sur les flux E/S.

9.2.1. Absence de solution pour certaines combinaisons de concentrations

L'objectif est ici d'étudier en quoi les mesures expérimentales, utilisées pour déterminer les contraintes sur les flux E/S du modèle, ont un impact sur les distributions prédites par les trajectoires solutions.

Caractérisation des concentrations « solutions » des périodes

J'ai réalisé une analyse par « grille » similaire à celle présentée lors de la caractérisation des combinaisons de concentrations solutions de la période $4,5\text{h} \rightarrow 6\text{h}$ (p. 194). Pour chaque période p_i, différentes combinaisons de concentrations sont testées afin de déterminer si elles permettent d'obtenir une solution du système. Dans l'analyse que je présente ici, une combinaison de concentrations est définie à la fois par les concentrations en début de période (t_i) et par celles en fin de période (t_{i+1}). Ce faisant, je peux déterminer les concentrations initiales à t_i qui sont solutions de la période p_i, tout en évaluant les concentrations atteignables au temps t_{i+1} compte tenu de ces concentrations initiales.

Pour chaque temps t de la culture modélisée, et pour chaque composé, j'ai calculé un intervalle des concentrations autorisées, défini par la moyenne des mesures expérimentales plus ou moins un écart-type. Ces intervalles correspondent aux intervalles « cibles » qui sont calculés dans l'approche « trajectoire solution » (avec le paramètre d fixé à 1). L'intervalle des concentrations de chaque composé à chaque temps est discrétisé en 5 valeurs de concentrations équidistantes (allant de la valeur minimale à la valeur maximale de l'intervalle). Les intervalles associés au pyruvate ne sont pas discrétisés, et chaque intervalle est remplacé par une unique valeur qui correspond à la moyenne des concentrations de pyruvate mesurées au temps t_i. Au final, on dispose de 5 valeurs de concentrations pour la biomasse, le glucose, l'acétate et le lactate, et ce pour chaque temps de la culture.

Pour une période p_i étudiée, une combinaison de concentrations correspond à une valeur de chaque composé au temps t_i, et une valeur de chaque composé au temps t_{i+1}. L'analyse d'une période est ainsi réalisée à partir de 390 625 combinaisons différentes (5^8). Pour chaque combinaison, j'ai utilisé la méthode FBA afin de déterminer s'il existait au moins une carte de flux solution permettant de modéliser l'évolution des concentrations.

La figure 9.6 donne un exemple des résultats obtenus pour la période $1,5\text{h} \rightarrow 3\text{h}$. Le graphique du glucose montre que lorsque la concentration initiale est « élevée » (relativement à l'intervalle des concentrations autorisées pour le temps $t = 3\text{h}$), il existe toujours une solution. De plus, cette concentration « élevée » au début de la période permet d'atteindre n'importe quelle concentration de glucose à la fin de la période (une concentration « élevée » en glucose permet aussi d'atteindre n'importe quelle concentration pour les autres composés, données non représentées). Par contre, lorsque la concentration initiale de glucose est « faible », seules des concentrations « moyennes » à « faibles » peuvent être atteintes au temps final. Comme le glucose est le composé qui a clairement le plus d'impact sur l'existence ou non d'une solution, je vais poursuivre cette analyse en ne considérant que ce composé.

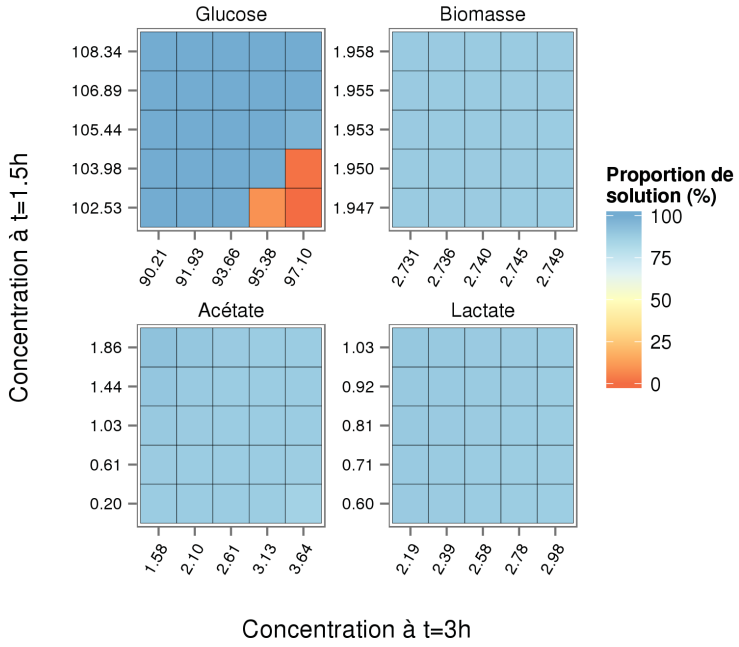


Figure 9.6. – Proportion de combinaisons de concentrations pour lesquelles une solution existe pour la période $1,5\text{h} \rightarrow 3\text{h}$, en fonction des concentrations utilisées aux temps 1,5 et 3h. Chaque graphique représente la dépendance entre l'existence d'une solution et les concentrations initiales et finales d'un composé. L'ordonnée indique les concentrations au temps 1,5h, l'abscisse les concentrations au temps 3h. Les concentrations sont exprimées en g/L pour la biomasse, et en mmol/L pour les autres composés. La proportion de solutions est indiquée par le gradient de couleur.

La figure 9.7 montre, pour toutes les périodes, la dépendance entre l'existence d'une solution, la concentration initiale ($t = i$), et la concentration finale de glucose ($t = i + 1$). Cette figure apporte une grande quantité d'informations. Pour la suite, je vais utiliser les termes de « concentration élevée », « moyenne » et « faible » pour décrire le positionnement des concentrations du glucose dans l'intervalle des concentrations autorisées à chaque temps. Selon le graphique, autrement dit selon la période considérée, on peut identifier plusieurs situations :

— De manière générale, la présence de « carrés » de couleur rouge indique qu'il existe une dépendance entre la concentration de glucose initiale et la concentration de glucose atteignable à la fin de la période. Cette dépendance existe pour 8 des 10 périodes étudiées.

— Une « ligne » entièrement constituée de carrés de couleur rouge indique que certaines concentrations initiales ne permettent pas d'obtenir de solution, peu importe la concentration finale ciblée. Ceci est le cas de la période 3h $\rightarrow$ 6h, où aucune solution n'existe lorsque la concentration en glucose à $t = 3$h est « faible ».

— Une « colonne » entièrement constituée de carrés de couleur rouge indique que les concentrations initiales ne peuvent conduire qu'à certaines concentrations finales, peu importe la concentration initiale considérée. C'est notamment le cas de la période 0h $\rightarrow$ 1,5h, où seules des concentrations finales « faibles » sont atteignables.

— Un « triangle » formé de carrés de couleur rouge indique que la diminution de la concentration en glucose initiale implique une diminution de la plage de concentrations finales atteignables.

— Enfin, les carrés qui présentent un niveau de solution intermédiaire (couleurs bleu clair à orangé) indiquent que, compte tenu de la combinaison des concentrations initiale et finale de glucose considérées, l'existence d'une solution dépend des concentrations des autres composés.

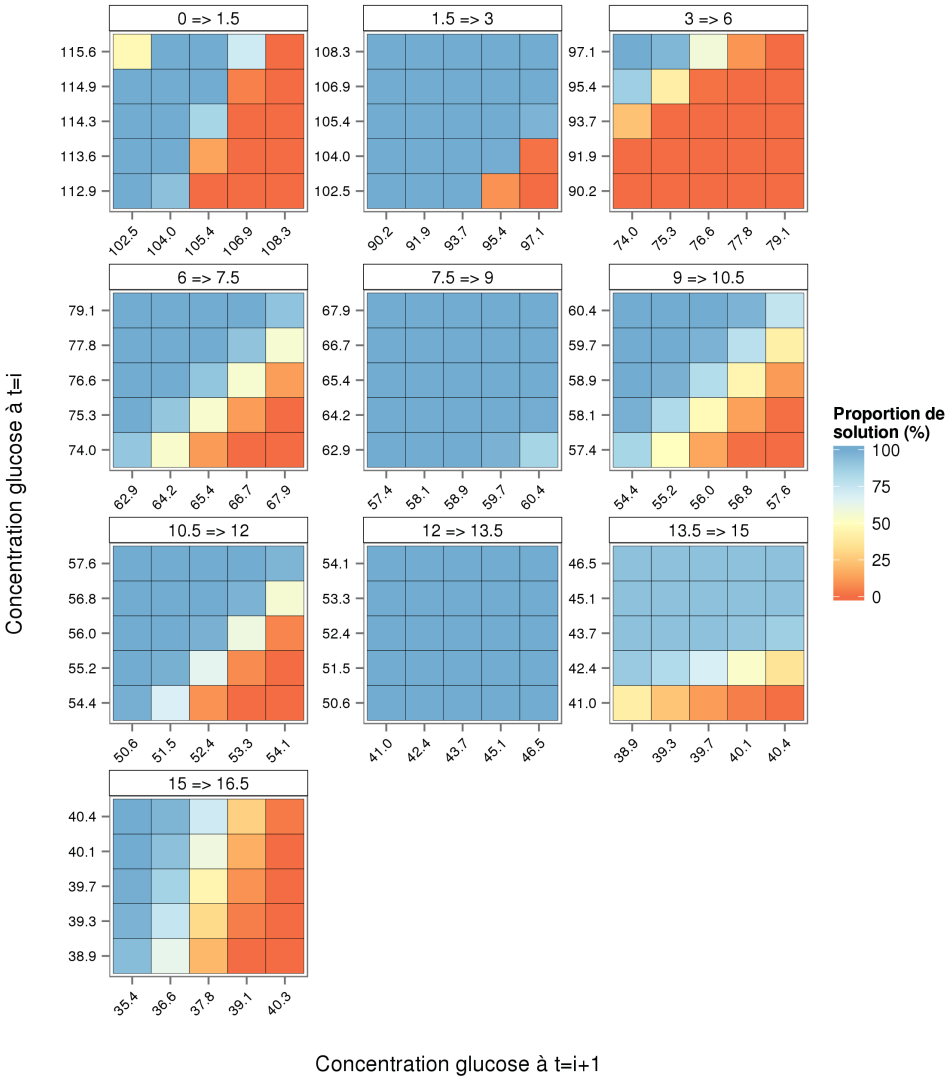


Figure 9.7. – Proportion de combinaisons de concentrations pour lesquelles une solution existe pour les différentes périodes de culture, en fonction de la concentration de glucose aux temps initial et final. Chaque graphique correspond à une période de culture. L'ordonnée indique les concentrations initiales de la période ($t = i$), l'abscisse les concentrations finales ($t = i + 1$). Les concentrations sont exprimées en mmol/L. La proportion de solutions est indiquée par le gradient de couleur.

Comparaison avec les concentrations prédites par les trajectoires solutions

Cette caractérisation « des concentrations solutions » peut maintenant être utilisée pour analyser les distributions de concentrations prédites par l'approche à base de trajectoires solutions. Afin de faciliter la lecture, la figure 9.8 redonne les distributions de concentrations de glucose prédites à l'issue de 1 000 trajectoires solutions.

Période 0h $\rightarrow$ 1, 5h. Dans l'approche à base de trajectoires solutions, les concentrations en glucose à $t = 0$ sont choisies par tirage aléatoire au sein d'une loi normale (centrée sur la moyenne des mesures expérimentales) et tronquée à un écart-type. D'après la caractérisation des concentrations solutions, les concentrations atteignables à l'issue de la période 0h $\rightarrow 1,5$h sont en majorité des concentrations « faibles », en particulier si l'on prend en considération que les concentrations à $t = 0$ sont situées autour de la concentration « moyenne ».

Périodes 1, 5h $\rightarrow$ 3h, 3h $\rightarrow$ 6h et 6h $\rightarrow$ 7, 5h. Étant donné que les concentrations prédites à l'issue de la période 0h $\rightarrow 1,5$h sont majoritairement « faibles », des concentrations « faibles » à « moyennement élevées » sont théoriquement atteignables à $t = 3$h. Cependant, seules les concentrations « moyennes » à « élevées » à $t = 3h$ sont solutions pour la période 3h $\rightarrow$ 6h. Comme l'approche par trajectoire solution instaure un lien de faisabilité entre les concentrations de chaque temps, seules les concentrations « moyennes » prédites à l'issue de la période $1,5$h $\rightarrow$ 3h sont solutions de la période 3h $\rightarrow$ 6h. Ces concentrations solutions à $t = 3$h conduisent à des concentrations « faibles » à $t = 6$h, qui elles-mêmes conduisent à des concentrations « faibles » à $t = 7,5$h.

Période 7, 5h $\rightarrow$ 9h. La période $7,5$h $\rightarrow$ 9h permet d'obtenir une solution pour de nombreuses combinaisons de concentrations solutions, y compris lorsque les concentrations à $t = 7,5$h sont « faibles ». En conséquence, les valeurs à $t = 9$h prédites par l'approche sont plus étalées que dans les périodes précédentes. Notons toutefois qu'aucune valeur de glucose « élevée », pourtant théoriquement atteignable, n'est prédite.

Périodes suivantes. Les concentrations prédites à l'issue des trajectoires solutions sont globalement « faibles » à $t = 10,5$ et $t = 12$h. On observe ensuite une plus grande variabilité des valeurs (relativement aux intervalles concentrations) à $t = 13,5$ et $t = 15$h, suivi d'un retour à des concentrations de glucose « faibles » à $t = 16,5$h.

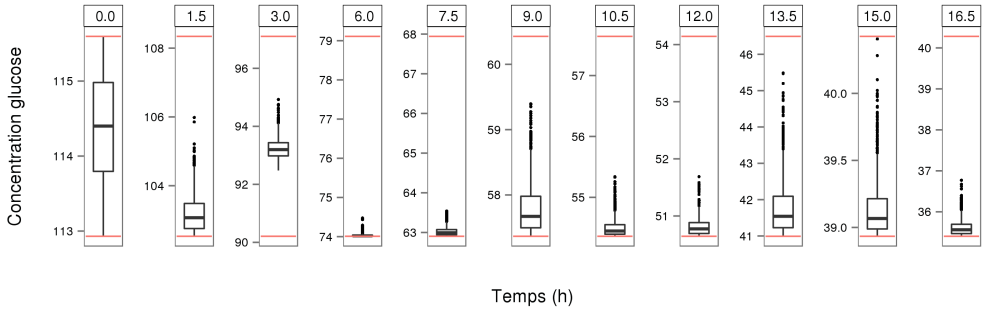


Figure 9.8. – Distribution des concentrations en glucose prédites lors de la génération de trajectoires solutions allant de $t = 0$ à $t = 16,5$h, sans prise en compte des mesures au temps $t = 4,5h$. La concentration, en ordonnée, est exprimée en mmol/L. Chaque graphique correspond à un temps différent, et présente une échelle des concentrations adaptées. Les boîtes à moustaches représentent la distribution des valeurs prédites par les trajectoires solutions. Les lignes horizontales de couleur rouge représentent les bornes minimales et maximales des intervalles de concentrations cibles.

Absence de solution pour certaines combinaisons de concentrations : bilan

Dans cette première analyse, il s'agit de comprendre en quoi les mesures expérimentales et leur variabilité ont un impact sur les concentrations prédites à l'issue des trajectoires solutions.

Compte tenu (i) du modèle, c'est-à-dire de la description du réseau métabolique utilisée, et (ii) des intervalles de confiance estimés à partir des mesures expérimentales, il s'avère que certaines portions des intervalles de concentrations ne permettent pas d'obtenir des cartes de flux solutions. Ainsi, même en considérant les périodes indépendamment les unes des autres, les concentrations moyennes des mesures observées ne peuvent pas être prédites à certains temps. Le fait que l'approche ajoute une contrainte de faisabilité entre les différentes périodes modélisées réduit d'autant plus l'amplitude possible des concentrations solutions.

À la lumière de cette caractérisation des concentrations « solutions », le positionnement « excentré » des concentrations prédites à partir des trajectoires solutions est globalement cohérent.

Toutefois, les concentrations de glucose prédites, notamment aux temps 9, 13,5 et 15h, sont moins variables que ce qui est théoriquement possible. Ce phénomène est à relier avec l'approche par échantillonnage utilisée : pour une trajectoire donnée, l'espace des solutions de chaque période est échantillonné de manière homogène, et l'une des solutions échantillonnées est choisie au hasard pour prolonger la trajectoire. Or, si le « volume » de l'espace des solutions correspond majoritairement à des solutions qui conduisent à de fortes consommations de glucose, alors la solution retenue aura plus de chance d'être une carte « fortement consommatrice » de glucose. Si cette carte permet néanmoins de prolonger la trajectoire jusqu'au temps final de la dynamique, la carte peut être considérée comme valable compte tenu des contraintes imposées sur le système, et la trajectoire solution est retenue par l'algorithme de l'approche.

L'obtention de prédictions de concentrations plus variables repose donc sur la modification de la forme de l'espace des solutions. Dans la sous-section suivante, je vais montrer que la variabilité des concentrations prédites peut être sensiblement augmentée en limitant le phénomène de « gaspillage énergétique », cela en ajoutant quelques contraintes raisonnables supplémentaires sur le système.

9.2.2. Gaspillage énergétique

Le terme de *gaspillage énergétique* fait ici référence à des répartitions de flux où une part relativement importante de l'énergie apportée par les substrats est dissipée par quelques flux du système, sans être utilisée pour produire de la biomasse ou des métabolites secondaires. Ce phénomène peut avoir un impact sur les concentrations prédites par les trajectoires solutions : plus le gaspillage sera important, moins il restera d'énergie disponible pour la synthèse des composés (biomasse et métabolites extracellulaires).

Dans cette partie, je vais discuter des réactions qui sont en grande partie responsables du gaspillage énergétique dans le système : d'une part la « pseudo- » réaction ATPm qui modélise la consommation d'ATP nécessaire pour maintenir l'intégrité des cellules, et d'autre part les couples de réactions PFKA–FBP et PPC–PPCk, qui forment des cycles futiles dans le système.

ATP de maintenance

Lors de la croissance d'une bactérie, les substrats consommés apportent l'énergie nécessaire à la vie de l'organisme. Cette énergie peut être divisée en deux parts, selon l'usage qui en est fait par la cellule. Une première partie de l'énergie est utilisée afin de croitre, notamment pour synthétiser les molécules nécessaires à la production de nouvelles cellules. La deuxième partie, qui n'est pas associée à la croissance, représente les dépenses énergétiques nécessaires pour maintenir l'intégrité des cellules déjà existantes, ainsi que d'autres fonctions comme la mobilité.

Ces deux usages sont pris en compte dans le modèle du métabolisme central de *C. glutamicum*. Les besoins énergétiques associés à la croissance sont décrits dans la pseudo-réaction de production de biomasse (« BIOM »). Les besoins énergétiques associés à la maintenance cellulaire sont quant à eux modélisés par la pseudo-réaction « ATPm », qui correspond à la transformation de l'ATP en ADP :

$$\text{ATPm} : \text{ATP} \Rightarrow \text{ADP} + \text{Pi}$$

Compte tenu des données disponibles dans la littérature, le flux ATPm a été contraint de manière *constante* (contrainte identique à toutes les périodes de temps) à une valeur minimale de 5 mmol/g biom./h (cf. p. 156).

Par ailleurs, aucune contrainte n'a été spécifiquement appliquée sur la valeur maximale. La valeur maximale possible est alors de 100 mmol/g biom./h, contrainte appliquée par défaut sur tous les flux du système afin de borner l'espace des solutions. Cette valeur maximale correspond à un flux d'ATP de maintenance largement supérieur à ce qui peut être rencontré dans la littérature.

L'idée était ici de ne pas poser d'hypothèse sur la quantité maximale d'énergie nécessaire à la maintenance des cellules, cultivées en condition de limitation en biotine. Cependant, le fait (*i*) de ne pas poser de contrainte « réaliste » sur la vitesse maximale du flux ATPm et (*ii*) d'utiliser une approche par échantillonnage de l'espace des solutions conduit potentiellement à échantillonner et choisir des cartes de flux qui présentent une forte dissipation d'énergie *via* la réaction ATPm. Ce phénomène est illustré par la figure 9.9, qui montre la distribution des valeurs pour le flux ATPm, estimée à partir de 1 000 trajectoires solutions. On peut y voir que certaines cartes échantillonnées présentent un flux ATPm élevé, en particulier pour la première période modélisée.

A posteriori, il semble donc plus pertinent de fixer une limite réaliste sur la vitesse maximale du flux ATPm. Compte tenu des valeurs rencontrées dans la littérature, une valeur maximale de 15 mmol/g biom./h pourrait par exemple être fixée ([Stouthamer 1975, Russell 1979, Mulder 1986, Vallino 1993, Varma 1994, Borodina 2005, Coze 2013]). Ce faisant, des cartes de flux présentant des valeurs supérieures à la valeur minimale de 5 mmol/g biom./h pourront toujours être échantillonnées, mais des cartes présentant une vitesse supérieure à 15 mmol/g biom./h ne seront pas incluses dans l'espace des solutions, et ne seront donc pas utilisées pour générer des trajectoires solutions et prédire les concentrations en métabolites extracellulaires.

Cycles futiles

Dans le modèle à base de contraintes qui décrit le métabolisme central de *C. glutamicum*, les couples de réactions PFKA–FBP et PPC–PPCk s'apparentent à des cycles futiles. La figure 9.10 présente l'environnement métabolique de ces deux cycles. Chaque couple de réactions forme une boucle, où le substrat de l'une des réactions est le produit de l'autre (et *vice-versa*), et chaque tour de boucle dissipe une molécule énergétique (ATP ou GTP).

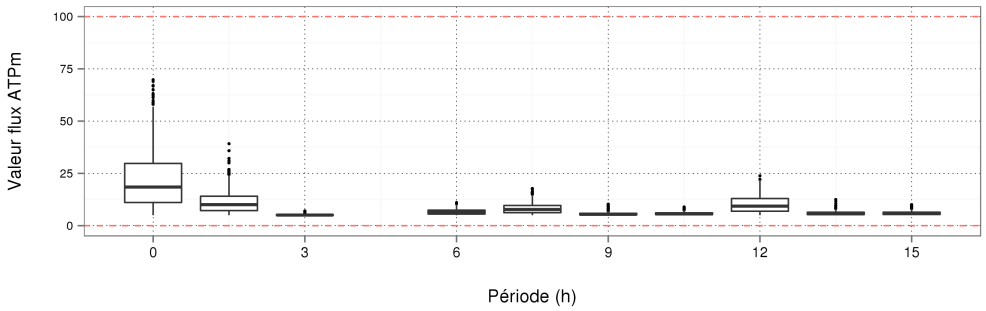


Figure 9.9. – Distribution des valeurs du flux ATPm lorsqu'aucune contrainte spécifique n'est appliquée. La valeur du flux spécifique, exprimée en mmol/g biom./h, est indiquée en ordonnée. Chaque boîte à moustache correspond à une période modélisée, dont le temps de début est donné en abscisse. Les droites horizontales en pointillé de couleur rouge indiquent les contraintes d'inégalité « standards » appliquées sur le flux.

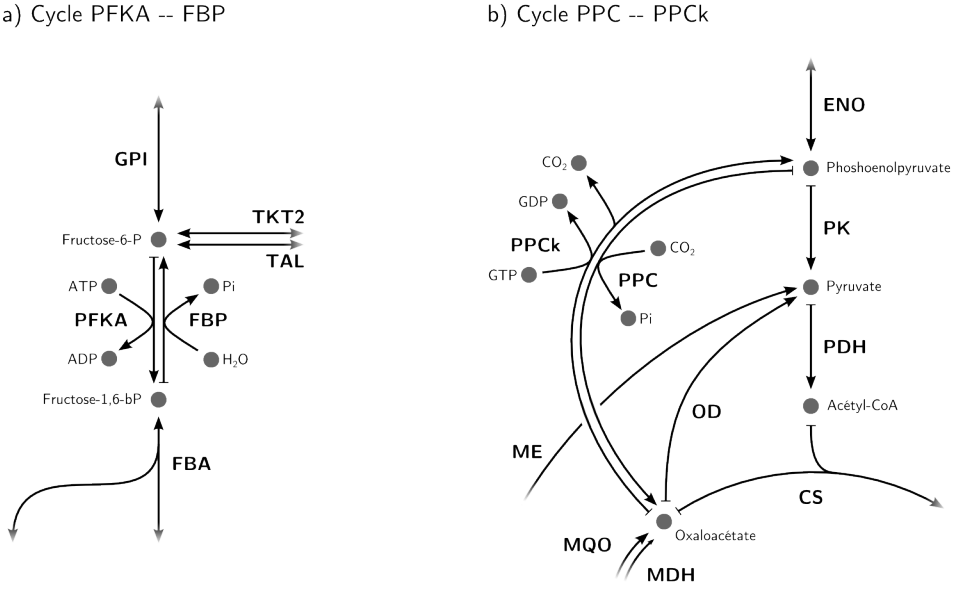


Figure 9.10. – Représentation des cycles futiles formés par les couples de réactions **a)** PFKA–FBP et **b)** PPC–PPCk. Chaque flèche représente une réaction, chaque point représente un métabolite. Une réaction réversible est indiquée par une double flèche. Le nom abrégé de chaque réaction est indiqué en gras, le nom complet de chaque réaction est disponible en annexe (annexe B, page 253).

Comme ces réactions forment des cycles, leurs vitesses ne sont pas limitées par les flux environnants. Par exemple, dans le couple de réactions PFKA–FBP, le couple peut se suffire à lui-même pour maintenir la balance des masses au niveau des métabolites fructose-6-phosphate et frutose-1,6-bisphosphate. En conséquence, les contraintes *variables* appliquées sur les flux E/S n'ont que peu d'effet sur les vitesses des réactions au sein de ces cycles futiles. Par exemple, dans une situation où les contraintes *variables* conduisent à une solution avec un flux GPI de $v_{\text{GPI}} = 6$ mmol/g biom./h, toutes les vitesses du cycle PFKA–PFB qui conduisent à un flux net traversant le cycle égal à 6 ($v_{\text{FBP}} - v_{\text{PFKA}} = 6$) sont solutions, et ce peu importe les valeurs brutes des flux dans le cycle (flux « cyclique »). Dans le modèle utilisé, seules les contraintes « par défaut » sont appliquées sur les flux qui forment ces cycles futiles : chaque flux peut donc varier entre 0 et 100 mmol/g biom./h. Ces cycles peuvent alors représenter une dissipation importante d'énergie, sous la forme d'ATP pour le cycle PFKA–FBP, et sous la forme de GTP pour le cycle PPC–PPCk. Ce phénomène de cycle futile est illustré par la figure 9.11, qui montre la distribution des valeurs du flux net (flux traversant le cycle) et du flux cyclique (flux brut dans le cycle futile).

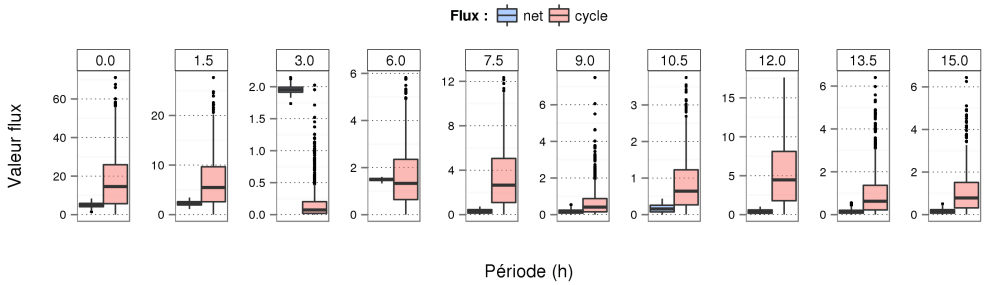


Figure 9.11. – Distributions des valeurs de flux nets et du flux cyclique dans le cycle futile formé par les réactions PFKA–FBP, à chaque période. La valeur des flux, exprimée en mmol/g biom./h, est indiquée en ordonnée. Chaque graphique correspond à une période modélisée. Le temps de début des périodes est indiqué au-dessus des graphiques. Pour chaque graphique, deux boîtes à moustaches sont représentées : la distribution du flux net, en bleu, et la distribution du flux « cyclique », en rouge. Le flux net est calculé en prenant la valeur absolue de la différence entre les flux PFKA et FBP : flux net = abs(PFKA − FBP). Le flux cyclique est calculé en prenant la valeur absolue de la différence entre le flux PFKA et le flux net : flux cyclique = abs(PFKA − flux net).

Une première manière de limiter les pertes énergétiques dues aux cycles futiles serait de simplement les supprimer. Pour cela, chaque cycle pourrait être décrit à l'aide d'une unique réaction réversible. Par exemple, le cycle PFKA–FBP pourrait être décrit en conservant uniquement la réaction PFKA, et en rendant cette réaction réversible afin de conserver la possibilité de réaliser la gluconéogenèse (synthèse du glucose à partir de composés non-glucidiques, phénomène observé lors de croissance sur acétate ou lactate, par exemple, [Wendisch 2000]). Cependant, cette simplification du réseau conduirait à une production d'ATP lorsque la réaction PFKA serait utilisée dans le sens inverse (fructose-1,6-bisphosphate vers fructose-6-phosphate), ce qui n'est pas le cas dans la réalité.

Par ailleurs, ces cycles *a priori* futiles existent potentiellement dans les cellules de *C. glutamicum* : le gène *fbp*, codant l'enzyme fructose 1,6-bisphosphatase qui catalyse la réaction FBP, est exprimé quelque soit la condition de culture [Rittmann 2003], et des cycles futiles entre les métabolites phospho*enol*pyruvate, pyruvate, oxaloacétate et malate ont été observés dans plusieurs études [Petersen 2000, Wendisch 2000].

Dans le modèle, une autre manière de limiter la perte énergétique par les cycles futiles est de conserver ces cycles dans la description du réseau, mais de limiter la vitesse des réactions qui y participent afin de maintenir un certain réalisme biologique. Comme les répartitions et les vitesses des flux changent selon la période, l'utilisation de contraintes *constantes* est peu adaptée. Une approche plus pertinente serait d'appliquer des contraintes *variables* spécifiques de chaque période et de chaque trajectoire. On peut, par exemple, imaginer la procédure suivante pour évaluer les contraintes *variables* à appliquer sur les flux PFKA et FBP pour une période donnée :

1. Compte tenu de la description du réseau, des contraintes *constantes*, et des contraintes *variables* appliquées sur les flux E/S, utiliser la méthode FVA pour évaluer les valeurs minimales et maximales atteignables par le flux GPI (flux réversible qui se situe directement en amont du cycle PFKA–FBP).
2. En déduire la valeur absolue maximale $v_{\text{GPI},abs}$ atteignable pour le flux GPI :

$$v_{\text{GPI},abs} = \text{abs}(v_{\text{GPI},min}, v_{\text{GPI},max})$$

3. Contraindre les flux PFKA et FBP à une valeur maximale égale à x fois la valeur absolue maximale $v_{\text{GPI},abs}$. Par exemple, avec $x = 3$:

$$v_{\text{PFKA},max} = 3 \times v_{\text{GPI},abs}$$
$$v_{\text{FBP},max} = 3 \times v_{\text{GPI},abs}$$

De cette manière, avec une valeur de x appropriée, le phénomène de dissipation d'énergie par le flux cyclique (boucle entre PFKA et FBP) est limité, et cette limitation n'a par ailleurs pas d'impact sur les valeurs possibles du flux net qui traverse le cycle.

Ajout de contraintes raisonnables

Afin de tester en quoi l'ajout de contraintes « raisonnables » sur les flux dissipateurs d'énergie pouvait améliorer les distributions des concentrations prédites, j'ai généré plusieurs jeux de trajectoires solutions :

— Jeu n° 1 : jeu témoin de 1 000 trajectoires solutions, similaire aux autres jeux générés jusqu'alors (pas de contraintes supplémentaires).

— Jeu n° 2 : jeu de 1 000 trajectoires où le flux ATPm est contraint à une valeur maximale de 15 mmol/g biom./h.

— Jeu n° 3 : jeu de 1 000 trajectoires où (i) le flux ATPm est contraint à une valeur maximale de 15 mmol/g biom./h, et (ii) les flux qui participent aux boucles futiles sont limités par des contraintes variables calculées dynamiquement pour chaque période de chaque trajectoire. Les flux PFKA et FBP sont contraints à une valeur maximale égale à 3 fois la valeur maximale absolue du flux GPI, et les flux PPC et PPCk sont contraints à une valeur maximale égale à 3 fois la valeur maximale du flux CS.

La figure 9.12 présente les distributions des concentrations en glucose prédites à l'issue de ces trois jeux de trajectoires. De manière générale, on peut voir que le fait de limiter la dissipation d'énergie permet d'obtenir des concentrations de glucose qui sont sensiblement plus variables. Comme on pouvait s'y attendre, c'est l'utilisation combinée d'une contrainte *constante* sur la valeur maximale du flux ATPm et de contraintes *variables* sur les cycles futiles qui permet d'obtenir l'amélioration la plus importante. J'utilise ici le terme d'*amélioration* puisque nous avons vu précédemment que, même si le positionnement des distributions de concentrations prédites était

cohérent avec les concentrations atteignables compte tenu des mesures et du réseau, les distributions semblaient moins variables que ce qui était théoriquement possible.

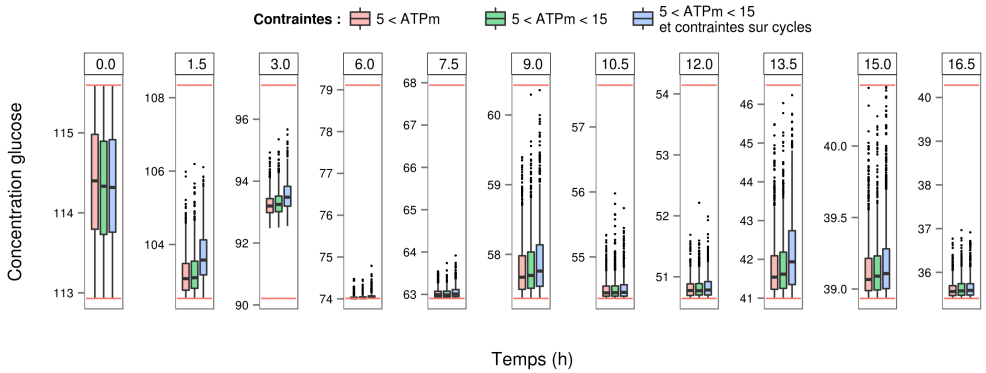


Figure 9.12. – Distribution des concentrations en glucose prédites en fonction des contraintes raisonnables ajoutées sur les flux dissipateurs d'énergie. La concentration, en ordonnée, est exprimée en mmol/L. Chaque graphique correspond à un temps différent, et contient trois distributions de concentrations : concentrations obtenues sans ajout de contrainte (en rouge), obtenues avec contrainte sur le flux ATPm (en vert), obtenues avec contraintes sur le flux ATPm et sur les flux des cycles futiles. Les lignes horizontales de couleur rouge représentent les bornes minimales et maximales des intervalles de concentrations cibles.

La figure 9.13 montre l'impact de l'ajout des contraintes raisonnables sur les distributions de valeurs de 6 flux du système (comparaison entre le jeu n° 1 et le jeu n° 3). Les flux présentés sur ce graphique ont été normalisés selon la formule donnée dans le chapitre précédent (équation (7.6), p. 144), que je redonne ici :

$$\begin{aligned} flux_{k,norm} &= flux_{k,spe} \cdot NC_k \cdot 1/flux_{carb} \\ \text{avec} \quad flux_{carb} &= \sum_{io=1}^{n} \left[max \begin{pmatrix} v_{io} \\ 0 \end{pmatrix} \cdot NC_{io} \right] \end{aligned} \tag{9.1}$$

où,

- $flux_{k,norm}$ et $flux_{k,spe}$ représentent respectivement le flux normalisé et le flux spécifique ;
- NC_k représente le nombre d'atomes de carbone impliqués dans le flux $flux_k$;
- v_{io} représente un flux E/S du système ;
- NC_{io} représente le nombre d'atomes de carbone impliqués dans la réaction correspondant au flux v_{io}.

Les flux normalisés expriment ainsi la quantité de carbone impliqué dans chaque réaction, relativisé par rapport à la quantité totale de carbone consommé par le système pendant la période de temps considérée.

De manière intéressante, les résultats montrent que l'augmentation de la variabilité des concentrations en composés extracellulaires prédits s'accompagne d'une réduction de la variabilité des flux du système. Cela peut s'expliquer par le fait que l'ajout de contraintes réduit la taille des espaces des solutions, ce qui se traduit par une étendue moins importante des valeurs de flux possibles. Il est aussi intéressant de noter que les médianes des flux sont modifiées. Ceci indique que la réduction de l'espace des solutions est réalisée de façon « asymétrique » vis-à-vis des étendues des valeurs possibles. Étant donné que les contraintes ajoutées conduisent à une description plus réaliste du système, on peut supposer que les distributions de flux obtenues à partir de ces contraintes sont elles-mêmes plus proches de la réalité.

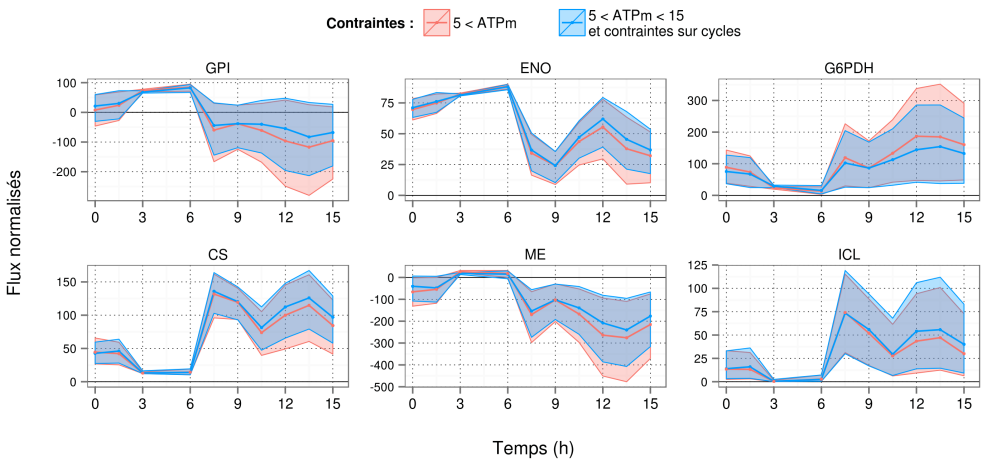


Figure 9.13. – Dynamique et variabilité des flux normalisés lors de l'utilisation des contraintes supplémentaires sur les flux dissipateurs d'énergie. Les flux ont été normalisés selon l'équation (9.1), et sont exprimés en (mmol de carbone)/(100 mmol de carbone carbone consommé). La médiane d'un flux est représentée par une ligne colorée continue, l'intervalle de confiance à 75% est représenté par une aire colorée. Les valeurs obtenues lorsqu'aucune contrainte n'est appliquée sur les flux dissipateurs (jeu n° 1) sont représentées en rouge, celles obtenues lorsque des contraintes supplémentaires sont appliquées (jeu n° 3) sont représentées en bleu.

Gaspillage énergétique : bilan

J'ai ici discuté des flux responsables d'un phénomène de gaspillage énergétique. Afin de limiter l'échantillonnage de solutions qui présentent une dissipation importante de l'énergie, j'ai proposé d'ajouter quelques contraintes raisonnables (compte tenu de la réalité biologique) : d'une part, appliquer une contrainte *constante* sur la valeur maximale du flux ATPm, et d'autre part d'ajouter des contraintes *variables* sur les flux impliqués dans les cycles futiles PFKA–FBP et PPC–PPCk. L'ajout de ces contraintes permet de prédire des concentrations en métabolites extracellulaires plus variables, tout en prédisant des flux internes moins variables.

Notons ici que le fait d'ajouter ces contraintes raisonnables permet par ailleurs de générer des trajectoires solutions plus rapidement : le temps moyen de génération d'une trajectoire passe ainsi d'environ 7 minutes et 45 secondes à environ 2 minutes (à type de processeur et charge de calcul similaire). Cette amélioration est à attribuer au fait que l'algorithme trouve plus facilement des solutions pour la période 3h → 6h.

9.2.3. Analyse des distributions de concentrations prédites : bilan

Le propos de cette section est d'identifier *pourquoi* les distributions des concentrations prédites à l'issue des trajectoires solutions était excentrées par rapport aux intervalles de confiance estimés à partir des mesures expérimentales.

Toutes les concentrations ne sont pas atteignables. La caractérisation des concentrations « solutions » de chaque période a permis d'apporter une première réponse : compte tenu des données et du modèle, toutes les concentrations incluses dans les intervalles de « confiance » ne sont pas atteignables. En conséquence, le positionnement excentré des concentrations prédites s'explique par le fait que les concentrations solutions ne sont pas distribuées de manière homogène sur les intervalles de concentrations cibles utilisés. Le fait que l'approche ajoute une contrainte de faisabilité entre les différentes périodes modélisées réduit d'autant plus l'amplitude possible des concentrations solutions.

Des incohérences entre le modèle et les mesures expérimentales ont déjà été identifiées lors de l'étude du temps 4,5h. Il s'avère en fait que des incohérences existent aussi pour les autres périodes de temps modélisées. Des causes possibles ont déjà été abordées : des mesures éventuellement erronées au niveau des concentrations du glucose à $t = 1,5$ et $t = 3\text{h}$; au moins un autre métabolite est potentiellement produit puis reconsommé par les cellules ; et les besoins en précurseurs et en énergie pour former de la biomasse sont probablement différents de ceux utilisés dans le modèle. Le fait que les concentrations expérimentales conduisent à des combinaisons de concentrations sans solution dans pratiquement toutes les périodes de temps semble indiquer que les mesures de glucose ne sont probablement pas en cause.

Notons ici que la caractérisation des combinaisons de concentrations solutions peut être un moyen efficace, en étude préliminaire, pour tester la cohérence entre un modèle MBC et des mesures expérimentales. En effet, cette approche ne nécessite pas de s'intéresser aux flux internes du système, ni de poser une quelconque hypothèse biologique quant à l'optimalité du réseau métabolique étudié. Une limite de cette approche est l'explosion du nombre de combinaisons à tester lorsque le nombre de composés produits et consommés augmente. Une manière d'éviter cette explosion combinatoire

est de choisir uniquement les composés qui jouent un rôle important dans le système. Ainsi, dans ma situation, je n'ai pas pris en compte le pyruvate, puisque ce composé ne représente qu'un faible flux de carbone. Une autre manière de limiter l'explosion combinatoire est d'utiliser un nombre variable de niveaux de concentration selon le composé considéré, en fonction de l'étendue des valeurs de concentrations associées au composé. Compte tenu du faible impact des composés autre que le glucose, j'aurai pu utiliser pour ces composés 3 niveaux de concentration au lieu de 5.

Certaines cartes prédites sont peu réalistes d'un point de vue énergétique. La comparaison des concentrations prédites avec les concentrations « solutions » atteignables m'a ensuite permis d'identifier que la variabilité des valeurs prédites était plus faible que ce qui était théoriquement possible d'obtenir.

Dans un deuxième temps, je me suis alors intéressé à une cause de cette faible variabilité : la prédiction de cartes de flux qui présentent une utilisation clairement non optimale de l'énergie fournie par les substrats. Le fait que (*i*) les flux dissipateurs d'énergie ne soient pas contraints de manière réaliste, et (*ii*) que l'approche repose sur l'échantillonnage d'espaces des solutions, peut conduire à sélectionner des cartes très « dépensières » en énergie si celles-ci sont solutions des périodes considérées et si elles permettent par ailleurs de poursuivre la construction des trajectoires solutions jusqu'au temps final.

Afin d'améliorer le réalisme des trajectoires solutions prédites, j'ai proposé d'appliquer des contraintes supplémentaires sur le système. Ce faisant, les concentrations prédites à l'issue de trajectoires présentent une variabilité plus cohérente avec les concentrations *a priori* atteignables. L'ajout de contraintes sur les flux dissipateurs d'énergie conduit par ailleurs à la prédiction de flux qui sont sensiblement moins variables.

Remarquons ici que la réaction ATPm utilisée dans le modèle n'est pas réellement présente dans les cellules : cette réaction est en fait une pseudo-réaction qui permet de considérer l'aspect énergétique de la maintenance cellulaire, sans devoir décrire l'ensemble des réactions qui sont réellement responsables des dépenses d'énergie. Si le système étudié était un modèle « cellule complète », c'est-à-dire un modèle décrivant l'ensemble des réactions biochimiques de l'organisme, la pseudo-réaction de dissipation d'ATP ne devrait théoriquement plus être nécessaire.

Remarquons aussi que, *in vivo*, l'activité de la fructose 1,6-bisphosphatase, qui catalyse la réaction FBP, est inhibée par la présence d'AMP ou de phospho*enol*pyruvate, ce qui limite alors l'importance du cycle futile PFKA–FBP [Rittmann 2003]. Comme les modèles à base de contraintes ne permettent pas d'utiliser les concentrations en métabolites internes, ce type d'information ne peut être utilisé dans le modèle. Le problème des cycles *a priori* futiles illustre ainsi l'une des limites des modèles à base de contraintes.

Les concentrations prédites peuvent être utilisées pour examiner l'effet de contraintes supplémentaires. Au sein du réseau étudié, il y a d'autres ensembles de réactions qui forment des cycles. C'est le cas des couples de réactions LDH–LLD et MDH–MQO, dont la description est donnée dans la table ci-dessous :

Réaction	Description
LDH	pyruvate + NADH ⇒ lactate + NAD
LLD	lactate + MQ ⇒ pyruvate + MQH_2
MDH	oxaloacétate + NADH ⇔ malate + NAD
MQO	malate + MQ ⇒ oxaloacétate + MQH_2

Pour ces deux cycles, la réalisation d'un tour de boucle équivaut à la réaction catalysée par la NADH déshydrogénase (réaction NDH : NADH + MQ ⇒ NAD + MQH_2), dans laquelle le potentiel énergétique d'une molécule de NADH est transféré pour former une molécule de ménaquinol (MQH_2), qui peut ensuite être utilisée pour la respiration. Bien que ces cycles ne soient pas *futiles* — dans le sens où ils n'entrainent pas de dissipation énergétique — il pourrait être intéressant d'examiner l'impact d'une limitation de la vitesse de ces réactions sur les répartitions de flux et les concentrations prédites.

Dans le métabolisme étudié, le cycle LDH–LLD se situe juste à proximité du flux d'échange de lactate avec le milieu extérieur. La réaction PQO, qui est l'une des deux réactions capables de produire de l'acétate, participe à la production de ménaquinol (PQO : pyruvate + MQ ⇒ acétate + MQH_2 + CO_2). Des contraintes « réalistes » sur le flux NDH et les flux cycliques dans les boucles LDH–LLD et MDH–MQO, pourraient réduire le potentiel de production de MQH_2. Une « solution » possible pour maintenir le pool de MQH_2

pourrait alors être de favoriser la réaction PQO, et donc la production d'acétate. On pourrait par exemple s'attendre à ce que l'utilisation de contraintes supplémentaires sur ces flux (LLD, LDH, MDH, MQO, NDH) conduise à prédire des distributions de concentrations en acétate plus élevées que celles actuellement prédites dans les premières périodes de la dynamique. Par ailleurs, la ré-oxydation du NADH en NAD est nécessaire afin de permettre le fonctionnement de la glycolyse (voie métabolique productrice de NADH). Le maintien du pool de NAD oxydé (imposé par la contrainte d'état stationnaire du système) pourrait donc « forcer » le système à produire du lactate intracellulaire, qui, puisqu'il ne peut pas être totalement reconsommé du fait de la limitation du flux cyclique, devra être excrété. Bien que ces résultats ne permettraient pas d'expliquer directement l'effet de la biotine, ils pourraient donner des éléments de réponse sur les phénomènes d'excrétion d'acides organiques observés expérimentalement.

L'utilisation des concentrations expérimentales à la fois pour (*i*) générer les trajectoires et (*ii*) évaluer les trajectoires générées peut être discutable. Toutefois, comme les mesures expérimentales sont utilisées comme *intervalles* cibles, et non comme *valeurs exactes* cibles, je pense que l'analyse du positionnement et de la variabilité des concentrations prédites vis-à-vis des intervalles initialement définis est justifiable. Comme je viens de l'illustrer, cela permet notamment d'identifier des parties du modèle qui nécessiteraient des améliorations, et d'évaluer les impacts de ces ajustements.

On peut aussi considérer les intervalles cibles comme des informations supplémentaires qui représentent la part inconnue — ou tout du moins non explicable — du système modélisé. Cette part inconnue peut par exemple correspondre aux divers phénomènes de régulation qui ont lieu pendant la culture, ou encore à des portions du métabolisme qui n'ont pas été modélisées. Les intervalles cibles permettent ainsi de « borner » l'espace des comportements possibles du système. *Idéalement*, au fil des ajustements raisonnables du modèle, les prédictions générées ne devraient plus avoir besoin de « s'appuyer » sur ces intervalles cibles pour prédire des concentrations cohérentes avec les mesures expérimentales : la part d'information connue et décrite dans le modèle est alors suffisante pour proposer une explication complète du phénomène réel étudié.

Chapitre 10

Conclusion, perspectives

10.1. Considérations sur l'approche actuelle

L'approche par trajectoire solution proposée dans cette thèse utilise une contrainte de faisabilité entre les périodes de la dynamique afin de ne sélectionner que certaines solutions de chaque période. Pour être retenue, une solution à une période p_i doit s'inscrire dans une succession de solutions — une trajectoire — à l'issue de laquelle l'évolution des concentrations en composés est cohérente avec les mesures expérimentales. Une évolution cohérente est matérialisée par le fait que les concentrations prédites doivent se situer dans des intervalles de concentrations cibles, définis à partir des mesures expérimentales.

D'un point de vue biologique, le fait de réaliser plusieurs mesures biologiques n'est pas anodin : en faisant plusieurs mesures on souhaite d'une part estimer une valeur moyenne, c'est-à-dire estimer la valeur que l'on aurait *a priori* le plus de chance de rencontrer si l'on réalisait d'autres mesures (*quelle est la valeur la plus représentative ?*), et d'autre part on souhaite évaluer le degré de confiance à accorder à cette valeur (*à quel point cette valeur moyenne est-elle précise ?*). On peut aussi utiliser ces notions de moyenne et de variabilité afin d'évaluer la validité de chaque mesure (*y a-t-il une valeur clairement éloignée des autres ?*).

Dans l'approche utilisée, aucune sélection particulière n'est appliquée lors du choix d'une solution, en dehors du fait que la solution choisie doit être solution de la période traitée et permettre de trouver des solutions pour toutes les périodes suivantes. En particulier, aucune considération n'est donnée au fait que les concentrations finalement prédites soient proches ou éloignées de la moyenne des mesures : l'espace des solutions contraint garantit que chaque solution échantillonnée conduit à des concentrations

dans l'intervalle cible à la fin de la période, mais il n'y a pas de « pression de sélection » pour préférer certaines solutions à d'autres.

Le fait qu'il n'y ait pas de sélection vis-à-vis du « positionnement » des concentrations qui seront prédites apporte certains avantages. En particulier, les concentrations finalement prédites peuvent être comparées aux intervalles des concentrations mesurées : cela permet de déterminer si des ajustements du modèle sont potentiellement nécessaires afin de prédire des concentrations plus proches du centre des intervalles, puis d'évaluer l'impact de ces ajustements sur les nouvelles distributions de concentrations prédites.

Je pense avoir illustré cet intérêt dans la section 9.2 (p. 201). Dans un premier temps, il s'avère que toute l'étendue des intervalles de concentrations mesurées n'est pas atteignable, ce qui révèle que certains comportements métaboliques *a priori* réels ne sont pas prédictibles compte tenu des connaissances utilisées. Dans un deuxième temps, on observe que la majorité des comportements rapportés par les trajectoires solutions conduisent à des distributions moins variables que ce qui est théoriquement possible, et que ces distributions de concentrations sont par ailleurs positionnées dans la partie basse des intervalles de concentrations cibles. Étant donné l'approche utilisée, on peut alors supposer que ces distributions de concentrations prédites sont dues au fait que certains comportements autorisés — autrement dit, certaines solutions échantillonnées — n'ont en fait pas lieu dans la dynamique réelle du réseau étudié. Dans un troisième temps, afin d'améliorer le réalisme du modèle, on peut alors proposer et tester différents ajustements dans la description du système. L'ajout de quelques contraintes raisonnables sur les réactions dissipatrices d'énergie a ainsi permis d'améliorer la représentativité des distributions de concentrations prédites, liée à des prédictions de flux internes sensiblement différentes. En définitive, le fait de ne pas sélectionner les solutions sur la base du positionnement des concentrations dans les intervalles autorise un processus itératif d'amélioration, de raffinement du modèle.

Cependant, l'idéal d'un modèle à partir duquel les trajectoires solutions conduiraient à des distributions de concentrations qui seraient centrées sur les moyennes expérimentales et qui présenteraient des dispersions similaires aux variabilités des mesures peut être une quête longue. La similarité de la variabilité comme but à atteindre nécessite par ailleurs de supposer que la variabilité mesurée soit uniquement due à la variabilité biologique (variabilité dans le fonctionnement du système biologique modélisé), et que

la variabilité liée aux processus de mesure soit nulle (variabilité non explicable par le modèle). Cette quête serait particulièrement longue si l'on prend en compte que les concentrations en métabolites ne peuvent être directement considérées dans les modèles à base de contraintes, et que donc les phénomènes de régulation ne peuvent être introduits directement dans la mécanique de la méthode.

Certaines approches, comme la méthode DFBA, cherchent à *prédire* ou *simuler* le comportement d'un système, c'est-à-dire à décrire à la fois (i) le fonctionnement du système et (ii) les conséquences de ce fonctionnement à un instant t sur le fonctionnement aux instants suivants. D'autres approches, comme la méthode MFA, se restreignent uniquement à analyser le comportement du système, en déduisant son fonctionnement à partir d'un ensemble de faits observés et supposés.

L'approche par trajectoire solution que je propose se situe quelque part entre ces deux extrémités. Les intervalles de concentrations cibles représentent des faits observés qui contraignent les comportements possibles du système et contribuent à l'analyse du système. Dans ce sens, l'approche s'apparente à une méthode d'analyse. Les intervalles autorisent aussi une certaine liberté dans les dynamiques possibles, et les dynamiques finalement obtenues dépendent des fonctionnements modélisés à chaque période et des conséquences de ces fonctionnements sur les périodes suivantes. Dans ce sens l'approche peut alors s'apparenter à une méthode de simulation ou méthode « prédictive ».

Finalement, si Lewis et coll. proposaient aujourd'hui leur classification des méthodes à base de contraintes [Lewis 2012], l'approche par trajectoire solution développée dans ce travail de thèse pourrai représenter une nouvelle méthode parmi celles situées dans le groupe « *Monte Carlo sampling* » de la branche « non biaisée » de l'arbre. Ce positionnement est illustré par la figure 10.1 (page suivante) .

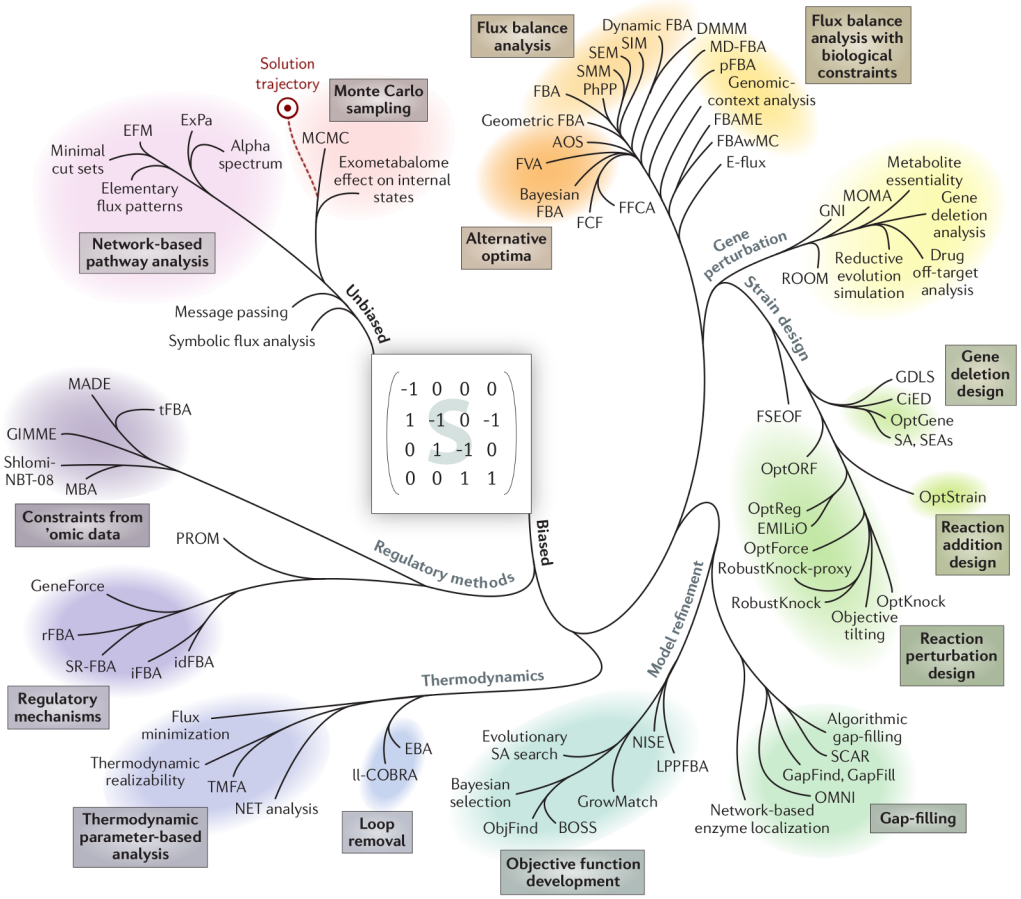


Figure 10.1. – Représentation « phylogénique » des approches MBC. Figure issue de [Lewis 2012], modifiée en faisant apparaître la méthode par trajectoire solution. Le positionnement de l'approche par trajectoire solution vis-à-vis de cette classification est indiqué par un point rouge (en haut, dans la partie gauche de la figure).

10.2. Éléments de réflexion pour la conception de nouvelles approches

Dans cette dernière section, je propose quelques éléments de réflexion pour la conception d'autres approches, qui aborderaient différemment la problématique de la prise en compte de la variabilité expérimentale lorsque la dynamique d'un réseau métabolique est modélisée avec un modèle à base de contraintes.

10.2.1. Approche analytique de la dynamique

L'algorithme proposé dans l'approche par trajectoire solution est un algorithme itératif : les périodes sont traitées les unes après les autres, et les prédictions de concentrations obtenues à l'issue d'une période p_i contraignent les solutions possibles pour la période suivante p_{i+1}. Par construction, cet algorithme ne garantit pas que les trajectoires générées conduisent à des concentrations qui soient centrées sur les moyennes expérimentales : pour une période donnée, et considérée de manière indépendante (sans prendre en compte la contrainte de faisabilité entre les périodes), la probabilité d'échantillonner une solution particulière ne dépend que de la forme de l'espace des solutions échantillonné. Cette approche non biaisée permet d'affiner le modèle en comparant les distributions de concentrations prédites avec celles qui ont été observées.

Si l'on souhaite directement estimer les répartitions de flux à l'issue desquelles les concentrations prédites seraient centrées sur les moyennes expérimentales, d'autres approches plus adaptées pourraient être envisagées. Pour obtenir de telles répartitions de flux, la dimension « simulation » de l'approche par trajectoire solution doit être minimisée afin de renforcer l'aspect « analytique ». Ainsi, il ne s'agit plus de « laisser » les trajectoires solutions conduire à n'importe quelles concentrations dans les intervalles de concentrations cibles, il s'agit maintenant de privilégier les solutions qui prédisent des concentrations proches des moyennes. En conséquence, les moyennes des mesures ne pourraient plus être utilisées comme élément de comparaison *a posteriori*, elles seraient directement utilisées comme facteur discriminant pour la sélection de « bonnes » trajectoires solutions.

Afin d'obtenir des successions de cartes de flux qui (i) conduisent à des distributions de concentrations centrées autour de la moyenne (cohérence des valeurs prédites à chaque instant), tout en (ii) conservant la dépendance de faisabilité entre les périodes (cohérence des valeurs prédites sur l'ensemble de la dynamique), on peut par exemple proposer la procédure suivante :

1. **Obtenir de « bonnes » cartes pour chaque période p_i, en traitant chaque période de manière indépendante**. Chaque période est définie sur l'intervalle de temps allant de $t = i$ à $t = i+1$. Il s'agit ici de traiter la première partie du problème, à savoir obtenir des cartes de flux qui soient valables lorsque les concentrations initiales (à $t = i$) et finales (à $t = i+1$) se situent autour des valeurs moyennes. Afin de conserver la cohérence avec la variabilité des données expérimentales, il faudrait toutefois autoriser une certaine variabilité autour des valeurs moyennes. Au final, chaque solution identifiée serait associée à une combinaison de concentrations, avec une valeur de concentration pour chaque composé au temps $t = i$, et une autre valeur pour chaque composé au temps $t = i + 1$. Je proposerai plus loin deux manières d'obtenir de telles cartes.
2. **Rechercher, parmi les cartes obtenues pour chaque période p_i, celles qui permettent de construire de « bonnes » trajectoires solutions**. Il s'agit ici de traiter la seconde partie du problème, à savoir identifier des successions de cartes qui, une fois mises bout à bout, conduiraient à des cinétiques de concentrations qui soient proches des cinétiques moyennes observées. Comme les périodes ont été traitées de manière indépendante pendant l'étape n° 1, il n'y a pas nécessairement d'égalité entre (i) la concentration d'un composé c au temps t_i associée à une carte prédite pour la période p_{i-1} (concentration finale du composé pour la période p_{i-1}), et (ii) la concentration de ce même composé c au même temps t_i, mais cette fois issue d'une carte prédite pour la période p_i (concentration initiale du composé pour la période p_i). En conséquence, la construction d'une bonne trajectoire solution consisterait en fait à retenir une carte de chaque période, telle que, pour chaque composé c :
   - les concentrations associées aux cartes retenues soient les plus proches possible de la moyenne observée à chaque temps ;
   - la distance entre (i) la concentration du composé c au temps t_i issue de la carte p_{i-1} et (ii) celle issue de la carte suivante p_i soit « petite ».

Au niveau de l'étape n° 1, on peut proposer deux manières différentes d'obtenir des solutions.

Une première manière est de tirer au hasard des combinaisons de concentrations initiales et finales. Chaque combinaison de concentrations serait utilisée pour calculer les contraintes d'égalité à appliquer sur les flux E/S du système, et l'espace des solutions ainsi défini serait ensuite échantillonné. Ce faisant, on obtiendrait un échantillon de cartes pour chaque combinaison de concentrations. Les concentrations pourraient par exemple être tirées au sein de lois normales de moyenne et d'écart-type égaux à ceux des mesures expérimentales.

Une seconde manière de procéder pour obtenir des cartes solutions est de ne réaliser qu'un seul échantillonnage des solutions pour chaque période. Pour cela, les contraintes d'inégalité sur les flux E/S seraient calculées de manière à ce que n'importe quelle combinaison de concentrations initiales et finales (concentration possible bornée par exemple par la moyenne des mesures plus ou moins un écart-type) puisse correspondre à des flux E/S qui soient théoriquement représentés dans l'espace des solutions. Un unique flux de consommation peut correspondre à plusieurs combinaisons de concentrations initiales et finales du composé. Les combinaisons de concentration $[c_{0h} = 20\,;\ c_{1h} = 15$ mmol/L], et $[c_{0h} = 18\,;\ c_{1h} = 13$ mmol/L], sont par exemple toutes les deux cohérentes avec un flux de consommation $v_c = 5$ mmol/h. Il s'agirait donc de déterminer les combinaisons de concentrations qui pourraient correspondre aux flux E/S de chaque solution échantillonnée.

L'étape n° 2 consiste a rechercher, parmi les solutions échantillonnées à chaque période, celles qui permettent de construire de « bonnes » trajectoires solutions. L'étude de la variabilité des flux nécessite de construire une population de trajectoires. En conséquence, à la construction de chaque nouvelle trajectoire, il sera nécessaire de ne pas prendre en compte les cartes déjà « attribuées » à des trajectoires existantes.

En définitive, il me semble que ce type d'approche permettrait d'obtenir des populations de trajectoires qui (*i*) seraient cohérentes avec les moyennes des mesures expérimentales et (*ii*) prendraient en compte la variabilité des mesures. La répartition et la variabilité des flux du système pourraient ensuite être analysées afin de caractériser la dynamique du système étudié. Notons ici que cette procédure n'a pas été testée, et qu'en conséquence il serait nécessaire d'approfondir la question afin d'évaluer sa faisabilité.

10.2.2. Approche exploitant la méthode DFBA pour identifier des contraintes sur les flux

Les systèmes métaboliques étudiés sont souvent sous-déterminés, notamment du fait que les données ne sont pas en quantité suffisante pour totalement contraindre les comportements possibles du système. En conséquence, une approche purement prédictive est difficile à concevoir sans l'utilisation d'hypothèses « directrices » qui guideraient fortement la dynamique du comportement au cours du temps. Dans l'approche DFBA, l'hypothèse directrice est que le système optimise un comportement prédéterminé, défini par la fonction *objectif*, et le comportement typiquement optimisé est la production de biomasse.

Pour être pertinente, l'utilisation d'une fonction *objectif* requiert d'une part que le comportement a *priori* optimisé soit connu, et d'autre part que le système soit suffisamment contraint pour conduire à des prédictions raisonnables. Ainsi, l'optimisation de la production de biomasse conduira à des prédictions irréalistes si certains flux, par exemple les flux de consommation des substrats ou encore les flux liés à la chaîne respiratoire, ne sont pas contraints à des valeurs maximales.

La figure 10.2 illustre cette situation. Dans cet exemple « extrême », la méthode DFBA est utilisée pour modélisation de la croissance de *C. glutamicum* en condition de limitation en biotine. Les informations utilisées pour cette dynamique sont uniquement les concentrations en substrat mesurées au début de la culture, le système est une version simplifiée de celle utilisée au fil de cette thèse, et le comportement optimisé correspond à la maximisation de la production de biomasse. On peut voir sur cette figure que, dans la dynamique prédite, la totalité du glucose est consommée dès la première période de la culture, et est converti en biomasse. Ceci illustre le fait que la connaissance du « but » optimisé n'est donc clairement pas suffisante pour modéliser correctement la dynamique d'un système par la méthode DFBA. L'enjeu est aussi de déterminer les contraintes à appliquer afin que l'optimisation de la fonction *objectif* conduise à une dynamique réaliste.

Comment déterminer des contraintes à appliquer sur le système qui permettraient la prédiction d'une dynamique réaliste ?

En dehors de l'utilisation de connaissances déjà disponibles, une manière de déterminer des jeux de contraintes cohérentes serait d'utiliser le principe

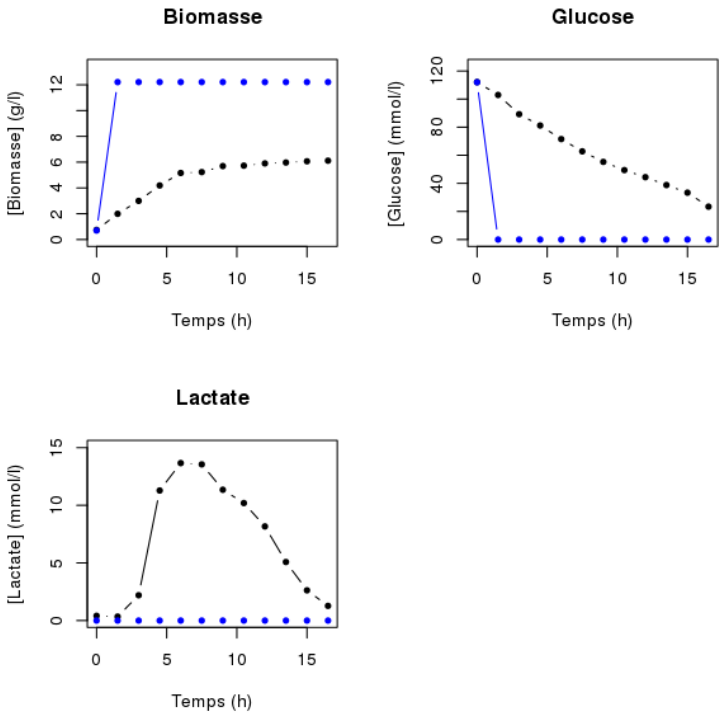


Figure 10.2. – Cinétiques des concentrations prédites à l'issue d'une simulation DFBA lorsqu'aucune contrainte sur les flux n'est appliquée. La concentration, en ordonnée, est exprimée en g/L pour la biomasse, et en mmol/L pour les autres composés. Les points de couleur bleue représentent les concentrations prédites, tandis que les points de couleur noire représentent les concentrations réellement mesurées.

de la méthode DFBA afin de tester différentes combinaisons de contraintes, et d'évaluer leurs impacts sur les cinétiques de concentrations prédites. On peut par exemple proposer la procédure suivante, en trois étapes :

1. Étant donné un ensemble de flux du système à contraindre, choisir une valeur maximale à appliquer sur chacun de ces flux.
2. Utiliser la méthode DFBA pour prédire la dynamique du système, étant données les contraintes choisies à l'étape n° 1, et une fonction *objectif* optimisée (par exemple la maximisation de la production de biomasse).
3. Évaluer le réalisme des cinétiques des concentrations prédites en les comparant avec celles issues des mesures expérimentales. Un « bon » jeu de contraintes permettra de prédire des cinétiques similaires.

Afin de guider le choix des valeurs des contraintes, ces trois étapes pourraient, par exemple, être intégrées dans un algorithme évolutionnaire :

A. Étant donné les flux à contraindre, générer une population de n jeux de contraintes *constantes* (un jeu correspondant à une contrainte sur la valeur maximale de chaque flux, ce qui correspond à l'étape n° 1) ;
B. Générer la dynamique pour chacun des jeux de contraintes (étape n° 2) ;
C. Évaluer chaque dynamique (étape n° 3), et attribuer un score à chaque jeu de contraintes ;
D. Ne retenir que les m jeux de contraintes qui ont permis d'obtenir les meilleurs scores ;
E. Générer $n - m$ nouveau jeux de contraintes, en utilisant des valeurs modifiées des contraintes utilisées dans les m meilleurs jeux ;
F. Revenir ensuite à l'étape B pour évaluer la nouvelle population de n jeux de contraintes, constituées des m meilleurs jeux retenus à l'étape D précédente, et des $n - m$ nouveaux jeux générés à l'étape E précédente.

Le choix des flux à contraindre pourrait être facilité par une analyse des dépendances entre les différents flux du système (compte tenu de la topologie du réseau). En ne sélectionnant que les flux dont les valeurs sont relativement indépendantes, cela permettrait de réduire le nombre de contraintes à évaluer au cours de la procédure.

Le score attribué à chaque jeu de contraintes pourrait être calculé en fonction de la distance entre les valeurs de concentrations prédites et celles mesurées. Un score pourrait par exemple correspondre à la somme des écarts au carré calculés pour chaque composé c_i à chaque temps t :

$$score = \sum_i \sum_t (c_{i,m,t} - \bar{c}_{i,e,t})^2$$

où $c_{i,m,t}$ correspond à la concentration en composé c_i prédite au temps t par la dynamique, et $\bar{c}_{i,e,t}$ correspond à la moyenne des concentrations expérimentales du composé c_i au temps t. En utilisant cette formule, plus un score sera faible et plus le jeu de contrainte pourra être considéré comme valable pour modéliser la dynamique du système.

Bien que la prise en compte de la variabilité expérimentale ne soit pas indispensable dans cette approche, elle pourrait être utilisée afin de pondérer les écarts au carré en fonction de la variabilité des mesures expérimentales, par exemple :

$$score = \sum_i \sum_t \frac{(c_{i,m,t} - \bar{c}_{i,e,t})^2}{\sigma^2_{i,e,t}}$$

où $\sigma_{i,e,t}$ correspond à l'écart-type des mesures expérimentales du composé c_i au temps t.

De cette manière, pour un composé c_i et un temps t donnés, une concentration prédite $c_{i,m,t}$ éloignée de, par exemple, 5 mmol/L de la moyenne des mesures aura un impact plus important sur le score si la variabilité expérimentale est faible (par exemple, si $\sigma_{i,e,t} = 0{,}5$ mmol/L), que si la variabilité des mesures est importantes (par exemple, si $\sigma_{i,e,t} = 5$ mmol/L).

En utilisant un prototype de cette approche, j'ai pu déterminer des jeux de contraintes *constantes* permettant d'obtenir des dynamiques — autrement dit, des successions de cartes — qui prédisent des cinétiques de concentrations se « rapprochant » des cinétiques mesurées expérimentalement. La figure 10.3 montre par exemple les cinétiques obtenues lorsque les contraintes *constantes* données dans la table 10.1 sont utilisées (p. 235). Comme on peut le voir sur la figure, les cinétiques prédites sont encore loin d'être réalistes, notamment parce que la totalité des substrats a été consommée à l'issue de la période 9h $\rightarrow 10,5$h, ce qui conduit à une absence de solution pour les périodes suivantes (la consommation de substrat est nécessaire ne serait ce que pour fournir l'énergie de maintenance des cellules).

Il est toutefois intéressant de noter que l'utilisation de seulement quelques contraintes sur les flux liés à la chaîne respiratoire permet de prédire des phénomènes de production suivis de reconsommation d'acides organiques.

Le fonctionnement des cellules de *C. glutamicum* au cours de la croissance en condition de limitation en biotine n'est clairement pas constant au fil de la culture. De fait, il pourrait être intéressant d'envisager non plus uniquement des contraintes *constantes* sur l'ensemble de la culture, mais aussi des contraintes *constantes par partie*, c'est-à-dire des contraintes dont la valeur changerait par exemple selon la phase de culture considérée. Ces contraintes seraient alors à rapprocher des phénomènes de régulation qui ont lieu au fil de la culture.

À supposer que des jeux de contraintes « cohérentes » puissent ainsi être identifiés, les successions de cartes de flux issues des dynamiques pourraient ensuite être analysées afin d'étudier la dynamique du métabolisme. Les valeurs des contraintes *a priori* cohérentes devraient aussi être confrontées avec les connaissances biologiques disponibles, voire avec d'autres mesures expérimentales, afin d'attester leur pertinence biologique.

Table 10.1. – Exemple de contraintes sur la valeur maximale des flux permettant d'obtenir une dynamique qui « se rapproche » des mesures expérimentales.

Flux	Valeur maximale (mmol/g biom./h)	Flux	Valeur maximale (mmol/g biom./h)
LDH	2,34	PPC	3,38
LLD	1,68	MQO	1,76
PDH	0,69	NDH	2,82

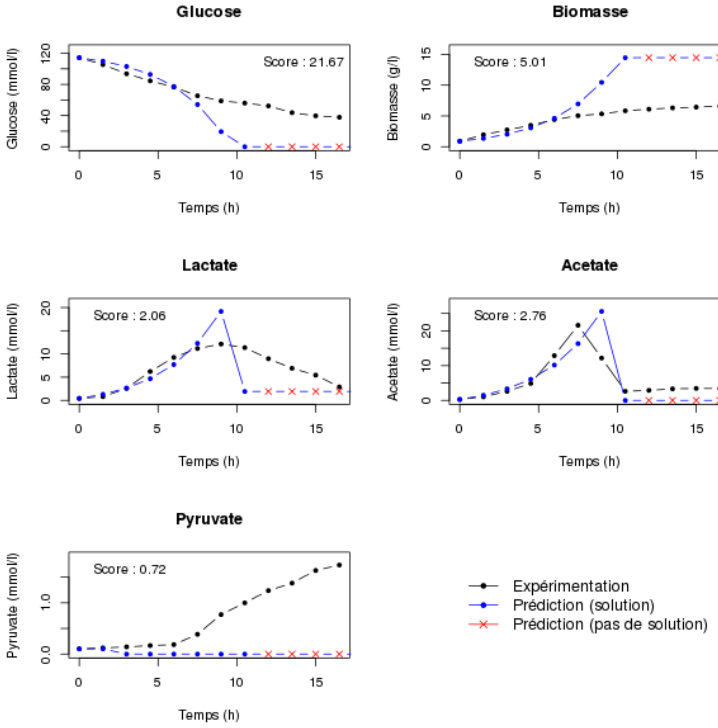


Figure 10.3. – Cinétiques des concentrations prédites à l'issue d'une simulation DFBA lorsque des contraintes « cohérentes » sont appliquées. La concentration, en ordonnée, est exprimée en g/L pour la biomasse, et en mmol/L pour les autres composés. Les points bleus représentent les concentrations prédites, les points noirs représentent les concentrations réellement mesurées. Les croix de couleur rouge indiquent que la dynamique ne peut pas être prolongée de $t = 12$ à $t = 16,5$h. Les contraintes appliquées sont données dans la table 10.1.

Conclusion générale

La biologie des systèmes s'intéresse aux interactions rencontrées dans les systèmes biologiques et à la manière dont ces interactions donnent naissances à des fonctions et processus biologiques. À terme, un idéal à atteindre est une compréhension intégrée des divers phénomènes ayant lieu au sein des êtres vivants. Devant l'ampleur de cette quête, les modèles formels sont des outils de choix pour accompagner les cycles de recherche et d'intégration des connaissances.

Un premier palier dans la compréhension intégrée du vivant peut correspondre à la modélisation de chaque réseau biologique, considéré séparément des autres. Un deuxième palier peut être la prise en compte de la dynamique au sein de ces systèmes. Un palier suivant serait l'étude simultanée de plusieurs systèmes. Vis-à-vis de ces considérations, ce travail de thèse propose plusieurs contributions.

Apports conceptuels

Même si la modélisation passe par une étape de simplification, il n'en reste pas moins important de garder à l'esprit la réalité des systèmes étudiés. Dans la première partie de ce manuscrit, j'ai présenté les acteurs et les actions essentiels au fonctionnement des réseaux métabolique et des réseaux de régulation, ainsi qu'une manière de structurer ces diverses « briques » du vivant.

Dans la seconde partie, j'ai donné quelques éléments de réflexion sur les enjeux de l'étude simultanée d'un réseau métabolique et d'un réseau de régulation, et j'ai considéré le positionnement de quatre formalismes de modélisation vis-à-vis d'une telle étude intégrée.

Apports méthodologiques

À l'issue de ce travail de thèse, je propose une approche, basée sur le formalisme des modèles à base de contraintes, pour étudier la dynamique d'un système métabolique.

En associant l'échantillonnage de l'espace des solutions avec l'utilisation d'une contrainte de « faisabilité » entre les périodes de temps considérées, cette approche permet de modéliser la dynamique d'un métabolisme en prenant en compte la variabilité des mesures expérimentales. La contrainte

de faisabilité entre les périodes permet de garantir que chaque « trajectoire solution » correspond à une succession de cartes de flux qui conduit à des cinétiques de concentrations cohérentes avec les mesures expérimentales.

Les populations de trajectoires solutions générées autorisent différents types d'analyses. D'une part, les répartitions de flux prédites peuvent être utilisées afin d'estimer les répartitions de matières les plus plausibles au sein du réseau étudié. D'autre part, la distribution des concentrations prédites permet d'évaluer le modèle utilisé pour étudier le réseau métabolique.

Le fait que cette approche soit basée sur le formalisme de la modélisation à base de contraintes permet, moyennant l'utilisation de l'hypothèse d'état stationnaire du système, d'une part d'étudier des réseaux métaboliques de taille relativement grande, et d'autre part d'utiliser des données expérimentales qui sont aisément mesurables, par exemple les concentrations en biomasse et en métabolites extracellulaires.

Apports biologiques

Les projets de recherche de mon équipe d'accueil étant tournés vers l'étude du fonctionnement et de la régulation du métabolisme de *Corynebacterium glutamicum* en conditions de stress, l'approche par « trajectoires solutions » a été utilisée afin d'étudier la dynamique de cette bactérie, lorsqu'elle est cultivée en condition de limitation en biotine.

Les résultats obtenus ont permis d'une part d'attester du fonctionnement de la méthode, et d'autre part de soulever plusieurs hypothèses quant aux phénomènes biologiques qui ont lieu pendant cette condition particulière de croissance.

Perspectives

En collaboration avec des scientifiques issus des champs disciplinaires des sciences formelles, il pourrait être intéressant d'examiner si le problème abordé dans l'approche par « trajectoire solution » pourrait être formulé de façon différente. En particulier, il s'agirait d'évaluer si la recherche de trajectoire solution, actuellement réalisée de manière itérative, pourrait être réalisée en une seule itération, en considérant l'ensemble de la culture.

On pourrait par ailleurs reprocher que les successions de cartes prédites ne soient pas nécessairement celles qui conduisent à des concentrations proches des concentrations moyennes observées expérimentalement. Afin d'obtenir de telles cartes de flux, tout en conservant la dépendance entre les périodes de temps considérées, j'ai proposé quelques éléments de réflexions pour la conception d'une autre approche. Il pourrait être intéressant d'évaluer ces propositions et, le cas échéant, de tester une telle approche.

La réalisation de quelques expériences supplémentaires permettrait quant à elle d'(in)valider les hypothèses émises sur le fonctionnement du métabolisme de *C. glutamicum*.

Annexes

Annexe A

Requêtes sur la base de données EcoCyc

Cette annexe liste les requêtes effectuées sur la base de données EcoCyc afin d'obtenir les données présentées dans le chapitre 3 (p. 51).

Les requêtes utilisent un langage de requête nommé BioVelo [Latendresse 2010], qui est implémenté dans le logiciel Pathway Tools [Karp 2002]. Les requêtes peuvent être directement saisies sur le site web d'EcoCyc, à l'adresse `http://ecocyc.org/query.shtml` (onglet « *free form advanced query page* »).

Pour chaque requête, un résumé du résultat est donné (résultats obtenus le 12 novembre 2014).

Requête n° 1. Obtenir la liste des gènes référencés dans la base, et qui codent au moins un produit de gène (ARN, polypeptide).

```
html-table-headers(
[(gene, produit):
  gene <- ecoli^^Genes,
    produit := gene^Product,
    #produit > 0
]
)
```

Résultat ⇒ liste de 4 489 gènes.

Requête n° 2. Obtenir la liste des gènes dont au moins un des produits catalyse une réaction chimique ou un transport.

```
html-table-headers(
[(gene, produit):
  gene <- ecoli^^Genes,
    produit := gene^Product,
    #produit > 0,
    enzyme := (gene-to-proteins gene),
    #enzyme > 0
]
)
```

Résultat ⇒ liste de 1 781 gènes.

Requête n° 3. Obtenir la liste des entités biologiques qui catalysent au moins une réaction chimique ou un transport (ARN, protéines et complexes).

```
html-table-headers(
{(prot):
  gene <- ecoli^^Genes,
    produit := gene^Product,
    #produit > 0,
    enzyme := (gene-to-proteins gene),
    #enzyme > 0,
    react <- (gene-to-reactions gene),
      prot <- (reaction-to-proteins react)
}
)
```

Résultat ⇒ liste de 1 509 entités.

Requête n° 4. Obtenir la liste des réactions et des transports catalysés par les entités biologiques obtenues avec la requête précédente.

```
html-table-headers(
{(react):
  gene <- ecoli^^Genes,
    produit := gene^Product,
    #produit > 0,
    enzyme := (gene-to-proteins gene),
    #enzyme > 0,
      react <- (gene-to-reactions gene)
}
)
```

Résultat ⇒ liste de 1 841 processus.

Requête n° 5. Obtenir la liste des gènes qui codent un régulateur transcriptionnel.

```
html-table-headers(
{(geneRegulateur):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulateur <- (genes-regulating-gene gene)
}
)
```

Résultat ⇒ liste de 249 gènes.

Requête n° 6. Obtenir la liste des gènes qui codent un régulateur transcriptionnel. Ne retenir que les régulateurs transcriptionnels qui peuvent former un complexe avec un métabolite.

```
html-table-headers(
{(geneRegulateur):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulateur <- (genes-regulating-gene gene),
      mono <- geneRegulateur^Product,
        comp <- (containers-of mono),
          #[
            react:
            react <- mono^appears-in-left-side-of
              ++ mono^appears-in-right-side-of
              ++ comp^appears-in-left-side-of
              ++ comp^appears-in-right-side-of,
            react isa "Binding-Reactions"
          ] > 0
}
)
```

Résultat ⇒ liste de 85 gènes. Il faut ajouter à cette liste le gène *rpiR*, qui code le régulateur transcriptionnel AlsR. AlsR peut former un complexe avec le métabolite D-allose. Ce gène n'est pas présent dans la liste à cause d'une erreur dans la base EcoCyc, au niveau de l'assignation de la réaction de formation du complexe.

Requête n° 7. Obtenir la liste des métabolites qui sont impliqués dans la formation d'un complexe avec les régulateurs transcriptionnels obtenus lors de la requête précédente.

```
html-table-headers(
{(substrates):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulateur <- (genes-regulating-gene gene),
      mono <- geneRegulateur^Product,
        comp <- (containers-of mono),
          #[
            react:
            react <- mono^appears-in-left-side-of
              ++ mono^appears-in-right-side-of
              ++ comp^appears-in-left-side-of
              ++ comp^appears-in-right-side-of,
            react isa "Binding-Reactions"
          ] > 0,
          reactWProt <- mono^appears-in-left-side-of
            ++ mono^appears-in-right-side-of
            ++ comp^appears-in-left-side-of
            ++ comp^appears-in-right-side-of,
            substrates <- [
              s: s <- reactWProt^Substrates,
              s isa "Compounds"
            ]
}
)
```

Résultat ⇒ liste de 84 métabolites, auxquels il faut ajouter le D-allose qui forme un complexe avec le régulateur AlsR.

Requête n° 8. Obtenir la liste des gènes qui sont régulés par un régulateur transcriptionnel qui est capable de former un complexe avec un métabolite.

```
html-table-headers(
{(geneRegulated):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulateur <- (genes-regulating-gene gene),
      mono <- geneRegulateur^Product,
        comp <- (containers-of mono),
          #[
            react:
            react <- mono^appears-in-left-side-of
              ++ mono^appears-in-right-side-of
              ++ comp^appears-in-left-side-of
              ++ comp^appears-in-right-side-of,
            react isa "Binding-Reactions"
          ] > 0,
          geneRegulated <- (genes-regulated-by-gene geneRegulateur)
}
)
```

Résultat ⇒ liste de 1 203 gènes, auxquels il faut ajouter les 6 gènes régulés par AlsR (*rpiR*, *alsA*, *alsB*, *alsC*, *alsE* et *rpiB*).

Requête n° 9. Obtenir la liste des gènes qui sont régulés par un régulateur transcriptionnel capable de former un complexe avec un métabolite. Ne retenir que les gènes dont au moins un des produits catalyse une réaction chimique ou un transport.

```
html-table-headers(
{(geneRegulated):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulateur <- (genes-regulating-gene gene),
      mono <- geneRegulateur^Product,
        comp <- (containers-of mono),
          #[
            react:
            react <- mono^appears-in-left-side-of
              ++ mono^appears-in-right-side-of
              ++ comp^appears-in-left-side-of
              ++ comp^appears-in-right-side-of,
            react isa "Binding-Reactions"
          ] > 0,
          geneRegulated <- (genes-regulated-by-gene geneRegulateur),
            enzyme <- (gene-to-proteins geneRegulated),
            #[enzyme > 0]
}
)
```

Résultat ⇒ liste de 737 gènes, auxquels il faut ajouter 5 gènes régulés par AlsR et impliqués dans le métabolisme (*alsA*, *alsB*, *alsC*, *alsE* et *rpiB*).

Requête n° 10. Obtenir la liste des gènes qui codent un régulateur transcriptionnel capable de former un complexe avec un métabolite. Ne retenir que les gènes dont le produit régule des gènes codant des entités impliqués dans le métabolisme.

```
html-table-headers(
{(geneRegulateur):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulateur <- (genes-regulating-gene gene),
      mono <- geneRegulateur^Product,
        comp <- (containers-of mono),
          #[
            react:
            react <- mono^appears-in-left-side-of
              ++ mono^appears-in-right-side-of
              ++ comp^appears-in-left-side-of
              ++ comp^appears-in-right-side-of,
            react isa "Binding-Reactions"
          ] > 0,
          geneRegulated <- (genes-regulated-by-gene geneRegulateur),
            enzyme <- (gene-to-proteins geneRegulated),
            #[enzyme > 0]
}
)
```

Résultat ⇒ liste de 82 gènes, auxquels il faut ajouter le régulateur AlsR.

Requête n° 11. Obtenir la liste des gènes qui sont régulés par un régulateur transcriptionnel.

```
html-table-headers(
{(geneRegulated):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulated <- (genes-regulated-by-gene gene)
}
)
```

Résultat ⇒ liste de 1 873 gènes.

Requête n° 12. Obtenir la liste des gènes qui sont régulés par un régulateur transcriptionnel et qui codent une entité catalysant une réaction ou un transport.

```
html-table-headers(
{(geneRegulated):
  gene <- ecoli^^genes,
    produit := gene^Product,
    #produit > 0,
    geneRegulated <- (genes-regulated-by-gene gene),
      enzyme := (gene-to-proteins geneRegulated),
      #enzyme > 0
}
)
```

Résultat ⇒ liste de 1 017 gènes.

Requête n° 13. Obtenir la liste des métabolites qui sont impliqués dans une réaction ou un transport catalysé par un produit de gène connu.

```
html-table-headers(
{(met):
  gene <- ecoli^^Genes,
    produit := gene^Product,
    #produit > 0,
    enzyme := (gene-to-proteins gene),
    #enzyme > 0,
      react <- (gene-to-reactions gene),
        met <- react^left ++ react^right,
          met isa "Compounds"
}
)
```

Résultat ⇒ liste de 1 413 métabolites.

Requête n° 14. Obtenir la liste des métabolites qui sont impliqués dans une réaction ou un transport référencé dans EcoCyc.

```
html-table-headers(
{(comp):
  react <- ecoli^^Reactions,
    comp <- (reaction-to-compounds react)
}
)
```

Résultat ⇒ liste de 2 215 métabolites. La différence avec la requête précédente est due au fait que certaines réactions et certains transports référencés dans EcoCyc ne sont pas associées à des catalyseurs.

Annexe B

Données supplémentaires de l'article

Cette annexe contient les données supplémentaires associées à l'article du chapitre 7 (p. 129).

Table 1 : liste des composés décrits dans le modèle

Compound	Molecular formula	Full name
13PDG	$C_3H_4O_{10}P_2$	1,3-bisphospho-D-glycerate
2PG	$C_3H_4O_7P$	2-phospho-D-glycerate
3PG	$C_3H_4O_7P$	3-phospho-D-glycerate
ACCOA	$C_{23}H_{34}N_7O_{17}P_3S$	acetyl-CoA
ACE	$C_2H_3O_2$	acetate
ACEP	$C_2H_3O_5P$	acetyl phosphate
ACO	$C_6H_3O_6$	cis-aconitate
ADP	$C_{10}H_{12}N_5O_{10}P_2$	adenosine-diphosphate
AKG	$C_5H_4O_5$	α-ketoglutarate
ATP	$C_{10}H_{12}N_5O_{13}P_3$	adenosine-triphosphate
CIT	$C_6H_5O_7$	citrate
CO2	CO_2	carbon dioxide
COA	$C_{21}H_{32}N_7O_{16}P_3S$	coenzyme A
D6PGC	$C_6H_{10}O_{10}P$	D-gluconate 6-phosphate
D6PGL	$C_6H_9O_9P$	6-phospho D-glucono-1,5-lactone
E4P	$C_4H_7O_7P$	D-erythrose 4-phosphate
F1P	$C_6H_{11}O_9P$	fructose 1-phosphate
F6P	$C_6H_{11}O_9P$	D-fructose 6-phosphate
FDP	$C_6H_{10}O_{12}P_2$	fructose 1,6-bisphosphate
FUMA	$C_4H_2O_4$	fumarate
G6P	$C_6H_{11}O_9P$	D-glucose 6-phosphate
GDP	$C_{10}H_{12}N_5O_{11}P_2$	guanosine-diphosphate
GLC	$C_6H_{12}O_6$	D-glucose

GLX	C_2HO_3	glyoxylate
GTP	$C_{10}H_{12}N_5O_{14}P_3$	guanosine-triphosphate
H	H	proton (for cellular respiration)
H2O	H_2O	water
ICIT	$C_6H_5O_7$	isocitrate
LAC	$C_3H_5O_3$	L-lactate
MAL	$C_4H_4O_5$	malate
MQ	$C_{51}H_{72}O_2$	menaquinone-8
MQH2	$C_{51}H_{74}O_2$	menaquinol-8 (reduced)
NAD	$C_{21}H_{26}N_7O_{14}P_2$	nicotinamide adenine dinucleotide
NADH	$C_{21}H_{27}N_7O_{14}P_2$	nicotinamide adenine dinucleotide (reduced)
NADP	$C_{21}H_{25}N_7O_{17}P_3$	nicotinamide adenine dinucleotide phosphate
NADPH	$C_{21}H_{26}N_7O_{17}P_3$	nicotinamide adenine dinucleotide phosphate (reduced)
O2	O_2	dioxygen
OA	$C_4H_2O_5$	oxaloacetate
PEP	$C_3H_2O_6P$	phosphoenolpyruvate
PI	HO_4P	inorganic phosphate
PYR	$C_3H_3O_3$	pyruvate
R5P	$C_5H_9O_8P$	D-ribose 5-phosphate
RL5P	$C_5H_9O_8P$	D-ribulose 5-phosphate
S7P	$C_7H_{13}O_{10}P$	D-sedoheptulose 7-phosphate
SUCC	$C_4H_4O_4$	succinate
SUCCOA	$C_{25}H_{35}N_7O_{19}P_3S$	succinyl-CoA
T3P1	$C_3H_5O_6P$	D-glyceraldehyde 3-phosphate
T3P2	$C_3H_5O_6P$	dihydroxyacetone phosphate
X5P	$C_5H_9O_8P$	D-xylulose 5-phosphate
ACEex	$C_2H_3O_2$	acetate (extracellular)
Biomass		biomass (extracellullar)
CO2ex	CO_2	carbon dioxide (extracellular)
FRCex	$C_6H_{12}O_6$	D-fructose (extracellular)
GLCex	$C_6H_{12}O_6$	D-glucose (extracellular)
H2Oex	H_2O	water (extracellular)
LACex	$C_3H_5O_3$	L-lactate (extracellular)
O2ex	O_2	dioxygen (extracellular)
PIex	HO_4P	inorganic phosphate (extracellular)
PYRex	$C_3H_3O_3$	pyruvate (extracellular)

Table 2 : liste des réactions décrites dans le modèle

La table qui suit présente l'ensemble des réactions décrites dans le modèle. Les colonnes apportent les informations suivantes :

— *Metabolic process* : la voie ou le processus métabolique concerné.

— *Flux model* : le nom de la réaction dans le modèle.

— *Gene name* : le nom du ou des gènes codant l'enzyme qui catalyse la réaction.

— *N° access.* : le numéro d'accession du ou des gènes dans le génome de référence (RefSeq : NC_006958.1).

— *Enzyme name* : le nom de l'enzyme qui catalyse la réaction.

— *N° EC* : le numéro EC de la réaction.

— *Reaction* : la réaction telle que décrite dans le modèle. Une flèche à sens unique « $\Rightarrow$ » indique que la réaction est considérée comme irréversible, tandis qu'une flèche à double sens » $\Leftrightarrow$ » indique que la réaction est réversible. Seuls les protons associés à la chaîne respiratoire sont pris en compte.

— ΔG^m : l'enthalpie libre de la réaction, estimée pour des concentrations en métabolites de 1 mM. Les valeurs ont été obtenues en 2013 avec l'outil *eQuilibrator* [Noor 2012].

— *Reference* : liste de publications apportant des informations sur la réaction considérée.

Metabolic process	Flux model	Gene name	N° access.	Enzyme name	N° EC	Reaction	ΔG^m (kJ/-mol)	Reference
glycolysis	GLK	*glk*	cg2399	glucose kinase	2.7.1.2	GLC + ATP ⇒ G6P + ADP	−24.9	[Park 2000]
	GPI	*pgi*	cg0973	glucose 6-P isomerase	5.3.1.9	G6P ⇔ F6P	3.3	
	PFKA	*pfkA*	cg1409	6-phosphofructokinase	2.7.1.11	F6P + ATP ⇒ FDP + ADP	−17.4	
	FBP	*fbp*	cg1157	fructose 1,6-biphosphatase	3.1.3.11	FDP + H2O ⇒ F6P + PI	−37.3	[Rittmann 2003]
	PFKB	*pfkb*	cg2119	1-phosphofructokinase	2.7.1.56	F1P + ATP ⇒ FDP + ADP + PI	−16.1	
	FBA	*fba*	cg3068	fructose 1,6-biP aldolase	4.1.2.13	FDP ⇔ T3P1 + T3P2	.2	[von der Osten 1989]
	TPI	*tpi*	cg1789	triose-P isomerase	5.3.1.1	T3P2 ⇔ T3P1	7.7	[Eikmanns 1992]
	GAP	*gap*	cg1791	glyceralhdehyde 3-P dehydrogenase	1.2.1.12	T3P1 + NAD + PI ⇔ 13PDG + NADH	25.0	[Eikmanns 1992]
	PGK	*pgk*	cg1790	phosphoglycerate kinase	2.7.2.3	13PDG + ADP ⇔ 3PG + ATP	−7.8	[Eikmanns 1992], [Reddy 2014]
	GPM	*gpm*	cg0482	phosphoglycerate mutase	5.4.2.1	3PG ⇔ 2PG	5.7	
	ENO	*eno*	cg1111	phosphopyruvate hydratase	4.2.1.11	2PG ⇔ PEP + H2O	−2.0	[Petersen 2000]
	PK	*pyk*	cg2291	pyruvate kinase	2.7.1.40	PEP + ADP ⇒ PYR + ATP	−30.6	[Jetten 1994a], [Jetten 1994b], [Petersen 2000], [Petersen 2001], [Netzer 2004], [Sauer 2005], [Becker 2008], [Sawada 2010]
	PDH	*aceE*, *aceF*, *lpd*	cg2466, cg2421, cg0441	pyruvate dehydrogenase	1.2.1.-	PYR + COA + NAD ⇒ ACCOA + NADH + CO2	−23.5	[Cocaign 1993], [Petersen 2001], [Sauer 2005], [Schreiner 2005b], [Wieschalka 2013]
pentose phosphate pathway	G6PDH	*zwf*	cg1778	glucose 6-P 1-dehydrogenase	1.1.1.49	G6P + NADP ⇔ D6PGL + NADPH	3.4	
	PGL	*devB*	cg1780	6-phosphogluconolactonase	3.1.1.31	D6PGL + H2O ⇒ D6PGC	−26.1	
	GND	*gnd*	cg1643	6-phosphogluconate dehydrogenase	1.1.1.44	D6PGC + NADP ⇔ RL5P + CO2 + NADPH	−8.5	
	RPI	*rpi*	cg2658	ribose-5-P isomerase	5.3.1.6	RL5P ⇔ R5P	−0.5	
	RPE	*rpe*	cg1801	ribulose-p 3-epimerase	5.1.3.1	RL5P ⇔ X5P	.0	
	TKT1	*tkt*	cg1774	transketolase	2.2.1.1	R5P + X5P ⇔ T3P1 + S7P	5.0	
	TKT2	*tkt*	cg1774	transketolase	2.2.1.1	X5P + E4P ⇔ T3P1 + F6P	−4.7	
	TAL	*tal*	cg1776	transaldolase	2.2.1.2	T3P1 + S7P ⇔ F6P + E4P	−4.7	
TCA cycle	CS	*gltA*	cg0949	citrate synthase	2.3.3.1	OA + ACCOA + H2O ⇒ CIT + COA	−40.3	[Eikmanns 1994], [van Ooyen 2011]
	ACN1	*acn*	cg1737	aconitate hydratase	4.2.1.3	CIT ⇔ ACO + H2O	8.9	
	ACN2	*acn*	cg1737	aconitate hydratase	4.2.1.3	ACO + H2O ⇔ ICIT	−2.3	
	ICD	*icd*	cg0766	isocitrate dehydrogenase	1.1.1.42	ICIT + NADP ⇒ AKG + NADPH + CO2	−14.5	[Eikmanns 1995]
	ODH	*odhA*, *lpd*	cg1280, cg0441	oxaloglutarate dehydrogenase	1.2.4.2	AKG + NAD + COA ⇒ SUCCOA + CO2 + NADH	−31.7	[Hoffelder 2010]
	SCS	*sucC*, *sucD*	cg2837, cg2836	succinyl-CoA synthetase	6.2.1.5	SUCCOA + ADP + PI ⇔ SUCC + COA + ATP	6.1	[Cho 2010]
	SDH	*sdhA*, *sdhB*, *sdhC*	cg0446, cg0447, cg0445	succinate dehydrogenase	1.3.5.4	SUCC + MQ + 2 X H ⇔ FUMA + MQH2	−11.2	[Bott 2003], [Kurokawa 2005], [Inui 2007], [Bussmann 2009], [Wieschalka 2013]

	FUM	*fum*	cg1145	fumarate hydratase	4.2.1.2	FUMA + H2O ⇔ MAL	−3.6	[Wieschalka 2013]
	MQO	*mqo*	cg2192	malate :quinone oxidoreductase	1.1.5.4	MAL + MQ ⇒ OA + MQH2	−53.9	[Molenaar 1998], [Molenaar 2000], [Bott 2003], [Mitsuhashi 2006]
	MDH	*mdh*	cg2613	malate dehydrogenase	1.1.1.37	OA + NADH ⇔ MAL + NAD	−32.1	[Molenaar 2000, Genda 2003]
plerotic reactions	PPC	*ppc*	cg1787	phosphoenolpyruvate carboxylase	4.1.1.31	PEP + CO2 ⇒ OA + PI	−43.2	[Jetten 1994b], [Peters-Wendisch 1997], [Peters-Wendisch 1998], [Petersen 2000], [Petersen 2001], [Sauer 2005], [Sato 2008], [Sawada 2010]
	PPCk	*pck*	cg3169	phosphoenolpyruvate carboxykinase	4.1.1.32	OA + GTP ⇒ PEP + GDP + CO2	−4.6	[Jetten 1994b], [Peters-Wendisch 1998], [Petersen 2000], [Petersen 2001], [Riedel 2001], [Gerstmeir 2003], [Sauer 2005], [Sawada 2010]
	PC	*pyc*	cg0791	pyruvate carboxylase	6.4.1.1	PYR + ATP + CO2 ⇒ OA + ADP + PI	−12.6	[Jetten 1994b], [Peters-Wendisch 1997], [Peters-Wendisch 1998], [Uy 1999], [Petersen 2000], [Peters-Wendisch 2001], [Petersen 2001], [Aoki 2005], [Sauer 2005], [Hasegawa 2008], [Sato 2008], [Cao 2014]
	OD	*odx*	cg1458	oxaloacetate decarboxylase	4.1.1.3	OA ⇒ PYR + CO2	−42.2	[Jetten 1995], [Peters-Wendisch 1998], [Petersen 2000], [Petersen 2001], [Sauer 2005], [Klaffl 2010]
	ME	*malE*	cg3335	malic enzyme	1.1.1.40	MAL + NADP ⇔ PYR + NADPH + CO2	−7.2	[Peters-Wendisch 1997], [Peters-Wendisch 1998], [Gourdon 2000], [Petersen 2000], [Petersen 2001], [Netzer 2004], [Sauer 2005]
	ICL	*aceA*	cg2560	isocitrate lyase	4.1.3.1	ICIT ⇒ GLX + SUCC	−19.7	
	MS	*aceB*	cg2559	malate synthase	2.3.3.9	ACCOA + GLX + H2O ⇒ MAL + COA	−44.4	[Reinscheid 1994]
acetate	PQO	*pqo*	cg2891	pyruvate :quinone oxydoreductase	1.2.5.1	PYR + MQ + H2O ⇒ ACE + MQH2 + CO2	−142.8	[Bott 2003], [Schreiner 2005a], [Schreiner 2006], [Wieschalka 2013]
	PTA	*pta*	cg3048	phosphate acetyltransferase	2.3.1.8	ACCOA + PI ⇔ ACEP + COA	1.7	[Gerstmeir 2003], [Veit 2009]
	AK	*ack*	cg3047	acetate kinase	2.7.2.1	ACEP + ADP ⇔ ACE + ATP	−3.8	[Gerstmeir 2003], [Veit 2009]
lactate	LDH	*ldh*	cg3219	L-lactate dehydrogenase	1.1.1.27	PYR + NADH ⇒ LAC + NAD	−27.2	[Hasegawa 2008], [Dietrich 2009], [Wieschalka 2013]
	LLD	*lld*	cg3227	quinone-dependent L-lactate dehydrogenase	1.1.2.3	LAC + MQ ⇒ PYR + MQH2	−53.9	[Bott 2003]
transport	PTSGX	*ptsG*, *ptsI*, *ptsH*	cg1537, cg2117, cg2121	PTS glucose		GLCex + PEP ⇒ G6P + PYR	−55.5	[Cocaign-Bousquet 1996], [Parche 2001], [Lindner 2011], [Ikeda 2012]
	PTSFX	*ptsF*, *ptsI*, *ptsH*	cg2120, cg2117, cg2121	PTS fructose		FRCex + PEP ⇒ FRC + PYR	−48.8	[Ikeda 2012]

	IOLTX ACEX LACX PYRX CO2X O2X PIX H2OX	*iolT1* *mctC*	cg0223 cg0953	glucose permease acetate permease		GLCex ⇒ GLC ACEex ⇔ ACE LACex ⇔ LAC PYRex ⇔ PYR CO2ex ⇔ CO2 O2ex ⇔ O2 PIex ⇔ PI H2Oex ⇔ H2O		[Lindner 2011], [Ikeda 2011], [Ikeda 2012] [Ebbighausen 1991], [Jolkver 2009]
respiration	CYTbd	*cydA*, *cydB*	cg1301, cg1300	cytochrome bd-type menaquinol oxidase	1.10.3.-	MQH2 + 0.5 x O2 ⇒ MQ + 2 x H + H2O		[Bott 2003]
	CYTaa	*qrcA*, *qrcB*, *qrcC*, *ctaC*, *ctaD*, *ctaE*, *ctaF*	cg2404, cg2403, cg2405, cg2409, cg2780, cg2406, cg2408	cytochrome bc_1-aa_3 supercomplex	1.10.3.-	MQH2 + 0.5 x O2 ⇒ MQ + 6 x H + H2O		[Bott 2003]
	ATPase	*atpA*, *atpB*, *atpC*, *atpD*, *atpE*, *atpF*, *atpG*, *atpH*	cg1366, cg1362, cg1369, cg1368, cg1363, cg1364, cg1367, cg1365	ATP synthase		ADP + PI + 3 x H ⇒ ATP		[Cocaign-Bousquet 1996], [Bott 2003], [Klapa 2003], [Shimizu 2003], [Varela 2003], [Varela 2004], [Shirai 2005]
	NDH	*ndh*	cg1656	NADH dehydrogenase	1.6.5.-	NADH + MQ ⇒ NAD + MQH2		[Bott 2003], [Nantapong 2005]
ATP regen.	NDK	*ndk*	cg2603	nucleoside diphosphate kinase	2.7.4.6	GTP + ADP ⇔ GDP + ATP		
APT maint.	ATPm			ATP maintenance		ATP ⇒ ADP + PI		[Stouthamer 1975], [Russell 1979], [Mulder 1986], [Vallino 1993], [Varma 1994], [Varela 2004], [Borodina 2005], [Coze 2013]
growth	BIOM			biomass production		0.318 x G6P + 0.126 x F6P + 0.684 x R5P + 0.198 x E4P + 0.062 x T3P1 + 0.872 x 3PG + 0.458 x PEP + 2.809 x PYR + 3.122 x ACCOA + 1.389 x AKG + 1.168 x OA + 29.20 x ATP + 11.51 x NADPH + 0.39 x NADH ⇒ 1 x Biomass + 3.122 x COA + 29.20 x ADP + 29.20 x PI + 11.51 x NADP + 0.39 x NAD		[Cocaign-Bousquet 1996], [Petersen 2000], [Zhao 2002], [Wittmann 2004a], [Wittmann 2004b], [Ozcan 2007], [Kjeldsen 2009], [Shinfuku 2009], [Bansal-Mutalik 2011]

Table 3 : mesures expérimentales

Les tables qui suivent donnent les concentrations en métabolites extracellulaires et en biomasse utilisées.

	Glucose (mmol/L)			Acetate (mmol/L)		
Time (h)	rep1	rep2	rep3	rep1	rep2	rep3
0.0	114.887	115.170	112.733	0.067	0.799	0,000
1.5	102.559	108.370	105.379	0.083	1.349	1.649
3.0	93.445	90.320	97.202	1.499	2.814	3.530
4.5	84.220	81.516	88.127	3.097	4.063	7.494
6.0	75.472	79.474	74.734	10.142	14.471	14.038
7.5	64.765	68.206	63.299	22.265	20.883	21.682
9.0	57.310	60.363	59.020	15.487	11.774	9.242
10.5	56.944	54.107	56.955	2.714	1.494	3.680
12.0	53.958	50.505	52.720	2.998	1.982	3.697
13.5	40.581	45.587	45.060	3.547	2.781	3.664
15.0	40.359	38.882	39.798	3.747	2.831	3.83
16.5	40.614	36.889	36.024	3.414	3.097	3.73

	Lactate (mmol/L)			Pyruvate (mmol/L)		
Time (h)	rep1	rep2	rep3	rep1	rep2	rep3
0.0	0.433	0.433	0.433	0.102	0.102	0.102
1.5	0.566	0.966	0.91	0.114	0.136	0.114
3.0	2.131	2.764	2.853	0.125	0.148	0.148
4.5	5.839	6.028	6.827	0.159	0.170	0.170
6.0	9.669	9.458	8.637	0.216	0.148	0.193
7.5	12.633	10.846	10.069	0.477	0.329	0.352
9.0	13.099	11.823	11.49	0.92	0.772	0.625
10.5	11.423	11.024	11.667	1.158	0.874	0.965
12.0	8.792	8.492	9.647	1.272	1.226	1.215
13.5	7.216	7.449	6.061	1.522	1.431	1.204
15.0	5.639	6.483	4.241	1.806	1.647	1.442
16.5	3.408	4.674	0.622	1.851	1.828	1.533

	Biomass (g. CDW/L)		
Time (h)	rep1	rep2	rep3
0.0	0.900	0.900	0.900
1.5	1.958	1.947	1.953
3.0	2.743	2.730	2.748
4.5	3.514	3.513	3.430
6.0	4.430	4.416	4.392
7.5	5.013	5.083	5.043
9.0	5.293	5.388	5.350
10.5	5.812	5.827	5.885
12.0	6.090	6.066	6.166
13.5	6.333	6.301	6.307
15.0	6.499	6.298	6.498
16.5	6.670	6.567	6.610

Annexe C

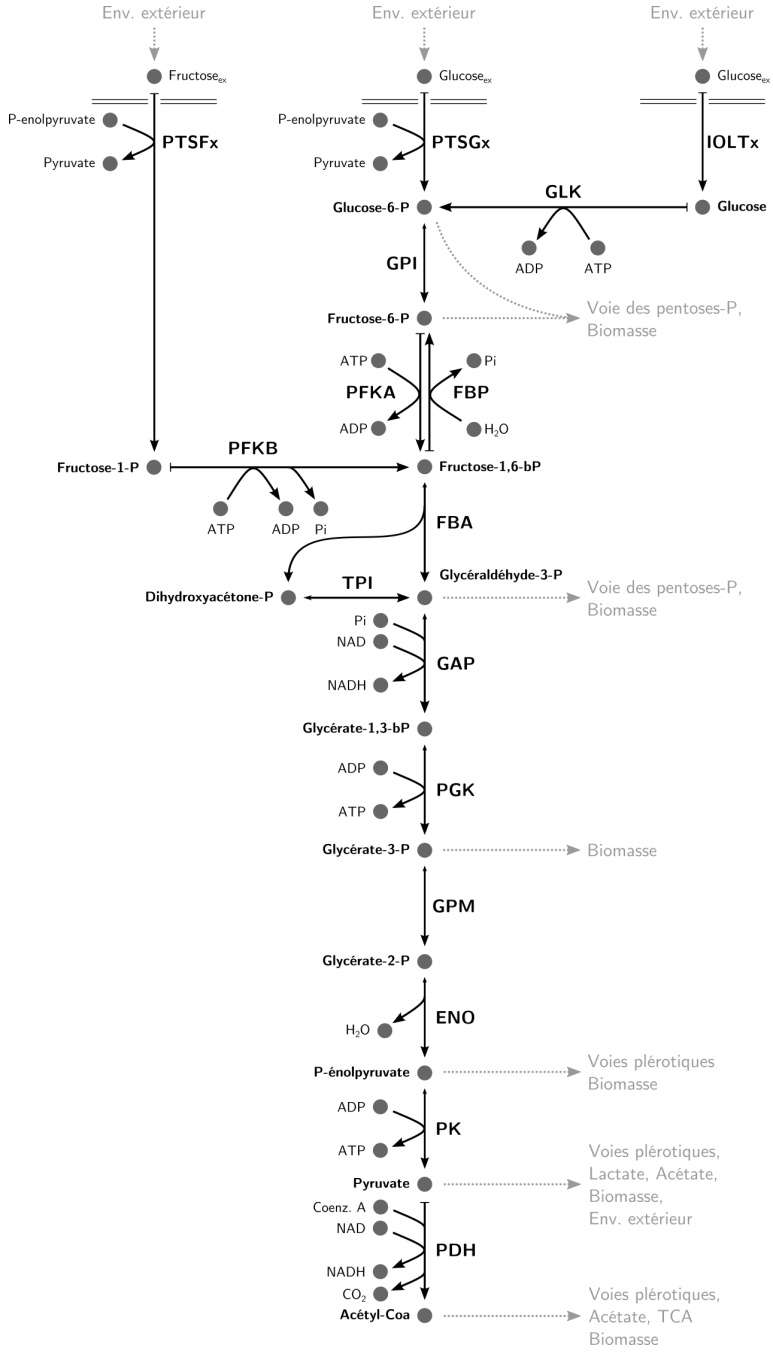
Schémas du réseau métabolique

Cette annexe donne, en complément de la description textuelle du réseau de l'annexe B (p. 253), des schémas qui illustrent différentes parties du réseau étudié. Pour chaque schéma :

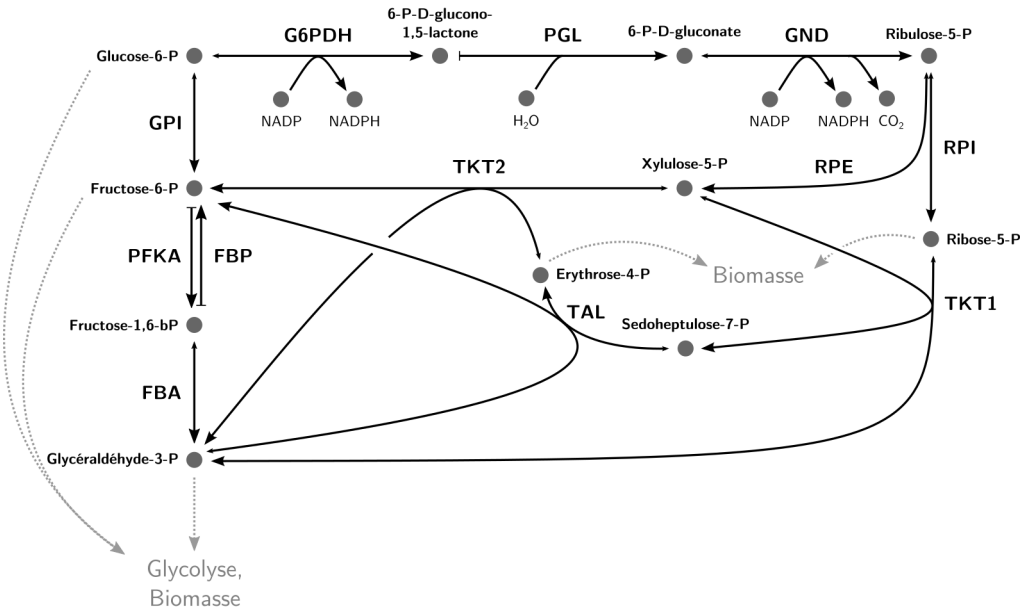
— le nom et la description complète des réactions sont disponibles en annexe B,

— les liens avec d'autres parties du réseau et avec l'environnement extérieur sont indiqués en grisé.

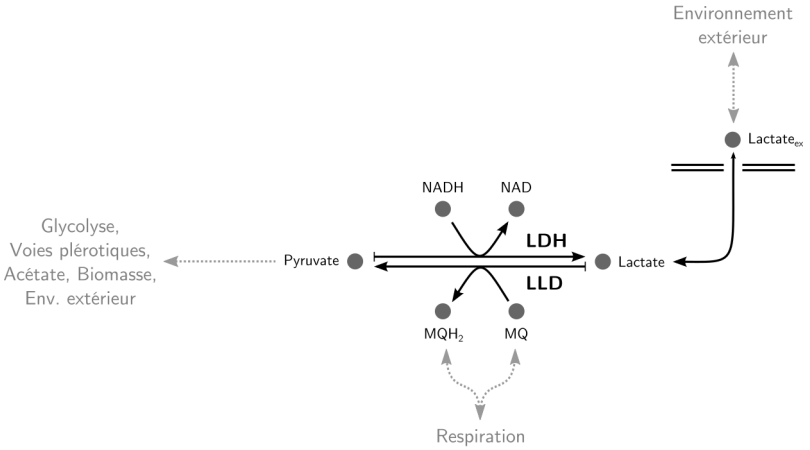
Glycolyse


Env. extérieur
Env. extérieur
Env. extérieur
Fructose$_{ex}$
Glucose$_{ex}$
Glucose$_{ex}$
P-enolpyruvate
Pyruvate
PTSFx
P-enolpyruvate
Pyruvate
PTSGx
IOLTx
GLK
Glucose-6-P
Glucose
ADP
ATP
GPI
Fructose-6-P
Voie des pentoses-P, Biomasse
ATP
Pi
PFKA
FBP
ADP
H_2O
PFKB
Fructose-1-P
Fructose-1,6-bP
ATP
ADP
Pi
FBA
TPI
Dihydroxyacétone-P
Glycéraldéhyde-3-P
Voie des pentoses-P, Biomasse
Pi
NAD
GAP
NADH
Glycérate-1,3-bP
ADP
PGK
ATP
Glycérate-3-P
Biomasse
GPM
Glycérate-2-P
H_2O
ENO
P-énolpyruvate
Voies plérotiques Biomasse
ADP
PK
ATP
Pyruvate
Voies plérotiques, Lactate, Acétate, Biomasse, Env. extérieur
Coenz. A
NAD
PDH
NADH
CO_2
Acétyl-Coa
Voies plérotiques, Acétate, TCA Biomasse

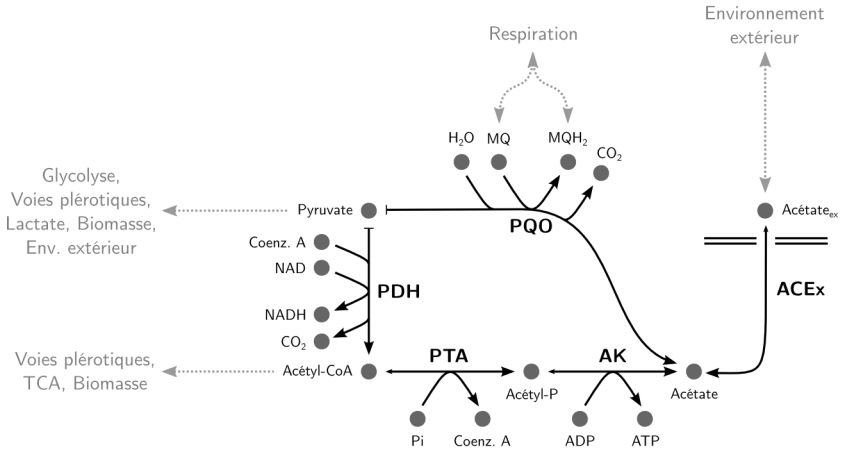
Voie des pentoses phosphates



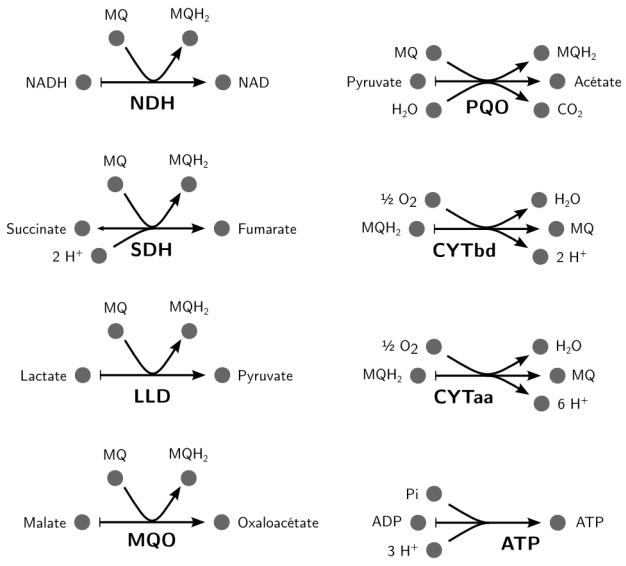
Métabolisme du lactate



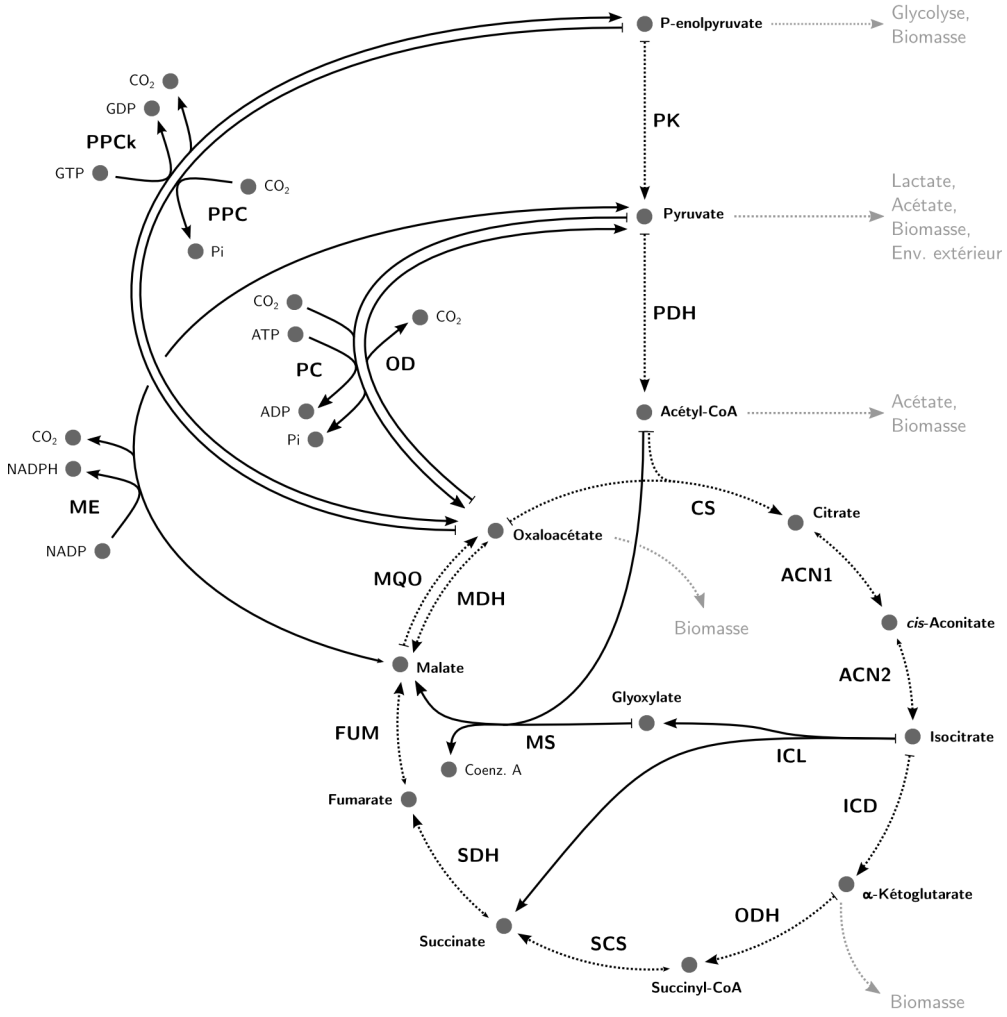
Métabolisme de l'acétate



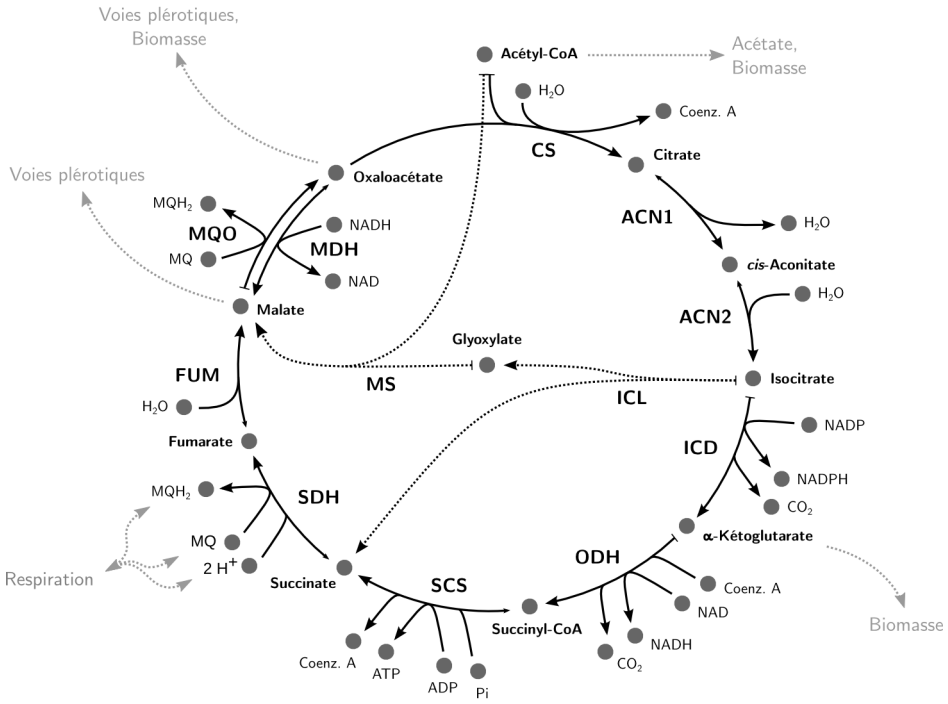
Réactions impliquées dans la respiration



Réactions « plérotiques »



Cycle de Krebs (TCA)



Références bibliographiques

[Almaas 2004] E. Almaas, B. Kovács, T. Vicsek, Z.N. Oltvai et A. Barabási. *Global organization of metabolic fluxes in the bacterium Escherichia coli.* Nature, vol. 427, no. 6977, pages 839–843, février 2004. *Cité en page 90*

[Antoniewicz 2006] M.R. Antoniewicz, J.K. Kelleher et G. Stephanopoulos. *Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements.* Metabolic Engineering, vol. 8, no. 4, pages 324–337, juillet 2006. *Cité en pages 115 et 132*

[Antoniewicz 2013] M.R. Antoniewicz. *Dynamic metabolic flux analysis–tools for probing transient states of metabolic networks.* Current Opinion in Biotechnology, vol. 24, no. 6, pages 973–978, décembre 2013. *Cité en page 116*

[Antoniewicz 2015] M.R. Antoniewicz. *Methods and advances in metabolic flux analysis : a mini-review.* Journal of Industrial Microbiology and Biotechnology, pages 1–9, janvier 2015. Publication anticipée en ligne. doi : 10.1007/s10295-015-1585-x. *Cité en page 114*

[Aoki 2005] R. Aoki, M. Wada, N. Takesue, K. Tanaka et A. Yokota. *Enhanced glutamic acid production by a H+-ATPase-defective mutant of Corynebacterium glutamicum.* Bioscience, Biotechnology, and Biochemistry, vol. 69, no. 8, pages 1466–1472, août 2005. *Cité en page 257*

[Bansal-Mutalik 2011] R. Bansal-Mutalik et H. Nikaido. *Quantitative lipid composition of cell envelopes of Corynebacterium glutamicum elucidated through reverse micelle extraction.* Proceedings of the National Academy of Sciences of the United States of America, vol. 108, no. 37, pages 15360–15365, septembre 2011. *Cité en page 258*

[Beard 2005] D.A. Beard et H. Qian. *Thermodynamic-based computational profiling of cellular regulatory control in hepatocyte metabolism.* American Journal of Physiology. Endocrinology and metabolism, vol. 288, no. 3, pages E633–E644, mars 2005. *Cité en page 92*

[Becker 2008] J. Becker, C. Klopprogge et C. Wittmann. *Metabolic responses to pyruvate kinase deletion in lysine producing Corynebacterium glutamicum.* Microbial Cell Factories, vol. 7, page 8, janvier 2008. *Cité en page 256*

[Bernot 2004] G. Bernot, J.P. Comet, A. Richard et J. Guespin. *Application of formal methods to biological regulatory networks : extending Thomas' asynchronous logical approach with temporal logic.* Journal of Theoretical Biology, vol. 229, no. 3, pages 339–347, août 2004. *Cité en page 70*

[Berthoumieux 2013] S. Berthoumieux, M. Brilli, D. Kahn, H. de Jong et E. Cinquemani. *On the identifiability of metabolic network models.* Journal of Mathematical Biology, vol. 67, no. 6-7, pages 1795–1832, décembre 2013. *Cité en page 76*

[Bordbar 2014] A. Bordbar, J.M. Monk, Z.A. King et B.O. Palsson. *Constraint-based models predict metabolic and associated cellular functions.* Nature reviews. Genetics, vol. 15, no. 2, pages 107–120, février 2014. *Cité en pages 90, 99 et 131*

[Borodina 2005] I. Borodina, P. Krabben et J. Nielsen. *Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism.* Genome Research, vol. 15, no. 6, pages 820–829, juin 2005. *Cité en pages 210 et 258*

[Bott 2003] M. Bott et A. Niebisch. *The respiratory chain of Corynebacterium glutamicum.* Journal of Biotechnology, vol. 104, no. 1-3, pages 129–153, septembre 2003. *Cité en pages 156, 256, 257 et 258*

[Bussmann 2009] M. Bussmann, D. Emer, S. Hasenbein, S. Degraf, B.J. Eikmanns et M. Bott. *Transcriptional control of the succinate dehydrogenase operon sdhCAB of Corynebacterium glutamicum by the cAMP-dependent regulator GlxR and the LuxR-type regulator RamA.* Journal of Biotechnology, vol. 143, no. 3, pages 173–182, septembre 2009. *Cité en page 256*

[Calik 2011] P. Calik, M. Sahin, H. Taspinar, E.S. Soyaslan et B. Inankur. *Dynamic flux balance analysis for pharmaceutical protein production by Pichia pastoris : human growth hormone.* Enzyme and Microbial Technology, vol. 48, no. 3, pages 209–216, mars 2011. *Cité en page 120*

[Cao 2014] Y. Cao, Z. Duan et Z. Shi. *Effect of biotin on transcription levels of key enzymes and glutamate efflux in glutamate fermentation by Corynebacterium glutamicum.* World Journal of Microbio-

logy and Biotechnology, vol. 30, no. 2, pages 461–468, février 2014. *Cité en page 257*

[Chaouiya 2007] C. Chaouiya. *Petri net modelling of biological networks.* Briefings in Bioinformatics, vol. 8, no. 4, pages 210–219, juillet 2007. *Cité en page 78*

[Cho 2010] H.Y. Cho, S.G. Lee, J.E. Hyeon et S.O. Han. *Identification and characterization of a transcriptional regulator, SucR, that influences sucCD transcription in Corynebacterium glutamicum.* Biochemical and Biophysical Research Communications, vol. 401, no. 2, pages 300–305, octobre 2010. *Cité en page 256*

[Chubukov 2013] V. Chubukov, M. Uhr, L. Le Chat, R.J. Kleijn, M. Jules, H. Link, S. Aymerich, J. Stelling et U. Sauer. *Transcriptional regulation is insufficient to explain substrate-induced flux changes in Bacillus subtilis.* Molecular Systems Biology, vol. 9, no. 709, page 709, janvier 2013. *Cité en page 57*

[Chubukov 2014] V. Chubukov, L. Gerosa, K. Kochanowski et U. Sauer. *Coordination of microbial metabolism.* Nature reviews. Microbiology, vol. 12, no. 5, pages 327–340, mai 2014. *Cité en page 56*

[Cocaign-Bousquet 1996] M. Cocaign-Bousquet, A. Guyonvarch et N.D. Lindley. *Growth rate-dependent modulation of carbon flux through central metabolism and the kinetic consequences for glucose-limited chemostat cultures of Corynebacterium glutamicum.* Applied and Environmental Microbiology, vol. 62, no. 2, pages 429–436, février 1996. *Cité en pages 156, 158, 193, 257 et 258*

[Cocaign 1993] M. Cocaign, C. Monnet et N.D. Lindley. *Batch kinetics of Corynebacterium glutamicum during growth on various carbon substrates : use of substrate mixtures to localise metabolic bottlenecks.* Applied Microbiology and Biotechnology, vol. 40, no. 4, pages 526–530, décembre 1993. *Cité en page 256*

[Colijn 2009] C. Colijn, A. Brandes, J. Zucker, D.S. Lun, B. Weiner, M.R. Farhat, T.Y. Cheng, D.B. Moody, M. Murray et J.E. Galagan. *Interpreting expression data with metabolic flux models : predicting Mycobacterium tuberculosis mycolic acid production.* PLoS Computational Biology, vol. 5, no. 8, page e1000489, août 2009. *Cité en page 57*

[Covert 2001] M.W. Covert, C.H. Schilling et B. Palsson. *Regulation of gene expression in flux balance models of metabolism.* Journal of

Theoretical Biology, vol. 213, no. 1, pages 73–88, novembre 2001. *Cité en page 92*

[Covert 2002] M.W. Covert et B.O. Palsson. *Transcriptional regulation in constraints-based metabolic models of Escherichia coli.* The Journal of Biological Chemistry, vol. 277, no. 31, pages 28058–28064, août 2002. *Cité en page 90*

[Covert 2003] M.W. Covert et B.O. Palsson. *Constraints-based models : regulation of gene expression reduces the steady-state solution space.* Journal of Theoretical Biology, vol. 221, no. 3, pages 309–325, avril 2003. *Cité en page 58*

[Covert 2008] M.W. Covert, N. Xiao, T.J. Chen et J.R. Karr. *Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli.* Bioinformatics, vol. 24, no. 18, pages 2044–2050, septembre 2008. *Cité en page 91*

[Coze 2013] F. Coze, F. Gilard, G. Tcherkez, M.J. Virolle et A. Guyonvarch. *Carbon-flux distribution within Streptomyces coelicolor metabolism : a comparison between the actinorhodin-producing strain M145 and its non-producing derivative M1146.* PloS ONE, vol. 8, no. 12, page e84151, janvier 2013. *Cité en pages 90, 115, 210 et 258*

[Cozzone 2005] A.J. Cozzone et M. El-Mansi. *Control of isocitrate dehydrogenase catalytic activity by protein phosphorylation in Escherichia coli.* Journal of Molecular Microbiology and Biotechnology, vol. 9, no. 3-4, pages 132–146, janvier 2005. *Cité en page 92*

[de Jong 2002] H. de Jong. *Modeling and simulation of genetic regulatory systems : a literature review.* Journal of Computational Biology, vol. 9, no. 1, pages 67–103, janvier 2002. *Cité en page 66*

[de Jong 2004] H. de Jong, J.L. Gouzé, C. Hernandez, M. Page, T. Sari et J. Geiselmann. *Qualitative simulation of genetic regulatory networks using piecewise-linear models.* Bulletin of Mathematical Biology, vol. 66, no. 2, pages 301–340, mars 2004. *Cité en page 74*

[de Oliveira Dal'Molin 2013] C.G. de Oliveira Dal'Molin et L.K. Nielsen. *Plant genome-scale metabolic reconstruction and modelling.* Current Opinion in Biotechnology, vol. 24, no. 2, pages 271–277, avril 2013. *Cité en page 90*

[De Vos 2011] D. De Vos, F.J. Bruggeman, H.V. Westerhoff et B.M. Bakker. *How molecular competition influences fluxes in gene expression networks.* PloS ONE, vol. 6, no. 12, page e28494, janvier 2011. *Cité en page 33*

[D'Huys 2012] P.J. D'Huys, I. Lule, D. Vercammen, J. Anné, J.F. Van Impe et K. Bernaerts. *Genome-scale metabolic flux analysis of Streptomyces lividans growing on a complex medium.* Journal of Biotechnology, vol. 161, no. 1, pages 1–13, septembre 2012. *Cité en pages 112, 113 et 133*

[Dietrich 2009] C. Dietrich, A. Nato, B. Bost, P. Le Maréchal et A. Guyonvarch. *Regulation of ldh expression during biotin-limited growth of Corynebacterium glutamicum.* Microbiology, vol. 155, no. 4, pages 1360–1375, avril 2009. *Cité en pages 144, 148, 159 et 257*

[Dominguez 1998] H. Dominguez, C. Rollin, A. Guyonvarch, J.L. Guerquin-Kern, M. Cocaign-Bousquet et N.D. Lindley. *Carbon-flux distribution in the central metabolic pathways of Corynebacterium glutamicum during growth on fructose.* European Journal of Biochemistry, vol. 254, no. 1, pages 96–102, mai 1998. *Cité en page 200*

[Durot 2009] M. Durot, P.Y. Bourguignon et V. Schachter. *Genome-scale models of bacterial metabolism : reconstruction and applications.* FEMS Microbiology Reviews, vol. 33, no. 1, pages 164–190, janvier 2009. *Cité en page 100*

[Ebbighausen 1991] H. Ebbighausen, B. Weil et R. Krämer. *Carrier-mediated acetate uptake in Corynebacterium glutamicum.* Archives of Microbiology, vol. 155, no. 5, pages 505–510, avril 1991. *Cité en page 258*

[Edwards 2001] J.S. Edwards, R.U. Ibarra et B.O. Palsson. *In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.* Nature Biotechnology, vol. 19, no. 2, pages 125–130, février 2001. *Cité en pages 104 et 106*

[Eikmanns 1992] B.J. Eikmanns. *Identification, sequence analysis, and expression of a Corynebacterium glutamicum gene cluster encoding the three glycolytic enzymes glyceraldehyde-3-phosphate dehydrogenase, 3-phosphoglycerate kinase, and triosephosphate isomerase.* Journal of Bacteriology, vol. 174, no. 19, pages 6076–6086, octobre 1992. *Cité en page 256*

[Eikmanns 1994] B.J. Eikmanns, N. Thum-Schmitz, L. Eggeling, K.U. Lüdtke et H. Sahm. *Nucleotide sequence, expression and transcriptional analysis of the Corynebacterium glutamicum gltA gene encoding citrate synthase.* Microbiology, vol. 140, no. 8, pages 1817–1828, août 1994. *Cité en page 256*

[Eikmanns 1995] B.J. Eikmanns, D. Rittmann et H. Sahm. *Cloning, sequence analysis, expression, and inactivation of the Corynebacterium glutamicum icd gene encoding isocitrate dehydrogenase and biochemical characterization of the enzyme.* Journal of Bacteriology, vol. 177, no. 3, pages 774–782, février 1995. *Cité en page 256*

[Famili 2003] I. Famili, J. Forster, J. Nielsen et B.O. Palsson. *Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network.* Proceedings of the National Academy of Sciences of the United States of America, vol. 100, no. 23, pages 13134–13139, novembre 2003. *Cité en page 104*

[Fuchs 2011] G. Fuchs. *Alternative pathways of carbon dioxide fixation : insights into the early evolution of life ?* Annual Review of Microbiology, vol. 65, pages 631–658, janvier 2011. *Cité en page 191*

[Gebert 2007] J. Gebert, N. Radde et G.W. Weber. *Modeling gene regulatory networks with piecewise linear differential equations*. European Journal of Operational Research, vol. 181, no. 3, pages 1148–1165, septembre 2007. *Cité en page 74*

[Genda 2003] T. Genda, T. Nakamatsu et H. Ozak. *Purification and characterization of malate dehydrogenase from Corynebacterium glutamicum.* Journal of Bioscience and Bioengineering, vol. 95, no. 6, pages 562–566, janvier 2003. *Cité en page 257*

[Gerstmeir 2003] R. Gerstmeir, V.F. Wendisch, S. Schnicke, H. Ruan, M. Farwick, D. Reinscheid et B.J. Eikmanns. *Acetate metabolism and its regulation in Corynebacterium glutamicum.* Journal of Biotechnology, vol. 104, no. 1-3, pages 99–122, septembre 2003. *Cité en page 257*

[Gianchandani 2009] E.P. Gianchandani, A.R. Joyce, B.O. Palsson et J.A. Papin. *Functional states of the genome-scale Escherichia coli transcriptional regulatory system.* PLoS Computational Biology, vol. 5, no. 6, page e1000403, juin 2009. *Cité en page 82*

[Gourdon 2000] P. Gourdon, M.F. Baucher, N.D. Lindley et A. Guyonvarch. *Cloning of the malic enzyme gene from Corynebacterium glutamicum and role of the enzyme in lactate metabolism.* Applied and environmental microbiology, vol. 66, no. 7, pages 2981–7, juillet 2000. *Cité en page 257*

[Hasegawa 2008] T. Hasegawa, K.I. Hashimoto, H. Kawasaki et T. Nakamatsu. *Changes in enzyme activities at the pyruvate node in*

glutamate-overproducing Corynebacterium glutamicum. Journal of Bioscience and Bioengineering, vol. 105, no. 1, pages 12–19, janvier 2008. *Cité en page 257*

[Hoffelder 2010] M. Hoffelder, K. Raasch, J. van Ooyen et L. Eggeling. *The E2 domain of OdhA of Corynebacterium glutamicum has succinyltransferase activity dependent on lipoyl residues of the acetyltransferase AceF.* Journal of Bacteriology, vol. 192, no. 19, pages 5203–5211, octobre 2010. *Cité en page 256*

[Ibarra 2002] R.U. Ibarra, J.S. Edwards et B.O. Palsson. *Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.* Nature, vol. 420, no. 6912, pages 186–189, novembre 2002. *Cité en page 106*

[Ikeda 2011] M. Ikeda, Y. Mizuno, S.i. Awane, M. Hayashi, S. Mitsuhashi et S. Takeno. *Identification and application of a different glucose uptake system that functions as an alternative to the phosphotransferase system in Corynebacterium glutamicum.* Applied Microbiology and Biotechnology, vol. 90, no. 4, pages 1443–1451, mai 2011. *Cité en page 258*

[Ikeda 2012] M. Ikeda. *Sugar transport systems in Corynebacterium glutamicum : features and applications to strain development.* Applied Microbiology and Biotechnology, vol. 96, no. 5, pages 1191–1200, décembre 2012. *Cité en pages 158, 257 et 258*

[Inui 2007] M. Inui, M. Suda, S. Okino, H. Nonaka, L.G. Puskás, A.a. Vertès et H. Yukawa. *Transcriptional profiling of Corynebacterium glutamicum metabolism during organic acid production under oxygen deprivation conditions.* Microbiology, vol. 153, no. 8, pages 2491–2504, août 2007. *Cité en page 256*

[Jacob 1961] F. Jacob et J. Monod. *Genetic regulatory mechanisms in the synthesis of proteins.* Journal of Molecular Biology, vol. 3, pages 318–356, juin 1961. *Cité en page 56*

[Jäger 1996] W. Jäger, P.G. Peters-Wendisch, J. Kalinowski et A. Pühler. *A Corynebacterium glutamicum gene encoding a two-domain protein similar to biotin carboxylases and biotin-carboxyl-carrier proteins.* Archives of Microbiology, vol. 166, no. 2, pages 76–82, août 1996. *Cité en page 193*

[Jetten 1994a] M.S. Jetten, M.E. Gubler, S.H. Lee et A.J. Sinskey. *Structural and functional analysis of pyruvate kinase from Corynebacte-*

rium glutamicum. Applied and Environmental Microbiology, vol. 60, no. 7, pages 2501–2507, juillet 1994. *Cité en page 256*

[Jetten 1994b] M.S.M. Jetten, G.A. Pitoc, M.T. Follettie et A.J. Sinskey. *Regulation of phospho(enol)-pyruvate-and oxaloacetate-converting enzymes in Corynebacterium glutamicum.* Applied Microbiology and Biotechnology, vol. 41, no. 1, pages 47–52, mars 1994. *Cité en pages 256 et 257*

[Jetten 1995] M.S. Jetten et A.J. Sinskey. *Purification and properties of oxaloacetate decarboxylase from Corynebacterium glutamicum.* Antonie van Leeuwenhoek, vol. 67, no. 2, pages 221–227, janvier 1995. *Cité en page 257*

[Jolkver 2009] E. Jolkver, D. Emer, S. Ballan, R. Krämer, B.J. Eikmanns et K. Marin. *Identification and characterization of a bacterial transport system for the uptake of pyruvate, propionate, and acetate in Corynebacterium glutamicum.* Journal of Bacteriology, vol. 191, no. 3, pages 940–948, février 2009. *Cité en page 258*

[Karlebach 2008] G. Karlebach et R. Shamir. *Modelling and analysis of gene regulatory networks.* Nature reviews. Molecular Cell Biology, vol. 9, no. 10, pages 770–780, octobre 2008. *Cité en page 66*

[Karp 2000] P.D. Karp. *An ontology for biological function based on molecular interactions.* Bioinformatics, vol. 16, no. 3, pages 269–285, mars 2000. *Cité en page 9*

[Karp 2002] P.D. Karp, S. Paley et P. Romero. *The Pathway Tools software.* Bioinformatics, vol. 18 Suppl 1, pages S225–S232, janvier 2002. *Cité en page 243*

[Karp 2010] P.D. Karp, S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, T.J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I.M. Keseler et R. Caspi. *Pathway Tools version 13.0 : integrated software for pathway/genome informatics and systems biology.* Briefings in Bioinformatics, vol. 11, no. 1, pages 40–79, janvier 2010. *Cité en page 156*

[Karr 2012] J.R. Karr, J.C. Sanghvi, D.N. Macklin, M.V. Gutschow, J.M. Jacobs, B. Bolival, N. Assad-Garcia, J.I. Glass et M.W. Covert. *A whole-cell computational model predicts phenotype from genotype.* Cell, vol. 150, no. 2, pages 389–401, juillet 2012. *Cité en pages 3 et 91*

[Kauffman 1969] S.A. Kauffman. *Metabolic stability and epigenesis in ran-*

domly constructed genetic nets. Journal of Theoretical Biology, vol. 22, no. 3, pages 437–467, mars 1969. *Cité en page 66*

[Kauffman 2003] K.J. Kauffman, P. Prakash et J.S. Edwards. *Advances in flux balance analysis.* Current Opinion in Biotechnology, vol. 14, no. 5, pages 491–496, octobre 2003. *Cité en pages 103 et 132*

[Keseler 2013] I.M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A.M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz Rascado, Q. Ong, S. Paley, I. Schröder, A.G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R.P. Gunsalus, I. Paulsen et P.D. Karp. *EcoCyc : fusing model organism databases with systems biology.* Nucleic Acids Research, vol. 41, no. Database issue, pages D605–D612, janvier 2013. *Cité en page 54*

[Kim 2012] T.Y. Kim, S.B. Sohn, Y.B. Kim, W.J. Kim et S.Y. Lee. *Recent advances in reconstruction and applications of genome-scale metabolic models.* Current Opinion in Biotechnology, vol. 23, no. 4, pages 617–623, août 2012. *Cité en page 90*

[Kimura 2005] E. Kimura. *L-Glutamate production.* In L. Eggeling et M. Bott, editeurs, Handbook of *Corynebacterium glutamicum*, pages 439–457. CRC Press, USA, mars 2005. *Cité en pages 153 et 194*

[Kinoshita 1957] S. Kinoshita, S. Udaka et M. Shimono. *Studies on the amino acid fermentation. Production of L-glutamic acid by various microorganisms.* The Journal of General and Applied Microbiology, vol. 3, no. 3, pages 193–205, juin 1957. *Cité en page 193*

[Kjeldsen 2009] K.R. Kjeldsen et J. Nielsen. *In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network.* Biotechnology and Bioengineering, vol. 102, no. 2, pages 583–597, février 2009. *Cité en page 258*

[Klaffl 2010] S. Klaffl et B.J. Eikmanns. *Genetic and functional analysis of the soluble oxaloacetate decarboxylase from Corynebacterium glutamicum.* Journal of Bacteriology, vol. 192, no. 10, pages 2604–2612, mai 2010. *Cité en page 257*

[Klapa 2003] M.I. Klapa, J.C. Aon et G. Stephanopoulos. *Systematic quantification of complex metabolic flux networks using stable isotopes and mass spectrometry.* European Journal of Biochemistry, vol. 270, no. 17, pages 3525–3542, septembre 2003. *Cité en pages 152 et 258*

[Koshland 1956] D.E. Koshland. *Molecular geometry in enzyme action.*

Journal of Cellular Physiology, vol. 47, no. Suppl 1, pages 217–234, mai 1956. *Cité en pages 24 et 28*

[Kurokawa 2005] T. Kurokawa et J. Sakamoto. *Purification and characterization of succinate :menaquinone oxidoreductase from Corynebacterium glutamicum.* Archives of Microbiology, vol. 183, no. 5, pages 317–324, août 2005. *Cité en page 256*

[Latendresse 2010] M. Latendresse et P.D. Karp. *An advanced web query interface for biological databases.* Database (Oxford), vol. 2010, page baq006, janvier 2010. *Cité en page 243*

[Lee 2008] J.M. Lee, J. Min Lee, E.P. Gianchandani, J.A. Eddy et J.A. Papin. *Dynamic analysis of integrated signaling, metabolic, and regulatory networks.* PLoS Computational Biology, vol. 4, no. 5, page e1000086, mai 2008. *Cité en pages 58, 91 et 126*

[Leighty 2011] R.W. Leighty et M.R. Antoniewicz. *Dynamic metabolic flux analysis (DMFA) : a framework for determining fluxes at metabolic non-steady state.* Metabolic Engineering, vol. 13, no. 6, pages 745–755, novembre 2011. *Cité en page 133*

[Lequeux 2010] G. Lequeux, J. Beauprez, J. Maertens, E. Van Horen, W. Soetaert, E. Vandamme et P.A. Vanrolleghem. *Dynamic metabolic flux analysis demonstrated on cultures where the limiting substrate is changed from carbon to nitrogen and vice versa.* Journal of Biomedicine and Biotechnology, vol. 2010, août 2010. *Cité en page 137*

[Lewis 2012] N.E. Lewis, H. Nagarajan et B.O. Palsson. *Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.* Nature reviews. Microbiology, vol. 10, no. 4, pages 291–305, avril 2012. *Cité en pages 99, 101, 225 et 226*

[Lindner 2011] S.N. Lindner, G.M. Seibold, A. Henrich, R. Krämer et V.F. Wendisch. *Phosphotransferase system-independent glucose utilization in Corynebacterium glutamicum by inositol permeases and glucokinases.* Applied and Environmental Microbiology, vol. 77, no. 11, pages 3571–3581, juin 2011. *Cité en pages 158, 257 et 258*

[Llaneras 2007a] F. Llaneras et J. Picó. *An interval approach for dealing with flux distributions and elementary modes activity patterns.* Journal of Theoretical Biology, vol. 246, no. 2, pages 290–308, mai 2007. *Cité en page 108*

[Llaneras 2007b] F. Llaneras et J. Picó. *A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are*

uncertain and/or insufficient. BMC Bioinformatics, vol. 8, page 421, janvier 2007. *Cité en pages 120 et 133*

[Llaneras 2009] F. Llaneras, A. Sala et J. Picó. *A possibilistic framework for constraint-based metabolic flux analysis.* BMC Systems Biology, vol. 3, page 79, juillet 2009. *Cité en pages 115 et 132*

[Llaneras 2012] F. Llaneras, A. Sala et J. Picó. *Dynamic estimations of metabolic fluxes with constraint-based models and possibility theory.* Journal of Process Control, vol. 22, no. 10, pages 1946–1955, décembre 2012. *Cité en pages 126 et 133*

[Luo 2006] R.Y. Luo, S. Liao, G.Y. Tao, Y.Y. Li, S. Zeng, Y.X. Li et Q. Luo. *Dynamic analysis of optimality in myocardial energy metabolism under normal and ischemic conditions.* Molecular Systems Biology, vol. 2, no. 1, janvier 2006. *Cité en page 126*

[Machado 2011] D. Machado, R.S. Costa, M. Rocha, E.C. Ferreira, B. Tidor et I. Rocha. *Modeling formalisms in systems biology.* AMB Express, vol. 1, no. 1, page 45, janvier 2011. *Cité en pages 57, 59 et 65*

[Mahadevan 2002] R. Mahadevan, J.S. Edwards et F.J. Doyle. *Dynamic flux balance analysis of diauxic growth in Escherichia coli.* Biophysical Journal, vol. 83, no. 3, pages 1331–1340, septembre 2002. *Cité en pages 124 et 126*

[Mahadevan 2003] R. Mahadevan et C. Schilling. *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metabolic Engineering, vol. 5, no. 4, pages 264–276, octobre 2003. *Cité en pages 107 et 132*

[Mashego 2006] M.R. Mashego, W.M. van Gulik, J.L. Vinke, D. Visser et J.J. Heijnen. *In vivo kinetics with rapid perturbation experiments in Saccharomyces cerevisiae using a second-generation BioScope.* Metabolic Engineering, vol. 8, no. 4, pages 370–383, juillet 2006. *Cité en page 92*

[McAdams 1998] H.H. McAdams et A. Arkin. *Simulation of prokaryotic genetic circuits.* Annual Review of Biophysics and Biomolecular Structure, vol. 27, pages 199–224, janvier 1998. *Cité en page 93*

[Mitsuhashi 2006] S. Mitsuhashi, M. Hayashi, J. Ohnishi et M. Ikeda. *Disruption of malate :quinone oxidoreductase increases L-lysine production by Corynebacterium glutamicum.* Bioscience, Biotechnology, and Biochemistry, vol. 70, no. 11, pages 2803–2806, novembre 2006. *Cité en page 257*

[Molenaar 1998] D. Molenaar, M.E. van der Rest et S. Petrović. *Biochemical and genetic characterization of the membrane-associated malate dehydrogenase (acceptor) from Corynebacterium glutamicum.* European Journal of Biochemistry, vol. 254, no. 2, pages 395–403, juin 1998. *Cité en page 257*

[Molenaar 2000] D. Molenaar, M.E. van der Rest, A. Drysch et R. Yücel. *Functions of the membrane-associated and cytoplasmic malate dehydrogenases in the citric acid cycle of Corynebacterium glutamicum.* Journal of Bacteriology, vol. 182, no. 24, pages 6884–6891, décembre 2000. *Cité en page 257*

[Mulder 1986] M.M. Mulder, M.J. Teixeira de Mattos, P.W. Postma et K. van Dam. *Energetic consequences of multiple K+ uptake systems in Escherichia coli.* Biochimica et Biophysica Acta, vol. 851, no. 2, pages 223–228, septembre 1986. *Cité en pages 210 et 258*

[Nantapong 2005] N. Nantapong, A. Otofuji, C.T. Migita, O. Adachi, H. Toyama et K. Matsushita. *Electron transfer ability from NADH to menaquinone and from NADPH to oxygen of type II NADH dehydrogenase of Corynebacterium glutamicum.* Bioscience, Biotechnology, and Biochemistry, vol. 69, no. 1, pages 149–159, janvier 2005. *Cité en page 258*

[Neidhardt 1990] F.C. Neidhardt, J.L. Ingraham et M. Schaechter. Physiology of the bacterial cell : a molecular approach. Sinauer Associates Inc, USA, 1990. *Cité en pages 89 et 158*

[Netzer 2004] R. Netzer, M. Krause, D. Rittmann, P.G. Peters-Wendisch, L. Eggeling, V.F. Wendisch et H. Sahm. *Roles of pyruvate kinase and malic enzyme in Corynebacterium glutamicum for growth on carbon sources requiring gluconeogenesis.* Archives of Microbiology, vol. 182, no. 5, pages 354–363, novembre 2004. *Cité en pages 256 et 257*

[Niklas 2011] J. Niklas, E. Schräder, V. Sandig, T. Noll et E. Heinzle. *Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line AGE1.HN using time resolved metabolic flux analysis.* Bioprocess and Biosystems Engineering, vol. 34, no. 5, pages 533–545, juin 2011. *Cité en page 120*

[Noor 2012] E. Noor, A. Bar-Even, A. Flamholz, Y. Lubling, D. Davidi et R. Milo. *An integrated open framework for thermodynamics of reactions that combines accuracy and coverage.* Bioinformatics, vol. 28, no. 15, pages 2037–2044, août 2012. *Cité en pages 156 et 255*

[Nyström 2004] T. Nyström. *Growth versus maintenance : a trade-off dictated by RNA polymerase availability and sigma factor competition ?* Molecular Microbiology, vol. 54, no. 4, pages 855–862, novembre 2004. *Cité en page 33*

[Orth 2010] J.D. Orth, I. Thiele et B.O. Palsson. *What is flux balance analysis ?* Nature Biotechnology, vol. 28, no. 3, pages 245–248, mars 2010. *Cité en page 103*

[Ozcan 2007] N. Ozcan, C.S. Ejsing, A. Shevchenko, A. Lipski, S. Morbach et R. Krämer. *Osmolality, temperature, and membrane lipid composition modulate the activity of betaine transporter BetP in Corynebacterium glutamicum.* Journal of Bacteriology, vol. 189, no. 20, pages 7485–7496, octobre 2007. *Cité en page 258*

[Palsson 2006] B.O. Palsson. Systems biology : properties of reconstructed networks. Cambridge university press, USA, 2006. *Cité en pages 103 et 110*

[Parche 2001] S. Parche, A. Burkovski, G.A. Sprenger, B. Weil, R. Krämer et F. Titgemeyer. *Corynebacterium glutamicum : a dissection of the PTS.* Journal of Molecular Microbiology and Biotechnology, vol. 3, no. 3, pages 423–428, juillet 2001. *Cité en pages 158 et 257*

[Park 2000] S.Y. Park, H.K. Kim, S.K. Yoo, T.K. Oh et J.K. Lee. *Characterization of glk, a gene coding for glucose kinase of Corynebacterium glutamicum.* FEMS Microbiology Letters, vol. 188, no. 2, pages 209–215, juillet 2000. *Cité en page 256*

[Patti 2012] G.J. Patti, O. Yanes et G. Siuzdak. *Metabolomics : the apogee of the omics trilogy.* Nature reviews. Molecular cell biology, vol. 13, no. 4, pages 263–269, avril 2012. *Cité en page 57*

[Peters-Wendisch 1997] P.G. Peters-Wendisch, V.F. Wendisch, S. Paul, B.J. Eikmanns et H. Sahm. *Pyruvate carboxylase as an anaplerotic enzyme in Corynebacterium glutamicum.* Microbiology, vol. 143, no. 4, pages 1095–1103, avril 1997. *Cité en page 257*

[Peters-Wendisch 1998] P.G. Peters-Wendisch, C. Kreutzer, J. Kalinowski, M. Pátek, H. Sahm et B.J. Eikmanns. *Pyruvate carboxylase from Corynebacterium glutamicum : characterization, expression and inactivation of the pyc gene.* Microbiology, vol. 144, no. 4, pages 915–927, avril 1998. *Cité en pages 158, 193 et 257*

[Peters-Wendisch 2001] P.G. Peters-Wendisch, B. Schiel, V.F. Wendisch, E. Katsoulidis, B. Möckel, H. Sahm et B.J. Eikmanns. *Pyruvate*

carboxylase is a major bottleneck for glutamate and lysine production by Corynebacterium glutamicum. Journal of Molecular Microbiology and Biotechnology, vol. 3, no. 2, pages 295–300, avril 2001. *Cité en page 257*

[Petersen 2000] S. Petersen, a.a. de Graaf, L. Eggeling, M. Möllney, W. Wiechert et H. Sahm. *In vivo quantification of parallel and bidirectional fluxes in the anaplerosis of Corynebacterium glutamicum.* The Journal of Biological Chemistry, vol. 275, no. 46, pages 35932–35941, novembre 2000. *Cité en pages 152, 213, 256, 257 et 258*

[Petersen 2001] S. Petersen, C. Mack, a.a. de Graaf, C. Riedel, B.J. Eikmanns et H. Sahm. *Metabolic consequences of altered phosphoenolpyruvate carboxykinase activity in Corynebacterium glutamicum reveal anaplerotic regulation mechanisms in vivo.* Metabolic Engineering, vol. 3, no. 4, pages 344–361, octobre 2001. *Cité en pages 256 et 257*

[Petri 1962] C.A. Petri. *Kommunikation mit Automaten.* Thèse de doctorat, Fakultät für Mathematik und Physik, Bonn, 1962. *Cité en page 78*

[Plummer 2006] M. Plummer, N. Best, K. Cowles et K. Vines. *CODA : convergence diagnosis and output analysis for MCMC.* R News, vol. 6, no. 1, pages 7–11, 2006. *Cité en pages 172 et 176*

[Price 2004a] N.D. Price, J.L. Reed et B.O. Palsson. *Genome-scale models of microbial cells : evaluating the consequences of constraints.* Nature Reviews. Microbiology, vol. 2, no. 11, pages 886–897, novembre 2004. *Cité en page 90*

[Price 2004b] N.D. Price, J. Schellenberger et B.O. Palsson. *Uniform sampling of steady-state flux spaces : means to design experiments and to interpret enzymopathies.* Biophysical Journal, vol. 87, no. 4, pages 2172–2186, octobre 2004. *Cité en pages 110, 113, 133 et 151*

[R Core Team 2014] R Core Team. *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. *Cité en page 136*

[Ramakrishna 2001] R. Ramakrishna, J.S. Edwards, A. McCulloch et B.O. Palsson. *Flux-balance analysis of mitochondrial energy metabolism : consequences of systemic stoichiometric constraints.* American Journal of Physiology. Regulatory, integrative and comparative physiology, vol. 280, no. 3, pages R695–R704, mars 2001. *Cité en page 104*

[Raman 2009] K. Raman et N. Chandra. *Flux balance analysis of biological systems : applications and challenges.* Briefings in Bioinformatics, vol. 10, no. 4, pages 435–449, juillet 2009. *Cité en page 103*

[Reddy 2014] G.K. Reddy et V.F. Wendisch. *Characterization of 3-phosphoglycerate kinase from Corynebacterium glutamicum and its impact on amino acid production.* BMC Microbiology, vol. 14, no. 1, page 54, mars 2014. *Cité en page 256*

[Reed 2004] J.L. Reed et B.O. Palsson. *Genome-scale in silico models of E. coli have multiple equivalent phenotypic states : assessment of correlated reaction subsets that comprise network states.* Genome Research, vol. 14, no. 9, pages 1797–1805, septembre 2004. *Cité en page 107*

[Reinscheid 1994] D.J. Reinscheid, B.J. Eikmanns et H. Sahm. *Malate synthase from Corynebacterium glutamicum : sequence analysis of the gene and biochemical characterization of the enzyme.* Microbiology, vol. 140, no. 11, pages 3099–3108, novembre 1994. *Cité en page 257*

[Riedel 2001] C. Riedel, D. Rittmann, P. Dangel, B. Möckel, S. Petersen, H. Sahm et B.J. Eikmanns. *Characterization of the phosphoenolpyruvate carboxykinase gene from Corynebacterium glutamicum and significance of the enzyme for growth and amino acid production.* Journal of Molecular Microbiology and Biotechnology, vol. 3, no. 4, pages 573–583, octobre 2001. *Cité en page 257*

[Rittmann 2003] D. Rittmann, S. Schaffer, V.F. Wendisch et H. Sahm. *Fructose-1,6-bisphosphatase from Corynebacterium glutamicum : expression and deletion of the fbp gene and biochemical characterization of the enzyme.* Archives of Microbiology, vol. 180, no. 4, pages 285–292, octobre 2003. *Cité en pages 213, 220 et 256*

[Rollin 1995] C. Rollin, V. Morgant, A. Guyonvarch et J.L. Guerquin-Kern. *13C-NMR studies of Corynebacterium melassecola metabolic pathways.* European Journal of Biochemistry, vol. 227, no. 1-2, pages 488–493, janvier 1995. *Cité en page 152*

[Russell 1979] J.B. Russell et R.L. Baldwin. *Comparison of maintenance energy expenditures and growth yields among several rumen bacteria grown on continuous culture.* Applied and Environmental Microbiology, vol. 37, no. 3, pages 537–543, mars 1979. *Cité en pages 210 et 258*

[Sainz 2003] J. Sainz, F. Pizarro, J.R. Pérez-Correa et E. Agosin. *Modeling of yeast metabolism and process dynamics in batch fermentation.* Biotechnology and Bioengineering, vol. 81, no. 7, pages 818–828, mars 2003. *Cité en page 126*

[Sato 2008] H. Sato, K. Orishimo, T. Shirai, T. Hirasawa, K. Nagahisa, H. Shimizu et M. Wachi. *Distinct roles of two anaplerotic pathways in glutamate production induced by biotin limitation in Corynebacterium glutamicum.* Journal of Bioscience and Bioengineering, vol. 106, no. 1, pages 51–58, juillet 2008. *Cité en page 257*

[Sauer 2005] U. Sauer et B.J. Eikmanns. *The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria.* FEMS Microbiology Reviews, vol. 29, no. 4, pages 765–794, septembre 2005. *Cité en pages 256 et 257*

[Sawada 2010] K. Sawada, S. Zen-in, M. Wada et A. Yokota. *Metabolic changes in a pyruvate kinase gene deletion mutant of Corynebacterium glutamicum ATCC 13032.* Metabolic Engineering, vol. 12, no. 4, pages 401–407, juillet 2010. *Cité en pages 256 et 257*

[Schellenberger 2009] J. Schellenberger et B.O. Palsson. *Use of randomized sampling for analysis of metabolic networks.* The Journal of Biological Chemistry, vol. 284, no. 9, pages 5457–5461, février 2009. *Cité en pages 110 et 161*

[Schilling 2002] C.H. Schilling, M.W. Covert, I. Famili, G.M. Church, J.S. Edwards et B.O. Palsson. *Genome-scale metabolic model of Helicobacter pylori 26695.* Journal of Bacteriology, vol. 184, no. 16, pages 4582–4593, août 2002. *Cité en page 106*

[Schreiner 2005a] M.E. Schreiner et B.J. Eikmanns. *Pyruvate :quinone oxidoreductase from Corynebacterium glutamicum : purification and biochemical characterization.* Journal of Bacteriology, vol. 187, no. 3, pages 862–871, février 2005. *Cité en page 257*

[Schreiner 2005b] M.E. Schreiner, D. Fiur, J. Holátko, M. Pátek et B.J. Eikmanns. *E1 enzyme of the pyruvate dehydrogenase complex in Corynebacterium glutamicum : molecular analysis of the gene and phylogenetic aspects.* Journal of Bacteriology, vol. 187, no. 17, pages 6005–6018, septembre 2005. *Cité en page 256*

[Schreiner 2006] M.E. Schreiner, C. Riedel, J. Holátko, M. Pátek et B.J. Eikmanns. *Pyruvate :quinone oxidoreductase in Corynebacterium glutamicum : molecular analysis of the pqo gene, significance of the*

enzyme, and phylogenetic aspects. Journal of Bacteriology, vol. 188, no. 4, pages 1341–1350, février 2006. *Cité en page 257*

[Schuetz 2007] R. Schuetz, L. Kuepfer et U. Sauer. *Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.* Molecular Systems Biology, vol. 3, no. 119, page 119, janvier 2007. *Cité en page 106*

[Schuetz 2012] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann et U. Sauer. *Multidimensional optimality of microbial metabolism.* Science, vol. 336, no. 6081, pages 601–604, mai 2012. *Cité en page 106*

[Schuster 2008] S. Schuster, T. Pfeiffer et D.a. Fell. *Is maximization of molar yield in metabolic networks favoured by evolution ?* Journal of Theoretical Biology, vol. 252, no. 3, pages 497–504, juin 2008. *Cité en page 106*

[Segal 2003] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller et N. Friedman. *Module networks : identifying regulatory modules and their condition-specific regulators from gene expression data.* Nature Genetics, vol. 34, no. 2, pages 166–176, juin 2003. *Cité en page 60*

[Segrè 2002] D. Segrè, D. Vitkup et G.M. Church. *Analysis of optimality in natural and perturbed metabolic networks.* Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 23, pages 15112–15117, novembre 2002. *Cité en page 104*

[Shimizu 2003] H. Shimizu, H. Tanaka, A. Nakato, K. Nagahisa, E. Kimura et S. Shioya. *Effects of the changes in enzyme activities on metabolic flux redistribution around the 2-oxoglutarate branch in glutamate production by Corynebacterium glutamicum.* Bioprocess and Biosystems Engineering, vol. 25, no. 5, pages 291–298, mars 2003. *Cité en pages 158 et 258*

[Shinfuku 2009] Y. Shinfuku, N. Sorpitiporn, M. Sono, C. Furusawa, T. Hirasawa et H. Shimizu. *Development and experimental verification of a genome-scale metabolic model for Corynebacterium glutamicum.* Microbial Cell Factories, vol. 8, page 43, janvier 2009. *Cité en pages 106 et 258*

[Shirai 2005] T. Shirai, A. Nakato, N. Izutani, K. Nagahisa, S. Shioya, E. Kimura, Y. Kawarabayasi, A. Yamagishi, T. Gojobori et H. Shimizu. *Comparative study of flux redistribution of metabolic pathway in glutamate production by two coryneform bacte-*

ria. Metabolic Engineering, vol. 7, no. 2, pages 59–69, mars 2005. *Cité en pages 158 et 258*

[Shlomi 2005] T. Shlomi, O. Berkman et E. Ruppin. *Regulatory on/off minimization of metabolic flux changes after genetic perturbations.* Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 21, pages 7695–7700, mai 2005. *Cité en page 104*

[Simão 2005] E. Simão, E. Remy, D. Thieffry et C. Chaouiya. *Qualitative modelling of regulated metabolic pathways : application to the tryptophan biosynthesis in E.coli.* Bioinformatics, vol. 21 Suppl 2, no. 2, pages ii190–ii196, septembre 2005. *Cité en page 58*

[Smallbone 2010] K. Smallbone, E. Simeonidis, N. Swainston et P. Mendes. *Towards a genome-scale kinetic model of cellular metabolism.* BMC Systems Biology, vol. 4, page 6, janvier 2010. *Cité en page 92*

[Smith 1984] R.L. Smith. *Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions.* Operations Research, vol. 32, no. 6, pages 1296–1308, 1984. *Cité en page 164*

[Snoussi 1989] E.H. Snoussi. *Qualitative dynamics of piecewise-linear differential equations : a discrete mapping approach.* Dynamics and Stability of Systems, vol. 4, no. 3-4, pages 565–583, janvier 1989. *Cité en page 68*

[Soetaert 2009] K. Soetaert, K. Van den Meersche et D. van Oevelen. *limSolve : solving linear inverse models*, 2009. Package R, version 1.5.1. *Cité en pages 136 et 164*

[Stouthamer 1975] A.H. Stouthamer et C.W. Bettenhaussen. *Determination of the efficiency of oxidative phosphorylation in continuous cultures of Aerobacter aerogenes.* Archives of Microbiology, vol. 102, no. 3, pages 187–192, mars 1975. *Cité en pages 210 et 258*

[Thiele 2005] I. Thiele, N.D. Price, T.D. Vo et B.O. Palsson. *Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet.* The Journal of Biological Chemistry, vol. 280, no. 12, pages 11683–11695, mars 2005. *Cité en pages 112 et 113*

[Thomas 1991] R. Thomas. *Regulatory networks seen as asynchronous automata : A logical description.* Journal of Theoretical Biology, vol. 153, no. 1, pages 1–23, novembre 1991. *Cité en page 66*

[Usuda 2010] Y. Usuda, Y. Nishio, S. Iwatani, S.J. Van Dien, A. Imaizumi, K. Shimbo, N. Kageyama, D. Iwahata, H. Miyano et K. Matsui. *Dynamic modeling of Escherichia coli metabolic and regulatory systems*

for amino-acid production. Journal of Biotechnology, vol. 147, no. 1, pages 17–30, mai 2010. *Cité en page 76*

[Uy 1999] D. Uy, S. Delaunay, J.m. Engasser et J.l. Goergen. *A method for the determination of pyruvate carboxylase activity during the glutamic acid fermentation with Corynebacterium glutamicum.* Journal of Microbiological Methods, vol. 39, no. 1, pages 91–96, décembre 1999. *Cité en page 257*

[Vallino 1993] J.J. Vallino et G. Stephanopoulos. *Metabolic flux distributions in Corynebacterium glutamicum during growth and lysine overproduction.* Biotechnology and Bioengineering, vol. 41, no. 6, pages 633–646, mars 1993. *Cité en pages 210 et 258*

[Van den Meersche 2009] K. Van den Meersche, K. Soetaert et D. van Oevelen. *xsample() : An R function for sampling linear inverse problems.* Journal of Statistical Software, Code Snippets, vol. 30, no. April, pages 1–15, 2009. *Cité en pages 110, 136 et 164*

[van Oevelen 2010] D. van Oevelen, K. Van den Meersche, F. Meysman, K. Soetaert, J.J. Middelburg et A.F. Vezina. *Quantifying food web flows using linear inverse models.* Ecosystems, vol. 13, no. 1, pages 32–45, 2010. *Cité en page 112*

[van Ooyen 2011] J. van Ooyen, D. Emer, M. Bussmann, M. Bott, B.J. Eikmanns et L. Eggeling. *Citrate synthase in Corynebacterium glutamicum is encoded by two gltA transcripts which are controlled by RamA, RamB, and GlxR.* Journal of Biotechnology, vol. 154, no. 2-3, pages 140–148, juillet 2011. *Cité en page 256*

[Varela 2003] C. Varela, E. Agosin, M. Baez, M. Klapa et G. Stephanopoulos. *Metabolic flux redistribution in Corynebacterium glutamicum in response to osmotic stress.* Applied Microbiology and Biotechnology, vol. 60, no. 5, pages 547–555, janvier 2003. *Cité en pages 158 et 258*

[Varela 2004] C.A. Varela, M.E. Baez et E. Agosin. *Osmotic stress response : quantification of cell maintenance and metabolic fluxes in a lysine-overproducing strain of Corynebacterium glutamicum.* Applied and Environmental Microbiology, vol. 70, no. 7, pages 4222–4229, juillet 2004. *Cité en pages 156, 158 et 258*

[Varma 1993] A. Varma, B.W. Boesch et B.O. Palsson. *Biochemical production capabilities of Escherichia coli.* Biotechnology and Bioengineering, vol. 42, no. 1, pages 59–73, juin 1993. *Cité en page 104*

[Varma 1994] A. Varma et B.O. Palsson. *Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.* Applied and Environmental Microbiology, vol. 60, no. 10, pages 3724–3731, octobre 1994. *Cité en pages 104, 124, 126, 137, 210 et 258*

[Veit 2009] A. Veit, D. Rittmann, T. Georgi, J.W. Youn, B.J. Eikmanns et V.F. Wendisch. *Pathway identification combining metabolic flux and functional genomics analyses : acetate and propionate activation by Corynebacterium glutamicum.* Journal of Biotechnology, vol. 140, no. 1-2, pages 75–83, mars 2009. *Cité en page 257*

[von der Osten 1989] C.H. von der Osten, C.F. Barbas, C.H. Wong et A.J. Sinskey. *Molecular cloning, nucleotide sequence and fine-structural analysis of the Corynebacterium glutamicum fda gene : structural comparison of C. glutamicum fructose-1,6-biphosphate aldolase to class I and class II aldolases.* Molecular Microbiology, vol. 3, no. 11, pages 1625–1637, novembre 1989. *Cité en page 256*

[Walsh 1984] K. Walsh et D.E. Koshland. *Determination of flux through the branch point of two metabolic cycles. The tricarboxylic acid cycle and the glyoxylate shunt.* The Journal of Biological Chemistry, vol. 259, no. 15, pages 9646–9654, août 1984. *Cité en page 152*

[Weaver 2014] D.S. Weaver, I.M. Keseler, A. Mackie, I.T. Paulsen et P.D. Karp. *A genome-scale metabolic flux model of Escherichia coli K-12 derived from the EcoCyc database.* BMC Systems Biology, vol. 8, page 79, janvier 2014. *Cité en page 90*

[Wendisch 2000] V.F. Wendisch, A.A. de Graaf, H. Sahm et B.J. Eikmanns. *Quantitative determination of metabolic fluxes during coutilization of two carbon sources : comparative analyses with Corynebacterium glutamicum during growth on acetate and/or glucose.* Journal of Bacteriology, vol. 182, no. 11, juin 2000. *Cité en pages 152 et 213*

[Wiback 2004] S.J. Wiback, I. Famili, H.J. Greenberg et B.O. Palsson. *Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space.* Journal of Theoretical Biology, vol. 228, no. 4, pages 437–447, juin 2004. *Cité en pages 110 et 133*

[Wiechert 2001] W. Wiechert. *13C metabolic flux analysis.* Metabolic Engineering, vol. 3, no. 3, pages 195–206, juillet 2001. *Cité en pages 114 et 115*

[Wieschalka 2013] S. Wieschalka, B. Blombach, M. Bott et B.J. Eikmanns. *Bio-based production of organic acids with Corynebacterium gluta-*

micum. Microbial Biotechnology, vol. 6, no. 2, pages 87–102, mars 2013. *Cité en pages 256 et 257*

[Willenbrock 2004] H. Willenbrock et D.W. Ussery. *Chromatin architecture and gene expression in Escherichia coli.* Genome Biology, vol. 5, no. 12, page 252, janvier 2004. *Cité en page 13*

[Wittmann 2004a] C. Wittmann, P. Kiefer et O. Zelder. *Metabolic fluxes in Corynebacterium glutamicum during lysine production with sucrose as carbon source.* Applied and Environmental Microbiology, vol. 70, no. 12, pages 7277–7287, décembre 2004. *Cité en page 258*

[Wittmann 2004b] C. Wittmann, H.M. Kim et E. Heinzle. *Metabolic network analysis of lysine producing Corynebacterium glutamicum at a miniaturized scale.* Biotechnology and Bioengineering, vol. 87, no. 1, pages 1–6, juillet 2004. *Cité en page 258*

[Xi 2011] Y. Xi, Y.P.P. Chen, C. Qian et F. Wang. *Comparative study of computational methods to detect the correlated reaction sets in biochemical networks.* Briefings in Bioinformatics, vol. 12, no. 2, pages 132–150, mars 2011. *Cité en page 113*

[Yeang 2006] C.H. Yeang et M. Vingron. *A joint model of regulatory and metabolic networks.* BMC Bioinformatics, vol. 7, page 332, janvier 2006. *Cité en page 60*

[Zarecki 2014] R. Zarecki, M.a. Oberhardt, K. Yizhak, A. Wagner, E. Shtifman Segal, S. Freilich, C.S. Henry, U. Gophna et E. Ruppin. *Maximal sum of metabolic exchange fluxes outperforms biomass yield as a predictor of growth rate of microorganisms.* PloS ONE, vol. 9, no. 5, page e98372, janvier 2014. *Cité en page 106*

[Zhao 2002] Y. Zhao et Y. Lin. *Flux distribution and partitioning in Corynebacterium glutamicum grown at different specific growth rates.* Process Biochemistry, vol. 37, pages 775–785, 2002. *Cité en page 258*

[Zhuang 2011] K. Zhuang, M. Izallalen, P. Mouser, H. Richter, C. Risso, R. Mahadevan et D.R. Lovley. *Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments.* The ISME Journal, vol. 5, no. 2, pages 305–316, février 2011. *Cité en pages 90 et 126*

[Zubay 1973] G. Zubay. *In vitro synthesis of protein in microbial systems.* Annual Review of Genetics, vol. 7, no. 5, pages 267–287, janvier 1973. *Cité en page 93*

Résumé

À l'issue de ce travail de thèse, je propose une approche basée sur le formalisme des modèles à base de contraintes, pour étudier la dynamique d'un système métabolique. En associant l'échantillonnage de l'espace des solutions avec l'utilisation d'une contrainte de « faisabilité » entre les périodes de temps considérées, cette approche permet de modéliser la dynamique d'un système métabolique en prenant en compte la variabilité des mesures expérimentales. La contrainte de faisabilité entre les périodes permet de garantir que chaque « trajectoire solution » correspond à une succession de cartes de flux qui conduit à des cinétiques de concentrations cohérentes avec les mesures expérimentales. Les populations de trajectoires solutions générées autorisent différents types d'analyses. D'une part, les répartitions de flux prédites peuvent être utilisées afin d'estimer les répartitions de flux les plus plausibles au sein du réseau étudié. D'autre part, la distribution des concentrations prédites permet d'évaluer le modèle utilisé pour étudier le réseau métabolique. Le fait que cette approche soit basée sur le formalisme de la modélisation à base de contraintes permet, moyennant l'utilisation de l'hypothèse d'état stationnaire du système, d'étudier des réseaux métaboliques de taille relativement grande, et d'utiliser des données expérimentales qui sont aisément mesurables, par exemple les concentrations en biomasse et en métabolites extracellulaires. Cette approche par « trajectoires solutions » a été utilisée afin d'étudier la dynamique du métabolisme de *Corynebacterium glutamicum*, lorsqu'elle est cultivée en condition de limitation en biotine. Les résultats obtenus ont permis d'une part d'attester du fonctionnement de la méthode, et d'autre part de proposer plusieurs hypothèses quant aux phénomènes biologiques qui ont lieu pendant cette condition particulière de croissance.

Abstract

In this thesis, I propose an approach based on the formalism of constraint-based models to study the dynamics of a metabolic system. By combining the sampling of the solutions space and the use of a "feasibility" constraint between the considered time periods, this approach allows to model the dynamic of a metabolic system taking into account the variability of experimental measurements. The feasibility constraint between time periods ensures that each "solution trajectory" corresponds to a succession of flux maps which leads to some kinetics of concentrations that are consistent with the experimental measurements. The generation of a population of solution trajectories allows several analyses. On the one hand, the predicted flux maps can be used to estimate the most plausible flux within the network studied. On the other hand, the distribution of predicted concentrations enables to assess the model used for studying the metabolic network. The fact that this approach is based on the formalism of constraint-based modeling allows, using the steady-state assumption of the system, to study metabolic networks of relatively large size, and to use experimental data that are easily measurable, such as biomass concentration and extracellular metabolites concentration. This approach by "solution trajectories" has been used to study the dynamics of the metabolism of *Corynebacterium glutamicum*, when grown under biotin-limited condition. The results allowed, first, to attest the functioning of the method, and second, to propose several hypotheses about biological phenomena that take place during this particular growth condition.