Université d'Evry, Val d'Essonne
Ecole doctorale *Des génomes aux organismes*

THESE DE DOCTORAT

Présentée pour l'obtention du titre de
Docteur de l'université d'Evry Val d'Essonne
Par

Gilles VIEIRA

Etude de la diversité métabolique dans l'espèce *Escherichia coli*.

A l'aide de réseaux et de modèles du métabolisme à l'échelle de l'organisme.

Soutenance publique le 5 décembre 2011 devant le jury composé de :

Marie-France SAGOT	Rapporteur
Jean-Charles PORTAIS	Rapporteur
Bruno BOST	Examinateur
Jean-Loup FAULON	Examinateur
Jacques VAN-HELDEN	Examinateur
Vincent SCHACHTER	Membre
David VALLENET	Encadrant
Claudine MEDIGUE	Directrice de thèse

Travail réalisé au sein du Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme. CEA – Genoscope, UMR 8030 de génomique métabolique

Résumé

Il existe plusieurs façons de concevoir l'étude des différences métaboliques chez les microorganismes. On peut soit s'intéresser à des variations ponctuelles (i.e. la présence ou non de certains acteurs moléculaires), soit s'intéresser à des variations des capacités métaboliques des organismes. Derrière ces analyses se cachent deux niveaux d'études : le premier se situe au niveau d'un ensemble restreint de gènes et d'activités enzymatiques et permet d'établir des relations de causalité entre ces deux éléments. Le second est à l'échelle la cellule et permet d'avoir une vue d'ensemble des capacités de croissance de l'organisme dans différentes conditions.

Cependant, la complexité et l'interdépendance des processus métaboliques, et surtout le manque de réseaux reconstruits suffisamment détaillés rendent difficile l'explicitation des liens entre caractéristiques phénotypiques de croissance et le génotype lors des analyses à l'échelle de la cellule.

En effet les avancées technologiques, dont le séquençage intégral du génome d'un organisme, l'accumulation des connaissances biologiques référencées dans des bases de données et enfin les avancées méthodologiques, permettent de reconstruire des réseaux et des modèles du métabolisme à l'échelle de la cellule. Malheureusement, pour obtenir des réseaux de qualité, il est encore aujourd'hui indispensable de passer par un processus de curation manuelle, long et fastidieux.

Dans cette thèse, nous proposons une nouvelle stratégie de reconstruction de réseaux et de modèles du métabolisme à l'échelle globale. Cette stratégie s'applique à un nombre quelconque d'organismes à condition qu'ils soient de la même espèce (ou proches d'un point de vue phylogénétique) et qu'il existe un réseau métabolique de référence de bonne qualité pour au moins l'un d'entre eux. Le point clé de cette stratégie repose sur l'utilisation et la propagation automatisée des connaissances déjà acquises sur les organismes étudiés. Nous avons appliqué cette stratégie pour reconstruire et étudier les réseaux métaboliques de 23 *Escherichia coli* et 6 *Shigellas*. Ces souches couvrent les différents groupes phylogénétiques des *E. coli* et contiennent des représentants de différents types de pathogénicité.

Les réseaux ainsi reconstruits présentent une nette amélioration en comparaison des réseaux reconstruits par les stratégies habituelles. L'analyse de ces réseaux a mis en évidence un fort lien entre l'évolution du métabolisme et l'histoire évolutive des souches. Pourtant, tous les processus métaboliques ne sont pas impactés de la même manière. On observe que certains processus sont fortement conservés alors que d'autres semblent subir moins de contraintes sélectives.

Nous avons ensuite converti ces réseaux en modèles métaboliques pour explorer les capacités physiologiques des différentes souches. Nous avons comparé nos prédictions de croissance à des expériences de croissance ainsi qu'aux résultats du modèle de référence. Dans le meilleur des cas, les modèles reconstruits ont un nombre de bonnes prédictions supérieur au modèle de référence et dans le pire des cas, ce nombre est voisin de celui du modèle de référence. Cette observation est le résultat de l'ajout de nouvelles voies métaboliques fonctionnelles (i.e, en accords avec les observations expérimentales) dans les modèles reconstruits.

Enfin nous avons préparé nos modèles pour l'intégration de données biologiques hétérogènes, données contenant des résultats expérimentaux, dont des concentrations d'enzyme et des résultats de simulation de flux provenant d'un modèle cinétique.

Au final le travail réalisé propose une nouvelle stratégie de reconstruction de réseaux et de modèles du métabolisme à l'échelle de la cellule qui permet d'étudier le lien entre l'évolution et les capacités métaboliques des organismes étudiés.

Tables des matières

Liste des tables

Listes des figures

Listes des annexes

Lexique

	A
ACM	Analyse des Correspondances Multiple
ADP	Adénosine Diphosphate
AND	Acide Désoxyribonucléique
ANR	Agence Nationale de Recherche
ARN	Acide Ribonucléique
ATP	Adénosine Triphosphate
ATPM	Réaction de maintenance énergétique
	B
BBH	Best Bidirectionnal Hit
bmax	borne maximale du flux
bmin	borne minimale du flux
bnumbers	nom des identifiants des gènes dans iAF1260 du type b****
	C
CAS	Chemical Abstracts Service
CBM	constraint-based model
cbmCom	Partie commune du modèle reconstruit avec iAF1260
CDS	CoDing Sequence
CO2	dioxyde de carbone
COBRA	COnstraint-Based Reconstruction and Analysis
cycSpe	Partie spécifique du modèle issue des cycs recosntruits
	D
DAEC	Diffusely adherent E. coli
DO	Densité Optique
	E
E. coli	Escherichia coli
EAEC	Enteroaggregative *E. coli*
EC number	Enzyme Comission number
EHEC	Enterohaemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
EPEC	Enteropathogenic *E. coli*
ETEC	Enterotoxigenic *E. coli*
ExPEc	Extraintestinal Pathogenic *E. coli*
	F
FBA	Flux Balance Analysis
FVA	Flux Variance Analysis
	G
GEM	Genome-based Modeling
Gla	Milieu minimum gluconate
Glc	Milieu minimum glucose

GO_id	Gene identifiant
GPR	Gene Protein Reaction link
gprEval	Evaluation des GPRs
	I
iAF1260	Modèle de référence d'*E coli*
iMAT	integrative Metabolic Analysis Tool
InPEc	Intestinal Pathogenic *E. coli*
IS	Insertion Sequence
	K
km	constante d'affinité de l'enzyme pour son substrat
	L
LB	Lysogeny broth
	M
MP_id	MetaCyc Pathway identifiant
MR_id	MetaCyc Reaction identifiant
	N
NAD	Nicotinamide adénine dinucléotide (oxydent)
NADH	Nicotinamide adénine dinucléotide (réducteur)
NMEC	Neonatal meningitidis E. coli
	O
O_id	Organisme identifiant
	P
pf	fichier d'entrée de Pathologic
PGDB	Pathway Genome DataBase
PkGDB	Prokaryote Genom DataBase
PTS	phosphotransferase system
	R
R	Milieu riche
Rss	Milieu riche sans sucre
	S
S	matrice stoechiométrique
SBML	System biology markup language
	T
THF	Tetrahydrofuran
	U
UPEC	Uropathogenic *E. coli*
	V
v	vecteur de flux
vmax	Vitesse maximale d'une enzyme, in vitro avec excès de substrat
	X
Xref	Référence croisée

Introduction

L'évolution des technologies en biologie permet maintenant de travailler à l'échelle de l'organisme, notamment grâce à l'évolution du séquençage. Cependant connaître le génotype d'un organisme n'assure pas une compréhension directe du fonctionnement de celui-ci et il faut généralement un pont, comme le métabolisme, pour relier le génotype au phénotype. La quantité de données nécessaire pour travailler à l'échelle de la cellule est telle qu'il a fallut faire appel à l'informatique. L'association entre la biologie et l'informatique n'est pas intuitive et le lecteur peut se demander comment lier ces deux disciplines. Pour montrer les intérêts d'une approche pluridisciplinaire dans la compréhension des phénomènes biologiques, nous allons, au cours des quatre parties de ce chapitre, détailler les éléments étudiés aussi bien d'un point de vue biologique qu'informatique. Ils montreront comment la connaissance biologique peut être modélisée en un objet mathématique. La première partie s'intéressera au métabolisme des organismes, elle donnera les éléments qui permettront d'en comprendre les principaux acteurs et mécanismes. La deuxième partie portera sur son pendant mathématique, notamment les différentes manières de formaliser, représenter et modéliser un même concept biologique. La troisième partie s'attachera à la biodiversité des microorganismes, elle expliquera l'intérêt des organismes modèles et donnera les détails de l'espèce *Escherichia coli* qui a été retenue pour les travaux de cette thèse. Enfin la quatrième et dernière partie de cette introduction sera l'occasion de présenter au lecteur les différents processus de reconstruction des modèles du métabolisme et leur utilisation, elle en donnera les limites ainsi que les problématiques soulevées.

Le métabolisme : in vivo

1 Généralités

Le mot métabolisme provient du grec *metabolê* qui signifie *changement*. Cette traduction caractérise en effet de manière appropriée les activités réalisées par le métabolisme et ses effets sur le monde extérieur. Un organisme vivant est capable de croître et de se reproduire. Pour se faire, il doit assimiler et convertir les substances présentes dans son environnement. A l'aide de processus chimiques et biochimiques, les différentes molécules consommées subissent des changements de forme ou de nature qui généreront la matière organique et l'énergie nécessaires à la vie de l'organisme.

D'une manière générale, le métabolisme désigne l'ensemble des acteurs et des processus biochimiques de dégradation, de synthèse de molécules biologiques et de production d'énergie chimique. La consommation et la sécrétion des molécules biologiques - appelées métabolites - entrainent des variations de leur concentration aussi bien à l'intérieur qu'à l'extérieur de l'organisme. Ces variations s'accompagnent parfois de *changements* morphologiques, physiologiques ou environnementaux, qui peuvent être aussi bien microscopiques que macroscopiques. La diversité du métabolisme est très importante, c'est pourquoi, dans les quelques paragraphes suivants, nous avons donné des exemples d'effets du métabolisme dans différents organismes, aussi bien chez les microorganismes que chez les plantes ou l'Homme.

Le métabolisme des triglycérides est un exemple de modification morphologique à l'échelle de la cellule. Issus de la dégradation des lipides, ces triglycérides entrainent une modification de la taille des adipocytes. En cas de forte concentration en lipide, l'organisme synthétise des triglycérides qui sont stockés dans la vacuole des adipocytes. Cette accumulation provoque une augmentation de la taille de la cellule. A contrario en cas de manque de lipide, les triglycérides sont convertis en lipides entrainant une diminution de la taille de l'adipocyte (Figure 1).

Figure 1: Adipocyte humain qui accumule les triglycérides.
(Source http://www.vulgaris-medical.com/)

On peut également citer comme modification morphologique, le tubercule de pomme de terre qui est une accumulation d'amidon à l'échelle macroscopique. La disponibilité en métabolite joue un rôle important sur la physiologie de l'organisme

et souvent une petite variation de concentration d'un seul métabolite peut avoir des conséquences importantes sur la physiologie et la morphologie d'un organisme. L'un des changements physiologiques les plus connus est la glycémie qui est liée à la quantité de sucre présente dans le sang. Lorsque cette concentration en sucre est trop faible l'individu est en état de manque et peut subir une crise d'hypoglycémie. Similairement, la concentration d'adrénaline a un effet direct sur le rythme cardiaque. Une augmentation de sa concentration en accélère le rythme.
Les organismes puisent dans l'environnement les métabolites nécessaires à leur vie et rejettent les métabolites dont ils n'ont plus l'utilité. Ces échanges continus modifient la composition de l'environnement. Ce changement peut être minime car les productions des différents organismes compensent leurs consommations respectives, comme dans le cas des écosystèmes qui sont à l'équilibre (Looman 1976). L'environnement peut aussi subir de profonds changements. Dans un composteur, les matières végétales (branches, feuilles, herbes, etc.) sont décomposées en matière organique par des microorganismes. Cette matière organique composée principalement d'humus est d'un aspect et d'une morphologie totalement différents des matières végétales d'origine.
Nous terminons notre série d'exemples par un composant du métabolisme qui a eu un impact à l'échelle de la planète et est en partie responsable de la biodiversité animale observée aujourd'hui. Ces effets sont dus à une réaction apparue il y a deux milliards d'années (Knoll 2003) et qui consiste à créer de la matière organique à partir d'eau, de dioxyde de carbone et de lumière, cette réaction est connue sous le nom de photosynthèse.

$$6CO_2 + 12H_2O \xrightarrow{Lumière} C_6H_{12}O_6 + 6O_2 + 6H_2O$$

Avant l'apparition de la photosynthèse, l'atmosphère ne contenait pratiquement pas de dioxygène et environ 10% de CO_2. Cette composition conférait à l'atmosphère un fort pouvoir réducteur. L'importante production de dioxygène par les organismes photosynthétiques au cours des millénaires a profondément modifié cette composition. Elle a réduit le CO_2 à l'état de trace et augmenté le niveau de dioxygène à 21%, donnant un pouvoir désormais oxydant à l'atmosphère. Ce changement de pouvoir oxydo-réducteur a permis l'apparition des mécanismes de respiration qui interviennent dans la synthèse d'énergie.
Tous ces exemples de changement (morphologique, physiologique, environnemental) ne sont qu'une infime partie de la diversité du métabolisme. Depuis la découverte de la glycolyse (Romano & T Conway 1996) qui est l'une des voies de synthèse d'énergie universelle les plus anciennement décrite, de nombreuses autres voies ont été identifiées. On peut citer la fermentation que Pasteur définissait comme « la vie sans air » et qui permet la production d'énergie dans des milieux anaérobiques. Chez les plantes on estime à plus d'un million le nombre de molécules synthétisables (Saito & Matsuda 2010). Cette diversité permet aux organismes (Natronomonas *pharaonis* (Falb et al. 2005; Gonzalez et al. 2010)) de vivre dans des conditions extrêmes comme dans les lacs salins d'Egypte, alors que l'environnement est extrêmement basique (pH supérieur à 11). Cette diversité permet une adaptation face aux modifications de l'environnement à tel point que des microorganismes peuvent dégrader des molécules non naturelles, créées par l'Homme (xénobiotique).

Cette diversité métabolique est depuis longtemps utilisée et mise à profit par l'Homme, aussi bien dans son alimentation, que sa santé et son bien être. Bien que

le travail de cette thèse ne porte pas sur l'application par l'Homme du métabolisme, il nous semble pertinent de donner quelques exemples d'utilisation du métabolisme pour en montrer l'importance. Nous venons de mentionner la fermentation, qui est un processus fréquemment exploité dans notre alimentation. Celui-ci entre dans la composition du pain, des alcools (bière, vin, etc.) et des produits laitiers (fromage, yaourt). Ce processus, exécuté par des bactéries et des levures, est utilisé depuis l'antiquité alors que les notions mêmes de métabolisme et de microorganismes sont beaucoup plus récentes (XVII^{ème} siècle).
La découverte des microorganismes pathogènes a suscité des interrogations concernant l'équilibre des populations et les mécanismes naturels de défense. Ernest Duchesne étudia une culture mixte de bactéries *Escherichia coli* et de champignons *Penicillium glaucum* et observa que ce dernier décime la population de *E. coli* (Duchesne 1912). L'étude de cette famille de champignon conduit Flemming à la découverte d'un antibiotique : la pénicilline (J. Wong 2003). Les antibiotiques sont synthétisés naturellement par certains organismes et l'Homme est capable maintenant de les fabriquer synthétiquement. Malheureusement les capacités d'adaptation des organismes ont permis l'apparition de voies de dégradation des antibiotiques (Henriques Normark & Normark 2002). Par conséquent de plus en plus de microorganismes deviennent résistants, ce qui implique de sérieux risques sanitaires.

Nos connaissances sur le métabolisme sont récentes par rapport à son utilisation par l'Homme. Ce n'est en effet que depuis une dizaine d'années que nous commençons à avoir une vue d'ensemble du métabolisme et de ses constituants chez les bactéries (P D Karp et al. 1999).

2 Les composants du métabolisme

Le terme métabolisme englobe tout ce qui est en relation avec la transformation chimique des molécules biochimiques : les acteurs et les moteurs moléculaires ainsi que leur organisation et hiérarchisation, qui sont des concepts plus abstraits, aident à la classification et la compréhension du métabolisme.

.2.1 Métabolites

Toutes les molécules organiques – c'est-à-dire qui contiennent au moins un atome de carbone - intervenant dans le métabolisme sont appelées des métabolites. Certaines molécules non organiques mais essentielles à la vie des organismes sont également considérées comme des métabolites : le phosphate PO_4^{3-}, le dioxygène et l'azote. D'autres atomes comme le fer, le magnésium, le potassium, le sodium, le calcium ou le chlore, sont également impliqués dans le métabolisme. Ces atomes ne sont cependant pas comptés comme des métabolites bien que leur rôle de capteur ou d'activateur soit essentiel. Des études sur la bactérie *Escherichia coli* B/r, ont permis de définir sa composition en biomasse Table 1 et (F. C. Neidhardt & Umbarger 1996) :

	% de la masse sèche	Nombre de molécules par cellule	Nombre de molécules différentes
Protéine	55	2350000	1850
ARN	20,5	255480	> 660
ADN	3,1	2,1	1
Lipide	9,1	22000000	Indisponible
Lipopolysaccharide	3,4	1430000	1
Peptidoglycan	2,5	1	1
Glycogène	2,5	4300	1
Polyamines	0,4	6700000	2
Métabolites, cofacteurs, ions	3,5	Indisponible	>> 800

Table 1 : Estimation des composants de la biomasse de la souche *E. coli* B/r.
Données issues de (F. Neidhardt 1996)

Il est très difficile d'estimer le nombre de métabolites, on sait qu'il en existe bien plus d'un million de types différents. Même si des dosages permettent d'estimer la quantité de quelques métabolites bien connus, on est, pour le moment, incapable d'estimer la quantité de tous les métabolites et ce, même grossièrement.
Cependant des études ont permis d'estimer le ratio entre les principaux atomes (carbone, oxygène, azote) des métabolites dans différents organismes Table 2 et (Oliveira et al. 2005; von Stockar & J. Liu 1999).

Organismes	Formule
Bactérie	CH(1,66)O(0,41)N(0,21)
Algue	CH(1,63)O(0,44)N(0,09)
Levure	CH(1,65)O(0,54)N(0,10)

Table 2: Ratio des différents principaux atomes dans différents types d'organismes.

On peut également citer comme atomes importants le phosphore et le souffre dont les ratios sont de l'ordre de 10^{-2} et 10^{-3}. Ces valeurs sont des moyennes qui dépendent de nombreux paramètres et principalement de la composition du milieu.
Certains métabolites sont essentiels au développement et à la vie de l'organisme, on les appelle métabolites primaires. Par oppositions, les métabolites non essentiels, qui peuvent se trouver dans des compartiments particuliers ou alors qui interviennent ponctuellement, sont appelés métabolites secondaires.

.2.2 Réactions

On désigne par réaction métabolique les transformations chimiques que subissent les métabolites. Une réaction modifie des métabolites substrats en métabolites produits et est généralement représentée par son équation bilan () :

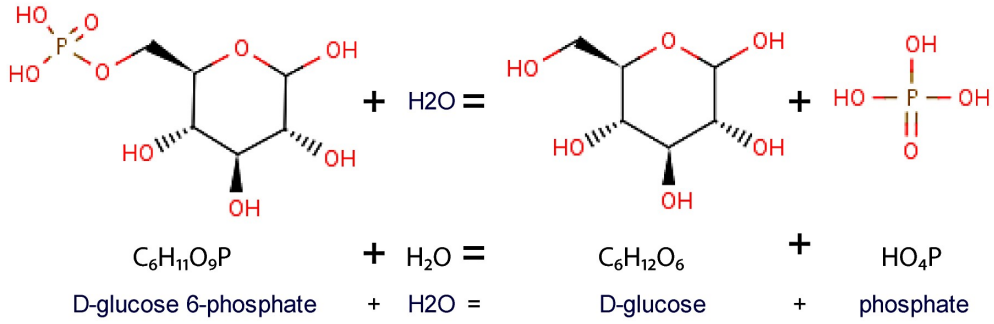


Figure 2: Equation bilan de la réaction "*glucose-6-phosphatase*".
Formule semi-développée (haut) formule brute (milieu) et nom des métabolites (bas). (Source http://www.brenda-enzymes.info/php/result_flat.php4?ecno=3.1.3.9)

Par convention les substrats sont placés à gauche de l'égalité et les produits à droite. Toutes les réactions suivent la loi de conservation de la matière, c'est à dire que le nombre d'atomes des substrats est égal aux nombres d'atomes des produits. Dans l'équation précédente le bilan des atomes pour une molécule de chacun des substrats est de 6 carbones, 13 hydrogènes, 10 oxygènes et un phosphore ($C_6H_{13}O_{10}P$) et il est identique au bilan des atomes pour une molécule de chaque produit ($C_6H_{13}O_{10}P$). Certaines réactions nécessitent d'utiliser plusieurs fois le même substrat ou donnent plusieurs fois le même produit. Le nombre de molécules mises en jeu est appelé coefficient stœchiométrique. Il désigne la proportion relative de métabolites transformés dans la réaction (Figure 3).

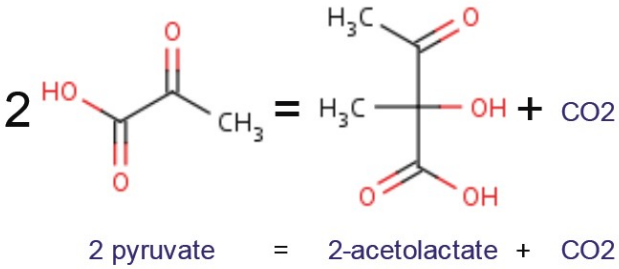


Figure 3: Réactions nécessitant plusieurs molécules d'un même substrat.
L'acétolactate synthase nécessite 2 molécules de *pyruvates* pour former une molécule de *2-acetolactate*.

Le signe égal, dans la *glucose-6-phosphatase*, signifie que la réaction est réversible. C'est à dire que la réaction peut transformer le D-glucose 6-phosphate en D-glucose et phosphate ou l'inverse, produire le D-glucose 6-phosphate à partir de D-glucose. Le D-glucose est dans ce cas un substrat et non plus un produit. Dans certains cas, notamment les réactions dont l'équilibre thermodynamique est fortement déplacé vers le produit, il est impossible d'effectuer la réaction dans les deux sens : une telle réaction est dite irréversible.

Les métabolites intervenant dans une réaction peuvent être séparés en deux catégories : les métabolites principaux et les cofacteurs. Les métabolites principaux désignent les métabolites d'intérêt transformés au cours de la réaction. Les cofacteurs quand à eux désignent des métabolites qui vont apporter un élément biochimique aux métabolites principaux. Il existe trois grands types de cofacteurs, ceux qui apportent de l'énergie et qui sont généralement composés d'un groupement phosphate comme l'ATP ; ceux impliqués dans les échanges d'électron tel que le couple NAD/NADH ; enfin ceux qui vont apporter ou prendre un groupement

chimique dont le THF. Comme les cofacteurs apportent ou récupèrent des éléments biochimiques ils sont souvent caractérisés par le couple avec/sans l'élément (Figure 4).

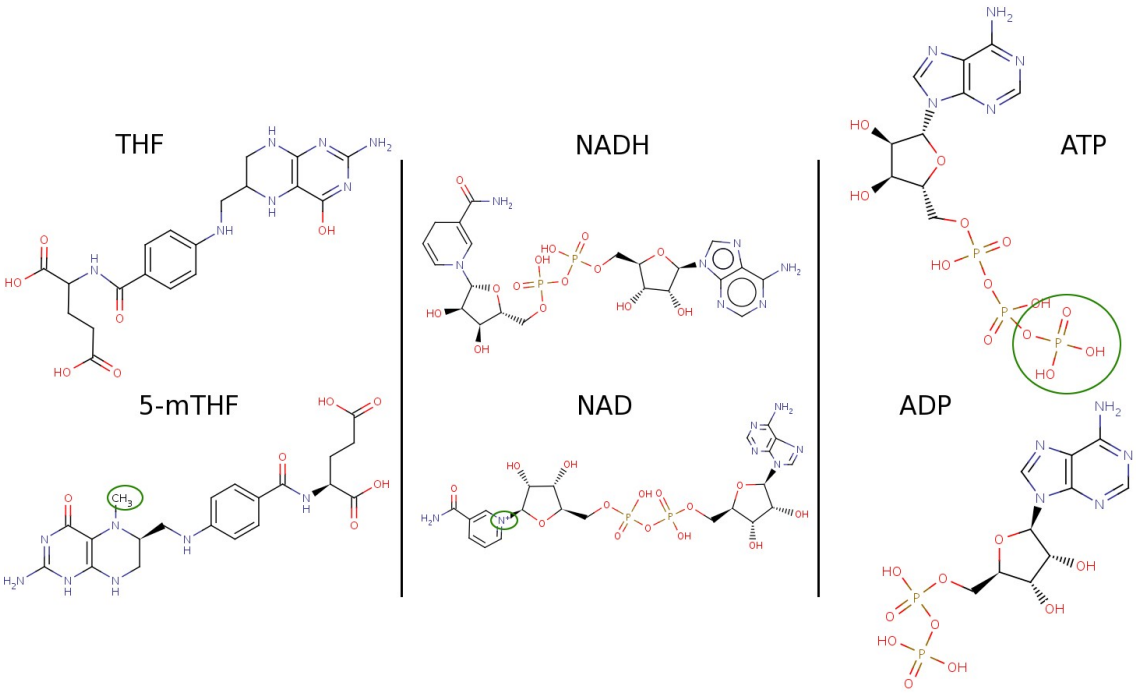


Figure 4: Couples de cofacteurs.
A gauche le couple 5-mTHF/THF pour le transfert du groupe méthyle. Au centre le couple NADH/NAD spécialisé dans le transfert de proton. A droite le couple ATP/ADP spécialisé dans l'apport d'énergie (liaison phosphate).

Les cofacteurs ont généralement une structure complexe et sont souvent synthétisables par l'organisme. Lors de sa synthèse la molécule habituellement cofacteur devient le métabolite principal. Il existe des processus biochimiques où un métabolite principal intervient aussi en tant que cofacteur, par exemple durant la synthèse d'ADP.

.2.3 Enzymes

Les réactions chimiques et biochimiques sont sujettes aux principes de la thermodynamique et généralement, après transformation, les molécules produites ont un niveau d'énergie plus faible. L'initiation de la réaction nécessite de passer la barrière de l'énergie d'activation qui représente l'énergie requise pour casser les liaisons des substrats, avant de créer celles des produits. Certaines réactions, dites spontanées, peuvent passer naturellement cette barrière. Pour les autres il existe des moteurs moléculaires protéiques ou ribonucléiques (ribozymes) (Guerrier-Takada et al. 1983) appelés enzymes. La première enzyme fut isolée par Anselme Payen et Jean-François Persoz en 1833 (Gay-Lussac et al. 1833), il s'agit de l'amylase initialement appelée diastase (du Grec *διαστασις*, "séparation").
Ce n'est que quarante après, en 1877 que le terme enzyme (du Grec *ενζυμον*, "levain") est utilisé par Wilhelm Kühne (Verein & Heidelberg 1877), pour désigner dans un premier temps le processus de fermentation.
Les enzymes sont constituées d'au moins un site de fixation du substrat et d'un site actif où se produira la réaction. Elles sont des catalyseurs, autrement dit la molécule

enzymatique est identique en début et en fin de réaction, mais peut varier surtout au niveau de sa forme durant la réaction (Figure 5A).

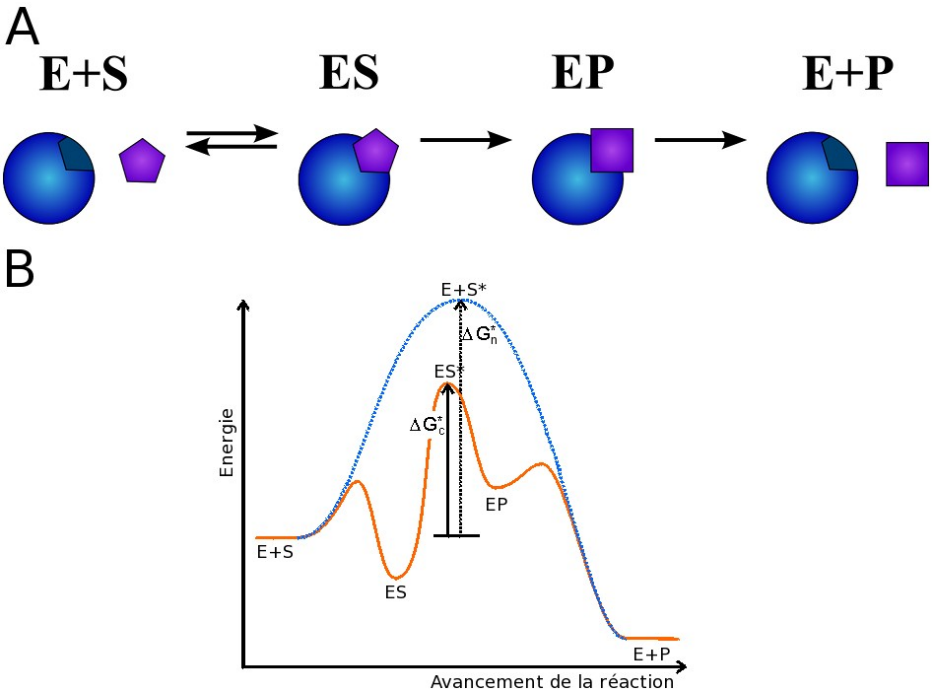


Figure 5 : Fonctionnement d'une enzyme et évolution de l'énergie au cours d'une réaction catalysée et non catalysée.

A : L'enzyme E et le substrat S forment un complexe ES, S est alors converti en produit P et forme le complexe EP, enfin le produit P se dissocie de l'enzyme. B : enzyme (E), substrat (S), substrat à l'état transitoire (S*), produit (P), complexe enzyme substrat (ES), complexe enzyme substrat transitoire (ES*), complexe enzyme produit (EP), énergie d'activation catalysée (ΔG^*_c) et non catalysée (ΔG^*_n) (source http://www.lsbu.ac.uk/biology/enztech/mechan.html).

Le complexe formé par l'enzyme et le substrat permet d'abaisser l'énergie d'activation nécessaire pour déclencher la réaction (Figure 5B) et ce, en facilitant la rencontre entre les différents substrats ou en stabilisant ces derniers par des liaisons électrochimiques. Cet abaissement facilite la réaction. Les enzymes sont appelées les moteurs du métabolisme puisqu'elles accélèrent significativement des réactions qui sont naturellement très lentes Table 3.

L'enzyme est spécifique du substrat et de la réaction réalisée. Seule la transformation du bon substrat sera accélérée et avec uniquement la réaction précise. Ceci confère un énorme avantage à la réaction par rapport à toutes les réactions annexes qui pourraient arriver sur ce substrat, et permet ainsi à la cellule de contrôler les réactions qui ont lieu.

	Vitesse non enzymatique (s^{-1})	Vitesse enzymatique (s^{-1})	Facteur d'accroissement
Chymotrypsine	4.10^{-9}	4.10^{-2}	10^{7}
Lysozyme	3.10^{-9}	5.10^{-1}	2.10^{8}
Triose phosphate isomérase	6.10^{-7}	2.10^{3}	3.10^{9}
Fumarase	2.10^{-8}	2.10^{3}	10^{11}
Uréase	3.10^{-10}	3.10^{4}	10^{14}
Désaminase d'adénosine	10^{-12}	10^{2}	10^{14}
Phosphatase alcaline	10^{-15}	10^{2}	10^{17}

Table 3: Effet des enzymes sur la vitesse des réactions. Données issues de (Horton et al. 1994).
Ces données sont à titre d'exemples et ne représentent pas les valeurs extrêmes.

L'activité enzymatique est caractérisée par un identifiant à quatre numéros appelés numéro EC pour Enzyme Commission. Ce numéro est défini par *l'International Union of Biochemistery and Molecular Biology* (IUBMB site : http://www.iubmb.org/). Il est basé sur la fonction de l'enzyme (premier numéro) et sa spécificité (les trois numéros suivants). Les deuxième et troisième numéros servent à préciser la cible de l'enzyme notamment les groupes fonctionnels chimiques impliqués. Le dernier numéro lui est spécifique des substrats et cofacteurs, par exemple ce numéro peut dépasser les 300 pour certaines oxydoréductases. Ceci contraste avec le premier numéro puisqu'il n'existe que 6 grandes familles d'enzyme (Table 4).

Nombre	EC	Type de réactions catalysées
1.-.-.-	Oxydoréductases	Réactions d'oxydoréduction
2.-.-.-	Transférases	Réactions de transfert de groupes fonctionnels
3.-.-.-	Hydrolases	Réactions d'hydrolyse d'un substrat en deux produits
4.-.-.-	Lyases	Réactions de coupure de liaisons covalentes par un procédé autre que l'oxydation ou l'hydrolyse
5.-.-.-	Isomérases	Réactions de réarrangement intramoléculaire, p. ex. isomérisation
6.-.-.-	Ligases	Réactions de jonction covalente de deux molécules utilisant l'hydrolyse d'ATP

Table 4 : Premier numéro de la classification enzymatique.
Il existe 6 grandes familles d'enzymes, chacune correspond à une activité biochimique spécifique.

2.3.1 Variation de l'activité enzymatique.

La vitesse d'une réaction est définie comme le nombre de molécules modifiées par unité de temps et s'exprime en $mol.L^{-1}.s^{-1}$. Cette vitesse de conversion de matière est nommée le flux de la réaction. Ce flux va dépendre de plusieurs facteurs, séparables en deux catégories : les contraintes liées à la thermodynamique d'une part et celles liées aux enzymes d'autre part.
Nous n'aborderons pas en détail l'aspect thermodynamique des réactions. Toute réaction possède une enthalpie libre ($\Delta_r G$) qui décrit le sens d'évolution de la réaction. Si cette valeur est négative la réaction ira du substrat vers les produits et inversement si cette valeur est positive. Dans le cas ou cette valeur est nulle le flux de la réaction devient nul. L'enthalpie libre dépend de la température, de la pression et des quantités de produits et substrats. A pression et température constantes, l'enthalpie libre ne dépendra que du ratio entre les produits et les substrats. On comprend facilement que la variation et la disponibilité des différents

acteurs métaboliques va influencer l'enthalpie libre et par conséquence la vitesse de la réaction. Il existe deux grands types de contrôles enzymatiques : le premier va agir sur la disponibilité en enzyme, le deuxième sur ses capacités catalytiques.
Les enzymes protéiques ou nucléotidiques sont produites via des processus de transcription et de traduction des gènes. L'organisme par des processus de régulation peut activer, désactiver ou amplifier chacune de ces étapes. Il peut également agir sur la dégradation de chacun des produits de ces étapes. Ces processus se situent bien en amont du métabolisme et sortent du cadre de nos travaux : ils ne seront pas plus détaillés (Berg et al. 2006; Horton et al. 1994).
Une fois l'enzyme produite, elle peut subir des modifications chimiques telles que l'ajout de groupements chimiques fonctionnels, notamment un phosphate. Cette phosphorylation peut faire passer l'enzyme d'un état actif à un état inactif ou inversement. Une grande partie des modifications de l'activité enzymatique est le fait de métabolites et petites molécules qui interagissent directement avec l'enzyme pour en activer ou inhiber l'activité. La Figure 6 schématise les effets de molécules activatrices et inhibitrices sur la conformation de l'enzyme.
Trois grands types de mécanismes d'inhibition sont généralement observés: inhibition compétitive, non-compétitive et mixte. L'inhibition compétitive est provoquée par une molécule non transformable par l'enzyme, mais capable de se fixer sur le site de fixation du substrat. Ceci entraine une compétition entre le substrat et l'inhibiteur pour se fixer sur l'enzyme. Lors d'une inhibition non compétitive, l'inhibiteur ne se fixe plus sur le site de fixation du substrat mais sur un autre site et induit un changement de conformation de l'enzyme. Ce changement de forme verrouille l'accès au site actif ou bloque la modification du substrat déjà présent dans le site de fixation. Enfin, l'inhibition mixte reprend les deux effets précédents.

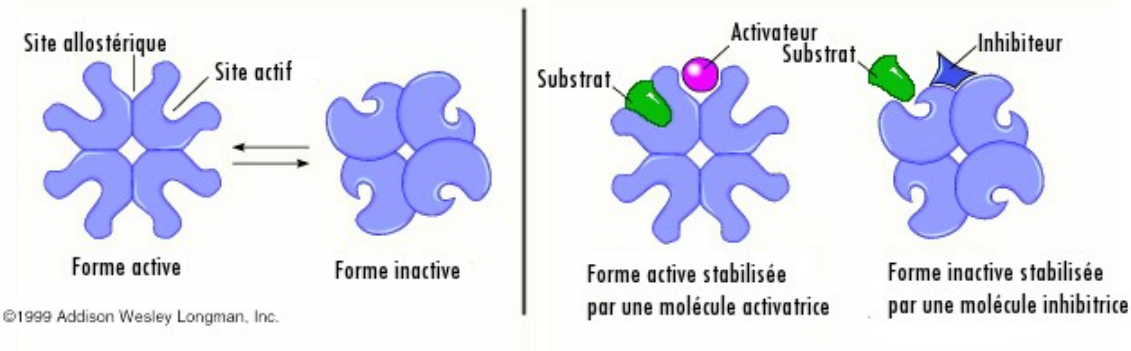


Figure 6 : Résumé du principe d'activation/inhibition par facteur moléculaire.
La forme active/inactive de l'enzyme est stabilisée par la fixation sur le site allostérique d'une molécule activatrice/inhibitrice. La modification de la structure de l'enzyme donne/bloque l'accès du substrat au site actif.

Nous n'avons évoqué ici que le cas des enzymes mono-substrat. Cependant une grande majorité des réactions impliquent plusieurs substrats et donnent plusieurs produits. Ces enzymes multi-substrats sont régies par des règles similaires.

3 Organisation

Le métabolisme d'un organisme est composé d'un nombre élevé de réactions qui dans la plupart des cas dépasse largement le millier. Dans la bactérie *E. coli* on dénombre aujourd'hui plus de 1750 réactions (Keseler et al. 2009). Ces réactions transforment de nombreux métabolites (environ 2000 chez *E. coli* (Keseler et al. 2009) en incluant les lipides et glycans). Comme nous l'avons déjà vu dans la partie

2.1, le métabolisme consiste à modifier successivement ces métabolites, par conséquent un métabolite produit d'une réaction, devient le substrat d'une autre. Il est possible de relier les différents enchaînements de réactions grâce aux métabolites qu'elles ont en communs : on obtient alors un réseau métabolique. Un même métabolite pouvant être substrat de plusieurs réactions, le réseau métabolique s'avère particulièrement dense. Il est de ce fait difficilement représentable, comme l'illustre la Figure 7.

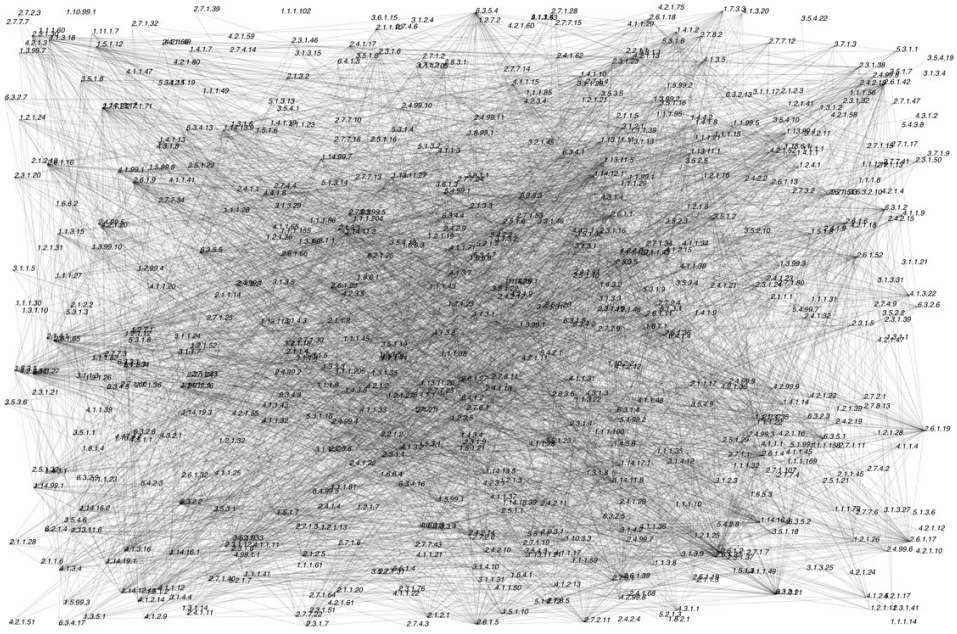


Figure 7 : Sous réseau de 604 réactions.
Les nœuds représentent les réactions et les arêtes les métabolites. D'après les connaissances actuelles, il existe 4062 liens possibles entre les 604 réactions et ce malgré le retrait des métabolites les plus partagés comme l'hydrogène, l'eau et les cofacteurs. (Source http://www.beilstein-institut.de/escec2007/proceedings/Tipton/T ipton.html)

.3.1 Processus et voies métaboliques

Afin de mieux appréhender le fonctionnement du réseau métabolique, il est nécessaire de subdiviser l'ensemble des réactions en fonction de leur objectif. Plusieurs milliers de réactions peuvent avoir lieu au sein d'un organisme et certaines de ces réactions vont partager des métabolites. Grâce à ce partage il est possible d'identifier une série de réactions qui remplit une fonction métabolique déterminée, par exemple passer d'un métabolite précurseur à un métabolite final. Cet enchaînement de réactions est appelé une voie métabolique. Ces voies métaboliques peuvent être classées dans des processus en fonction de leur objectif : dégradation ou synthèse de composés. Nous reviendrons sur ces processus. On peut regrouper ces processus suivant le type des métabolites produits (acide aminé, nucléotide, énergie, etc.). Cette séparation est représentée par la Figure 8. Si les sous-ensembles sont bien définis et bien séparables les uns des autres, il ne faut pas oublier qu'ils sont fortement interconnectés entre eux.

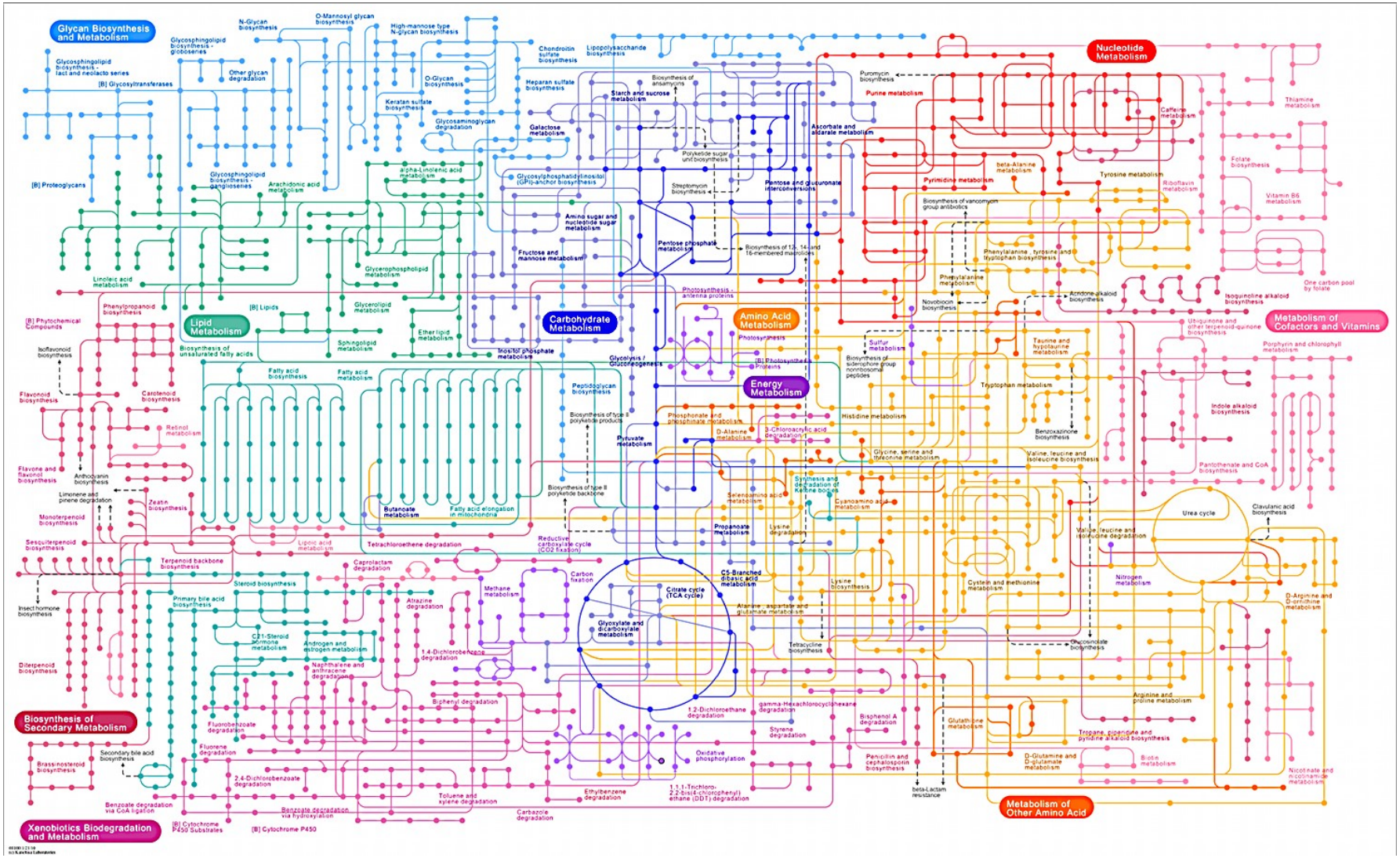


Figure 8 : Réseau métabolique, avec coloration fonctionnelle des voies métaboliques,
Issu de la base de données métabolique KEGG (Kanehisa et al. 2007). (Source http://www.genome.jp/kegg/kegg2.html)

.3.2 Processus métaboliques

De manière générale on désignera par processus métabolique, un ensemble de réactions visant la même fonction (synthèse/dégradation) ou un ensemble de réactions qui s'organisent autour d'un type de métabolite (Table 5). Ces processus métaboliques peuvent être composés d'une ou plusieurs voies métaboliques.

.

Type de métabolite	Description
Carbohydrate (bleu foncé)	Métabolites de la forme C*m*(H_2O)*n*
Energie (violet)	Métabolistes liés à la production d'énergie: Phosphorylation, oxydation, photosynthèse.
Lipide (vert foncé)	Lipides, acides gras et des membranes lipidiques
Nucléotide (rouge)	Métabolite liés aux bases puriques et pyrimidiques
Acide Aminé (orange)	Acides aminés principaux.
Acide Aminé secondaire (orange)	Acides aminés secondaires.
Glycan (bleu clair)	Poly et oligosaccharides
Cofacteurs et vitamines (rose)	Cofacteurs (NAD,ATP etc.) et vitamines (A etc.)
Terpenoïdes et policétides (vert clair)	Métabolites multi-cycliques
Métabolites secondaires (fuchsia)	Métabolites qui ne sont pas essentiels à la vie de l'organisme
Xénobiotique (marron)	Métabolisme des molécules étrangères à la cellule et toxique

Table 5 : Les principaux types de métabolites.
Issu de KEGG (Kanehisa et al. 2007)

Sur la Figure 8, les couleurs correspondent aux différents types de métabolites, bien qu'un effort de disposition soit réalisé, on observe un chevauchement des différentes couleurs dû aux interdépendances très fortes entre les différents processus. Ces processus peuvent être à leur tour divisés en voies métaboliques, c'est à dire une succession de réactions qui permet de passer d'un métabolite précurseur à un métabolite final. Nous avons déjà évoqué la glycolyse qui est une voie de transformation du glucose en pyruvate. Généralement dans une voie métabolique on distingue les métabolites d'intérêts - qui relient les réactions entre elles - des autres métabolites et cofacteurs.
La présence des voies au sein des organismes est très variable. Elle dépend du type d'organisme étudié (plante, animaux, microorganisme, etc.) et de leur mode de vie (autotrophe/auxotrophe, aérobique/anaérobique, parasite/symbionte etc.). A cela vient s'ajouter les différentes variantes d'une voie qui permettent à partir du même précurseur d'obtenir le même métabolite final en passant par des métabolites intermédiaires différents, comme l'illustre la Figure 9.

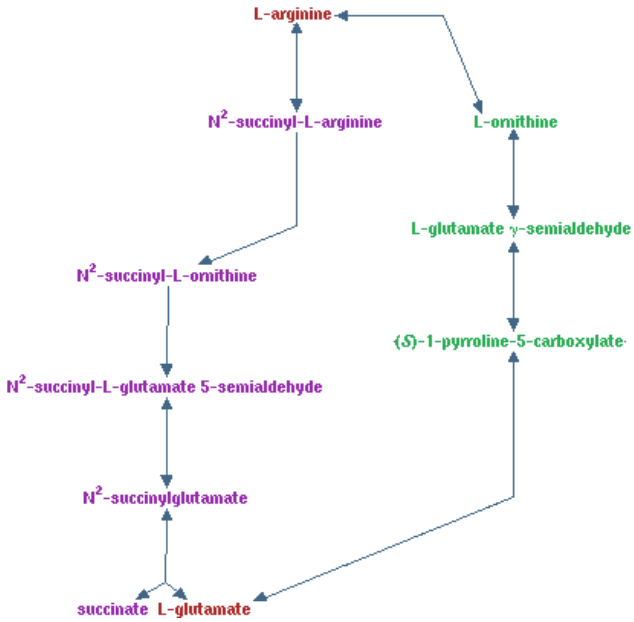


Figure 9 : Voies de dégradation alternatives de l'arginine.
Deux voies permettant de dégrader le L-argine en L-glutamate, soit en passant par les réactions colorées en violet soit par celles colorées en vert.

La présence de différentes voies métaboliques qui effectuent la même conversion de métabolites au sein d'un organisme peut paraître redondante. Il faut garder à l'esprit que l'environnement est en constant changement et l'utilisation de métabolites secondaires différents offre une plasticité et une adaptation rapide de l'organisme vis à vis de ces changements (B L Schneider et al. 1998). Ces voies sont généralement séparées en deux grandes familles : les voies de dégradation qui permettent à partir de métabolites divers d'obtenir des métabolites précurseurs ; et les voies de synthèse qui à partir des métabolites précurseurs fabriquent les métabolites essentiels à l'organisme.

3.2.1 Dégradation

Lorsqu'un ou plusieurs précurseurs indispensables aux voies du métabolisme central sont indisponibles dans l'environnement, l'organisme utilise généralement des voies de dégradation permettant de transformer les composés présents dans l'environnement en ces précurseurs centraux. Ces voies cataboliques sont nombreuses et concernent tous les genres de métabolites. La base de données MetaCyc (Peter D Karp, Monica Riley, et al. 2002) recense plus de 750 voies de dégradation différentes (Table 6).
Certaines de ces voies, comme la glycolyse, sont présentes dans un grand nombre d'organismes : on parle alors du métabolisme central. Il comprend également les voies de fermentation qui génèrent de l'énergie dans des environnements anaérobiques. On inclut également le cycle de Krebs qui se situe généralement après la glycolyse et produit de nombreux métabolites essentiels. D'autres voies de dégradation sont spécifiques de certains organismes, ce qui leur confère une niche exclusive.

Classe du métabolite dégradé	Nombre de voies	Classe du métabolite dégradé	Nombre de voies
Composés aromatiques	166	Autres	25
Acide aminés	118	Nucléosides et Nucléotides	25
Nutriments inorganiques	90	Hormones	24
Métabolites secondaires	83	Acide gras et lipides	20
Carbohydrates	58	Alcool	16
Amines et polyamines	48	Aldéhydes	12
Carboxylases	40	Composés polymériques	10
Composés chlorés	39	Cofacteurs	3
Composés en C1	26	Protéines	3

Table 6: Nombre de voies métaboliques en fonction du genre du métabolite dégradé.
Données issues de (Peter D Karp, Monica Riley, et al. 2002).

3.2.2 Biosynthèse

Les organismes ont pratiquement tous besoin des mêmes métabolites primaires pour vivre, comme nous l'avons expliqué dans la partie 2.1, et il est rare que tous ces métabolites soient naturellement disponibles dans l'environnement. Des voies métaboliques assurent la synthèse de ces métabolites essentiels manquants à partir de précurseurs centraux. Toujours dans la partie 2.1 nous avons évoqué la diversité des métabolites secondaires, qui se traduit par un nombre important de voies anaboliques : il dépasse le millier dans MetaCyc (Peter D Karp, Monica Riley, et al. 2002) Table 7 ; ce nombre est cependant fortement sous évalué puisque très peu de données sur les plantes sont à ce jour intégrées à cette base.

Classe du métabolite dégradé	Nombre de voies	Classe du métabolite dégradé	Nombre de voies
Métabolites secondaires	426	Amines et polyamines	36
Cofacteurs	179	Nucléosides et nucléotides	30
Acides gras et lipides	115	Autres	30
Acides aminés	109	Composés aromatiques	24
Carbohydrates	87	Sidérophores	17
Structures cellulaires	45	Régulations	5
Hormones	43	Aminoacyl t-ARN	4

Table 7: Nombre de voies métaboliques en fonction du genre du métabolite synthétisé.
Données issues de (Peter D Karp, Monica Riley, et al. 2002).

De même que pour les voies de dégradation, deux grandes catégories de voies de biosynthèse se distinguent : celles présentes chez pratiquement tous les organismes et celles, plus éparses, présentes dans quelques organismes spécifiques. Les voies les plus communes synthétisent généralement des métabolites essentiels à la vie de l'organisme, que ce soit des lipides (constituants essentiels des membranes), des nucléotides (essentiels pour la synthèse d'ADN/ARN) ou encore des acides aminés (pour la synthèse de protéines). Les voies moins conservées synthétisent généralement des molécules conférant un avantage particulier dans l'environnement de l'organisme : des mécanismes d'attaque (antibiotique) ou de défense biochimique sécrétion de molécule modifiant l'environnement afin de le rendre moins hostile.
La diversité métabolique des voies de synthèse et de dégradation permet d'effectuer des actions précises. Cette spécificité offre des comportements différents vis à vis de métabolites identiques. Reprenons la pénicilline déjà évoquée : ce champignon est mortel pour certaines bactéries, cependant il n'est pas dangereux pour l'homme.

C’est pourquoi cette famille de champignons mortels pour les bactéries est utilisée dans notre alimentation notamment pour la production de fromages bleus (Figure 10).

Figure 10: *Penicillium* sur boîte de culture (point blanc) à droite et dans un fromage (points bleus) à gauche.
Sur la culture on voit la plage de lyse sur la population de *Staphylocoque* (traits blancs) qui sont incapables de vivre en présence de *Penicillium* (Photo de C.L. Case, Skyline College).

Comme pour les voies de dégradation, il existe de nombreuses variantes des voies de biosynthèse, qui permettent une adaptation rapide en fonction des variations de l’environnement.

4 Le métabolisme et la biodiversité

Il est pour le moment impossible d’estimer précisément la diversité métabolique : on considère qu’il existe plus d’un millier de métabolites qui interviennent dans le métabolisme primaire, et ce nombre explose dans le métabolisme secondaire, puisqu’on suppose qu’il met en jeu plusieurs centaines de milliers de métabolites (Villas-Bas et al. 2007). Ces métabolites secondaires ont souvent une structure complexe et ont un rôle très précis : on y retrouve les antibiotiques comme la pénicilline, des hormones et de multiples drogues. Ces métabolites sont les produits finaux de voies de synthèses encore inconnues, qui se sont mises en places au cours de l’évolution, et qui ont conféré ou qui confèrent toujours un avantage sélectif. Ces avantages peuvent être de nature diverse. Ces voies métaboliques peuvent améliorer les capacités de production de la biomasse. Elles peuvent conférer une protection contre les métaboliques toxiques, ou au contraire produire des toxines pour supprimer les organismes en compétition pour les ressources disponibles dans l’environnement. La diversité des capacités métaboliques permet aussi de s’adapter plus facilement aux modifications environnementales et aux conditions extrêmes.

.4.1 Plasticité, évolution et adaptation du métabolisme

Le métabolisme possède une certaine variabilité et redondance, qui lui permettent de s’adapter à la constante évolution des ressources disponibles dans l’environnement. La flexibilité du métabolisme peut s’expliquer par des voies parallèles (Figure 9) ou au sein d’une même voie grâce à des réactions qui diffèrent par leurs métabolites

secondaires (Figure 11). La redondance peut être assurée au niveau de la réaction grâce aux isozymes : ce sont des enzymes différentes mais qui effectuent la même transformation biochimique.

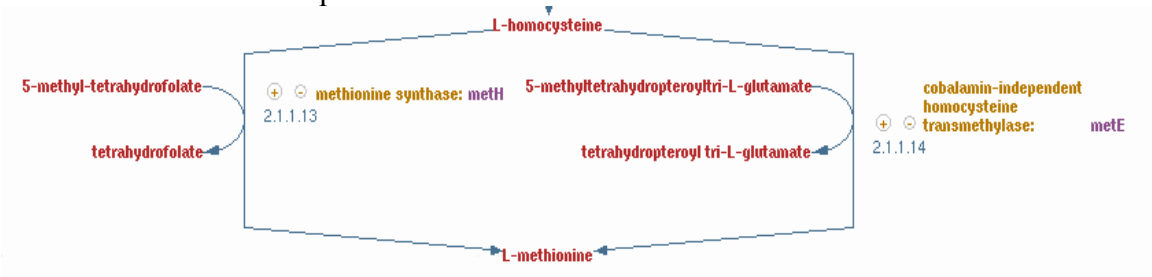


Figure 11 : 2 Réactions équivalentes.
Deux réactions codées par des gènes différents interviennent dans la voie de biosynthèse de l'homosérine et la méthionine. Les 2 réactions ont le même substrat d'intérêt la L-homocystéine et le même produit d'intérêt la L-méthionine. La différence se fait au niveau des couples de cofacteurs : à gauche le couple 5-mTHF/THF et à droite le couple 5-mTHF-glutamate/THF-glutamate.

A l'origine les organismes possédaient un ensemble de gènes restreint par rapport à ce que nous observons aujourd'hui (Jensen 1976). C'est à partir de cet ensemble réduit que se sont mis en place des réactions avec les métabolites d'intérêt identiques, et des voies métaboliques alternatives. Cette capacité à utiliser différents métabolites pour effectuer la même fonction est appelée la plasticité métabolique.
Le métabolisme est transformé par l'évolution à l'aide de deux grands types d'évènements : (1) des modifications apportées à l'activité même des enzymes et (2) des modifications apportées à l'organisation des enzymes en voies métaboliques.

4.1.1 Modification structurelle de l'enzyme

Les mutations du gène codant une enzyme peuvent modifier l'efficacité, la spécificité et la fonction de l'enzyme.
La modification de l'efficacité d'une enzyme agit uniquement sur la cinétique de la réaction qu'elle catalyse. Ce type de modification ne modifie en rien l'équation bilan de la réaction.

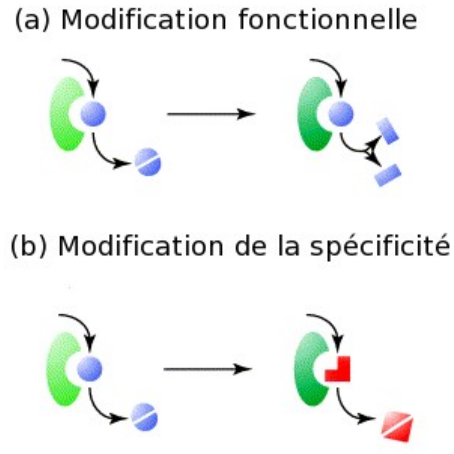


Figure 12 : Evolution structurelle des enzymes.
Les modifications des enzymes peuvent être (a) fonctionnelles : à partir d'un même substrat, le produit sera différent ou (b) porter sur sa spécificité : le substrat est différent mais la fonction réalisée par l'enzyme reste la même (figure issue de (Schmidt et al. 2003)).

Des simulations informatiques ont montré qu'à partir d'un ensemble réduit d'enzymes multifonctions très généralistes, qui peuvent soit se spécialiser en ce scindant, soit se dupliquer, il est possible d'arriver à un réseau métabolique avec des métabolites « hubs » et un grand nombre d'enzymes hautement spécialisées (T. Pfeiffer et al. 2005). Après duplication, l'une des enzymes va garder son caractère générique, offrent plus de liberté au duplicata en levant des contraintes sélectives ce qui rend possible sa spécialisation. Une fois l'enzyme spécialisée, il devient difficile de la faire évoluer ce qui explique le grand nombre d'enzymes spécialisées dans la simulation. Ces enzymes spécialisées peuvent quand même évoluer (Schmidt et al. 2003) soit en modifiant les produits de la réaction alors même que les substrats restent identiques (Figure 12a), ou alors en modifiant la spécificité du substrat (Figure 12b). Le changement de substrat peut simplement concerner des cofacteurs ou des métabolites secondaires, mais également les substrats principaux.

4.1.2 Evolution des voies métaboliques

Les modifications de spécificité ou de fonction enzymatique ne constituent qu'une partie des processus permettant l'acquisition de nouvelles capacités métaboliques. L'apparition ou la modification de réactions entrainent des changements des voies métaboliques qui sont des suites de réactions. Ces nouvelles voies métaboliques peuvent être dues à des modifications au sein de l'organisme ou grâce à l'incorporation d'éléments externes.
Par éléments externes on entend un matériel génétique provenant d'un autre organisme qui va être assimilé puis utilisé. On désigne cet événement par le terme de transfert horizontal. Le nouveau matériel génétique peut contenir des gènes permettant d'effectuer des transformations biochimiques auparavant impossibles (Pál, Papp & Lercher 2005a).
Les modifications des voies internes à l'organisme peuvent être divisées en cinq types comme l'illustre la Figure 13.

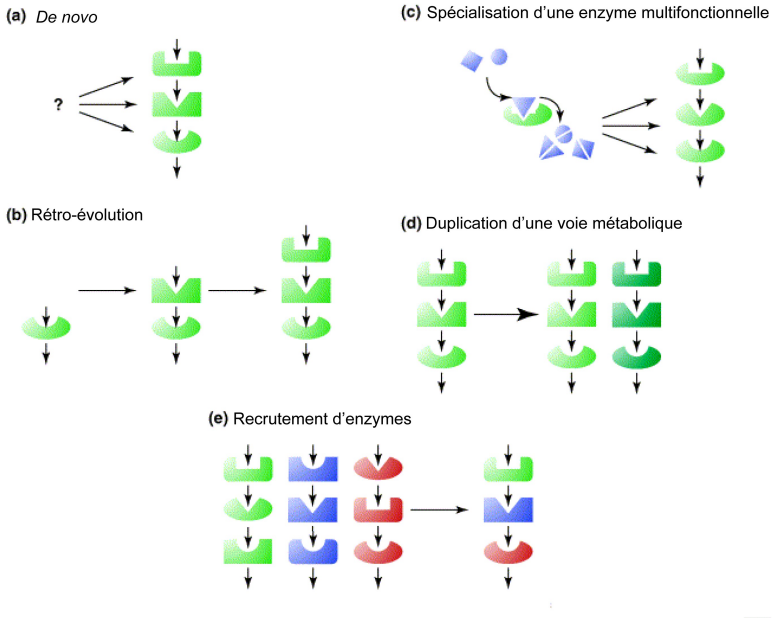


Figure 13: Les différents scénarii d'évolution des voies métaboliques.
Figure issue de (Schmidt et al. 2003)

La création *de novo* d'une voie métabolique (Figure 13a) reflète l'absence de connaissance ; on suppose qu'à un moment donné des enzymes indépendantes ont permis un enchaînement de réactions biochimiques d'un substrat à un produit final. La rétro-évolution (Figure 13b) consiste en un recrutement d'enzymes étape par étape en partant du produit final vers un métabolite initial. Un recrutement d'enzyme en partant du métabolite initial est également possible. La fragmentation d'une enzyme multifonctionnelle en autant d'enzymes que de fonctions (Figure 13c) est aussi une source de nouvelles voies métaboliques. La duplication d'un segment d'ADN peut entraîner la duplication d'une voie métabolique (Figure 13d). Ce dédoublement permet de conserver la fonction d'origine avec l'une des voies et d'explorer des nouvelles possibilités métaboliques avec l'autre. Enfin le dernier mode d'évolution est le recrutement d'enzyme ou « patchwork » (Figure 13e) : des enzymes de différentes voies métaboliques vont être utilisées successivement pour former une nouvelle voie métabolique. L'évolution par recrutement rappelle que les voies métaboliques sont fortement entrelacées et interdépendantes d'où certaines difficultés –sur lesquelles nous reviendrons – lors de la schématisation et modélisation du réseau métabolique.

Nous avons vu comment peuvent apparaître de nouvelles voies ou enzymes, cependant le métabolisme peut aussi perdre des fonctionnalités si celles-ci ne sont plus requises.

Le métabolisme : in silico

La modélisation utilise les mathématiques pour représenter de façon abstraite la réalité. Cette façon de raisonner et de concevoir l'objet de l'analyse de manière abstraite n'est pas triviale et nécessite une approche différente de celle de la biologie expérimentale. Pour mieux comprendre la modélisation et les modèles du métabolisme, nous allons nous attarder sur la notion générale de modèle pour ensuite expliquer les différentes façons de modéliser les sous-systèmes du métabolisme ainsi que le réseau métabolique dans sa totalité.

1 Notion de modèle

D'après Georges E. P. Box, statisticien du début du XX^{ème} siècle *« tous les modèles sont faux, mais certains sont utiles ; la véritable question est de savoir à partir de quel approximation ne sont ils plus utilisables ?*» (Box & Draper 1987). Dans cette citation se trouvent deux fondements de la modélisation :
1- Un modèle est une approximation de la réalité et par conséquent, il est et sera, toujours incomplet.
2- Un modèle est basé sur des hypothèses ; il faut trouver des hypothèses suffisamment simples et pertinentes pour que les résultats soient exploitables.
Avant de débuter la conception d'un modèle, il est crucial de bien définir ce que l'on souhaite modéliser. Cela peut paraître simple ; cependant c'est un exercice délicat dont dépend tout le processus de modélisation. Reprenons la réaction *glucose-6-phosphatase* (). Nous pouvons l'étudier de différentes façons et chacune de ces façons peut donner lieu à une modélisation différente. On peut se focaliser sur sa fonction *phosphatase* qui consiste à dissocier un groupement phosphate d'une molécule. On peut s'intéresser aux métabolites qui interviennent dans la réaction (c'est le cas de l'équation bilan). On peut envisager de suivre le devenir de chacun des atomes des métabolites. Il est aussi possible de travailler sur la dynamique de la réaction, etc. Une fois que l'objet de l'étude est bien défini on peut choisir un type de modèle.
L'une des principales difficultés lors de la conception du modèle est de formuler les hypothèses qui vont le régir. Celles-ci doivent être, à la fois simples pour être exploitables et suffisamment détaillées pour avoir un sens vis à vis de la réalité. Il en résulte que pour une même réalité, il existe autant, voire plus, de modèles que d'hypothèses.
Le dernier point important est l'interprétation des résultats d'une simulation d'un modèle. Les résultats obtenus sont toujours à mettre en relation avec les hypothèses qui ont permis la construction du modèle. Dans le cas où les simulations sont cohérentes avec des résultats expérimentaux, il faut en déduire que les hypothèses du modèle peuvent expliquer l'objet de l'étude et non pas que les hypothèses sont la réalité.

L'introduction de la modélisation en biologie est récente comparée à des domaines comme la physique où certains principes remontent à l'antiquité. Les premières applications de la modélisation en biologie étaient dans le domaine de l'écologie et des populations. Ceci s'explique par le fait que ce sont des caractères observables à l'œil nu contrairement à la biologie moléculaire et à la génétique. Suite aux travaux

de Gregor Mendel en 1865 sur l'observation de la morphologie des pois (Edelson 2001) l'un des premiers modèles mathématiques et biologiques a vue le jour en 1908 : le principe de Hardy-Weinberg (Hardy 1908). Il considère un gène avec deux formes alléliques *a* et *A* avec une fréquence respective *p* et *q*. A partir de ces fréquences on peut écrire la table de croisement :

	A (p)	*a (q)*
A (p)	*AA* p^2	*Aa (pq)*
a (q)	*Aa (pq)*	*Aa* (p^2)

Table 8: Table de croisement des formes alléliques.
Il en ressort les fréquences suivantes, $f(AA)=p^2$, $f(aa)=q^2$ et $f(Aa)=2pq$.

La loi de distribution des formes alléliques qui en découle est $p^2+q^2+2pq=1$. A partir de cette équation on peut prédire l'évolution des fréquences des différentes formes allèles.

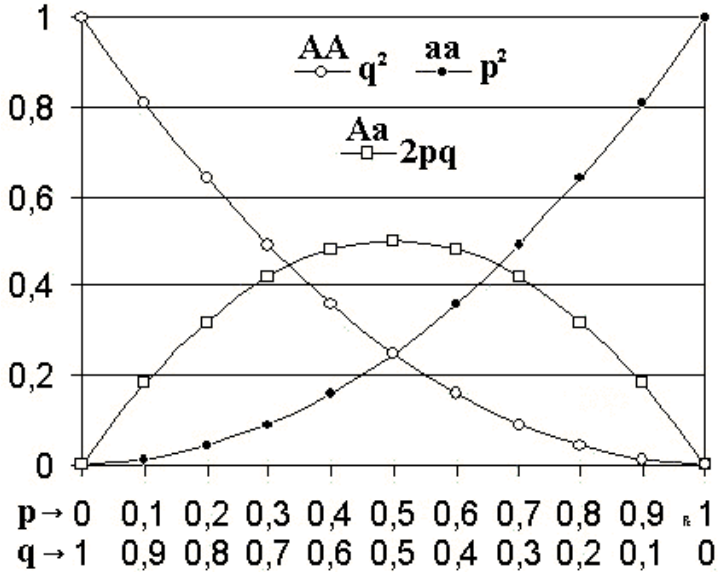


Figure 14: Evolution des proportions des différents allèles d'après le principe de Hardy-Weinberg.
Suivant le ratio *p/q* une forme peut devenir majoritaire ou les deux formes peuvent s'équilibrer

Ce modèle a mis en évidence que dans le cas où un gène possède deux formes alléliques et qu'aucun événement de recombinaison ou mutation n'intervient, si une des deux formes est plus fréquente que l'autre alors cette forme restera la seule présente sur le long terme (Figure 14).

Si l'utilisation tardive des modèles en biologie est principalement dûe au caractère microscopique des acteurs biologiques et à leur récente découverte, les nouvelles technologies permettent d'identifier et de donner un rôle à de plus en plus d'acteurs. Ceci permet d'en déduire de nombreux processus nouveaux. Cependant la complexité des mécanismes biologiques rend l'isolation et la modélisation des processus très délicates. C'est pourquoi une partie de la biologie est consacrée à l'étude des systèmes et des interactions. Cette discipline en plein essor depuis les années 2000 est appelée biologie des systèmes (Hiroaki Kitano 2002). Elle combine à la fois la biologie moléculaire et cellulaire en les associant à la physique, la chimie et aux mathématiques. Les travaux réalisés durant ma thèse s'inscrivent dans cette nouvelle discipline.

2 Les modélisations du métabolisme

Appliquée au métabolisme la biologie des systèmes permet de travailler à différents niveaux, allant de la réaction en passant par la voie métabolique et jusqu'au niveau du métabolisme dans sa globalité. Il existe une large gamme de modèles adaptés au métabolisme et comme nous l'avons dit, le bon modèle dépendra des questions posées et des connaissances disponibles. Ces modèles peuvent être classés en trois grandes familles en fonction de la dynamique, des paramètres et de la stœchiométrie (Figure 15) (Jörg Stelling 2004).

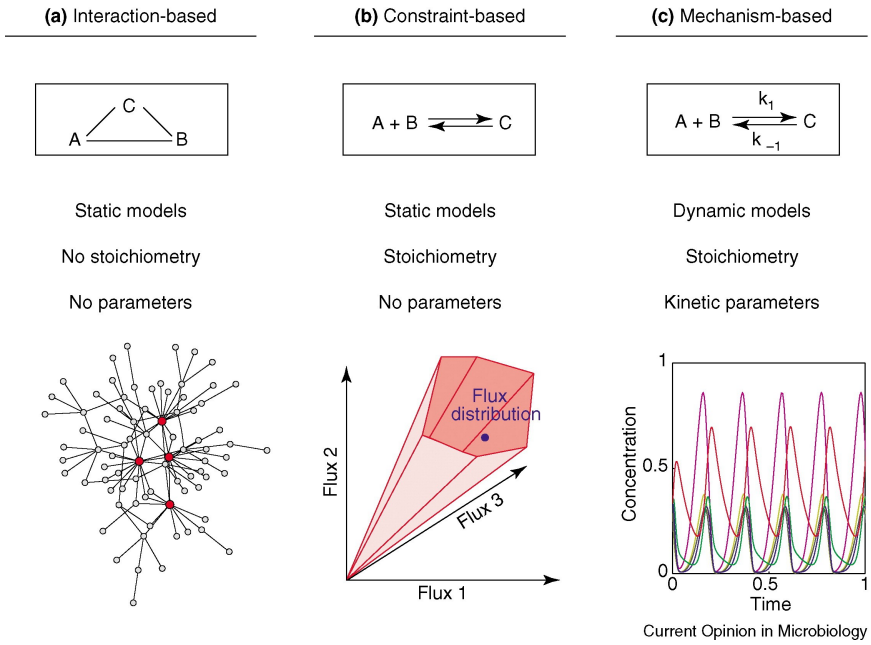


Figure 15: Les différents types de modèles appliqués au métabolisme.
(a) Interaction entre les différents acteurs (b) Contraintes qui utilisent la topologie du réseau, la stœchiométrie et la réversibilité des réactions. (c) Mécanisme détaillée des réactions. Les schémas en bas illustrent les résultats obtenus. (a) représentation de *hubs* (en rouge) dans un réseau d'interaction *free-scale*. (b) Le cône des distributions de flux admissibles. (c) Evolution des concentrations au cours du temps. Figure issue de (Jörg Stelling 2004).

.2.1 Les graphes métaboliques

L'un des moyens les plus simples et plus rapides de modéliser le réseau métabolique est l'objet mathématique graphe. L'utilisation des graphes permet de profiter de la théorie des graphes qui est développée en mathématiques depuis le XVIII^{ème} siècle et également d'effectuer des analyses sur la topologie et d'autres propriétés structurelles du réseau.

2.1.1 Les différents types de graphes

Un graphe est une structure mathématique composée de nœuds reliés par des arêtes. De façon plus rigoureuse un graphe G est un ensemble de couple (E, V) où E est l'ensemble des arêtes et V l'ensemble des nœuds. Il existe trois principaux types de graphes (Figure 16) qui sont utilisés pour modéliser le métabolisme et le choix du graphe dépendra des questions et des phénomènes étudiés.

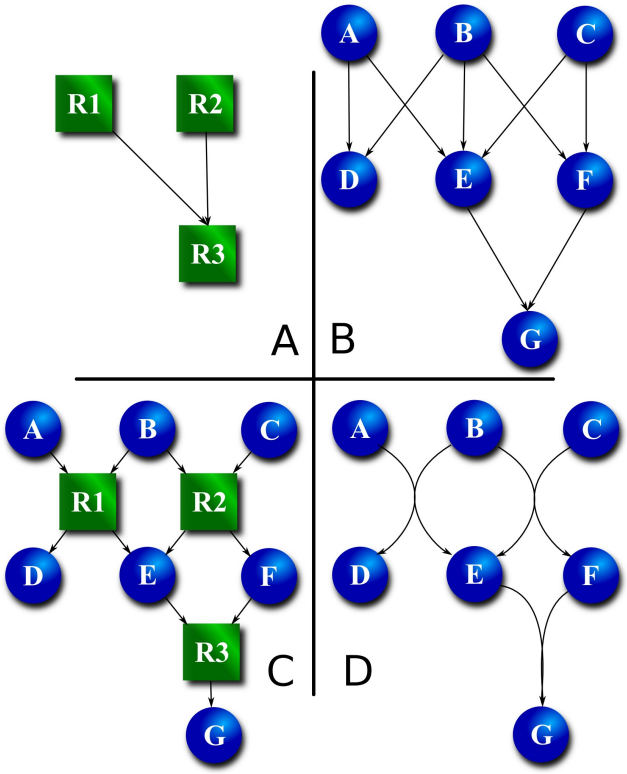


Figure 16 : Les principaux types de graphes appliqués au métabolisme.
A) Graphe des interactions entre réactions, les nœuds sont des réactions et deux réactions partageant un métabolite sont reliées par une arête. B) Graphe des interactions entre les métabolites, les nœuds sont des métabolites et deux métabolites partageant une réaction sont reliées par une arête. C) Graphe biparti avec en bleu les métabolites et en vert les réactions, un métabolite ne peut être relié qu'à des réactions et inversement. D) Hypergraphe, les nœuds sont des métabolites et les arêtes sont des réactions, une arête peut relier plus de deux nœuds.

Les graphes simples s'intéressent à un seul acteur métabolique à la fois : il peut s'agir des réactions, des enzymes ou des métabolites. Le principe de ces graphes est de lier des acteurs partageant une caractéristique, dans le cas d'un graphe de réactions, deux réactions sont liées par une arête si le métabolite substrat d'une réaction est produit de l'autre réaction. Dans la Figure 16A la réaction *R1* produit le métabolite *E* qui est consommé par la réaction *R3* d'où un lien entre *R1* et *R3*. De même on peut faire le graphe métabolite. Le métabolite *D* est produit par la réaction *R1* à partir du métabolite *A*, d'où le lien *(A,B)* sur la Figure 16B. On peut utiliser un graphe biparti, c'est à dire un graphe où les nœuds peuvent appartenir à deux ensembles distincts *U* et *V* tel que une arête ait une extrémité dans *U* et l'autre dans *V*. Appliquer au métabolisme l'ensemble *U* peut représenter l'ensemble des réactions et l'ensemble *V*, l'ensemble des métabolites. La Figure 16C représente un graphe biparti. La comparaison entre un graphe biparti et les graphes précédents montre un gain d'information sur les relations entre métabolites et réactions (Figure 17).

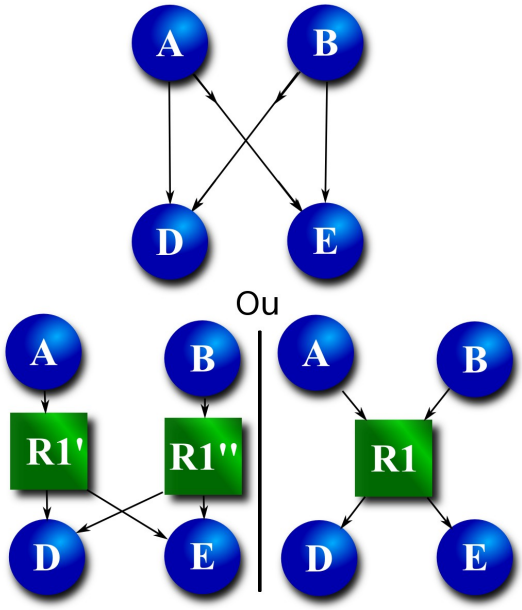


Figure 17 : Différence d'information entre les types de graphes.
A partir du sous graphe métabolite (Figure 16B), on ne peut pas différencier les deux graphes bipartis.

Le graphe métabolique biparti permet de conserver la relation entre les substrats, la réaction et les produits. On peut facilement passer d'un graphe biparti à un graphe de réactions ou de métabolites, mais l'inverse est impossible. En effet conserver un seul type d'acteur métabolique est une simplification avec perte définitive de connaissances. Le dernier type de graphe est l'hypergraphe noté H. D'un point de vue mathématique H est un couple d'arêtes et de nœuds (V, E) où $V=\{v_1, v_2,\ldots,v_m\}$ et $E=\{e_1,e_2,\ldots,e_m\}$ avec $E_i \subseteq V$, pour $i=1,2,\ldots,m$. Autrement dit une arête peut relier plus de deux nœuds (Figure 16D). Usuellement les nœuds sont les métabolites et les arêtes les réactions. Bien que l'hypergraphe et le graphe biparti soient différents les informations contenues sont sensiblement identiques, ainsi on retrouve les liens entre les substrats, les produits et les réactions.

Tous les graphes représentés sont dirigés, ce qui revient à considérer que les réactions sont irréversibles. Il existe deux manières de représenter une réaction réversible : en utilisant des arêtes non orientées au risque de poser des ambigüités sur les couples de substrats et produit, ou en dupliquant la réaction en deux réactions, une dirigée dans un sens et une autre dans l'autre sens. Encore une fois le choix de l'une ou l'autre méthode dépend des objectifs de la modélisation.

2.1.2 Les analyses

L'analyse des graphes métaboliques n'est pas utilisée dans ce mémoire mais nous en donnons les principaux points. Les graphes métaboliques sont des graphes dans lesquels les arêtes et les nœuds représentent des phénomènes biologiques ; on peut les analyser avec les mesures usuelles des graphes : la distance, la centralité, le degré, le diamètre et des coefficients d'agglomérations. Ces mesures servent à évaluer l'homogénéité du réseau métabolique, repérer des nœuds avec beaucoup d'interactions. Il est également possible de s'intéresser à la topologie et organisation du réseau. Ainsi on a observé que le métabolisme pouvait être représenté par un graphe qui suit une loi de puissance (Albert-László Barabási & Albert 1999). Ce qui signifie qu'un faible nombre de nœuds sera très connecté et la majorité des nœuds

seront faiblement connectés. Cette structure particulière permet des décompositions en modules et en organisations hiérarchiques (Erzsébet Ravasz 2009; E Ravasz et al. 2002). La Figure 18 reprend les principales structures topologiques observables dans le métabolisme.

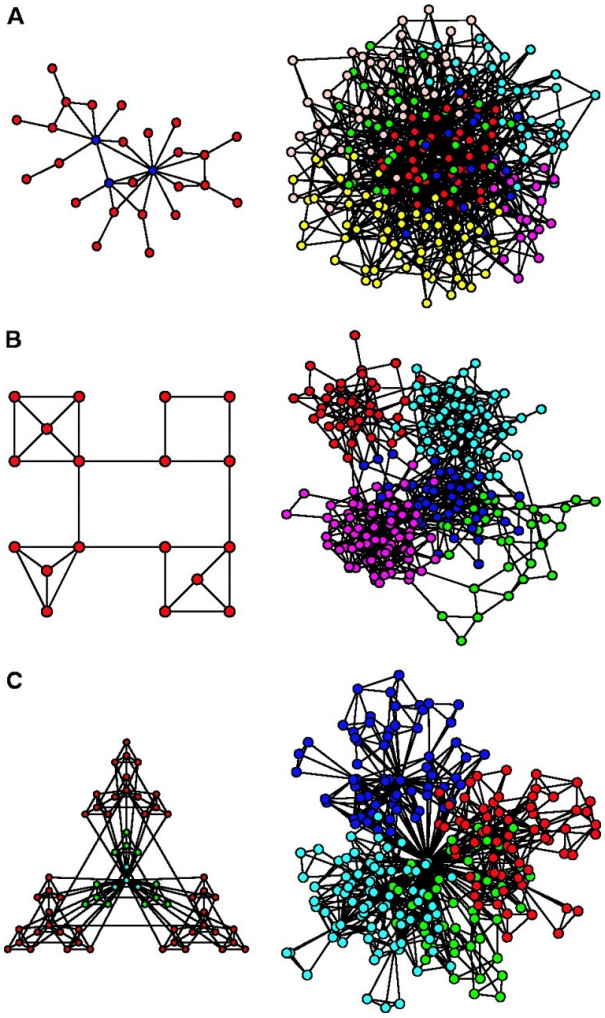


Figure 18 : Structures des graphes métaboliques.
Structures des graphes métaboliques. (A) Illustration schématique (gauche) d'un réseau *free-scale*, dont le degré de distribution suit une loi de puissance. Dans ce cas quelques nœuds hautement connectés ou hubs (en bleu), jouent un rôle important pour relier l'ensemble des réseaux. A droite un réseau suivant les mêmes règles mais avec 256 nœuds on ne distingue aucun module. (B) Schématisation d'un réseau modulaire (gauche), on distingue quatre modules dans lesquels les nœuds sont fortement reliés et reliés entre eux par de rares liens. A droite un réseau de 256 nœuds représentant quatre modules. (C) Réseau *free-scale* hiérarchique et modulaire (à gauche). La hiérarchie est représentée de façon croissante, bleu, vert et rouge. A droite le même type de réseau mais pour 256 nœuds. Figure issus de (E Ravasz et al. 2002).

L'une des difficultés est d'assigner un rôle biologique à une propriété topologique. Si les modules peuvent être reliées aux voies métaboliques (Erzsébet Ravasz 2009) et à des aspects de robustesse du réseau, il semble que les hypothèses posées dans un premier temps sont partiellement erronées (Lima-Mendez & Jacques van Helden 2009).

.2.2 Le modèle cinétique

Les graphes mettent en relation les métabolites et les réactions d'une manière statique et ne prennent pas en compte la stœchiométrie. Une autre approche de la

modélisation du métabolisme est l'utilisation de modèles cinétiques. Leur objectif est de décrire les mécanismes et la vitesse de conversion des métabolites par les enzymes. Ces modèles apportent une dimension temporelle puisqu'ils décrivent l'évolution des concentrations au cours du temps.

2.2.1 Cinétique des réactions enzymatiques

Michaelis et Menten ont proposé une équation cinétique simple capable d'expliquer le fonctionnement d'une enzyme (Michaelis & Menten 1913). Cette équation empirique établit une relation entre la concentration en substrat [S] et les principales caractéristiques de l'enzyme : sa vitesse maximale v_{max} et son affinité avec le substrat K_m :

$$v([S]) = v_{max} \cdot \frac{[S]}{K_m + [S]} \quad (1)$$

Lors de conditions optimales, à concentration d'enzyme fixée et en présence d'un excès de substrat, la réaction atteint sa vitesse maximale v_{max}. Ce paramètre est spécifique d'une enzyme, deux isozymes n'auront pas la même v_{max}. De plus son ordre de grandeur varie de façon importante entre différentes enzymes, certaines convertissent ~0.5 molécule par seconde tandis que d'autres dépassent les 500 000 molécules par seconde (Stephanopoulos et al. 1998). Le second paramètre est la constante d'affinité ou constante de Michaelis K_m. Elle représente la concentration de substrat lorsque la vitesse instantanée est égale à la moitié de la vitesse maximale : $v(S)={}^1/_2 v_{max}$ (voir Figure 19).
La courbe de la Figure 19 peut être décomposée en une partie à forte augmentation et une partie asymptotique. A faible concentration en substrat toutes les enzymes ne sont pas utilisées et les substrats nouvellement ajoutés vont être immédiatement modifiés par les enzymes libres, ce qui augmente la vitesse de la réaction. A forte concentration en substrats, pratiquement toutes les enzymes sont utilisées et la vitesse de la réaction a atteint son maximum.

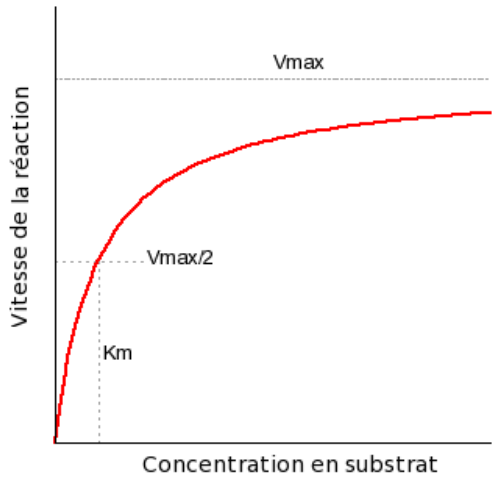


Figure 19: Vitesse d'une réaction enzymatique.
Elle est une fonction de la concentration de substrat pour une concentration en enzyme fixe.

L'équation cinétique de Michaelis-Menten est la simplification et la concaténation des différentes étapes décrites dans la Figure 5. En considérant que la libération du

produit après transformation est immédiate et irréversible, on obtient l 'équation suivante (Briggs George Edward & Haldane John Burdon Sanderson 1925):

$$E + S \underset{k_{-1}}{\overset{k_1}{\leftrightarrow}} ES \xrightarrow{k_2} E + P \quad (2)$$

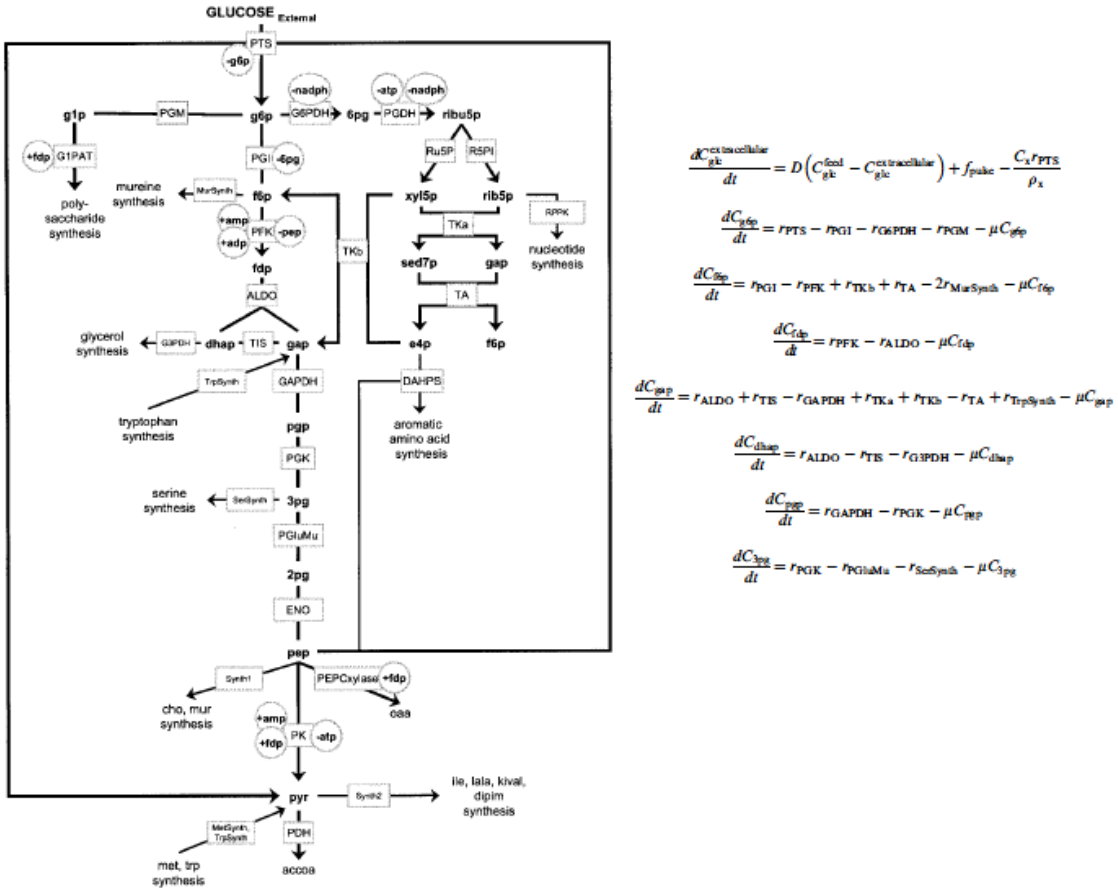
En supposant que la réaction est à l'état stationnaire ([ES] constant) et en introduisant la concentration totale en enzyme E_{tot}, qui est la somme des enzymes libres E et complexées ES, il est possible d'estimer les paramètres de l'équation de Michaelis-Menten :

$$v_{max} = k_2.[E_{tot}] \quad (3) \text{ et } K_m = \frac{k_{-1} + k_2}{k_1} \quad (4).$$

L'équation de Michaelis-Menten peut être adaptée au cas des enzymes multi-substrats et elle peut prendre en compte des effets d'activation ou d'inhibition et même des effets allostériques. De cette palette d'équations, on peut décrire l'évolution des concentrations des différents métabolites au court du temps et ainsi estimer la vitesse des réactions et de consommation/production de métabolites. Pour calculer ces vitesses on modélise chacune des réactions par une équation différentielle qui sera une fonction des concentrations et du temps, et qui dépendra des paramètres cinétiques de l'enzyme (v_{max}, k_m, etc.). On peut aussi utiliser des approches stochastiques (Gillespie 2007) pour simuler le comportement d'un ensemble de réactions. La réaction est décomposée en étapes élémentaires, comme la fixation du substrat sur l'enzyme. A chacune de ces étapes est associée une probabilité et à chaque pas de temps une des étapes va être tirée au sort.

De tels modèles ne se limitent pas à une étude temporelle des variations de concentrations. Ils permettent des analyses plus théoriques sur les états d'équilibres, les bifurcations, ainsi que sur la robustesse et la sensibilité du système (Di Ventura et al. 2006).

Ce cadre de modélisation n'est pas appliqué qu'au métabolisme mais également à la modélisation de la régulation de l'expression des gènes et de leurs produits. Ainsi certains modèles vont s'intéresser à un seul gène et sa régulation (Gonze et al. 2003), tandis que d'autres peuvent porter sur un ensemble de réactions comme le métabolisme central d'*Escherichia coli* (Figure 20) (Chassagnole et al. 2002).



$$\frac{dC_{glc}^{extracellular}}{dt} = D\left(C_{glc}^{feed} - C_{glc}^{extracellular}\right) + f_{pulse} - \frac{C_x r_{PTS}}{\rho_x}$$

$$\frac{dC_{g6p}}{dt} = r_{PTS} - r_{PGI} - r_{G6PDH} - r_{PGM} - \mu C_{g6p}$$

$$\frac{dC_{f6p}}{dt} = r_{PGI} - r_{PFK} + r_{TKb} + r_{TA} - 2r_{MurSynth} - \mu C_{f6p}$$

$$\frac{dC_{fdp}}{dt} = r_{PFK} - r_{ALDO} - \mu C_{fdp}$$

$$\frac{dC_{gap}}{dt} = r_{ALDO} + r_{TIS} - r_{GAPDH} + r_{TKa} + r_{TKb} - r_{TA} + r_{TrpSynth} - \mu C_{gap}$$

$$\frac{dC_{dhap}}{dt} = r_{ALDO} - r_{TIS} - r_{G3PDH} - \mu C_{dhap}$$

$$\frac{dC_{pgp}}{dt} = r_{GAPDH} - r_{PGK} - \mu C_{pgp}$$

$$\frac{dC_{3pg}}{dt} = r_{PGK} - r_{PGluMu} - r_{SerSynth} - \mu C_{3pg}$$

Figure 20: Modèle cinétique du métabolisme central chez *E. coli.*
A gauche la vue schématique du modèle à droite une partie des équations différentielles du modèle

La complexité de ce genre de modèle nécessite l'utilisation de l'informatique pour résoudre le système d'équation, et des modèles plus évolués peuvent dépasser les capacités de calcul des machines actuelles. Une autre limitation de ce genre de modèles est l'estimation des paramètres. Dans le cas des enzymes connues et synthétisables, on peut estimer la constante d'affinité et la vitesse maximale, est décrire leur cinétique. Ce n'est malheureusement pas le cas de la majorité des enzymes et des paramètres. Il faut donc tenir compte des incertitudes sur les paramètres et les équations cinétiques, auxquelles s'ajoutent des approximations engendrées par les hypothèses nécessaires à la construction du modèle.

.2.3 Le contrôle métabolique

Il est possible de quantifier les effets et les dépendances entre les différents acteurs (enzymes, métabolites ou flux des réactions) qui interviennent dans un modèle cinétique. Lorsque le système est à l'état stationnaire il est possible de réaliser une analyse du contrôle métabolique (Heinrich & Rapoport 1974; Kacser & Burns 1973). Elle permet d'estimer un coefficient de contrôle qui représente l'influence de la variation d'un des acteurs sur l'ensemble des acteurs ainsi que sur le comportement global du système. Ce genre d'analyse appliqué à des voies métaboliques, a remis en cause l'hypothèse selon laquelle c'est l'étape limitante qui régit le système. Bien que certains acteurs aient un coefficient de contrôle plus élevé que d'autres, c'est l'ensemble des différents acteurs d'une voie métabolique qui vont, au final, contrôler le flux de celle-ci (D A Fell 1992). De plus la modification de la concentration d'une enzyme aura un effet sur son coefficient de contrôle mais

également sur l'ensemble des coefficients de contrôle. Cela s'explique par le fait que peu importe le réseau modélisé, la somme des coefficients de contrôle est égale à 1. Par conséquent, la modification d'un coefficient de contrôle entraine forcément la modification d'au moins un autre coefficient de contrôle. L'analyse du contrôle métabolique peut être décomposée en modules (Schuster et al. 1993), obtenus grâce aux analyses topologiques comme nous l'avons vu précédemment. Ce découpage permet de réduire la complexité du système; cependant le nombre de paramètres à estimer reste trop important dans la plupart des cas. Ceci limite l'utilisation de ce type d'analyse à des exemples de comportements théoriques (Hornberg et al. 2007).

.2.4 Le modèle stœchiométrique à base de contraintes

Il existe un cadre de modélisation non paramétrique qui se situe entre la modélisation simple et statique à base de graphe et la modélisation complexe et dynamique à base d'équations différentielles. Il s'agit de la modélisation à base de contraintes ou constraint-based model (CBM) en anglais (Price et al. 2004). Les CBMs se focalisent uniquement sur le flux des réactions. Chaque réaction du réseau métabolique possède une valeur numérique qui représente la quantité de matière qui la traverse. Cet ensemble de valeurs est appelé distribution de flux ; il est représenté sous forme d'un vecteur généralement appelé v (Figure 21). Sans contraintes l'espace des solutions est infini, v peut prendre n'importe quelle valeur.

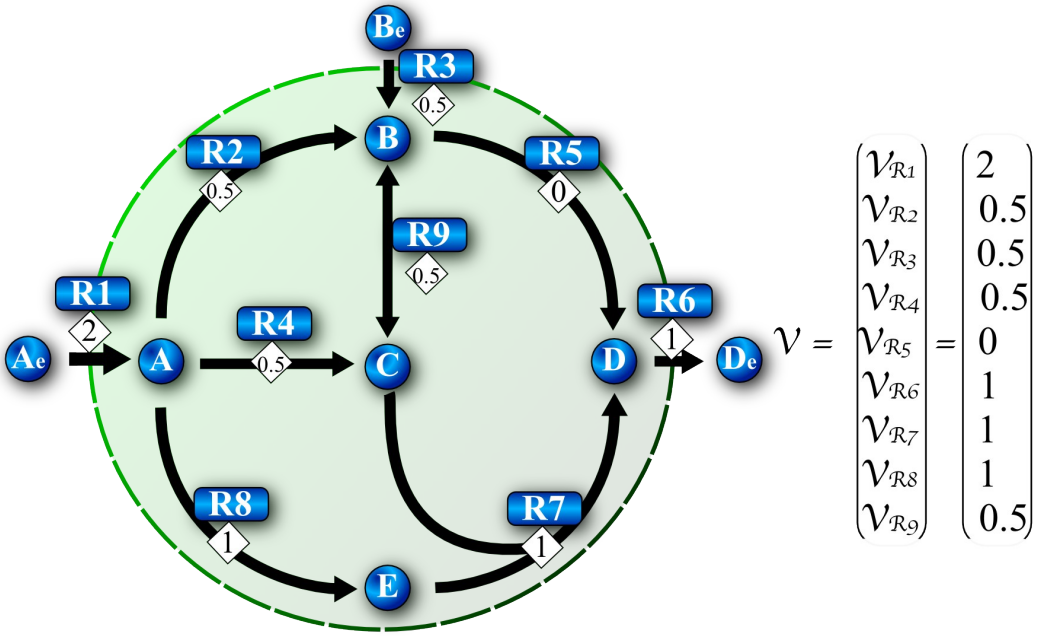


Figure 21 : Réseau métabolique et distribution de flux.
Le réseau métabolique (gauche) est composé de 9 réactions (rectangles) et 8 métabolites (cercles). La zone verte représente la bactérie et les métabolites avec le suffixe « *e* » indiquent des métabolites disponibles dans l'environnement. Les réaction R1, R3 et R6 permettent les échanges entre la bactérie et l'environnement : ce sont des transporteurs. Dans le losange blanc se trouve la valeur en unité arbitraire de chacun des flux. A droite la représentation vectorielle, chaque réaction est représentée par son flux v_{Rx}.v. Ce vecteur représente une distribution de flux qui sera calculée par l'intermédiaire du CBM.

Pour une solution particulière, le vecteur de flux sera constant au court du temps; à l'inverse des modèles cinétiques le flux de matière qui traverse une réaction ne varie pas, il s'agit d'une moyenne dans l'intervalle de temps. En effet bien qu'il n'y ait pas de description temporelle du système, on considère que les CBMs modélisent

des phénomènes de l'ordre de la minute. Les mécanismes d'ajustement enzymatique, décrits dans la partie 2.3, sont d'une vitesse inférieure à la seconde. Ils ont donc le temps de se produire et de s'équilibrer dans ce laps de temps : ils n'impactent pas les simulations des CBMs. Par contre les phénomènes de dilution ou division cellulaire rentrent dans ce cadre de modélisation. Un des avantages est que cet ordre de temps est également celui des observations et résultats expérimentaux qui sont utilisés pour construire et simuler les modèles. Si les concentrations en métabolites ne varient pas, il est important de comprendre que l'absence de variation de la concentration n'est pas synonyme d'absence de production ou de consommation des métabolites ; cela signifie que la somme de flux de production d'un métabolite est équivalente à la somme des flux le consommant. L'hypothèse selon laquelle la variation de concentration est nulle se nomme *hypothèse de stationnarité*. La contrainte directement issue de cette hypothèse est la contrainte principale des CBMs : la conservation de la matière. Lors d'une simulation d'un CBM il n'y a pas de création et d'accumulation de matière, mais uniquement des échanges et des transformations biochimiques.

2.4.1 Aspects mathématiques

Si l'objet mathématique qui représente le réseau métabolique est le graphe, celui qui représente le CBM est une matrice. Ses colonnes représentent des métabolites et ses lignes les réactions. Cette matrice contient le coefficient stœchiométrique des métabolites impliqués dans les réactions ; de ce fait elle est appelée *matrice stœchiométrique* et elle est généralement désignée par la lettre *S*. Pour un couple (métabolite, réaction) dans la matrice stœchiométrique la valeur sera nulle lorsque le métabolite n'intervient pas dans la réaction. Si le métabolite est consommé par la réaction, la valeur sera son coefficient précédé du signe moins ; enfin si le métabolite est produit, la valeur dans la matrice sera simplement son coefficient (Figure 22). On constate dans la Figure 22 que par rapport au réseau de la Figure 21, trois flux ont fait leur apparition, de plus ils ne respectent pas l'équilibre des équations bilans. Ces réactions artefactuelles sont appelées flux d'échanges et modélisent la disponibilité des métabolites dans l'environnement. Ils nous donnent un contrôle sur le modèle, puisque nous pouvons décider de la présence ou de l'absence d'un métabolite simplement en autorisant ou en annulant son flux d'échanges.
Il existe aussi d'autres réactions artefactuelles qui ne respectent pas cette contrainte. Ces réactions sont des approximations de phénomènes biologiques tels que le maintient énergétique de la cellule ou la création de biomasse qui représente la dilution des métabolites dans la cellule en intervenant lors de la division.
Les distributions de flux *V*, compatibles avec la contrainte d'état stationnaire, sont solutions de l'équation suivante $S.v=0$ (pour tous $v \in V$). Cette contrainte permet de faire passer l'espace des solutions de $\mathbf{R}^n$ à un sous espace de dimension inférieur. Elle n'est pas la seule contrainte capable de diminuer l'espace des solutions et l'un des buts de la modélisation CBM est d'ajouter de nouveaux types de contraintes pour réduire cet espace au maximum.

$$\frac{d}{dt}\begin{pmatrix} C_A \\ C_B \\ C_C \\ C_D \\ C_E \\ C_{Ae} \\ C_{Be} \\ C_{De} \end{pmatrix} \quad (1) \quad = 0 \Leftrightarrow \begin{pmatrix} \nu_{R_1} & -\nu_{R_2} & -\nu_{R_4} & -\nu_{R_8} \\ \nu_{R_2} & \nu_{R_3} & -\nu_{R_4} & -\nu_{R_9} \\ \nu_{R_4} & -\nu_{R_7} & \nu_{R_9} & \\ \nu_{R_4} & -\nu_{R_6} & \nu_{R_7} & \\ & -\nu_{R_7} & \nu_{R_8} & \\ & -\nu_{R_1} & \nu_{Ex_{Ae}} & \\ & -\nu_{R_3} & \nu_{Ex_{Be}} & \\ & \nu_{R_6} & -\nu_{Ex_{De}} & \end{pmatrix} \quad (2) \quad = 0$$

$$\Leftrightarrow \left(\begin{array}{cccccccc} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right) \cdot \left(\begin{array}{c} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \\ \nu_6 \\ \nu_7 \\ \nu_8 \\ \nu_9 \\ \hline \nu_{Ex_{Ae}} \\ \nu_{Ex_{Be}} \\ \nu_{Ex_{De}} \end{array}\right) \quad (3) \quad = 0$$

$$\Leftrightarrow S.\nu = 0 \quad (4)$$

Figure 22 : Matrice stœchiométrique et conservation de la matière.

Nous avons extrait la matrice stœchiométrique du réseau de la Figure 22. Nous avons ajouté les réactions d'échanges notées Ex suivi du nom du métabolite (sous la ligne pleine de la matrice). A l'état stationnaire, la variation des concentrations est nulle (1), ce qui signifie que la somme des flux produisant les métabolites est équivalente à celle les consommant (2). Soit le produit de la matrice stœchiométrique par le vecteur des flux est nul (3). On obtient la contrainte *S.v=0* (4).

La contrainte d'état stationnaire représente principalement les relations entre les flux. La valeur du flux sera identique (modulo le coefficient stœchiométrique) le long d'un enchainement linéaire de réactions ; il sera divisé à un embranchement de façon à ce que la somme des flux entrant soit équivalent à la somme des flux sortant (toujours au coefficient stœchiométrique près).

Il existe d'autres contraintes qui peuvent être spécifiques des flux et des réactions. On peut citer plusieurs exemples : la réversibilité de la réaction et le sens de l'activité de l'enzyme, ou encore l'introduction de valeurs minimales et maximales pour les différents flux. L'ajout de ces informations et de l'inégalité sur les bornes des flux ($v_{min} \leq v \leq v_{max}$) permettent de restreindre encore plus l'espace des solutions. Sans entrer dans les détails, l'ensemble des solutions ne sera plus un espace vectoriel, mais un polytope convexe avec des propriétés linéaires. Bien que l'espace soit borné, il existe une infinité de solutions. Afin de réduire encore cet espace, un pan d'études sur les CBMs est consacré à l'optimisation d'une fonction dite objective ; généralement il s'agit de la fonction de biomasse R_{bm}. Ces

optimisations font appel aux techniques de programmation linéaire, et se présentent de la manière suivante :

$$maximiser\ v_{bm},\ tel\ que :$$
$$S.v=0$$
$$v_{\min} \leq v \leq v_{\max}$$

Ce qui se lit : optimiser la production de biomasse, sous la contrainte de l'état stationnaire et en respectant les bornes minimales et maximales des flux. Cette méthode d'optimisation est expliquée plus en détail dans la partie 4.1

Il existe d'autres façons de réduire l'espace de solution, puisque les CBMs permettent d'intégrer différents types de contraintes telsque celles issues des données de types –omiques sur lesquelles nous reviendrons plus tard. D'autres données peuvent également être converties en contraintes. Il s'agit par exemple des connaissances sur le réseau de régulation (Covert & B. Ø. Palsson 2002) ou de signalisation (J. M. Lee, Min Lee, et al. 2008) ainsi que des données de thermodynamique (Beard et al. 2004). Cependant l'ajout de ces contraintes est plus difficile puisque elles impliquent des valeurs entières, de plus la linéarité du problème n'est pas préservée.

3 Outils de reconstruction des réseaux/modèles à l'échelle de la cellule

.3.1 Historique

Si la modélisation cinétique des réactions date des années 1910 (Michaelis & Menten 1913), il aura fallut près d'un siècle pour voir les premiers modèles métaboliques à l'échelle de la cellule (J S Edwards & B O Palsson 2000). Ce délai s'explique par le temps requit pour accumuler suffisamment de connaissance d'une part, et le temps nécessaire au développement des outils et méthodes de reconstruction d'autre part. Dés le début du XIX^{ème} siècle on commença à décrire le fonctionnement des voies métaboliques. Cependant le manque de méthodes pour identifier la totalité des molécules mises en jeu laissait généralement ces voies au niveau d'hypothèses. Les récentes évolutions technologiques telles que la spectrographie de masse ou le séquençage (Figure 23), permettent d'identifier et de donner un rôle à de plus en plus d'acteurs au sein des voies.

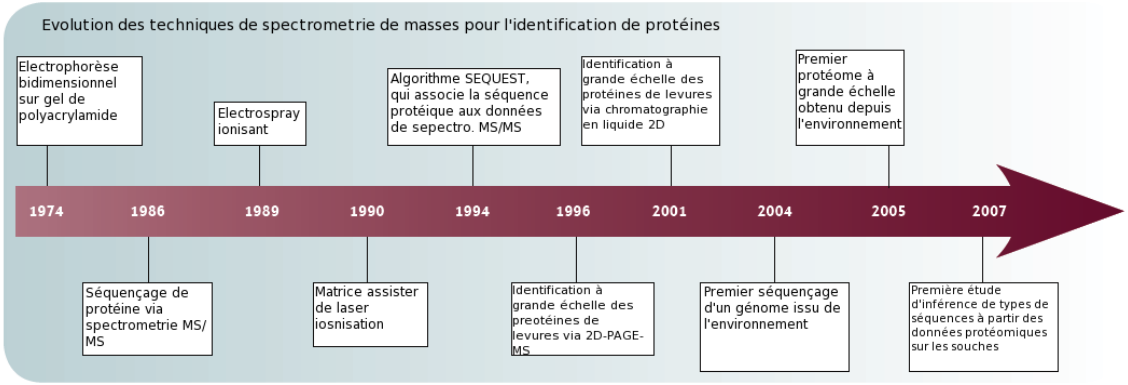


Figure 23 : Evolution des méthodes d'identification des protéines.
En 40 ans l'identification des protéines est passée de simple gel à des traitements automatisés de masse. Figure issue de (VerBerkmoes et al. 2009)

Parallèlement les développements des outils de séquençage ont permis de passer du décodage de quelques gènes à des génomes entiers (Blattner et al. 1997). La mise en

relation entre les protéines identifiées par spectrométrie et les données de séquençage, permet de prouver l'existence de certaines enzymes de l'organisme et ainsi de valider les hypothèses sur les voies métaboliques.
La quantité de données ainsi générée est rapidement devenue inexploitable par la seule recherche d'information dans la littérature et les publications. Afin de rendre l'accès aux connaissances plus simple et plus rapide la conception de bases de données appliquées à la biologie a été initiée. Elles se focalisent généralement sur un aspect de la biologie comme le génome ou le métabolisme. Les bases de données génomiques contiennent les annotations des gènes et peuvent être centrées sur un organisme comme EcoCyc pour *E. coli* (P D Karp et al. 1999) ou SGD pour *Saccharomyces* (Cherry et al. 1998). Elles peuvent rassembler des ensembles de génomes sur un type d'organisme, comme la levure dans le cas de CYGD (Güldener et al. 2005), ou encore un ensemble de génomes comme MicroScope, GenBank, EntrezGene et Genome Reviews (Baker et al. 2000; Benson et al. 2010; Maglott et al. 2011; Vallenet et al. 2009). Il existe aussi une diversité dans les bases métaboliques généralistes dont les plus couramment utilisées sont KEEG (Kanehisa et al. 2007), MetaCyc (Peter D Karp, Monica Riley, et al. 2002), SEED (DeJongh et al. 2007). Il existe aussi des bases de données spécifique des enzymes BRENDA (I. Schomburg et al. 2004) et ENZYME (Bairoch 2000) ou des transporteurs métaboliques, Transport DB (Ren et al. 2007). La Table 9 donne les principales bases de données biologiques.

Bases de données d'annotations génomiques		
DDBJ	http://www.ddbj.nig.ac.jp/	Base de données générale sur les séquences nucléotidiques
EMBL	http://www.ebi.ac.uk/embl/	Base de données générale sur les séquences nucléotidiques
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/	Base de données générale sur les séquences nucléotidiques
Integr8	http://www.ebi.ac.uk/integr8/	Informations intégrées sur des génomes complets
CMR	http://cmr.jcvi.org/	Informations intégrées sur des génomes complets de procaryotes
IMG	http://img.jgi.doe.gov/	Système intégré d'information et d'analyses sur les génomes microbiens
MicroScope	http://www.genoscope.cns.fr/agc/microscope/	Système intégré d'information et d'analyses sur les génomes microbiens
SEED	http://seed-viewer.theseed.org/	Système intégré d'information et d'analyses basé sur des sous processus biologiques
Bases de données protéiques et enzymatiques		
BRENDA	http://www.brenda-enzymes.info/	Système d'information sur les fonctions enzymatiques regroupant des données expertes et de la littérature
ENZYME	http://www.expasy.ch/enzyme/	Base de données sur la nomenclature des enzymes, associant information extensive aux numéros EC.
UniProt	http://www.ebi.ac.uk/uniprot/	Ressource universel protéique regroupant séquences et annotations de SwissProt (manuel) ou trEMBL (automatique)
TransportDB	http://www.membranetransport.org/	Prédiction des protéines transmembranaires pour les organismes entièrement séquencés.
PSORTdb	http://db.psort.org/	Base de données sur la localisation des protéines d'après des données expérimentales
Prolinks	http://prolinks.mbi.ucla.edu/	Base de données sur des liens fonctionnels entre protéines prédits
STRING	http://string.embl.de/	Base de données sur des liens fonctionnels entre protéines prédits et observés
Base de données métabolique		
CheBI	http://www.ebi.ac.uk/chebi/	Base de données sur les petites molécules du métabolisme
Pubchem	http://pubchem.ncbi.nlm.nih.gov/	Base de données sur les petites molécules
LipidMaps	http://www.lipidmaps.org/	Base de données sur les lipides du métabolisme
Reactome	http://www.reactome.org/	Base de données experte sur les voies métaboliques
KEGG	http://www.genome.jp/kegg/	Ensemble de bases de données sur les métabolites, réactions et voies métaboliques

Base de données métabolique avec lien gène/protéine		
MicroCyc	http://www.genoscope.cns.fr/agc/microscope/metabolism/microcyc.php	Collection de voies métaboliques et génomes microbiens
BioCyc	http://www.biocyc.org/	Collection de voies métaboliques et génomes microbiens
UniPathway	http://www.grenoble.prabi.fr/obiwarehouse/unipathway/	Collection experte de voies métaboliques liées à UniProt
UM-BBD	http://umbbd.msi.umn.edu/	Base de données microbienne sur les réactions biocatalytiques et les voies de dégradation
Bases de données pour les résultats expérimentaux		
IntAct	http://www.ebi.ac.uk/intact/	Base de données sur les interactions protéiques
DIP	http://dip.doe-mbi.ucla.edu/	Base de données expertes sur les interactions protéiques
Array Express	http://www.ebi.ac.uk/aerep/	Base de données publique sur les microarray
GEO	http://www.ncbi.nlm.nih.gov/geo/	Base de données publique sur les microarray
ASAP	http://asap.ahabs.wisc.edu/	Base de données de résultats d'expériences de génomique fonctionnelle
Bases de données espèce-centrique		
E. coli multi-omics DB	http://ecoli.iab.keio.ac.jp/	Jeu de données de transcriptomique, protéomique, métabolique et fluxomique sur *E. coli* K12
Systomonas	http://www.systomonas.de/	Jeu de données de transcriptomique, protéomique, métabolique et fluxomique sur *pseudomonads*
PubMed	http://www.pubmed.org/	Base de données bibliographique
Bases de données pour les modèles métaboliques		
BiGG	http://bigg.ucsd.edu/	Base de données des modèles à base de contraintes
BioModels	http://www.ebi.ac.uk/biomodels/	Base de données de modèles mathématiques et systèmes biologiques

Table 9 : Bases de données principales utilisables dans les processus de reconstruction de modèles du métabolisme.
Tableau modifié issu de (Maxime Durot et al. 2009).

L'exploitation des apports et des limites des différents types de modèles : stochastiques, cinétiques ou à base de contraintes (A. Arkin et al. 1998; Chassagnole et al. 2002; Jörg Stelling 2004; A Varma, Boesch & B O Palsson 1993b), ainsi que des différents processus biologiques modélisés (A. Arkin et al. 1998; Chassagnole et al. 2002; Van Dien & Keasling 1998; P. Wong et al. 1997) et enfin des évaluations des capacités métaboliques (Majewski & Domach 1990; Papoutsakis & C. L. Meyer 1985; A Varma, Boesch & B O Palsson 1993a; A Varma, Boesch & B O Palsson 1993b) réalisés ces quinze dernières années ont permis de concevoir un cadre de modélisation adapté au métabolisme à l'échelle de la cellule. C'est en 1999 que l'un des premiers modèles du métabolisme permettant la caractérisation des capacités métaboliques de l'organisme modèle *E. coli* a été conçu (J S Edwards & B O Palsson 2000). Ce modèle a depuis été amélioré plusieurs fois (Jennifer L Reed et al. 2003; Feist et al. 2007).

.3.2 Processus de reconstruction

La reconstruction du métabolisme à l'échelle de l'organisme utilise la séquence du génome qui peut être soit récupérée dans l'une des bases citées précédemment soit séquencée pour l'occasion. Il n'existe pas une seule et unique méthode de reconstruction des modèles, cependant il est possible de décomposer le processus en quatre parties (Figure 24).

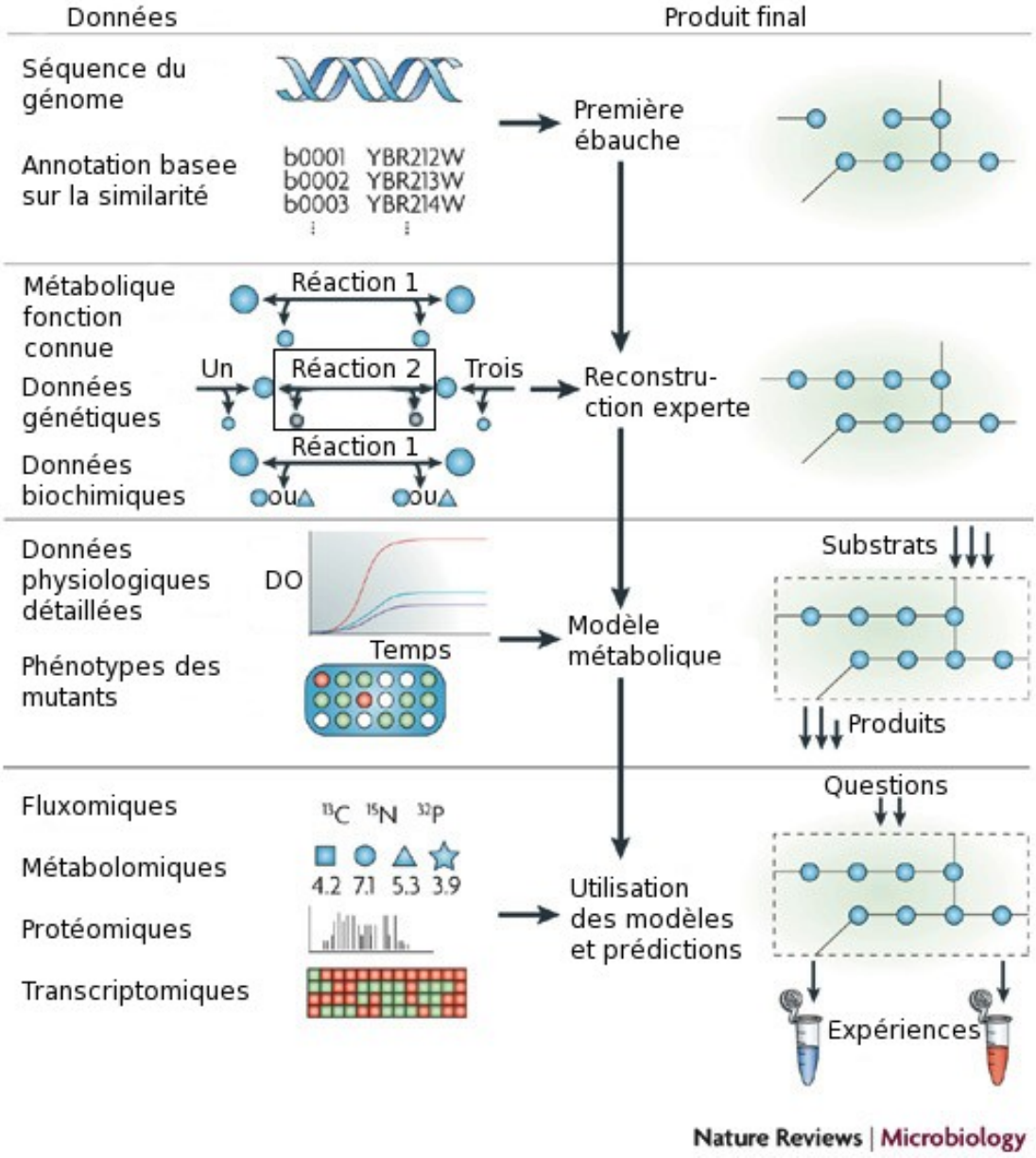


Figure 24 : Les principales étapes de reconstruction des modèles du métabolisme à l'échelle de la cellule.

Le processus de reconstruction peut être divisé en quatre phases successives. 1) reconstruction d'un réseau métabolique. 2) Amélioration du réseau et passage au modèle. 3) Amélioration du modèle. 4) intégration de données expérimentales. A partir de la deuxième phase, le processus est dépendant de données expérimentales qui vont permettre de raffiner le modèle. Chacun des produits des phases peut répondre à des questions biologiques de plus en plus complexes. Figure issue de (Feist et al. 2009).

3.2.1 De la séquence aux activités enzymatiques

La première étape consiste à identifier les gènes présents dans le génome nouvellement séquencé et en déduire une liste d'activités enzymatiques (Figure 25). Cette étape peut être automatisée et certaines plate-formes se sont spécialisées dans ce domaine (voir la Table 9). Dans un premier temps on repère sur la nouvelle séquence tous les cadres de lecture (open Reading frame ORF en anglais), qui codent potentiellement pour des CDSs. Une fonction peut être assignée à ceux-ci par homologie plus ou moins stricte (taux de similitude et longueur de l'alignement) avec les séquences des bases de données génomiques de référence (Table 9). Une limite découle immédiatement de cette méthode : une fonction non référencée ou sans séquence associée ne peut être retrouvée. L'annotation obtenue fournit une liste d'identifiants des produits de gène qui, associée aux bases de données enzymatiques, permettent d'en déduire les sous unités des enzymes, les isozymes et les complexes enzymatiques. Certaines plate-formes d'annotation permettent d'effectuer une expertise manuelle dans le but d'ajouter de nouvelles fonctions qui ne peuvent être obtenues automatiquement.

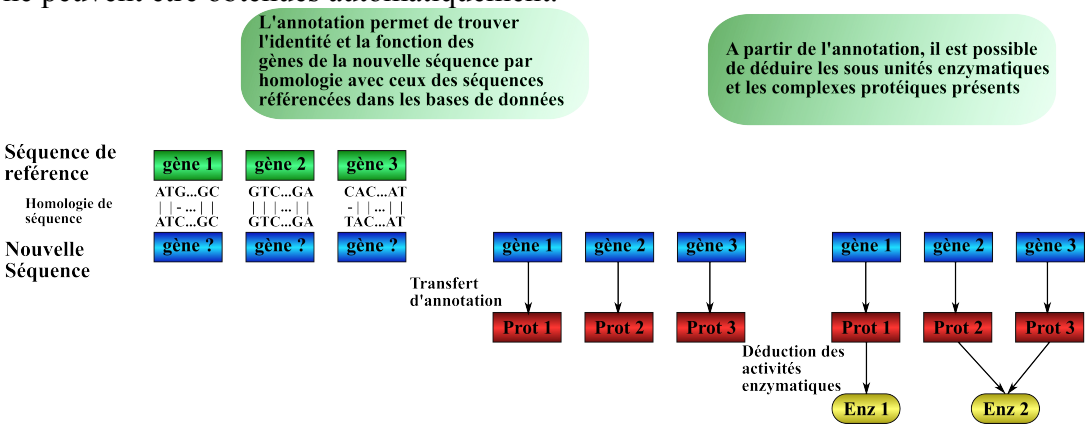


Figure 25 : De la séquence aux activités enzymatiques.
Par homologie de séquence l'annotation va être propagée des séquences de référence vers la nouvelle séquence. A partir de l'annotation ainsi obtenue, et grâce aux bases de données enzymatiques, il est possible d'en déduire les enzymes, isozymes et complexes enzymatiques qui sont potentiellement présents dans l'organisme.

3.2.2 Des activités enzymatiques au réseau métabolique

La deuxième étape du processus de reconstruction consiste à associer les réactions biochimiques aux enzymes. Cette étape peut être automatisée ; cependant une expertise est nécessaire pour obtenir un réseau plus complet (Figure 26).

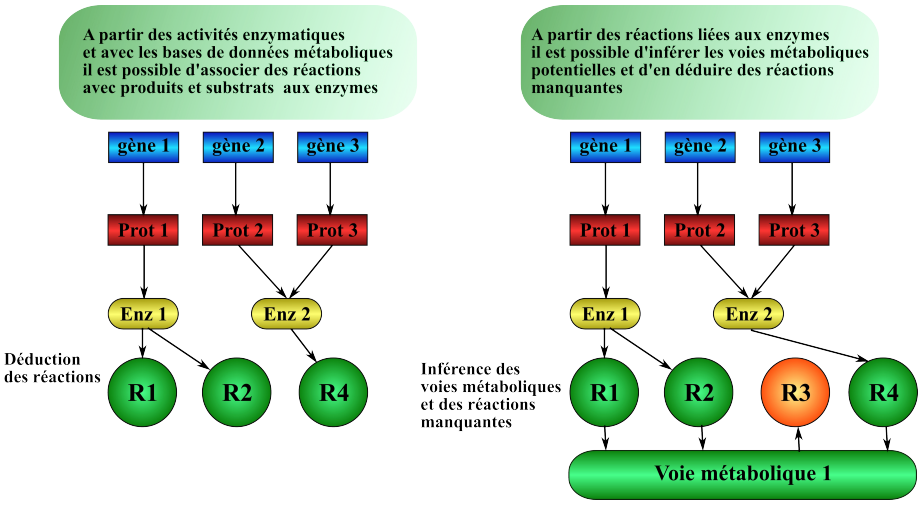


Figure 26 : Création du réseau métabolique.
A partir de l'ensemble des activités enzymatiques déduites de l'annotation et grâce aux bases de données métaboliques, un premier ensemble de réactions est défini. Celui-ci peut être amélioré par l'application de processus de raffinement automatique et/ou manuelle.

Les bases de données enzymatiques et métaboliques (voir Table 9), contiennent un ensemble de réactions et transporteurs métaboliques dont la présence est prouvée dans au moins un organisme. Ces bases regroupent les liens entre les numéros EC ou TC pour les transporteurs et la ou les réactions correspondantes. Malheureusement la biodiversité fait que la spécificité du substrat ou l'activité enzymatique peuvent varier entre des enzymes avec le même numéro EC ou TC. Par conséquent la réaction catalysée par une enzyme dans l'organisme de référence peut différer de la réaction catalysée par l'enzyme homologue dans l'organisme dont on reconstruit le réseau métabolique. De plus certains organismes sont compartimentés. La localisation cellulaire est importante puisque certaines réactions ont lieu dans des compartiments spécifiques. On peut citer la réaction de photosynthèse chez les plantes qui a lieu dans les membranes des thylacoïdes.
Il existe plusieurs outils qui permettent la reconstruction d'une ébauche de réseau dont les principaux sont PathwayTools (Peter D Karp, S. Paley, et al. 2002), GEM System (Arakawa et al. 2006), metaShark (Pinney et al. 2005) et d'autres moins répandus (Borodina & J. Nielsen 2005; Goesmann et al. 2002; Notebaart et al. 2006). L'ensemble de réactions ainsi obtenu est incomplet en terme de contenu puisqu'il ne comprend que les réactions pour les quelles l'enzyme et les gènes ont été identifiés sur le génome. Il est possible d'effectuer des améliorations automatiques et/ou manuelles pour augmenter le contenu du réseau métabolique. Les deux méthodes sont orthogonales puisque les processus automatiques permettent de retrouver rapidement des réactions référencées manquantes alors que la curation manuelle est longue et fastidieuse : elle vise l'ajout de connaissance spécifique à l'organisme.
Certaines données spécifiques sur l'organisme peuvent être trouvées dans différentes bases de données , mais elles sont le plus souvent dans des livres (Dickinson & Schweizer 2004; Heuner & Swanson 2008; Mobley et al. 2001; F. Neidhardt 1996) et des publications (articles ou revues). Les informations contenues sont de divers types comme la direction des réactions, la localisation des protéines (Huh et al. 2003) ou la spécificité en substrat et cofacteur.
Les algorithmes d'amélioration du réseau sont généralement basés sur la complétion des voies métaboliques (SMILEY (Jennifer L Reed et al. 2006), GapFind/GapFill

(Satish Kumar et al. 2007) et PathoLogic (Green & P. Karp 2004)). L'hypothèse sous jacente est que si une majorité des réactions, ou des réactions spécifiques d'une voie métabolique décrite dans une base de référence, est présente dans l'organisme alors les réactions manquantes de cette voie doivent probablement être présentes. Plusieurs raisons peuvent expliquer l'absence de ces réactions : une enzyme inconnue, la réaction est spontanée, une isoenzyme etc. Une autre partie des méthodes de raffinements des réseaux s'intéresse aux sens des réactions et leur faisabilité (Henry et al. 2006; Kümmel et al. 2006).

3.2.3 Du réseau métabolique au modèle métabolique

Avant d'effectuer des simulations et des prédictions, il est nécessaire de convertir le réseau métabolique en un objet mathématique (Borodina & J. Nielsen 2005). C'est une étape cruciale et délicate, qui nécessite généralement l'ajout d'hypothèses. Certaines d'entre elles représentent des connaissances, d'autres sont parfois sans véritable fondement biologique mais obligatoire au bon fonctionnement du modèle. Ces réactions représentent le manque de connaissances, comme par exemple une réaction puits c'est à dire une réaction dont la seule fonction est de consommer en permanence du substrat. Il existe deux points importants lors de la conversion en modèle : définir une fonction de maintient de la cellule et vérifier la stœchiométrie des réactions. Une fois ces deux points réalisés le modèle peut être utilisé pour explorer les propriétés physiologiques et les capacités de production de l'organisme. Des outils permettent de manipuler, d'améliorer et simuler le modèle CBM (Klamt et al. 2007; S. Y. Lee et al. 2005; Price et al. 2004). L'un des principaux est la COnstraint-Based Reconstruction and Analysis (COBRA) toolbox (Price et al. 2004) qui est composé de méthodes de vérifications des modèles et de simulations.
Les CBMs sont basés sur la stœchiométrie des réactions : par conséquence celle-ci doit être la plus précise et la plus fidèle à la réalité possible. Dans cette optique des algorithmes capables de détecter les incohérences stœchiométriques des réactions ont été mis au point (Gevorgyan et al. 2008; Pharkya et al. 2004). L'un des inconvénients des bases de données métaboliques est que les équations bilans sont généralement données à pH neutre alors que suivant les compartiments cellulaires le pH peut être plus acide ou plus basique. Certaines bases de données comme BRENDA (I. Schomburg et al. 2004) détaillent l'équation bilan dans différentes conditions de pH. Il existe également, en complément des bases de données, des logiciels capables d'estimer l'état de protonation des métabolites suivant le pH (Milletti et al. 2010).
La fonction de biomasse est la représentation mathématique de l'ensemble des métabolites que l'organisme doit synthétiser pour se maintenir et se multiplier. Nous avons vu précédemment (partie 2.1 sur les métabolites) qu'il est possible d'estimer la composition en biomasse de certains organismes. On peut décomposer cette biomasse en briques élémentaires comme par exemple les nucléotides qui constituent l'ADN, les acides aminés qui forment les protéines et les lipides des bicouches lipidiques. Chacune de ces briques va se voir affecter d'un coefficient stœchiométrique et l'ensemble des couples briques/coefficients formera la fonction de biomasse. Cette fonction est déduite à partir de résultats expérimentaux pour quelques organismes, elle reste cependant une approximation et dans l'idéal elle devrait être définie pour chaque condition d'expérience.
En plus des réactions artificielles dues au manque d'information, il est également nécessaire de rajouter des réactions artefactuelles qui permettent de contrôler le

modèle. La grande majorité de ces réactions sont des flux d'échanges qui représentent les composés disponibles dans l'environnement. La fonction de biomasse est également une des fonctions artefactuelles. Le modèle obtenu est normalement capable de produire les précurseurs de biomasse sur un milieu riche ; cependant, tout comme l'ébauche de réseau, il est nécessaire d'améliorer celui-ci si l'on désire un modèle plus fidèle à la réalité biologique.

3.2.4 Amélioration du modèle par l'ajout de données expérimentales

Il existe deux principales sources de données biologiques pour raffiner les modèles : des données de croissance sur de nombreux milieux et des données sur l'essentialité des gènes via des mutants. La première permet l'ajout d'information dans le modèle et la seconde permet une correction des informations déjà disponible.
La prédiction de phénotypes consiste à estimer les capacités du modèle à utiliser les métabolites externes dans différents environnements définis au préalable. Cette analyse est effectuée de manière pratiquement systématique lors de la reconstruction de modèles (Duarte et al. 2004; Maxime Durot et al. 2008; J. Lee, Yun, et al. 2008; Oh et al. 2007; Feist et al. 2007). Ces prédictions sont ensuite comparées à des résultats expérimentaux de croissances sur les mêmes environnements. Il est possible de s'intéresser aux capacités métaboliques *in-silico* de manière qualitative en se focalisant sur la production des précurseurs de la biomasse (Imieliński et al. 2005). Il est également possible de quantifier ces capacités en utilisant la fonction de biomasse et en mesurant son flux de matière (Price et al. 2004). Il existe des méthodes expérimentales pour déterminer à grande échelle les sources de carbones utilisées par une bactérie, dont les BIOLOGs (Mauchline & Keevil 1991). Ce sont des plaques de cultures contenant quatre-vingt quinze sources de carbones différentes et un témoin. La confrontation entre les données expérimentales et les prédictions permet de mettre en évidence certaines inconsistances du modèle. Il existe deux types d'incohérence entre le modèle et les données expérimentales de croissances : soit l'expérience montre une croissance mais pas le modèle, on parle de faux négatifs, ou le modèle prédit une croissance mais pas l'expérience, et dans ce cas on parle faux positifs.
Il existe trois principales raisons aux faux négatifs. L'absence du transporteur de la source de carbone, l'absence d'une ou plusieurs réactions de la voie de dégradation de la source de carbone, et enfin l'absence d'information sur les voies de dégradation de la source de carbone. Quant aux faux positifs il existe deux principales raisons : des phénomènes de régulations qui ne sont pas pris en compte dans le modèle et des effets de sur-prédictions des réactions inférées par le processus de reconstruction. Bien que majoritairement en cas d'incohérence l'erreur provienne du modèle, il se peut dans certains cas que la prédiction soit juste et que l'incohérence soit le résultat d'erreurs expérimentales.
Les données d'essentialités des gènes contrairement au phénotype de croissance n'ont pas vocation à identifier de potentiels trous dans le réseau métaboliques. Elles servent à explorer les liens gènes-protéines-réactions (GPRs) et définir dans un milieu donné quels sont les gènes indispensables à la vie de l'organisme (Figure 27).
Il est important de comprendre que l'essentialité des gènes et des réactions associées est dépendante de l'environnement, même si certains gènes sont essentiels dans plusieurs environnements. Cela s'explique facilement par l'utilisation des métabolites ; si dans l'environnement la seule source de carbone disponible est le glucose, l'inactivation d'une réaction de la voie de dégradation du glucose aura un

effet important alors que l'inactivation de la voie de dégradation d'une autre source n'aura pas d'effet. Certaines réactions notamment celles qui produisent des métabolites indispensables de la biomasse sont essentielles, peu importe l'environnement.

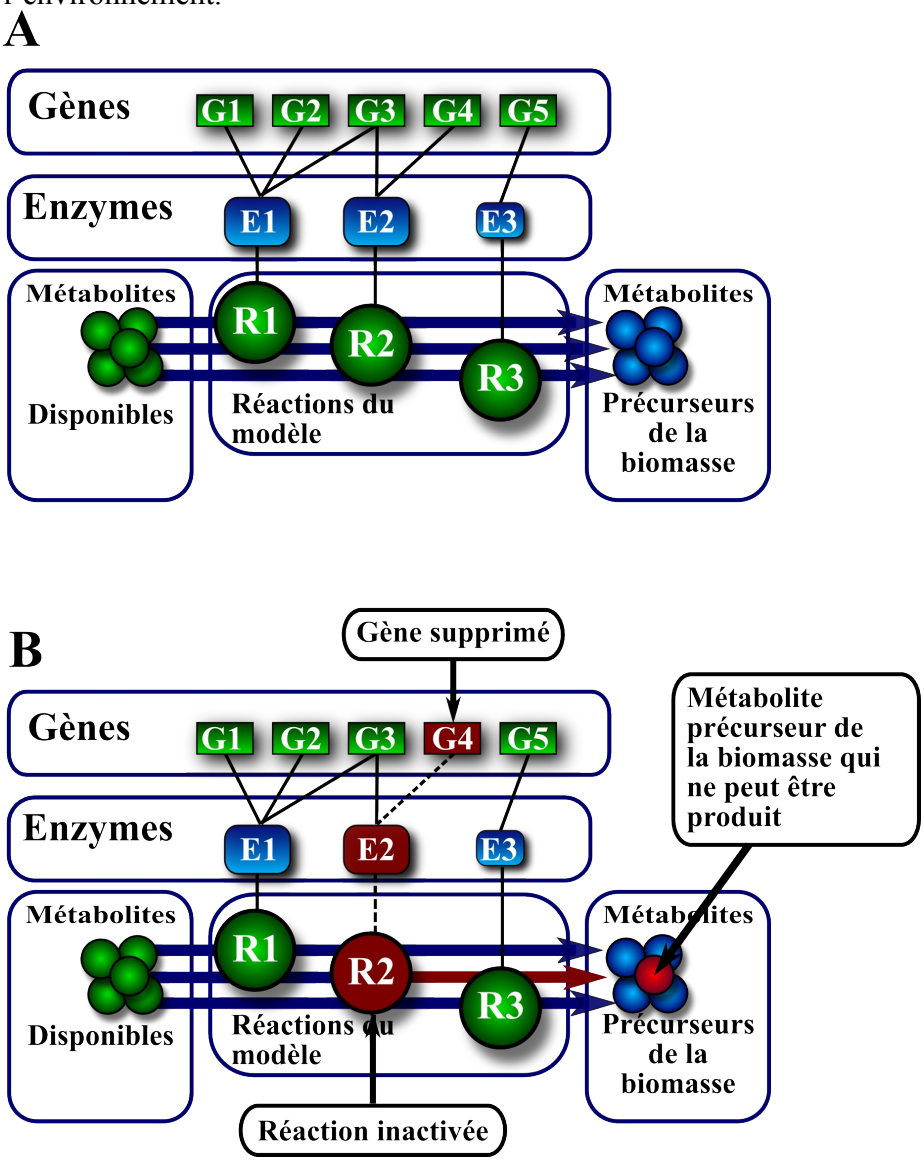


Figure 27 : Lien Gène-Protéine-Réaction (GPR) et essentialité.
Les GPRs sont les liens qui permettent de relier les réactions aux gènes du génome, ce qui rend la différenciation entre les isozymes et les complexes enzymatiques possible. En A dans un milieu donné il est possible de produire l'ensemble des précurseurs de la biomasse. En B la suppression du gène G4 inactive la réaction R2 et bloque la production d'un des précurseurs. Le gène G4 est donc essentiel à la vie de l'organisme dans cet environnement.

Connaissant les métabolites de l'environnement et les métabolites précurseurs de la biomasse on peut, à l'aide du modèle, en déduire les réactions qui doivent avoir lieu et, par les GPRs, en déduire les gènes essentiels. Ces prédictions peuvent être ensuite comparées à des données d'essentialités chez l'organisme par l'intermédiaire d'une banque de mutants, comme il en existe pour différents organismes de référence (Akerley et al. 2002; T. Baba et al. 2006; de Berardinis et al. 2008; Kitagawa et al. 2005; Kobayashi et al. 2003; Liberati et al. 2006; Suzuki et al. 2006). Il existe deux cas d'incohérence entre l'expérience et la simulation : le premier cas correspond aux faux essentiels c'est à dire que le gène est prédit comme

essentiel or l'expérience montre que l'organisme peut vivre sans ce gène. Le deuxième cas correspond aux faux facultatifs, c'est à dire des gènes prédits comme non essentiels par le modèle mais qui sont indispensables à la survie de l'organisme. Ces deux types d'erreurs mettent en lumière des erreurs d'association entre les gènes, les complexes et les réactions. Ceci permet dans certains cas de corriger les GPRs en modifiant les complexes enzymatiques ou en ajoutant des isozymes. Lorsqu'une erreur est identifiée il peut malheureusement exister une ou plusieurs solutions qui corrigent le modèle. Identifier ces solutions peut vite devenir fastidieux en raison de la combinatoire des associations entre les gènes, les complexes et les enzymes. Des algorithmes dont AutoGPR (Maxime Durot 2009) sont capables de proposer une liste de scénarii possibles, et de rendre le modèle cohérent avec les résultats expérimentaux. Dans le cas d'un seul scénario il est possible de corriger le modèle, dans les cas de plusieurs scenarii il est nécessaire de procéder à de nouvelles expériences pour valider ou invalider les différents scénarii.

4 Utilisation des modèles du métabolisme à l'échelle de la cellule

.4.1 Propriétés des réseaux

Les CBMs s'intéressent à la quantité de matière qui passe au travers des réactions. Etudier la répartition de la matière, appelé flux de distribution, est donc la suite logique à la construction d'un CBM (Almaas et al. 2004; Jennifer L Reed & B. Ø. Palsson 2004; Wiback et al. 2004).
L'analyse de l'équilibre des flux ou FBA est l'une des utilisations les plus courantes des CBMs (Figure 28). Ce genre d'analyse consiste à estimer l'ensemble des valeurs de flux qui sont compatibles avec les contraintes et qui résolvent une fonction objective (Kauffman et al. 2003).

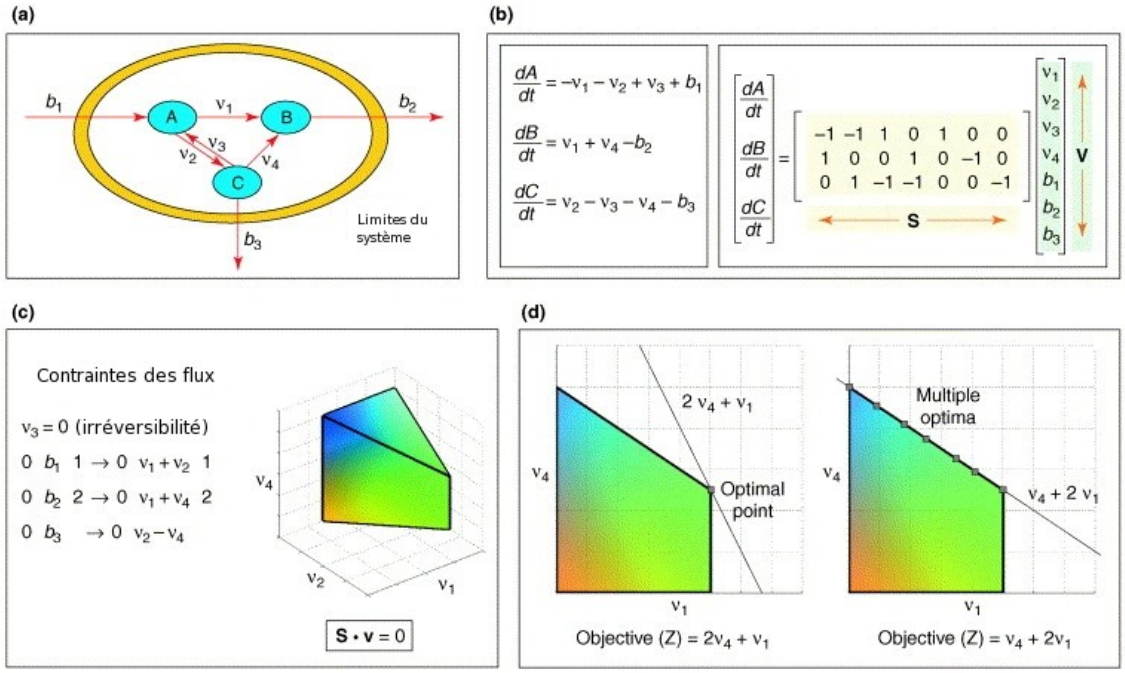


Figure 28 : Méthodologie de l'analyse de la distribution des flux.
(a) Un modèle qui comprend trois métabolites (A, B et C) avec trois réactions v_i dont une irréversible et trois flux d'échange b_i. (b) Equation de conservation de la matière avec $S.v=0$. Contraintes sur les flux (irréversibilité), ce qui va donner le cône des flux admissibles. (d) Optimisation suivant différentes fonctions objectives Z. Dans certains cas un seul point optimal existe ; la plus par du temps la solution sera un ensemble de points, le long d'une arrête du cône.

Il existe diverses manières d'étudier la distribution de flux. On peut estimer par échantillonnage la distribution des flux des différentes réactions et dans différentes conditions (Almaas et al. 2004). Ces travaux ont mis en évidence un squelette composé de quelques réactions avec un flux important tandis que la majorité des réactions ont des flux hautement variables. Il est aussi possible d'estimer la variabilité des flux en calculant les bornes maximales et minimales d'un flux indépendamment des autres Cette technique est appelée analyse de la variabilité des flux ou FVA en anglais (R Mahadevan & C H Schilling 2003). A l'inverse on peut s'intéresser aux liens et dépendances entre les flux des réactions qui varient de façon similaire ou opposées lors des modifications de l'environnement (Burgard et al.

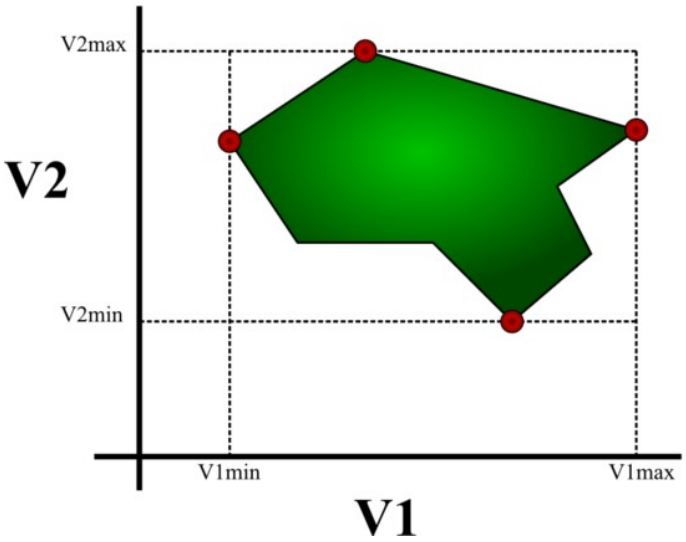
2004). Ce couplage des flux peut être mis en relation avec d'autres données biologiques comme le réseau de régulation (Notebaart et al. 2008; Jennifer L Reed & B. Ø. Palsson 2004). Il est même possible d'observer et d'étudier l'évolution des gènes transférés horizontalement (Pál, Papp & Lercher 2005a; Pál, Papp & Lercher 2005b).
Il est aussi possible de s'intéresser aux chemins entre différents métabolites. L'étude des voies métaboliques recherche d'un côté les chemins minimaux et respectant les contraintes (Klamt & Jörg Stelling 2003; Papin et al. 2004), et de l'autre leur redondance via les voies métaboliques alternatives (Papin et al. 2002). L'énumération de l'ensemble des routes minimales ou alternatives reste limitée à des sous réseaux à cause de la combinatoire qui pour le moment dépasse les capacités de calcul des machines. Cependant des progrès algorithmiques sur le calcul, des voies (Terzer & Jörg Stelling 2008) ou les porjets méthodologiques sur la découpe en sous réseaux (Verwoerd 2011), laissent présager le calcul des voies sur le réseau complet dans un avenir proche.

.4.2 La prédiction des phénotypes de croissance

Nous avons vu que la prédiction de phénotypes de croissance associée aux données expérimentales permet d'améliorer le modèle. Après l'optimisation du modèle il est possible d'estimer à partir des métabolites externes, les métabolites produits, la valeur des flux de croissances, la production des métabolites précurseurs ou encore d'étudier le bilan énergétique. L'analyse de l'équilibre des flux est une méthode dont le but est de prédire de manière quantitative les phénotypes de croissances. Cette méthode est dans la continuité des travaux du milieu des années 80 de Papoutsakis (Papoutsakis 1984) puis Fell (D A Fell & Small 1986) qui ont introduit la programmation linéaire pour calculer des flux de production de métabolites. Savinell et Palsson effectuèrent des analyses détaillées et développèrent l'aspect théorique de la FBA (Savinell & B O Palsson 1992a; Savinell & B O Palsson 1992b). La FBA a été conçu dans le but d'effectuer des prédictions quantitatives de la fonction de biomasse (Amit Varma & Bernhard O. Palsson 1994). Elle repose sur une hypothèse très forte selon laquelle, au cours de l'évolution, l'organisme a optimisé son métabolisme pour produire de la biomasse. Bien que discutable (Schuster et al. 2008) cette hypothèse s'est avérée correcte dans certaines conditions (J S Edwards et al. 2001; Rafael U. Ibarra et al. 2002).
L'une des limites du FBA est qu'il peut exister un grand nombre de distributions en accord avec les contraintes et l'hypothèse d'optimalité de la production de biomasse.

.4.3 Prédiction de la variabilité des flux.

L'étude de ces solutions alternatives (R Mahadevan & C H Schilling 2003) permet de mieux comprendre les redondances du réseau métabolique et les différents modes de fonctionnements du réseau métabolique, mais elle reste délicate. L'une des méthodes est l'analyse de la variabilité des flux ou FVA (Flux Variability Analysis, Figure 29). Elle consiste à estimer les valeurs extrêmes des flux plutôt qu'une solution particulière. Le calcul est effectué flux par flux et non plus globalement comme dans la FBA.



$$Cas\ 1$$
$$Max\ \nu_i,$$
$$sujet.\ \ S.\nu = 0,$$
$$f^{Tv} = Z_{obj},$$
$$0 \leq \nu \leq \nu_{max}\ pour\ i = 1...n$$

$$Cas\ 2$$
$$Min\ \nu_i,$$
$$sujet.\ \ S.\nu = 0,$$
$$f^{Tv} = Z_{obj},$$
$$0 \leq \nu \leq \nu_{max}\ pour\ i = 1...n$$

Figure 29 : Schéma de la FVA.
A gauche la représentation graphique de la FVA avec en vert l'espace de solution et en rouge les valeurs extrêmes des flux. A droite la formalisation du problème, qui consiste à maximiser (minimiser) un flux, sous contraintes d'état stationnaire, d'une fonction objective Z_{obj} et des contraintes du modèle.

La FVA permet d'obtenir l'intervalle de toutes les valeurs admissibles pour un flux donné ; cependant il est important de comprendre que puisque le calcul est réalisé indépendamment pour chaque flux, l'ensemble des couples v_1 et v_2, ne sont pas compris dans l'espace de solutions. Par exemple sur la Figure 29 on voit que le couple (v_{1min}, v_{2min}) n'est pas compris dans l'espace de solution.

.4.4 Délétion de gènes

Les GPRs sont des liens abstraits qui relient les réactions au génome. Des méthodes qui modifient la FBA prennent en compte et estiment les effets de la délétion d'un ou plusieurs gènes et donc des GPRs, sur les capacités de production du modèle métabolique. L'objectif de ce genre d'approche est d'estimer l'essentialité des gènes d'une part et d'améliorer le rendement de certains flux, en particulier la biomasse d'autre part. Les algorithmes de prédiction de flux simulant des délétions de gènes se basent sur l'hypothèse que la variation des flux doit être minimale comparée aux flux des prédictions dans des conditions standards. Il existe différentes façons et donc approches qui suivent cette hypothèse, parmi lesquelles deux se démarquent. La première consiste à minimiser l'ajustement de chacun des flux et est appelée MOMA pour Minimization Of Metabolic Adjustment (Segrè et al. 2002) (Figure 30).

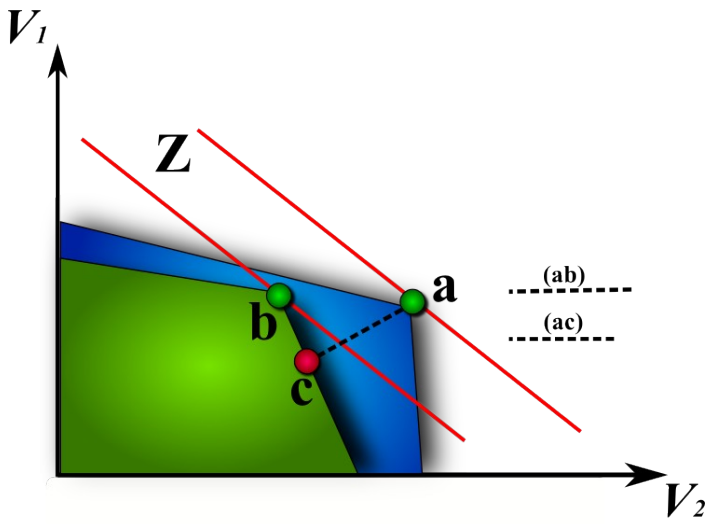


Figure 30 : Illustration du principe de l'algorithme MOMA pour un système à deux réactions. En bleu l'espace initial de solutions, en vert celui après la perturbation génétique. Soit Z la fonction objective dans une FBA classique, la solution initiale est (a) avant modification et (b) après. Moma propose comme solution alternative le point (c) car c'est le point appartenant à l'espace des solutions vert le plus proche de (a). De ce fait, la perturbation engendrée par la modification génétique est minimale.

L'autre approche, ROOM pour Regulatory On/Off Minimization (Shlomi et al. 2005), consiste à trouver la solution optimale tout en minimisant le nombre de flux qui vont être modifiés par rapport aux conditions standards. Les deux manières de procéder donnent des résultats plus proches des données expérimentales qu'une FBA réalisée sur réseau avec le gène supprimé. Ces meilleurs résultats peuvent s'expliquer par le fait que la FBA repose sur l'hypothèse de la lente optimisation du métabolisme durant l'évolution et non sur un changement récent et ponctuel. Ces méthodes ont ouvert la voie à plusieurs algorithmes utilisés dans la bio-ingénierie dont la série des Opt-. (OptKnock (Burgard et al. 2003), OptStrain (Pharkya et al. 2004), OptReg (Pharkya & C. D. Maranas 2006) et OptGene (Patil et al. 2005)). Plus généralement les CBMs sont un outil de la bio-ingénierie et la biologie synthétique : ces domaines étant hors du cadre de la thèse, ils ne seront pas plus détaillés.

.4.5 Intégration de données biologiques

Les nouvelles technologies permettent d'obtenir de plus en plus de données et de plus en plus rapidement. Cependant l'analyse de ces données suit difficilement le rythme de production et la mise en rapport de différents types de données n'est pas triviale. La biologie des systèmes met en relation cet ensemble de données – *omiques* et les modèles à base de contraintes qui sont un excellent cadre d'intégration de par la flexibilité des données utilisables et la modélisation à l'échelle de l'organisme.

4.5.1 Génomique

Les données de génomique concernent le génome, sa séquence et les annotations. Ces dernières ne se limitent pas uniquement aux gènes mais également aux régions en aval et en amont qui contiennent les sites de fixations de facteurs de transcription régulant l'expression des gènes. La principale utilisation de la génomique dans les modèles du métabolisme à lieu lors du processus de reconstruction du réseau métabolique et de la création des liens GPRs. L'apport d'information n'est pas à

sens unique et le modèle métabolique peut servir dans une moindre mesure à compléter les connaissances sur le génome. Le réseau reconstruit à partir de l'annotation seule est incomplet ; pour compléter celui-ci nous avons vu qu'il existe différentes méthodes capables de trouver des réactions manquantes (Green & P. Karp 2004; Satish Kumar et al. 2007). Des méthodes sont développées pour relier ses réactions à des gènes grâce aux enzymes. Ces méthodes sont délicates à mettre en œuvre puisqu'elles nécessitent de connaître l'ensemble des ORFs de l'organisme (Brasch et al. 2004). Elles nécessitent également diverses connaissances telles que le contexte génomique (Kharchenko et al. 2006), les profiles phylogénétiques (L. Chen & Vitkup 2006) ou encore des données d'expressions (Kharchenko et al. 2004). Des méthodes comme CANOE (A. Smith article en cours) essaient d'intégrer les différentes informations pour assigner à des gènes sans annotation des enzymes orphelines de gènes.

4.5.2 Transcriptomique

La transcriptomique fournit des informations sur la présence et la quantité relative des ARN transcrits à partir des gènes. Depuis le milieu des années 90, de nombreuses données ont été générées sur un grand nombre d'organismes et dans de multiples conditions et environnements (Hardiman 2004). La présence ou l'absence de transcrit peut être traduit en contraintes sur la présence ou non des réactions et donc sur le fonctionnement ou le blocage des voies métaboliques (Bumgarner & Yeung 2009; ter Kuile & H V Westerhoff 2001). Ce lien entre transcrit et réaction repose sur l'hypothèse que la présence d'un transcrit implique l'activité enzymatique et son absence implique l'absence d'activité enzymatique. Cette hypothèse est globalement juste, mais il existe plusieurs facteurs qui peuvent l'invalider. Tout d'abord la régulation peut se faire sur le transcrit du gène : on parle alors de régulation post-transcriptionnelle. Ensuite l'absence de transcrit peut être dûe à des aléas expérimentales. Enfin dans le cas des transcrits en très faible quantité il est difficile de savoir si le nombre de transcrits est suffisant pour être effectif dans l'organisme.

4.5.3 Protéomique

La protéomique concerne le devenir des transcrits qui sont traduits en protéines. L'objectif de la protéomique est d'estimer l'ensemble des protéines présentes dans l'organisme ainsi que leur quantité. Les CBMs utilisent principalement les informations sur les protéines enzymatiques. Les deux méthodes les plus utilisées dans ce domaine sont les électrophorèses sur gel en deux dimensions et la spectrométrie de masse (T. Liu et al. 2007; Patterson & Aebersold 2003). D'autres méthodes beaucoup plus lourdes à base de western-blot ont été développées (Ghaemmaghami et al. 2003). En parallèle, des protocoles expérimentaux ont permis de se focaliser sur la localisation des protéines et de leurs quantités dans les différents organites (Jerby et al. 2010). On peut appliquer les hypothèses du transcriptome au protéome : la présence d'une protéine enzymatique implique la présence de la réaction dans les conditions de l'étude. Un des soucis de la protéomique est la différence de concentration entre les protéines : les protéines les plus abondantes masquent les protéines en faible quantité. Ces différences de concentrations rendent difficile la classification en fonction de l'expression (nulle, faible, normale ou sur exprimée). De plus la capacité de conversion des enzymes varie également dans une large gamme de valeurs et par conséquence l'activité

enzymatique finale est un mixte entre la concentration et l'efficacité de l'enzyme. Il existe cependant des tentatives pour dépasser cette limite en attribuant un score à chacune des protéines (Hardiman 2004) .

4.5.4 Métabolomique

La métabolomique s'intéresse à l'ensemble des métabolites de l'organisme et principalement la variation des constituants du métabolome en fonction des modifications de l'environnement et des conditions d'expérimentation. Le métabolome est le résultat des transformations des métabolites effectuées par les enzymes du protéome. Estimer la totalité des métabolites d'une cellule est relativement récent et les méthodes et protocoles expérimentaux sont encore en plein développement. Les méthodes les plus couramment utilisées sont la spectrométrie de masse, la spectrométrie par résonance magnétique nucléaire et la spectrométrie par vibration (Bumgarner & Yeung 2009; ter Kuile & H V Westerhoff 2001). Récemment d'autres approches ont vu le jour utilisant des nanotechnologies (T. Liu et al. 2007; Patterson & Aebersold 2003). L'une des difficultés majeures de la métabolomique est la diversité des métabolites, difficulté à laquelle s'ajoute la gamme de valeurs des variations des concentrations. Ces difficultés ne semblent cependant pas freiner l'application de la métabolomique qui va des microbes (Ghaemmaghami et al. 2003) à l'humain (Markuszewski et al. 2005). Le fait d'avoir directement accès aux métabolites trouve des applications dans la pharmacologie et toxicologie (Robertson 2005) et également dans l'agroalimentaire (Gibney et al. 2005). La concentration d'un métabolite est facilement exploitable dans un modèle cinétique, mais nécessite de fortes hypothèses ou des manipulations pour être intégrée dans un CBM.

4.5.5 Fluxomique

La suite de l'étude du métabolome, est celle du déplacement de matière à travers les flux ou fluxomique. Depuis la fin des années 90 grâce au carbone 13 et à la RMN on est capable d'estimer les flux des principales voies du métabolisme central (U Sauer et al. 1999). Depuis, les améliorations aussi bien du point de vue du matériel que des calculs et des méthodes expérimentales ont permis de mesurer les flux sur des ensembles de mutants (Fischer & Uwe Sauer 2003). Il est également possible de s'intéresser aux variations des flux en fonction des modifications de l'environnement (Uwe Sauer 2004). Contrairement au métabolome, la fluxomique est directement utilisable par les CBMs en apportant des contraintes sur la valeur de certains flux. Cependant la fluxomique reste pour le moment un processus fastidieux à mettre en place.

.4.6 Gestion des modèles

La reconstruction des modèles du métabolisme à l'échelle de la cellule est un exercice récent qui pour le moment découle d'un long processus manuel. De ce fait et à l'inverse des plate-formes génomique, il n'existe pas de structures qui recense l'ensemble ou au moins une partie des CBMs existants, et qui propose différents outils de comparaisons et d'analyses. Néanmoins, il existe BiGG (Schellenberger et al. 2010) la base de données principale concernant les modèles à base de contraintes ; mais elle ne regroupe que les modèles reconstruits au sein de leur équipe. Les autres modèles ne sont pas centralisés et il n'existe pas non plus de consensus sur la forme sous laquelle ils sont délivrés. Une base de données

communautaire pour les modèles biologiques est disponible, mais elle est consacrée aux modèles cinétiques : *BioModels Database* (Le Novere 2006). Bien que le cadre de modélisation ne soit pas le même, cette base de donnée impose un format de fichier qui peut être utilisé pour les CBMs.

4.6.1 SBML

Le format en question est le *System Biology Markup Language* ou SBML (Hucka et al. 2003), dont la première version est apparue en 2003 et qui est actuellement en version 3. Initialement prévu pour les modèles cinétiques sa structure modulaire lui permet de s'adapter à un grand nombre de types de modèles. Cette flexibilité est issue de la structure même du langage qui est une succession de listes optionnelles (Table 10). Dans un SBML le début d'une définition de liste est signalé par une balise qui sera toujours de la forme '<*ListOfX*>' où *X* est remplacé par le nom de la liste. Cette définition se terminera toujours par une balise de la forme '</*ListOfX*>'. Une entité de la liste *X* commencera toujours par la balise '<*X*>' et se finira toujours par la balise '</*X*>'.

	Début de la définition du modèle
Liste des fonctions	Fonctions qui peuvent être utilisées lors des simulations
Liste des définitions d'unité	Unités qui sont utilisées dans le modèle
Liste des compartiments	Les conteneurs (physique ou virtuel) où les espèces sont localisées
Listes des espèces	Entités du même type qui interviennent dans les réactions (ions protéines)
Liste des paramètres	Quantités constantes ou variables, locales ou globales qui vont régir le modèle
Liste des tâches initiales	Conditions initiales du modèle
Liste des règles	Expression mathématique symbolisant les interactions et effet des différentes variables.
Listes des contraintes	Expression mathématique utilisant les paramètres et variables dont le résultat booléen permet de détecter le bon fonctionnement du modèle et potentiellement diagnostiquer les erreurs
Listes des réactions	Description des transformations et interactions des espèces entre elles
Listes des évènements	Actions qui peuvent modifier instantanément une ou plusieurs variables lorsqu'une condition est satisfaite
	Fin de la définition du modèle

Table 10 : Les différentes listes optionnelles du format SBML

Les CBMs utilisent les listes compartiments, espèces et réactions. Comme le frmat SBML est utilisé principalement pour les modèles cinétiques et contient des champs pour les paramètres cinétiques, il n'existe pas de champs spécifiques des CBMs. Cependant il est possible d'inclure un champ *note* dans lequel nous allons introduire toutes ces informations.
Le fichier SBML modifié, pour être compatible avec les CBMs, se compose de 4 blocs.
Le premier constitué d'une ligne concerne l'identification du modèle :

```
<model id="iEcoliK12MG1655" name="E. coli K-12 MG1655">
```

Le deuxième bloc contient la liste des compartiments, normalement la définition d'un compartiment contient un volume cet aspect n'étant pas pris en compte dans les CBMs, cet élément a été supprimé.

```
<listOfCompartments>
```

```
                    <compartment id="c" name=”cytosole”/>
                    …
                </listOfCompartments>
```

Le bloc suivant est consacré à la définition des espèces qui dans les CBMs sont les métabolites.

```
<listOfSpecies>
    <species metaid="M_lys_DASH_L_LBRACKET_c_RBRACKET_"
    id="M_lys_DASH_L_LBRACKET_c_RBRACKET_" name="L-Lysine"
    compartment="c">
      <notes>
        <html:p>
          FORMULA: C6H15N2O2
        </html:p>
      </notes>
    </species>
    …
</listOfSpecies>
```

Un métabolite est défini par un identifiant unique, un nom et son compartiment, ce qui permet de distinguer les métabolites lors des réactions de transports ou les réactions se produisant à cheval sur deux compartiments. Un premier ajout spécifique des CBMs est présent dans la liste des métabolites : il est définit après la balise <*note*>. Cet élément est la formule chimique avec le mot clé *FORMULA*.

Le dernier bloc est la liste de l’ensemble des réactions : c’est également le bloc le plus complexe.

```
<listOfReactions>
    <reaction id="R_LYSabcpp" name="L-lysine transport via ABC system (periplasm)"
    reversible="false" >
      <notes>
        <html:p>
          Equation: lys-L[p] + h2o[c] + atp[c] --> pi[c] + lys-L[c] + h[c] + adp[c]
        </html:p>
        <html:p>
          GENE ASSOCIATION: (GO2224481) and (GO2224478) and (GO2224479) and
        (GO2224477)
        </html:p>
      </notes>
      <listOfReactants>
        <speciesReference species="M_lys_DASH_L_LBRACKET_p_RBRACKET_"
        stoichiometry="1"/>
        <speciesReference species="M_h2o_LBRACKET_c_RBRACKET_" stoichiometry="1"/>
        <speciesReference species="M_atp_LBRACKET_c_RBRACKET_" stoichiometry="1"/>
      </listOfReactants>
      <listOfProducts>
        <speciesReference species="M_pi_LBRACKET_c_RBRACKET_" stoichiometry="1"/>
        <speciesReference species="M_lys_DASH_L_LBRACKET_c_RBRACKET_"
        stoichiometry="1"/>
        <speciesReference species="M_h_LBRACKET_c_RBRACKET_" stoichiometry="1"/>
        <speciesReference species="M_adp_LBRACKET_c_RBRACKET_" stoichiometry="1"/>
      </listOfProducts>
    </reaction>
    …
</listOfReactions>
```

Une réaction est définie à l’aide d’un identifiant unique, son nom et une information sur la réversibilité via une valeur booléenne. Trois sous-blocs vont définir la réaction. Le premier, compris dans la balise <*note*>, contient les informations spécifiques du CBM ; chacune des informations est précédée par un mot clé qui définit l’information contenue. Le premier mot clé est *Equation*, il est succédé par l’équation bilan de la réaction. Les métabolites sont localisés dans cette équation (les crochets contiennent l’identifiant du compartiment). La réversibilité est indiquée par le symbole ‘<=>’ et l’irréversibilité par ‘-->’. Le second mot clé,

GENE ASSOCIATION, est suivi de la GPR. Le sous-bloc suivant compris dans la balise '*<listOfReactants>*' contient l'ensemble des métabolites substrats localisés de la réaction. Chaque substrat localisé sera définit par son identifiant et son coefficient stœchiométrique. Le dernier sous-bloc correspond aux métabolites produits localisés de la réaction et est compris dans la balise '*<listOfProducts>*' ; tout comme pour les substrats cette liste contient des métabolites localisés identifiés par leur identifiant ainsi que leur stœchiométrie dans la réaction.
Le format SBML profite d'une importante communauté et de bibliothèques compatibles avec la plupart des langages de programmation (perl, java, etc.) et des logiciels de modélisation (Matlab, R, etc.).

4.6.2 NemoStudio

Il existe cependant une initiative pour centraliser et simuler les modèles du métabolisme à base de contraintes. Cette web plate-forme nommée NemoStudio à trois objectifs (F Le Fèvre et al. 2009) : premièrement manipuler les modèles pour prédire le flux de biomasse, estimer la productibilité de certains métabolites ou encore simuler l'effet d'une délétion. Deuxièmement, associer le modèle avec des bases de données métaboliques comme KEGG ou EcoCyc. Enfin comparer les résultats des simulations avec des résultats expérimentaux.
NemoStudio permet aussi d'exporter les modèles au format SBML. La plate-forme repose sur une base de données relationnelle, nommée CycSim (Figure 31). Cette base a été conçue avec un point de vue modélisation, ce qui a comme conséquence la création de quatre classes de tables.
L'objectif de ce paragraphe n'est pas d'expliciter les tables et champs de cette base de données mais d'en donner les fondements et les principales caractéristiques. La première classe est dédiée à la définition d'un organisme : un nom, un identifiant, une description et une liste de gènes. La deuxième est la définition du modèle, avec un nom, un identifiant, une description et une liste d'activations. Si pour le moment un organisme est relié à un modèle, il est possible de relier un organisme à plusieurs modèles ; par exemple deux modèles provenant de travaux différents, ou bien l'historique des améliorations et des versions d'un même modèle. L'inverse est également possible : en prévision de l'évolution des CBMs, la possibilité d'intégrer des méta-modèles a été pris en compte (un méta-modèle est un modèle comprenant plusieurs organismes). L'activation est un concept abstrait qui relie les réactions au modèle. Il existe différents types d'activations : GPR, spontanée et inconnue. Elles sont spécifiques du modèle et leur présence indique que la réaction associée peut être active dans le modèle. La troisième classe de table, concerne les données des CBMs : à l'instar des bases métaboliques généralistes, elle référence les différentes réactions et métabolites, indépendamment des modèles. Une particularité par rapport aux bases de données usuelles : les réactions n'utilisent pas de métabolites, mais des métabolites localisés. La localisation des métabolites est essentielle en modélisation : pour avoir lieu, l'enzyme et les métabolites doivent être présents dans le bon compartiment. Cette particularité impose la présence de compartiments dans les modèles. Les métabolites ne sont pas directement reliés aux modèles, mais dépendent des associations {activations, réactions} et {compartiments, modèle}. La dernière classe concerne les références croisées vers d'autres ressources. Si à première vue cette classe paraît secondaire, j'aurai l'occasion aux cours de ce manuscrit de montrer l'importance de ces liens.

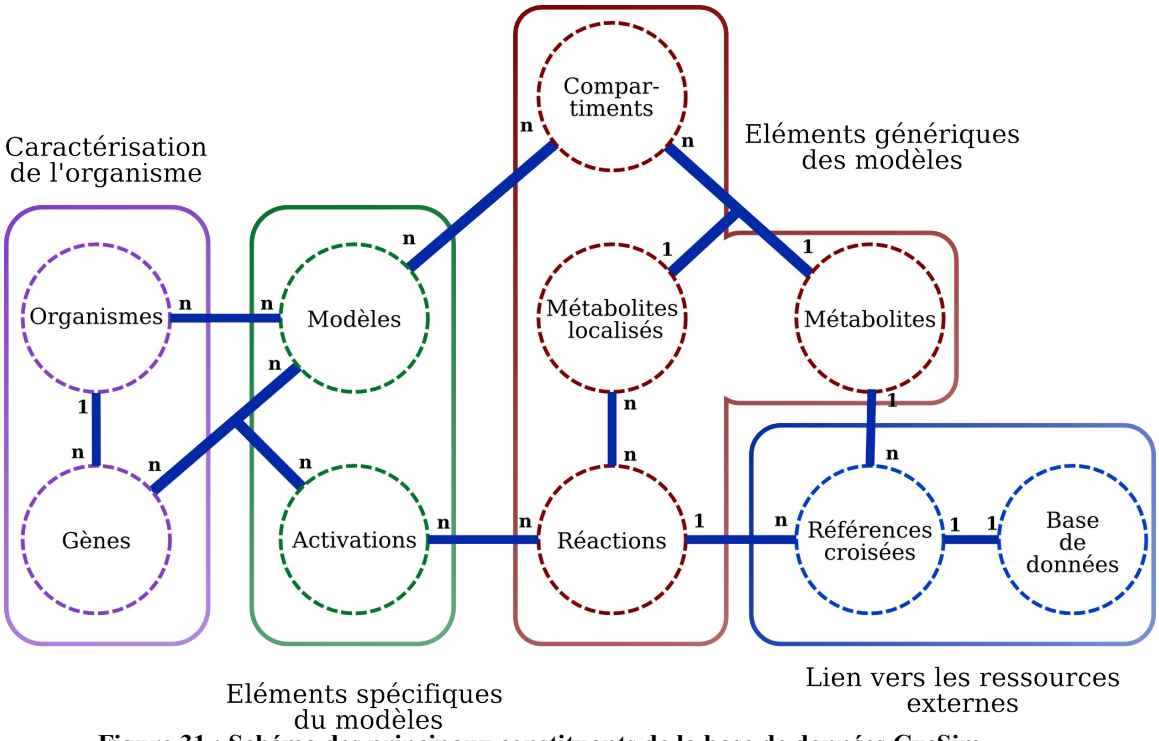


Figure 31 : Schéma des principaux constituants de la base de données CycSim.
La base peut être divisée en quatre sous-parties : en violet les éléments qui caractérisent un organisme (nom, identifiant, description, gènes) ; en vert, son pendant modélisé (un nom, un identifiant, une description et une liste d'activations qui définissent les réactions pouvant avoir lieu dans le modèles) ; en rouge les éléments communs à l'ensemble des modèles (les compartiments cellulaires, les réactions, les métabolites et leur localisation) ; enfin en bleu, les liens vers les différentes ressources. Les *1* et *n* indiquent le nombre d'associations possible entre deux tables.

Escherichia coli

5 Généralité et organisme modèle

La plupart des connaissances actuelles en biologie, des propriétés les plus simples aux mécanismes les plus complexes, ont été découvertes dans des organismes dits organismes modèles. Il en existe pour les différents règnes du vivant et sous familles : procaryote, eucaryote, plante, animal, mammifère, insecte, batracien (Table 11).

Virus			*Phage lambda*
Procaryote			*E. coli*
			B. subtilis
Eucaryote	Unicellulaire		S. cerevisiae
	Multicellulaire	Invertébré	*C. elegans*
			Drosophila melanogaster
		Vertébré	*Xenopus levi*
			Mus musculus
		Plante	*A. thaliana*

Table 11 : Liste des différents organismes modèles.
La liste est non exhaustive.

Les organismes modèles sont à l'origine des organismes présentant des caractéristiques communes : ils sont utilisables en laboratoire (c'est à dire qu'on peut recréer des environnements artificiels pour les faire vivre), ils ont des cycles de division relativement courts, et ils ne présentent aucun risque infectieux ou de contamination. On peut également ajouter qu'ils sont facilement manipulables en particulier pour les modifications génétiques.
Les organismes modèles sont à l'origine de grandes avancées en biologie ; puisqu'ils focalisent l'étude sur un même organisme, ils facilitent la mise en relation des différents processus découverts. Travailler sur un organisme modèle présente de nombreux avantages (Fields & Johnston 2005) : la somme des connaissances sur celui-ci permet d'aller toujours plus loin dans la compréhension du vivant, ce qui explique pourquoi justement les grandes découvertes fondamentales ont étés faites sur ces organismes (cycle cellulaire, synthèse d'ADN, voies métaboliques essentielles (Romano & T Conway 1996)). Il existe une communauté importante derrière chacun d'eux, ce qui rend possible plus d'échanges et plus de collaborations. Enfin de nombreuses méthodes et protocoles expérimentaux sont disponibles : par exemple la banque des mutants *Keio* qui est une collection de simples délétions de gènes chez *E. coli* (Yamamoto et al. 2009). Les organismes modèles ne présentent pas que des avantages car ils entrainent un biais important des connaissances biologiques ; la quantité d'informations et d'outils disponibles pour ces organismes provoquent un effet « boule de neige » ; par commodité lors d'une nouvelle étude on choisira un organisme modèle, produisant des nouvelles données qui inciteront d'autres équipes à travailler sur cet organisme etc. Si on possède des organismes modèles pour différentes sortes d'organismes vivants, la question de leur représentativité au sein de leurs espèces, et plus généralement au sein de la diversité biologique, est souvent remise en cause. Cependant il est

impossible de répondre à cette question ; on est incapable d'estimer cette biodiversité. De plus il reste suffisamment d'inconnu dans ces organismes pour justifier l'intérêt qu'ils suscitent : par exemple chez *E. coli* K-12 MG1655 encore 20% des gènes ont une fonction inconnue (Touchon et al. 2009).
La reconstruction du métabolisme à l'échelle de la cellule nécessite une quantité de données que seul un organisme modèle peut apporter. Ce domaine est relativement jeune et pour le moment, même s'il existe des exemples de reconstruction de modèles métaboliques d'organismes multicellulaires (Jerby et al. 2010), la plupart des travaux portent sur des organismes unicellulaires. C'est pourquoi nous avons choisi l'espèce bactérienne *Escherichia coli* dont l'organisme modèle est la souche K-12MG1655.

6 *E. coli* comme organisme modèle, historique

Escherichia coli (Figure 32) est l'un des organismes modèles les plus étudiés, ce qui en fait la bactérie la mieux connue. C'est Theodor Escherich, pédiatre Allemand, qui isola pour la première fois la bactérie en 1885. Identifiée dans les fèces d'individus sains, il nomma la bactérie *Bacterium coli*, coli étant pour colon son habitat. Rebaptisée *Bacillus coli* dix ans après, c'est finalement en 1919 qu'elle obtient son nom définitif : *Escherichia coli*, en l'honneur d'Escherich (Castellani & Chalmers 1919).

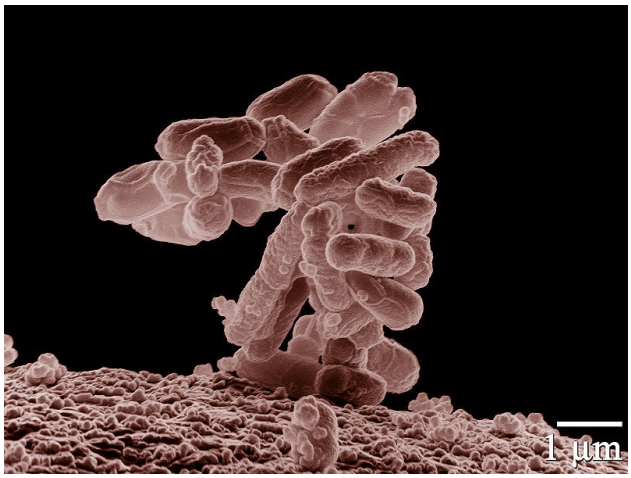


Figure 32 : E. coli au microscope électronique.
Source Agricultural Research Service **http://www.ars.usda.gov/**

La souche de référence de *E. coli* est la souche K-12MG1655 : c'est une souche de laboratoire, qui est incapable de vivre à l'état naturel ; elle a permis de grandes avancées dans la génétique, la biologie moléculaire et cellulaire, et dans le métabolisme. Si dans un premier temps elle était considérée comme représentative de l'espèce, actuellement avec l'augmentation considérable des souches disponibles et de la diversité qu'elles apportent, cette hypothèse est maintenant remise en cause (Touchon et al. 2009). Cela ne signifie pas que K-12 MG1655 n'est pas représentative de l'espèce mais simplement qu'une seule souche ne peut pas représenter l'ensemble de la diversité des *E. coli* (Hobman et al. 2007). Si les souches peuvent avoir différentes capacités, elles sont toutes des bacilles à Gram négatif, non sporulant, aérobique ou anaérobique, elles possèdent un habitat primaire (tube digestif des animaux) et un habitat secondaire (les sols et l'eau).

.6.1 Mode et cycles de vie de *E. coli*

Des prélèvements et analyses d'échantillons ont révélé la présence d'espèce *E. coli*, sur un peu près toute la surface de la terre, que ce soit des zones habitées par l'Homme ou plus sauvages (Skurnik et al. 2006). De même, ces hôtes présentent des caractéristiques très différentes ; ainsi on retrouve des *E. coli* dans des animaux peu importe leur régime alimentaire (omnivore, carnivore etc.), l'anatomie intestinale et la durée du transit. Cependant ces éléments ont un effet sur leur présence et leur prévalence au sein de la flore intestinale. La prévalence est le nombre d'individus de la souche comparé au nombre d'individus total. Sa prévalence dans les selles humaines est de 100% et peut descendre à 10 % pour les reptiles (Gordon & Cowling 2003; Skurnik et al. 2006). Les souches sont généralement spécifiques d'un type d'hôte ; chez l'Homme c'est l'une des premières espèces à coloniser le tube digestif des nouveau-nés (Penders et al. 2006, p.penders). Chez l'adulte sain cette colonisation est restreinte à quelques souches, généralement en dessous de cinq (Sears & Brownlee 1952). Une de ces souches va devenir résidente et sera présente pendant plusieurs mois voire des années. Les autres souches sont dites transitoires et ont une durée de présence de l'ordre du jour ou de la semaine ; une fois qu'une souche est devenue résidente, il est difficile pour une autre souche de prendre sa place (Sears et al. 1950; Cooke et al. 1972). Toutes ces valeurs sont données à titre indicatif, en effet la grande variabilité des régimes alimentaires de l'Homme, les conditions d'hygiène et de préparation des aliments, vont avoir un effet sur le nombre de souches différentes ingérées. Et on peut facilement comprendre que plus un individu va être en contact avec des souches différentes plus il a de chance d'avoir un grand nombre de souches transitoires et par la même occasion, de souches résidentes.

Chez l'être humain les *E. coli* vivent la majorité du temps dans le tube digestif et en particulièrement dans le colon. Plus précisément, elle réside dans le mucus qui protège la couche épithéliale de l'intestin ; environnement dans lequel l'oxygène est en faible quantité : on parle de micro-aérobie. Parmi toutes les sources de carbones disponibles dans l'intestin, les *E. coli* utilisent préférentiellement le *gluconate*, mais elles sont capables d'utiliser de nombreuses autres sources dont le *ribose*, le *fucose*, le *mannose* etc. (D.-E. Chang et al. 2004). La présence d'*E. coli* dans l'intestin de l'hôte est bénéfique à ce dernier : en effet la colonisation, en plus de stimuler la défense immunitaire, protège l'hôte d'agents infectieux (Vollaard & Clasener 1994; Macpherson et al. 2000).

E. coli a été identifiée dans les selles, et peut se retrouver dans l'environnement de cette manière. Hors de son hôte elle doit être capable de vivre, se reproduire pour pouvoir coloniser un nouvel hôte. Dans ce nouvel environnement elle rencontre des conditions de vie différentes, par exemple le milieu n'est plus micro-aérobique. Bien que cela soit difficile à estimer, on considère que près de la moitié des *E. coli* vivent dans l'environnement et l'autre moitié dans des êtres vivants. Cette capacité d'adaptation à deux environnements totalement différents est le reflet des capacités métaboliques et de la flexibilité de son réseau métabolique, ce qui en fait un très bon candidat pour les thématiques de ce travail de thèse.

.6.2 Diversité et phylogénie

Les capacités d'adaptation sont le résultat de la plasticité génomique des *E. coli*. En effet les évènements de mutations, recombinaisons et transfert horizontaux ne sont pas des événements rares (Touchon et al. 2009; Tenaillon et al. 2010). Ceux-ci

interviennent de façon ponctuelle. Par conséquent il est possible de retrouver l'histoire évolutive des différentes souches. Les *E. coli* se répartissent dans cinq groupes phylogénétiques : A, B1, B2, D, E et F (Tenaillon et al. 2010) et Figure 33. Ces groupes ne sont pas figés et évoluent en même temps que les connaissances : depuis peu le groupe D est scindé en un groupe D et un groupe E (Jaureguy et al. 2008). Au sein de l'arbre phylogénétique des *E. coli*, se trouve une autre espèce, les *Shigella,* sur laquelle je reviendrai dans la partie suivante (Figure 34). Cette présence dans l'arbre prouve une origine commune et une divergence évolutive récente entre ces deux espèces. La proximité des espèces du genre *Escherichia* rend souvent difficile le classement des souches, et ce sont des marqueurs métaboliques qui permettent de différencier les espèces *E. coli*, *E. albertii* et *E. fergusonii*. Parfois ce sont les divergences de séquences qui permettent d'affilier les souches à des espèces. Ainsi en 2009 des souches, considérées comme des *E. coli* et partageant les différents critères caractéristiques des *E. coli*, ont finalement été réassignées dans cinq nouveaux clades (Walk et al. 2009).
Le nombre de souches de *E. coli* varie d'un individu à un autre est d'une région à une autre. Il en est de même pour la proportion des différents groupes phylogénétiques (Table 12). Elle peut varier du simple au triple entre les différents pays ; de plus on constate qu'elle n'est pas fixe avec le temps ; ainsi en France on est passé d'une majorité de souches du groupe phylogénéitque A en 1980 à une répartition pratiquement uniforme entre les groupes A, B1, B2 et D en 2000.

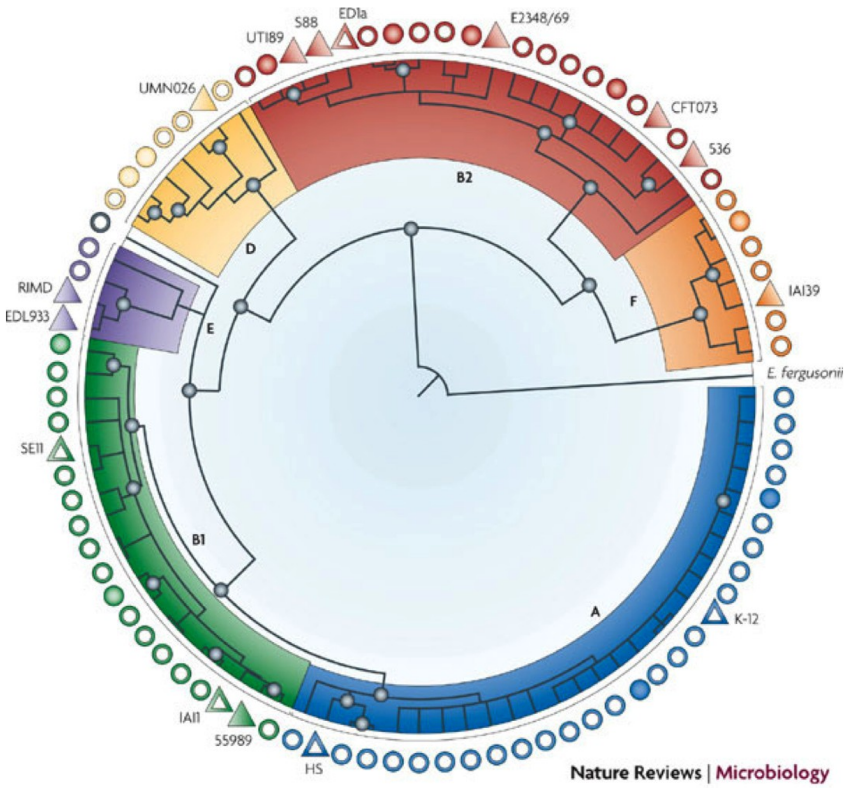


Figure 33 : Arbre phylogénétique des *E. coli*.
L'arbre est basé sur 8 gènes de ménage présent dans 72 souches. Les triangles symbolisent les 15 souches de référence de *E. coli*. Les symboles vides représentent des souches commensales. Les points sur les nœuds indiquent que le nœud est supporté à 80% (d'après (Tenaillon et al. 2010))

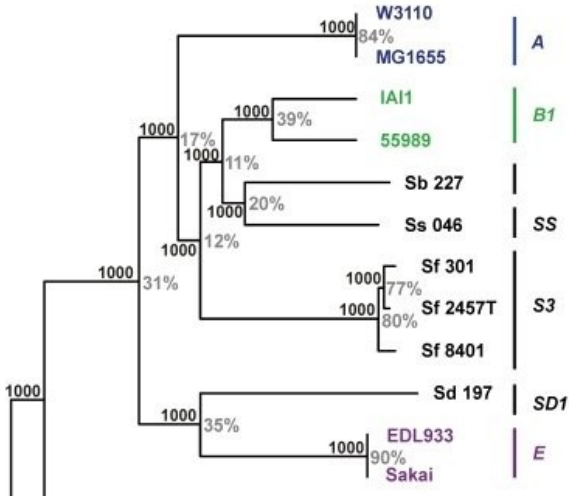


Figure 34 : Sous arbre phylogénétique des *E. coli* et *Shigella*.
Les *Shigellas* sont inclus au milieu des *E. coli* d'après (Touchon et al. 2009).

Population	Groupe phylogénétique			
	A	B1	B2	D
France (1980)	61	12.5	10.5	16
Croatie	35	32	19	14
Mali	24	58	2	16
Bénin	50	32.5	17.5	0
Pakistan	47	18	12	23
Guyane Française	63.5	20.5	3	13
Bolivie	77	10	5	8
Colombie	57	3.5	25	14.5
France (2000)	25.5	21	29.5	24
Suède	29	11	46	14
USA	20.5	12.5	48	19
Japon	28	0	44	28
Australie	19.5	12.5	45	23

Table 12 : Prévalence des groupes phylogénétiques chez les êtres humains.
D'après (Tenaillon et al. 2010).

Si les groupes phylogénétiques sont basés sur des séquences nucléotidiques communes, la diversité, elle, provient des séquences qui ne sont pas partagées par toutes les souches. On estime à un peu moins de 2000 le nombre de gènes communs entre toutes les souches d'*E. coli*, ce qui est élevé quand on considère les 4300 gènes de la souche K-12 MG1655. Pourtant, ce chiffre ne représente que 11% du nombre total de gènes différents chez les souches d'*E. coli* aujourd'hui séquencées, soit plus de 18000 gènes (Touchon et al. 2009).
C'est parmi les 16000 gènes variables que se trouvent les différentes capacités d'adaptation et de colonisation des *E. coli*. C'est aussi parmi eux que se trouvent les gènes responsables du caractère pathogène de certaines souches. (Tenaillon et al. 2010)

.6.3 Commensalisme et pathogénicité

La majorité des souches de *E. coli* sont commensales, et vivent sans importuner leur hôte. L'autre partie des souches sont pathogènes et entrainent des problèmes de santé chez leur hôte humain ou animal. Si certaines infections sont bénignes, *E. coli* provoque la mort de plus de 2.5 millions de personnes par an, ce qui fait de cette

espèce l'un des agents les plus mortels qui existent (Thomas A Russo & James R Johnson 2003). Il n'existe pas un mode unique de virulence chez les *E. coli*, et une souche peut très bien rester dans l'hôte sans provoquer de symptôme. L'agressivité de la bactérie va dépendre de l'état de l'hôte et de ses défenses immunitaires : c'est pourquoi on dit que les *E. coli* sont des pathogènes opportunistes. La séparation entre *E. coli* commensales et pathogènes n'est pas stricte, et la nature de l'hôte est importante : la souche APEC O1 est un pathogène des oiseaux (*Avian Pathogenic E. coli*), tandis que la souche du sérotype O157:H7 est spécifique des Hommes. Il existe deux classes de pathogènes : les pathogènes intestinales ou InPEc (*Intestinal Pathogenic E. coli*) et les pathogènes hors de l'intestin ou ExPEc (*Extra-intestinal Pathogenic E. coli*). Chacune de ces classes est divisée en pathovars, aux nombres de deux pour les ExPEc et de 6 pour les souches InPEc (Nataro & Kaper 1998; T A Russo & J R Johnson 2000). Les souches ExPEc sont séparées en fonction de leur lieu d'infection (Croxen & Finlay 2010) : ainsi on trouve les *E. coli* uropathogènes ou UPEC (*UroPathogenic E. coli*) qui vont provoquer des infections dans le système urinaire (rein et vessie), et les *E. coli* méningitiques du nouveau-né ou NMEC (*Neonatal Meningitidis E. coli*) qui agissent dans le cerveau. Les deux pathovars peuvent être trouvés dans le sang puisque c'est ainsi que les bactéries atteignent leur cible. Les ExPEc peuvent aussi provoquer des septicémies et des infections respiratoires.

Trois des pathovars de InPEc agissent dans le gros intestin (Figure 35): Les *E. coli* entéro–aggrégatives ou EAEC *(EnteroAggregative E. coli*), les *E. coli* entérohémorragiques ou EHEC (*EnteroHaemorrhagic E. coli*), les *E. coli* entéroinvasifs ou EIEC (*EnteroInvasive E. coli*) et les *Shigella*. Dans l'intestin grêle, on retrouve les EAEC et les trois autres pathovars (Figure 35): Les *E. coli* entérotoxinogènes ou ETEC (*EnteroToxigenic E. coli*), les *E. coli* entéropathogènes ou *EPEC* (*EnteroPathogenic E. coli*) et les *E. coli* à adhérence diffuse ou DAEC (*Diffusely Adherent E. coli*) ; Les symptômes des EPEC, ETEC et DAEC sont de violentes diarrhées.

Les EIEC forment le seul pathovar qui est toujours virulent contrairement aux autres, il en est de même pour les *Shigella*. Dans la partie précédente j'ai évoqué les *Shigella* qui sont présentes dans l'arbre phylogénétique des *E. coli*. L'une des principales différences entre ces deux espèces est le mode de vie : si les *E. coli* sont prototrophes, les *Shigella* sont auxotrophes et sont des parasites intra-cellulaires, c'est à dire qu'elles vivent à l'intérieur des cellules de l'hôte et utilisent les métabolites essentiels produits par celui-ci pour survivre.

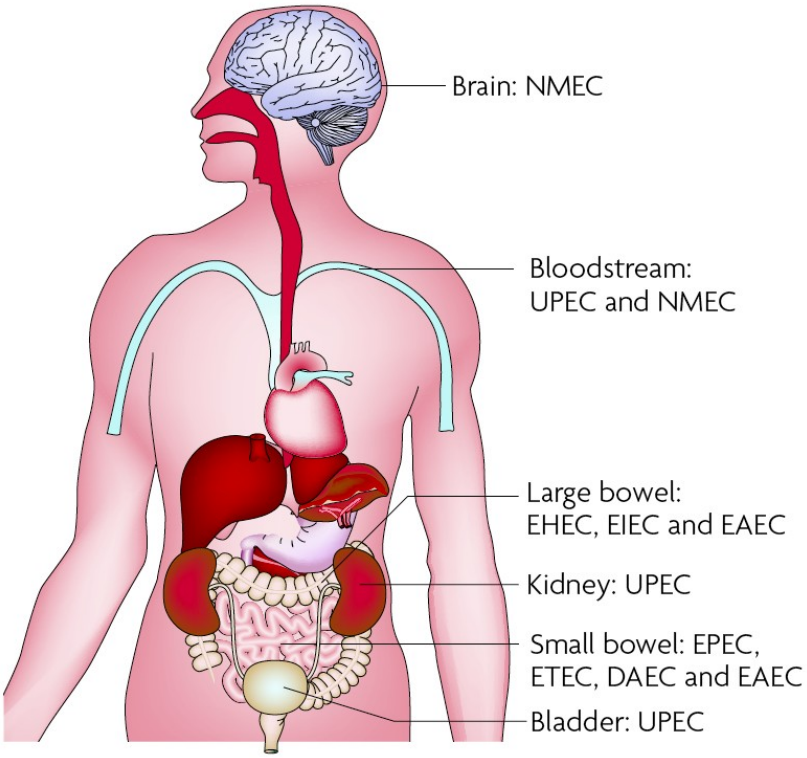


Figure 35 : Zones d'infections des différents pathovars.
D'après (Croxen & Finlay 2010).

D'un côte les souches d'*E. coli* sont divisées en groupes phylogénétiques, de l'autre elles sont divisées en pathovars. Naturellement on se demande s'il existe un lien entre les deux. Aucun élément ne permet de relier ces deux aspects ; on note néanmoins certaines tendances : ainsi très peu de souches des groupes A et B1 sont pathogènes et une grande partie des souches B2 sont des ExPEc (B Picard et al. 1999; Jaureguy et al. 2008; Touchon et al. 2009). Ces travaux ont montré qu'il n'existe pas de gènes qui soient spécifiques du commensalisme ou de la pathogénicité ; toutefois certains gènes, appelés facteurs de virulence, sont connus pour les ExPEc. Cette absence d'éléments a fait naître l'hypothèse selon laquelle la virulence est un effet secondaire de l'adaptation et qu'elle confère un avantage à la manière de la compétition entre *Staphylocoque* et *Penicilium* (page 17). Ainsi les toxines émises ne ciblent pas l'hôte mais des bactéries concurrentes pour la colonisation ; malheureusement l'hôte est quand même affecté par celles-ci (Le Gall et al. 2007).

Problématiques et objectifs de la thèse.

De plus en plus de génomes séquencés et annotés d'*E. coli* sont disponibles. Les comparaisons génomiques font ressortir une forte conservation des gènes liés au métabolisme. Cependant la diversité génomique est tellement importante qu'il est difficile d'extraire un signal fort et un lien clair entre le génotype et le phénotype. A côté de ces génomes, différents types de données expérimentales commencent à apparaître. La mise en relation de ces données est un exercice délicat sans cadre approprié.
Parmi les cadres possibles, les modèles à base de contraintes semblent être de bons candidats de par leur flexibilité aux données, surtout depuis qu'il est possible de les reconstruire et de manipuler à l'échelle de la cellule.
Grâce aux connaissances accumulées sur certains organismes, on est capable de reconstruire des réseaux du métabolisme à l'échelle de la bactérie. Ils permettent de savoir quelles sont les réactions, les voies de dégradation et de synthèse qui sont présentes dans l'organisme. Le caractère descriptif des réseaux empêche cependant de savoir si ces voies sont fonctionnelles. Pour estimer les capacités métaboliques et la quantité de matières qui traverse les réactions, il est nécessaire de convertir le réseau métabolique en un objet mathématique : le modèle métabolique.
Que ce soit la reconstruction du réseau ou sa conversion en modèle, le temps et les ressources requises sont considérables ; ils sont de l'ordre du couple d'années.
Il existe des moyens de reconstructions automatiques pour les réseaux ; ils diminuent le temps requis. Malheureusement le résultat obtenu est un réseau d'une qualité bien en deçà de ce qui est produit par une expertise manuelle.

Nous souhaitons étudier la diversité métabolique des *E. coli*, pour établir des liens entre le génome, et le phénotype et ce, par l'intermédiaire de réseaux et modèles du métabolisme à l'échelle de la bactérie. Ils permettront d'étudier les différences de constitutions et les différentes capacités métaboliques d'une vingtaine d'*E. coli*. Ils formeront également un cadre pour l'intégration et la mise en relation de données expérimentales hétérogènes dans des modèles à base de contraintes.

Pour cela, nous devons, d'une part, reconstruire les réseaux et modèles du métabolisme global, et d'autre part analyser et comparer les réseaux et modèles obtenus.

Nous avons choisi d'orienter nos travaux autour de deux axes :

- Le premier, méthodologique, concerne la mise en place d'un processus constitué de modules successifs et dont l'objectif est la reconstruction automatique et rapide des réseaux puis des modèles avec un critère de qualité élevée. Ces processus doivent également assurer une homogénéisation des données pour pouvoir passer du génome, au modèle, en passant par le réseau, et cela sans la moindre ambigüité.

- Le second porte sur l'analyse de la diversité métabolique, avec l'aide de données expérimentales. Cet axe s'intéresse à l'analyse des différences et des similitudes entre les souches, par l'intermédiaire des réseaux et des modèles, afin d'expliciter les liens entre génomes, capacités métaboliques et phénotypes.

Chapitre I :
Reconstruction et analyses des réseaux métaboliques

1 Article sur les réseaux métaboliques

La première partie de ce chapitre reprend l'article qui présente la méthode de reconstruction et les différentes analyses réalisées sur ceux-ci.
Les figures supplémentaires S1, S2 et S3 sont données en annexe (respectivement Annexe 1, Annexe 2, Annexe 3). Les annexes 1 et 2 étant trop volumineuses, elles sont disponibles à ces adresses :
Annexe1 http://www.genoscope.cns.fr/agc/doc/GVthese/Annexe1-Complete.xls
Annexe 2 http://www.genoscope.cns.fr/agc/doc/GVthese/Annexe2-Complete.xls

JOURNAL OF BACTERIOLOGY, Mar. 2011, p. 1461–1472 Vol. 193, No. 6
0021-9193/11/$12.00 doi:10.1128/JB.01192-10

Core and Panmetabolism in *Escherichia coli*[V]†

Gilles Vieira,[1]* Victor Sabarly,[2,3] Pierre-Yves Bourguignon,[1]‡ Maxime Durot,[1] Fran¸cois Le F`evre,[1] Damien Mornico,[1] David Vallenet,[1] Odile Bouvet,[2] Erick Denamur,[2] Vincent Schachter,[1]§ and Claudine M´edigue[1]

CNRS UMR 8030, Universit´e d'Evry, CEA, IG, Genoscope, 2 rue Gaston Cr´emieux, CP5706, F-91057 Evry Cedex, France[1]; INSERM U722 and Universit´e Paris Diderot, 16 rue Henri Huchard, 75018 Paris, France[2]; and INRA, UMR de G´en´etique V´eg´etale, INRA/CNRS/Universit´e Paris-Sud/AgroParistech, Ferme du Moulon, F-91190 Gif sur Yvette, France[3]

Received 5 October 2010/Accepted 3 January 2011

Escherichia coli **exhibits a wide range of lifestyles encompassing commensalism and various pathogenic behaviors which its highly dynamic genome contributes to develop. How environmental and host factors shape the genetic structure of *E. coli* strains remains, however, largely unknown. Following a previous study of *E. coli* genomic diversity, we investigated its diversity at the metabolic level by building and analyzing the genome-scale metabolic networks of 29 *E. coli* strains (8 commensal and 21 pathogenic strains, including 6 *Shigella* strains). Using a tailor-made reconstruction strategy, we significantly improved the completeness and accuracy of the metabolic networks over default automatic reconstruction processes. Among the 1,545 reactions forming *E. coli* panmetabolism, 885 reactions were common to all strains. This high proportion of core reactions (57%) was found to be in sharp contrast to the low proportion (13%) of core genes in the *E. coli* pangenome, suggesting less diversity of metabolic functions compared to that of all gene functions. Core reactions were significantly overrepresented among biosynthetic reactions compared to the more variable degradation processes. Differences between metabolic networks were found to follow *E. coli* phylogeny rather than pathogenic phenotypes, except for *Shigella* networks, which were significantly more distant from the others. This suggests that most metabolic changes in non-*Shigella* strains were not driven by their pathogenic phenotypes. Using a supervised method, we were yet able to identify small sets of reactions related to pathogenicity or commensalism. The quality of our reconstructed networks also makes them reliable bases for building metabolic models.**

Escherichia coli is a versatile species encompassing commensal organisms, as well as intraintestinal *E. coli* (InPEc) and extraintestinal *E. coli* (ExPEc) pathogens (27, 49). This variety of lifestyles has been seen as a consequence of the huge *E. coli* genome plasticity (51). However, linking genomic elements to phenotypic behaviors is not trivial because several layers of biological processes separate genes from their phenotypic effects, and in extreme cases, the evolutionary path can lead either to the functional convergence of distinct sets of genes or to the functional divergence of an initially common set of genes. Consequently, in order to establish links between ge- nomes and phenotypes, one needs an integrative layer. A re- cent study on a set of 20 *E. coli* strains (51) has shown that a large fraction of the shared genomic elements with known function is related to metabolism. Because it is now feasible to reconstruct metabolic networks at the genome scale (7, 13, 16, 26), these metabolic networks can, in principle, be used as functional bridges between genomic diversity and phenotypic differences. Currently, such reconstructions are performed automatically from the annotation of input genomes, using algorithms that match these annotations with the contents of reference metabolic databases (13, 16).

In this work, we studied the metabolic diversity of the *E. coli* species from an evolutionary point of view, with a focus on (i) the extent of metabolic diversity compared to that of genomic diversity, (ii) the correlation between metabolic diversity and phylogeny, and (iii) the metabolic functions associated with pathogenicity.

To these ends, we reconstructed and compared the meta- bolic networks of 29 strains of *E. coli*, for which genome se- quences and annotations were available (51). This set of strains comprises 23 *E. coli* strains covering all main phylogenetic groups (A, B1, B2, D, E, and F) (11) and various pathogenic or nonpathogenic behaviors (commensal, ExPEc, InPEc), as well as 6 *Shigella* strains, which are human obligate intraintestinal pathogens belonging to the *E. coli* species (15, 44). To obtain metabolic networks suitable for comparative analyses, we first developed a high-quality automated reconstruction process which builds homogenized genome annotations and combines metabolic evidence from the EcoCyc and MetaCyc databases (7, 28a). This reconstruction process is also able to infer en- zyme complexes by similarity with K-12 MG1655 complexes. In a second step, we defined the core and variable parts of *E. coli* metabolic networks and analyzed their metabolic roles. We then confronted differences in metabolic networks with *E. coli* phylogeny and phenotypes to assess which factors influenced most changes in *E. coli* metabolism. As differences were

* Corresponding author. Mailing address: Laboratoire d'Analyses Bioinformatiques pour la G´enomique et le M´etabolisme, CEA/IG/ Genoscope, 2 rue Gaston Cr´emieux, CP5706, F-91057 Evry Cedex, France. Phone: 33 1 60 87 36 07. Fax: 33 1 60 87 25 14. E-mail: gvieira@genoscope.cns.fr.

† Supplemental material for this article may be found at http://jb.asm.org/.

‡ Present address: Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, D-04103 Leipzig, Germany.

§ Present address: TOTAL Gas and Power, 2 place Jean Miller, La D´efense 6, F-92078 Paris La D´efense Cedex, France.

[V] Published ahead of print on 14 January 2011.

found to be uncorrelated with phenotypes, we finally performed a supervised search for metabolic differences specific to *E. coli* pathogenic phenotypes.

MATERIALS AND METHODS

Reannotation of *E. coli* genomes. Building upon a previous annotation work performed for 20 *E. coli*/*Shigella* strains in the context of the ColiScope project (51) with the MicroScope platform (52), we added nine newly published *E. coli* genomes (strains ATCC 8739, E24377A [45], SE11 [38], LF82 [35], O127:H6 E2348/69 [23], O157:H7 EC4115, HS [45], 042 [9], and SMS-3-5 [17]). All of these publicly available genomes were reannotated using the following process. First, all genomes were integrated in the ColiScope database using MICheck, a method which enables rapid verification of sets of annotated genes and frameshifts in previously published bacterial genomes (10). Second, functional annotations of our previously annotated *E. coli* genes were automatically transferred in the new strains to genes showing very strong sequence similarity (85% identity on at least 80% of the length of the smallest protein). The remaining genes, i.e., those without any ortholog in any ColiScope genome, were left with their original functional annotations. All genome annotations are available through the MicroScope web platform (http://www.genoscope.cns.fr/agc/microscope/coliscope).

Metabolic network reconstruction. Our metabolic network reconstruction process is mostly based on Pathway Tools (version 14.0), which is the BioCyc reconstruction software (28), and its associated metabolic database, MetaCyc (7). We used as input all genome annotations coming from our reannotation process, including genes, pseudogenes, partial genes, and insertion sequence-like and prophage-like elements.

By default, Pathway Tools associates genes with metabolic reactions from MetaCyc by examining gene ontology terms, gene product names, and EC number terms found in the genome annotation. Those reactions will be denoted matched reactions. Due to wrongly formatted or unspecific EC numbers or insufficiently explicit textual annotations, Pathway Tools may in some cases either overpredict or miss enzymatic reactions. To improve the accuracy of this gene-reaction association step, we exploited the expert curation done in the EcoCyc metabolic database for *E. coli* K-12 MG1655 (28a) by transferring gene-reaction associations found in EcoCyc to orthologous genes in the other strains. For this, we mapped genes from K-12 MG1655 to genes of each *E. coli* strain using the best bidirectional hit (BBH), computed by BLAST (2), with similarity rates above 70% and overlap above 80% of the shorter gene length. Direct associations between each gene having an ortholog in K-12 MG1655 and the corresponding EcoCyc reactions were then specified in a dictionary file given as an additional input to Pathway Tools. Pathway Tools was finally executed using this file and the homogenized genome annotations. All reconstructed networks are available from the Metacoli project website (http://www.genoscope.cns.fr/agc/metacoli) and are included in the MicroCyc repository (http://www.genoscope.cns.fr/agc/microcyc).

Since Pathway Tools infers full metabolic pathways (28), some reactions lacking an associated gene were retrieved on the basis of their presence in an inferred pathway. These purely inferred reactions were left in the MicroCyc databases to allow users to examine complete metabolic pathways but were removed for all comparative analyses done in this work.

Similarly, reactions associated only with pseudogenes were kept in the MicroCyc databases but were removed from our comparative analyses.

The occurrences of all reactions (gene-associated, inferred, pseudogene-associated, and spontaneous reactions) can be found in Table S1 in the supplemental material.

Inference of complexes. Even though BioCyc databases are able to represent protein complexes, the Pathway Tools reconstruction software does not automatically infer them. Benefiting from the protein complexes stored in EcoCyc for *E. coli* K-12 MG1655, we inferred by homology complexes for all strains using the following procedure.

First, for each protein complex experimentally identified in *E. coli* K-12 MG1655 and extracted from EcoCyc, we recursively analyzed its composition in terms of subunits. An equivalent subunit was inferred in the studied *E. coli* strain if and only if we could find in its genome an orthologous polypeptide using BBH computed by BLAST (2). Second, when an orthologous complex could be inferred, the functional annotations of the K-12 MG1655 complex were transferred to the reconstructed protein complex. Third, the functional annotations associated initially with each subunit of the complex were deleted if they were shared with the reconstructed complex. This final step ensures that the enzymatic function is held only by the complex, if appropriate. This procedure was implemented using the CyClone application programming interface (31), and all complexes are directly stored with the metabolic networks in the MicroCyc repository (http://www.genoscope.cns.fr/agc/microcyc). The list of inconsistencies raised during the complex reconstruction process (i.e., complexes with missing subunits) is available in Table S2 in the supplemental material.

Computation of pan- and core genome/metabolism. To compute pan- and core genomes, we considered genes that were not pseudogene, partial gene, insertion sequence-like, or prophage-like elements. We clustered genes using the orthoMCL program (version 1.4) (32) for proteins with similarities above 70% and overlap above 70%. We obtained 14,986 clusters of genes that we called the pangenome and 1,957 clusters encompassing at least one gene from each strain that we called the core genome. To evaluate how core and pangenomes evolve when strains are added or removed, we computed them as a function of the number of strains for 5,000 random input orders of strains.

Similar analyses were conducted on metabolic networks. Core metabolism was defined as the set of reactions present in all strains, and panmetabolism was defined as the set of all reactions of all strains. Core metabolism was composed of 885 reactions, and panmetabolism contained 1,545 reactions. Evolution of the sizes of core and panmetabolism was studied by computing them for 10,000 random input orders of metabolic networks.

Computation of genetic distances and phylogenetic tree. We computed the phylogenetic tree using a six-step procedure. (i) First, we built a modified core genome including pseudogenes and the genome of an outgroup reference organism, *Escherichia fergusonii* (29). Gene homologies were determined by nucleotide sequence comparisons of genes with similarities of >80% and coverage of >80%. This modified core genome gathered a set of 1,388 common genes. (ii) We performed multiple alignments on the sequences of these core genes using the MUSCLE program (version 3.6) (14). (iii) Sequence blocks of good alignment were then selected with the GBLOCKS program (version 0.91) (8). (iv) We concatenated those blocks to build one long sequence for each organism. (v) We reconstructed the phylogenetic tree on the basis of these long sequences with the PHYML program (version 3.1) (20), using maximum likelihood and a GTR+gamma model. The genetic distance was directly derived from the branch length of the generated tree. (vi) Finally, 100 bootstrap experiments were performed on the previous step to assess the robustness of the tree topology.

Computation of metabolic distances. We defined the metabolic distance between two metabolic networks to be the number of distinct gene-associated reactions between them. We computed it using reaction occurrence vectors: each component of this vector corresponds to a reaction of panmetabolism and specifies whether the reaction is present (value = 1) or absent (value = 0) in the considered metabolic network. Metabolic distance is therefore directly computed as the Manhattan distance between reaction occurrence vectors, $D(x, y) = \sum_{i=1}^{n} |x_i - y_i|$, for reaction i in reaction occurrence vectors x and y of length n. Using this distance, we created a metabolic tree by neighbor joining with R (46) and the R package ape (40).

MCA. Factorial multiple-correspondence analysis (MCA) is a projection technique that provides a low-dimensional graphical representation of a set of elements by capturing the maximal amount of variability from the variables describing those elements. We conducted an MCA on the reconstructed metabolic networks for the 23 *E. coli* non-*Shigella* strains using R (46) and the package FactoMineR (30). We took as active variables the occurrence of reactions from panmetabolism. Considering the first two eigenvalues was sufficient to explain 34% of the data set diversity. We extracted reactions which had a significant contribution effect on the first two dimensions using the dimdesc function with a multiple-test correction (Bonferroni correction) and a P value lower than 0.05.

Compactness and separation measures. We computed two measures to assess the compactness and separation of phylogenetic and phenotypic groups according to the metabolic distance. We first defined a center for each group by taking the mean of the occurrence vectors of all groups' strains. Group compactness was then defined as the average metabolic distance between the group center and all groups' strains. Separation between two groups was defined as the metabolic distance between the group centers. Both measures were computed in R (46) using the package clv (http://CRAN.R-project.org/package=clv).

Classification tree analysis. We used classification and regression tree analysis (CART) (6), a supervised method, to determine which combinations of reactions separate strains according to their pathogenicity. We used the R (46) package rpart (3) with the Gini index as the criterion of homogeneity to build the trees. We removed reactions from the core metabolism which carry no discriminating information and grouped together reactions with the same occurrence in the strains (called the occurrence profile). We obtained 155 different profiles. We computed three different groups of CARTs: commensal versus other pheno-types, ExPEc versus other phenotypes, and InPEc versus other phenotypes. We

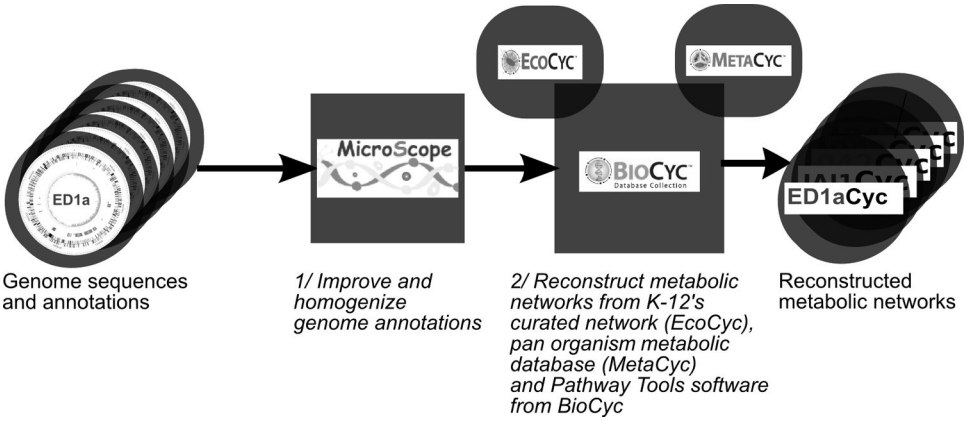


FIG. 1. Metabolic network reconstruction process. Genome annotations are homogenized using the MicroScope platform. Then, metabolic networks are reconstructed with BioCyc software tools using the reference metabolic database EcoCyc to benefit from expert curation on the K-12 MG1655 strain and infer enzymatic complexes and the panorganism metabolic database MetaCyc to retrieve non-K-12 MG1655 reactions.

focused on groups of reactions belonging to the first nodes of the most homogeneous trees.

RESULTS

Reconstruction of metabolic networks. In order to link phenotypic and genomic diversity through metabolism, one needs to accurately pinpoint similarities and differences in metabolic function in the set of strains under scrutiny. Although several tools are provided to automatically reconstruct metabolic networks from genome annotation only (7, 13, 16, 26), their level of accuracy and the completeness of the resulting networks are usually not sufficient to allow detailed downstream analyses, unless manual curation is carried out (13, 16). Here, we exploited the proximity of all strains to the well-studied *E. coli* K-12 MG1655 strain to develop a more efficient automatic reconstruction process. To improve the accuracy of the default BioCyc reconstruction process, our reconstruction strategy uses improved genome annotations: EcoCyc, the highly curated metabolic database for *E. coli* K-12 MG1655 (28a), and Pathway Tools, the BioCyc metabolic reconstruction software (28). This strategy was applied in two steps (Fig. 1).

First, annotations of all *E. coli* genomes were improved and homogenized. In the context of the ColiScope project, an important manual annotation work of the newly sequenced *E. coli* strains was performed on genes and regions not found in K-12 MG1655, thus allowing, at the end of the process, the reannotation of orthologs in the previously available *E. coli* and *Shigella* genomes (51). In the current study, nine new *E. coli* strains have been added to the ColiScope project within the MicroScope platform (52), and their genomes were reanno- tated in terms of both syntactic prediction and functional an- notations on the basis of orthologs available in the ColiScope project (see Materials and Methods). This reannotation pro- cess revealed some inaccurate or missed gene annotations in these new strains and allowed us to standardize the defini- tion and identification of pseudogenes. As a result, a set of consistent functional annotations for all 29 genomes was obtained and made available at the following URL: http: //www.genoscope.cns.fr/agc/microscope/coliscope.

In the second step, we translated all genome annotations, encompassing genes, pseudogenes, and partial genes, into metabolic networks by first identifying metabolic reactions from EcoCyc for genes having orthologs in the K-12 MG1655 ge- nome and then executing Pathway Tools with MetaCyc to translate the annotations of the remaining genes (see Ma- terials and Methods for the detailed procedure). Using the highly curated EcoCyc database as the main pivot to recon- struct the metabolism of all *E. coli* species significantly im- proves the translation efficiency, as shown afterwards, since it prevents Pathway Tools from performing false predictions for genes orthologous to K-12 genes. Previous pivot-based recon- struction methodologies have already been applied to other organisms (37, 50) but were often unable to predict reactions absent from the pivot organisms. Here, our strategy also takes advantage of the panorganism MetaCyc database (7) to con- sider reactions beyond those present in K-12 MG1655. All of our reconstructed networks can be browsed, queried, and down- loaded from the MicroCyc website (http://www.genoscope.cns.fr /agc/microcyc).

Pathway Tools infers full metabolic pathways (28); there- fore, some reactions with no associated gene are retrieved on the basis of their sole occurrence in an inferred pathway. No direct evidence supports these inferred reactions, which often serve as candidates to fill missing biochemical activities (19). Since we kept our reconstruction process fully automatic and performed no further curation on the inferred pathways, we separated these inferred reactions from matched reactions (re- actions associated with genes).

To evaluate the benefits of our optimized strategy, we reconstructed the networks using three increasing levels of improvements and compared their respective qualities. The three levels of reconstruction were done using (i) raw genome annotations directly extracted from the GenBank database and the default Pathway Tools process (strategy a), (ii) updated genome annotations from ColiScope and the default Pathway Tools process (strategy b), and (ii) updated genome annota- tions from ColiScope and the combined EcoCyc/Pathway Tools process (strategy c, our optimized reconstruction pro- cess). We estimated the quality of the reconstructed networks with the following criteria: number of matched reactions in the

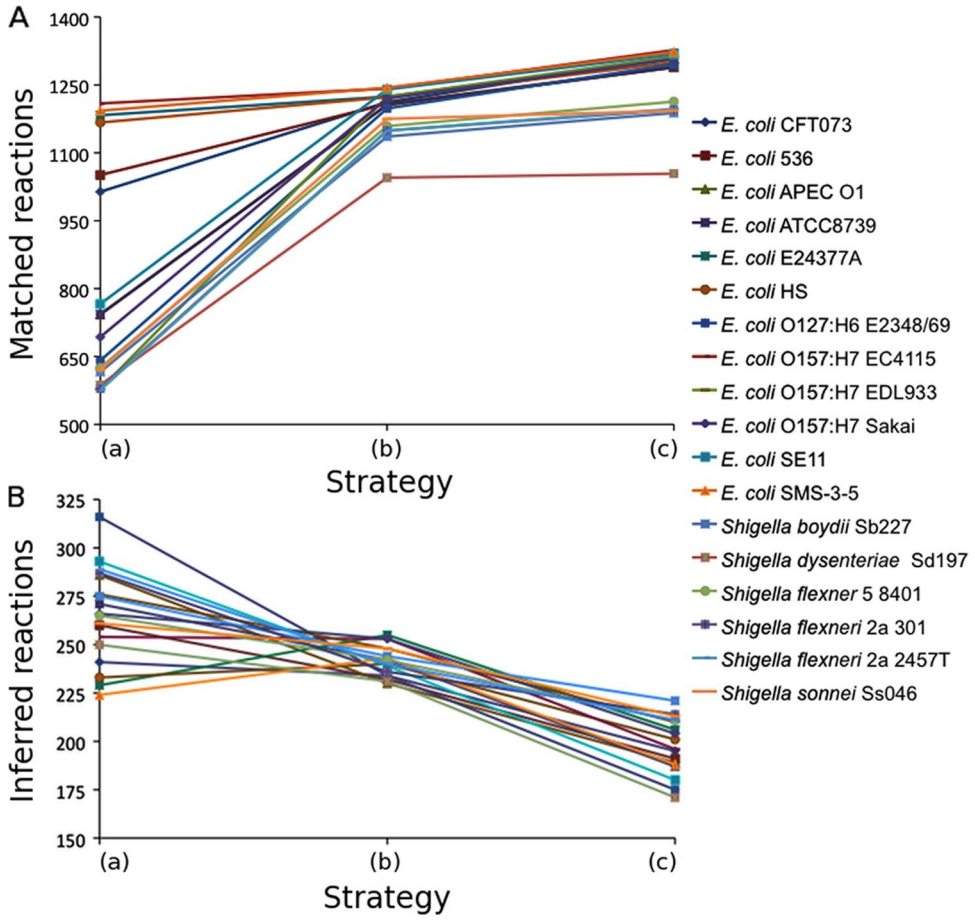


FIG. 2. Number of matched (A) and inferred (B) reactions for each network according to the reconstruction strategy. Strategy a, use of raw genome annotations directly extracted from GenBank database and the default Pathway Tools process; strategy b, use of updated genome annotations from ColiScope and the default Pathway Tools process; strategy c, use of updated genome annotations from ColiScope and the combined EcoCyc/Pathway Tools process.

networks to assess their comprehensiveness, number of inferred reactions to estimate their levels of confidence, and number and completion of metabolic pathways.

Genome annotation quality directly impacted the number of matched reactions (Fig. 2A, strategy b versus strategy a). On average, homogenization of genome annotations increased the number of matched reactions in each strain by an average of 31% and decreased the number of inferred reactions by an average of 10% (Fig. 2B). In the case of *E. coli* O157:H7 EDL933, the number of matched reactions increased more than 2-fold, jumping from 578 to 1,224 reactions. The use of EcoCyc as a pivot (Fig. 2, strategy c versus strategy b) resulted in networks with a small increase in size (2%, on average). The number of matched reactions slightly increased (5%, on average), while the number of inferred reactions considerably decreased (22%, on average). This shows that our process manages to transfer some of the curation performed in EcoCyc to the other reconstructed networks, mainly preventing the inference of wrong reactions.

As regards metabolic pathways, we observed that their total number decreased when improved genome annotations were used (451 versus 386 pathways, on average, for strategies a and b, respectively). This effect is mostly the consequence of the removal of falsely inferred reactions (on the basis of erroneous annotations and erroneous EC number-reaction associations), which triggered the inclusion of wrong pathways. Strategy c, however, slightly increased the number of pathways (2.5% increase with 396 pathways, on average), adding curated path- ways from EcoCyc and also removing some false-positive path- ways. The completion of pathways also improved when we optimized the reconstruction strategy. Starting from 45% of pathways with holes in strategy a, this proportion decreased to 42% in strategy b and reached 34% in strategy c. Furthermore, more than 44% of the pathways with holes in strategy c in- cluded only one hole. The improvement was most noticeable when we employed EcoCyc as a pivot, suggesting again that curation done on a reference metabolic network can be effi- ciently adapted to closely related organisms.

Table 1 shows the main characteristics of the final metabolic networks. On average, they include 1,491 reactions (1,274 matched, 217 inferred), with small variations occurring around that number: 1,300 to 1,564 (1,054 to 1,338 for matched reactions). The reaction count is slightly lower for *Shigella* strains (1,437 total, on average) than for non-*Shigella* strains (1,504 total, on average), a trend that is even stronger when inferred reactions and those associated with pseudogenes are

TABLE 1. Main characteristics of the reconstructed metabolic networks

Strain	Phylogenetic group	Phenotype	No. of genes	No. of reactions				No. of metabolites	No. of pathways
				Total	With gene	With pseudogene	Without gene		
Escherichia coli									
ATCC 8739	A	Commensal	4,411	1,499	1,301	11	187	1,454	347
HS	A	Commensal	4,541	1,510	1,300	9	201	1,443	349
K-12 MG1655	A	Commensal	4,182	1,439	1,269	4	166	1,385	340
K-12 W3110	A	Commensal	4,394	1,461	1,273	7	181	1,425	344
55989	B1	InPEc	4,961	1,473	1,268	6	199	1,440	348
E24377A	B1	InPEc	5,346	1,521	1,308	7	206	1,473	351
IAI1	B1	Commensal	4,412	1,486	1,271	3	212	1,450	351
SE11	B1	Commensal	5,071	1,504	1,318	4	182	1,451	345
536	B2	ExPEc	4,654	1,499	1,290	18	191	1,452	344
APEC O1	B2	ExPEc	4,874	1,482	1,289	4	189	1,392	340
CFT073	B2	ExPEc	5,396	1,532	1,312	25	195	1,456	345
ED1a	B2	Commensal	5,103	1,507	1,292	11	204	1,361	340
LF82	B2	InPEc	4,584	1,483	1,299	4	180	1,378	332
O127:H6 E2348/69	B2	InPEc	4,944	1,485	1,296	14	175	1,423	336
S88	B2	ExPEc	4,848	1,503	1,288	6	209	1,433	343
UTI89	B2	ExPEc	5,305	1,512	1,314	4	194	1,464	346
042	D	InPEc	5,031	1,509	1,311	6	192	1,463	343
UMN026	D	ExPEc	5,046	1,564	1,338	3	223	1,452	352
O157:H7 EC4115	E	InPEc	5,784	1,534	1,327	11	196	1,446	344
O157:H7 EDL933	E	InPEc	5,267	1,531	1,313	8	210	1,445	346
O157:H7 Sakai	E	InPEc	5,431	1,524	1,307	13	204	1,459	344
IAI39	F	ExPEc	4,740	1,531	1,307	10	214	1,484	352
SMS-3-5	F	Commensal	5,128	1,514	1,323	3	188	1,457	347
Shigella									
S. boydii Sb227	S1	Shigellosis	4,717	1,461	1,188	52	221	1,413	332
S. dysenteriae Sd197	SD1	Shigellosis	4,867	1,300	1,054	75	171	1,238	304
S. flexneri 2a 2457T	S3	Shigellosis	4,339	1,475	1,213	51	211	1,425	340
S. flexneri 2a 301	S3	Shigellosis	4,675	1,472	1,195	69	214	1,433	338
S. flexneri 5 8401	S3	Shigellosis	4,393	1,480	1,197	66	211	1,426	337
S. sonnei Ss046	SS	Shigellosis	4,938	1,434	1,193	28	213	1,358	337

removed (1,173 versus 1,301 gene-associated reactions for *Shigella* and non-*Shigella* strains, respectively). *Shigella* strains actually exhibit a significantly higher number of pseudogenes than non-*Shigella* strains, an observation that is consistent with their evolution to become obligate pathogens (44).

We included in the networks enzymatic complexes gener- ated by similarity with strain K-12 MG1655 complexes de- scribed in EcoCyc (see Materials and Methods). Among the 712 homomeric complexes found in EcoCyc, 707 (99%) could be transferred to at least another strain (missing complexes were associated with pseudogenes in EcoCyc) and 458 (65%) were common to all networks. Among the 285 heteromeric complexes from EcoCyc, 278 (97%) were created for at least one strain and 107 (38%) were common to all strains. When *Shigella* strains were removed, the number of common heteromeric complexes reaches 157 (55%). We found in the networks an average of 237 complete heteromeric complexes and an average of 31 heteromeric complexes for which only part of the subunits could be identified. Since we could not automatically identify the reason for the subunit absence (possible reasons include missing gene, annotation error, or another gene with an equivalent product) and since we had evidence for at least a part of the complex, we decided to keep the reactions linked to these incomplete complexes. The names and compositions of all these complexes can be found in Table S2 in the supplemental material.

Using a unified source of genome annotations and a com- mon reconstruction process for all metabolic networks limits the biases originating from the reconstruction process, thus making our networks reliably comparable. In order to focus on the most reliable reactions, we performed our comparative analyses using the set of gene-associated reactions (matched reactions) and discarded reactions associated only with pseu- dogenes or with no gene.

Core and variable parts of metabolism. We separated metabolic reactions into three categories according to their occurrence in strains: panmetabolism, core metabolism, and variable metabolism (see Materials and Methods and Table S1 in the supplemental material). Panmetabolism is the set of all reactions of all strains, i.e., the global metabolic network of *E. coli* species. Core metabolism is the set of reactions common to all strains. Variable metabolism is the difference between pan- and core metabolism, i.e., the set of reactions that are missing from at least one strain.

Panmetabolism included 1,545 reactions. Among them, 885 reactions belonged to core metabolism (57% of the number for panmetabolism) and 660 reactions belonged to variable metabolism (43% of the number for panmetabolism). In each strain, these 885 core reactions represented the major part of the metabolic network (59%, on average), with only 416 reac- tions, on average, belonging to variable metabolism. The oc- currence of variable reactions was not uniformly distributed

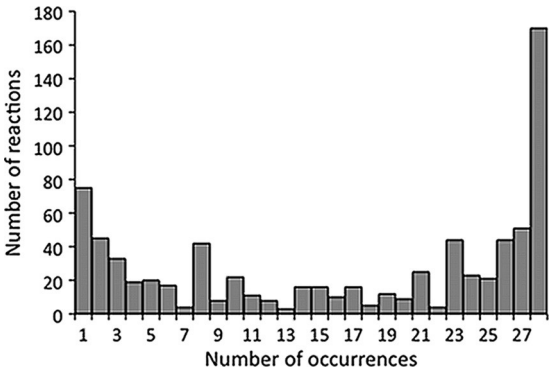


FIG. 3. Distribution of reaction occurrences in strains for reactions not in core metabolism.

among the strains, and its distribution exhibited a U-like shape (Fig. 3): variable reactions tended to be either common to all but a few strains or specific to one or a few strains. Relatively few reactions were shared by medium-size subsets of strains. A peak was yet visible at eight occurrences: these were mainly reactions specific to the eight strains of the B2 group.

When *Shigella* strains were removed, panmetabolism remained nearly identical (1,543 reactions), while core metabolism increased to 1,065 reactions (69% of the number for panmetabolism). This showed that *E. coli* reconstructed networks are well conserved and that *Shigella* has mostly lost reactions since its divergence (22). A set of 180 reactions was therefore absent from *Shigella* core metabolism. It may well include metabolic functions that were no longer required for *Shigella* strains to live in their current habitats (*Shigella* has a parasitic lifestyle) and were thereby lost in these strains. These lost reactions include, for instance, the D-allose degradation pathway (18) and about 10 pathways involved in aromatic compound (e.g., phenylethylamine and phenylacetate) degradation or in amino acid (e.g., histidine) degradation. Lost core reactions were also found among biosynthesis pathways linked to amino acid, nucleotide, and fatty acid anabolism.

Missing reactions from our networks reflected to some extent the auxotrophies found experimentally for *Shigella* strains (1). We observed, for instance, that the nicotinic acid biosynthesis pathway lacks the essential L-aspartate oxidase activity (genes *ndaA* and *ndaB* [42, 43]) in all *Shigella* strains except *Shigella dysenteriae* Sd197, a result that corroborates exactly the auxotrophies for NAD experimentally determined in a previous work (1). Similarly, the absence of homoserine *O*-transsuccinylase (*metA* gene [54]) in *Shigella flexneri* 2a strain 301 may explain the methionine auxotrophy reported for some *S. flexneri* strains in the same work. A few other reported auxotrophies could not, however, be interpreted by simply looking at reaction presence/absence. Turning these metabolic networks into mathematical models of metabolism may help with investigating these cases, as several modeling methods are available to study growth environments in a more systematic manner (13, 16).

The core metabolism/panmetabolism ratio was in sharp contrast to the core metabolism/panmetabolism ratio for the genome (see Materials and Methods for details on core and pangenome computation). For our set of strains, the core genome represented only 13% of the pangenome (1,957 common clusters over 14,986 clusters) (Fig. 4A), a ratio much smaller than that for core metabolism. In addition, an assessment of the variation of the sizes of panmetabolism and pangenome as a function of the number of strains (Fig. 4B) showed that the size of panmetabolism approached a plateau at 29 strains, whereas the pangenome size was still steadily increasing. These results suggest that diversity is more limited within *E. coli* metabolic networks than it is within all gene functions. Two main interpretations can be hypothesized from this observa- tion. First, this estimation of metabolic diversity is limited to the set of reactions already known. Consequently, panmetabo- lism may lack many unknown reactions, especially those spe- cific to poorly studied organisms. In contrast, the pangenome is more confidently estimated since most genes, even those whose functions remain unknown, are detected on genomes. Because of this limitation, adding new strains to the study would not significantly expand panmetabolism if the strain-specific reactions are unknown, which is often the case for

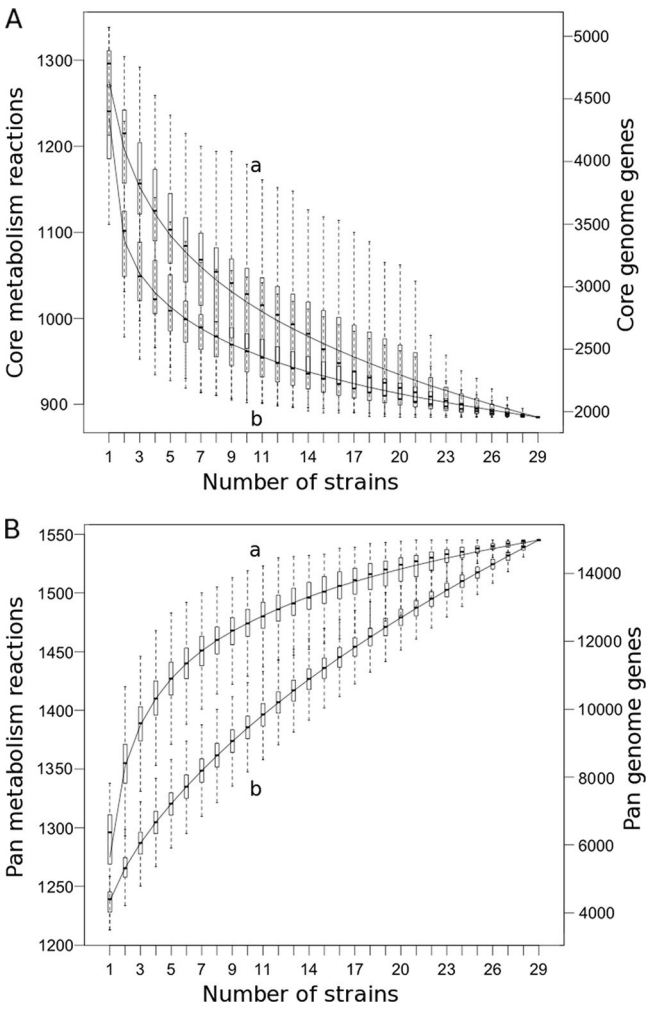


FIG. 4. Evolution of *E. coli* core metabolism (A, curve a), core genome (A, curve b), panmetabolism (B, curve a), and pangenome (B, curve b), according to the number of included strains. Boxes delimit the first and third quartiles of 10,000 different input orders of meta- bolic networks and 5,000 different input orders of genomes.

TABLE 2. Distribution of reactions of core, variable, and panmetabolism across metabolic processes, as defined in BioCyc databases

No. of metabolic occurrences [a]	Core	Variable	Pan
Biosynthesis	508	236	744
Degradation	200	224	424
Detoxification	9	5	14
Energy metabolism	68	29	97
Transport pathways	2	2	4
Other	262	231	493
Total	885	660	1,545

[a] Some reactions occur in distinct metabolic processes; therefore, the sum of occurrences is higher than the total number of reactions.

newly sequenced organisms. This observation has actually motivated several initiatives which focus on the search for novel enzymatic activities rather than on the mere sequencing of additional genomes (4, 5). Second, genes coding for enzymatic functions may vary less than those coding other functions. Diversity in metabolism could be traced back to a relatively small number of distinct enzymes; genomic diversity may involve nonenzymatic processes such as regulation, which contributes to another level of metabolic diversity via the control of metabolism (33).

We next examined in more detail how core and variable reactions were distributed among metabolic categories (Table 2). Interestingly, the proportion of core reactions was significantly higher in biosynthetic processes (68%) than in other metabolic categories (the Fisher exact test, $P < 10^{-15}$). This contrasts with degradation processes, which contain a significantly lower proportion of core reactions (29%) than other metabolic categories (the Fisher exact test, $P < 10^{-15}$). Biosynthesis reactions actually constitute the majority of reactions from core metabolism (57%, 508 reactions). This result can be interpreted by the fact that, when environments are changing, metabolic functions closely related to metabolites from the environment (e.g., degradation pathways) are more likely to vary than biosynthetic reactions, which usually use ubiquitous basic metabolites as precursors. A similar effect has been observed in a previous study among the functions of horizontally transferred genes (i.e., variable genes), which were found to be involved more often in transport and peripheral degradation pathways than in central biosynthetic processes (39).

Reactions involved in sucrose degradation are a good illustration of variable metabolism. The ability to use sucrose as a sole carbon source is a highly variable phenotype in enterobacteria. Among commensal strains, *E. coli* K-12 MG1655, K-12 W3110, HS, ATCC 8739, and SMS-3-5 cannot utilize sucrose, whereas the IAI1 and SE11 strains can. This phenotype is also highly variable for *E. coli* pathogenic strains. Chromosomal genes associated with sucrose degradation are organized in a cluster of two operons coding for a non-phosphotransferase system permease (*cscB* gene) and a fructokinase (*cscK* gene) in the first operon and a sucrose hydrolase (*cscA* gene) in the second operon, with both being controlled by an adjacent repressor (*cscR* gene) (24). This cluster is integrated next to a tRNA-Arg gene, and the codon adaptation index (CAI) of the cluster genes is among the lowest of all *E. coli* genes (among the 8% of genes with the lowest CAI), suggesting acquisition of the *csc* genes by horizontal gene transfer.

Structure of *E. coli* metabolic diversity. To study how metabolic diversity is distributed within the *E. coli* species, we analyzed the metabolic distances, defined by the number of distinct reactions between two strains (see Materials and Methods), between strains. We first grouped strains according to metabolic distance and obtained the tree shown in Fig. 5A. Overall, strain groups matched phylogenetic groups relatively well. Group B2, D, E, and F strains clearly clustered according to their groups. The F group is a new group composed of strains previously included in the D one (25), a fact that was visible from the genomic point of view (Fig. 5B) but also from the metabolic one (Fig. 5A). Strains from the A and B1 groups are, however, mixed together. Group A and B1 strains are actually phylogenetically close (Fig. 5B), and the evolutionary distance between them may be too small to imply a significant difference in their metabolic networks.

All *Shigella* strains were markedly more distant from the other strains (Fig. 5A). *Shigella* strains have evolved from multiple distinct phylogenetic groups (15, 44), and this effect is still visible from the strain phylogenetic tree, since they are spread among *E. coli* groups (Fig. 5B). However, the high metabolic distances that separate them from other strains have blurred this signal, suggesting that evolution of their metabolism has been rapid.

To further study the link between metabolism and genetic diversity, we directly compared metabolic and genetic dis- tances for all pairs of strains (Fig. 6; see Materials and Meth- ods). A Mantel test performed on this pair of distances showed that they are significantly correlated (P ::s 0.01), yet they have a relatively large dispersion due to *Shigella* (linear regression, $r^2 = 0.15$). When the focus is on non-*Shigella* strains, linear regression between the two distances significantly improved (linear regression, $r^2 = 0.54$), showing that metabolic distance increases with genetic distance. Strains of the same phyloge- netic groups (blue symbols in Fig. 6) were separated by sets of 50 to 150 reactions, and this number did not vary with genetic distance. Metabolic distances between non-*Shigella* strains from distinct phylogenetic groups were slightly higher but still in the range 100 to 250 reactions. Here again, group A and B1 strains behaved as if they formed a single phylogenetic group, and their genetic and metabolic distances were compa- rable to intragroup distances: sets of 75 to 125 reactions (set of leftmost black symbols in Fig. 6).

As observed above, for similar genetic distances, *Shigella* metabolic networks were markedly more distant from other networks than were non-*Shigella* metabolic networks. Furthermore, metabolic distances between *Shigella* strains were comparable to metabolic distances between *Shigella* and non-*Shi- gella* strains, while the distance between *Shigella* strains from the same phylogenetic group (i.e., those of the S3 *Shigella* group) was equal to the intragroup *E. coli* metabolic distance. This suggests that their metabolic networks have quickly evolved by genetic drift (11) and that most metabolic differences were not common to all *Shigella* strains. Among the 176 pseudoreac- tions (linked only to pseudogenes) found in at least one *Shi- gella* strain, none were pseudoreactions in all 6 *Shigella* strains and 92 were pseudoreactions in only one *Shigella* strain. Nev-

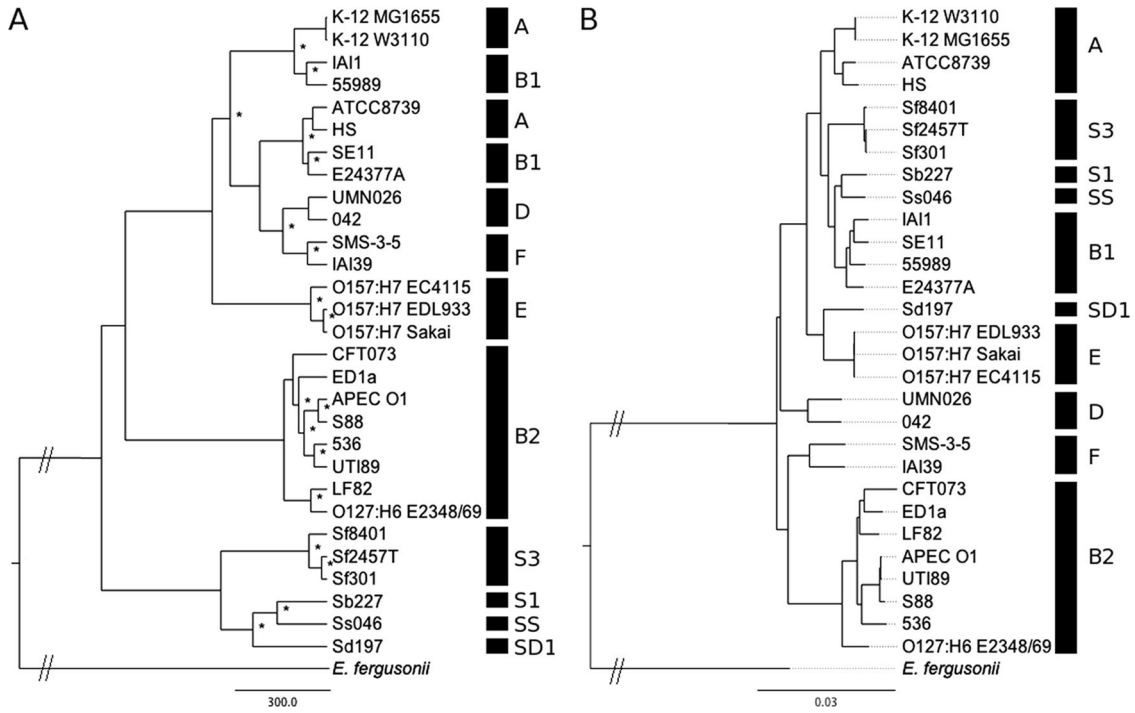


FIG. 5. Evolution tree of *E. coli* according to metabolic (A) and genetic (B) distances. *, nodes with a bootstrap value greater than 70% for the metabolic tree. All nodes of the genetic tree have bootstrap values greater than 70%. Phylogenetic groups are defined according to references 44 and 11 for *E. coli* and *Shigella* strains, respectively.

ertheless, convergent inactivation of a few metabolic charac- ters has been reported, indicating adaptive evolution (1, 11, 34, 42). This could be a consequence of its parasitic lifestyle, which removes requirements for some degradation/biosynthesis pathways, as mentioned above.

In order to examine in more detail metabolic diversity within non-*Shigella* strains, we performed an MCA (see Materials and Methods) on reaction occurrences (Fig. 7). The first two factorial axes accounted for 34% of all variability. There were more than a hundred reactions with a significant contribution (see Materials and Methods) to the first axis. Half of the reactions with a high contribution were involved in biosynthetic processes, especially lipid biosynthesis (71% of them). Another 23% of high-contribution reactions were associated with degradation, in particular, aromatic compound degradation (37% of them). Most of the remaining reactions were not part of any pathway. Similarly, we observed on the second axis that 57% of high-contribution reactions were linked to biosynthetic processes (with 82% of them being lipid biosynthesis), and 25% were associated with degradation (with 42% of them being aromatic compound degradation).

The large number of reactions with high contributions on each of these axes made our MCA robust to addition or removal of reactions. Moreover, when the MCA was computed while discarding dozens of reactions with the best contributions, only minor changes to the distribution of strains were observed (data not shown).

In agreement with observations on metabolic distances, Fig. 7A shows that phylogenetic groups were relatively well separated by the first two axes of the MCA for all except strains of groups A and B1, which are mixed. Group F strains were separated from group D strains on both axes, confirming the existence of metabolic differences between them. Such a clear separation supports the separation of group F strains from group D strains (25).

When strains were grouped according to their phenotypes (commensal, ExPEc, or InPEc; Table 1 and Fig. 7B), no clear separation could be seen from the MCA. Indeed, reaction occurrence in strains seemed to be poorly correlated with strain phenotypes. In order to compare more robustly pheno- typic and phylogenetic groups with metabolic distances, we computed compactness (mean distance between group centers and group members) and separation (distance between two group centers) measures for all groups and all pairs of groups using the metabolic distances (Table 3) (see Materials and Methods). These two measures globally evaluate the closeness of strains within a group and their separation between two groups, according to the chosen distance (21). Compactness measures confirmed that strains grouped by phylogeny were markedly closer to each other than strains grouped by pheno- type (26 to 68 for phylogenetic groups versus 92 to 138 for phenotypic groups). Furthermore, when compactness mea- sures are compared with separation measures, phylogenetic groups appeared to be globally distinct, except for the A and B1 groups, which here again showed overlap. Metabolic sepa- ration between phenotypic groups was, in contrast, not signif- icantly higher than within-group distances. Strains from phe- notypic groups were nearly as distant from each other than from strains of other phenotypic groups. Therefore, pathogenicity phenotypes did not appear to drive large changes in reaction occurrence in these strains.

As the presence of small sets of specific reactions can, how-

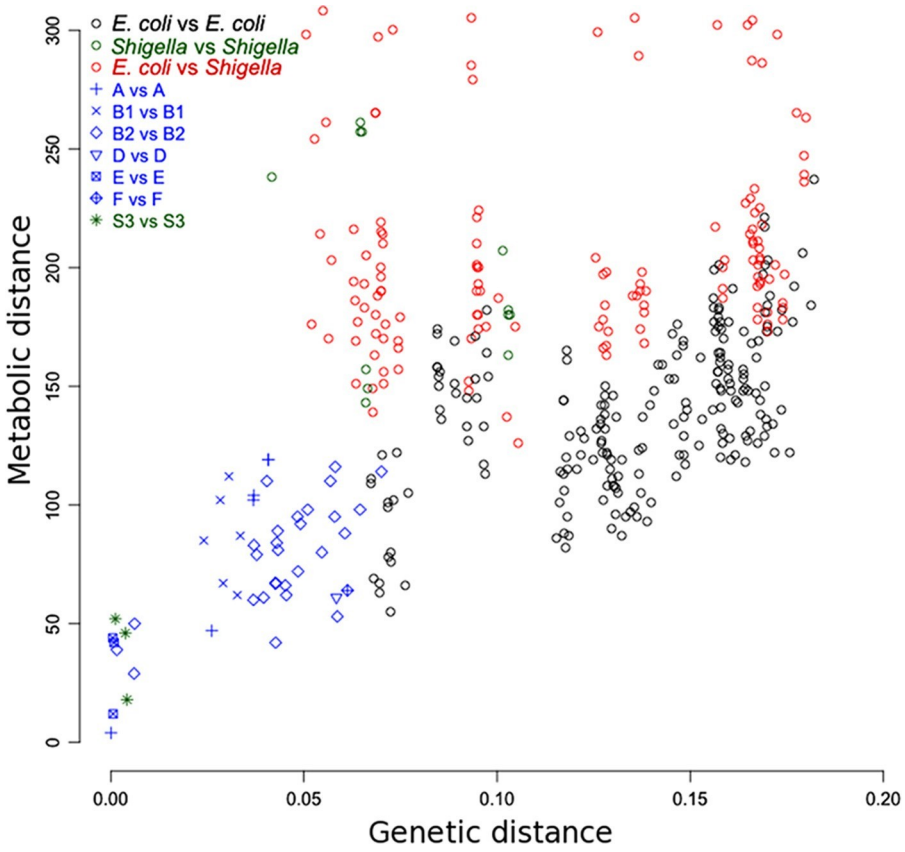


FIG. 6. Plot of genetic distances (*x* axis) versus metabolic distances (*y* axis) for all pairs of strains, colored according to strain phylogenetic groups. Blue, both strains in each pair are non-*Shigella* strains from the same phylogenetic group; black, strains are from distinct groups but both are non-*Shigella*; green, both are *Shigella* strains; red, strains are from distinct groups, with one being *Shigella* and the other being non-*Shigella*.

ever, induce notable changes in phenotypes, we looked in more detail for specific differences between networks grouped by pathogenicity. As no reaction was found to be completely specific to any pathogenicity phenotype, we used a supervised method able to slightly relax the specificity constraint and find such characteristic sets of reactions (CART; see Materials and Methods). We applied this method to each pathogenicity phenotype (commensal, InPEc, and ExPEc).

We observed that most commensal strains (except ED1a and SMS-3-5) possess reactions able to degrade phenylacetate and phenylethylamine (12) (*paa* transcription unit), which are absent from InPEc and ExPEc strains (except E24377A and 55989). E24377A and SMS-3-5 were further separated from the commensal strains by the presence of a plasmid-encoded toxin (PET) serine precursor (gene *sat*), which is known to be an important virulence factor (47) associated with both intes- tinal and extraintestinal infections.

ExPEc strains were mainly characterized by the absence of psicose and psicoselysine degradation pathways (*frl* transcription unit). They also specifically possess a putative transporter of capsular polysaccharide (gene *kpsT*), a virulent element used by the virulent strain *E. coli* K1 during neonatal septice- mia and meningitis (41, 53).

Reactions characteristic of InPEc strains could be less clearly identified. Most of them are putative reactions, like a maleylacetoacetate isomerase (a locus similar to *maiA* in *Salmonella*), and another one has a high similarity with glutathione *S*-transferase and a cobalamine adenosyltransferase (gene *glmL*).

Results from this analysis can be found in Table S3 in the supplemental material.

DISCUSSION

Establishing a link between genomes and phenotypes is difficult because several layers of biological processes intervene between genes and their phenotypic effects. Metabolism is one of these layers, and thanks to automated metabolic reconstruc- tion tools, it can be studied at the genome scale for sequenced organisms. However, identifying sound metabolic differences between distinct organisms and assessing diversity within a set of metabolic networks, as was done in this work, require suf- ficiently detailed metabolic networks that standard automated methods usually do not produce without curation (13). Here, we were able to improve an automated reconstruction strategy by leveraging the proximity of all strains with *E. coli* K-12 MG1655, whose genome and metabolism are incomparably well-known. As a result, we provide high-quality metabolic net- works for 29 *E. coli* strains, including 6 *Shigella* strains, all of which are suitable for comparative analyses (available at http://www.genoscope.cns.fr/agc/metacoli/). Most noteworthy, a large improvement in network completeness was achieved by updat-

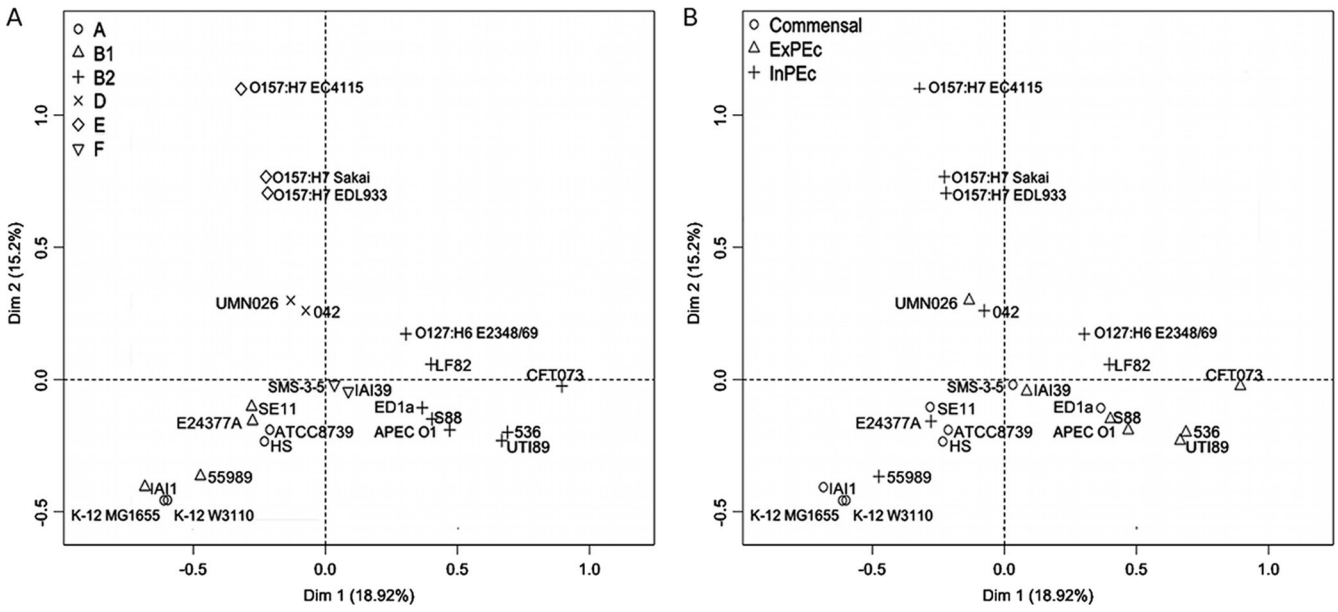


FIG. 7. Plot of the first two axes of MCA of reaction occurrences in *E. coli* non-*Shigella* strains, labeled according to phylogenetic groups (A) and phenotypes (B). MCA was performed on reactions associated with genes. The distance between strains can be interpreted as the most significant dissimilarities between their reaction absence/presence profiles.

ing and homogenizing genome annotations for all *E. coli* strains, while using EcoCyc as a primary reconstruction pivot allowed the transfer of some of the manual curation done on K-12 MG1655 metabolism and thereby limit the proportion of falsely inferred reactions. We were also able to infer enzyme complexes similar to those known in K-12 MG1655.

The reconstructed networks were composed of a majority of *E. coli* core reactions and relatively few variable reactions. Moreover, examining the evolution of the size of panmetabolism as a function of the number of networks indicates that reconstructing the metabolism of new strains will add only little diversity to the current panmetabolism. The size of panmetabolism is yet likely to be underestimated, as many reactions remain unknown. Characterizing missing enzyme activities in the current strains will most probably contribute to expanding the knowledge of panmetabolism at least as much as sequencing and annotating new strains do.

We observed that biosynthetic reactions were mostly part of core metabolism and that degradation processes, on the other hand, were mainly found in variable metabolism. This can be interpreted by the fact that the selection pressure acting on biosynthetic processes is likely to be similar for all strains, as these processes, which take as inputs common central meta- bolic precursors, are only weakly influenced by the environ- ment. Conversely, degradation processes are directly linked to compounds from the environment, and their selection there- fore depends on the environment and strain lifestyles (39).

This evolutionary interpretation is supported by the large metabolic differences separating the six *Shigella* strains from the others. These strains, whose parasitic lifestyles make large

TABLE 3. Compactness and separation measures for phylogenetic and phenotypic groups, according to the metabolic distance

Phylogenetic group or phenotype	Compactness	Separation											
		A	B1	B2	D	E	F	S1	S3	SD1	Commensal	ExPEc	InPEc
A	62												
B1	64	50											
B2	68	149	148										
D	30	128	118	134									
E	22	153	142	162	93								
F	32	117	110	103	97	131							
S1	NA[a]	188	191	196	197	205	187						
S3	26	181	169	201	169	193	182	150					
SD1	NA	282	280	303	308	290	297	238	258				
SS	NA	161	156	221	189	199	188	163	181	207			
Commensal	92												
ExPEc	88										119		
InPEc	102										92	114	
Shigellosis	138										159	199	177

[a] NA, not applicable; the group has only one member.

parts of *E. coli* panmetabolism dispensable, have actually lost many reactions still present in all non-*Shigella* strains. These differences make their metabolic networks sufficiently distinct from the other *E. coli* networks to blur their phylogenetic origin (see metabolic tree in Fig. 5A).

When the *Shigella* strains were removed from the study, we observed that differences between metabolic networks were significantly correlated with the strains' phylogenies but not with their commensal/pathogenic phenotypes. This suggests that changes in metabolic networks occurred with strain diver- gence and were mostly not driven by strain phenotypes, as was yet the case for the *Shigella* phenotype.

The fact that *E. coli* commensal/pathogenic phenotypes do not globally influence their metabolic networks does not mean that no metabolic characteristic can be associated with them. First, the presence or absence of only a few enzymes may be related to these phenotypes. Using a supervised classification method, we were able to identify such cases, with some having already been described in literature. Second, diversity in metabolic behaviors does not originate from enzyme diversity only. Diversity in enzyme regulation and activity also influences metabolism and cannot be assessed by solely studying recon- structed metabolic networks. It involves, for instance, studying regulatory networks or experimentally measuring how metab- olism actually operates in each strain. Our reconstructed net- works represent a first step toward such investigations, as they form a solid basis on which to build the metabolic models needed to integrate and interpret such experimental data.

ACKNOWLEDGMENTS

This work is supported by a grant from the French National Research Agency (ANR) to the Metacoli project (contract number ANR-08-SYSC-011) and by MICROME, a collaborative project funded by the European Commission within its FP7 Program, contract number 222886-2. E.D. is partly supported by the Fondation pour le Recherche M´edicale. V. Sabarly is partly supported by D´el´egation G´eminale pour l'Armement.

REFERENCES

1. **Ahmed, Z. U., M. R. Sarker, and D. A. Sack.** 1988. Nutritional requirements of shigellae for growth in a minimal medium. Infect. Immun. **56:**1007–1009.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
3. **Atkinson, E. J., and T. M. Therneau.** 2000. An introduction to recursive partitioning. Technical report. Mayo Foundation, Rochester, MN.
4. **Baran, R., W. Reindl, and T. R. Northen.** 2009. Mass spectrometry based metabolomics and enzymatic assays for functional genomics. Curr. Opin. Microbiol. **12:**547–552.
5. **Beloqui, A., et al.** 2009. Reactome array: forging a link between metabolome and genome. Science **326:**252–257.
6. **Breiman, L., J. Friedman, R. Olshen, and C. Stone.** 1984. Classification and regression trees, new edition. Chapman & Hall/CRC, New York, NY.
7. **Caspi, R., et al.** 2010. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. **38:**D473–D479.
8. **Castresana, J.** 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17:**540–552.
9. **Chaudhuri, R. R., et al.** 2010. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. PLoS One **5:**e8801.
10. **Cruveiller, S., et al.** 2005. MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. Nucleic Acids Res. **33:**W471–W479.
11. **Denamur, E., B. Picard, and O. Tenaillon.** 2010. Population genetics of pathogenic *Escherichia coli*, p. 269–286. *In* D. A. Robinson, D. Falush, and E. J. Feil (ed.), Bacterial population genetics in infectious disease. Wiley-Blackwell, West Sussex, United Kingdom.
12. **Diaz, E., A. Ferrandez, M. A. Prieto, and J. L. Garcia.** 2001. Biodegradation of aromatic compounds by *Escherichia coli*. Microbiol. Mol. Biol. Rev. **65:** 523–569.
13. **Durot, M., P. Bourguignon, and V. Schachter.** 2009. Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol. Rev. **33:**164–190.
14. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accu- racy and high throughput. Nucleic Acids Res. **32:**1792–1797.
15. **Escobar-Pa´ramo, P., C. Giudicelli, C. Parsot, and E. Denamur.** 2003. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. J. Mol. Evol. **57:**140–148.
16. **Feist, A. M., et al.** 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. **3:**121.
17. **Fricke, W. F., et al.** 2008. Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. J. Bacteriol. **190:**6779–6794.
18. **Fukiya, S., H. Mizoguchi, T. Tobe, and H. Mori.** 2004. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. J. Bacteriol. **186:**3911–3921.
19. **Green, M. L., and P. D. Karp.** 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. BMC Bioinform. **5:**76.
20. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:**696–704.
21. **Handl, J., J. Knowles, and D. B. Kell.** 2005. Computational cluster validation in post-genomic data analysis. Bioinformatics **21:**3201–3212.
22. **Hershberg, R., H. Tang, and D. A. Petrov.** 2007. Reduced selection leads to accelerated gene loss in *Shigella*. Genome Biol. **8:**R164.
23. **Iguchi, A., et al.** 2009. Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. J. Bacteriol. **191:**347–354.
24. **Jahreis, K., et al.** 2002. Adaptation of sucrose metabolism in the *Escherichia coli* wild-type strain EC3132. J. Bacteriol. **184:**5307–5316.
25. **Jaureguy, F., et al.** 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics **9:**560.
26. **Kanehisa, M., et al.** 2007. KEGG for linking genomes to life and the environment. Nucleic Acids Res. **36:**D480–D484.
27. **Kaper, J. B., J. P. Nataro, and H. L. T. Mobley.** 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. **2:**123–140.
28. **Karp, P. D., et al.** 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief. Bioinform. **11:**40–79.

28a. **Keseler, I. M., C. Bonavides-Mart´ınez, J. Collado-Vides, S. Gama-Castro, R. P. Gunsalus, D. A. Johnson, M. Krummenacker, L. M. Nolan, S. Paley, I. T. Paulsen, M. Peralta-Gil, A. Santos-Zavaleta, A. G. Shearer, and P. D. Karp.** 2009. EcoCyc: a comprehensive view of Escherichia coli biology. Nucleic Acids Res. **37:**D464–D470.

29. **Lawrence, J. G., H. Ochman, and D. L. Hartl.** 1991. Molecular and evolutionary relationships among enteric bacteria. J. Gen. Microbiol. **137:**1911–1921.
30. **L^e, S., J. Josse, and F. Husson.** 2008. FactoMineR: an R package for multivariate analysis. J. Stat. Softw. **25:**1–18.
31. **Le F`evre, F., S. Smidtas, and V. Scha¨chter.** 2007. Cyclone: Java-based querying and computing with Pathway/Genome databases. Bioinformatics **23:** 1299–1300.
32. **Li, L., C. J. Stoeckert, and D. S. Roos.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13:**2178–2189.
33. **Maslov, S., S. Krishna, T. Y. Pang, and K. Sneppen.** 2009. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proc. Natl. Acad. Sci. U. S. A. **106:**9743–9748.
34. **Maurelli, A. T., R. E. Ferna´ndez, C. A. Bloch, C. K. Rode, and A. Fasano.** 1998. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **95:**3943–3948.
35. **Miquel, S., et al.** 2010. Complete genome sequence of Crohn's disease- associated adherent-invasive *E. coli* strain LF82. PLoS One **5:**e12714.
36. Reference deleted.
37. **Notebaart, R. A., F. H. J. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink.** 2006. Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinform. **7:**296.
38. **Oshima, K., et al.** 2008. Complete genome sequence and comparative anal- ysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. DNA Res. **15:**375–386.
39. **Pa´l, C., B. Papp, and M. J. Lercher.** 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat. Genet. **37:**1372–1375.
40. **Paradis, E., J. Claude, and K. Strimmer.** 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20:**289–290.
41. **Pavelka, M. S., L. F. Wright, and R. P. Silver.** 1991. Identification of two genes, kpsM and kpsT, in region 3 of the polysialic acid gene cluster of *Escherichia coli* K1. J. Bacteriol. **173:**4603–4610.
42. **Prunier, A., et al.** 2007. nadA and nadB of *Shigella flexneri* 5a are antiviru- lence loci responsible for the synthesis of quinolinate, a small molecule inhibitor of *Shigella* pathogenicity. Microbiology **153:**2363–2372.
43. **Prunier, A., R. Schuch, R. E. Ferna´ndez, and A. T. Maurelli.** 2007. Genetic structure of the nadA and nadB antivirulence loci in *Shigella* spp. J. Bacte- riol. **189:**6482–6486.
44. **Pupo, G. M., R. Lan, and P. R. Reeves.** 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc. Natl. Acad. Sci. U. S. A. **97:**10567–10572.
45. **Rasko, D. A., et al.** 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. **190:**6881–6893.
46. **R Development Core Team.** 2009. R: a language and environment for statistical computing. R Development Core Team, Vienna, Austria.

47. **Restieri, C., G. Garriss, M. Locas, and C. M. Dozois.** 2007. Autotransporter-encoding sequences are phylogenetically distributed among *Escherichia coli* clinical isolates and reference strains. Appl. Environ. Microbiol. **73:**1553–1562.

48. Reference deleted.

49. **Tenaillon, O., D. Skurnik, B. Picard, and E. Denamur.** 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. **8:**207–217.

50. **Teusink, B., et al.** 2005. In silico reconstruction of the metabolic pathways of Lactobacillus plantarum: comparing predictions of nutrient requirements with those from growth experiments. Appl. Environ. Microbiol. **71:**7253–7262.

51. **Touchon, M., et al.** 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. **5:**e1000344.

52. **Vallenet, D., et al.** 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. Database **2009:**bap021.

53. **Whitfield, C.** 2006. Biosynthesis and assembly of capsular polysaccharides in Escherichia coli. Annu. Rev. Biochem. **75:**39–68.

54. **Zagaglia, C., et al.** 1991. Virulence plasmids of enteroinvasive *Escherichia coli* and *Shigella flexneri* integrate into a specific site on the host chromo- some: integration greatly reduces expression of plasmid-carried virulence genes. Infect. Immun. **59:**792–799.

2 Approfondissements

Dans cette partie, nous allons approfondir certains points méthodologiques et certains résultats. Nous reviendrons plus en détail sur la mise en place de la méthodologie, des difficultés recontrées et les solutions apportées.

.2.1 Homogénéité des données

L'identification précise d'un gène, d'une protéine ou d'une réaction est essentielle durant le processus de reconstruction. S'il existe un nom usuel pour chaque entité biologique connue, l'association par ce nom n'est pas une méthode totalement fiable. En effet, la plupart des gènes, des protéines et des réactions possèdent des noms synonymes qui sont parfois utilisés à la place du nom usuel. *E. coli* possède près de 2500 gènes avec des synonymes, et par exemple, le gène nommé *rmlB* dans MicroCyc est désigné et référencé par un de ses synonymes dans EcoCyc *rfbB*, et il possède un autre synonyme *som*. La situation s'aggrave pour les éléments sans nom usuel, dans ce cas ils seront identifiés par des labels ou des identifiants : le gène avec le label ECK0295 dans MicroCyc ne possède pas de nom associé. Malheureusement, il n'existe pas de nomemclature sur les noms ou de consensus sur les identifiants, et chacune des bases de données ou des ressources utilisées est libre de créer et d'utiliser sa propre nomenclature. Ainsi un gène sans nom sera identifié par un identifiant du genre G---- chez EcoCyc (*gnumber*), b---- chez BiGG (*bnumber*), un identifiant du genre ECK---- (*Ecknumber*) ou encore par l'identifiant RefSeq de la séquence et la position du gène sur cette séquence.
Le gène qui code pour une thréonine déshydratase est associé à différents labels suivant la base de données : EG10990, b3117 ou ECK3106. Ceux-ci sont associés au même nom usuel, *tdcB* : nous pouvons automatiquement établir un lien entre les différentes ressources. Les deux exemples présentés précédemment demandent une expertise puisque dans le meilleur des cas, il faut vérifier les synonymes, et dans le pire effectuer une recherche de séquence homologue.
Le cas des enzymes est encore plus complexe, les noms de celles-ci sont souvent longs puisqu'ils désignent leur fonction biochimique et ils comprennent de nombreux synonymes. L'enzyme dont le numéro EC est 2.6.1.2 (1) se nomme usuellement alanine transaminase et possède 17 synonymes qui, suivant les bases de données, peuvent se trouver à la place du nom usuel (*glutamic-pyruvic, transaminase, glutamic-alanine transaminase, GPT, beta-alanine aminotransferase, alanine aminotransferase*, etc.). Cependant une recherche de *beta-alanine aminotransferase* dans KEGG renvoie également à l'enzyme avec le numéro EC 2.6.1.19 (2). La réaction catalysée par l'enzyme (1) possède comme identifiant dans KEGG R00258, dans BiGG ALATA_L et dans EcoCyc ALANINE-AMINOTRANSFERASE-RXN. Alors que celle catalysée par l'enzyme (2), a des identifiants différents (respectivement R01648, ABTA et GABATRANSAM-RXN). Il paraît évident que l'utilisation des noms et synonymes, au sein d'un processus de réconciliation automatique des bases de données est impossible.
Notre stratégie de reconstruction des réseaux métaboliques utilise comme base de données principale MetaCyc, nous avons eu à unifier les éléments suivants avec les autres ressources: les noms des gènes, les numéros EC, et le nom des enzymes. Néanmoins nous serons amenés à revenir sur le problème d'association et d'unification entre les différentes ressources dans la partie de reconstruction de modèles (Chapitre II partie 2.4).

.2.2 MicroScope

La plate-forme MicroScope et son interface web MaGe[1] (Magnifying Genomes) (Vallenet et al. 2009; Vallenet et al. 2006) sont des outils collaboratifs d'annotation et de comparaisons de génomes bactériens. Elle repose sur une base de données (PkGDB pour Prokaryotic Genome DataBase), où sont sauvegardés de nombreux résultats pré-calculés. L'un des points forts de cet outil est l'unification des données, qui associe à chaque élément un identifiant unique, ceci garantit que parmi les noms et synonymes on choisira toujours le même nom usuel. Les données qui alimentent la base servent à l'annotation automatique et à l'unification ; elles servent aussi à l'annotation manuelle experte. Cette dualité a été utilisée par le consortium ColiScope lors du séquençage et l'annotation de nouvelles souches d'*E. coli*. Le processus d'annotation automatique sur les 5 souches (S88, UMN026, IAI1, ED1a, 55989 et IAI39), a permis d'annoter près de 22 000 gènes. A ceux-ci viennent s'ajouter 7 864 gènes annotés manuellement. Le résultat du long et fastidieux travail d'annotation manuelle est ensuite propagé aux 15 souches d'*E. coli* déjà disponibles, puis aux autres souches séquencées et rendues publiques depuis.
Depuis sa création MicroScope s'est enrichi de nouvelles fonctionnalités et notamment un ensemble de modules consacrés au métabolisme. Le module MicroCyc est au métabolisme ce que MicroScope est au génome, un ensemble d'outils et de résultats pré-calculés qui repose sur une base de données (PGDB pour Pathway Genome DataBase). MicroCyc s'inscrit dans la continuité de MicroScope et leurs deux bases de données sous-jacentes sont reliées grâce à des identifiants communs notamment ceux des gènes. MicroCyc repose sur l'outil Pathway-Tools pour générer les PGDB à partir des génomes. C'est en complément de ce processus que se place notre nouvelle stratégie.

.2.3 Reconstruction

Parmi les différentes bases de données et outils de reconstructions, Pathway-Tools et les PGDB au format cyc nous ont semblé les plus adaptés à notre situation; nous les avons préférés à leur principal concurrent KEGG et cela pour plusieurs raisons. Bien que KEGG possède une base de données métabolique généraliste comme BioCyc, ses méthodes d'association des réactions aux gènes par l'annotation sont moins développées et plus limitées. Cette association repose en effet uniquement sur le numéro EC complet contenu dans l'annotation. Malheureusement toutes les enzymes n'ont pas de numéro EC, ou certains numéros sont incomplets. La base MetaCyc recense 114 numéros EC partiels qui sont associés à 1583 réactions différentes, ce sont autant de réactions qui échappent au processus de reconstruction de KEGG. A cela s'ajoute les actualisations et la non-exhaustivité des numéros EC qui sont autant de sources d'erreurs dans les réseaux métaboliques reconstruits. Le processus de reconstruction BioCyc est plus souple et prend en compte les différentes informations de l'annotation fonctionnelle. Ainsi en complément du numéro EC, il est possible d'utiliser le nom de l'enzyme comme critère d'association entre le gène, l'enzyme et la réaction. L'atout majeur de Pathway-Tools est l'algorithme Pathologic (Green & P. Karp 2004) et la reconstruction de voies métaboliques. Il permet d'inférer des réactions sans gène associé et il structure les données afin d'avoir une vue fonctionnelle du réseau métabolique et une hiérarchisation des fonctions métaboliques. La contrepartie est cependant assez lourde puisque tel quel le processus a tendance à sur-prédire les voies métaboliques et par transitivité les réactions sans

[1] http://www.genoscope.cns.fr/agc/microscope/home/index.php

gène associé. L'inférence de réactions à partir de l'annotation textuelle connaît des limites et entraîne des erreurs dans le cas où le nom de l'enzyme est générique (c'est à dire lorsque le substrat n'est pas précisé). Par exemple, l'enzyme *alcool déshydrogénase* peut correspondre à une multitude de réactions puisqu'on dénombre plus de 200 métabolites de type alcools différents dans la base de données MetaCyc. A cela vient s'ajouter les problèmes de synonymes et de mauvaises annotations qui vont aboutir elles aussi à des réactions incorrectes. En parallèle, le processus d'inférence peut prédire des voies métaboliques qui ne sont pas présentes dans l'organisme. Un trop faible nombre de réactions et un manque de connaissance peuvent aboutir à la prédiction de faux positifs (réactions prédites dans le réseau métabolique alors que dans la réalité celles-ci n'en font pas partie). Il est extrêmement délicat de discerner un faux positif, d'une véritable réaction, sans données expérimentales supplémentaires ; c'est pourquoi il est important d'utiliser le maximum de connaissances disponibles lors de la reconstruction. Les outils de KEGG et BioCyc sont tous limités aux cas des enzymes et réactions référencées ; malheureusement un grand nombre d'activités enzymatiques ne sont pas référencées et restent éparpillées dans la littérature. De même, les activités enzymatiques nouvellement découvertes mettent également un certain temps avant d'obtenir un numéro EC et d'être référencées dans les différentes bases de données.

2.3.1 Les Cycs et MicroCyc

La base de données de BioCyc repose sur une ontologie développée autour des fonctions biologiques (Peter D. Karp 2000). Une ontologie est définie comme une spécification et une conceptualisation conçues dans le but d'être réutilisables dans de nombreuses applications et implémentations (Guarino 1998). Elle décrit formellement des concepts et explicite les relations existantes entre eux. Appliquée au métabolisme, l'ontologie va permettre de hiérarchiser les différents acteurs (réactions, enzymes, métabolites, voies métaboliques etc.) et expliciter les liens entre ces acteurs (une enzyme catalyse une réaction, par exemple) Figure 36.

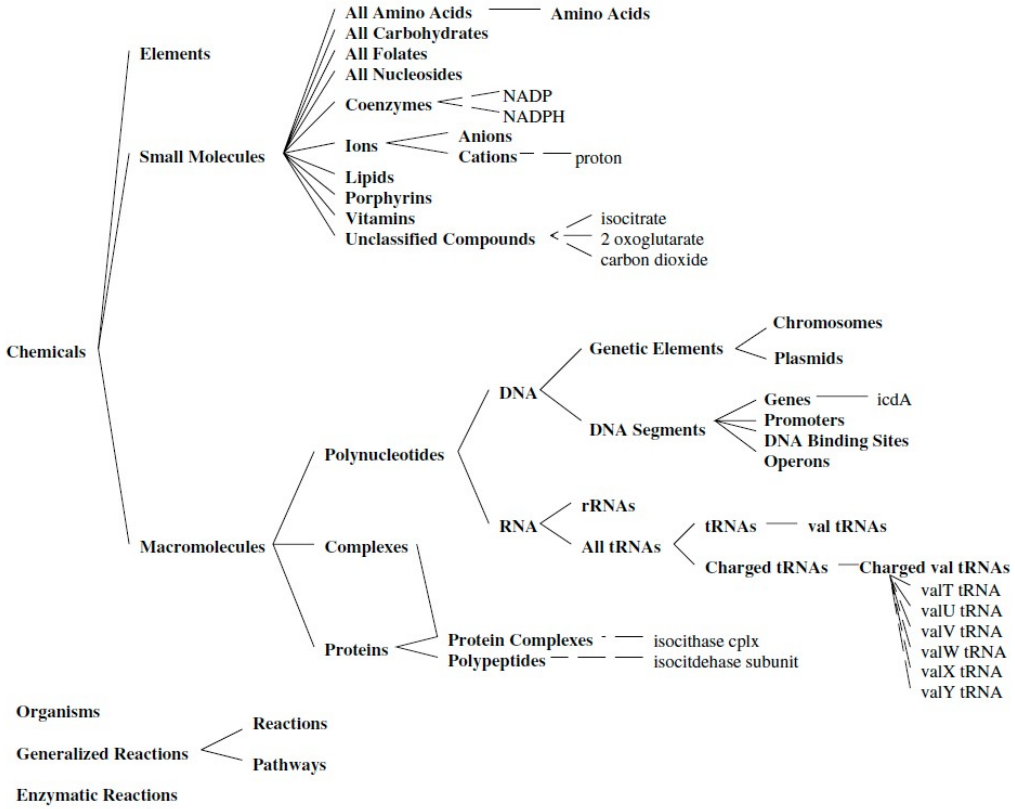


Figure 36 Organisation hiérarchique d'EcoCyc
Les principales classes (en gras), de quelques instances (normal) et de leurs relations.

Les relations « classe::sous-classes » sont en traits plein tandis que les relations « classes::instances » sont en pointillés

Cette ontologie est plus détaillée que celle utilisée dans KEGG qui, par exemple, ne prend pas en compte la notion de cofacteur pour une réaction enzymatique. L'ensemble des classes, instances et relations est encodé dans *Frame knowledge Representation System* appelé *OCELOT* et est utilisé par l'outil de visualisation Pathway-Tools (Peter D Karp, S. Paley, et al. 2002).

2.3.2 La base de données MicroCyc.

Pathway-Tools possède des outils de visualisation et de recherche avancés qui présentent pour notre étude un défaut majeur : il est très difficile d'accéder directement aux données à partir de scripts ou programmes autre que Pathway-Tools. Pour contourner cet inconvénient nous nous sommes servis de l'ontologie de BioCyc pour créer une base de données relationnelle. La base de données métabolique MicroCyc ne reprend pas en intégralité cette ontologie. En effet son objectif n'est pas d'être un duplicata d'*OCELOT* en base de données relationnelle, mais simplement une base de données qui contient les classes et les attributs d'intérêts pour un accès rapide et simplifié lors de nos analyses.

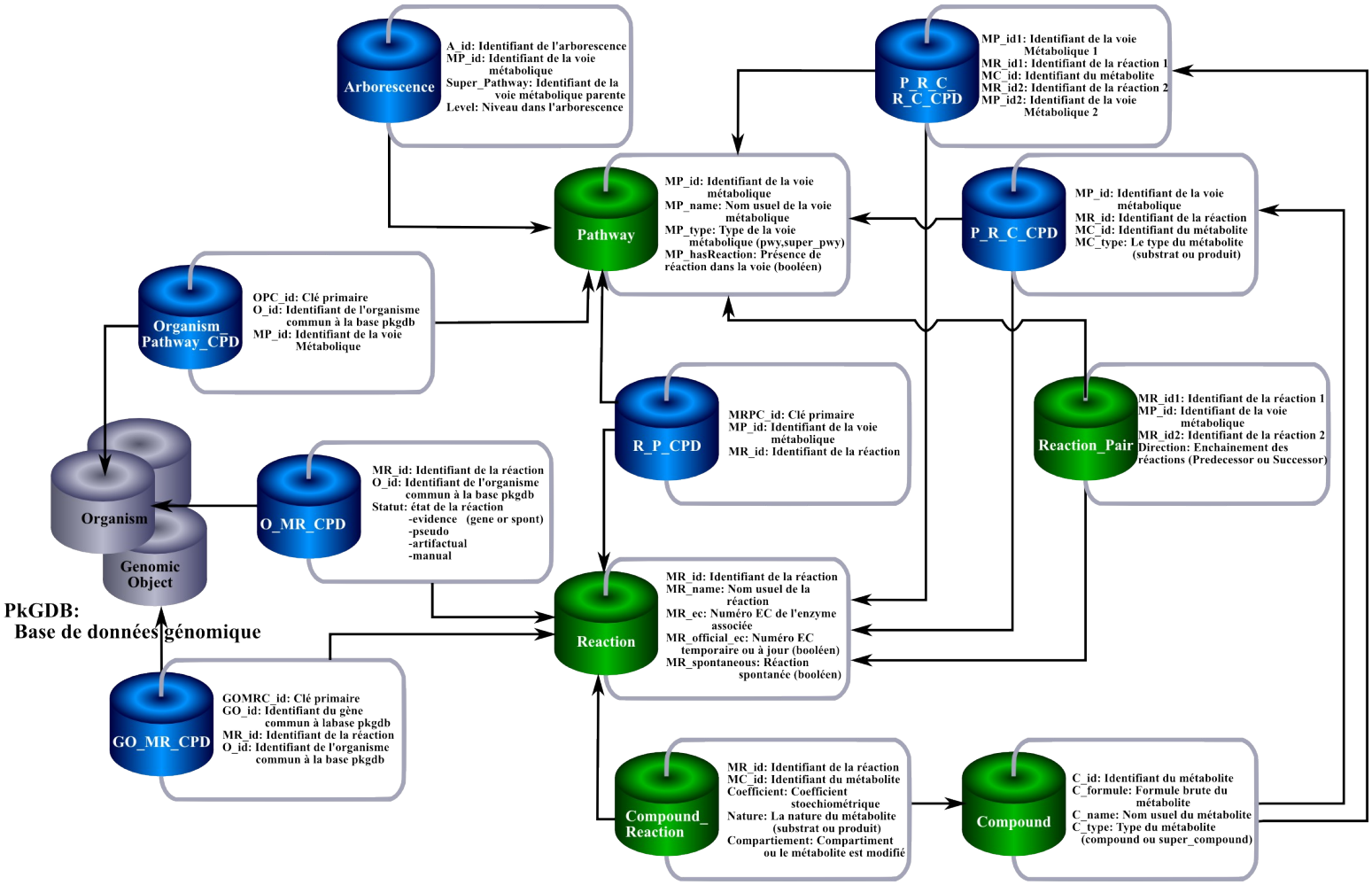


Figure 37: Illustration synthétique de la base de données MicroCyc.

En vert les tables représentant des classes de BioCyc et en bleu les tables de correspondance. En gris la base de données génomique de MicroScope. MicroCyc et MicroScope sont reliés par les identifiants des organismes (O_id) et les identifiants de gènes (GO_id).

La Figure 37 représente le schéma relationnel de MicroCyc, elle est composée de cinq tables principales : les trois acteurs principaux du métabolisme (réactions, voies métaboliques et métabolites) ainsi que deux dérivés (les métabolites au sein d'une réaction et les paires de réactions au sein d'une voie métabolique). Elle est également composée de sept tables de correspondance qui permettent de relier les différents acteurs entre eux et à la base génomique tout en apportant des informations complémentaires.
Les tables concernant les acteurs principaux ([métabolites], [réactions] et [voies métaboliques]) sont toutes constituées d'un champ *identifiant unique*, et d'un champ *nom usuel*. Les tables contiennent aussi des champs spécifiques : *Formule* et *type* (instance ou générique) pour les [métabolites] ; *Numéro EC*, *état du numéro EC*, et *spontanéité* pour les [réactions] ; *Type de la voie* (simple ou ensemble de voies), *possède une réaction* pour les [voies métaboliques]. Ce dernier champ permet de regrouper des voies métaboliques au sein du même ensemble lorsque celles-ci ne sont pas consécutives et qu'elles nécessitent une réaction afin de transformer le produit de la première voie en substrat de la suivante.
La table [Compound_Reaction], était à l'origine, une simple table de correspondance entre la table [Reaction] et la table [Compound]. Cependant ce lien n'était pas suffisant pour pouvoir décrire une réaction. Nous l'avons transformée en table principale par l'ajout d'un champ *Compartiment* et un champ *Stœchiométrie*.
De même la table [R_P_CPD] (R pour Reaction, P pour Pathway et CPD pour Correspondence) qui relie les réactions aux voies métaboliques, n'est pas suffisante pour décrire sans ambigüité l'enchaînement de réactions de la voie. C'est pourquoi nous avons ajouté la table [Reaction_Pair], qui indique les paires de réactions successives au sein d'une voie métabolique. Nous avons enfin ajouté une table de correspondance [O_MR_CPD*]* (O pour Organism MR pour MetaCyc Reaction et CPD pour Correspondence) qui permet de donner un statut aux réactions au sein d'un organisme. Nous considérons quatre cas de figure sur l'état de la réaction qui sont définis dans la Table 13.

Etat	Description
Evidence	La réaction est associée à au moins un gène ou est spontanée
Pseudo	La réaction est associée un pseudogène, elle est considérée comme non fonctionnelle
Artefactual	La réaction n'est ni spontanée ni associée à un gène (qu'il soit complet ou pseudo).
Manual	La réaction est rajoutée manuellement après le processus

Table 13: Les différents états possibles d'une réaction dans un organisme.

La table [GO_MR_CPD] (GO pour Gene Onthology, MR pour MetaCyc Reaction et CPD pour Correspondence) est importante puisqu'elle fait le lien entre la base génomique et la base métabolique de la plate-forme. L'identifiant unique du gène dans un organisme (*GO_id*) est associé à l'identifiant de la réaction (*MR_id*), ce qui assure une continuité du génome au métabolisme en toute transparence. Cette continuité est très importante puisqu'elle permet à la fois de faciliter les associations par homologie lors des processus de reconstruction, et de faire les comparaisons entre les différents réseaux en permettant des allers-retours entre les différents niveaux d'études (génome, enzyme, réaction et voie métabolique). Dans cette optique l'une des dernières améliorations de la partie génomique, est l'ajout d'un champ *MR_id* directement dans la fiche d'annotation des gènes.

.2.4 Méthodologie de reconstruction

2.4.1 Préparation des données

La préparation des données d'entrées nécessaires à l'utilisation de Pathway-Tools est en partie prise en charge par MicroScope grâce à l'annotation fonctionnelle et aux résultats enregistrés ; cependant ils ne sont pas utilisables directement. Pour reconstruire le réseau métabolique d'une souche, Pathway-Tools requiert comme donnée principale la séquence du génome au format *fasta* et un fichier contenant les informations relatives aux gènes au format propre de PathoLogic ou format *pf*. Les fichiers *pf* sont constitués d'une succession de blocs qui correspondent aux gènes et dont la structure est indiquée dans la Table 14. La création des fichiers *pf* est une étape importante pour trois raisons : 1) c'est l'élément qui nous permet d'introduire les connaissances diverses dans le processus de reconstruction. 2) c'est à ce moment que sont créés les liens entre MicroCyc et MicroScope. 3) l'utilisation des mêmes identifiants et noms unifie les données entre les différents réseaux reconstruits. C'est pourquoi ces fichiers *pf* sont réalisés avec le plus grand soin.
Les informations requises pour compléter les champs sont pour la plus part directement accessibles dans MicroScope. Ainsi, une simple requête permet d'obtenir l'identifiant unique du gène (*ID*), son nom (*NAME*) ses synonymes (*SYNONYM*), sa position (*STARTBASE, ENDBASE*), le produit du gène (*PRODUCT-TYPE*) et l'annotation fonctionnelle (*GENE-COMMENT*).

Nom	Description
ID	Identifiant unique du gène. L'identifiant du gène est le même que celui de la base de données MicroScope
NAME	Nom unique du gène
STARTBASE	Position du début du gène sur la séquence fournit en parallèle lors de la reconstruction
ENDBASE	Position de la fin du gène sur la séquence fournit en parallèle lors de la reconstruction
PRODUCT-TYPE	Produit du gène (protéine, ARN, …)
SYNONYM*	Synonyme du gène
GENE-COMMENT	Annotation fonctionnelle: Nom de la fonction métabolique et numéro EC
*FUNCTION**	Nom de la fonction
*EC**	Numéro EC
METACYC*	Identifiant de la réaction dans la base de données MetaCyc

Table 14 : Champs requis lors de la création d'un gène au format *pf*.
Le caractère « * » désigne des champs qui peuvent être multiples. Les champs en italique et en gras sont exclusifs : soit l'identifiant MetaCyc est connu soit on possède des informations sur la fonction et le numéro EC ; dans les cas où l'on dispose de la fonction, du numéro EC et de l'identifiant MetaCyc, la priorité est donnée à ce dernier.

Nous avons décidé d'utiliser comme *ID*, l'identifiant du gène de la base de données de MicroScope. Avec le travail d'homogénéisation de l'annotation, on aurait pu utiliser le nom usuel à la place de l'identifiant, puisque ce nom est identique pour tous les gènes homologues : on considère comme homologues des gènes avec une identité supérieure à 70% sur un recouvrement d'au moins 80% de la longueur de la séquence en acides aminés des protéines qu'ils codent. C'est lors de la création des fichiers *pf* que la mise en place du pivot intervient. Pour rappel, le pivot désigne le réseau métabolique de référence. Il sert de canevas pour la reconstruction des autres

réseaux ; l'identification des éléments communs entre le pivot et la souche en reconstruction fournit une base solide, et limite l'inférence de faux positifs, lors de l'utilisation de l'algorithme Pathologic. Appliqué aux souches d'*E. coli*, le pivot le plus évident est EcoCyc, soit le réseau de la souche K-12 MG1655. Pour chacune des souches à reconstruire, ou souches cibles, nous allons rechercher quels sont ses gènes homologues à ceux de K-12 MG1655. Puis pour chaque gène ainsi identifié, nous allons sauvegarder les associations gènes/réactions d'EcoCyc, dans le fichier *pf* spécifique de CFT073. Par exemple la souche CFT073 et le gène *ttdB* Figure 38. La fiche du gène (Figure 38B) donne les informations nécessaires à la constitution du bloc gène : le nom, les synonymes, les positions de début et de fin, le produit (champ *FUNCTION),* type du produit et le numéro EC. Comme le gène est homologue à celui de K-12 MG1655 (identité = 100% Figure 38A) les champs *function* et *EC* ne seront pas remplis. A la place, au champ *METACYC* est attribué la valeur de l'identifiant de la réaction MetaCyc « LTARTDEHYDRA-RXN » : son nom usuel est *L-tartrate dehydratase*, et elle est associée au gène *ttdB* dans le pivot. Ce champ *METACYC* permet de contourner le processus standard d'association par numéro EC ou nom du produit: il impose une association stricte entre le gène et la réaction. Ce raccourci, appliqué à tous les gènes homologues de la souche K-12 MG1655, nous permet de nous affranchir d'éventuelles erreurs, aussi bien dans l'annotation que durant le processus d'association gène/réaction. Dans le cas des gènes sans homologue dans K-12 MG1655, le champ *EC* sera rempli par le numéro EC s'il existe dans l'annotation initiale, et le champ *function* sera rempli par le produit de la fiche du gène.

A: Fiche d'homologie du gène ttdB entre CT073 et K-12 MG1655

Organism	Label	Gene	Product	maxLrap	minLrap	Ident %	Eval
Escherichia coli str. K-12 substr. MG1655	b3062	ttdB	L-tartrate dehydratase, beta subunit	0.99505	1	100	2.68445e-117

B: Fiche du gène chez CFT073

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms
CDS	3647965	3648570	606 (201aa)	+1	no	ttdB	ygjB
Product	L-tartrate dehydratase, beta subunit						
EC number	4.2.1.32						
PubMedId	3297921, 8371115, 93381464						
Product Type	e : enzyme						
Localization	1 : Unknown						
Class	1b : Function experimentally demonstrated in the studied species						
BioProcess	6.7 : Fermentation ;						
Roles	1.3.5 : Fermentation ;						

Figure 38 : Homologie et fiche du gène *ttdB*.
A) le gène *ttdB* de CFT073 et de K-12 MG1655 sont identiques à 100%, ce qui en fait des gènes homologues. B) La fiche du gène *ttdB* chez CFT073 qui contient les informations nécessaires à la formation d'un bloc gène dans un fichier *pf.*

Pour chaque organisme un couple de fichiers (un *fasta* et un *pf*) est produit et alimentera le logiciel Pathway-Tools. Toujours dans l'optique d'optimiser l'inférence des réactions nous avons décidé d'ajouter un dictionnaire associant à chaque numéro EC le nom de l'enzyme correspondant. Ceci nous permet de garder l'association

numéro EC/nom de l'enzyme à jour indépendamment de BioCyc qui possède son propre cycle d'actualisation.

2.4.2 Complexes

L'ontologie de BioCyc considère plusieurs types de macromolécules, les *polynucléotides* comprenant les molécules d'ADNs et d'ARNs, les *protéines* qui peuvent être des polypeptides ou des complexes de polypeptides, et les *complexes* qui sont des ensembles de protéines ou de complexes protéiques. Les polypeptides sont associés d'une part aux gènes (qui sont une sous classe de polynucléotides) et d'autre part aux complexes protéiques (Figure 39). Au sommet de l'organisation hiérarchique des complexes protéiques, se trouve l'*enzymatic-reaction*, autrement dit l'enzyme qui va catalyser la réaction (Figure 39). Cette structure n'est pas sans rappeler la GPR (partie 3.2.4 de l'introduction sur la modélisation), cependant, il existe deux différences : 1) La GPR ne prend pas en considération le nombre de sous-unités. 2) la GPR peut contenir des gènes régulateurs indispensables à la réaction, mais qui ne participent pas au complexe enzymatique. Nous avons déjà évoqué l'importance des GPRs dans les modèles du métabolisme à échelle de la cellule (toujours dans la partie 3.2.4), et c'est l'une des motivations de l'ajout des complexes dans le processus de reconstruction des réseaux. A cela il faut ajouter deux autres motivations : la propagation des connaissances et l'utilisation des complexes pour découvrir des différences (inconsistances ou véritables différences) entre nos réseaux métaboliques reconstruits.

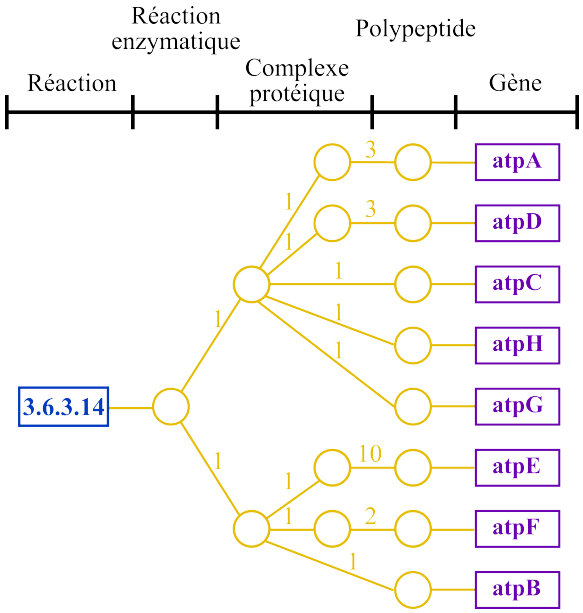


Figure 39 : Schéma gène-réaction.
Les 5 classes de BioCyc sont représentées : Les gènes, les polypeptides, les complexes protéiques, la réaction enzymatique et la réaction. Les chiffres indiquent le nombre d'unités nécessaires pour faire un sous-complexe.

Les complexes enzymatiques sont « organisme spécifique » et ne sont pas pris en considération par le logiciel Pathway-Tools lors de la reconstruction du réseau métabolique d'un nouvel organisme. Cependant comme nous l'avons évoqué précédemment, il nous a semblé pertinent d'ajouter la reconstruction des complexes lors du processus globale de la reconstruction du réseau métabolique. Toujours dans un souci de classification des différences entre le réseau de référence et celui

reconstruit, nous avons anticipé les inconsistances possibles lors de la reconstruction des complexes (Table 15).

Complexes		Référence	
		Tous avec un homologue	Sans homologue
Cible	Tous avec un homologue	Les mêmes protéines et enzymes sont impliquées	Enzyme incomplète
	Sans homologue	-Spécifique du réseau reconstruit -Possibilité d'avoir une isozyme	Possibilité de substitution dans le complexe

Table 15: Informations apportées par l'utilisation des complexes.
Il existe quatre cas 1) tous les complexes et sous-complexes d'une enzyme ont un homologue, 2) le réseau de référence présente des complexes sans homologues et donc la réaction est probablement non fonctionnelle ou un complexe inconnu du réseau cible compense le complexe manquant. 3) Il existe dans le réseau reconstruit des complexes sans homologue, il peut s'agir d'isozymes ou de complexes spécifiques de la cible. 4) Il existe des complexes spécifiques dans le réseau de référence et dans la cible : soit ce sont des substitutions, soit des isozymes.

Nous avons également différencié dans le cas des complexes sans homologues, les cas où l'ensemble des sous-unités est absent des cas où au moins une des sous-unités est présente. Cette classification des inconsistances a été mise en place pour faciliter d'éventuels travaux de curation manuelle des réseaux.
Pour cela à la fin du processus de reconstruction un tableau bilan est créé (Annexe 1). Il permet de voir rapidement les raisons d'absence d'un complexe et la liste des gènes des complexes de la souche pivot sans homologue dans la souche cible ; à noter que les gènes sont identifiés par les *ECKnumbers* puisque ce travail repose sur les gènes d'*EcoCyc*. Le logiciel Pathway-Tools ne possède pas de fonctionnalité pour la reconstruction des complexes, et le seul moyen d'intégration passe par une expertise manuelle dans une interface graphique, ce qui empêche toute automatisation. Pour pallier ce manque, nous avons développé un programme « tiers » en java: il insère automatiquement les complexes évalués positivement dans les Cycs nouvellement créés. Comme précisé dans l'article (1), c'est par homologie de séquence avec le pivot que les complexes de la cible sont évalués ; par conséquence l'inférence des complexes se limite à ceux du pivot. Nous appelons complexe complet, un complexe, dont toutes les sous-unités sont associées à des gènes fonctionnels (ce qui exclut les pseudogènes et les gènes partiels). Tous les complexes complets d'EcoCyc sont retrouvés au moins une fois complet dans un des réseaux reconstruits. On retrouve en moyenne dans nos réseaux 87% des hétéro-complexes complets (complexes formés de protéines différentes) ; 8% pour lesquels au moins une sous-unité est absente, et enfin 5% pour lesquels aucunes sous-unités n'a été identifiées. Ce nombre élevé d'hétéro-complexes trouvés par réseau, entraîne naturellement un fort taux de conservation des complexes dans nos réseaux : plus de la moitié (56%) d'entre eux sont communs à tous les réseaux reconstruits, et ce pourcentage atteint 80% pour les hétéro-complexes complets présents dans au moins 20 des 23 réseaux. (Figure 40).

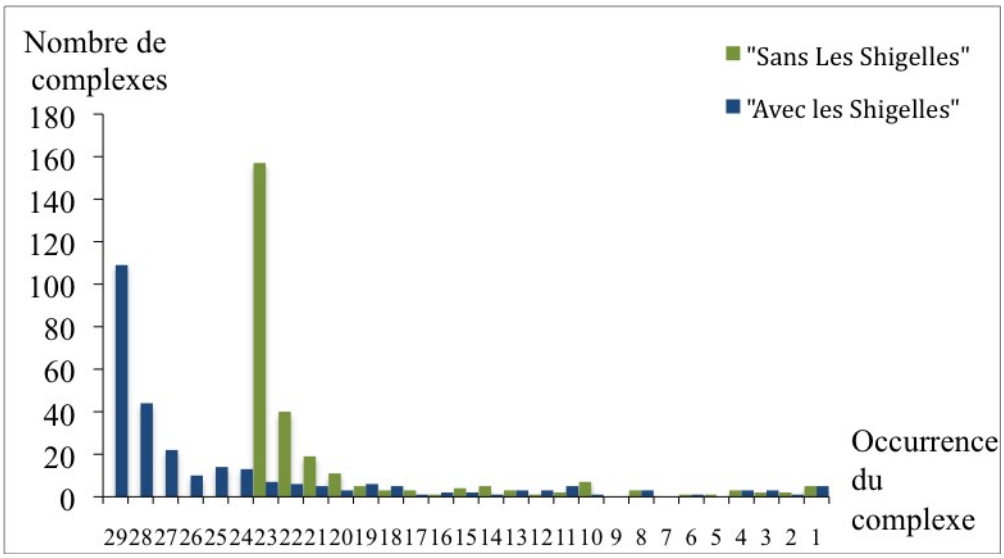


Figure 40: Occurrences des complexes dans les différents réseaux.
En bleu les données pour les 29 réseaux et en vert les données pour les 23 *E. coli*. Dans les deux cas la majorité des hétéro-complexes est présente dans tous les réseaux et très peu d'entre eux sont absents de plus de cinq réseaux.

Ce taux élevé s'explique principalement par un biais de connaissances, les voies et réactions ubiquitaires, sont celles les plus étudiées et dont les mécanismes sont les mieux décrits. Parmi ces descriptions, on trouve les produits de gènes et les complexes enzymatiques, il est donc normal de retrouver ces informations dans la majorité des réseaux reconstruits. La différence de résultats avec et sans *Shigella* (Figure 40), s'explique simplement par leur mode de vie. En tant que parasites, elles utilisent un ensemble de métabolites essentiels directement produits par leur hôte. Ceci entraîne une levée de certaines pressions évolutives, conduisant à l'apparition de pseudogènes, et la disparition des complexes associés ; des exemples sont donnés dans l'article 1 p1466.

2.4.3 Extraction et préparation des données métaboliques

Une fois la reconstruction du réseau métabolique terminée et l'inférence des complexes effectuée, une dernière étape intervient pour préparer nos données avant les analyses. Comme nous l'avons précisé dans la partie 2.3.2, les outils proposés par BioCyc permettent de visualiser et d'explorer graphiquement les réseaux métaboliques, mais ils sont peu et mal adaptés aux analyses détaillées. Il est possible d'interroger dynamiquement les réseaux de Pathway Tools par l'intermédiaire d'une API. Nous avons développez et mis en place un programme java, qui par l'API récupère les informations des Cycs une fois chargés dans le logiciel Pathway-Tools, et remplit la base de données relationnelle MicroCyc. C'est sur cette base de données que repose l'intégralité des travaux réalisés sur les réseaux métaboliques.

3 Première applications de la nouvelle stratégie de reconstruction

Nous avons déployé les processus qui permettent d'appliquer la stratégie de reconstruction à différents organismes. Avant de procéder à l'analyse présentée dans l'article, nous avons travaillé sur un sous-ensemble des souches. Après le succès de l'application de nos méthodes sur une dizaine de souches, nous sommes passéà une échelle supérieure en nombre de souches.

Nous avons expliqué dans la partie de l'introduction consacrée à la reconstruction des modèles (partie 3.2), que lors de l'initiation du projet de thèse, il n'y avait pas de méthode de reconstruction automatique de réseaux métaboliques de souches proches ou appartenant à la même espèce. Ce constat s'explique principalement par le manque de précisions des processus de reconstruction existants. Pour évaluer si notre stratégie est capable de capturer la diversité métabolique, nous avons reconstruit les réseaux de 17 *E. coli* uniquement (Table 16). Le processus utilisait à ce moment la version de 11.5 de BioCyc, ce qui explique la différence pour certains chiffres clés entre ces premiers travaux et l'article présenté au début de ce chapitre.

Souches	Pathogénicité	Groupe phylogénétique	Nombre de réactions
ED1a	Commensal	**B2**	1534
HS		**A**	1545
IAI1		**B1**	1546
K-12 MG1655		**A**	1555
K-12 W3110		**A**	1556
042	Pathogène de l'intestin	**D**	1528
55989		**B1**	1546
O127:H6 E2348/69		**B2**	1527
O157:H7 EDL 933		**E**	1527
O157:H7 sakai		**E**	1527
536	Pathogène hors de l'intestin	**B2**	1536
		B2	1524
CFT073		**B2**	1552
IAI39		**B2**	1547
S88		**B2**	1527
UMN026		**D**	1554
UTI89		**B2**	1547

Table 16: Premier ensemble de souches, sur lesquelles fut appliqué le nouveau processus de reconstruction des réseaux métaboliques.
Les souches sont réparties parmi les différents groupes phylogénétiques, et parmi les différents types de pathogénicité.

Nous avons analysé les similitudes et les différences entre les réseaux reconstruits pour évaluer la diversité métabolique récupérée par notre stratégie. Nous avons introduit le concept de *« core métabolisme »* comme l'ensemble des réactions (ou voies métaboliques) qui sont présentes dans l'ensemble des réseaux métaboliques étudiés et le concept de *« pan métabolisme »* comme l'ensemble des réactions (ou voies métaboliques) de l'ensemble des réseaux : il comprend donc le core métabolisme et la partie variable du métabolisme. La Table 17 résume la décomposition en pan et core métabolisme de nos réseaux.

	Réactions	Voies

			métaboliques	
Core	1419	83%	228	82%
Variable	284	17%	49	18%
Pan	1703		277	

Table 17: Core, variabilité et pan métabolisme pour les 17 souches de *E. coli*.

Si la grande majorité des réactions (83%) ou des voies métaboliques (82%) est commun à tous les réseaux, on observe quand même une certaine variabilité. L'utilisation d'un réseau pivot entraîne un biais, et ce malgré toutes les précautions prises dans notre nouvelle stratégie. La taille du réseau reconstruit est égale à celle du réseau pivot, auquel on ajoute des réactions spécifiques (A) et auquel on supprime des éléments spécifiques du pivot (B). De manière général l'ensemble A est plus petit que l'ensemble B : d'un côté le pivot est issu d'un long travail d'expertise qui s'appuie sur un grand nombre d'observations et d'expériences. Ainsi, Il existe dans EcoCyc des réactions dont la présence est prouvée expérimentalement, mais dont le ou les gènes sont toujours inconnus ; l'absence de preuves expérimentales, dans les différentes souches dont le réseau est reconstruit et qui sont moins étudiées, explique l'absence de certaines de ces réactions dans leurs modèles. De l'autre côté le nombre d'éléments spécifiques d'un modèle reconstruit est toujours sous-estimé, à cause du manque de connaissance relatif à cet organisme et d'une manière générale sur le métabolisme. Ce constat est visible dans la Table 16, où les deux souches de *E. coli* K-12 possèdent les réseaux les plus grands. Ce résultat nous a fait craindre que les réseaux métaboliques reconstruits soient de simples sous réseaux des souches K-12. Il n'en est rien, puisque le pan métabolisme est supérieur en nombre de réactions et de voies métaboliques à celui des réseaux des souches K-12. On dénombre ainsi près de 150 réactions et 20 voies métaboliques qui ne sont pas présentent dans le réseau pivot.
De plus nous avons observé que les réactions appartenant à la partie variable du métabolisme sont aléatoirement réparties dans les différentes réseaux Figure 41.

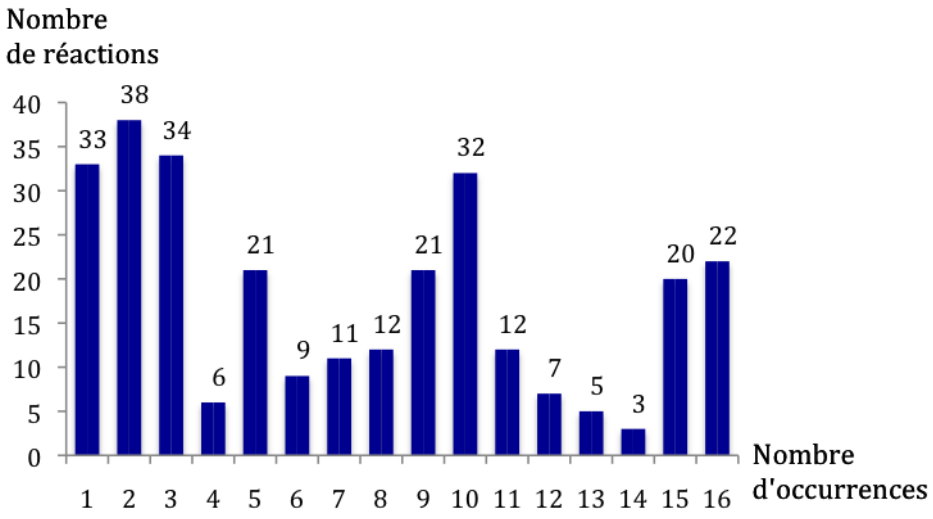


Figure 41 : Occurrences des réactions.
On distingue trois zones de fortes occurrences de réactions ; Les occurrences de 1 à 3 représentent les réactions qui sont spécifiques d'un faible nombre de réseaux ; les occurrences à 15-16 représentent des réactions que l'on retrouve dans pratiquement tous les réseaux ; le pic central aurait pu être lié au groupe phylogénétique ou pathogénique, mais aucun signal significatif n'a été observé sur la provenance de ces réactions.

L'ensemble des réseaux métaboliques présentant suffisamment de diversité, nous avons exploré cette diversité. Les résultats obtenus avec ce premier ensemble de réseaux métaboliques n'ont pas vocation à être pleinement exploités. Ils nous ont servi à observer des tendances et comportements globaux dans le but d'orienter et préparer l'analyse plus détaillée présentée dans l'article (partie 1 de ce chapitre).
Dans une précédente partie nous avons mentionné les quatre statuts possibles pour une réaction (Table 13 paragraphe 2.3.2). Ce statut permet de filtrer les réactions pour conserver, tout au long de nos différentes études, les réactions dont la présence ne fait aucun doute, c'est à dire celles associées au statut *evidence*. Deux principaux éléments justifient ce choix : tout d'abord l'un des objectifs de la thèse est de relier les phénotypes et génotypes par l'intermédiaire du métabolisme, et seules les réactions liées à des gènes permettent d'établir un tel lien. Ensuite, puisque parmi les réactions avec le statut *artefactual*, il est impossible de différencier un faux positif d'une réaction dont le gène est inconnu, nous préférons ne pas les prendre en compte pour éviter toute conclusion erronée. Les réactions spontanées ont le statut *évidence*, car si elles ne sont pas liées à des gènes, leurs substrats sont disponibles dans le réseau puisque ces métabolites sont utilisés par des réactions liées à des gènes.
Ce jeu de données est l'un des premiers de ce genre dans l'univers des réseaux métaboliques, j'ai dû concevoir une série d'analyses adaptées. La génomique est la ressource initiale du processus de reconstruction et le réseau métabolique est également à l'échelle de la cellule, je me suis inspiré de certaines méthodes de génomiques comapratives pour les transposer au métabolisme et à ce qu'on pourrait appeler de la métabolomique comparative. Je me suis intéressé aux comparaisons intra-espèce et j'ai par analogie au core et pan génome, introduit les notions de core et pan métabolisme telles que je les ai définies précédemment. Avant de rechercher des éléments métaboliques précis qui relient le phénotype au génotype j'ai regardé si le core et pan génome et le core et pan métabolisme présentent des caractéristiques communes. Que ce soit pour le génome ou le métabolisme, le pan augmente en fonction du nombre de souches. Le coefficient de la pente est constant pour le pan génome tandis que celui du pan métabolisme diminue à partir d'une dizaine de réseaux : le nombre de nouvelles réactions apportées par chaque nouveau réseau est de plus en plus faible. La plus grande différence entre ces deux univers est la proportion du « core » par rapport au « pan » : si dans le cas des génomes cette proportion n'atteint pas les 20% dans le cas du métabolisme elle dépasse les 80%. Le pan métabolisme est sous-estimé : si on est capable de prédire pratiquement l'ensemble des gènes d'un génome, on est pour le moment incapable d'estimer l'ensemble des réactions d'un organisme. En effet s'il existe différentes méthodes de prédiction de gènes à partir des séquences même si leur fonction reste inconnue, il est aujourd'hui impossible de déterminer toutes les réactions à partir des séquences. Néanmoins ce manque de connaissance sur les réactions métaboliques, ne peut expliquer à lui seul, l'écart des proportions observées. Cette constatation montre une très forte conservation du métabolisme, et par conséquent une forte conservation des gènes liés aux réactions, que l'on peut supposer supérieure aux gènes codant pour d'autres fonctions biologiques non liées au métabolisme. J'ai ensuite regardé si toutes les fonctions et processus métaboliques étaient équivalents en termes de conservation. Le nombre de réseaux métaboliques est trop limité dans cet ensemble de test, c'est pourquoi cette analyse a été réalisée sur l'ensemble des réseaux utilisés dans l'article (partie 1) et la partie 3.1 de ce chapitre.
J'ai par la suite envisagé différentes façons d'étudier la diversité de nos réseaux, et j'ai volontairement évité les analyses topologiques sur les graphes métaboliques. En

effet, la proportion importante du core métabolome me laissait craindre peu de différence quant à la structure même des réseaux ainsi que dans leurs organisations modulaires et hiérarchiques.
Le caractère infectieux opportuniste d'*E. coli* et sa plasticité génomique, ne me permettaient pas d'avoir de fort apriori sur la diversité contenu dans nos réseaux métaboliques. J'ai donc choisi parmi un ensemble de méthodes d'analyse statistique, celle qui nous paraissait le mieux correspondre à cette situation : l'Analyse des Correspondances Multiples (ACM). L'ACM fait partie des analyses descriptives multidimensionnelles appelées méthodes factorielles, et plus précisément elle se situe dans le prolongement de l'analyse en composantes principales. A partir du contenu en réactions portant le statut *evidence*, l'ACM permet de représenter graphiquement la proximité des réseaux métaboliques suivant leur contenu réactionnel. Pour cela, les réseaux sont distribués le long d'axes, eux-mêmes construits à partir de la contribution des différentes réactions. Cette contribution est différente pour chacune des réactions et donne l'importance de sa présence ou absence dans la diversité métabolique. A l'instar de l'analyse du core et pan métabolisme, l'application de l'ACM à l'ensemble des réseaux de tests avait pour vocation de détecter les tendances, si elles existent, de la diversité métabolique. Il est clairement apparu (Figure 42) que les réseaux se regroupent en fonction de leur groupe phylogénétique ; de plus les groupes sont parfaitement indépendants. Le premier axe explique plus de 20% de la diversité métabolique. Il sépare totalement les réseaux métaboliques des souches du groupe phylogénétique B2, des autres réseaux. Le second axe lui sépare les réseaux des souches A et B1 des D et E. Il faut noter que les souches des groupes A et B1 sont majoritairement commensales tandis que celles des groupes E et D sont majoritairement pathogènes.

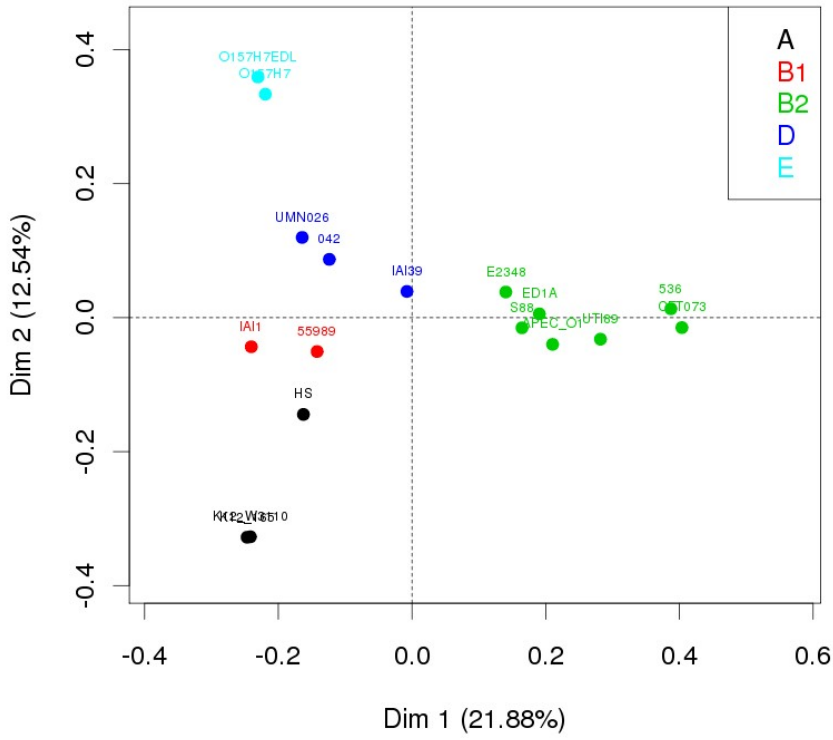


Figure 42: ACM sur les 17 réseaux métaboliques initiaux.
Les réseaux sont colorés suivant leur groupe phylogénétique.

Les résultats précédents montrent que la stratégie développée pour automatiser le processus de reconstruction de réseaux métaboliques est effective. Ainsi, bien que les

réseaux produits soient issus de souches proches, ils possèdent suffisamment de différences pour une analyse détaillée.
L'utilisation d'un réseau pivot comme canevas, entraîne néanmoins un biais : plus une souche est proche de la référence, plus son réseau sera complet. On observe ce phénomène avec la souche K-12 W3110, qui à quelques différences génétiques près, est une copie de la souche de référence et cette souche possède le plus grand réseau reconstruit (voir Table 16). Malgré ce biais, les réseaux restent homogènes en taille ce qui permet des analyses plus approfondies. Les premières tendances observées font état d'une forte conservation du métabolisme ainsi que d'un fort lien entre la phylogénie et la diversité métabolique.

.3.1 Résultat complémentaire de l'article Réseaux.

Fort des résultats préliminaires, j'ai décidé d'étendre mon ensemble de réseaux à six nouvelles souches de *E. coli* et six *Shigella*. L'intégration du processus de reconstruction au sein de la plate-forme MicroScope et la mise à jour régulière des données de la base me permettent, non seulement d'ajouter des réseaux à mon étude, mais aussi de mettre à jour les réseaux déjà existants. Cette mise à jour porte sur des améliorations génomiques : actualisations des annotations des souches déjà présentes et homogénéisation de celles apportées par de nouvelles souches. Elle comprend aussi des améliorations métaboliques, notamment avec, d'une part la mise à jour de notre dictionnaire numéro EC/nom de l'enzyme et d'autre part, l'utilisation de la dernière version de BioCyc disponible. Concrètement, ce changement de version s'accompagne d'un important gain d'information puisqu'entre la version 11.5 (jeu de test) et la version 14 (jeu d'analyse) la base de données EcoCyc est passée de 1550 *reaction frame* à 1815 (les *réaction frame* correspondent aux réactions d'EcoCyc). Lors de l'extraction des réactions j'ai recentré mes activités sur les réactions enzymatiques et métaboliques, car au fur et à mesure des versions, BioCyc s'est doté de plus en plus d'éléments et notamment les réactions de signalisation et de régulations. Ainsi parmi les 1815 *reaction frames*, 1397 correspondent à des réactions enzymatiques.
Je vais maintenant revenir sur quelques points de l'article qui mérite de plus amples informations. Tout d'abord, je vais donner un supplément d'information sur la comparaison des différentes façons de reconstruire les réseaux métaboliques. Pour mémoire, les trois différentes stratégies de reconstruction sont : a) Pathway-Tools directement appliqué aux génomes annotés issus des banques de données. b) Pathway-Tools appliqué aux génomes qui ont subi le processus d'annotation de MicroScope et c) Pathway-Tools avec le dictionnaire enzymatique, le réseau métabolique de référence EcoCyc et le processus d'annotation de MicroScope appliquée aux génomes. Pour éviter un biais introduit par les différentes versions des bases de données et des logiciels, j'ai lancé les différentes stratégies de reconstruction les unes à la suite des autres. L'application des trois stratégies n'a pu être effectuée sur toutes nos souches car certaines ont été directement annotées par MicroScope, et donc bénéficient directement de l'homogénéisation de l'annotation. Pour compléter l'article, je vais détailler le gain sur les voies métaboliques (Figure 43) ; les résultats sur le gain en réaction sont disponibles dans l'article (partie 1 de ce chapitre). La complétion des voies métaboliques est directement reliée aux éléments faux positifs : il est impossible de savoir si le trou dans la voie est dû à un gène inconnu, ou s'il s'agit d'une sur prédiction : un faux positif.
L'effort d'homogénéisation d'annotation semble diminuer le nombre de voies métaboliques inférées entre la stratégie (a) et (b) ; cette diminution est en fait une

amélioration des réseaux : la diminution des annotations incomplètes ou erronées du cas (a) empêche l'inférence de réactions « faux positives » qui à leur tour empêchent l'inférence de voies métaboliques erronées. On constate également que, toujours dans la stratégie (a), près de la moitié des voies sont incomplètes. Si l'homogénéisation de l'annotation permet de réduire le nombre d'erreurs, l'ajout du réseau pivot permet de récupérer des voies métaboliques spécifiques de l'espèce. Ces voies sont issues d'une expertise manuelle, on peut être confiant quant à leur présence comparée aux sur-prédictions de la stratégie (a).

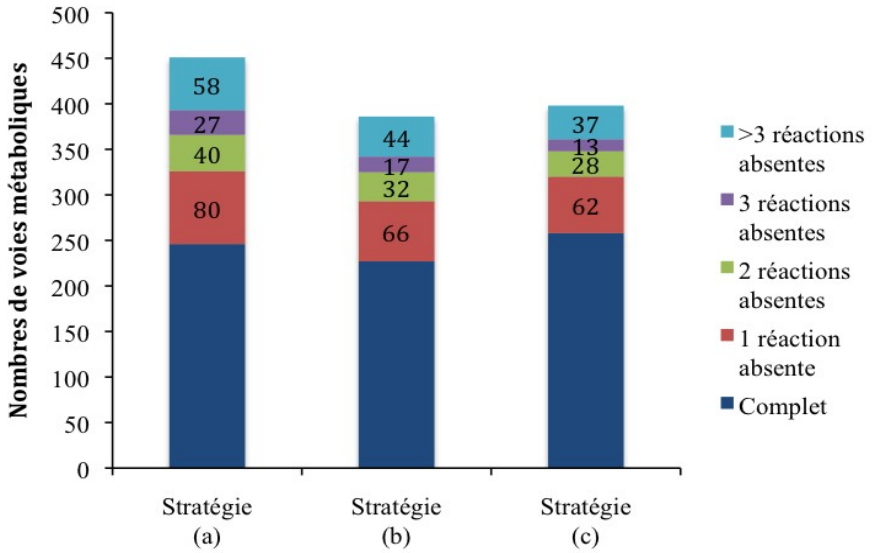


Figure 43: Complétion des voies métaboliques.
Les chiffres donnent le nombre de voies métaboliques de chacune des catégories. Entre les stratégies (a) et (b) le nombre total de voies métaboliques a diminué, tout comme le nombre de trous dans les voies métaboliques.

L'analyse de l'ACM se limite aux deux premiers axes, puisque comme le montre la Figure 44 dès le troisième axe, on observe un décrochage de la contribution de l'axe : celui-ci est pratiquement divisé par deux.

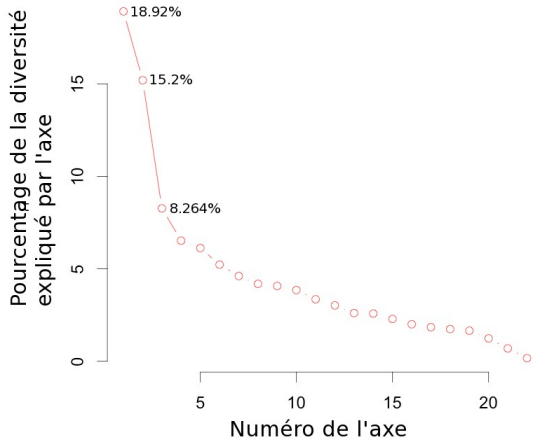


Figure 44: Diversité métabolique expliquée par les différents axes de l'ACM.
On observe un décrochage des valeurs dès le troisième axe.

Le troisième axe apporte cependant un fait remarquable : il sépare le groupe phylogénétique A en deux sous-ensembles (A_1 et A_2) et de même pour le groupe phylogénétique B1 ($B1_1$ et $B1_2$). .Ainsi les deux souches K-12, qui appartiennent au sous-groupe A_1 sont regroupées avec les souches 55989 et IAI1 qui font parties du sous-groupe $B1_1$, nous nommerons cet ensemble (e1). Les autres souches du sous-groupe A_2, HS et ATCC8739, sont groupées avec les souches du sous-groupe $B1_2$ à savoir FE11 et E24377A, et forment l'ensemble (e2). Ces regroupements de souches sont en accord avec l'arbre métabolique, où ces deux ensembles (e1) et (e2) sont dans des sous arbres différents (Figure 5 de l'article 1). J'ai regardé les différences entre ces deux groupes mais aucun élément métabolique (réaction ou voie métabolique) réellement significatif n'a été décelé.
L'analyse de la pathogénicité des *E. coli*, s'est révélée peu efficace, une analyse sans apriori ne fournit aucun résultat. J'ai utilisé une méthode statistique supervisée appelée « arbres de régression ». Ceci m'a permis de mettre en évidence quelques réactions. Cependant il s'agit plus d'un *« proof of concept »* méthodologique, que d'une analyse qui à vocation d'obtenir des résultats précis et ce, pour deux raisons. Premièrement, les connaissances métaboliques actuelles ne sont pas suffisantes ; parmi les réactions mises en évidence certaines étaient « putatives ». Deuxièmement, le nombre de réseaux est insuffisant au regard du nombre de réactions de la partie variable du métabolisme. Néanmoins, les méthodes proposant des combinaisons de présence et d'absence de réactions donnent de meilleurs résultats que les méthodes basées sur le dénombrement des occurrences.
Les résultats obtenus dans l'article, montrent que si pour la phylogénie une vingtaine de réseaux suffit à établir des relations entre évolution et métabolisme, ce nombre est trop faible pour étudier la pathogénicité de part la diversité des mécanismes et le caractère opportuniste des *E. coli*. Puisque d'autres génomes d'*E. coli* sont disponibles, j'ai décidé de relancer le processus sur ces nouveaux génomes.

4 Application à un plus grand ensemble de souches

La dernière partie des travaux sur la reconstruction des réseaux métaboliques est consacrée au passage au « haut-débit » et à l'estimation des limites de l'efficacité du pivot en fonction de la proximité des organismes. Nous savons que le processus est effectif pour une vingtaine de souches, nous sommes donc passé à un ordre de grandeur supérieur, c'est à dire plus d'une centaine de souches. Lors du passage de l'ensemble des réseaux tests, à l'ensemble des réseaux de l'article, nous avions profité des mises à jour de BioCyc ; il en est de même pour cette nouvelle reconstruction de réseau, Ainsi nous avons rajouté 84 nouvelles *E. coli* dont certaines sont caractéristiques des différents clades. A ces souches viennent s'ajouter 8 souches « non *E. coli »* : 3 *Salmonella*, 2 *E. alberti*, 3 *E. fergusonii*. L'ensemble des reconstructions s'est déroulé sans erreurs. Nous disposons donc au final de 107 souches d'*E. coli* et de 14 souches d'espéces proches. La composition des réseaux reconstruits est disponible en annexe (Annexe 1). Le résumé du nombre de réactions associées à des gènes est donné dans la Figure 45. L'ensemble des réseaux reconstruits à partir des génomes des *E. coli* présente une homogénéité en taille qui varie de l'ordre d'une centaine de réactions. La seule exception est la souche e267 dont le réseau est quelque peu différent en taille et se rapproche plus de la taille des réseaux des *Shigella.*

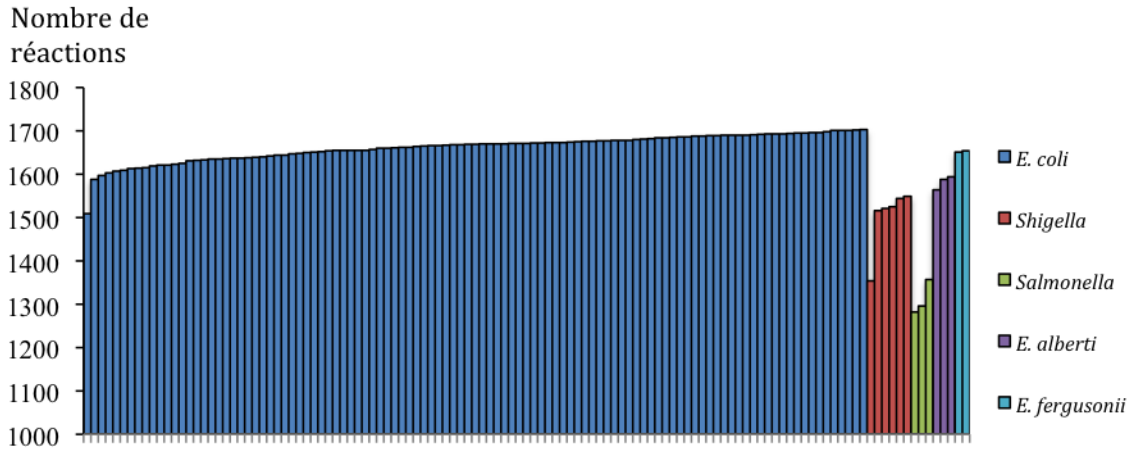


Figure 45: Nombre de réactions des 121 réseaux métaboliques.
En bleu les réseaux des *E. coli*, en rouge les *Shigella*, en vert les *Salmonella*, en violet les *E. alberti* et en bleu claire les *E. fergusonii*.

On peut observer un lien entre la taille du réseau reconstruit et la distance génétique. Dans la phylogénie des espèces représentées, les *Shigella* sont inclus dans le sous arbre des *E. coli*. Le côté parasitique des *Shigella* est responsable de la différence de taille. Parmi les autres souches, les plus proches parents des *E. coli* sont les *E. fergusonii* qui possèdent les réseaux les plus grands après les *E. coli*. Puis viennent les *E. alberti* qui ont des réseaux plus petits que les *E. fergusonii,* enfin, notre « outgroup » est composé des *Salmonella* qui ont les réseaux les plus petits (Table 18).

	Réactions			
	Total	Avec gène	Pseudo	Sans gène
Salmonella	1529	1312	2	215
Shigella	1788	1502	77	210
E. alberti	1806	1582	14	210
E. fergusonii	1892	1653	3	237
E. coli	1883	1662	17	204

Table 18: Moyenne sur les différents types de réactions regroupées par espèces.
On constate une forte différence de taille entre les souches de types *Escherichia* et les *Salmonella*. C'est principalement le nombre de réaction avec gène qui différencie la taille des réseaux.

Il ne faut pas en déduire que les réseaux des *E. coli* sont les réseaux les plus grands, Entre le réseau de référence EcoCyc et celui de référence de *Salmonella enterica* serovar Typhimurium str. LT2, il n'y a que 80 réactions enzymatiques de différence. Ces réseaux mettent en évidence les deux limites de la nouvelle stratégie : celle-ci est performante lorsqu'il s'agit de souches de la même espèce, l'ajout du pivot devient inopérant dans le cas de souches provenant d'espèces différentes. Ainsi, une méthode automatisée donnera des réseaux d'une qualité moindre qu'une reconstruction experte.
A l'instar de l'article, nous avons calculé l'évolution du core et pan métabolisme pour les 107 réseaux d'*E. coli* (Figure 46). Cependant, en raison de la qualité moindre des nouveaux génomes, j'ai pris en compte les pseudos réactions. Les génomes utilisés précédemment sont finalisés, tandis que ces nouveaux génomes sont, pour la plus part, non finalisés et sous la forme de contigs qui contiennent de nombreux pseudogènes qui sont vraisemblablement dus à des erreurs de séquençages. Le pan métabolisme est constitué de 2032 réactions et le core de 1345 réactions soit 66%. A

noter que sans l'inclusion des « pseudo-réactions » le core a une valeur de seulement 909 réactions. Comme mentionné dans les conclusions de l'article, la proportion du core a diminué mais elle reste toujours supérieure à la moitié du pan.

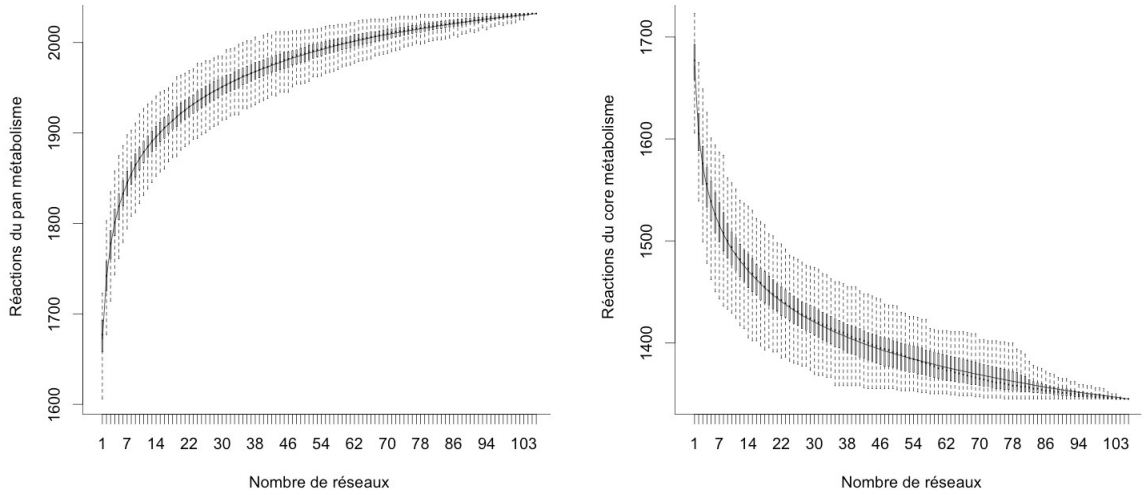


Figure 46: Pan, core métabolisme.
L'évolution est calculée pour 1000 différents ordres aléatoires des réseaux.

L'analyse des courbes et notamment de la deuxième partie de celles-ci, révèle une stabilisation du nombre de réactions. Ceci rejoint les conclusions de l'article : cette saturation est en partie due à une diversité métabolique plus faible que la diversité génomique. Une autre raison est le manque d'information spécifique sur les différentes souches. J'ai examiné plus en détail le lien entre la taille du pan métabolisme et le nombre de réactions. Pour cela, j'ai calculé la différence entre le nombre de réactions aux réseaux à $n+1$ et n, et obtenu la courbe de la Figure 47. Très rapidement l'ajout d'un nouveau réseau apporte peu de nouvelles réactions. Ainsi dès 9 réseaux, l'apport de chaque nouveau réseau est de moins de 10 réactions. A partir de 15 réseaux on passe sous le seuil des 5 nouvelles réactions par réseau. Enfin à 66 réseaux, tout nouvel ajout de réseaux apporte moins d'une réaction. J'ai de la même manière regardé la stabilisation du core et de la partie variable du métabolisme. Le core métabolisme diminue de moins de 10 réactions après 7 réseaux, de moins de 5 réactions après 13 réseaux et de moins d'une réaction après 55 réseaux.

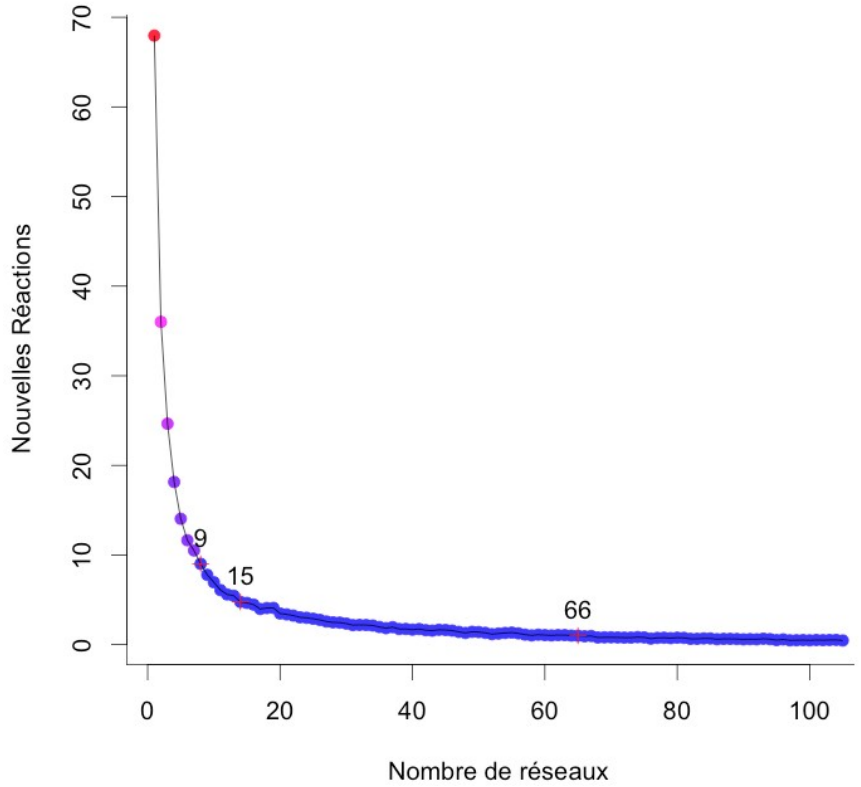


Figure 47: Evolution de l'apport de nouvelles réactions en fonction du nombre de réseaux.
Plus la couleur est froide moins l'ajout de nouvelles réactions est importantes. Les croix rouges et les chiffres au dessus indiquent à partir de combien de réseaux le nombre de nouvelles réactions est inférieur à 10 (à 9 réseaux), 5 (à15 réseaux) et 1 (à 66 réseaux) par réseaux ajoutés joutés.

Ces résultats montrent la limite d'utilisation du pivot et des connaissances métaboliques actuelles sur les *E. coli*. Une manière de lever cette limite serait de prendre une nouvelle souche de *E. coli*, de préférence éloignée de la souche *K*-12 MG1655 dans l'arbre phylogénétique ; sur cette nouvelle souche il faudrait effectuer un nouveau travaille de recherche fondamentale pour créer un nouveau réseau métabolique de référence. Puis modifier notre stratégie pour prendre en compte ces deux réseaux pivots.

J'ai examiné la distribution des occurrences des réactions dans les 107 réseaux reconstruits ; comme durant les précédentes analyses, j'ai constaté une distribution en forme de U (Figure 48), ce qui signifie qu'une large partie des réactions est soit spécifique d'un faible ensemble de réseaux soit à l'inverse commune à la majorité des réseaux. Entre ces deux extrêmes aucune tendance sur les réactions ou réseaux impactés n'est observée.

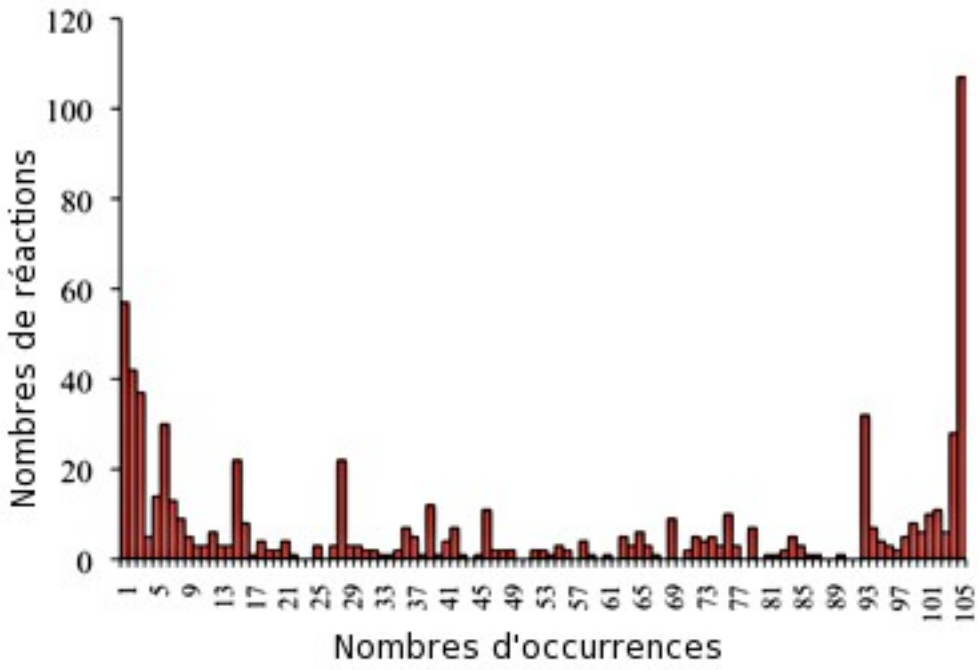


Figure 48: Occurrences des réactions pour les souches de *E. coli.*
Le core métabolisme est de 1345 réactions.

Enfin j'ai réalisé l'arbre métabolique sur l'ensemble des 121 réseaux (Figure 49). J'ai utilisé les mêmes réactions que pour les calculs précédents : les réactions associées à des gènes et des pseudogènes. L'arbre obtenu (Figure 49) est cohérent avec celui de l'article, et est en accord avec l'arbre phylogénétique. La principale différence reste la même dans les deux arbres métaboliques : la position des *Shigella* dans l'arbre. Elles sont toujours sur un sous-arbre différent des *E. coli* alors que dans l'arbre phylogénétique elles sont situées parmi les *E. coli.* Les nouveaux réseaux métaboliques s'intègrent parmi les anciens, à l'instar des génomes dans l'arbre phylogénétique. Les réseaux issus de souches du même groupe phylogénétique ont tendance à se regrouper dans l'arbre métabolique, tandis que les réseaux appartenant à des souches qui ne sont pas de *E. coli* sont en marge des deux arbres.

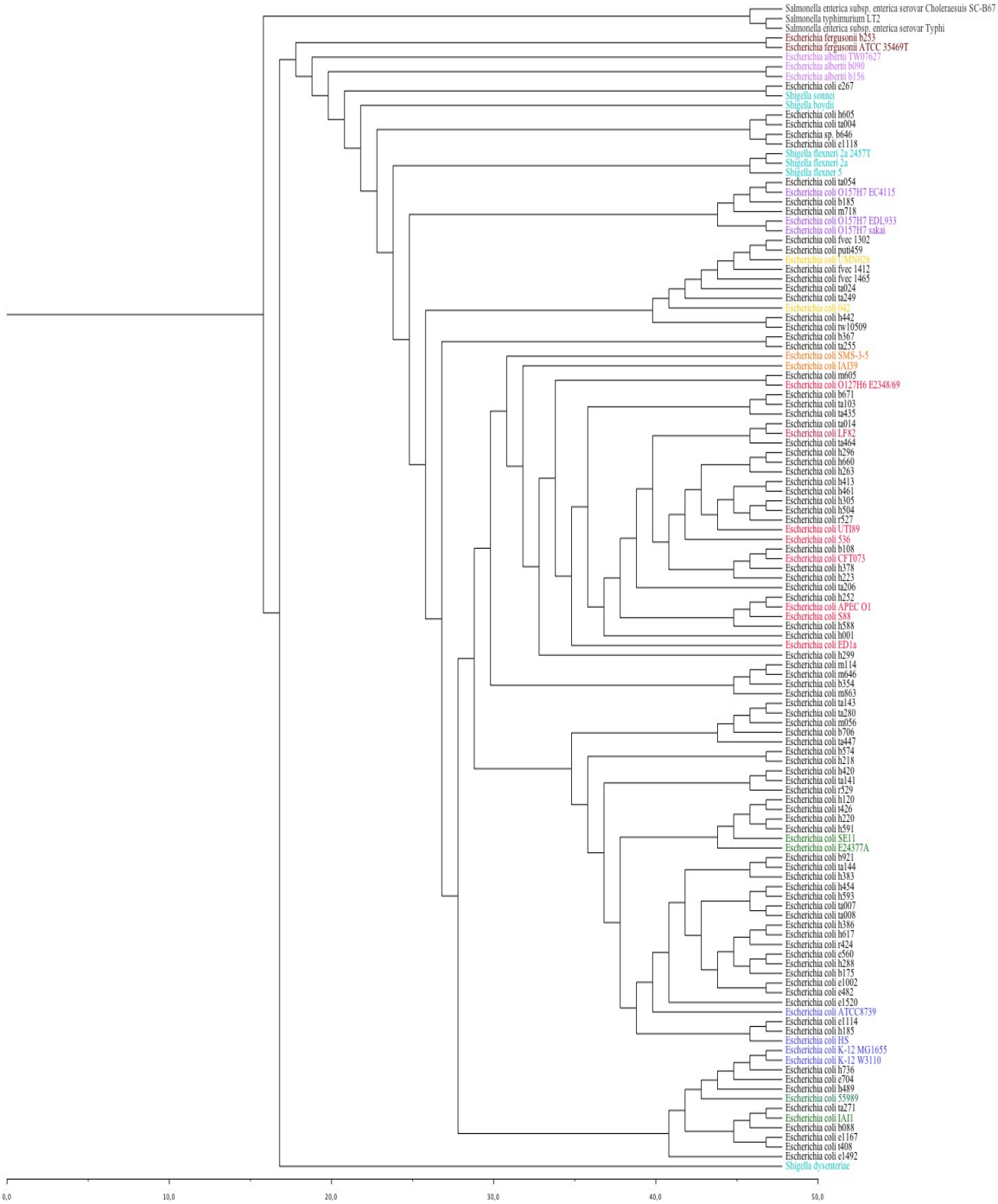


Figure 49 : Arbre métabolique des 121 réseaux métaboliques.

Il a été construit par neighbor-joining sur l'ensemble des réactions avec gènes (incluant les pseudogènes) ; Nous avons coloré les réseaux remarquables : Bleu clair les *Shigella*, bleu foncé groupe phylogénétique A, vert le groupe phylogénétique B1, rouge le groupe phylogénétique B2, jaune groupe phylogénétique D, violet foncé groupe phylogénétique E, orange groupe phylogénétique F, violet clair *E. albetii*, rouge foncé *E. fergusonii*, et gris *Salmonella*.

5 Conclusions

Lors de l'initialisation du projet de thèse et en regard de la diversité génomique des *E. coli*, l'élaboration d'une stratégie de reconstruction automatique des réseaux métaboliques à l'échelle de la cellule, comprenant suffisamment de diversité, nous paraissait possible. Afin d'y parvenir, nous avons élaboré cette stratégie avec, comme préoccupation principale, l'utilisation des connaissances métaboliques déjà produites que ce soit celles référencées dans les bases de données métaboliques ou celles issues spécifiquement des organismes, grâce aux réseaux métaboliques de référence.

.5.1 Apports de la méthodologie

La stratégie développée pour la reconstruction des réseaux métaboliques est basée sur la propagation des connaissances. Celle-ci prend deux axes différents : d'une part la propagation d'une connaissance hautement détaillée sur un organisme précis, et d'autre part une connaissance plus éparse répartie entre divers organismes. L'union de ces deux axes permet d'obtenir des réseaux avec une base commune, sur laquelle va venir s'ajouter les connaissances spécifiques à l'organisme dont on souhaite reconstruire le réseau. Les réseaux ainsi reconstruits possèdent ainsi un cœur fiable dont la présence des réactions et des voies métaboliques est sure. Cette base solide à comme effet de grandement diminuer le nombre de faux positifs parmi les réactions et les voies métaboliques inférées par les algorithmes des processus de reconstruction. Au final, à partir de cette base commune et grâce à la propagation des connaissances souches spécifiques, nous avons obtenus des réseaux de plus grandes tailles et avec un ratio de réactions associées à des gènes plus important que dans les réseaux issus des méthodes de reconstruction dite traditionnelles, et ces réseaux reconstruits avec cette nouvelle stratégie présentent suffisamment de diversité pour effectuer une analyse approfondie.

Il existe cependant plusieurs limites à cette stratégie ; tout d'abord l'obligation d'avoir un réseau de référence de qualité. Si pour les organismes modèles, les réseaux existent et sont mis à jour plus ou moins régulièrement, ce n'est pas le cas pour la majorité des organismes. Dès lors, appliquer cette stratégie à une nouvelle espèce nécessitera la création d'un réseau de référence. C'est le cas de l'espèce *Acinetobacter* dont la création d'un réseau de référence pour la souche *baylyi* a nécessité plusieurs années de travail se basant sur des analyses et expériences *in-vivo* et *in-silico*.

Si les problèmes liés à la qualité de l'annotation ont été résolus en grande partie, le processus reste tributaire de la qualité et de l'état de finalisation des séquences du génome des organismes. Enfin, la méthode est naturellement dépendante de l'état des connaissances métaboliques contenues dans les bases de données. Ces limites ne sont pas spécifiques à notre stratégie.

Le processus permet une mise à jour simple et rapide de l'ensemble des réseaux reconstruits lorsqu'une des ressources (séquence, annotations, base de données enzymatiques ou métaboliques) est mise à jour. Cette capacité est importante : on constate qu'au fur et à mesure des nouvelles versions de MetaCyc et d'autres ressources, la taille moyenne des réseaux augmente.

L'ajout du réseau pivot permet un gain de qualité et de complétion des réseaux. Il introduit toutefois un biais dans la composition des réseaux : tant que les souches sont proches de la souche de référence le gain est significatif, mais plus la souche dont on reconstruit le réseau est éloignée de cette souche de référence, moins l'effet du pivot sera visible pour au final être négligeable.

.5.2 Diversité métabolique

La diversité métabolique provient, pour une partie, des éléments métaboliques présents dans le réseau de référence mais absents dans d'autres réseaux. L'autre apport de la diversité est l'utilisation de la base de données métabolique généraliste. Elle permet de récupérer des éléments métaboliques qui ne sont pas présents dans le réseau de référence. Cette diversité s'est révélée exploitable dès l'analyse réalisée avec le sous-ensemble de réseaux de test, et les tendances observées sur un faible nombre de réseaux sont toujours valables sur un ensemble plus grand.

5.2.1 La différence entre le génome et le métabolome

Bien qu'issue des génomes, la diversité métabolique et la diversité génomique différent dans leur comportement. Alors que le core génome compte pour une très faible fraction du pan génome puisqu'il est de l'ordre de 10%, le core métabolisme atteint plus de 60% du pan métabolisme. La plus forte conservation des gènes liés au métabolisme était attendue, puisque des analyses sur le pan génome des *E coli* ont montré que les éléments génétiques les plus variables sont les IS et les prophages ; à contrario les éléments les plus conservés sont ceux dont la fonction est connue. La conservation était également attendue de part les connaissances métaboliques actuelles. Je l'ai évoqué dans l'introduction : il existe de nombreuses voies métaboliques ubiquitaires aussi bien dans l'anabolisme que dans le catabolisme, dont la présence est essentielle à la vie des organismes. L'évolution du pan génome et du pan métabolome en fonction du nombre de génomes/réseaux métaboliques diffère également. Alors que le pan génome est en constante augmentation, le pan métabolisme lui sature assez rapidement. Ainsi à partir d'une vingtaine de réseaux l'augmentation devient beaucoup moins prononcée et à partir de 70 réseaux, l'augmentation est de moins d'une réaction par nouveau réseau ajouté. Une autre différence entre les deux est d'ordre méthodologique ; à ce jour on est incapable d'évaluer l'intégralité de ces deux domaines. Néanmoins dans le domaine de la génomique on arrive maintenant à estimer la quasi-totalité des CDSs d'un organisme. Malheureusement, une partie de ces CDSs restent sans fonction précise ; fonction qui pourrait être métabolique. Par conséquence il est impossible d'estimer précisément les capacités métaboliques totales de nos réseaux. Le core métabolisme obtenu dans cette étude diminuera au fur et à mesure que des nouveaux réseaux seront ajoutés ; il paraît cependant peu probable que le ratio core/pan métabolisme diminue au niveau du ratio core/pan génome.

5.2.2 Diversité intra/inter espèces

Tous les processus métaboliques ne sont pas sujets aux mêmes contraintes. Cela est visible dans la composition du core et de la partie variable du métabolisme. Ainsi la proportion de processus impliquée dans la biosynthèse et l'énergie est beaucoup plus importante dans le core métabolisme. A l'inverse les processus de dégradation sont beaucoup plus représentés dans la partie variable du métabolisme. Cela peut se comprendre facilement par la nécessité de produire des métabolites essentiels pour la maintenance et la multiplication des cellules. Autour de ce core métabolique centré sur la synthèse et l'énergie viennent se greffer des voies de dégradation : certaines présentent sur tous les réseaux sont relativement anciennes, d'autres acquises par transfert horizontaux sont beaucoup plus récentes et se retrouvent uniquement sur un faible nombre de réseaux. Un autre élément impact cette diversité des processus de dégradation : le mode de vie de l'organisme. Ainsi, les *Shigella* qui sont des parasites,

possèdent des réseaux métaboliques plus petits, et surtout des pertes de voies de dégradation. Si ces pertes sont indépendantes d'un réseau à l'autre, elles touchent principalement les voies de dégradation. En effet l'utilisation des métabolites produits par l'hôte lève les contraintes sur des voies métaboliques qui dès lors vont disparaître au grès de la dérive génétique.
J'ai observé une différence inter-espèce puisque les *E. coli* sont regroupées dans le même sous-arbre métabolique. Si les différences métaboliques qui expliquent l'arbre métabolique sont bien réelles, elles restent largement surestimées étant donné le biais introduit par la stratégie ; les souches non *E. coli* ont des réseaux de moins bonne qualité.

5.2.3 Diversité et évolution

L'étude de la diversité métabolique sans apriori avec l'ACM a révélé un fort lien entre celle-ci et les groupes phylogénétiques. Et bien que nous ayons vu que les processus de dégradation sont les processus les plus variables, ceux sont les processus de biosynthèse notamment la synthèse des lipides qui sont en majorité responsables du lien avec la phylogénie. Cela peut paraître contradictoire à première vue, mais s'explique facilement ; si effectivement les voies de dégradations sont les moins conservées, elles sont aussi les plus éparses. A contrario, les voies de synthèses de lipides sont spécifiques d'un sous-ensemble de souches, ce qui donne un signal fort dans l'ACM. Des voies de dégradations sont tout de même présentes parmi les éléments responsables de la répartition des souches dans l'ACM : il s'agit des voies de dégradations des composés aromatiques. Les liens explicités par l'ACM ne suffisent pas à la déduction de l'histoire évolutive du métabolisme au regard de l'histoire évolutive décrite par la phylogénie. C'est pourquoi nous avons pratiqué d'autres analyses, qui montrent une corrélation entre la distance génétique et la distance métabolique. Pourtant l'arbre phylogénétique est calculé sur des gènes communs tandis que l'arbre métabolique est calculé en se basant sur les gènes différents. Cette corrélation rend possible l'utilisation du métabolisme comme un moyen de réduire la complexité génomique. En effet, au niveau génomique la majorité des éléments non-communs sont spécifiques d'un très faible nombre de souches et surtout ceux sont des IS ou prophages ; au niveau du métabolisme ces éléments sont filtrés, ce qui a pour effet d'augmenter le signal des éléments partagés par un nombre conséquent d'organisme.
Nous avons volontairement fait l'impasse sur les différences qui sont spécifiques d'un unique réseau pour deux raisons : la diversité étant sous-estimée et certains génomes n'étant pas finalisés, nous ne pouvons être sûr de ces différences, ensuite nous avons peu d'information sur les souches individuellement, alors que nous avons des renseignements sur des groupes de souches, comme leur groupe phylogénétique et leur type de pathogénicité.

5.2.4 Diversité et pathogénicité

E. coli revient régulièrement dans l'actualité à cause des différentes maladies que cette espèce peut entrainer. C'est donc tout naturellement que nous avons abordé la question du métabolisme et de la pathogénicité. Malheureusement en l'état actuel, nous n'avons pu établir de liens précis. Cela ne signifie pas qu'il n'existe pas d'élément métabolique responsable des maladies. Cependant les rares éléments mis en évidence sont soit putatifs, soit connus, mais avec des mécanismes d'actions inconnus. J'ai envisagé d'effectuer une analyse de type « arbre de régression » sur la

centaine de réseaux reconstruits. Cependant, le manque d'information sur la pathogénicité de certaines souches et les résultats obtenus dans les différentes études laissent supposer qu'au terme d'une longue et fastidieuse analyse, le meilleur résultat soit un ensemble de réactions putatives.

Le travail effectué sur la reconstruction métabolique est un travail précurseur, et plusieurs axes peuvent être envisagés pour continuer ce travail. Tout d'abord il semble pertinent de continuer l'analyse des réseaux et d'étudier la diversité pour essayer d'en déduire l'histoire évolutive des réseaux. Nous avons commencé l'exploration de l'histoire évolutive en reconstruisant les réseaux ancestraux par parcimonie ; les réactions apparaissent ou disparaissent au niveau de la racine ou alors au niveau des feuilles. Entre ces deux zones les réactions non communes ont la même probabilité de présence ou d'absence.

Un autre axe de développement peut être l'amélioration des connaissances, par un travail de curation sur un ou plusieurs réseaux, pour en faire des réseaux de références. Comme nous l'avons déjà dit, c'est un travail long et fastidieux, mais qui peut être utilisé à nouveau par notre processus de reconstruction.

Un autre moyen d'améliorer les connaissances peut être l'utilisation de méthodes informatiques et statistiques pour aider à l'identification de nouvelles réactions, ou bien proposer des gènes candidats aux réactions sans gène ; par exemple *CANOE (*A. Smith en révision*)*.

Il serait intéressant d'appliquer notre stratégie à une autre espèce*, E. coli* est la bactérie la plus étudiée ; il serait donc intéressant de comparer les résultats que j'ai obtenus avec ceux obtenus sur une autre espèce. Par exemple la base de données MicroScope contient plus d'une vingtaine de génomes séquencés d'*Acinetobacter* ainsi qu'un réseau de référence pour la souche *baylyi* sp1.

Parmi toutes les options qui se présentaient, j'ai choisi de prolonger l'étude du métabolisme à l'échelle de la bactérie, en passant par les modèles du métabolisme à base de contraintes. Le réseau métabolique peut être représenté par un graphe, celui-ci comme décrit dans l'introduction est un objet statique. Le passage aux modèles métaboliques nous permet d'estimer les différentes capacités métaboliques de nos souches ; capacité que l'on peut directement relier aux gènes et ainsi continuer sur notre objectif d'établir des liens entre phénotype et génotype par l'intermédiaire du métabolisme.

Chapitre II : Reconstruction et analyses des modèles à haut débit

La reconstruction d'un modèle métabolique à base de contraintes à l'échelle d'une bactérie nécessite des connaissances encore plus précises que celles requises pour la reconstruction des réseaux métaboliques. Chacun des éléments métaboliques (réactions, métabolites, coefficient stœchiométrique, compartimentation etc.) doit être identifié et surtout parfaitement défini.
Par exemple, la réaction catalysée par l'enzyme *alcool déshydrogénase* (numéro EC 1.1.1.1) qui transforme un alcool en aldéhyde, n'est pas une réaction suffisamment précise pour être intégrée à un modèle. Le problème vient de la classe métabolique alcool qui représente l'ensemble des métabolites qui contiennent le groupe fonctionnel *–OH* : éthanol, butanol, propanol etc. Les modèles requièrent d'expliciter chacun des alcools substrats et chacun des aldéhydes produits.
S'il existe des méthodes automatiques de reconstruction de plusieurs réseaux métaboliques en parallèle (Peter D Karp, S. Paley, et al. 2002), ou des méthodes de reconstruction de modèles au niveau de sous-processus métaboliques (DeJongh et al. 2007) il n'existe pas d'approche pour la reconstruction de modèles métaboliques à l'échelle de l'organisme. La conception d'un modèle métabolique repose généralement sur l'existence d'un réseau métabolique de bonne qualité. Avant ces travaux, il n'existait pas de méthodes pour reconstruire rapidement des réseaux métaboliques de bonne qualité ; il n'y avait donc pas de cadre favorable au développement des processus automatiques pour la reconstruction des modèles.

1 Reconstruction d'un modèle métabolique à l'échelle de la cellule

Avant d'expliquer comment j'ai reconstruit de manière semie-automatique et rapidement les modèles métaboliques globaux des 23 souches d'*E. coli* étudiées dans l'article (Chapitre I partie 1), je vais, dans un premier temps, revenir sur les différences entre un réseau et un modèle métabolique. Ces explications permettront une meilleure compréhension des différentes étapes et choix réalisés tout au long de cette partie. Je parle de reconstruction rapide et semie-automatique, car le prétraitement des données est actuellement impossible à automatiser et nécessite une longue expertise manuelle. Le processus de reconstruction est lui totalement automatisé. Pour rester cohérent avec la partie précédente, j'ai apporté un soin particulier à l'homogénéisation des données, ainsi qu'aux liens entre les différentes bases de données nécessaires à la création des modèles : le passage du modèle au génome en passant par le réseau métabolique, est transparent et sans ambigüité.

.1.1 Différences entre un réseau métabolique et un modèle métabolique

Les différences et le passage du réseau au modèle sont détaillés dans le chapitre d'introduction (le métabolisme : *in silico* partie 3.2.3); j'en rappelle ici simplement les points principaux. Si le réseau métabolique est un ensemble de réactions, de métabolites et de voies métaboliques dont la fonction est avant tout de centraliser les connaissances métaboliques à un instant donné, le modèle lui, est un objet

mathématique qu'il est possible de manipuler pour effectuer des simulations et des prédictions. Les modèles à bases de contraintes (CBMs) sont une catégorie de modèles mathématiques applicables à l'étude du métabolisme. Ils reposent sur l'évaluation des flux de matière qui traversent chacune des réactions sous des contraintes telles que la loi de conservation des masses, la disponibilité de certains métabolites ou encore le sens de conversion des métabolites au sein des réactions. La conservation de la matière est liée à la matrice stœchiométrique et par extension aux équations bilans : le nombre d'atomes consommés doit être égal au nombre d'atomes produits. Il est donc impératif d'avoir des réactions équilibrées. Cependant dans la plupart des bases de données métaboliques, dont MicroCyc (Vallenet et al. 2009) sur laquelle reposent mes réseaux, il existe des problèmes d'équilibres, les principaux étant l'état de protonation et la présence de molécules d'eau. Il faut également connaître la formule chimique de chacun des métabolites qui interviennent dans la réaction, et donc exclure tous les composés génériques dont la formule comprendrait un radical et des groupes fonctionnels (*R-OH* pour un alcool) ; à défaut il faut expliciter chacun des métabolites qui peuvent correspondre à la formule générique (éthanol, butanol, propanol etc.).
A partir des équations bilan des réactions, le cœur mathématique du modèle est créé : la matrice stœchiométrique.
Enfin il est obligatoire d'ajouter à l'ensemble des réactions biochimiques, des réactions artificielles afin de contrôler le modèle. Ces réactions d'échanges représentent des capacités d'assimilation et de production de l'organisme.
En complément de la matrice stœchiométrique, il est nécessaire de définir un ensemble de contraintes : la réversibilité des réactions, la fonction de biomasse, la fonction de maintenance énergétique, les limites de certains flux, etc. Ces contraintes appliquées à la matrice stœchiométrique vont définir le modèle sur lequel nous allons pouvoir effectuer des prédictions et des simulations.

.1.2 Similitudes et différences des processus de reconstruction des réseaux et des modèles

Il existe pour les réseaux métaboliques plusieurs bases de données, mais le manque de rigueur sur la stœchiométrie des réactions et les formules des métabolites, nécessite un prétraitement des données pour qu'elles soient utilisables au niveau des modèles. Il existe tout de même une base de données consacrée aux modèles globaux du métabolisme : BiGG (Schellenberger et al. 2010). Créée en même temps que les travaux de cette thèse, cette base de données n'est pas généraliste ; au contraire, elle contient uniquement les réactions et métabolites des modèles déjà reconstruits par leur équipe : *E. coli* K-12 MG1655, *S. cerevisiae* et *H. sapiens*. A cette base de données on peut ajouter les réactions issues du modèle d'*Acinetobacter baylyi*. Toutes les réactions des réseaux reconstruits qui ne sont pas dans ces quatre modèles doivent être vérifiées manuellement avant d'être intégrées aux nouveaux modèles, puisque, pour le moment, les différentes équipes qui reconstruisent des modèles ne s'occupent que d'un organisme à la fois et n'ont pas établi de lien entre les différents modèles et réseaux; à noter tout de même une approche récente de reconstruction automatisée des modèles globaux du métabolisme (Henry et al. 2010). Il existe cependant une méthodologie de reconstruction qui servira de base à mon processus ; cette méthodologie est expliquée dans le chapitre d'introduction (le métabolisme : *in silico* partie 3.2). La vocation initial d'un réseau n'est pas de prédire les capacités métaboliques de l'organisme étudié sur un milieu donné : il peut donc être incomplet

en omettant certains processus de synthèse ou de dégradation. Un modèle doit au contraire pouvoir produire tous les précurseurs de biomasse ou dégrader certains métabolites. La qualité d'un modèle se mesure, entre autre, au nombre de réactions isolées et au nombre de métabolites jamais produits ou jamais utilisés. Plus ceux-ci sont nombreux, plus la qualité du modèle est basse.
Pour pouvoir estimer les capacités du modèle il est nécessaire d'estimer la fonction de biomasse et de définir des environnements. Le modèle peut être reconstruit successivement au réseau métabolique ou en parallèle ; si pour *Acinetobacter* le réseau et le modèle sont issus d'une même reconstruction, assurant un passage de l'un à l'autre, le réseau et le modèle d'*E. coli* ont été réalisés indépendamment: dans ce cas le passage du réseau au modèle n'est pas trivial et surtout nécessite un important travail d'homogénéisation, qu'il est possible d'aider informatiquement mais qui ne peut être entièrement automatisé.

2 Nouvelle stratégie de reconstruction

Si pour la reconstruction des réseaux métaboliques, la stratégie décrite repose sur des logiciels existants (Peter D Karp, S. Paley, et al. 2002) pour les modèles, il a fallu entièrement dessiner le processus. Pour cela j'ai défini au préalable les objectifs et les spécifications des modèles. Fort des résultats obtenus avec la reconstruction des réseaux métaboliques, j'ai décidé d'utiliser une méthodologie similaire : repartir d'une base solide et fiable autour de laquelle j'ajouterai des nouveaux éléments (Figure 50). Cette base solide repose sur le modèle de référence d'*E. coli* : *iAF1260* (Feist et al. 2007); il correspond à la souche K-12 MG1655. Ce modèle a la même vocation qu'EcoCyc (Keseler et al. 2011) dans le processus de reconstruction des réseaux métaboliques : il est le pivot de la méthode. Sur ce canevas j'effectue deux opérations :

1) repérer les réactions et métabolites qui sont communs d'*iAF1260*.

2) intégrer les éléments qui n'existent pas dans ce pivot.

L'intégration des nouvelles données est cependant délicate ; en effet, il est indispensable de vérifier manuellement chacun des métabolites et des réactions (formule, stœchiométrie, etc.). Enfin il faut s'assurer de la cohérence entre les éléments extraits du pivot et ceux nouvellement créés.

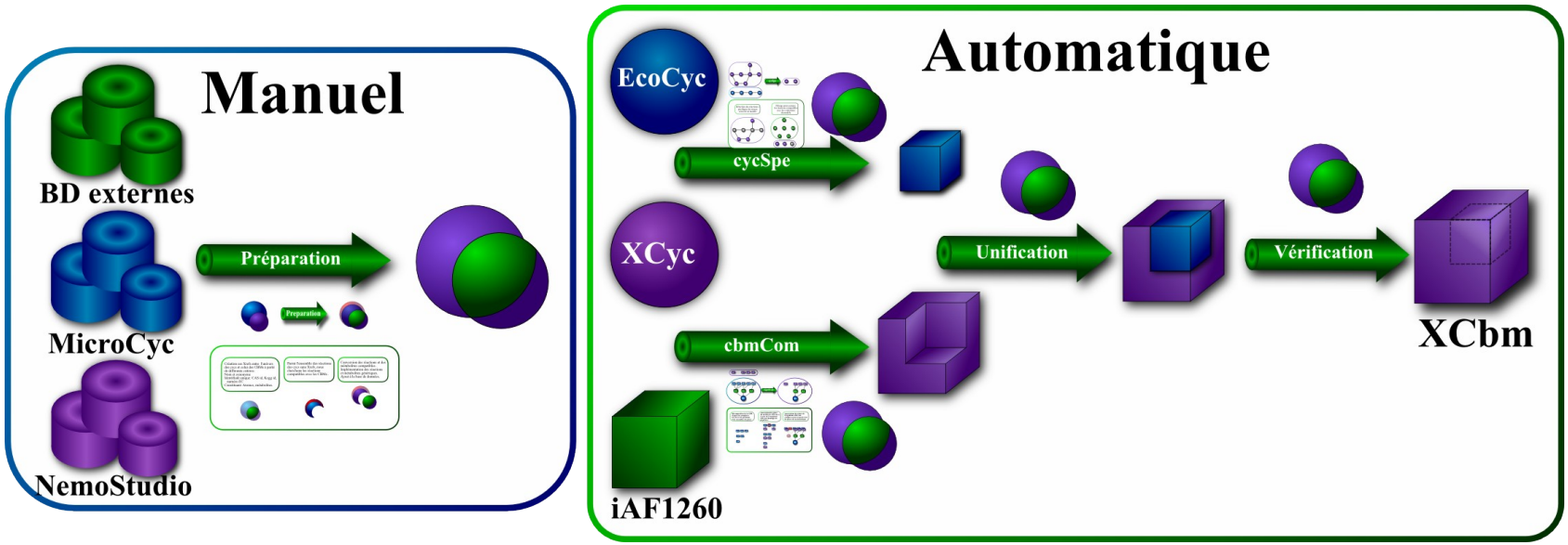


Figure 50: Processus de reconstruction des modèles du métabolisme à base de contraintes.
Le processus est composé d'une partie manuelle et une partie automatique. La partie manuelle prépare les données qui sont ensuite utilisées dans la partie automatique. Cette dernière infère les réactions et métabolites à partir d'un modèle pivot, des données préparées et du réseau à convertir.

.2.1 Objectifs globaux

Les objectifs de mes travaux sont doubles: d'une part, je souhaite mettre en place une nouvelle méthodologie pour tout ce qui touche au processus de reconstruction des modèles : l'automatisation, la continuité avec les plate-formes déjà existantes et la compatibilité avec les méthodes dédiées aux CBMs. D'autre part, je développe un axe scientifique qui lui se focalise sur le contenu des modèles, leur qualité, leur capacité à produire la biomasse ou à utiliser certains métabolites comme sources de carbone. Ce double objectif est donc dédié d'une part à une communauté qui cherche à reconstruire ses propres modèles et d'autre part, à une communauté qui s'intéresse au métabolisme des *E. coli* et qui utilisera les modèles produits pendant ma thèse.

2.1.1 Objectifs méthodologiques

NemoStudio (F Le Fèvre et al. 2009) est aux modèles ce que MicroCyc est au réseauxx (Vallenet et al. 2009), une plate-forme de visualisation et de manipulation de modèles métaboliques à l'échelle de la cellule. Développé en parallèle du modèle d'*Acinetobacter* (Maxime Durot et al. 2008), NemoStudio permet d'effectuer des simulations et des prédictions de croissances, ainsi que des comparaisons avec des données expérimentales. Les modèles reconstruits doivent être compatibles avec NemoStudio et sa base de données. Parmi eux, cinq vont être utilisés dans le cadre de l'ANR MetaColi. Ils vont servir à l'intégration de données hétérogènes, certaines issues d'expériences *in vivo*, d'autres de données *in silico* de modèles cinétiques. Pour faciliter ces échanges, les modèles doivent être compatibles avec les normes SBML (Hucka et al. 2003), qui est un langage de programmation spécifique de la biologie des systèmes. La méthodologie usuelle de reconstruction des CBMs s'effectue par des boucles de curation manuelle et de comparaison à des données expérimentales. Cette méthodologie n'est pas adaptée à notre situation ; aussi, étant donné le caractère novateur de mon approche, il est difficile de prévoir l'ensemble des difficultés techniques. Afin de faciliter la correction de ces difficultés, sans modifier l'ensemble du processus, celui-ci doit être modulaire. La difficulté des précédents travaux non automatisés pour obtenir un modèle fonctionnel laisse supposer qu'il faudra modifier et relancer plusieurs fois ce processus ; celui-ci doit donc être rapide. L'utilisation de différentes ressources et bases de données, aussi bien en entrée qu'en sortie du processus, peut être une source d'erreurs. Il est donc important que le passage d'une base à l'autre se fasse sans ambigüité. Par exemple, le terme *UDCPDP* désigne une réaction, tandis que le terme *udcpdp* désigne un métabolite : en cas de recherche, sans restriction sur les majuscules, il est possible d'associer un métabolite avec une réaction ou inversement, ce qui est bien sur, une erreur ; ces associations entre deux bases de données sont appelées références-croisées ou *Xref*. Enfin le processus doit être réutilisable sur d'autres organismes, et pour cela il doit être découplé des données initiales.

2.1.2 Objectifs scientifiques

Du point de vue scientifique, l'objectif principal est bien sûr la création de modèles à bases de contraintes fonctionnels pour les 23 souches d'*E. coli*. Par fonctionnel, je sous-entends des modèles capables de produire de la biomasse sur au moins quatre milieux définis dans la partie 4.1.2. En complément de ces milieux, les modèles doivent aussi pouvoir utiliser un maximum de sources de carbone assimilables *in vivo*. Parmi ces sources figurent des nouveaux métabolites qui ne sont pas présents dans

iAF1260. Sans atteindre la qualité des CBMs reconstruits manuellement, les modèles reconstruits automatiquement doivent se rapprocher de leurs résultats. A l'instar des réseaux, les modèles doivent présenter une certaine diversité, c'est à dire posséder suffisamment de différences et trouver une méthode pour les expliciter. Un autre objectif majeur est l'unification des données : comme je l'ai déjà évoqué dans l'introduction, il n'existe pas de dénomination unique pour les éléments biologiques. Cette lacune est à l'origine de mauvaises associations (faux positif) ; elle est également responsable de l'absence de certains éléments (faux négatif). Unifier toutes les ressources est impossible dans le cadre d'une thèse ; néanmoins, j'ai réalisé un important travail pour relier, par références croisées, les bases de données MicroScope & MicroCyc à la base de données NemoStudio. Non seulement cela facilite les comparaisons entre les différents niveaux d'études (génomes, réseaux, modèles), mais cela me permet aussi d'identifier toutes les réactions et métabolites des réseaux sans équivalent dans la base de données des modèles, éléments qu'il faudra convertir en métabolites et réactions compatibles avec les modèles.

.2.2 Conception d'un processus de reconstruction de modèles avec pivot

Lors de la conception du processus de reconstruction et de la définition des étapes, plusieurs alternatives se sont présentées. A chaque fois la solution choisie est celle qui permet de respecter le plus possible les objectifs fixés. La première décision importante est la séparation du prétraitement des données de la partie processus de reconstruction. Cette division rend le processus de reconstruction plus générique : c'est la partie de prétraitement des données qui sera à modifier en fonction des données disponibles. Le processus de reconstruction est divisé en quatre parties. Les deux premières parties sont indépendantes l'une de l'autre : elles permettent d'une part, de récupérer la diversité métabolique engendrée par les réseaux reconstruits et d'autre part, de récupérer les éléments homologues du pivot. Les deux autres parties servent à unifier, homogénéiser et vérifier les métabolites et réactions inférés lors des parties précédentes. La partie spécifique du modèle provient des réseaux reconstruits (ou cyc) : ce module est nommé *cycSpe* pour *cyc spécifique*. L'autre ensemble de réactions et de métabolites est issu du modèle à base de contraintes (ou CBM), *iAF1260*, et se nomme *cbmCom* pour *CBM commun*. Un ensemble de fonctions est dédié à la recherche et la récupération des réactions spontanées et des réactions artefactuelles d'*iAF1260* qui sont indispensables au bon fonctionnement du modèle. Un autre ensemble de fonctions vérifie l'intégralité des composants du modèle ; il s'assure des associations des GPRs en cas de redondance entre les éléments issus *cbmCom* et *cycSpe*. Il vérifie également si toutes les réactions, et ce, quelque soit le module d'inférence sont bien présentes dans le modèle reconstruit. Ces deux ensembles de fonctions sont regroupés au sein du troisième module d'unification.

.2.3 Module cbmCom

Le premier module que je vais détailler est celui qui concerne le pivot. A partir des gènes de la souche, et du modèle pivot ce module doit créer un sous-modèle qui contient l'ensemble des gènes, GPRs, réactions et métabolites localisés homologues du pivot. Il comprend une fonction principale : *gprEval.*
Il repose sur une hypothèse très forte mais justifiée. Je considère que le modèle de K-12 MG1655, *iAF1260*, et que le réseau de K-12 MG1655, EcoCyc, sont équivalents en contenu. Le modèle *iAF1260* est pourtant composé de 2082 réactions et le réseau

d'EcoCyc n'en compte que 1715. Cette différence de composition s'explique simplement par les contraintes de modélisation des CBMs, c'est à dire l'implémentation des réactions génériques et la suppression des réactions et des voies métaboliques qui sont isolées au sein du réseau. A cela, on peut ajouter le fait que le réseau et le modèle sont de très bonne qualité et représentent le même organisme : avec les connaissances actuelles, il est impossible d'obtenir une équivalence parfaite entre les deux.
J'ai modifié à plusieurs reprises le module *cbmCom* pour arriver à inférer l'ensemble des éléments d'*iAF1260* nécessaires à mes modèles. Dans un premier temps je me suis focalisé sur l'estimation des GPRs homologues et des réactions qui leur sont associées.
J'appelle sous-GPR chacun de ces groupes : ils correspondent à des isozymes. Autrement dit chacun des groupes code pour un complexe capable d'effectuer l'activité catalytique et donc la réaction. Le modèle de K-12 MG1655 est composé de 1944 GPRs dont 956 sont différentes. J'ai créé une fonction nommée *gprEval* pour *GPR évaluation*, dont le rôle est d'estimé à partir d'un génome cible et des GPRs du modèle pivot, les GPRs homologues du modèle cible.

Pour rappel, la GPR est une formule logique composée de gènes qui sont associés par des *et* ou des *ou*. Par exemple le transporteur du potassium a comme GPR *((b1250) (b1291, b3290, b1363) (b3747) (b1291, b3290, b3849))*. La GPR peut être décomposée en sous parties parenthèsées ; ainsi la GPR précédente peut être divisée en quatre parties (*b1250)* ou *(b1291* et *b3290* et *b1363)* ou *(b3747)* ou *(b1291* et *b3290* et *b3849)*.

La fonction *gprEval* est itérative et prend en entrée un modèle de référence et un organisme cible. Pour chaque réaction du modèle avec une GPR (Figure 51), trois étapes sont appliquées:
1) décomposition de la GPR en sous-GPRs.
2) Pour chaque gène de la sous-GPR, recherche de son homologue dans le génome cible.
3) Si tous les gènes de la sous-GPR ont un homologue, la fonction crée une sous-GPR homologue, puis elle joint toutes les sous-GPRs homologues dans une GPR homologue qui fera partie du nouveau modèle.
La recherche de gènes homologues se fait par l'intermédiaire des homologies de séquences et des *Best Bidirectional Hits*. Dans un processus indépendant il faudrait réaliser les alignements de séquences entre notre pivot et notre cible. Cependant ce processus est développé dans l'environnement de MicroScope qui contient en mémoire de nombreux résultats concernant les souches d'*E. coli* à reconstruire : ces alignements sont déjà réalisés et stockés dans la bases de données, ce qui rend le processus plus rapide. Malheureusement les gènes du modèle *iAF1260* ont comme identifiant des *bnumbers* et MicroScope utilise des *GO_id* : j'ai donc dû utiliser une table de correspondance entre le *bnumbers* d'*iAF1260* et les *GO_id* d'*E. coli* K-12MG1655. *GprEval* appliquée au génome de la souche K-12 MG1655, évalue positivement l'ensemble de 956 GPRs différentes et des 1944 réactions associées : l'ensemble des sous-GPRs d'*iAF1260* ont un homologue dans notre modèle reconstruit de K-12 MG1655.

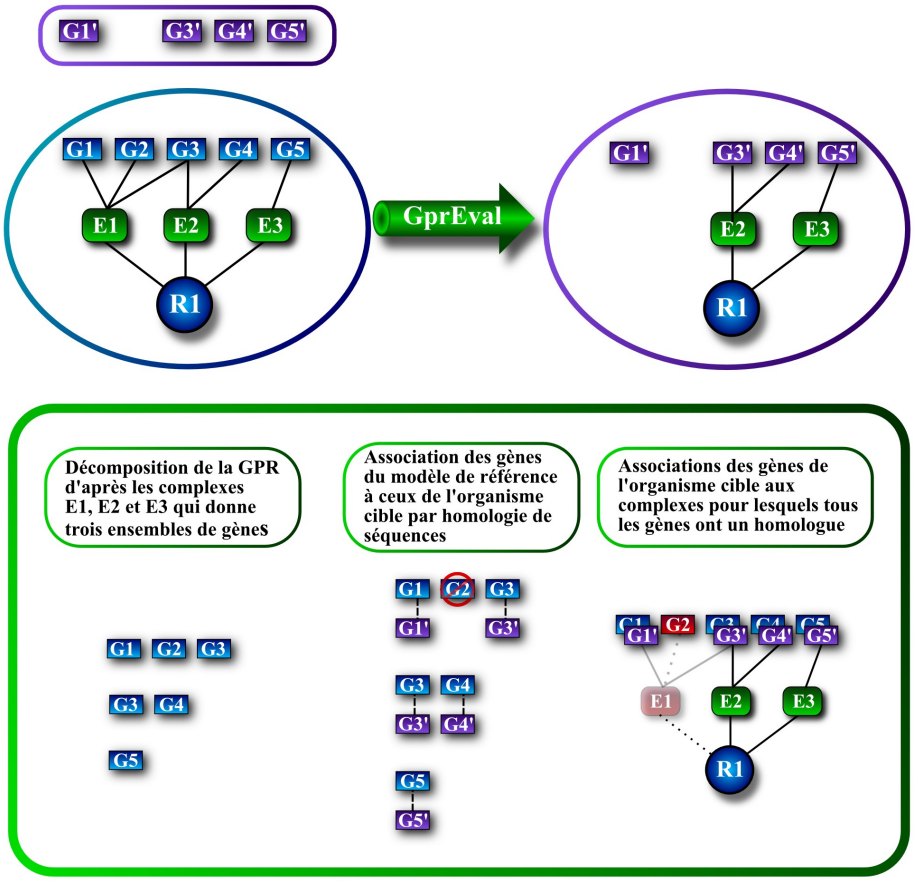


Figure 51 : Fonction *gprEval* d'évaluation des GPRs :
A partir du modèle pivot (ellipse bleu) et du génome de l'organisme cible (rectangle violet) *gprEval* infère les GPRs du nouveau modèle (ellipse violette). Dans un premier temps la GPR de la réaction R1 ((G1 et G2 et G3) ou (G3 et G4) ou G5) est décomposée en sous-ensembles (G1, G2, G3), (G3, G4) et le groupe G5. Pour chaque sous-ensemble, on regarde si tous les gènes ont un homologue dans le génome de la cible. Si c'est le cas une nouvelle GPR est créée dans le nouveau modèle associant les gènes homologues à la réaction. Si tous les gènes n'ont pas d'homologue alors la GPR reconstruite ne contiendra pas ce sous-ensemble : ((G3' et G4') ou G5').

Les modèles reconstruits uniquement avec *gprEval* ne sont pas capables de produire un flux de biomasse, et ce, même pour la souche K-12 MG1655. La raison est simple : d'une part on ignore les gènes codants pour une partie des réactions qui sont essentielles, d'autre part *gprEval* ne peut inférer les réactions artificielles qui représentent un manque de connaissance du métabolisme des *E. coli*. En excluant les réactions d'échange, *iAF1260* comprend 138 réactions sans gène ce qui représente 6.6% des réactions. J'ai décidé d'inclure ces réactions mais pas de manière systématique. J'ai choisi d'ajouter ces réactions si et seulement si l'ensemble des métabolites substrats est présent dans le modèle en reconstruction. Dans le cas des réactions réversibles, la réaction est ajoutée si l'ensemble des métabolites produits ou l'ensemble des métabolites substrats sont présents.
Il existe dans *iAF1260* une GPR artificielle qui relie des réactions au gène artificiel *s0001* ; ce gène et cette GPR indiquent que la réaction est spontanée. Il existe 29 réactions spontanées dans le modèle de référence. Ces réactions sont ajoutées suivant les mêmes critères que les réactions sans gène associé.
Bien que les conditions d'ajouts des réactions sans gène et des réactions spontanées soient identiques, j'ai conçu des fonctions différentes pour deux raisons : d'une part, je trace le type d'inférence, réaction avec GPR, spontanée ou sans gène, dans la base

de données, en sauvegardant le nom de la fonction qui a permis de récupérer la réaction. D'autre part, ne connaissant pas l'efficacité du processus a priori, j'ai préféré isoler chaque type de réactions inférées au sein de fonctions différentes pour en modifier le fonctionnement sans altérer celui des autres.
Initialement les fonctions d'inférences des réactions spontanées et sans gène étaient exécutées à la suite de la fonction *gprEval* ; cependant, elles ont été déplacées dans le module d'unification des réactions issues des modules *cycSpe* et *cbmCom*. L'ajout de ces réactions est dépendant du contenu en métabolites au moment de l'exécution de la fonction et certains des métabolites peuvent être absents de l'ensemble généré par la fonction *cbmCom*, mais présents dans l'ensemble issu de la fonction *cycSpe*.
Il existe cinq réactions sans gène qui sont systématiquement ajoutées aux modèles. Elles sont précédées du préfix *DM_* dans le modèle pivot (*DM_4HBA, DM_5DRIB, DM_AACALD, DM_HMFURN* et *DM_OXAM*). Ces réactions sont artificielles et ne représentent aucune réaction biochimique, cependant elles sont indispensables au bon fonctionnement du modèle, elles servent à empêcher l'accumulation des métabolites suivants : *4-hydroxy-benzoate, 5'-déoxyribose, aminoacétaldehyde, 4-hydroxy-5-méthyl-3(2H)-furanone* et *oxamate.*
Après l'ajout de ces fonctions, j'ai relancé la reconstruction du modèle de la souche K-12 W3110 qui est pratiquement identique à K-12 MG1655. J'ai obtenu un modèle fonctionnel qui est la copie d'*iAF1260* : il contient l'intégralité des réactions avec GPRs, spontanées, sans gènes et les réactions artificielles.
Dans le meilleur des cas, le module *cbmCom* peut produire une copie d'*iAF1260*, mais plus généralement les modèles reconstruits uniquement par ce module, seront un sous ensemble d'*iAF1260*. En effet le but de celui-ci consiste à récupérer les éléments communs entre le pivot et la cible. Pour compléter les modèles j'ai développé le module *cycSpe*, chargé d'inférer les éléments spécifiques de la cible.
Néanmoins, le nombre restreint de réactions et métabolites dans la base de données NemoStudio, ne permet pas de reconstruire des modèles avec la diversité des réseaux métaboliques estimée lors du chapitre précédent. Différentes difficultés font de cette étape le point le plus délicat et le plus coûteux à mettre en place : il faut sélectionner et convertir les métabolites et réactions des réseaux, utilisables dans les CBMs. Avant de détailler la façon dont j'améliore mes modèles, je vais revenir plus longuement sur l'unification des CBMs et des réseaux.

.2.4 Homogénéisation des données

La qualité des modèles dépend de la capacité du processus à intégrer les données provenant de différentes ressources. Il faut pour cela disposer d'un module dédié à l'homogénéisation et à la vérification des données. Dans notre étude, ce module définit des références-croisées (Xrefs) entre 3 bases de données de référence : Microscope (niveau génomique), MicroCyc (niveau du réseau métabolique) et NemoStudio (niveau du modèle métabolique). Les auteurs de la ressource métabolique MetaCyc et du modèle *iAF1260* ont écrit un papier commun sur l'unification des deux ressources (Feist et al. 2007). Plus de 80% des métabolites du modèle sont associés à ceux des réseaux, ce qui est très bon. Concernant les réactions, l'association s'est avérée beaucoup moins efficace et le chiffre précis n'a pas été évoqué dans l'article. Dans les données extraites des travaux d'unification, j'ai recensé 918 références-croisées de réactions : 44% des réactions du modèle ont une référence-croisée avec une réaction du réseau. Pour ne pas dupliquer des métabolites ou des réactions et ainsi engendrer des incohérences, une partie de mon travail a consisté à compléter et à curer les références-croisées. Les méthodes automatiques

ayant montré leurs limites, j'ai utilisé des méthodes semi-automatiques pour créer des listes de métabolites et de réactions candidats, et en déduire de nouvelles références-croisées. Que ce soit pour les métabolites ou les réactions, j'ai appliqué la même méthodologie : j'ai défini une liste de critères qui définissent un métabolite (une réaction) ; ces critères peuvent être spécifiques d'un métabolite (réaction), ou d'un nombre restreint de métabolites (réactions). Je compare ensuite les éléments issus du modèle métabolique d'*iAF1260* à ceux du réseau métabolique d'EcoCyc : je regroupe par paires les éléments qui partagent les mêmes critères. La création des Xrefs à partir des paires nécessite une expertise manuelle afin d'éliminer les faux positifs. Les comparaisons basées sur des critères différents peuvent conduire à définir plusieurs fois la même association entre un(e) métabolite (réaction) du réseau et un(e) métabolite (réaction) du modèle. C'est ce faisceau d'évidence qui nous permet d'être confiant quant à la véracité de cette référence croisée.

2.4.1 Homogénéisation des métabolites

Les métabolites peuvent être identifiés de différentes manières : le nom, les synonymes, la formule chimique qui peut prendre différentes formes (brute, développée etc.). S'il n'existe pas de lien direct entre les identifiants des réseaux au format cyc et les identifiants CBMs, ils existent des identifiants indirects : il s'agit du numéro CAS et de l'identifiant KEGG. Le numéro CAS est un identifiant unique pour les produits chimiques et biochimiques issu de la banque de données *Chemical Abstract Service*[2]. Sur les 1039 métabolites d'*iAF1260,* 370 (36%) possèdent un numéro CAS dont 354 uniques (34%). Cette différence est dûe à l'instanciation de métabolites dit génériques ; ces métabolites sont le plus souvent donnés sous la forme d'un radical et d'un ou plusieurs groupements fonctionnels. Leur incorporation dans le modèle nécessite d'expliciter chacun des radicaux. Je me suis basé sur tous les champs définis précédemment dans les fonctions d'associations de métabolites. Les résultats, suivant le critère utilisé, sont donnés dans Table 19.

Dans un premier temps j'ai associé les métabolites par leur nom, la condition est une correspondance exacte entre les noms ou entre le nom et un des synonymes. Cette méthode a permis la création de 637 Xrefs. J'ai ensuite utilisé les identifiants indirectes KEGG et numéro CAS, ce qui a rajouté 18 nouvelles Xrefs. Ce faible nombre n'est pas surprenant, car les éléments les plus connus et les plus étudiés sont également les éléments les mieux définis : noms homogénéisés et multiples références croisées. J'ai ensuite comparé les formules des 384 métabolites restants aux 5947 métabolites de MetaCyc pour lesquels il existe une formule. Suivant le pH, un métabolite peut être chargé ou neutre. Généralement la formule chimique du métabolite est la formule neutre, cependant pour éviter les erreurs dûes à l'état de protonation du métabolite, j'ai préféré ignorer le nombre d'atomes d'hydrogène. J'ai regroupé les métabolites en fonction de la formule sans hydrogène, ce qui donne 234 formules différentes. J'ai comparé les 234 groupes contenant des métabolites du CBM et de MetaCyc : il en résulte 57 nouvelles Xrefs. Pour augmenter le nombre de correspondances, j'ai finalement décidé de regarder manuellement les 327 métabolites restants. J'ai effectué des recherches par mots clés en décomposant les noms des métabolites. J'ai également utilisé une nouvelle ressource : ChEBI (Chemical Entities of Biological Interest). Parmi les métabolites sans Xref plus de la moitié (171), sont des instances; la décomposition des noms m'a permis d'associer une classe de métabolite de MetaCyc à plusieurs métabolites du CBM. Par exemple, j'ai découpé le

[2]http://www.cas.org/

nom *Phosphatidylglycerol* en *phophatidyl* et *glycerol* puis j'ai recherché dans la base de données MetaCyc les métabolites qui correspondaient aux deux termes. J'ai trouvé une classe de métabolites correspondante, et finalement la famille des *Phosphatidylglycerol (n-CX:Y)* des CBMs a été associée à la classe des *L-1-phosphatidyl-glycerol* de MetaCyc. J'ai réussi à associer des familles de MetaCyc à toutes les implémentations présentes dans le CBM à l'exception des muréines, ce qui donne 155 nouvelles Xrefs. A ce stade je suis parvenu à obtenir 877 métabolites sur 1039 possédant au moins une Xref ; ceci représente 84% des métabolites d'*iAF1260* et dépasse le pourcentage de correspondance des précédents travaux. J'ai appliqué la même méthodologie de découpage des noms en sous-mots pour les 167 métabolites restants et j'ai ainsi associé 105 nouveaux métabolites. Un total de 982 (90%) métabolites du modèle *iAF1260* ont donc une référence-croisée avec MetaCyc. Enfin j'ai pris soin de vérifier chacune des Xrefs.

	Métabolites associés aux Xrefs		
	Générés	Nouveaux	Sommes cumulée
Nom	637	637	637
Identifiant	577	18	655
Formule	234	57	712
Instance	171	155	877
Mot clé	162	105	982

Table 19 : Amélioration du nombre de correspondance entre les métabolites d'*iAF1260* et MetaCyc.

Les champs en gras représentent les critères pour lesquels il est possible d'automatiser complètement l'association. Au final 982 métabolites sur les 1039 ont pu être associés.

2.4.2 Homogénéisation des réactions

A l'instar des métabolites il a fallu identifier les correspondances entre les réactions de MetaCyc et celles d'*iAF1260*. Pour cela j'ai défini différents critères pour faciliter la réconciliation entre les 2082 réactions d'*iAF1260* et les 8703 réactions de MetaCyc en utilisant le nom des réactions, le numéro EC de l'enzyme associée à la réaction, les métabolites substrats et les métabolites produits. Si les noms et synonymes permettent une association sans trop d'ambigüité, le numéro EC et les métabolites de la réaction ne le permettent pas. Nous l'avons vu, certains numéros EC peuvent être incomplets (chapitre I 2.1), ou bien des réactions n'ont pas de numéro EC. Des métabolites, notamment en excès, tels que l'eau ou l'hydrogène, sont omis dans la description de certaines réactions. Avant même de commencer la création des correspondances, une première difficulté s'est présentée. EcoCyc n'est pas totalement inclus dans MetaCyc : il existe près de 400 réactions d'EcoCyc sans équivalent dans l'autre base. Pour les trois quarts, il s'agit de transporteurs ou de réactions de signalisation que j'ai ajouté à l'ensemble de réactions de MetaCyc afin d'établir les associations correspondantes. L'ensemble des informations de correspondance en fonction des différents critères est résumé à la fin de cette section dans la Table 22.

Le premier critère d'association est, une fois encore, la correspondance exacte entre les noms ou entre le nom et un des synonymes. A ma grande surprise, seulement 141 associations ont été établies, ce qui est très faible. J'ai utilisé les numéros EC (Table 20), et une nouvelle fois le résultat s'est avéré en deçà de mes attentes. Ainsi seulement 954 réactions d'*iAF1260* ont un numéro EC (moins de la moitié des réactions) ; pire encore, 104 numéros EC correspondent à 543 réactions différentes.

Par ailleurs, MetaCyc contient 6702 réactions associées à des numéros EC, parmi lesquels 4303 sont différents et 1306 sont incomplets.

	EC iAF1260	EC MetaCyc
Total	954	6702
Différents	522	4304
Incomplet	0	1306
Doublons	104	682

Table 20 Information sur les classes enzymatiques d'*iAF1260* et MetaCyc.
Si une grande partie des réactions de MetaCyc ont un numéro EC (77%), moins de la majorité des réactions d'*iAF1260* ont un numéro EC (46%)

Pratiquement tous les numéros EC (505) différents d'*iAF1260* ont une correspondance chez MetaCyc. Les différentes instances de réactions et les différentes mises à jour des numéros EC ont généré un nombre de paires de réactions largement supérieur aux attentes : 4430 paires. Au final, après sélection, j'ai obtenu 818 Xrefs parmi lesquels j'ai retrouvé 129 références-croisées précédentes. L'étape suivante fut la création de paires de réactions à partir des gènes. Parmi les réactions restantes 1158 réactions d'*iAF1260* sont associées à 759 gènes différents. Une nouvelle fois j'ai été confronté au problème d'unification des ressources. En effet les identifiants de gène d'*iAF1260* sont des *bnumbers* et ceux d'EcoCyc des *gnumbers*. Bien qu'il existe une table de correspondances entre ces identifiants au sein de la base MicroCyc, l'ajout d'un tel intermédiaire peut être une source d'erreur et une correspondance erronée peut engendrer une mauvaise référence croisée. A partir des *bnumbers* et de la table de correspondances, j'ai récupéré 721 réactions de MetaCyc candidates et j'ai généré 1379 paires de réactions potentielles. L'association gène/réaction du modèle, n'est pas un homologue de l'association gène/réaction du réseau. Si dans le modèle l'association comprend l'ensemble des gènes indispensables à la réaction, dont ceux responsable de la régulation, l'association dans le réseau se limite aux gènes codants les protéines des enzymes : il est impossible de se baser sur une comparaison stricte et automatique pour établir un lien entre une GPR du modèle et une association gène/réaction du réseau. Après expertise manuelle, un total de 321 nouvelles références-croisées ont été créées.
A ce stade, un ensemble de 1124 références-croisées est obtenu : ce nombre est déjà supérieur aux données des précédents travaux. Cependant, il est toujours trop faible pour assurer une bonne reconstruction des modèles.
Le dernier critère d'association est relatif au contenu en métabolite des réactions. Plus précisément il utilise la similitude des substrats d'une part et des produits d'autre part. Cette partie reprend le travail précédent d'association des métabolites. J'ai obtenu 512 correspondances exactes de substrats d'une part et produits d'autre part, parmi lesquelles 486 se sont avérées être de véritables références croisées. Les autres paires sont des faux positifs soit pour des raisons liées à la stœchiométrie (nombre de protons impliqués lors du transport de métabolites), soit pour des questions de réversibilité de la réaction. Ces nouvelles associations m'ont permis d'associer 131 nouvelles réactions. Les correspondances partielles des métabolites impliqués dans les réactions ont été séparées en huit catégories résumées dans Table 21.

Nombre de métabolites sans correspondance		Nombre total de paires	Nombre de Xrefs
iAF1260	MetaCyc		
0	1	4	0
0	>1	2	0
1	0	375	339
1	1	69	26
1	>1	409	3
>1	0	2084	16
>1	1	1407	25
>1	>1	2535	100

Table 21 : nombre total de paires de réaction associées et nombre de paires valides en fonction du nombre de métabolites associés.

La première colonne donne le nombre de métabolites dans la réaction du modèle sans association avec un métabolite de la réaction du réseau. La deuxième colonne donne le nombre de métabolite de la réaction du réseau sans association avec un métabolite dans la réaction du modèle. « >1 » indique qu'il existe plus d'un métabolite non associé.

Enfin, j'ai regardé manuellement les 764 réactions restantes ; pour cela j'ai utilisé d'autres ressources elles que les bases de données KEGG, Enzymes, Brenda. J'ai décomposé les noms des réactions en mot clés pour effectuer des recherches sur les réactions utilisant des métabolites non ubiquitaires (c'est à dire intervenant dans peu de réactions). Cette analyse manuelle, longue et fastidieuse, a permis l'identification de 470 nouvelles références-croisées.

Après ces différentes étapes, j'ai créé des références-croisées pour 80% des réactions d'*iAF1260*, soit 2 fois plus que dans l'article original des travaux précédents (Feist et al. 2007). Si un bon taux de recouvrement entre MetaCyc et *iAF1260* est obtenu, les références-croisées doivent cependant être mises à jour à chaque modification de MetaCyc (ou EcoCyc) et du modèle. En effet, certaines réactions possédant le plus souvent un identifiant générique tels que *RXN-** ou *RXN0-**, sont mises à jour et se voient attribuer un nouvel identifiant définitif. L'actualisation de ces réactions est relativement simple car il suffit de modifier l'identifiant MetaCyc de la référence. Dans le cas de nouvelles réactions, il faut refaire les processus d'association que nous venons de voir.

L'intégralité des références-croisées générées est intégrée à la base de données NemoStudio (F Le Fèvre et al. 2009), et est utilisée dans différents modules du processus de reconstruction des modèles.

		Xrefs				Réactions associées aux Xrefs		
		Générées	Validées	Nouvelles	Somme cumulée	Générées	Nouvelles	Somme cumulée
Nom		179	141	141	141	141	141	141
Numéro EC		4430	796	677	818	769	640	803
Gene		1379	352	352	1170	321	321	1124
CorrespondanceMétabolite	totale	512	486	242	1322	480	131	1252
	partielle	6885	509	114	1436	475	66	1318
	Recherche Manuelle	470	470	470	1906	351	351	1669

Table 22 Résumé des Xrefs et réactions associées créées suivant le critère de sélection.
La colonne *Générées* représente l'ensemble des associations, faux positifs inclus. La colonne *Validées* est le nombre de Xrefs. La colonne *Nouvelles* donnent le nombre d'Xrefs uniques (une Xref peut être récupérée par différents critères). La colonne *sommes cumulées* donne le nombre total d'éléments uniques après application des différents critères.

.2.5 Module de pré-traitement

Pour automatiser le processus de reconstruction, j'ai choisi d'extérioriser l'intégration de toutes les données provenant des analyses manuelles. Deux types de données doivent être préparés : les références-croisées mentionnées dans la section précédente, et la conversion des réactions et des métabolites des réseaux qui n'existent pas dans NemoStudio (Figure 52). Il est important de mettre à jour les références-croisées en premier puisqu'elles sont utilisées pour associer les réactions des réseaux et repérer les réactions à convertir. La fonction *XrefUpdater* s'occupe de la mise à jour et de la création des nouvelles Xrefs. Elle prend comme argument quatre variables : l'identifiant NemoStudio, l'identifiant du réseau, le nom de la base de données (MicroCyc), et un commentaire : ce dernier a pour but de conserver l'origine de la référence-croisée. L'autre fonction *newReaction* prend en entrée une liste de réactions MetaCyc à convertir en réaction CBM. Pour chaque réaction de la liste, je vérifie si elle n'est pas présente dans une Xref. Si c'est le cas, je vais extraire de MicroCyc différentes informations : les métabolites substrats, les métabolites produits, les coefficients stœchiométriques, la réversibilité, la localisation et le nom de la réaction. Pour chaque métabolite je regarde s'il est impliqué dans une référence-croisée. Si ce n'est pas le cas, je crée un nouveau métabolite. Pour cela je récupère dans MicroCyc le nom et la formule du métabolite et je l'insère dans la base de données NemoStudio. Je vérifie ensuite la présence du métabolite dans le bon compartiment pour la réaction. Si le couple métabolite/compartiment n'existe pas dans la base de données, alors j'ajoute ce couple à la base. Enfin, toujours dans un souci de traçabilité, un champ commentaire est rempli avec la mention « *créé pour MetaColi »* que ce soit pour la création d'un nouveau métabolite ou d'une nouvelle réaction.

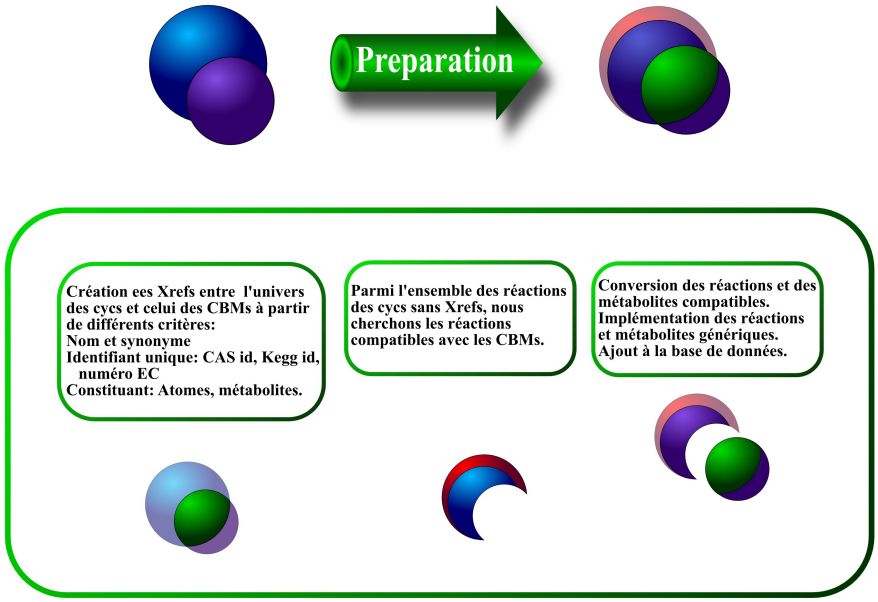


Figure 52 : Module de préparation des données.
Les données d'origines différentes conduisent à une expertise manuelle en deux temps : 1) Détecter les éléments équivalents entre les réseaux et les CBMs. 2) Parmi les éléments des réseaux sans équivalent, filtrer ceux qui peuvent être convertis.

Que ce soit pour la création des métabolites ou des réactions dans la base NemoStudio, une référence-croisée est ajoutée vers MicroCyc. De plus nous utilisons l'identifiant MicroCyc comme identifiant pour NemoStudio afin de ne pas complexifier les données finales.

.2.6 Module cycSpe

Le deuxième module, *cycSpe* a pour rôle d'identifier les réactions et les métabolites des réseaux métaboliques à convertir et intégrer au niveau des modèles métaboliques. Il prend en entrée le réseau du modèle en reconstruction et le réseau pivot, et donne en sortie un sous-modèle avec les réactions spécifiques du modèle en reconstruction.
Ce module est celui qui a subi le plus de modifications entre les premiers essais de reconstruction et la version finale. Comme je l'ai expliqué dans la partie 2.3, nous considérons qu'EcoCyc est l'équivalent du modèle *iAF1260*. Par conséquent l'information supplémentaire qu'il est possible d'introduire dans les différents modèles se trouve dans le *pan* métabolisme et plus précisément parmi les éléments absents d'EcoCyc. Ces métabolites et réactions n'existent pas dans la base de données NemoStudio : il faut donc les intégrer en s'assurant que les éléments soient compatibles avec les prérequis du modèle. Dans un premier temps, la recherche des références-croisées et la création des réactions et des métabolites ont été réalisées progressivement (i .e. reconstruction par reconstruction). Cependant la même réaction était évaluée pour chaque modèle : j'ai donc décidé d'extérioriser la fonction d'évaluation et de conversion des réactions, comme expliqué dans la section précédente. Dans un deuxième temps, j'ai essayé de convertir l'ensemble des 526 métabolites et 375 réactions du pan métabolisme qui n'apparaissent pas dans EcoCyc. Rapidement, il s'est avéré que la conversion automatique était impossible et ce, pour plusieurs raisons. En particulier, je me suis trouvé dans l'impossibilité de définir certains métabolites, c'est à dire qu'il m'a été impossible d'expliciter une instance ou une formule à celui-ci. C'est par exemple le cas de la classe métabolite *[protein] N^6-(dihydrolipoyl)lysine* : je n'ai trouvé aucune instance de métabolite correspondante.

De même pour les réactions, certaines étaient redondantes avec mon travail d'homogénéisation ; d'autres étaient des réactions impliquant des classes de métabolites, ou dans le pire des cas, des réactions dont les métabolites étaient non spécifiés. J'ai redéfini la façon de procéder, en ajoutant des critères pour sélectionner les métabolites admissibles : ils ne doivent pas posséder de références-croisées vers NemoStudio et ils doivent obligatoirement être des instances de métabolites (par opposition aux classes de métabolites tel que *[protein] N^6-(dihydrolipoyl)lysine*, ou un alcool*)*. Pour cela, j'ai défini un dictionnaire qui relie les classes métaboliques aux instances des métabolites et remplacé chaque classe par ses instances dans la liste finale des métabolites à convertir. Puis j'ai filtré cette liste et finalement j'ai vérifié manuellement chacun des 281 métabolites. En ce qui concerne les réactions j'ai appliqué un filtre sur le contenu en métabolites : seules les réactions dont tous les métabolites se trouvent soit dans les références-croisées, soit dans la liste des nouveaux métabolites à ajouter sont conservées. Un traitement spécial est appliqué aux réactions utilisant des classes de métabolites : dans ce cas on crée autant de réactions qu'il existe de paires d'instances. On travaille par paire d'instance puisque, à un métabolite substrat instancié correspond un unique métabolite produit par la réaction. Au final j'ai récupéré 180 nouvelles réactions et 9 réactions génériques qui ont fourni 14 réactions instanciées.
La création des nouvelles réactions a lieu dans le module de pré-traitement. De ce fait, on considère que toutes les réactions et tous les métabolites sans références-croisées à ce niveau de la reconstruction sont incompatibles avec le formalisme des modèles.
Pour chaque réaction inférée je vais regarder si elle est reliée à des gènes. Si c'est le cas, je crée une GPR et j'ajoute l'étiquette « *avec gène »* à la réaction. Si le complexe enzymatique est inconnu, la GPR associée à la réaction sera composée d'un ensemble de gènes séparés par des *ou*. S'il n'existe pas de gènes associés à cette réaction, je regarde si celle-ci est spontanée, et dans ce cas je l'étiquette avec le statut « *spontanée »*. Enfin si la réaction n'a ni gène et n'est pas spontanée, je la marque de l'étiquette « *inconnue »*.

L'algorithme final de *cycSpe* est schématisé dans la Figure 53 : pour chaque réseau métabolique, on regarde la liste des réactions absentes du réseau pivot. Dans cette liste, les réactions génériques sont substituées par leurs implémentations, puis pour chaque réaction, on regarde l'existence d'une référence-croisée vers NemoStudio. Si elle existe, la réaction est ajoutée à la liste des réactions du sous-modèle. Une fois la liste de nouvelles réactions constituée, une seconde étape est chargée d'ajouter les réactions au modèle. Pour chacune d'elles, on regarde son statut : associée à des gènes, spontanée ou inconnue. On regarde aussi les métabolites qui interviennent dans la réaction, plus précisément les couples (métabolites, compartiments) ; à chaque fois qu'un nouveau couple est identifié, il est ajouté au modèle.

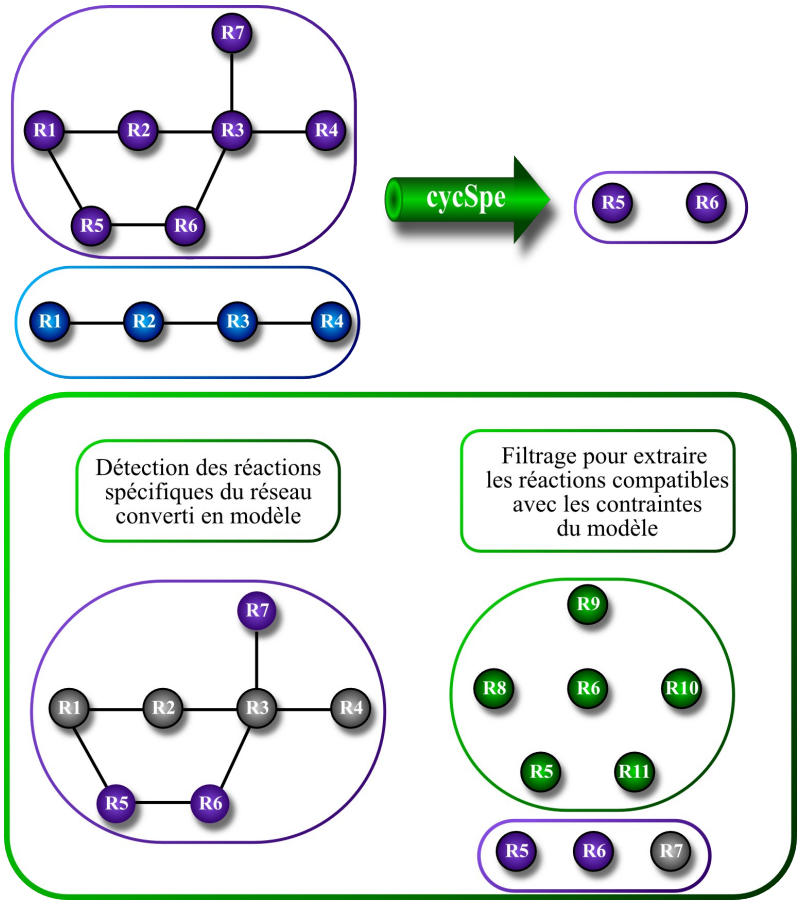


Figure 53 : Le module *cycSpe* compare le réseau pivot et le réseau de la cible pour en déduire les réactions spécifiques de ce dernier.

Le réseau à convertir (ellipses violettes) est comparé au réseau de référence (ellipses bleues). Un premier ensemble de réactions est déduit (R5, R6 et R7). Cet ensemble est comparé à la liste des réactions admissibles (ellipse verte) pour récupérer les réactions du sous modèle (R5 et R6).

En sortie, le module *cycSpe* produit un sous-modèle avec les réactions spécifiques de la cible ; couplé au sous-modèle produit par le module *cbmCom*, le résultat est un modèle très proche du modèle final. Malheureusement, l'utilisation de deux ressources différentes engendre des inconsistances, d'où la nécessité de rechercher et corriger ces erreurs.

.2.7 Module d'unification

Une fois les deux sous-modèles constitués, il faut compléter l'ensemble de réactions. A ce stade de la reconstruction, les réactions spontanées n'ont pas été prises en compte, de même que certaines réactions sans gène associé. Pour combler ces lacunes, j'établis la liste des métabolites présents dans le modèle cible. Pour chacune des réactions spontanées ou inconnues d'*iAF1260*, je regarde si tous les substrats sont présents dans la liste et dans ce cas, la réaction est créée dans le modèle avec une étiquette *« spontanée »* ou *« inconnue »* suivant l'étiquette de la réaction dans le pivot. Pour les réactions réversibles, si l'ensemble des substrats ou l'ensemble des produits est présent dans la liste alors la réaction est récupérée.
J'ai également mis en place une fonction pour forcer l'ajout de réactions à partir d'une liste spécifique ; j'en donnerai un exemple ultérieurement.

Dans la base de données NemoStudio les réactions sont indépendantes du modèle. C'est par l'intermédiaire de l'activation que le lien réaction/modèle est créé.
Il existe trois sortes d'activations :

- Par gène, lorsqu'une GPR est disponible
- Spontanée, si la réaction ne nécessite pas de catalyseur

Les sous-modèles issus de *cbmCom* et de *cycSpe* ne peuvent pas être simplement fusionnés. La comparaison des GPRs des réactions communes au modèle et au réseau a montré des incohérences. J'ai conçu une fonction pour les détecter, pour une réaction donnée, elle compare le type d'activation du modèle pivot à celui du modèle reconstruit
La Table 23 recense les différents cas d'incohérences possibles:

			iAF1260		
			GPR	Spontanée	Inconnue
Cyc	GPR	Réaction	+*		
		Activation	+gène*	+gène	+gène -inconnue
	Spontanée	Réaction	+		
		Activation	+spontanée		+spontanée
	Inconnue	Réaction	+		
		Activation	+inconnue	+spontanée -inconnue	

Table 23 : Tables des incohérences entre les sous-modèles issus de *cycSpe* et *cbmCom*.
Le symbole « + » indique l'ajout. Le symbole « -» la suppression de l'activation dont le nom est inscrit derrière.
Le symbole « * » indique que le cas des GPRs est particulier : lorsque la GPR du réseau contient des gènes supplémentaires, ils sont ajoutés ; dans le cas ou la réaction n'existait pas dans *cbmCom* elle est ajoutée.

Parmi les neuf cas possibles, sept peuvent entraîner des modifications des activations. Le cas le plus fréquent d'incohérence est une différence entre la GPR du pivot et du modèle cible. Deux raisons peuvent expliquer cette différence :

1. La GPR du modèle cible est évaluée positivement par le module *cbmCom* et elle contient un ou quelques gènes sans homologue dans la GPR du pivot. Ces gènes proviennent du complexe enzymatique dans le réseau cible et ils sont alorsajoutés à la GPR du modèle cible.

2. La GPR du modèle cible est différente de la GPR du modèle pivot, et ne provient pas du module *cbmCom*. La GPR du modèle cible est créee à partir des associations gènes/réactions du réseau cible.

Deux autres cas d'incohérences impliquent des réactions du réseau avec association gène/réaction :

1. L'activation de la réaction est étiquetée *spontanée* dans le pivot : la réaction se voit affecter d'une activation *par gène* en plus de l'activation *spontanée*.

2. Dans le cas d'une activation *inconnue* dans le pivot: celle-ci est substituée par une activation *par gène*.

Les réactions sans gène uniquement chez *iAF1260* ne peuvent être inférées par les modules *cycSpe* ou *cbmCom,* mais peuvent être associées par références croisées dans le module d'unification. Quand une telle situation se présente, j'ajoute la réaction au nouveau modèle et je lui associe le type activation qui convient (*inconnue* ou *spontanée*).
Les deux derniers cas concernent des ambigüités entre des activations *spontanées* et des activations *inconnues*. Dans les deux cas je privilégie l'activation *spontanée* qui se substitue à l'activation *inconnue*.
Enfin certaines réactions d'*iAF1260* sont des réactions artificielles paliant un manque de connaissance. Elles sont indispensables au bon fonctionnement du modèle, mais ne correspondent à aucun des critères précédents. Elles sont récupérées par une fonction spéciale qui ajoute les réactions passées en argument sans aucune vérification. Cette fonction doit être utilisée avec parcimonie et toute réaction ajoutée doit reposée sur une obligation fonctionnelle ou sur des hypothèse solides (par exemple des preuves expérimentales).
Après ces vérifications, j'ajoute un dernier type de réactions : les réactions artificielles d'échanges ou flux d'échanges. Ces flux modélisent l'apparition ou la dilution des métabolites dans l'environnement. Par conséquence pour chaque métabolite pour lequel il existe un transporteur vers l'environnement, je crée un flux d'échange. La liste des métabolites externes est rapidement obtenue, puisque NemoStudio comprend l'entité métabolite localisée : il suffit d'appliquer un filtre sur le compartiment.
Suite à toutes ces opérations le processus délivre des modèles qui doivent au moins produire de la biomasse sur milieu riche et sur milieu minimum glucose.

.2.8 Implémentation et utilisation

Contrairement au processus de reconstruction des réseaux, celui des modèles ne peut être totalement automatisé puisqu'une analyse manuelle des nouvelles données et une homogénéisation des ressources sont indispensables. Cette nécessité rend caduque tout intérêt au lancement systématique du processus de reconstruction après la mise à jour d'une des ressources : il sera donc exécuté sur demande.
L'implémentation est réalisée en java pour rester cohérent avec les autres outils de reconstruction et les programmes développés au sein de l'équipe. L'origine hétérogène des données et les différents modules m'ont conduit à découper le programme en trois familles de packages : le premier concerne l'extraction des données, le second les algorithmes de reconstruction, et le dernier l'exportation des données.
La reconstruction des modèles est extrêmement sensible aux données : j'ai donc choisi d'isoler tout celles qui concerne l'extraction et la préparation des données. De plus, en prévision d'évolutions possibles, j'ai créé un package pour chacune des sources de données, ce qui facilite l'ajout des nouvelles sources qui peuvent voir le jour.
Les packages concernant le cœur du processus sont au nombre de trois et peuvent être exécutés indépendamment. Le premier package concerne le module *cbmCom*, le second le module *cycSpe,* et le dernier est le module d'unification, chargé de la vérification et l'union des sous-modèles. L'exportation des données est gérée par le dernier groupe de package qui comprend un package d'ajout des éléments à la base de données NemoStudio, un package de suppression d'éléments de cette même base et un package qui permet l'exportation au format SBML. J'ai reconstruit les modèles sur un ordinateur portable macbook pro équipé d'un processeur double cœur à 2,53GHz et de 4Go de mémoire RAM. Le déroulement du processus complet, comprenant la

mise à jour des références-croisées, la création des nouvelles réactions, la reconstruction de 23 modèles et l'exportation des données au format SBML et dans NemoStudio, s'exécute en 24h.
La disponibilité des modèles au sein de la plate-forme NemoStudio est illustrée dans l'Annexe 5.

.2.9 Préparation des modèles et lien avec le modèle cinétique

Un des éléments importants des modèles à base de contraintes est la fonction de biomasse qui représente la consommation de métabolites par l'organisme pour son maintient et sa croissance. C'est elle qui sera optimisée lors des simulations notamment des FBAs dont le principe est expliqué dans le chapitre d'introduction sur la modélisation partie 4.1 ; sa composition est donc cruciale et doit être au plus proche de la réalité. Dans l'introduction nous avons vu qu'il est impossible d'estimer la quantité de tous les métabolites présents dans un organisme. Néanmoins il est possible d'estimer la masse sèche des principaux types de métabolites de la cellule ainsi que les composés élémentaires (les acides aminés pour les protéines ou les nucléotides pour la molécule d'ADN). La fonction de biomasse d'*iAF1260* a été conçue d'après la masse sèche de la souche de *E. coli* B/r (Table 24).

Composition de la biomasse de la souche *E. coli* B/r

Protéine [20] (55.0%)		
L-alanine	L-arginine	L-asparagine
L-aspartate	L-cystéine	L-glutamine
L-glutamate	glycine	L-histidine
L-isoleucine	L-leucine	L-lysine
L-methionine	L-phenylalanine	L-proline
L-sérine	L-threonine	L-tryptophan
L-tyrosine	L-valine	
RNA [4] (20.5%)		
ATP	CTP	GTP
UTP		
DNA [4] (3.1%)		
dATP	dCTP	dGTP
dTTP		
Inorganic ions [13] (1.0%)		
ammonium	calcium	chlorine
cobalt	cuivre	fer
magnésium	manganèse	molybdate
phosphore	potassium	sulfate
zinc		

Lipide [6] (9.1%)		
structure		
phosphatidylethanolamine	Phosphatidylglycerol	cardiolipin
acyl longueur : lien non saturé		
16:00	16:01	18:01
LPS (3.4%) [1]		
Core KDO2 lipide A		
Cofacteurs, Prosthétique Groupes et autres [10] (<2.9%)		
S-adenosylmethionine	FAD	coenzyme A
thiamine diphosphate	riboflavine	undecaprenyl pyrophosphate
pyridoxal 5′-phosphateb	folates	quinones
chorismate	enterobactine	glutathione
spermidine	vitamine B12	NAD(P) hèmes
putrescine		
Muréine[4] (2.5%)		
structure		
muréine disaccharide		
Longueur de chaine peptidique		
pentapéptide	tétrapéptide	tripéptide
Glycogène [1] (2.5%)		
glycogène		

Table 24 : Principaux types de métabolites et métabolites élémentaires de la souche *E. coli* B/r. En gras les types de métabolites, entre crochet le nombre de métabolites élémentaires et entre parenthèses le pourcentage de la masse sèche. Ces données sont issues de (Feist et al. 2007).

Ces données ont été complétées et modifiées notamment grâce aux données d'essentialité pour obtenir la fonction de biomasse finale. Cette dernière repose sur l'hypothèse que les souches K-12 MG1655 et B/r sont suffisamment proches pour que les métabolites élémentaires de la biomasse soient identiques. Avec les données d'essentialités sur *iAF1260* les auteurs du modèle ont défini une fonction de biomasse *core*, qui correspond au plus petit ensemble de métabolites élémentaires nécessaires. Composée de soixante trois métabolites, cette fonction *core* est secondée par une fonction de maintien de l'énergie qui représente la consommation en ATP.
J'ai considéré que les souches de mon étude sont suffisamment proches de *E. coli* B/r pour pouvoir utiliser la même fonction de biomasse et la même fonction de maintenance énergétique dans les modèles reconstruits.
En effet, bien que la fonction de *biomasse core* soit initialement conçue par des délétions, je suppose que cette fonction reste valide dans le cas d'ajout de nouveaux métabolites dans le modèle. De plus cette fonction a été réutilisée avec succès par d'autres équipes dans les modèles de *Salmonella.*
J'ai choisi d'exporter les modèles reconstruits au format SBML et de les stocker dans la base de données de NemoStudio. Le format SBML est actuellement le format le plus abouti pour la biologie des systèmes mais aussi le plus généraliste. Il est aussi bien adapté aux modèles cinétiques du métabolisme qu'aux modèles à base de contraintes. Comme il s'agit d'un langage à balises (Chapitre d'introduction partie 4.6.1), il est possible d'y ajouter tous les éléments jugés pertinents ; en particulier l'ensemble des références-croisées qui ont été expertisées manuellement.
Les modèles vont être en interaction avec des modèles cinétiques du métabolisme central (Chassagnole et al. 2002) et patrie 2.2 de l'introduction sur le métabolisme *in silico*), et ils vont aussi recevoir des contraintes issues de données de protéomique. Le langage SBML est particulièrement bien adapté pour fournir un support contenant l'ensemble des données, en ajoutant par exemple les *km* et les *vmax* des différentes réactions du modèle cinétique, et en associant les concentrations relatives des différentes protéines aux réactions correspondants à ces enzymes.
Il existe une bibliothèque SBML pour la plupart des langages de programmation : il est ainsi aisé de passer du programme codé en JAVA à des outils d'analyses dont la COBRA-Toolbox sous Matlab, ou le jumelage avec le modèle cinétique sous R : jumelage pour lequel j'ai dû associer les identifiants du modèle cinétique et du modèle à base de contraintes.
Sur les trente réactions que comprend le modèle cinétique, 6 sont des réactions d'échanges artificielles et ne peuvent être associées à des réactions du CBM.
Parmi les 24 réactions biochimiques, 3 réactions pouvaient porter à confusion lors de l'association :

- Pour deux réactions du modèle cinétique, il existe plusieurs réactions du CBM utilisant les mêmes métabolites.
- Pour la troisième réaction, elle effectue en une étape les modifications de deux réactions du CBM.

Dans le premier cas l'étude de la réversibilité et du sens de la réaction nous a permis de lever l'ambigüité. Dans le second, nous avons associé la réaction du modèle cinétique aux deux réactions du CBM.

3 Matériel et méthodes

Toutes les études des modèles sont réalisées sur un macbook pro double cœur (2.53Ghz) et 4 Go de mémoire ram sous OS X 10.6, pour des raisons de compatibilité

de bibliothèques ; tous les logiciels et les bibliothèques sont compilés en versions 32bit.
Le logiciel principal est Matlab[3] version 2009b puis 2010a.
Les modèles sont tous disponibles à travers le format SBML (Hucka et al. 2003). La bibliothèque libsbml[4] (en version 2.4, puis 4 et enfin 4.1) nous permet de les manipuler (import, export, tests de consistance). L'utilisation des versions successives est là pour s'assurer de la compatibilité en cas de nouvelle installation.
Le solveur utilisé lors des optimisations est *cplex*[5] version 12, il est relié à Matlab par la bibliothèque *cplex* version 12.2.
Le chargement et l'optimisation des modèles sont gérés par la COBRA-Toolbox en version 13.5 (Becker et al. 2007). Une version modifiée de la fonction de chargement a été développée afin de prendre en argument le modèle au format SBML et un modèle de référence ; elle permet d'ajouter directement les réactions de biomasse et de maintenance énergétique du modèle de référence au modèle chargé.
L'étude de la composition des modèles utilise 3 fonctions. La première, *findCommonMet* prend en argument deux modèles et retourne trois listes : l'ensemble des métabolites communs, l'ensemble des métabolites spécifiques du premier modèle et l'ensemble des métabolites spécifiques du deuxième modèle. La deuxième fonction est la version dédiée aux réactions : *findCommoRxn*. La dernière, *findCommonCell* prend en argument 2 listes (de métabolites ou de réactions) et donnent la liste des éléments communs et les listes des éléments spécifiques.
La recherche de connectivité du métabolisme à partir de la matrice stœchiométrique est effectuée par la fonction *findMetabolicGraphConnectedComponents* du package matlab suiteSparse[6] La variabilité des flux était initialement calculée par la méthode issue de (R Mahadevan & C H Schilling 2003) puis par la fonction *fastFVA* (Gudmundsson & Thiele 2010). La comparaison des variabilités de flux est réalisée par trois fonctions créées à cet effet : *similarityFluxes, compareModelFVA* et *compareMediaFVA*.
La première calcule le score de similitude entre deux résultats de *fastFVa*. La seconde utilise *similarityFluxes* pour comparer la variabilité des flux entre deux modèles sur un même milieu. La dernière compare la variabilité des flux pour un même modèle sur deux milieux différents.
L'ensemble des calculs sont lancées à partir de deux routines, l'une consacrée aux différences de constitutions des modèles, l'autre consacrée à la variabilité des flux.

4 Analyses des modèles reconstruits

Une fois le processus mis au point et implémenté, j'ai reconstruit les modèles à bases de contraintes des 23. *E. coli* utilisées pour l'analyse de la diversité métabolique (chapitre I partie 1). Avant d'évoquer en détail les différents modèles reconstruits, je vais étudier les différences entre le modèle pivot *iAF1260* et son équivalent produit par notre processus *K-12 MG16555Cbm*.

[3]Disponible à l'adresse suivante : http://www.mathworks.fr/
[4]Disponible à l'adresse suivante : http://sbml.org/Main_Page/
[5]Disponible à l'adresse suivante : http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/
[6]Disponible à l'adresse suivante : http://www.cise.ufl.edu/research/sparse/SuiteSparse/

.4.1 Les différences entre *iAF1260* et K-12MG1655Cbm

4.1.1 Différences de composition

La validation du processus de reconstruction par pivot est faite à partir de l'étude des différences entre le modèle de référence de K-12 MG1655 : *AF1260* et notre modèle reconstruit. La reconstruction de ce modèle est particulière puisqu'elle ne peut bénéficier du module *cycSpe*.
Le module *cbmCom* récupère les 1944 réactions du pivot associées à des GPRs. Les 29 réactions spontanées et les 108 réactions sans GPR sont également récupérées grâce aux références-croisées dans le module d'unification. Le modèle reconstruit est très proche de celui d'origine, néanmoins il existe quelques différences qu'il nous faut étudier pour estimer les biais de reconstruction de notre processus.
Si l'ensemble des réactions avec GPRs est retrouvé, 169 présentent des inconsistances entre les GPRs reconstruites et celles du modèle d'origine (8.6% de l'ensemble des GPRs). Ces incohérences proviennent de l'étape de vérification des GPRs entre les associations gènes/réactions du réseau et les GPRs du modèle reconstruit (Table 25). Dans 83% des cas, l'incohérence de la GPR est dûe à un manque de précision dans la base de données MicroCyc. La correction consiste à ajouter un ou plusieurs gènes dans la GPR. La localisation de la réaction est l'information la plus fréquemment absente : elle représente près de la moitié des cas d'incohérences (47% Table 25). Ceci crée des ambigüités entre la réaction qui a lieu dans le cytosol et son équivalent dans le périplasme (dont le nom comporte par convention le suffixe *pp* dans *iAF1260*). Si dans le modèle ces deux réactions sont distinctes et associées à des gènes différents, dans le réseau il n'existe qu'une seule réaction non localisée, à laquelle sont associés tous les gènes (Figure 54).

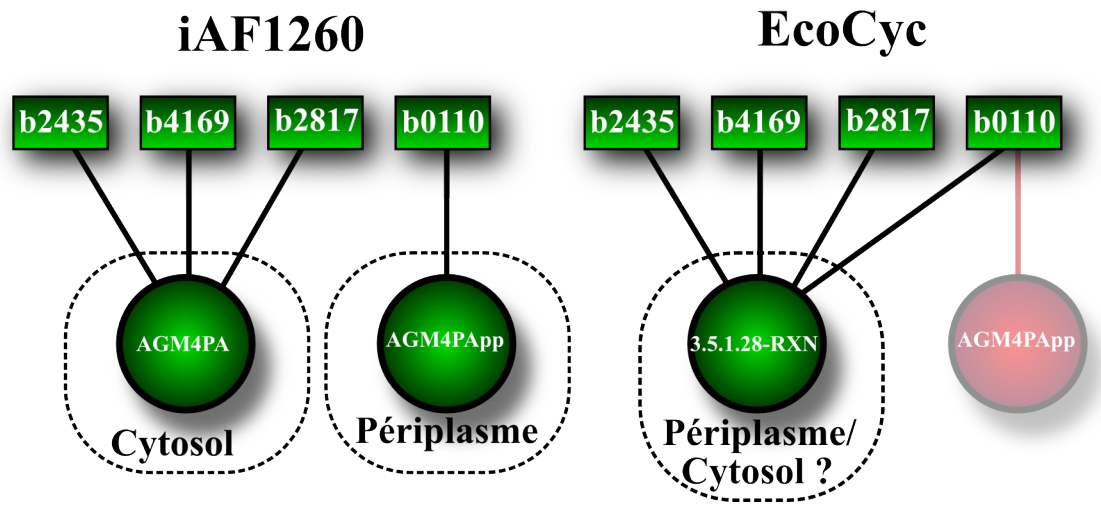


Figure 54 : Différences de GPR pour une même réaction entre *iAF1260* (modèle) et EcoCyc (réseau).
A gauche il existe une réaction par compartiment : celle qui a lieu dans le cytosol est encodée sur trois gènes (b2435, b4169 et b2817), et celle qui a lieu dans le périplasme est encodée par un gène (*bnumber* : b0110). A droite la même réaction dans EcoCyc : les quatre gènes sont associés à la même réaction.

Puisque ce genre d'incohérences est facilement identifiable et puisque nous n'avons pas d'information sur les liens gènes/réactions ⇔ localisations, nous conservons l'ensemble des gènes de la GPR (partie droite de la Figure 54).

Type d'erreur		Nombre de réactions	Pourcentage	
Gènes en plus		23	14%	
Réaction indissociable		5	3%	
Erreur de Xref		1	1%	
Erreur de Xref (EC)		1	1%	
Erreur de Xref (IC)		2	1%	
Générique		47	28%	
Générique et Gènes en plus		2	1%	
Localisation	Non générique	42	25%	47%
	Générique	38	22%	
Putative		8	5%	

Table 25 : Liste des incohérences des GPRs entre *iAF1260* et K-12 MG1655Cbm.
Xref (EC) signifie que l'erreur de Xref provient de numéro EC identique. Xref (IC) est une erreur de Xref dans les associations de la littérature. En gras les inconsistances causées par le manque d'information sur la localisation de la réaction dans EcoCyc. En italique les inconsistances dues aux implémentations des réactions génériques.

L'autre moitié des incohérences concerne des gènes supplémentaires dans la GPR issue des réseaux ; il en existe trois catégories distinctes.
Tout d'abord, le processus de reconstruction des réseaux métabolique peut inférer des réactions putatives par homologie de séquence sans que le gène en question soit le *« Best Bidirectional Hit »* BBH du gène dans l'oganisme pivot. Ces gènes donnent une nouvelle sous-GPR spécifique des modèles reconstruits : c'est le cas pour 4% des GPRs inconsistantes. Ne pouvant infirmer/confirmer cette sous-GPR sans preuve expérimentale, elle est conservée dans les différents modèles reconstruits.
La deuxième catégorie, qui représente 51% des cas (28%, 23% et 1 % dans la Table 25), concerne l'implémentation des réactions génériques d'EcoCyc. Celles-ci peuvent être instanciées en spécifiant les métabolites qu'elles transforment. Généralement l'ensemble des gènes des réactions instanciées est associé à la réaction générique. Or une partie des gènes de la GPRs peut concerner la spécificité au substrat. Néanmoins, si on pose l'hypothèse qu'une réaction instanciée peut, en réalité, utiliser différents substrats, alors il semble cohérent de conserver tous les gènes de la réaction générique dans les instances.
Enfin la dernière catégorie, qui représente plus de 14% des cas, concerne de nouvelles associations de gènes. La réaction d'EcoCyc est équivalente à la réaction d'*iAF1260* et bien qu'une partie des gènes soit commune, il existe dans le réseau des associations de gènes qui ne sont pas présentes dans le modèle. En l'absence de preuve *in-vivo*, notamment avec des mutants, je ne peux écarter aucun de ces gènes.

Cette comparaison m'a permis de relever trois erreurs d'associations entre les réactions d'EcoCyc et le modèle d'*iAF1260*. Une des erreurs est le résultat d'un numéro EC erroné. Dans *iAF1260* la *D-alanine-D-alanine dipeptidase* est associée au numéro EC 3.4.17.14 : ce numéro est normalement associé à la *D-alanyl-D-alanine carboxypeptidase* ; la référence-croisée a donc été corrigée.
Une autre erreur d'association de données déjà présentes dans NemoStudio concerne les réactions *RBFSa* et *RBFSb*. Il existe un lien entre RBFSa associée à la *riboflavin synthase*, et RBFSb associée à la *6,7-dimethyl-8-ribityllumazine synthase.* Si chez *E. coli* les deux réactions sont bien distinctes, dans le réseau des levures les 2 réactions sont codées par les mêmes gènes, d'où l'association incorrecte dans NemoStudio. Pour corriger ce genre d'incohérence, nous avons mis en place la possibilité de choisir

les ressources pour estimer des références-croisées. Elle peut être globale sur l'ensemble de MicroCyc ou local sur un réseau uniquement (par exemple EcoCyc).
La dernière erreur est une association erronée dont on ignore l'origine mais qui est, vraisemblablement, manuelle. Elle a été corrigée.
Le faible nombre d'erreurs d'association entre des réactions d'*iAF1260* et de MicroCyc met en évidence le soin apporté à l'expertise manuelle et à l'homogénéisation des données : sur plus de 1900 références-croisées créées, seulement 3 se sont avérées erronées. Ces références-croisées sont la base de la reconstruction et surtout le moyen de passer facilement du réseau au modèle ; il était donc très important d'avoir le minimum d'erreurs.
Il y a, en termes de réactions et de métabolites, très peu de différences entre *iAF1260* et K-12MG1655Cbm ; cependant notre objectif n'est pas l'amélioration du modèle initial, et à l'aide de la souche K-12 W3110, clone métabolique de la souche K-12 MG1655, nous verrons qurl est le gain potentiel engendré par notre stratégie de reconstruction.
On notera tout de même que le modèle de K-12 MG1655 reconstruit par nos soins, comporte une voie de dégradation secondaire pour le *galactose*. Elle utilise comme métabolite intermédiaire le *D-galactono-1,4-lactose*. Cette voie est composée de deux transporteurs et de deux réactions, qui permettent le passage du *D-galactono-1,4-lactose* du milieu externe au cytosol en passant par le périplasme. La première réaction convertit le *β-D-Galactose* en *D-galactono-1,4-lactose et* la seconde transforme ce dernier en *D-Galactonate*. Celui-ci est ensuite transformé en plusieurs étapes en *pyruvate* qui est un des métabolites les plus utilisés par l'organisme.

4.1.2 Définition des milieux et premières simulations

Avant d'effectuer différentes simulations sur le modèle reconstruit et de comparer les résultats obtenus avec ceux du modèle pivot, il est nécessaire de définir des contraintes, portant sur les environnements de croissances (milieux) et la fonction objective.
Comme développé au cours l'introduction dans la partie consacrée aux CBMs (partie 2.4), la principale contrainte est la conservation de la matière qui se traduit par :
Le produit de la matrice stœchiométrique par le vecteur des flux solutions de l'équation est nul soit $S.v=0$.
J'ai ensuite défini les contraintes sur les bornes des flux. Comme je ne dispose pas d'information précise sur les différents flux, j'ai choisi comme borne maximale (*bmax*) une valeur arbitraire largement supérieure aux valeurs des flux d'entrés,c'est à dire 1000. J'ai pris comme borne minimale (*bmin*) son opposée : -1000. Les flux d'entrés qui représentent les métabolites disponibles seront de l'ordre de la dizaine. Pour chacun des flux je vais associer une borne maximale et une borne minimale ; si la borne maximale est dans tous les cas *bmax*, la borne minimale dépend de la réversibilité de la réaction asscié au flux.
Dans le cas où la réaction est réversible, le flux sera compris entre *bmin* et *bmax.* Par convention, on considère que lorsque le flux est positif la réaction se déroule dans le sens de l'équation bilan c'est-à-dire de la gauche vers la droite ; lorsque le flux est négatif la réaction se produira dans l'autre sens. Dans le cas des réactions irréversibles, la réaction ne peut se dérouler que dans un sens : la borne minimale associée est nulle.

Un seul flux possède par défaut une borne minimale strictement positive : il s'agit de celui passant par la réaction de maintenance énergétique *ATPM*. La valeur estimée provient directement du modèle de référence et s'appuie sur différents résultats expérimentaux : elle est fixée à 8.9, toujours en unité arbitraire (Feist et al. 2007).
Il est ensuite important de définir les différents environnements, parmi lesquels on distingue les milieux riches des milieux minimums. Dans le cadre de modélisation des CBMs, on considère comme milieu riche un environnement où l'ensemble des métabolites importés ou exportés par l'organisme est présent. Par opposition, le milieu minimum sera un milieu contenant un ensemble minimum de métabolites permettant à la cellule de se développer, c'est à dire avoir un flux de biomasse non nul. Ces milieux sont composés de deux types de métabolites. Ceux indispensables et présents dans tous les milieux minimums : calcium, chlore, dioxyde de carbone, cobalt, cuivre, fer (fe^{2+} et fe^{3+}), eau, hydrogène, potassium, magnésium, manganèse, molybdate, sodium, phosphate, tungstène et zinc.
Les autres sont variables et comprennent une source de carbone, une source d'azote, une source de souffre et une source d'oxygène ; les sources par défaut sont données dans la Table 26.

Source	Métabolite par défaut
Carbone	Glucose
Azote	Ammonium
Souffre	Sulfate
Oxygène	Dioxygène

Table 26 : Métabolites par défaut du milieu minimum.
A ces sources variables viennent s'ajouter le calcium, le chlore, le dioxyde de carbone, le cobalt, le cuivre, le fer (fe2+ et fe3+), l'eau, l'hydrogène, le potassium, le magnésium, le manganèse, le molybdate, le sodium, le phosphate, le tungstène et le zinc.

Par convention on parlera d'un milieu minimum suivi du nom de la source particulière : par exemple le milieu minimum gluconate est un milieu contenant les métabolites indispensables et comme source de carbone le *gluconate* ; les autres sources variables sont celles par défaut (Table 26).
Les différents travaux sont réalisés sur quatre milieux, deux simples (les milieux minimums) et deux complexes (milieu riche et milieu riche sans sucre). Deux milieux sont choisis car ils représentent les conditions usuelles en laboratoire : le milieu riche et le milieu minimum glucose. Les deux autres milieux sont choisis d'après certaines caractéristiques de l'espèce *E. coli* (chapitre d'introduction sur les *E. coli*) : il s'agit du milieu minimum gluconate, l'une des sources de carbone préférentielles de cette espèce. L'autre est le milieu riche sans sucre, qui essaie de se rapprocher de l'urine, milieu impossible à modéliser de par sa grande variabilité.
Du point de vue de la modélisation, chaque métabolite de l'environnement est représenté par un flux d'échange :

- Un flux d'échange nul représente l'absence du métabolite dans le milieu.
- Un flux d'échange négatif signifie la présence du métabolite dans le milieu.

A partir de ces deux règles, nous pouvons définir les 4 milieux :

- Milieu riche : Tous les flux d'échanges sont négatifs.
- Milieu riche sans sucre : Tous les flux d'échanges des métabolites qui ne sont pas des sucres sont négatifs ; les flux d'échanges des métabolites de type sucre sont nuls.

- Milieu minimum glucose : La source de carbone est le glucose, l'ensemble des flux d'échanges des métabolites indispensables et des sources variables sont négatifs ; le reste des flux d'échanges est nul.

- Milieu minimum gluconate : La source de carbone est le gluconate et l'ensemble des flux d'échange des métabolites indispensables et des sources variables est négatif ; le reste des flux d'échange est nul.

La fonction objective est la fonction de biomasse du modèle pivot, elle est compatible avec la totalité des modèles reconstruits puisque chaque modèle comprend l'intégralité des métabolites précurseurs de la biomasse.

Avec les différents éléments définis ci-dessus, j'ai effectué les premières simulations sur le modèle reconstruit afin de tester sa capacité à produire de la biomasse dans différentes conditions. Pour cela j'ai effectué des analyses du type FBA (cf. chapitre d'introduction sur le métabolisme *in silico* partie 4.1). Ces simulations m'ont servi à mettre en place un cadre de modélisation en assurant la cohérence entre les contraintes, les environnements et la fonction objective.
Le premier milieu testé est le milieu riche. Grâce à l'attention et au soin porté à la création et mise en place du processus de reconstruction, notre modèle de K-12 MG1655 est non seulement capable de produire de la biomasse, mais en plus son flux de biomasse est légèrement supérieur à celui d'origine (28.6643 contre 28.6190). Cette augmentation est due aux réactions supplémentaires de notre modèle dont le flux n'est pas nul ; ceci montre que la voie secondaire de dégradation du *galactose* rajoutée est fonctionnelle et qu'elle contribue à la production de biomasse. Le milieu riche, de par son grand nombre de métabolites disponibles est le milieu le plus aisé pour obtenir de la biomasse puisqu'il présente le moins de contrainte au final. Les *E. coli* sont capables de pousser avec comme seule source de carbone le *glucose*. J'ai réalisé une FBA sur milieu minimum glucose. Cette fois-ci les flux de biomasse des deux modèles sont identiques avec une valeur de 0.9293. Ce résultat, bien que trente fois inférieur à celui sur milieu riche, montre que le modèle *K-12 MG16555Cbm* est aussi capable de pousser sur milieu minimum glucose. Nous détaillerons plus en détail les différentes capacités de croissance dans le prochain chapitre.

4.1.3 Comparaison de la variabilité des flux : Définition d'un score de similarité

Le modèle reconstruit étant fonctionnel, j'ai pratiqué une analyse de la variabilité des flux ou FVA (*Flux Variability Analysis,* voir le chapitre d'introduction sur le métabolisme *in silico* partie 4.3) afin d'étudier les bornes des différents flux. Cette analyse permet d'explorer l'espace des solutions des simulations, espace réduit à un vecteur de flux unique dans les analyses FBA.
Sur milieu riche la FBA appliquée au modèle *K-12 MG1655Cbm* propose une solution pour laquelle 775 réactions ont un flux non nul (772 pour *iAF1260*). La FVA apporte un supplément d'information puisque sur milieu riche, on dénombre 920 (901) réactions qui peuvent avoir un flux non nul, et 509 (506) réactions dont le flux est toujours différent de zéro. On constate que le faible nombre de réactions ajoutées entraîne des modifications sur une vingtaine de réactions entre les deux modèles.
Sur le milieu minimum glucose on dénombre autant de réactions (419) dans les deux modèles pour lesquelles les flux peuvent être non nuls.
J'ai naturellement voulu comparer la variabilité des flux entre mon modèle et celui de référence ; cependant il n'existe aucune situation similaire dans le domaine des

CBMs, et par conséquent, aucune façon de comparer cette variabilité. J'ai donc dû créer un cadre de comparaison de la variabilité des flux.
La comparaison des intervalles de flux doit prendre plusieurs éléments en compte (Figure 55), et en premier lieu le sens du flux. Si les flux des réactions irréversibles sont nuls ou positifs, pour les réactions réversibles ces flux peuvent, dans certains, cas passer d'une borne inférieure négative à une borne supérieur positive, ce qui traduit un changement de direction de la réaction. Il faut aussi tenir compte des plages de valeurs des intervalles qui doivent être suffisamment disjoint pour en déduire une différence ; c'est le cas lorsque l'un des deux est nul, ou bienqu'il y a un changement de sens du flux uniquement dans l'une des deux réactions. Dans d'autres cas la similitude ou la différence est plus délicate à évaluer, d'où la mise au point d'une méthode de calcul de similarité, qui consiste à estimer un voisinage aux bornes des intervalles des flux.
Soit deux flux v_1 et v_2 avec comme intervalles respectifs I_1 et I_2, et soit b_{min1} et b_{max1} les bornes de I_1 et b_{min2} et b_{max2} les bornes de I_2. Soit un intervalle Ib_{min1} centré sur b_{min1} et un autre intervalle Ib_{max1}, centré autour de b_{max1}.

$$
\begin{array}{l}
Si \\
(b_{min1} \geq 0 \;\&\; b_{min2} \geq 0) \;\&\; (b_{max1} > 0 \parallel b_{max2} > 0) \;\&\; (b_{min2} \in Ib_{min1}) \;\&\; (b_{max2} \in Ib_{max1}) \\
(b_{min1} < 0 \;\&\; b_{min2} < 0) \;\&\; (b_{max1} \leq 0 \parallel b_{max2} \leq 0) \;\&\; (b_{min2} \in Ib_{min1}) \;\&\; (b_{max2} \in Ib_{max1}) \\
(b_{min1} < 0 \;\&\; b_{min2} < 0) \;\&\; (b_{max1} > 0 \;\&\; b_{max2} > 0) \;\&\; (b_{min2} \in Ib_{min1}) \;\&\; (b_{max2} \in Ib_{max1}) \\
alors \\
I_1 \text{ et } I_2 \; sont \; similaires
\end{array}
$$

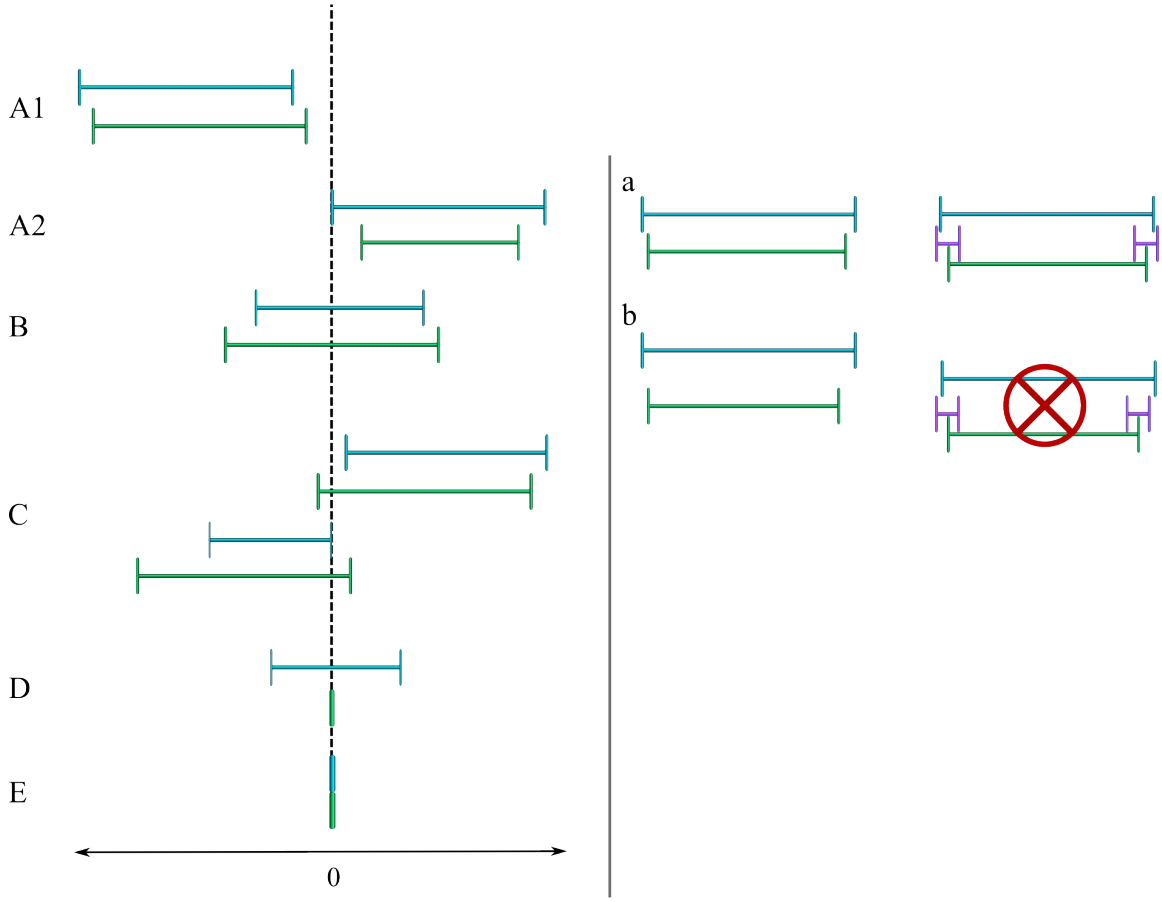


Figure 55 : Comparaisons des intervalles de variation d'un même flux entre deux conditions de simulation différentes.

A gauche sont présentés les différents cas possibles. A) les intervalles des deux flux sont du même signe, A1) tous deux positifs (zéro inclus) ou A2) tous deux négatifs (zéro inclus). B) Les deux intervalles ont leurs bornes inférieures strictement négatives et leurs bornes supérieures strictement positives. C) L'un des intervalles à une borne négative et l'autre positive tandis que l'autre intervalle a ses bornes de même signe. D) L'un des deux intervalles est nul. E) Les deux intervalles sont nuls. Dans les cas C et D, je considère automatiquement que les flux sont différents. Le cas E peut être considéré comme une similitude, mais les réactions n'intervenant pas dans la simulation, j'ai préféré les exclure. Les cas A et B sont ambigüe tels quels. A droite est représentée la méthode pour lever l'ambigüité des cas A et B. Pour chacune des bornes de l'intervalle vert, on calcule un nouvel intervalle violet ; celui-ci est centré sur la borne et dépendra de la longueur de l'intervalle vert, ainsi que d'un seuil. Dans le cas a) les bornes de l'intervalle bleu sont comprises dans les intervalles violets : on considère alors que l'intervalle vert et le bleu sont similaires. Dans le cas b) seule la borne inférieure de l'intervalle bleu est comprise dans l'intervalle violet : les deux intervalles sont considérés comme différents.

La grande différence des valeurs admissibles d'un même flux entre deux simulations rend caduque l'utilisation d'une valeur fixe à soustraire ou à additionner aux bornes, que ce soit pour un même modèle sur deux milieux différents (un facteur 30 sur le flux de biomasse entre le milieu riche et le milieu minimum glucose), ou deux modèles différents sur un même milieu. J'ai donc opté pour une valeur qui dépend de la longueur de l'intervalle, ainsi qu'un coefficient d'extension des bornes de l'intervalle.

$$n_1 = b_{min1} - b_{max1}$$
$$Ib_{min1} = [b_{min1} - \alpha.n_1; b_{min1} + \alpha.n_1]$$
$$Ib_{max1} = [b_{max1} - \alpha.n_1; b_{max1} + \alpha.n_1]$$

$$n_2 = b_{min2} - b_{max2}$$
$$Ib_{min2} = [b_{min2} - \alpha.n_2; b_{min2} + \alpha.n_2]$$
$$Ib_{max2} = [b_{max2} - \alpha.n_2; b_{max2} + \alpha.n_2]$$

La comparaison de deux modèles pose une difficulté supplémentaire puisqu'ils ne possèdent pas les mêmes réactions. J'ai décidé d'exclure de la comparaison les réactions spécifiques d'un des modèles. De même, j'exclue les réactions dont le flux est nul dans les deux conditions. Lors de la première série d'analyses une limitation technique est venue perturber les comparaisons. Les approximations engendrées par le solveur numérique lors du calcul de la variabilité des flux, notamment au voisinage de zéro, produit des valeurs positives ou négatives de l'ordre de 10^{-4} et moins. Bien que ces approximations soient minimes, elles peuvent faire passer les intervalles du cas A, au cas B ou C (Figure 55). Pour limiter les artefacts produits par l'optimisation, j'ai choisi un seuil au delà duquel on considère le flux comme nul.
J'ai défini comme score de similarité, le nombre d'intervalles similaires divisé par le nombre d'intervalles dont au moins un des deux intervalles est non nul.
L'utilisation d'un ratio rend le score comparable indépendamment du nombre de flux mis en jeu, nombre qui est beaucoup plus important sur milieu riche que sur milieu minimum.

$$score = \frac{nombre\ d'intervalles\ identiques}{nombre\ d'intervalles\ avec\ au\ mois\ un\ flux\ non\ nul}$$

J'ai réalisé différents tests pour observer l'effet du seuil de sensibilité du solveur et du coefficient d'extension des bornes de l'intervalle de variation du flux sur le score de similarité (Figure 56).
On observe que pour un coefficient d'extension des bornes inférieur à 35%, le score de similarité varie légèrement dans la plupart des cas, et la variation maximale reste inférieure à 5% dans les conditions les moins favorables. Le seuil à partir duquel le flux est considéré comme nul a lui plus d'importance, mais dans une seule situation : la comparaison de différents modèles sur milieu riche. Dans ces conditions, le score de similarité est doublé entre un seuil à 10^{-3} et un seuil à 10^{-5}. A la vue de ces résultats, j'ai décidé de prendre comme seuil 10^{-3} et comme coefficient 0.1. J'ai également été contraint d'arrondir les bornes des flux à 10^{-5}, du fait des approximations engendrées par les différentes optimisations du solveur.

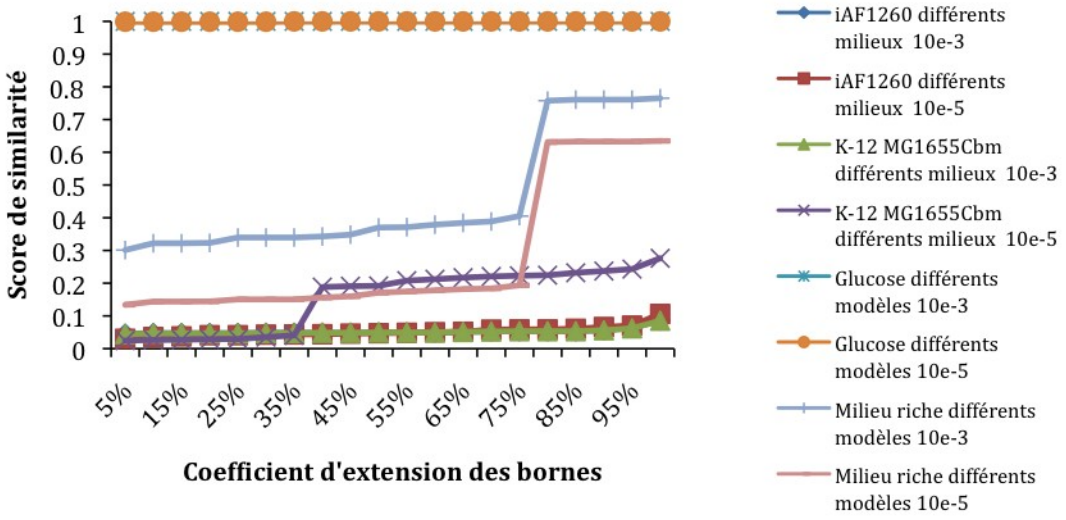


Figure 56 : Evolution du score de similarité pour différentes comparaisons.
J'ai testé différents seuils pour corriger les approximations du solveur et différents coefficients d'extension des bornes des intervalles. En bleu foncé et rouge, *iAF1260* sur milieu riche et milieu minimum glucose. En vert et violet, K-12 MG1655CBM sur milieu riche et milieu minimum glucose. En bleu clair et orange les deux modèles sur milieu minimum glucose. En gris et rose les deux modèles sur milieu riche.

La faible valeur du coefficient souligne un fait important : lorsque deux intervalles sont similaires, les bornes sont relativement proches ; à contrario si les deux intervalles sont différents, alors cette différence est suffisamment grande puisque l'on doit augmenter la taille de l'intervalle de plus de 80% pour les rendre similaires.
J'appelle paire similaire, une paire de flux dont les bornes des intervalles sont identiques ; une paire non nulle est une paire de flux dont au moins un des deux flux à des bornes différentes de 0 ; enfin une paire toujours non nulle est une paire de flux non nuls sur toutes les comparaisons d'une même étude.

4.1.4 Comparaison de la variabilité des flux : Application.

Pour cette analyse nous disposons des 4 milieux définis précédemment : milieu minimum glucose, milieu minimum gluconate, milieu riche et milieu riche sans sucre. Nous disposons également du modèle pivot *iAF1260* et du modèle reconstruit par notre processus *K-12 MG1655Cbm*.
Il existe deux façons d'envisager l'étude de la similarité des variations d'intervalles de flux :

(i) Un modèle sur deux milieux différents.

(ii) Deux modèles différents sur un milieu.

Dans le cas (i), nous nous intéressons à la variabilité au sein d'un même organisme entre deux conditions. Ce score est le reflet des capacités d'adaptation du métabolisme lorsque la souche change d'environnement. Puisque nous nous attachons aux valeurs maximales et minimales des flux, le score de similarité représente dans ce cas la variation maximum des flux qui traversent les réactions entre deux conditions et se lit : « entre deux milieux il y a au maximum un pourcentage *X* de flux dont le régime de fonctionnement est identique ».

Dans le cas (ii), le centre d'intérêt est la variabilité entre modèles et par extension entre souches. Le score de similarité reflète ici la variation des capacités métaboliques entre deux organismes. Il représente la variation maximum du flux d'une réaction entre deux modèles et se lit : « entre deux modèles et parmi les réactions communes il y a au maximum un pourcentage *X* de flux dont le régime de fonctionnement est identique ».

Appliqué aux quatre milieux et deux modèles (*iAF1260*, *K-12 MG1655Cbm*) le cas (i) donnent 2 fois 6 scores de similarités résumés dans la Table 27

		K-12 MG1655Cbm			
		Glucose	Gluconate	Milieu riche	Milieu riche sans sucre
iAF1260	Glucose		26.58%	4.39%	4.88%
	Gluconate	26.58%		4.32%	5%
	Milieu riche	4.65%	4.58%		19.91%
	Milieu riche sans sucre	4.88%	5%	32.42%	

Table 27 : Score de similarité de la variabilité des flux dans différents milieux.
Les valeurs dans la partie supérieur droite concerne le modèle reconstruit, et les valeurs dans la partie inférieur gauche celle du modèle pivot.

On observe dans la Table 27 qu'entre milieux de même type (simple ou complexe) le score de similarité est beaucoup plus élevé qu'entre milieux de types différents. Cette constatation était attendue sur milieu simple puisqu'à l'exception de la voie de dégradation de la source de carbone, les voies mises en jeu sont les mêmes : les voies de biosynthèse des métabolites précurseurs de biomasse. La similitude sur milieux complexe est plus surprenante, mais peut s'interprèter facilement. Il existe plusieurs sources de carbone (hors sucre), d'azote etc. et différents précurseurs de biomasses sont disponibles (acides aminés, nucléotides etc.) ; pour chaque source externe le modèle peut ainsi passer d'une utilisation exclusive d'une source à aucune utilisation de la source en passant par une utilisation modérée, et pour chaque précurseur il peut être synthétisé *«de novo »* ou importé depuis le milieu. Ces différentes possibilités font qu'une partie des flux, notamment de transport, ont des bornes minimales nulles et des bornes maximales égales à la limite de modélisation et sont donc totalement similaires entre les deux milieux complexes.
Toujours dans la Table 27 on remarque que les scores de similarité sont identiques entre les deux modèles, à l'exception des comparaisons comprenant le milieu riche; dans ces comparaisons, le score d'*iAF1260* est supérieur au score de *K-12 MG1655Cbm*. Cette différence résulte du seul effet de la voie de dégradation secondaire du *galactose* inférée dans le modèle reconstruit. En effet sur milieux minimum et milieu riche sans sucre, cette voie n'est pas utilisée et les scores de similarité sont donc identiques entre les deux modèles ; sur milieu riche la voie est utilisable, entrainant plus de capacités métaboliques et donc de variabilité en plus dans *K-12 MG1655Cbm*

Le cas (ii) appliqué aux deux modèles et aux quatre milieux, produit quatre comparaisons résumées dans la Table 28

	Score de similarité	Optimisation	
		iAF1260 flux de biomasse	*K-12 MG1655Cbm* flux de biomasse
Glucose	100%	0.929	0.929
Gluconate	100%	0.847	0.847
Milieu riche sans sucre	62.5%	25.633	25.742
Milieu riche	40.5%	28.619	28.664

Table 28 : Score de similarité, et optimisation de la biomasse dans différents milieux.
Les simulations sont effectuées sur deux milieux complexes : milieu riche et milieu riche sans sucre. Et sur deux milieux minimum : glucose et gluconate.

L'étude des scores de similarité montre qu'il n'existe aucune différence dans la variabilité des flux sur milieux minimum entre le modèle pivot *iAF1260* et le modèle reconstruit par notre processus *K-12 MG1655Cbm*. Ce constat était attendu puisque nous avions vu que la valeur du flux de biomasse optimal est la même pour les deux modèles, et les voies métaboliques utilisées sont les mêmes ; ces voies sont principalement les voies de biosynthèses des métabolites précurseurs de biomasse.
Sur les milieux complexes le score de similarité entre les deux modèles diminue. Il passe à 62.5% sur milieu riche sans sucre et à 40.5% sur milieu riche. Si pour le milieu riche la différence était attendue, pour le milieu sans sucre cette diminution du score est plus surprenante puisque la différence entre les deux modèles est une voie de dégradation du *galactose*. Il s'avère que cette voie peut être utilisée ; en effet le *galactitol* présent dans l'environnement peut être transformé en *galactose*.
On constate également que la similarité de la valeur du flux de biomasse ne semble pas liée au score de similarité de la variabilité des flux, mais sur seulement 4 valeurs il est pour le moment impossible d'estimer s'il existe ou non une corrélation entre le deux.
Ces observations mettent en avant deux points importants sur la comparaison des modèles CBMs

- La différence de variabilité des flux dépend de la complexité du milieu : plus celui-ci est simple, plus les intervalles des flux sont similaires et les flux traversent des réactions associées aux voies de biosynthèse des précurseurs de biomasse (sous condition que le modèle possède les voies adéquates).

- Des résultats d'analyses FBA proches n'impliquent pas une similitude de la variabilité des intervalles de flux ; autrement dit pour un optimum proche, l'espace de solutions peut être très différent. Ainsi l'ajout d'une voie de dégradation supplémentaire modifie les régimes extrêmes de fonctionnement du métabolisme.

Les comparaisons entre *iAF1260* et *K-12 MG1655Cbm* ont montré que le processus de reconstruction est capable de reproduire un modèle fidèle à celui d'origine et surtout un modèle fonctionnel. La reconstruction de ce modèle est particulière puisqu'il s'agit du même organisme que le modèle pivot et donc du même génome. La souche K-12 W3110 étant pratiquement identique à K-12 MG1655, son modèle reconstruit servira de témoin.

.4.2 Comparaisons des différents modèles reconstruits.

J'ai reconstruit les 23 modèles métaboliques des souches utilisées lors de l'analyse des réseaux métaboliques (chapitre I partie 1). Les 6 *Shigellas* ont volontairement été écartées de la reconstruction. Les différences métaboliques sont trop importantes et elles ne correspondent pas aux conditions de proximité du pivot. De plus, leur mode de vie parasitique n'était pas compatible avec les milieux définis pour le modèle de référence.

4.2.1 Différences de composition

Les différences et similitudes entre le modèle de K-12 MG1655 reconstruit et *iAF1260* ne sont pas représentatives de ce que l'on peut observer avec les autres modèles reconstruits (Figure 57). Ainsi, si on ne compte que 5 réactions en plus dans *K-12 MG1655Cbm*, on dénombre en moyenne 54 nouvelles réactions dans les autres modèles reconstruits. Le plus faible ajout (hors K-12 MG1655) concerne K-12 W3110 avec 22 nouvelles réactions, tandis que le plus grand ajout concerne CFT073 avec 78 réactions.
La souche K-12 W3110 est un clone métabolique de la souche K-12 MG1655 : il est donc normal que le modèle de K-12 W3110 contienne peu de nouvelles réactions par rapport au modèle de K12-MG1655 qui est la référence de l'espèce. Néanmoins ces nouvelles réactions montrent les limites d'*iAF1260*, qui comme tout modèle possède des lacunes et peut être amélioré.
Bien que ces réactions ne soient pas présentes dans *iAF1260*, cela ne signifie pas que les modèles reconstruits sont significativement plus grands. La taille moyenne des modèles est de 2391 réactions, flux d'échange compris, soit une dizaine de plus que chez *iAF1260* ; le nombre total de réactions différentes est de 2270 (2571 avec les flux d'échange). Le plus petit modèle est celui de la souche O127:H6 E2348/69 avec 2352 réactions (30 de moins qu'*iAF1260*) et le réseau le plus grand est celui d'UMN026 avec 2416 réactions (34 de plus qu'*iAF1260*).
Si UMN026 possède le réseau métabolique le plus grand et donc naturellement le modèle métabolique le plus grand, O127:H6 E2348/69 ne possèdent pas le réseau le plus petit alors que son modèle l'est. Si l'ordre de grandeur des réseaux et des modèles est cohérent, la perte de diversité métabolique au niveau des modèles n'impacte pas toutes les souches de la même manière.
Si certains modèles sont plus petits que le pivot alors qu'ils possèdent des réactions ajoutées par les différents modules, c'est tout simplement parce qu'il existe des réactions d'*iAF1260* pour lesquelles la GPR ne possède pas d'homologue dans l'organisme étudié (Figure 58). Ainsi les réactions, qui sont présentes dans l'ensemble des modèles, sont au nombre de 2244. Je m'attendais à retrouver l'ensemble des réactions utilisées par *iAF1260* sur milieu minimum glucose dans ce *« core »* réactionnel ; cependant sept réactions sont manquantes (Table 29).
Pour cinq d'entre elles il s'agit des instances d'un transporteur du périplasme vers le cytoplasme des bases azotées. Ce transporteur est absent de neuf modèles (ceux des souches 042, ATC 8739, E24377A, IAI1, HS, les 3 O157:H7 et UMN026). Aucun homologue au gène *nupG* (identifiant 1087153 de la base de données MicroScope), n'a été trouvé dans ces souches.
Pour les deux autres réactions absentes, il s'agit également d'une absence de gène homologue : ainsi *l'hydroxypyruvate isomérase* (numéro EC 5.3.1.22) n'apparaît pas dans la souche O157:H7 EC4115, cette réaction n'est pas directement impliquée dans une voie métabolique de dégradation ou de synthèse. Enfin, la dernière réaction

absente des modèles des souches de ED1a et IAI39, est la *dihydroxyacétone phosphotransférase* (numéro EC 2.7.1.121) impliquée dans une voie alternative de la dégradation du glycérol.

Transporteur périplasmique des nucléosides EC: aucun

H^+ + Nucléoside [périplasme] -> H^+ + Nucléoside [cytoplasme]

Hydroxypyruvate isomérase EC: 5.3.1.22

dihydroxyacétone + phosphénolpyruvate -> dihydroxyacétone phosphate + pyruvate

Dihydroxyacétone phosphotransférase EC: 2.7.1.121

hydroxypyruvate <=> tartronate semialdéhyde

Table 29 : Réactions absentes du core réactionnel des modèles.

Lorsque l'on compare le core réactionnel avec l'ensemble des flux non nuls issu de l'analyse FVA d'*iAF1260* sur milieu minimum glucose, on le trouve compris dans cet ensemble. Autrement dit les sept réactions absentes du core ne sont pas essentielles.

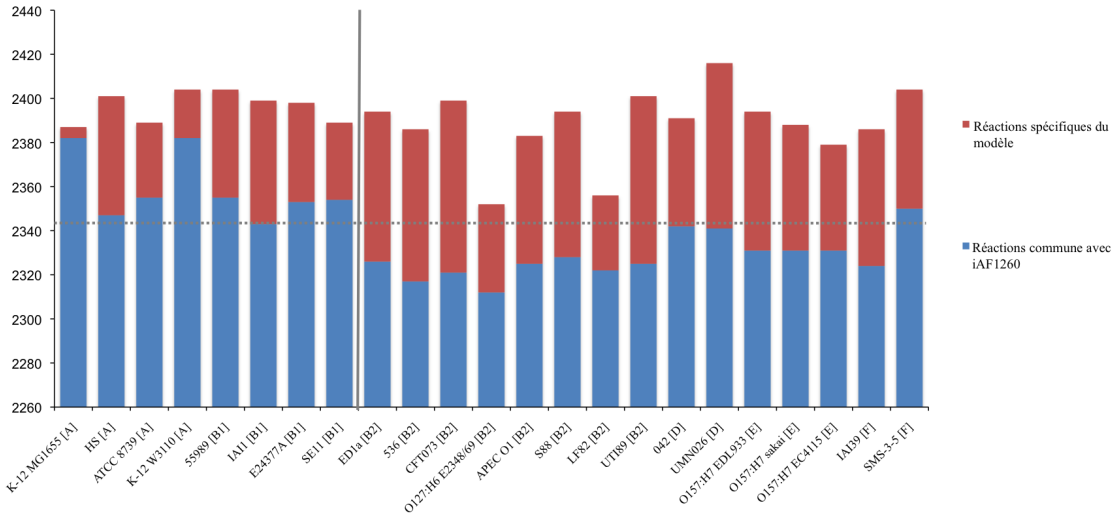


Figure 57 : Nombre de réactions communes et spécifiques des différents modèles avec le modèle référence.

Le modèle de référence comprend 2382 réactions. Les souches sont organisées suivant leur groupe phylogénétique, la barre verticale sépare les souches A et B1 des souches des autres phylogroupes. La barre en pointillé horizontale donne la valeur minimum de réactions communes pour les souches A et B1.

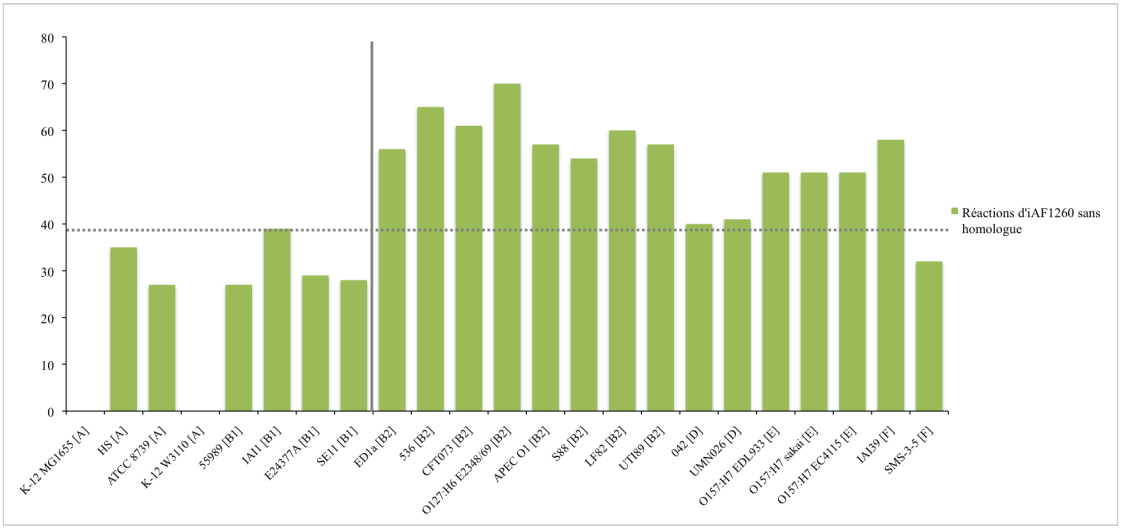


Figure 58 : Nombre de réactions spécifiques d'*iAF1260* suivant les différents modèles.
Les souches sont organisées suivant leur groupe phylogénétique ; la barre verticale sépare les souches A et B1 des souches des autres phylogroupes. La barre en pointillé horizontale donne la valeur maximum de réactions sans homologue pour les souches A et B1.

Sans surprise les deux modèles de K-12 (*K-12 MG1655Cbm* et *K-12 W3110Cbm*) sont les seuls qui possèdent l'ensemble des réactions du modèle pivot et qui ont le moins de réactions supplémentaires. On constate également un effet de la phylogénie sur le nombre de réactions d'*iAF1260* manquantes : les sept premiers modèles de la Figure 58 issus des groupes phylogénétiques A et B1, ont moins de 30 réactions manquantes contre plus de 30 pour les autres. Cette proximité est également visible sur la Figure 57, où ces mêmes modèles possèdent plus de réactions en commun avec *iAF1260* et moins de réactions spécifiques en moyenne que les modèles issus de souches provenant d'autres groupes phylogénétiques. Comme pour les réactions, j'ai comparé la présence et absence des métabolites ; les modèles reconstruits possèdent tous plus de métabolites que le modèle de référence avec une moyenne de 1069 métabolites par modèle (contre 1039). Les métabolites impliqués dans la fonction de biomasse sont tous présents dans tous les modèles. Si cela ne garantit pas le bon fonctionnement de ces derniers, cela permet de conserver la fonction de biomasse et d'être confiant quant aux résultats des simulations. Parmi les 1039 métabolites d'*iAF1260*, 1003 sont communs à tous les modèles, et 121 métabolites, absents de ce modèle, ont été nouvellement créés par le processus de reconstruction.

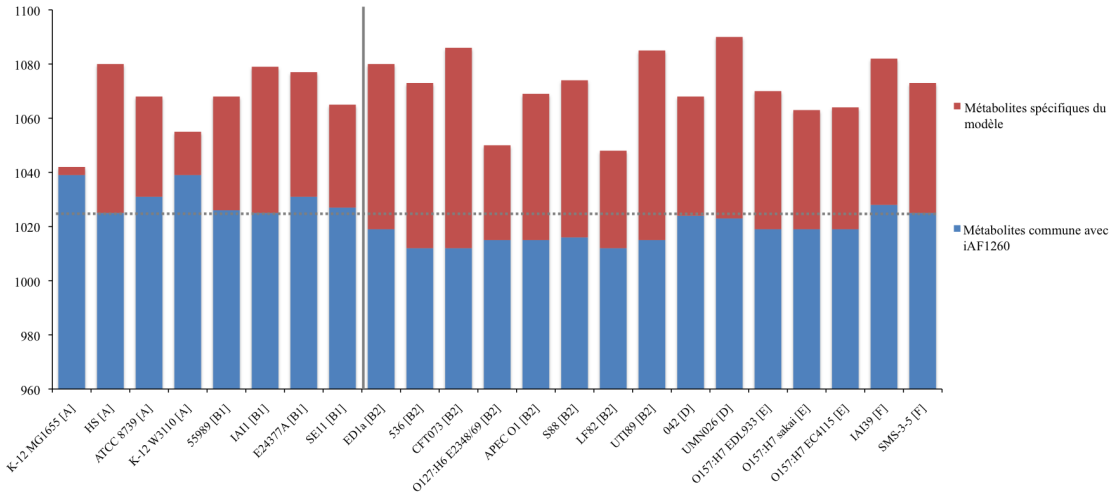


Figure 59 : Nombre de métabolites communs et spécifiques des différents modèles avec le modèle référence.

Le modèle de référence comprend 1039 métabolites. Les souches sont organisées suivant leur groupe phylogénétique, la barre verticale sépare les souches A et B1 des souches des autres phylogroupes. La barre en pointillé horizontale donne la valeur minimum de métabolites communs pour les souches A et B1.

Le nombre de métabolites communs avec *iAF1260* et le nombre de métabolites spécifiques de nos modèles sont cohérents avec ce que j'ai observé pour les réactions (Figure 59): les modèles avec le plus de réactions en commun avec *iAF1260* sont ceux avec le plus de métabolites en commun. L'effet de la phylogénie est également présent dans la diversité des métabolites. De même, les métabolites spécifiques d'*iAF1260* dans mes différents modèles suivent la même tendance que celle des réactions (Figure 60).

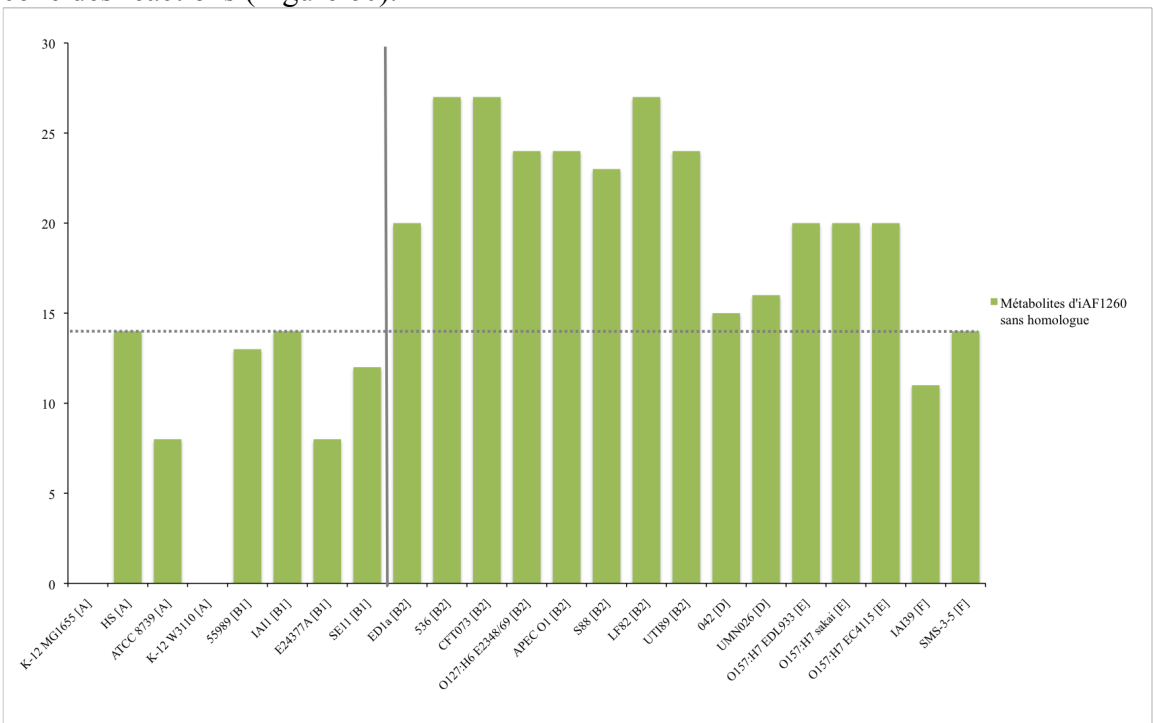


Figure 60 : Nombre de métabolites spécifiques d'*iAF1260* suivant les différents modèles.

Les souches sont organisées suivant leur groupe phylogénétique, la barre verticale sépare les souches A et B1 des souches des autres phylogroupes. La barre en pointillé horizontale donne la valeur maximum de métabolites sans homologue pour les souches A et B1.

Cette similitude entre les réactions et les métabolites indique que celles-ci utilisent des métabolites spécifiques, c'est-à-dire unique à la réaction, et non ubiquitaires (e.g. pyruvate, acetyl-coa, etc.). De ce fait, l'absence de la réaction entraîne automatiquement l'absence du métabolite dans le modèle. Le fait que les métabolites précurseurs de biomasse soient présents et que toutes les réactions essentielles sur milieu minimum glucose soient également là, montre que les réactions et les métabolites sans homologue dans *iAF1260* sont impliqués dans des voies métaboliques pour lesquelles il existe des chemins alternatifs.
Cette spécificité réaction/métabolite est également vérifiée pour les éléments métaboliques issus de MicroCyc et ajoutés aux modèles. On constate qu'en moyenne ces modèles possèdent 52 nouvelles réactions pour 48 nouveaux métabolites. Ce ratio d'environ un métabolite par réaction, indique que les ajouts ne concernent pas des éléments ubiquitaires, mais là aussi des voies et chemins alternatifs dans le métabolisme, tendant à montrer que nous sommes capables de capturer une partie de la spécificité du métabolisme de chaque organisme.
Avant de passer aux simulations, le dernier élément de comparaison concerne les transporteurs, qui est l'un des points les plus délicats des modèles du métabolisme. En effet en raison de la multi-spécificité des transporteurs, de la difficulté de leurs identifications et du peu de connaissance sur leurs GPRs, leur inclusion dans les modèles est difficile, et pourtant essentiel. De ce fait, on ne dénombre en moyenne deux nouveaux transporteurs par modèle, et en moyenne seulement un transporteur d'*iAF1260* absent. Le modèle *O127:H6 E2348/69Cbm* fait tout de même exception à cette observation : sept transporteurs sont absents ; ils sont du type ABC et utilisent de l'ATP. Les GPRs de ces réactions n'ont pu être validées par la fonction *gprEval* du module *cbmCom*.
Par exemple le transporteur du *α-L-arabinopyranose* est codé par trois gènes chez K-12 MG1655 ce qui donne la GPR (*araH* et *araG* et *araF*). Chez O127:H6 E2348/69 le gène *araH* est fragmenté : il est donc considéré comme non fonctionnel et par conséquent, la GPR associée ne peut être évaluée positivement (Figure 61). La complémentarité des deux fragments et la similarité avec le gène de K-12 MG1655, laissent supposer une possible erreur de séquençage.

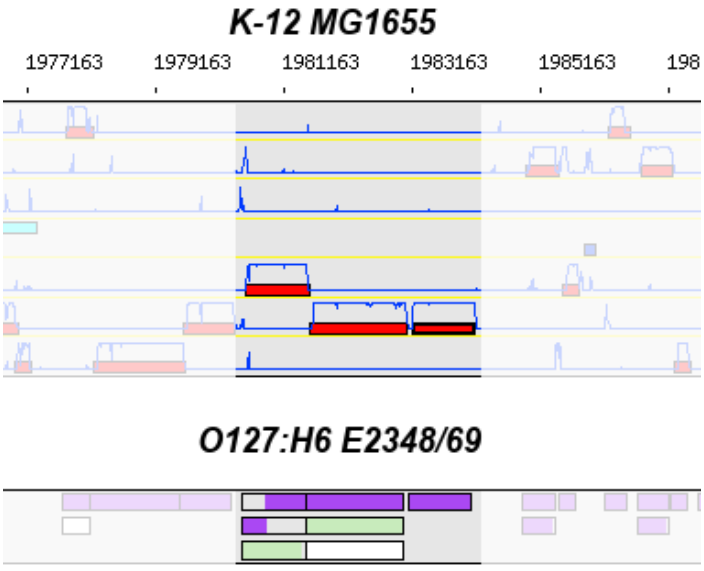


Figure 61 : Alignement de l'opéron *ara*.
L'opéron est composé de 3 gènes *araG araH* et *araF*, ils codent pour l'ABC transporteur du *α-L-arabinopyranose*. On constate que le premier gène *araH* est fragmenté chez O127:H6 E2348/69.

La présence de tous les métabolites précurseurs de biomasse, le faible nombre de transporteurs manquants et les chemins alternatifs impliqués dans les différences observées, laissent présager des modèles fonctionnels capables d'avoir un flux de biomasse non nul dans différents environnements.

4.2.2 Simulation et optimisation des modèles

Contre toute attente, la première version des modèles (*v0*) s'est avérée non fonctionnelle quel que soit le milieu de simulation : le flux de biomasse calculé par le solveur numérique pour toutes les combinaisons modèles et milieux possibles était nul. J'ai analysé le contenu en réactions de cette v0, et j'ai constaté l'absence des réactions artificielles qui servent de puits dans le modèle pivot. En effet ces cinq réactions ne sont pas étiquetées comme spontanées dans *iAF1260*, et leur équation bilan n'est pas équilibrée : elles sont non compatibles avec les critères de sélection automatique des réactions de mon module. J'ai créé une exception pour gérer ce genre de réactions et j'ai reconstruit une nouvelle version des modèles (*v1*). Cette fois-ci les modèles sont capables de produire de la biomasse dans différentes conditions, à l'exception du modèle *536Cbm*. Ce dernier est capable de produire de la biomasse sur milieu riche avec un flux de 28.17, contre 28.62 pour *iAF1260*. Par contre le flux de biomasse devient nul sur les milieux minimums à l'exception du milieu minimum arginine. En comparant le contenu en métabolites et en réactions, en particulier les réactions traversées par des flux non nuls sur milieu minimum des autres modèles, une réaction dont la présence est indispensable sur tous les milieux minimums sauf l'arginine a été isolée. Il s'agit de l'*ornithine carbamoyltransférase* ou OCBT qui est codée par deux gènes, *argF* et *argI* chez K12-MG1655 (Figure 62).

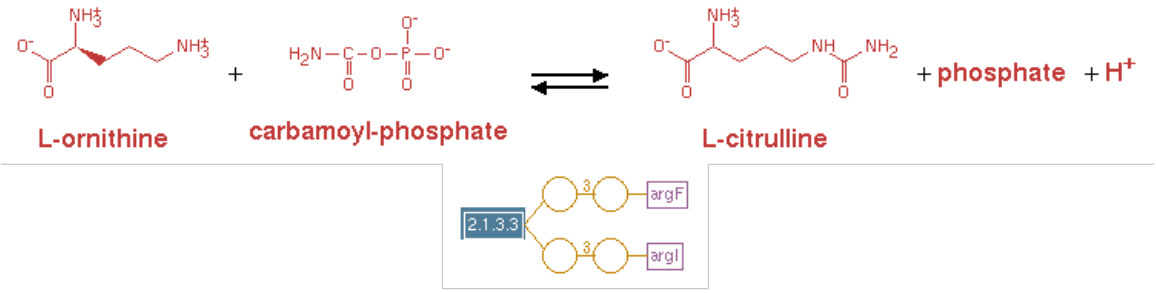


Figure 62 : L'OCBT, réaction impliquée dans la voie de synthèse de l'arginine.
Cette réaction est indispensable pour la synthèse de l'acide aminé arginine puisqu'il n'existe pas de voie alternative dans les souches d'*E. coli*. Cette réaction est encodée par deux gènes chez K-12 MG1655 : *argI* et *argF*.

Une étude des homologues des gènes *argF* et *argI* montre que ces derniers sont fragmentés dans le génome de la souche 536 : les deux parties du gène *argI* sont dans des cadres de lectures décalé ce qui laisse envisager une erreur de séquençage. Quant au gène *argF*, il est tronqué passant de 324 acides aminés chez K-12 MG1655 à 45 chez 536. La réaction OCBT ne peut être récupérée par le processus puisque la GPR correspondantene peut être évaluée positivement. Pour résoudre cette difficulté cette réaction étant essentielle, j'ai utilisé la fonction d'ajout sans vérification des réactions pour l'insérer dans *536Cbm*. J'ai donc généré une nouvelle version de mes modèles (*v2*) qui est celle utilisée pour les simulations.
J'utilise comme fonction de biomasse pour tous mes modèles celle définie dans *iAF1260*, de même, j'ai conservé la fonction de maintenance énergétique avec la même borne minimale (voir partie 2.9).

Dans cette *v2*, tous les modèles produisent de la biomasse sur milieu minimum glucose et gluconate, sur milieu riche et milieu riche sans sucre. Et c'est sans surprise que l'on retrouve un flux de biomasse dont la valeur est identique pour tous les modèles sur milieu minimum glucose ou gluconate, respectivement 0.9293 et 0.8475. On observe néanmoins de la variabilité sur les deux milieux complexes (Figure 63).

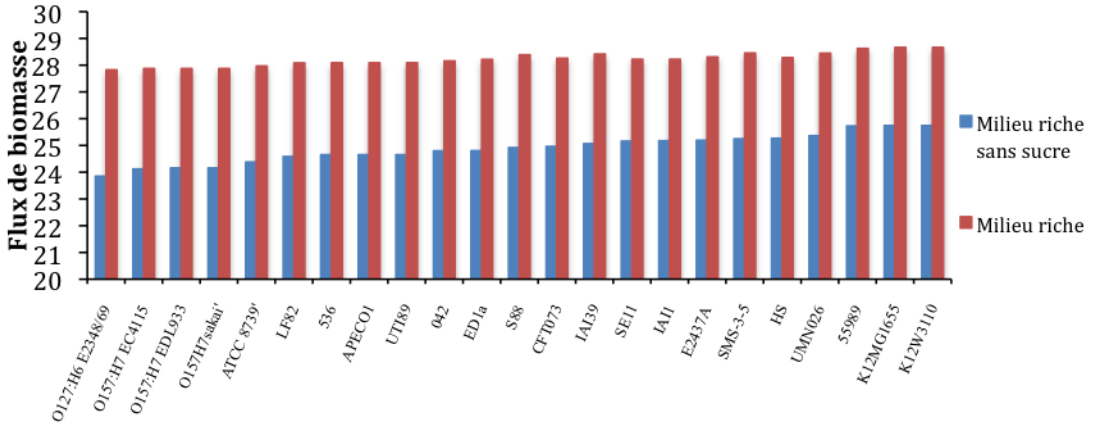


Figure 63 FBA réalisées sur 23 modèles et 2 milieux complexes.
En rouge les résultats d'optimisation des modèles sur milieu riche et en bleu sur milieu riche sans les sucres. Il existe peu de variation sur milieu riche, cette variation est légèrement plus grande sur milieu riche sans sucre.

Sur milieu riche, la valeur du flux de biomasse varie de 27.8246 pour O127:H6 E2348/69 à 28.6643 pour les deux souches de K-12, avec une moyenne de 28.22. *O127:H6 E2348/69Cbm* est le modèle avec le moins bon rendement : cette observation est cohérente avec le fait que ce soit le modèle comprenant le moins de réactions et de transporteurs, et donc le moins de capacités métaboliques. Sur milieu sans sucre, les valeurs des flux de biomasse vont de 23.85 à 25.743 avec une moyenne de 24.885. Si la différence de 0.84 entre le meilleur et le moins bon rendement sur milieu riche peut paraître faible, elle est cependant égale au rendement des modèles sur milieu minimum gluconate ; on peut donc considérer que la différence n'est pas négligeable.
La plus grande variabilité dans le milieu riche sans sucre s'explique parceque les sucres sont à la fois source de carbone et d'énergie, en contraignant les modèles sur ces deux points, ils doivent exploiter les voies alternatives ce qui met en évidence les différentes capacités métaboliques des modèles.
On constate que sur les milieux riches, les deux modèles les plus efficaces sont ceux des souches de K-12, ce qui est cohérent puisqu'il s'agit de la même souche que le modèle pivot qui a subi une expertise manuelle poussée.
J'ai ensuite étudié les différentes capacités de mes modèles. Pour cela j'ai estimé, comme pour *K-12 MG1655Cbm*, l'ensemble des milieux minimums sur lesquels les modèles prédisent un flux de biomasse non nul (Figure 64). En moyenne, les modèles donnent une croissance positive sur 165 milieux, ce qui est faible comparé aux 172 milieux d'*iAF1260*. De plus, les modèles capables d'utiliser le plus de sources de carbone sont les souches de K-12. Sachant que j'ai introduit de nouveaux métabolites et transporteurs, les résultats sont en deçà des attentes.

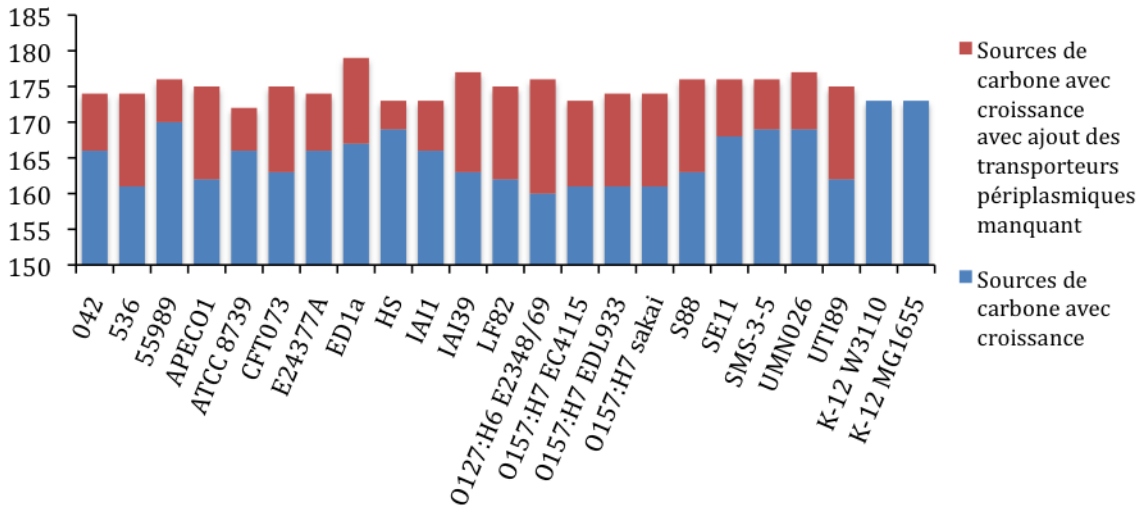


Figure 64 : Nombre de sources de carbone pour lesquelles les modèles prédisent un flux de biomasse non nul.

En bleu, résultats directement issus des modèles ; en rouge, résultat après ajout des transporteurs périplasmiques potentiellement absents.

J'ai vérifié la connectivité des différents modèles pour m'assurer que les réactions introduites s'intègrent à la base du modèle issue d'*iAF1260*. J'ai trouvé qu'à l'exception des modèles des K-12, tous les autres présentent des sous-parties non connectées au reste des réactions. Ce nombre de sous-parties varie de 5 pour le modèle *HSCbm* à 19 pour le modèle *O127:H6 E2348/69Cbm*, avec une moyenne de 12 sous-parties par modèle. Pour chaque modèle, la plus grande sous-partie comprend plus de 99% des réactions du modèle, tandis que la moyenne des autres sous-parties est de deux réactions.

J'ai, dans un premier temps, supposé que ces sous-parties étaient dues à l'introduction des nouvelles réactions provenant de MicroCyc ; il s'est avéré que toutes les réactions nouvellement créées font partie de la sous-partie principale : les réactions ajoutées par le module *cycSpe* sont toutes reliées à la composante principale des modèles.
J'ai recherché les causes des autres sous-parties : il s'est avéré que la totalité des sous-parties, quelque soit le modèle, est due à l'absence de plusieurs transporteurs périplasmiques. Contrairement au problème rencontré avec *536Cbm*, il ne s'agit pas d'hypothétiques erreurs de séquençage puisque dans la plupart des cas, le gène codant pour la réaction est absent. Une illustration est donnée dans la Figure 65.
J'ai déjà évoqué la difficulté d'estimer et d'inférer les transporteurs métaboliques. Comme ils permettent le passage de plusieurs composés, l'absence du transporteur principal d'un métabolite, ne signifie pas systématiquement l'absence du transport de celui-ci. On peut envisager divers scénarii, cependant il est impossible d'en réfuter sans un travail de curation approfondi des modèles et notamment la comparaison des simulations avec des résultats d'expériences.
On peut envisager plusieurs hypothèses qui expliquent cette absence de transporteur :

(i) Véritable absence du transporteur périplasmique.

(ii) Absence des transporteurs extracellulaire et périplasmique.

(iii) GPR erronée dans le modèle *iAF1260*.

(iv) Présence d'une isozyme.

(i) Il s'agit du cas le plus simple : le transporteur périplasmique n'existe pas dans l'organisme bien que le transporteur extracellulaire soit présent. (ii) Le transporteur périplasmique et le transporteur extracellulaire sont tous les deux absents de la souche, cependant suite à une GPR erronée, le transporteur extracellulaire est inféré par le processus de reconstruction. (iii) La GPR du transporteur chez *iAF1260* est fausse par conséquence l'évaluation positive de la GPR par le module *cbmCom* ne peut avoir lieu. (iv) La GPR dans le modèle *iAF1260* est correcte, mais dans les organismes en question, c'est une isozyme qui catalyse la réaction.
Avec les données actuelles, il est impossible de valider ou de réfuter les hypothèses formulées : nous ne pouvons donc pas appliquer de corrections aux modèles reconstruits.

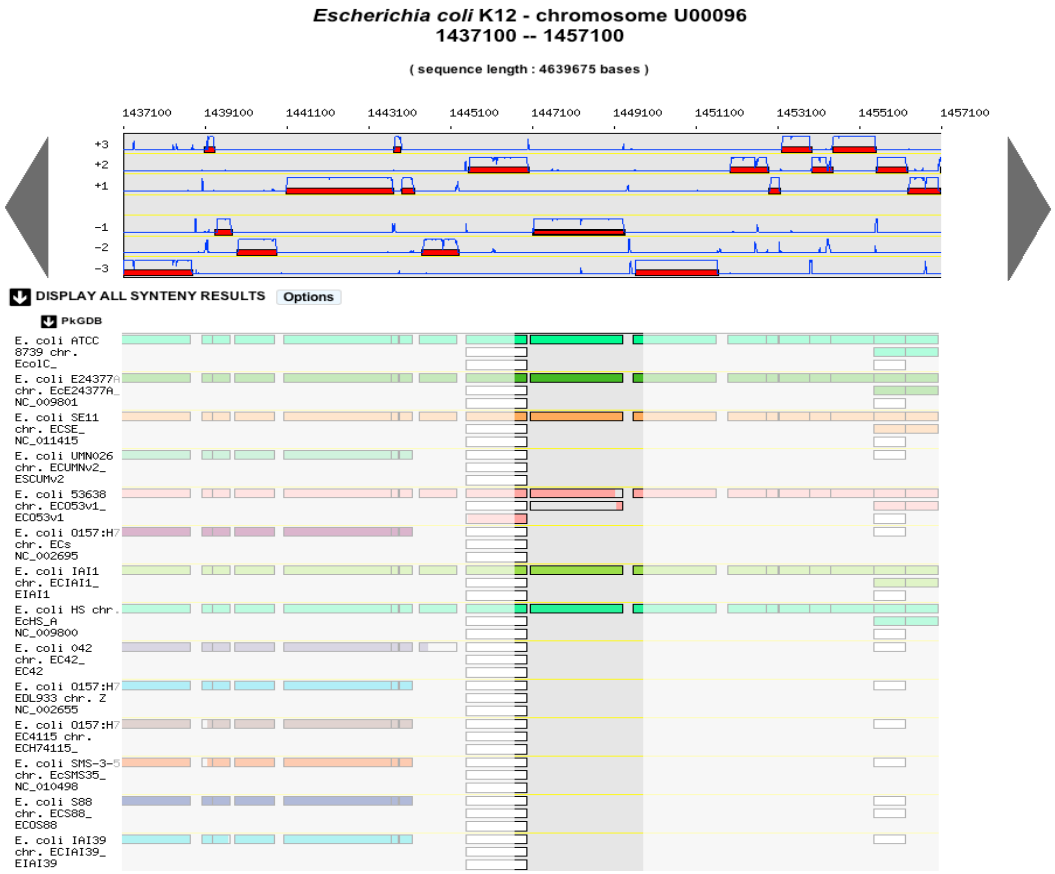


Figure 65 : Alignements multiples des *E. coli* pour le gène *tynA*.
Le gène est soit présent avec un très fort taux de similitude, soit absent de la souche.

Au final, ce sont 28 transporteurs périplasmiques associés à 18 GPRs qui vont être absents dans un ou plusieurs modèles. Avec l'ajout de ces transporteurs dans les différents modèles, j'ai réalisé une nouvelle estimation des sources de carbones assimilables (Figure 64). Avec cette modification, le résultat est plus proche de mes attentes : les modèles sont capables d'utiliser plus de source de carbone que le modèle initial puisqu'ils disposent de nouvelles capacités métaboliques conférées par les réactions converties de MicroCyc.

Cette observation explique en partie pourquoi sur milieux riches les deux modèles de K-12 sont les plus efficaces à produire de la biomasse ; en effet, avec en moyenne deux transporteurs extérieurs supplémentaires et 9 transporteurs périplasmiques absents, les modèles reconstruits ont une perte de sept sources de métabolites assimilables.

Pour le reste des travaux, j'utiliserai les modèles sans ces transporteurs. En effet puisqu'il m'est impossible, à ce stade, de prouver leur présence, je préfère éviter un ajustement arbitraire de mes modèles, ce qui me permet d'estimer les caractéristiques des modèles dans le pire des cas.

4.2.3 Variabilité des flux et utilisation des réactions dans les différents modèles

Les premiers calculs de la variabilité des flux étaient longs et coûteux en ressources, les améliorations algorithmiques des différentes méthodes dédiées aux CBMs permettent d'effectuer ce calcul beaucoup plus rapidement. Je peux donc, comme pour la comparaison *iAF1260* contre *K-12 MG1655Cbm*, regarder celle-ci par modèle mais surtout comparer la similarité entre différentes simulations. Pour cela, je vais comparer un modèle sur différents milieux, puis différents modèles sur le même milieu.
Pour faciliter le calcul des similitudes de flux, j'ai développé de nouvelles fonctions compatibles avec la COBRA-Toolbox. La variabilité des flux est calculée par la méthode *fastFVA* qui utilise le solveur numérique *glpk*, le tout exécuté dans le logiciel Matlab. La première fonction (*similarityFluxes*) permet de calculer le score de similitude à partir de deux résultats de *fastFVA* ; elle prend également en paramètre un seuil en deçà duquel le flux est considéré comme nul et le coefficient d'extension des bornes de l'intervalle du flux. Pour faciliter l'utilisation de cette fonction, j'ai développé deux autres fonctions (*compareModelFVA* et *compareMediaFVA*) dont le rôle est de réaliser deux FVA et de calculer le score de similitude. *CompareModelFVA*, comme son nom l'indique, se focalise sur un modèle et exécute un calcul de similarité de la variabilité des flux pour deux environnements différents donnés en paramètre. *CompareMediaFVA* réalise le même calcul pour un environnement précis et deux modèles différents. Ces trois fonctions donnent en résultat les mêmes variables : le score de similarité défini dans la partie 4.1.3, la liste des réactions avec un flux non nul dans au moins une des conditions testées, et la liste des réactions avec un intervalle de valeurs admissibles similaire dans les deux conditions testées.
L'ensemble des calculs de similarité ont utilisé la même valeur de seuil en deçà duquel le flux est considéré comme nul = 10^{-3}, et le même coefficient de fluctuation pour les bornes des intervalles des flux = *0.1*.

Modèle constant et environnements différents.

J'ai limité l'étude de la diversité en fonction des médias sur quatre environnements : milieu riche, milieu riche sans sucre, milieu minimum glucose et milieu minimum gluconate. Afin de faciliter la lecture des résultats, j'ai mis en place des abréviations : R désigne le milieu riche, RSS le milieu riche sans sucre, Glc le milieu minimum glucose, et Gla le milieu minimum gluconate. La paire de milieux testés sera désignée par leurs abréviations séparées par un « _ » (e.g. Glc_Gla pour une comparaison milieu minimum glucose et milieu minimal gluconate).
Les scores de similitude sont très faibles, compris entre 4% (R_Gla) et 29% (Glc_Gla). Au sein d'une même comparaison de couple de milieux, la variation du score en fonction des modèles est négligeable : le ratio de la variance sur la moyenne est inférieur à 0,01%. Les scores les plus élevés sont obtenus sur milieu de même type avec en moyenne : 15% R_RSS et 27% Glc_Gla. Le score devient très faible entre les milieux minimum et le milieu riche (respectivement 4.6% et 4.4% de similarité). Ce score augmente légèrement pour atteindre 7.9% pour chacun des milieux minimum contre le milieu sans sucre.

Les similitudes sur milieu minimum étaient attendues, puisque exceptée la voie de dégradation de la source de carbone, les voies de synthèse de métabolites essentiels mis en jeu son les mêmes. Le score de similitude sur les milieux riches est plus surprenant étant donné le nombre de capacités métaboliques utilisées. Le score de similitude sur Glc_Gla est tout de même en-dessous des attentes : il y a une moyenne de 412 flux non nuls sur Glc, 431 sur Gla, et 401 de ces flux sont communs aux deux conditions (Table 30). Puisque 97% des réactions avec un flux non nul sont communes aux milieux minimums, on pouvait s'attendre à une plus forte similarité dans la variabilité de ces flux. De cette observation, on peut en déduire qu'une petite différence parmi les capacités métaboliques mises en jeu (3%), peut entraîner une grande différence de l'espace de solution : seulement 28% des flux varient de la même manière.

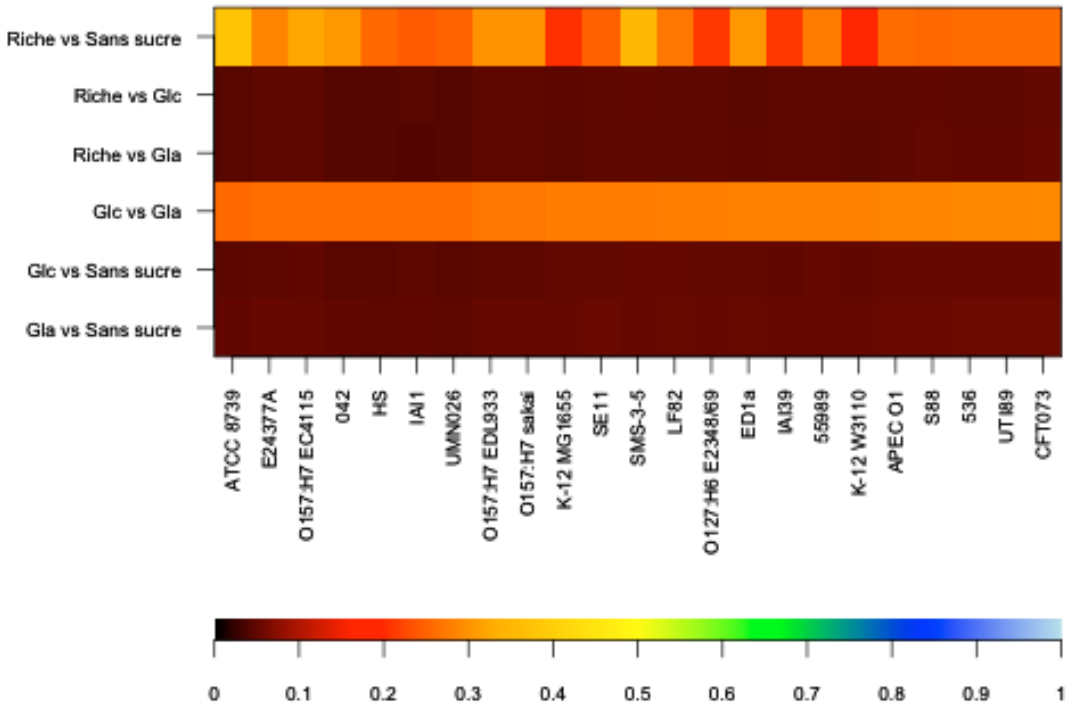


Figure 66 : Similarité de flux à modèle constant et environnements différents.
Milieu riche : milieu pour lequel l'ensemble des métabolites externes du modèle sont présents. Milieu sans sucre : milieu riche auquel tous les sucres ont été retiré. Milieu Glc : milieu minimum avec comme source de carbone le glucose. Milieu Gla : milieu minimum avec comme source de carbone le gluconate.

J'ai regardé parmi les flux similaires et non nuls ceux qui sont récurrents dans différentes conditions (Table 30). On constate une homogénéité dans le nombre de paires présentes dans toutes les comparaisons, que ce soit les paires similaires ou non nulles. Ainsi lors des comparaisons entre milieux de même type (Glc_Gla et R_Rss), ce nombre atteint la centaine de réactions ; il diminue à une quarantaine de réactions lors des comparaisons entre milieux de types différents (Glc_R, Glc_Rss. R_Gla et Gla_Rss).

	Riche vs Sans Sucre	Riche vs Glucose	Riche vs Gluconate	Glucose vs Gluconate	Glucose vs Sans sucre	Gluconate vs Sans Sucre	Toutes comparaisons
				Similaire			
Commun	134	39	38	96	39	43	38
Total	931	1014	1033	401	971	987	1104
				Non nul			
Commun	249	40	42	97	42	47	39
Total	1000	1079	1097	401	1066	1081	1167

Table 30 : Flux similaires et flux non nuls.

La première ligne donne le nombre de flux similaires présents dans tous les modèles. La deuxième ligne donne le nombre de flux similaires dans au moins un des modèles. La troisième ligne donne le nombre de flux non nuls présents dans tous les modèles. La dernière ligne donne le nombre de flux non nuls dans au moins un modèle.

Si le nombre de paires est proche dans les comparaisons sur milieux de même type, la nature des réactions qui les compose, est différente. Sur les milieux riches, 44% des réactions sont des réactions de transport et 21% des réactions d'échange ; ces pourcentages tombent respectivement à 28% et 3% sur milieux simples. J'ai regardé les paires communes des comparaisons Glc_Gla et R_RSS, en dehors des cofacteurs les réactions utilisent des acides aminés et des métabolites du métabolisme central (le pyruvate, acétate, etc.).

Les réactions spécifiques de l'analyse Glc_Gla utilisent beaucoup plus de cofacteurs différents : *NADH*, *COA*, *quinone*, *demethylmenaquinone*. Elles produisent également de nombreux acides aminés, mais la principale différence par rapport aux réactions de la comparaison R_RSS, est la synthèse des *ribo* et *désoxyribonucléotides*. Leur présence s'explique simplement par la disponibilité de ces métabolites sur milieux riches et leur absence sur milieux minimum. La similarité des flux de synthèse des nucléotides sur Glc_Gla est due au taux de croissance très proche entre milieu minimum glucose et Gluconate, ceci implique la production de séquences nucléotidiques à des vitesses identiques.

Les paires spécifiques de l'analyse R_RSS, sont pour les deux tiers des transporteurs ou des flux d'échanges. Les limites de ses flux sont proches de la valeur maximale (ou minimale) que j'ai imposée comme contrainte de simulation; ceci laisse supposer que sur ces deux milieux ces réactions sont des facteurs limitant le flux de biomasse. Ces observations montrent que si globalement les modèles se comportent de manière assez similaire, localement il existe une grande variation. On peut ajouter à ce constat le nombre total de paires similaires dans au moins une des comparaisons : il est très proche du nombre total de paires non nulles (Table 30). Autrement dit pour 95% des paires non nulles, dans au moins un milieu et un modèle (V), il existe dans un modèle m un flux v_m et une comparaison de milieux (e_1, e_2) tel que les intervalles du flux v_{me1} et v_{me2} sont similaires.

Modèles variables et environnements constants

En première partie de l'étude de la variabilité des différents modèles, je vais comparer l'utilisation des flux entre FVA et FBA (Table 31).

	Flux non nuls FBA	Flux actifs FVA	Flux non nuls FVA	Flux non nuls FBA	Flux actifs FVA	Flux non nuls FVA
		Riche			Riche sans sucre	
moyenne	769	892	660	732	842	650
min	787	871	641	722	827	630
max	746	921	670	744	863	660
		Minimum glucose			Minimum gluconate	
moyenne	414	415	305	414	434	295
min	409	408	304	409	427	294
max	422	422	307	424	441	298

Table 31 : Comparaison de l'activité des flux entre FVA et FBA.

Le nombre de flux mis en jeux sur milieux riches est toujours supérieur à celui sur milieux minimum. Le nombre de flux actifs sur FBA est toujours compris entre le nombre de flux non nuls sur FVA et le nombre de flux actifs sur FVA. Le premier représente les flux toujours actifs dans l'espace de solution, et le second, les flux qui peuvent être actifs dans une des solutions alternatives.

Les résultats sur FVA et FBA donnent le même profil de résultat : le milieu riche est celui qui utilise le plus de réactions différentes (769) ; c'est également le milieu sur lequel on a le plus d'écarts entre le nombre minimum de flux actifs (746 chez O157:H7 Sakai) et le nombre maximal (787 chez ED1a). Le milieu riche sans sucre est évidemment celui qui arrive juste après en nombre de flux actifs, avec en moyenne 732 ; cette fois-ci c'est UMN026, le modèle avec le moins de flux actifs (722), et K-12 MG1655 (744) qui en ont le plus.
Sur les milieux minimum, les résultats sont identiques avec en moyenne 414 flux actifs. L-Les valeurs extrêmes sont proches également proches sur ces deux milieux avec comme valeurs minimales 409 (sur glucose pour ATCC 8739 et sur gluconate pour O157:H7 EC4115) et comme valeurs maximales 422 paires sur glucose (pour SE11) et 424 paires sur gluconate (pour K-12 W3110).
Si sur le milieu riche le nombre de flux actifs varie d'une quarantaine, ce chiffre tombe à une vingtaine sur les autres milieux. Peu importe le milieu, le nombre de flux actifs sur FBA est toujours compris entre le nombre de flux actifs sur FVA et le nombre de flux non nuls de la FVA. Cette constatation montre une des limites de la FBA qui prédit la meilleure distribution de flux mais qui ne prend pas en compte l'ensemble des capacités du modèle ou, à l'inverse, les flux qui sont essentiels (toujours non nul).
Ainsi en moyenne on peut observer :

- Sur milieu riche 660 flux sont essentiels et 232 flux sont dispensables.
- Sur milieu riche sans sucre 650 flux non nuls et 192 flux sont dispensables
- Sur milieu minimum glucose 305 flux sont essentiels et 110 flux sont dispensables
- Sur milieu minimum gluconate 295 flux sont essentiels et 139 flux sont dispensables

Sur milieux minimum, on constate très peu de variabilité entre les modèles en ce qui concerne les paires de flux toujours actifs : la moyenne et la valeur minimale sont identiques. Ces résultats indiquent qu'environ deux tiers des réactions, qui peuvent être utilisées sur ces milieux minimums, sont essentielles

La FBA donne une information supplémentaire : sur milieu minimum glucose, parmi les 110 réactions non essentielles, 109 sont utilisées pour obtenir le flux de biomasse optimal. Ces réactions actives dans certaines conditions reflètent la flexibilité du métabolisme et les voies alternatives. Il est important de noter que les modèles ne prennent pas en compte la régulation, qui peut dans certains cas bloquer des voies (e.g. la régulation de l'opéron lactose).

La seconde partie de cette analyse concerne la variabilité et les similitudes des flux entre les différents modèles sur un même milieu. Le premier milieu testé est le milieu riche, puisque c'est un milieu plus favorable à la production de biomasse. Les comparaisons des 23 modèles produisent 253 scores de similarité qui sont représentés sur la Figure 67.

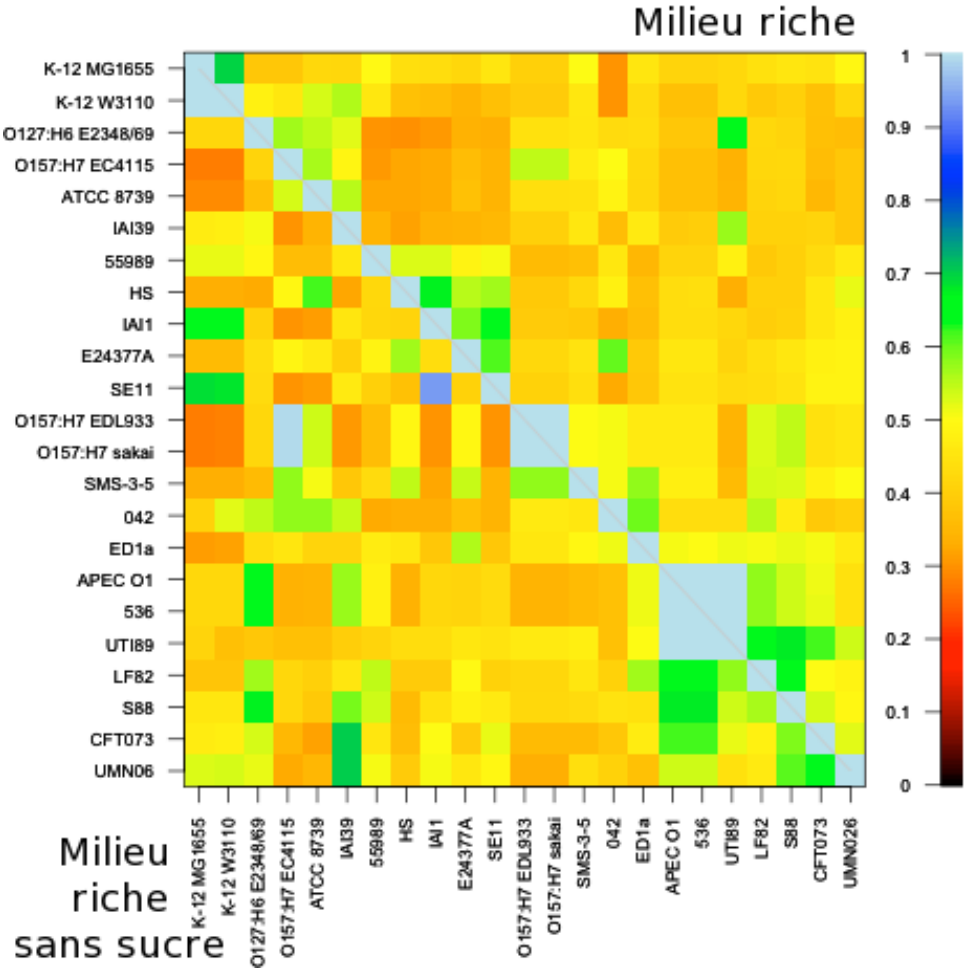


Figure 67 Score de similarité sur milieu riche et milieu riche sans sucre.
Les modèles sont organisés en fonction des similitudes sur milieu riche. Les comparaisons sur milieu riche sont dans le triangle supérieur droit et les comparaisons sur milieu riche sans sucre dans le triangle inférieur gauche.

On observe, sur cette figure, une forte dominante jaune associée à un taux de similarité compris entre 0.4 et 0.5, intervalle qui comprend à 49% des comparaisons. La moyenne est de 45.2% de similitude avec un score minimal de 0.3 et maximal de 1. La fréquence des scores est donnée dans la Table 32.

Intervalle de score	[0.1 ,0.2[	[0.2 ,0.3[	[0.3 ,0.4[	[0.4 ,0.5[	[0.5 ,0.6[	[0.6 ,0.7[	[0.7 ,0.8[	[0.8 ,0.9[	[0.9 ,1]

Occurrences	0	0	1	115	50	10	0	0	4

Table 32 : Occurrence des scores de similarité sur milieu riche.

On compte en moyenne 869 paires non nulles par comparaison de modèles (pour mémoire les réactions spécifiques d'un des deux modèles ne sont pas prises en compte). Ce nombre est du même ordre que le nombre de flux qui peuvent être actifs sur milieu riche (860). Il varie peu, de 842 (CFT073 versus K-12 W3110), à 889 (O157:H7 EC4115 versus O157:H7 Sakai et EDL933). Au total, il existe 916 paires non nulles, dont 907 sont similaires dans une des comparaisons.
Parmi ces paires similaires, 209 sont communes dans toutes les comparaisons : 89 sont des réactions de transport, 41 des réactions d'échange dont l'ensemble des échanges ferriques ; la majorité des 79 réactions restantes sont impliquées dans la synthèse des nucléotides, d'autres concernent la synthèse des cofacteurs, 26 sont impliquées dans le métabolisme alternatif du carbone, et 8 d'entre elles font parties de la glycolyse qui compte 22 réactions dans le modèle d'*iAF1260*.
Il existe 4 couples de modèles avec un score de similitude supérieur à 0.99 : (*536Cbm*, *APEC O1Cbm*), (*536Cbm*, *UTI89Cbm*), (*APEC O1Cbm*, *UTI89Cbm*) et (*O157:H7 EDL933*, *O157:H7 SakaiCbm*). Un fait remarquable concerne le lien entre les modèles des couples, ils appartiennent systématiquement au même groupe phylogénétique et au même type de pathogénicité ; ainsi *536Cbm*, *APEC O1Cbm* et *UTI89Cbm* sont du groupe phylogénétique B2 et sont des pathogènes extra-intestinal, et les souches O157:H7 sont du groupe phylogénétique E et sont des pathogènes intra-intestinal.
Un autre élément particulier concerne les trois modèles O157:H7 : ils ont exactement le même flux de biomasse sur FBA (27.873). Si le score de similarité entre O157:H7 Sakai et EDL933 est de 1, ce dernier descend à 0.545 lorsque l'on compare ces modèles à *O157:H7 EC4115Cbm*. Ce point est assez surprenant puisqu'il montre qu'à flux de biomasse optimal identique, la similitude de la variation des flux peut être identique ou être similaire à un peu plus de 50%.
Suite à cette observation, j'ai examiné le lien entre le flux de biomasse sur FBA et le score de similarité. La corrélation entre la différence du flux optimal de biomasse et le score de similarité est de -0.3 avec une p-value de 10^{-7}. Ceci montre encore une fois la différence entre l'optimisation de la production de biomasse et l'estimation des capacités métaboliques.
Sur milieu riche sans sucre, on observe une tendance plus orangée que sur milieu riche (Figure 67), ce qui se traduit par une moyenne de 45.8% de similitude entre les modèles reconstruits. Les fréquences des scores sont résumées dans la Table 33.

Intervalle de score	[0.1 ,0.2[	[0.2 ,0.3[	[0.3 ,0.4[	[0.4 ,0.5[	[0.5 ,0.6[	[0.6 ,0.7[	[0.7 ,0.8[	[0.8 ,0.9[	[0.9 ,1]
Occurrences	0	8	79	95	45	16	2	0	8

Table 33 Occurrence des scores de similarité sur milieu riche sans sucre.

Le nombre de paires de flux non nuls est de 909, ce qui est très proche du nombre obtenu sur milieu riche. Ce résultat n'est pas intuitif puisque sur ce milieu sans sucre, le nombre de sources de carbones est restreint.
Entre milieu riche et milieu riche sans sucre, 829 paires sont communes ; parmi les 80 paires spécifiques du milieu riche sans sucre, on dénombre 71% de transporteurs. Cet ensemble comprend aussi plusieurs réactions appartenant à la voie de dégradation du

glycolate et *glyoxylate*. Ces voies métaboliques sont des points d'entrées dans le cycle TCA et le cycle du glyoxylate.
On recense 880 paires de flux similaires sur au moins une des comparaisons, et 211 d'entre elles le sont sur toutes les comparaisons. Si ce denier chiffre est très proche de celui trouvé sur milieu riche, 47 réactions sont spécifiques du milieu sans sucre dont 24 transporteurs et 15 flux d'échanges, les 8 réactions restantes appartiennent au métabolisme alternatif du carbone.
On dénombre 18 couples de modèles avec un score de similarité supérieur à 0.65, dont 8 avec un score supérieur 0.9 ; ils sont présentés dans la Figure 68.

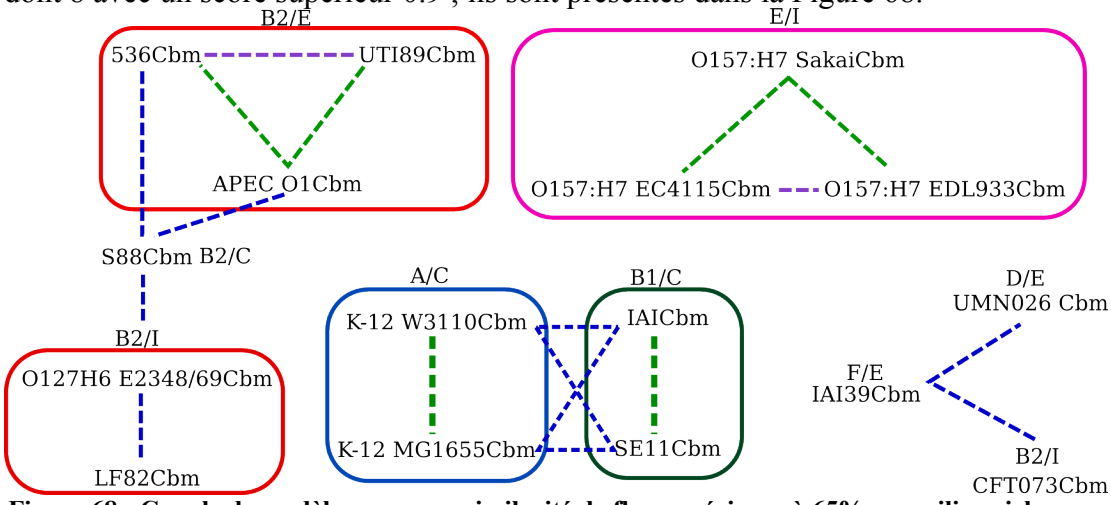


Figure 68 : Couple de modèles avec une similarité de flux supérieure à 65% sur milieu riche sans sucre.

Les couples x/y donnent le groupe phylogénétique (x), et le type de pathogénicité (y). Les lignes pointillées bleues indiquent un couple pour lequel le score est compris entre 0,65 et 0,9. Les lignes pointillées vertes indiquent les couples avec un score compris entre 0.9 et 0.99. Les lignes pointillées violettes sont les similitudes totales. Les ensembles de couples d'un même groupe phylogénétique et d'un même type de pathogénicité sont encadrés : en rouge pour les B2, violet pour les E, bleu pour les A et vert pour les B1.

On dénombre 6 couples pour lesquels les modèles n'appartiennent pas au même groupe phylogénétique ; 4 d'entre eux concernent des paires A/B1 qui du point de vue métabolique sont confondus (Chapitre I partie 1).
Si on considère que les souches des groupes phylogénétiques A et B1 forment un phylogroupe métabolique, alors on constate que parmi les 19 couples avec le plus grand score de similarité, 13 possèdent le même phylogroupe et le même comportement pathogène.
Parmi ces couples, nous retrouvons ceux des milieux riches, avec une différence importante chez les modèles des souches O157:H7. Les FBA pour ces trois souches donnent la même valeur pour le flux de biomasse, mais cette fois-ci, les 3 modèles ont des variations de flux similaires à plus de 99%.
Je me suis intéressé à cette différence de similarité entre milieu riche et milieu sans sucre pour les modèles des O157:H7. La principale raison est l'absence d'une réaction chez O157:H7 EC4115, la *dihydroxyacétone phosphotransférase* (DHAPT). Cette réaction intervient dans la dégradation du glycérol et produit du *phosphoénol pyruvate*, du *dihydroxyacétone phosphate* et du *pyruvate* qui sont trois métabolites impliqués dans la glycolyse. J'ai regardé les raisons de l'absence de cette réaction dans ce modèle. Chez *iAF1260,* la *DHAPT* possède une GPR composée uniquement d'une sous GPR comprenant cinq gènes ; chez MicroCyc le complexe associé à cette réaction ne comporte que 3 gènes qui sont inclus dans la GPR (réaction avec

l'identifiant 2.7.1.121-RXN). Les cinq gènes de la GPR ont des homologues chez Sakai et EDL933, chez EC4115 ; les 3 gènes communs à la GPR et au complexe du réseau sont manquants (Figure 69).

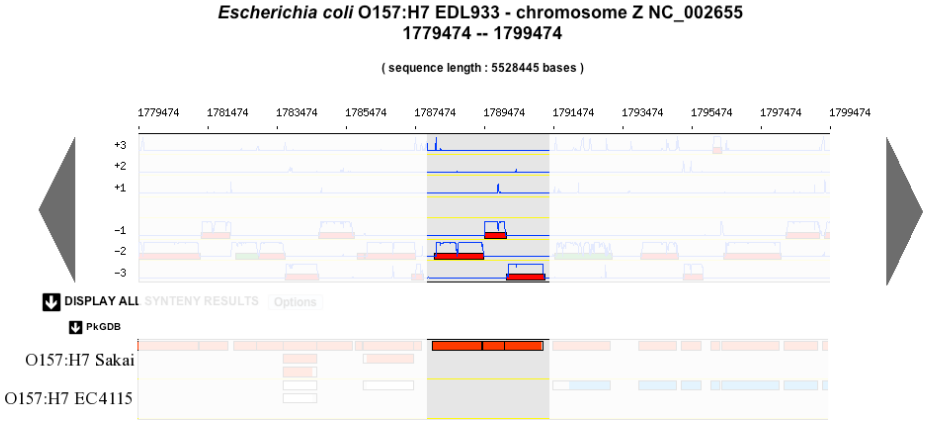


Figure 69 : Alignement de l'opéron dha.
Chez O157:H7 EDL933 et Sakai les 3 gènes *dhaH*, *dhaL* et *dhaK* sont présents. Chez O157:H7 EC4115 les trois gènes sont absents.

Cet exemple est intéressant, puisqu'il illustre l'importance des différences ponctuelles entre les modèles reconstruits et la difficulté des comparaisons des modèles; s'il existe, au final, peu de différences entre les modèles reconstruits, elles suffisent à induire des comportements distincts.
Cependant, suivant la manière d'étudier cette diversité, les conclusions peuvent être opposées ; ainsi on peut passer d'un résultat où les modèles sont strictement identiques à des résultats donnant ces mêmes modèles différents à 75%. Je reviendrai ultérieurement sur le problème de comparaison des modèles.
Les observations sur milieu riche et riche sans sucre montrent que les nombres de réactions qui peuvent être utilisées sont assez proches, mais le manque de sucre augmente le flux des transporteurs, flux qui deviennent égaux à la borne imposée dans la simulation.
Ces comparaisons ont mis en évidence un ensemble de réactions liées au métabolisme du carbone, que ce soit la glycolyse/glugogénèse et le métabolisme alternatif du carbone, (dont les variations sont identiques dans tous les modèles) sur milieu riche ou milieu riche sans sucre.

Passons maintenant à l'analyse des milieux simples, en commençant par le milieu minimum glucose. La comparaison entre *iAF1260* et *K-12 MG1655Cbm* avait donné un flux optimal de biomasse identique et un score de similitude de 1. On dénombre 376 paires de flux non nuls. Ces 376 paires sont similaires dans au moins une des comparaisons ; par contre seulement 255 paires sont similaires sur toutes les comparaisons. L'ensemble des scores est résumé dans le triangle supérieur gauche de la Figure 70, sur laquelle on observe une forte dominante bleue, c'est à dire des scores de similitude compris entre 0.8 et 1. En moyenne, les flux sont similaires à 91.25%, et les scores varient entre 0.71 pour le couple (O157:H7EC4115, UTI89) et 1 pour une vingtaine de couples. La fréquence des scores est donnée dans la Table 34.

Intervalle de score	[0.7 ,0.75[	[0.75 ,0.8[	[0.8 0,.85[	[0.85 ,0.9[	[0.9 ,0.95[	[0.95 ,1[	1
Occurrences	16	0	59	45	0	110	23

Table 34 Occurrence des scores de similarité sur minimum glucose.

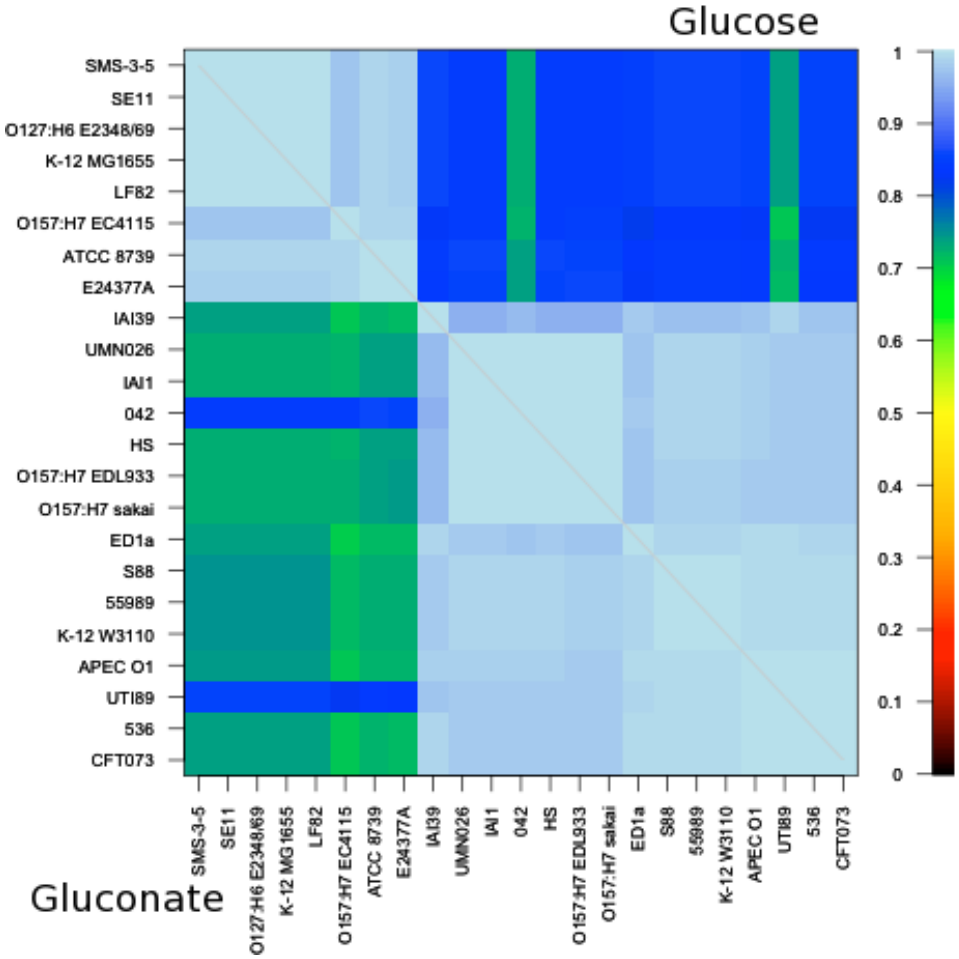


Figure 70 : Score de similarité sur milieu minimum glucose et gluconate.
Les modèles sont organisés en fonction des similitudes sur milieu minimum glucose. Les comparaisons sur milieu minimum glucose sont dans le triangle supérieur droit et les comparaisons sur milieu minimum gluconate dans le triangle inférieur gauche.

L'analyse de la Table 34 donne trois tendances distinctes. La première, t1, avec des scores compris entre 0.7 et 0.75 concerne les modèles *042Cbm* et *UTI89Cbm*. La deuxième, t2, compris entre 0.8 et 0.9 concerne 13 modèles : *IAI39Cbm*, *UMN026Cbm*, *IAI1Cbm*, *HSCbm*, *O157:H7 EDL933Cbm*, *O157:H7 SakaiCbm*, *ED1aCbm*, *S88Cbm*, *55989Cbm*, *K-12 W3110Cbm APEC O1Cbm*, *536CBM* et *CFT073Cbm*). Enfin la dernière, t3, est supérieure à 0.95 et comprend les 8 autres modèles : *SMS-3-5Cbm*, *SE11Cbm*, *O127:H6 E2348/69Cbm K-12 MG1655Cbm*, *LF82Cbm*, *O157:H7 EC4115Cbm*, *ATCC 8739Cbm* et *E24377ACbm*. Sur la Figure 70, on observe que les scores entre les modèles du groupe t1 et t2, sont identiques aux scores des comparaisons entre les modèles du groupe t1 ou du groupe t2. A l'inverse, les comparaisons entre modèles de t1 et de t3 donnent des scores beaucoup plus faibles que lors des comparaisons t2/t3.
A partir de cette observation, j'ai recherché les différences de paires similaires parmi ces trois ensembles en me focalisant sur les réactions communes à t1 et t2, et absentes de t3. Le métabolisme central de *E. coli* est bien connu, et j'ai supposé que les modèles le représentaient au mieux. C'est pourquoi j'ai regardé en priorité les réactions ajoutées par le module *cycSpe*. Parmi les paires différentes entre t1, t2 et t3,

on trouve des transporteurs périplasmiques (j'ai déjà discuté des raisons de ces absences dans la partie 4.2.3) ; cependant la similarité de ces paires est très variable entre les comparaisons et ne suit pas la composition des trois ensembles.
Une réaction est absente de tous les modèles de l'ensemble t3 : il s'agit d'une *formate déhydrogénase* avec l'identifiant 1.2.1.2-RXN dans MicroCyc ; cette réaction transforme le *formate* en *dioxyde de carbone* en utilisant le couple d'oxydo-réducteur *NAD/NADH*. Cette réaction ajoutée aux modèles par le module *cycSpe*, entraîne des modifications des bornes de flux ; d'autres réactions dont *PFL* et *FHL* produisent ou consomment du *formate.*
Une autre réaction est absente du groupe t1 ainsi que de certains modèles des groupes t2 et t3. Il s'agit de l'*acétate kinase,* avec comme identifiant MicroCyc *RXN-7985*. Cette réaction utilise les mêmes substrats et donne les mêmes produits que la réaction *PPAKr* d'*iAF1260* (Figure 71). La réaction issue de MicroCyc est réversible tandis que celle d'*iAF1260* est à sens unique. De plus ces deux réactions sont associées à des gènes différents.

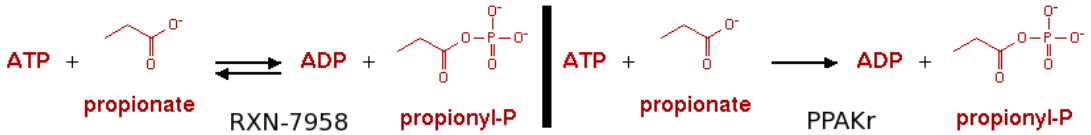


Figure 71 : Différence entre RXN-7958 et PPAKr
Il existe deux différences entre ces deux réactions : la première est la réversibilité et la deuxième la GPR.

Une dernière réaction importée par *cycSpe* a été mise en évidence: il s'agit de la *L-phénylalanine aminotransferase*, avec l'identifiant MicroCyc *RXN-10814*. Cette réaction est présente dans certains modèles de l'ensemble t1 et t2 et influe sur les réactions utilisant le phényle, en particulier la réaction d'*iAF120* : *PHETA1*. Comme pour le couple *PPAKr/RXN-7958*, la réaction issue de MicroCyc est irréversible tandis que celle issue d'*iAF1260* est réversible, et les gènes associés sont différents.
Parmi les paires similaires dans toutes les comparaisons, on retrouve sans surprise les flux d'échanges des métabolites qui composent le milieu minimum, mais aussi un quart des réactions de la glycolyse. On constate aussi une forte représentation des réactions liées aux acides aminés avec la totalité de celles impliquées dans le métabolisme de l'histidine, la moitié de celles du métabolisme de la thréonine et de la tyrosine, les deux tiers de celles impliquées dans le métabolisme de la cystéine et de la lysine, et trois cinquièmes des réactions du métabolisme de la valine, leucine et isoleucine. Enfin le métabolisme des bases puriques et pyrimidiques est aussi fortement représenté avec 20 des 24 réactions.
Ce résultat était attendu puisque les flux optimaux de biomasse sont identiques, ce qui impose des flux de création des composants de la biomasse identiques (acides aminés, nucléotides etc.).
La dernière comparaison porte sur le milieu minimum gluconate, sur lequel on dénombre 396 paires non nulles, et 375 d'entre elles sont présentes dans tous les modèles. Toutes ces paires non nulles sont similaires dans au moins une comparaison et 313 le sont dans toutes les comparaisons.
Sur la Figure 70 on peut observer une dominance bleue claire et une tendance verte. Ceci indique qu'une partie des scores est très élevée (>0.9) tandis que l'autre partie est proche des 0.65.
Puisque la FBA donne un flux de biomasse identique pour tous les modèles, je m'attendais à un score moyen de similitude proche de celui obtenu sur milieu

minimum glucose ; celui-ci est en fait bien plus faible : 0.83. Pourtant les valeurs maximales et minimales sont identiques dans les deux cas (respectivement 0.7 et 1).

Intervalle de score	[0.7 ,0.75[	[0.75 ,0.8[	[0.8 0,.85[	[0.85 ,0.9[	[0.9 ,0.95[	[0.9 ,9.5[	1
Occurrences	104	0	9	7	0	110	23

Table 35 Occurrence des scores de similarité sur minimum gluconate.

A l'instar du glucose, la distribution des fréquences des scores donne trois tendances (Table 35). On retrouve les trois groupes t1, t2 et t3 composés des mêmes modèles. Les scores entre modèles de t1 et t2 sont toujours identiques aux scores des comparaisons intra t1 ou intra t2 ; par contre les scores des comparaisons t2/t3 sont cette fois-ci inférieurs aux scores des comparaisons t1/t3. Cette observation m'a laissé supposer que les réactions responsables de la différence sur milieu minimum glucose doivent être les mêmes ; cependant la différence entre les scores des comparaisons inter-groupes laisse présager l'existence d'autres réactions responsables de ces différences.
Une réaction, qui a pour identifiant MicroCyc *TYROSINE-AMINOTRANSFERASE-RXN,* a été mise en évidence. Elle utilise les mêmes métabolites que la réaction *TYRTA* d'*iAF1260* et il s'agit là aussi d'une différence de GPR et de réversibilité.
Les paires des réactions toujours similaires sur milieu minimum gluconate sont pratiquement les mêmes que sur milieu minimum glucose. On retrouve des proportions identiques des réactions des différents processus métaboliques (métabolismes des acides aminés, des bases puriques et pyrimidiques etc.). Au niveau des paires de réactions différentes, on note bien sûr l'absence de toute la chaîne du transport du glucose, du flux d'échange aux premières étapes de la glycolyse, en passant par le système *PTS.* A l'inverse les paires similaires spécifiques du gluconate concernent l'importation de celui-ci (flux d'échange et transporteurs), et sa dégradation en *6-phospho-D-gluconate*. A cela s'ajoutent des réactions appartenant à des voies de dégradation où le gluconate est un produit intermédiaire : voie de dégradation de l'idonate et la voie de dégradation oxydative de glucose.

Il existe 59 paires de réactions qui sont similaires sur l'ensemble des couples de modèles et sur les quatre environnements (Table 36).

Processus	Réactions
Métabolisme de l'alanine et l'aspartate	2
Métabolisme alternatif du carbone	7
Cycle de l'acide citrique	1
Métabolisme des cofacteurs	2
Métabolisme du folate	2
Glycolyse/Gluconéogénèse	6
Transport et métabolisme des ions inorganique	1
Biosynthèse et recyclage des lipopolysaccharides	3
Métabolisme des membranes lipidique	1
Métabolisme des nucléotides	11
Oxydation phosphorylation	3
Transport, membrane interne	17
Transport, membrane externe	1
Transport, porine membrane externe	2

Table 36 : Processus et nombre de réactions des paires similaires sur toutes les expériences

Ce cœur de réactions qui présente les mêmes valeurs extrêmes, est composé pour un tiers des transporteurs, et pour un sixième de réactions de synthèse des nucléotides. Un dixième des réactions sont liées à la glycolyse ; cela peut paraître surprenant sur milieu minimum gluconate, cependant en regardant ces réactions, on constate qu'il s'agit de la dégradation du glycogène qui permet d'obtenir du glucose. Les transporteurs périplasmiques de ce dernier sont également présents dans cet ensemble de réactions, ce qui est normal sur milieu gluconate puisque le glucose périplasmique permet de faire rentrer le gluconate dans le cytoplasme..

La comparaison de la variation des flux permet d'aller au-delà de la simple comparaison du flux optimal obtenu par FBA. Elle permet de voir l'impact local des différences de composition des réseaux.
Ainsi, alors que la FBA, sur milieu minium donne le même résultat sur tous les modèles, on se rend compte que les espaces de solutions sont différents. Cette constatation implique trois remarques importantes. Premièrement, elle montre que la diversité récupérée par le processus de reconstruction est exploitable et modifie l'espace des solutions. Deuxièmement, que la solution de la FBA n'est pas une valeur représentative lorsque l'on veut tester les capacités des modèles ; deux flux de biomasse identiques en FBA ne permettent pas de conclure sur la similitude des capacités des modèles, mais simplement de voir si au mieux les modèles produisent de la biomasse à la même vitesse. De plus, l'analyse a montré qu'il n'existe pas de lien entre proximité de résultat FBA et similitude de la variabilité des flux. Enfin le dernier point concerne l'importance des différences ponctuelles : si du point de vue global, l'ajout d'une ou de quelques réactions n'a qu'un impact limité en amont et en aval de ces réactions, la flexibilité des flux est totalement modifiée. Ce point montre une des limites de l'analyse de la variabilité : si l'espace des solutions est différent entre deux modèles, cela ne signifie pas que les deux modèles utilisent des capacités différentes. Cependant avec les contraintes actuelles, il est impossible de réduire ces espaces de solutions et, par conséquence, de préciser l'état métabolique du modèle.

4.2.4 Bilan de la diversité des modèles

La méthodologie de reconstruction des modèles du métabolisme à l'échelle de la cellule est très proche de celle utilisée pour la reconstruction des réseaux métaboliques. Cependant les différences entre modèles et réseaux rendent l'exercice beaucoup plus délicat. S'il est possible d'intégrer dans les réseaux des éléments mal définis, il est impossible de les utiliser dans les modèles. Malheureusement, c'est parmi ces éléments que se trouve une grande partie de la diversité métabolique. Cet effet est visible sur le ratio du core et pan métabolisme ; celui-ci est de 57% pour les réseaux et de 86% pour les modèles. Cette différence de 30% est évidemment non négligeable, mais elle montre surtout le fossé entre ce qui y est connu, et ce qui est «parfaitement » connu.
Ce n'est pas pour autant que les modèles reconstruits vont être des « clones» du pivot, simplement la diversité qu'ils vont refléter n'est qu'une partie de la diversité réelle. Cette nuance de définition des connaissances n'est pas spécifique du cadre de modélisation ; c'est une limite qui s'impose à tous les types de modèles. Néanmoins l'accroissement des éléments biologiques bien définis permet aujourd'hui de travailler à l'échelle de la cellule avec des modèles biologiquement pertinents. Tous les

résultats obtenus dans cette partie ont été interprété en prenant en considération une diversité certes limitée, mais qui est fondée sur des éléments précis.
La base fonctionnelle des modèles est la même pour tous : il s'agit d'*iAF1260*. Lorsque l'on compare ce modèle aux modèles reconstruits, par l'intermédiaire du module *cbmCom*, on se rend compte qu'en moyenne 98.2% des réactions du pivot se retrouvent dans les modèles. Si le fait de travailler au sein de la même espèce explique un pourcentage élevé, ce dernier reste largement supérieur à celui attendu. Pour rappel dans les modèles, on descend en dessous de 80% de similitude. La raison principale de cette différence est le biais de connaissance évoqué au début de cette partie. Sans remettre en question la qualité du modèle pivot, qui est très bonne, on se rend compte que ce modèle synthétise les connaissances sur les réactions et le processus biologiques les plus connus, c'est pourquoi il subit régulièrement des mises à jour pour ajouter des processus moins répandus (Feist et al. 2007; Orth et al. 2011).
La souche K-12 W3110 est un clone métabolique de la souche K-12 MG1655 et donc d'*iAF1260* : le modèle de K-12 W3110 devrait être un clone du pivot. Pourtant *K-12 W3110Cbm* contient 22 réactions spécifiques; si cela ne représente que 0.9% des réactions, ce chiffre correspond pourtant à la moitié des réactions du pivot qui ne sont pas retrouvées dans les différents modèles.
L'une des valeurs ajoutées du processus de reconstruction provient du module *cycSpe*, puisqu'il est en charge de l'ajout des nouvelles réactions. En comptant les réactions d'échange, 8% des réactions des modèles ont été créé par celui-ci. Les modèles reconstruits sont capables de produire de la biomasse sur milieu riche et milieu minimum glucose, et aussi sur plus de 170 sources de carbone différentes. Les observations donnent les modèles de K-12 comme ceux qui utilisent le plus de sources différentes (une dizaine de source en plus). J'ai montré qu'il s'agissait d'un artefact de reconstruction qui malheureusement ne peut être corrigé sans preuve expérimentale. En corrigeant manuellement celui-ci, on a constaté que les modèles de K-12 sont finalement sous la moyenne du nombre de source de carbones utilisables par les modèles. Ce constat prouve que la diversité ajoutée à nos réseaux est fonctionnelle et apporte de nouvelles capacités.
L'un des points de comparaison des modèles est évidemment le flux de biomasse. Sur milieu riche, le modèle de K-12 MG1655 produit un flux de biomasse supérieur au pivot ; dans le meilleur des cas, les autres modèles égalent cette valeur, mais ils produiront majoritairement un flux inférieur. Là encore l'artefact de reconstruction et le biais de connaissance sont responsables de ces résultats. Sur milieu minimum glucose, les flux de biomasse sont identiques, ce qui implique que la diversité rajoutée n'a pas apporté de nouvelles réactions améliorant le rendement de conversion du glucose. L'analyse des résultats de FBA donne finalement peu d'information, et je me suis intéressé à une autre analyse fréquente des CBMs : la FVA.
Si le nombre de réactions différentes est faible, leurs impacts n'est pas négligeable. La variabilité des flux est évidemment sensible au milieu de simulation ; si les milieux de même nature (riche ou minimum) ont tendance à diminuer la variabilité entre les paires de flux, les comparaisons entre des milieux de différentes natures permettent de définir un cœur de réactions qui varie toujours dans le même intervalle.
L'analyse des FVA a montré les différences entre comportement local et global. Ainsi, si deux modèles ont des flux de biomasse optimaux identiques, cela ne signifie pas que leurs flux varient de la même manière. Une réaction en plus ou en moins peut avoir un effet local important en divisant le flux de matière, mais au niveau global, cette variation se traduira par une faible différence dans le score de similarité. L'effet de la présence ou absence d'une réaction est difficilement prévisible, elle dépend des

métabolites mis en jeu, et des réactions capables d'effectuer les mêmes transformations de substrat mais avec des cofacteurs différents. Evidemment le milieu de simulation joue également un rôle important.
Trois critères ont été considéré pour estimer la diversité des modèles : la composition en réaction, le flux de biomasse sur FBA et le score de similarité des variations des flux. On se rend compte que suivant le critère pris en compte, les conclusions concernant la similarité ou les différences entre modèles peuvent être différentes. Dans un cas les résultats peuvent donner une similitude complète des modèles ou, au contraire, une différence totale. Cela n'implique pas que l'un des critères est erroné, mais simplement que chaque critère répond à une question différente ; si ce qui m'intéresse est la production maximale d'un flux, savoir que les deux modèles n'ont pas le même nombre de réactions, ou que leurs flux ne varient pas de la même manière est inutile.
Puisqu'il n'existe pas de jeu de modèles identiques, il n'existe pas non plus de méthodes de comparaison de plusieurs modèles. Je me suis limité aux trois critères énoncés précédemment, car ils sont simples, facilement interprétables et ils montrent divers aspects des capacités métaboliques des modèles reconstruits.
La comparaison de ces capacités montre l'existence possible de différents modes de fonctionnement du métabolisme, mais dans l'état actuel il est impossible de restreindre suffisamment l'espace des solutions pour voir si deux modèles utilisent réellement le même mode de fonctionnement. Pour restreindre l'espace des solutions, il faut ajouter de nouvelles contraintes, qui cette fois-ci, ne soient pas issues de la structure du modèle ni de la conservation de la matière, mais qui proviennent d'observations réalisées sur les différentes souches de l'étude.

5 Conclusions

Les réseaux permettent d'estimer la capacité de croissance d'un organisme sur un milieu : si des voies de dégradation des composés de l'environnement, ou si un des métabolites précurseurs de biomasse est absent alors on peut dire que l'organisme n'est pas adapté à cet environnement. Dans le cas contraire, la conclusion possible avec les réseaux est : l'organisme peut vivre dans ce milieu. Le passage de « peu produire de la biomasse » à « produit un flux de biomasse d'une valeur de » est important et nécessite un travail de conversion du réseau en modèle mathématique. L'apport de la modélisation au métabolisme permet d'estimer le flux de matière qui traverse les réactions et ainsi en déduire les réactions actives et inactives. Au final on est capable de prédire si un modèle produit de la biomasse dans tel ou tel environnement. Si le réseau métabolique et le modèle métabolique sont des représentations d'une même réalité biologique, ils sont différents sur de nombreux aspects.
Je suis parti de cette observation pour concevoir le processus de reconstruction des modèles. Ainsi, il doit être identique à celui de la reconstruction des réseaux, mais chaque module du processus aura une nature et un rôle différents. J'ai donc repris le concept du *pivot*, en introduisant le modèle de référence *iAF1260*. Si dans la reconstruction des réseaux le pivot sert à identifier les réactions catalysées via la recherche de gènes homologues, dans le modèle, le pivot sert de base fonctionnelle, c'est à dire qu'en plus des réactions avec gènes, le processus récupère également les réactions (biologiques ou artificielles) nécessaires au bon fonctionnement des modèles. La partie spécifique des réseaux est apportée par les bases de données métaboliques généralistes ; malheureusement il n'existe pas d'équivalent en

modélisation. Je me suis servi des réseaux reconstruits et de leurs différences de composition pour créer artificiellement une base de données généraliste concernant les modèles métaboliques. Enfin, dans la reconstruction des réseaux une dernière étape d'intégration des complexes avait lieu : pour les modèles cette dernière étape consiste à vérifier l'ensemble des réactions et la cohérence interne du modèle.
Le processus est relativement simple, mais sa mise en œuvre est des plus complexes. Puisque le réseau et le modèle représentent la même chose, il existe des éléments communs. Les identifier devrait être simple ; en réalité cette identification nécessite un travail et un temps conséquent. Un des freins de l'automatisation des processus en biologie est le manque d'homogénéisation des données ; chaque ressource utilise sa propre nomenclature et il est souvent plus rentable de partir de zéro que d'essayer de relier plusieurs bases de données. Néanmoins pour avoir un suivi du génome au modèle, j'ai décidé d'unir deux ressources différentes. Ce travail a permis de relier 94.5% des métabolites du modèle aux métabolites des réseaux. Pour les réactions, le score est plus faible avec quand même 80.1% d'associations.
Durant les différentes étapes d'estimation des liens entre ces deux univers, il est apparu que plus une réaction ou un métabolite est référencé dans différente ressource, plus il possède de descriptions (nom, formule etc.) et plus le lien est facile à établir. On peut séparer ces éléments en trois catégories :

i) Ceux qui ne possèdent aucun des critères définis précisément.

ii) Ceux définis par un des critères et qui contiennent des informations moins détaillées pour les autres: par exemple un nom et un *numéro EC* incomplet.

iii) Ceux bien définis qui possèdent un nom unique, un identifiant de type *numéro EC* ou *CAS*, une formule chimique ou une équation bilan équilibrée etc.

Lors de la recherche des réactions spécifiques des modèles, j'ai dû convertir des réactions et des métabolites provenant des réseaux, en réactions compatibles avec les modèles. Lorsque ces éléments appartiennent à la catégorie iii) la conversion se fait sans ambigüité. Pour les réactions et métabolites de la catégorie i) le passage du réseau au modèle est, à quelques exceptions près, impossible. La catégorie ii) a donné lieu à une nouvelle expertise manuelle pour identifier les éléments qui peuvent être intégrés aux modèles. Le temps consacré à identifier, filtrer et unifier les données est actuellement l'un des facteurs qui bloque le passage de la reconstruction des réseaux et des modèles à haut débit. Bien qu'aujourd'hui à chaque mise à jour des bases de données le nombre de références-croisées augmente, celles-ci ne concernent que les nouveaux éléments et le retard pris ces dernières années est comblé par ajouts ponctuels dans le cadre d'études précises, comme c'est le cas dans mes travaux.
La reconstruction des 23 modèles est relativement rapide comparée à une reconstruction *« from scratch »*. Les modèles obtenus contiennent un nombre de métabolites et un nombre de réactions assez proches. A l'instar des réseaux, il existe un biais dû au pivot, mais l'absence de connaissance sur les autres modèles rend impossible son estimation ; on note tout de même que 94.5% des réactions du pivot sont présentes dans tous les modèles. La proportion de réactions communes au sein des modèles est trop importante comparée à ce que j'ai observé sur les réseaux ; cette constatation reflète un problème de connaissance. Pour les réseaux métaboliques on a constaté qu'à partir de 66 réseaux le nombre de nouvelles réactions passe en dessous

de 1 pour tout nouveau réseau ajouté ; pour les modèles, à 23 on est déjà proche de cette limite. Cette saturation de la diversité est le second goulot d'étranglement de la reconstruction des modèles du métabolisme.
La méthodologie mise en place dans mes travaux est capable de détecter et récupérer la diversité disponible à l'heure actuelle dans les bases de données dédiées aux CBMs. Le problème vient de l'alimentation de leur contenu. Le niveau de définition des éléments des CBMs impose pour le moment une étude au moins semie-manuelle, avant d'intégrer un nouvel élément. Il existe des projets qui ont pour objectif d'améliorer et de désengorger la reconstruction des CBMs. Je pense en particulier au projet européen Microme[1] qui reprend les éléments mis en place dans ces travaux, c'est à dire une prédiction automatique à laquelle il est possible d'ajouter des informations manuelles. Le travail de croisement des bases de données que j'ai effectué a d'ailleurs été utilisé comme ressource lors de la curation et de la définition des éléments constituants cette nouvelle base de données.
Bien que la diversité des modèles soit sous-évaluée, elle existe quand même. En effet, seulement trois modèles sont plus petits que le pivot, et la taille de ce dernier est largement inférieure à la moyenne du nombre de réactions par modèle. Pourtant ce sont bien les modèles de K-12 qui possèdent les meilleurs rendements sur milieu riche. Si il est fort probable que cela soit dû à une lacune en transporteurs périplasmiques, je suis incapable, avec les données actuelles, de valider ou infirmer cette hypothèse. Tout au long des comparaisons, les transporteurs ont été mis en évidence, ce qui est normal puisqu'ils sont les entrées et sorties des modèles. Ils restent un point sensible des modèles pour plusieurs raisons. Tout d'abord, l'effort d'identification des transporteurs a été produit sur les transporteurs des métabolites essentiels, laissant tous les transporteurs de métabolites secondaires moins bien définis. Ensuite, si certains transporteurs sont spécifiques, d'autres peuvent être multi-substrats, ce qui peut être contourné par l'ajout de réactions à la manière de l'implémentation des réactions génériques, à condition de connaître les différents substrats. Enfin on sait qu'il existe différents mécanismes de régulation des transporteurs qui ne sont pas pris en compte dans le cadre de modélisation.
Pour étudier la diversité, et comme il n'existe pas de travaux équivalents, j'ai définis trois critères : si le premier, le nombre de réactions, est intuitif, il n'est pas forcément le plus approprié.

	Contenu	FBA	Similarité
Milieu riche	Oui	Oui	Forte
Milieu riche sans sucre	Oui	Oui	Forte
Glucose	Oui	Non	Faible
Gluconate	Oui	Non	Faible

Table 37 : Conclusion des différentes analyses sur la diversité.
Chaque colonne répond à la question : les modèles sont ils différents ? Cette question générale n'est pas pertinente ; la première colonne répond à la question : les modèles ont ils des réactions différentes ? La deuxième répond à la question : dans les meilleures conditions est-ce que certains modèles ont un rendement plus élevé ? Enfin la dernière colonne répond à la question : est-ce que les flux varient de la même façon entre deux modèles.

Chacun des critères répond à une question précise et doit donc être interprété en fonction de celle-ci ; par conséquent les résultats des différents critères ne peuvent être comparés entre eux, et il ne faut pas être étonné si les résultats paraissent contradictoires (Table 37).

[1] www.microme.eu

La première question est : le contenu du modèle est il le même pour tous ? Les modèles, à une exception près (*O157:H7 SakaiCbm* et *EDL933Cbm*), sont tous différents les uns des autres et ne sont pas des copies du modèle pivot.
La deuxième question concerne les capacités de production des modèles : sont elles les mêmes pour tous les modèles ? J'ai introduit quatre milieux, deux simples et deux complexes pour pouvoir répondre à cette question. Il en résulte que sur milieu simple il est impossible de différencier les différents modèles. Sur milieu riche on observe des flux de biomasse avec des valeurs différentes ; ce point est important puisqu'il prouve que la diversité introduite dans les modèles reconstruits est fonctionnelle et impacte la production de biomasse.
La dernière question porte sur la variabilité des flux et des variations similaires entre les modèles : la variabilité des flux est elle la même entre deux milieux ou deux modèles ? Le but de cette analyse est de prolonger la FBA, qui donne une distribution optimale, mais qui ne reflète pas les capacités métaboliques et les régimes de fonctionnement du métabolisme. Le premier résultat de cette analyse montre la prédominance du milieu sur les similitudes ; peu importe le modèle, les scores de comparaison entre deux milieux différents sont pratiquement identiques.
Aucun lien entre la valeur de flux optimal de biomasse et la phylogénie ou la pathogénicité n'a été trouvé. Si, sur FBA, il était impossible de différencier les modèles sur milieu simple, l'analyse de la similarité des flux a mis en évidence trois ensembles de modèle sur ces milieux. Ces différences sont locales et concernent un nombre très restreint de réactions.
Sur milieu riche sans sucre, on a constaté que les couples de modèles avec le plus grand score de similitude sont des couples partageant le même groupe phylogénétique et le même type de pathogénicité. C'est d'ailleurs la première fois dans mes travaux que l'on note un lien entre pathogénicité et métabolisme.
Cette analyse a également mis en avant les inégalités d'effet de présence ou d'absence d'une seule réaction. Sur le milieu minimum glucose on a observé qu'une réaction faisait varier de moins d'1% le score de similarité puisqu'il existe des réactions ayant le même rôle ; sur milieu riche sans sucre, une seule et unique réaction est responsable d'une division par deux du score de similitude passant de 1 à 0.53.

Les critères que j'ai choisis ne sont pas les seuls possibles, cependant le but de mes travaux n'est pas de comparer les différents critères de la diversité des modèles métaboliques, raison pour laquelle je me suis limité à ces trois critères. On peut imaginer de nombreuses autres façons d'évaluer cette diversité. Par exemple, s'intéresser à la fréquence des valeurs au sein d'un intervalle de flux ; la FVA donne la borne minimale et maximale, mais il est peu probable que la valeur réelle du flux soit distribuée de manière uniforme dans cet intervalle. Si deux flux ont le même intervalle de variation, mais si le premier à tendance à être proche de la borne minimale et le deuxième proche de la borne maximale, peut-on dire que ces flux sont similaires ?
La méthodologie de reconstruction automatique des modèles, couplée à celle des réseaux, a montré sa capacité à reconstruire des modèles de haute qualité permettant de rendre compte de la diversité métabolique, observée et compatible avec les CBMs disponibles à l'heure actuelle. Les différentes analyses et comparaisons effectuées prouvent que les comportements des modèles sont bien différents et que chaque modèle possède son propre espace de solutions caractéristique des capacités métaboliques de l'organisme modélisé. Il est impossible pour le moment de dire si nos modèles utilisent des capacités différentes ou si ils utilisent les capacités

communes; les données issues uniquement des génomes annotés et des bases de données métaboliques ne permettent pas de réduire suffisamment l'espace des solutions. C'est pourquoi il est important d'introduire des données et résultats expérimentaux comme contraintes.

Chapitre III : Intégration de données hétérogènes.

La biologie des systèmes met en relation des résultats expérimentaux de types -*omiques* (fluxomique, protéomique, génomique, transcriptomique, métabolomique, etc.). L'intégration de ces données est un exercice délicat et complexe lorsque l'on travaille à l'échelle de la cellule. Néanmoins, la capacité des modèles à bases de contraintes (CBMs) à utiliser différentes natures de données, en fait un cadre d'intégration particulièrement bien adapté.
En théorie n'importe quel type de données peut être utilisé comme facteur limitant des flux des CBMs ; en pratique toutes les observations ne sont pas exploitables. Il existe deux raisons principales au manque d'utilisation de données biologiques dans les CBMs.
Premièrement, une limite est tout simplement liée à la disponibilité de données expérimentales. Si pour la souche de référence *E. coli* K-12 MG1655 et d'autres organismes modèles, il existe des jeux de données concernant la plupart des aspects biologiques (moléculaires, biochimiques et physiologiques). Pour les autres souches, il existe au mieux quelques jeux de données spécifiques. Un ensemble commun à plusieurs souches est extrêmement rare.
La seconde limite est d'ordre technique. Si certaines données sont faciles à intégrer, d'autres nécessitent un travail d'adaptation et de formalisme. Par exemple, les prédictions de croissances sur des mutants donnent des résultats simples (viable, pas viable) qui peuvent être reliés aux flux de biomasse d'une manière booléenne (flux non nul ou nul). Les données de concentration des métabolites ne sont pas utilisées dans les CBMs ; elles sont supposées constantes et n'interviennent pas dans le calcul du résultat. Pourtant, par l'intermédiaire de la thermodynamique, ces concentrations servent à contraindre le sens des réactions, en calculant l'entropie des métabolites, puis de la réaction pour en déduire son sens. Que l'intégration soit directe ou indirecte, il est nécessaire de convertir l'observation en contrainte. Dans le cas des mutants, cela consiste à contraindre les flux des réactions pour lesquelles le gène est essentiel à zéro. Dans le cas des concentrations cela consiste à rendre les valeurs de flux associées aux réactions positives ou négatives suivant le sens défini par le calcul thermodynamique.

1 MetaColi

Les études réalisées durant cette thèse sont exclusivement *in silico*, mais reposent sur des données *in vitro*, fournies par l'intermédiaire d'une collaboration au sein de l'ANR MetaColi.
Les objectifs à long terme de l'ANR MetaColi dépassent le cadre de mes travaux ; néanmoins ces derniers font partis des bases indispensables pour les atteindre. Le projet cherche à apporter une meilleure vision de l'adaptation du métabolisme des *E. coli* en s'intéressant à un nombre restreint de souches d'origine phylogénétique variée et de pathovars différents. Cette compréhension n'est possible que par l'intégration de données biologiques hétérogènes au sein de modèles métaboliques fonctionnels. Ces modèles sont de type global avec les CBMs, et local avec des modèles cinétiques du métabolisme central de *E. coli* dont les acteurs sont présents dans l'ensemble des souches. Le projet souhaite proposer de nouvelles méthodologies permettant d'intégrer des données expérimentales à différents niveaux du métabolisme. Cette

façon de procéder permet de prendre en compte les différences qualitatives, avec la présence ou absence des réactions, et les différences quantitatives, avec les flux de matière qui traversent chacune des réactions. Mon rôle au sein du projet est d'apporter les réseaux et les modèles métaboliques des différentes souches à l'échelle de la cellule, et de préparer ces modèles pour l'intégration de l'ensemble des ressources dans un cadre de simulation probabiliste.

2 Les données expérimentales

.2.1 Les souches

Le projet MetaColi se focalise sur 5 souches différentes qui sont réparties dans quatre groupes phylogénétiques (Table 38), et deux pathovars. Parmi les souches, on retrouve K-12 MG1655 la souche modèle isolée chez l'Homme il y a près d'un siècle ; c'est une commensale du groupe phylogénétique A. La deuxième souche, ED1a, a été isolée récemment en France. Cette souche du groupe B2 est contrairement à toute attente commensale. En effet, ED1a contient des îlots génomiques associés à la pathogénicité ; pourtant aucune infection ne lui est imputée. La troisième souche, IAI1, est une B1 proche de K-12 sur le plan métabolique (même sous-arbre métabolique). La souche EHEC est un pathogène intestinal responsable de l'épidémie de 1996 à Sakai (Japon) qui a fait près 5 727 victimes selon l'OMS. Cette souche du group E a hérité du nom de la ville : O157:H7 Sakai. La dernière souche, CFT073, est une autre B2 qui est une UPEC (pathogène du système urinaire) causant des pyélonéphrites aiguës.

Données expérimentales	**Souche**	**Groupe phylogénétique**	**Pathogénicité**	**Réactions**
Protéomique & Biolog	CFT073	B2	ExPEc	2399
	ED1a	B2	Commensal	2394
	IAI1	B1	Commensal	2399
	K-12 MG1655	A	Commensal	2387
	O157:H7 sakai	E	InPEc	2388
Biolog	042	D	InPEc	2391
	536	B2	ExPEc	2386
	55989	B1	InPEc	2404
	APEC O1	B2	ExPEc	2383
	HS	A	Commensal	2401
	O127:H6 E2348/69	B2	InPEc	2352
	O157:H7 EC4115	E	InPEc	2379
	O157:H7 EDL933	E	InPEc	2394
	S88	B2	ExPEc	2394
	UMN026	D	ExPEc	2416

Table 38 : Caractéristiques des souches utilisées.
Les données de Biologs sont disponibles pour 14 souches et les données de protéomiques pour 5 souches.

Pour certains des autres modèles reconstruits, je dispose des données expérimentales : ces dernières sont incluses dans l'analyse. Ainsi 9 souches supplémentaires vont être

étudiées pour le jeu de données Biolog, en complément des 5 souches initiales du projet MetaColi. Elles sont issues des différents groupes phylogénétiques et représentent les différents types de pathogénicité (Table 38).

.2.2 Les Biologs

La première série de données utilisées est la capacité de croissance sur différentes sources de carbones. Dans le chapitre précédent, j'ai montré que les modèles produisent de la biomasse sur plus d'une centaine de sources de carbone. Cette estimation reposant uniquement sur des prédictions *in silico*, la comparaison avec des expérimentations de croissance *in vitro* permet d'identifier des incohérences entre les modèles et la réalité.

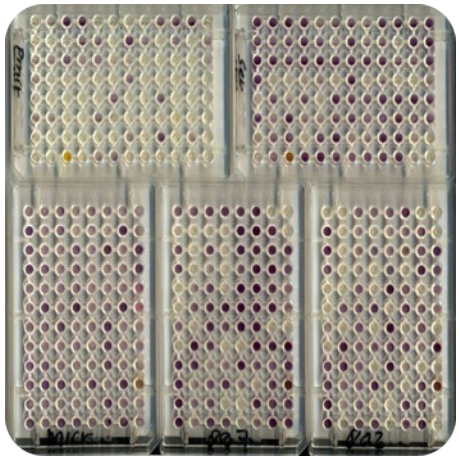


Figure 72 : Microplaques de Biolog.
Les microplaques de Biolog sont des plaques contenant 96 puits et 95 sources de carbones différentes. Elles permettent de tester rapidement la croissance de bactéries sur un grand nombre d'environnements.

Les données de croissance ont été réalisées sur des microplaques Biolog GN2 (AES Chemunex, Combourg, France). Chaque plaque est constituée de 96 puits (Figure 72), et chaque puits contient une source unique de carbone, à l'exception d'un puits qui sert de témoin. Ils contiennent également un colorant le tetrazolium, qui est un indicateur du métabolisme du carbone ; il est corrélé avec la croissance bactérienne. De plus, chaque puits contient les sources d'azote, de phosphate, de souffre etc. nécessaires à la croissance des bactéries ; cependant la liste complète des métabolites n'est pas disponible.
Deux types de résultat sont fournis à partir des Biologs : la mesure au point final et le coefficient de la courbe de croissance. La première est une mesure qualitative qui consiste à observer la différence de densité optique (DO) entre le puits contrôle et le puits de la source de carbone. La DO mesure l'absorbance de la lumière qui est proportionnelle à la surface cellulaire dans le milieu de culture. En partant du postulat que la surface cellulaire est constante pour une bactérie, l'absorption devient un indicateur de la quantité de bactéries. La méthode du point final présente un inconvénient majeur : il est parfois difficile de différencier une faible croissance de l'absence de croissance. Un seuil, à partir duquel on considère que la croissance est effective, a été estimé sur un ensemble de réplicats. Ce seuil, correspondant à une probabilité de 1% d'obtenir des faux positifs : il a été fixé à une valeur de 0.339 de DO, sachant que les valeurs varient entre -0.2 et 1.25. Ces valeurs de DO ont été données pour 13 souches.

Un deuxième type de résultat correspond à la courbe de croissance sur une durée de 18h. L'avantage de cette méthode est l'obtention de deux valeurs : le rendement et le taux maximal de croissance. Le rendement est la différence entre la DO au temps 0 et la DO au bout des 18h ; le taux de croissance est obtenu par régression polynomiale sur la courbe de croissance. Ces deux valeurs sont hautement corrélées, et donc seul le rendement est conservé. Les expériences ont été répétées 2 à 3 fois pour chaque combinaison (souche versus source de carbone) afin de déterminer un seuil sous lequel la croissance est considérée comme nulle (DO < 0.258). Au final je dispose de 3040 valeurs sur 14 souches, allant de -0.11 à 1.21.

.2.3 La protéomique

Les données de protéomique proviennent d'un long travail effectué par les collaborateurs et nécessite plusieurs étapes. Je vais en résumer le principe.
La première étape consiste à cultiver les cinq souches (K-12 MG1655, IAI1, ED1a, CFT073 et O157:H7 Sakai) sur quatre milieux. A l'instar des milieux de simulation, ils sont répartis en deux milieux minimums (glucose et gluconate) et deux milieux complexes (LB et urine humaine). Chaque combinaison milieu/souche est mise en culture dans plusieurs réplicats (2 à 6 fioles d'Elenmeyer). La culture est stoppée lorsque l'on atteint le milieu de la phase exponentielle, c'est à dire que le coefficient de la courbe de croissance commence à diminuer. Les cellules sont récupérées, lavées et centrifugées pour extraire les protéines (ces opérations sont répétées plusieurs fois).
La deuxième étape est une électrophorèse bidimensionnelle. Les protéines extraites précédemment vont subir des processus de précipitation/lavement. La première dimension consiste en une isoélectrofocalisation sur une gamme de pH allant de 4 à 7 ; les protéines vont migrer en fonction de leur charge électrique globale. La deuxième dimension est réalisée sur gels d'acrylamide pour séparer les protéines en fonction de leur taille. Un colorant, le bleu de Coomassie G-250, est ajouté pour colorer les protéines formant ainsi des spots. Chaque combinaison culture/extraction/électrophorèse est répétée au minimum trois fois. Un gel de référence est également créé à partir des extraits des 20 combinaisons souche/milieu (Figure 73). Les spots conservés sont ceux qui présentent le même profil d'absence ou de présence sur toutes les répétitions d'une condition expérimentale, et le profil doit être observable dans au moins dans 15 des 20 combinaisons. A partir de la superficie du spot, de l'intensité de la coloration et avec l'aide d'un logiciel d'analyse d'images, il est possible, par intégration, d'évaluer la quantité de protéines présente dans le spot.
L'étape suivante est l'identification des protéines contenues dans les spots. Pour cela une analyse au spectromètre de masse et un traitement informatique des résultats permet de prédire une liste de protéines candidates pour chacun d'eux. Cette liste est ensuite comparée aux prédictions faites à partir des données des souches contenues dans MicroScope à l'aide de deux logiciels différents.

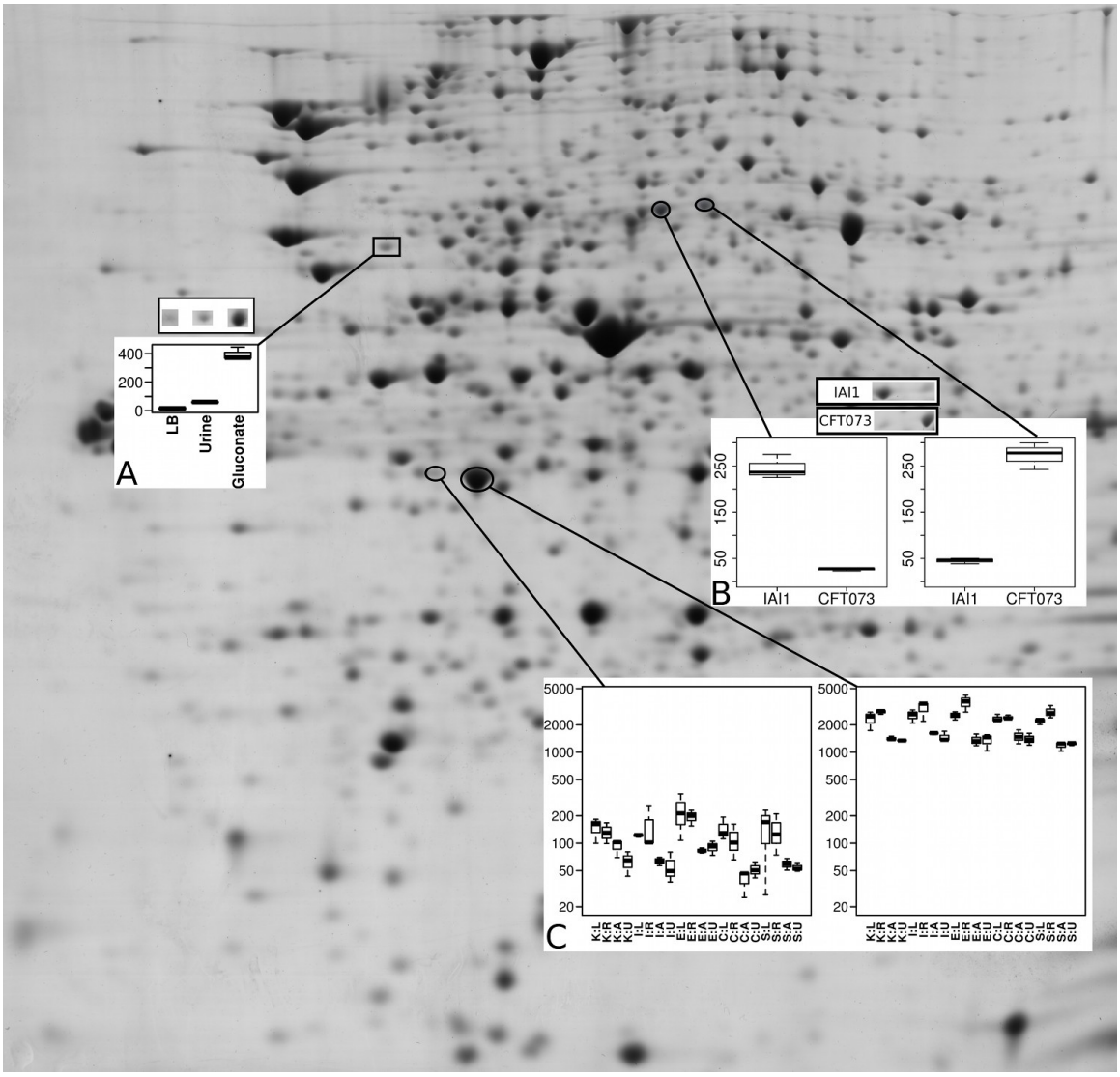


Figure 73 : Gel de référence obtenu à partir d'un mélange des extraits de culture des 20 combinaisons souche/milieu de croissance.

(A) Exemple de spot dont le volume varie selon le milieu de culture, ici pour la souche IAI1. (B) Deux spots alléliques migrent à deux endroits distincts du gel. Les deux spots sont présents sur le gel de référence. Par contre seul celui de gauche se retrouve sur les gels de IAI1 et celui de droite sur ceux de CFT073. Le logiciel d'analyse d'image détecte tout de même une faible coloration à la place des spots alléliques manquants, ce qui peut aboutir à des conclusions erronées quant à la concentration de la protéine. (C) Cas où deux spots correspondant à la même protéine sont retrouvés sur tous les gels. Cependant celui de gauche a une concentration 100 fois plus faible (l'ordonnée correspond au volume en échelle logarithmique), et dans ce cas n'est pas pris en compte dans la suite de l'analyse. Pour les diagrammes en boîte à moustaches, les boîtes vont du 1er au 3ème quartile et sont coupées par les médianes. Les segments en pointillés vont jusqu'aux valeurs extrêmes. (Image et légende issue de la thèse de V. Sabarly (Sabarly 2010))

Une avant dernière étape consiste à vérifier, associer et unifier les différents spots. En raison des variations alléliques entre les souches, différents spots peuvent correspondre à la même protéine ; un système de filtre a été mis en place pour conserver les cas sans ambiguïté et relier les spots contenant la même protéine.

Enfin, la dernière étape consiste à normaliser les résultats : si l'effet des répétitions est négligeable, les effets des souches, des milieux et des interactions souches/milieux

sont présents et biaisent les données. Après estimation des paramètres de correction, toutes les valeurs sont mises à jour.
Au final ces expériences ont généré un jeu de données unique, de plus de 13000 concentrations relatives.

3 Gestion des données

Dans le précédent chapitre, nous avons vu que les modèles sont disponibles sous forme de fichiers au format SBML, mais également dans la base de données de la plate-forme NemoStudio. Pour rester dans l'esprit des travaux précédents et permettre un accès rapide aux données, j'ai initié les fondations d'une base de données de faits biologiques. Cette base doit être compatible avec les données des bases relatives aux génomes (MicroScope), aux réseaux métaboliques (MicroCyc) et aux modèles du métabolisme (NemoStudio). Par exemple, l'identifiant de l'organisme doit être le même que celui de MicroScope ; de même pour tous les acteurs : les gènes doivent être associés à ceux de MicroScope, les protéines à celles de MicroCyc, etc.
Cette base répond à deux problématiques. D'une part l'organisation et la sauvegarde des résultats expérimentaux, d'autre part l'exploitation et l'intégration des données aux CBMs. Actuellement cette base de faits ne concerne que les données Biolog et les données de protéomique, mais elle reste ouverte à tout type de données.
Dans cette base, les modèles n'apparaissent pas afin de rester cohérent avec NemoStudio ; les résultats d'expériences concernent des organismes, et un organisme peut être représenté par plusieurs modèles. Associer l'expérience à l'organisme permet de ne pas dupliquer les données pour chacun des modèles.
Les expériences et modèles sont utilisés sur différents environnements, d'où la création de deux tables représentant les milieux [Media] et [Media_Met_Cpd] (Table 39).

[Media]		**[Media_Met_Cpd]**	
Media_id	Identifiant unique du milieu	*Media_id*	Identifiant du milieu
Media_name	Nom du milieu	*Mets_id*	identifiant NemoStudio du métabolite
Description	Information sur le milieu		
Media_type	liste (complexe, simple)	*Status*	Liste (allow, forbid)

Table 39 : Représentation d'un milieu dans la table [Media].
L'ensemble des métabolites listés dans le champ *allow* vont avoir un flux d'échange non nul. Le métabolite artefactual AllExf dans le champ *forbid* implique que tous les métabolites qui ne sont pas listés auront un flux d'échange nul.

La première table [Media] sert à définir le milieu, elle contient un champ *media_id* qui est l'identifiant unique du milieu, *media_name* le nom du milieu*, media_type* une liste qui spécifie si le milieu est simple (peu de métabolites) ou complexe (beaucoup de métabolites) et *description* qui donne des informations sur le milieu. La deuxième table [Media_Met_Cpd] sert à faire le lien entre le milieu et les métabolites. Elle possède les champs : *Media_id* qui fait référence à l'identifiant unique d'un milieu, *Met_id* qui est l'identifiant unique d'un métabolite de NemoStudio et le champ *Status* qui est une liste de deux valeurs possibles (allow, forbid) indiquant si le métabolite est présent dans le milieu et donc autorisé, ou absent et donc interdit. Si certains milieux, comme le LB, possèdent un vaste spectre de métabolites, la représentation de ces milieux riches en modélisation consiste à autoriser tous les flux pour chaque métabolite présent à l'extérieur du modèle. Ainsi d'un point de vue modélisation, deux milieux riches ne contiennent pas le même ensemble de métabolites. Pour rendre

le milieu indépendant du modèle et éviter de créer autant de fois un milieu qu'il existe de modèle, j'ai défini un métabolite artificiel nommé *AllMetExt.* Il désigne l'ensemble des métabolites non explicités dans un milieu ; l'exemple du milieu riche est donné dans la Table 40. Par défaut, on considère que sur milieu complexe, le statut de ce métabolite est *allow*, tandis que sur milieu simple il est *forbid*.

[Media]		[Media_mets]	
Media_id	Riche	*Media_id*	Riche
Media_name	Milieu riche		
Description	Milieu qui contient l'ensemble des métabolites assimilables par l'organisme	*Mets_id*	AllMetExt
Type	Complexe	*Status*	allow

Table 40 Définition du milieu riche.
Un seul métabolite correspond au milieu riche, le métabolite artificiel AllMetExt, ce qui signifie que toutes les bornes des flux d'échanges sont non nulles.

Une troisième table est également attachée à la définition des milieux bien que non utilisée dans ces travaux. L'étude au niveau du réseau métabolique des *Shigella* et des travaux réalisés ultérieurement m'ont conduit à prendre en considération l'auxotrophie dans la définition du milieu. Bien qu'il soit possible de créer un milieu complémenté pour chaque auxotrophie, il m'a semblé plus approprié de gérer ces cas en parallèle de la définition des milieux. Pour cela, la table [Orga_Auxo_Cpd] permet d'associer un organisme (champ *O_id*) à une auxotrophie (champ *Met_id*) ; l'organisme est défini par l'identifiant de MicroScope et l'auxotrophie par l'identifiant de métabolite dans NemoStudio. Cette table permet d'ajouter à la liste des métabolites d'un milieu, ceux qui sont indispensables du fait de l'auxotrophie : ainsi l'ensemble des métabolites du milieu minimum glucose d'un organisme auxotrophe pour le *tryptophane* peut être automatiquement complété par l'acide aminé correspondant.

3.1.1 Biolog

Les résultats de Biolog sont classés dans une table [Biolog_continu] composée des champs : *O_id* l'identifiant de l'organisme commun à MicroScope, *Media_id* l'identifiant du milieu issu de la table [Media], le champ *val_pf* qui contient la moyenne des valeurs des répétitions des expériences au point final pour un couple (*Media_id, O_id*), et le dernier champ est *val_cc* qui contient, pour un couple (*Media_id*, *O_id*), la moyenne des répétitions des valeurs de rendement. Initialement, les données étaient sauvegardées dans la table [Biolog_discret] comprenant les mêmes champs que [Biolog_continu] avec comme différence une valeur booléenne pour les champs *val_pf* et *val_cc*. Cependant, la perte d'information entraînée par la conversion des valeurs continues en valeurs discrètes était trop importante : par exemple, cela empêchait toute variation du seuil de DO pour conclure à la croissance ou non d'une souche.

3.1.2 Protéomique

Les données de protéomique sont stockées dans 4 tables différentes, dont une est facultative ; je reviendrai sur ce choix dans la partie 5.3. Les différentes protéines sont identifiées sur des gels : la première table, nommée [Gel], contient toutes les informations relatives à ces gels. Chacun d'eux possède un identifiant unique que nous conservons dans le champ *Gel_id*. Un gel est spécifique d'un organisme et d'un

milieu : ceci se traduit par un champ *O_id* qui assure le lien avec MicroScope et un champ *Media_id*, permettent la liaison avec la table [Media]. Un champ *description* permet d'ajouter des commentaires sur le gel. La table contient deux champs qui ne sont pas utiles à la modélisation, mais qui assurent un suivi des données : les champs *batch* qui indique le numéro de répétition et *Date* qui correspond à la date de l'expérience. Ces informations sont conservées dans le cas où une erreur serait détectée à posteriori sur une répétition, ou sur des expériences réalisées un jour précis.
La seconde table concerne les « spots » et leurs associations avec les organismes et les enzymes : [Spot_O_Prot_Cpd]. Suivant les souches, une même enzyme peut apparaître à des emplacements différents. Cette table assure la correspondance entre la souche (champ *O_id*), le spot (champ *Spot_id*) et la protéine (champ *Prot_id*). Les liens multiples, entre enzymes et spots m'ont conduit à définir une autre table [Prot_Gel_Vol_Cpd] qui donne directement l'information du volume (champ *Volume*) d'une enzyme (champ *Prot_id*) dans un gel (champ *Gel_id*). La dernière table permet de relier les enzymes aux gènes [Prot_GO_Cpd]. Elle est composée des champs *Prot_id*, *GO_id*, et *O_id,* qui font respectivement référence aux protéines, gènes et organismes.
Les données de protéomiques sont constituées de 13109 volumes, réparties sur 61 gels différents. Sur ces gels, 289 spots sont identifiés dans lesquels se trouvent 239 enzymes. Ces dernières sont identifiées par un nom, pour chaque enzyme je me suis assuré que son nom est bien le nom usuel dans MicroCyc. J'ai ensuite identifié les gènes pour chaque couple (organisme, enzyme) en utilisant les données contenues dans MicroCyc ; les valeurs sont résumées dans la Table 41.

Souches	K-12 MG1655	O157:H7 Sakai	CFT073	ED1a	IAI1
Nombre de gènes	221	205	217	212	219

Table 41 : Nombre de gènes associés à des protéines en fonction des souches

J'ai ensuite rempli les différentes tables en prenant soin à chaque fois d'utiliser les identifiants communs aux différentes bases de données.

4 Intégration des données de Biologs

Les Biologs sont habituellement utilisés dans le but d'améliorer les modèles : les prédictions d'un modèle sont comparées aux résultats expérimentaux pour expliciter les incohérences et identifier les voies de dégradations absentes du modèle (Chapitre d'introduction sur la reconstruction des modèles du métabolisme partie 4.2). Dans ce travail, j'utilise les données Biologs d'une manière différente : elles vont permettre d'évaluer la qualité des modèles et leurs capacités de dégradation. Dans cette comparaison, j'ai supposé que les différences ne sont pas dûes à l'absence d'une voie métabolique commune à une majorité des modèles. En effet, les différentes étapes d'homogénéisation de l'annotation, puis de reconstruction des réseaux et de conversion en modèles, minimisent le risque de « sous-prédiction », c'est-à-dire l'absence erronée de réactions qui sont communes à une majorité des réseaux et des voies métaboliques associées.
Deux jeux de données Biolog sont disponibles : le premier contenant les mesures au point final, le second contenant les rendements. Seul les résultats de rendement vont être détaillés dans ce manuscrit, puisqu'ils présentent une qualité supérieure aux autres données ; cependant les résultats au point final sont présentés en Annexe 6.
Dans les deux cas, j'ai transformé les valeurs continues en valeurs booléennes (pousse/ne pousse pas) avec le même processus. J'ai calculé la moyenne des répétions

de chaque couple (organisme, modèle) et comparé celle-ci aux seuils en dessous duquel on considère que la croissance est nulle. Ces seuils ont été estimés par les collaborateurs et sont fixés à 0.258 pour les mesures de rendement et à 0.339 pour les mesures de point final.
Le calcul des prédictions est effectué sous Matlab. Pour chaque source de carbone, un milieu minimum est créé avec les métabolites suivants (en plus de la source de carbone testée) : calcium, chlore, dioxyde de carbone, cobalt, cuivre, fer (fe^{2+} et fe^{3+}), eau, hydrogène, potassium, magnésium, manganèse, molybdate, sodium, phosphate, tungstène, zinc. Le flux de biomasse est obtenu par FBA en utilisant la COBRA-Toolbox (Becker et al. 2007) et le solveur numérique *cplex*. Afin de faciliter l'exécution de l'analyse, une fonction de définition des contraintes des modèles spécifiques des milieux minimum a été mise au point. En prenant en entrée un modèle et une source de carbone, elle contraint à 0 les bornes maximales et minimales des flux d'échanges, et autorise lés flux d'échanges des métabolites qui forment le milieu minimum.
Lors des comparaisons des observations expérimentales contre des prédictions, le terme « souche » fera référence aux observations expérimentales, tandis que le terme « modèle » fera référence aux prédictions.

.4.1 *iAF1260* vs *K-12 MG1655Cbm*

La première comparaison consiste naturellement à confronter les données Biolog de la souche de référence d'*E. coli* K-12 MG1655 aux deux modèles : celui de référence *iAF1260* et celui reconstruit dans cette étude, *K-12 MG1655Cbm* (l'ensemble de valeurs est disponible dans la feuille de calcul 1 de Annexe 7). Une croissance de la bactérie est observée sur 30 des 94 sources minimales Biolog. Ce nombre varie peu en fonction de la valeur du seuil : il doit être inférieur à 0.19 pour qu'une nouvelle source de carbone soit considérée comme assimilable. La division du seuil par 2 n'implique que 7 sources de carbone supplémentaires.
Le modèle de référence (*iAF1260*) prédit une croissance sur 44 sources et celui reconstruit (*K-12 MG1655Cbm*) sur 45 surces, soit 1.5 fois le nombre de sources observées. Parmi les 30 sources de carbone assimilables par la souche, une seule est non dégradable par des deux modèles : il s'agit du *methyl-pyruvate* (formule $C_4H_6O_3$). Ce métabolite est absent dans les deux modèles et également dans le réseau, Bien que référencé dans MicroCyc, le *methyl-pyruvate* n'est utilisé ou produit par aucune réaction.
Une autre source est cette fois-ci prédite comme dégradée uniquement par le modèle *K-12 MG1655Cbm :* il s'agit du *D-galactonolactone*. Cette différence était attendue puisque, comme nous l'avons déjà vu dans le chapitre précédent (partie 4.1), la seule différence entre les deux modèles est la présence de la voie de dégradation du *D-galactonolactone* qui produit du *pyruvate*. Nous avons vu que cette voie était fonctionnelle et nous avons maintenant la preuve expérimentale que cet ajout est un vrai positif. En retraçant l'origine de l'intégration de cette voie qui est également absente dans le réseau de référence, le processus de reconstruction du réseau a permis de prédire l'existence de la voie décrite dans MetaCyc par la présence d'une réaction annotée dans K-12 qui a fourni suffisamment d'évidence pour réaliser l'inférence de l'existence des autres réactions de la voie.
Les pourcentages de similitudes entre les observations et le modèle *K-12 MG1655Cbm* sont donnés dans la Table 42.

		Biolog	
		Croissance	Non
Model	Croissance	97%	27%
	Non	3%	73%

Table 42 : Pourcentage de similitudes entre l'observation et la prédiction du modèle *K-12 MG1655Cbm*.

Au final 79% des comparaisons prédictions/observations sont concordantes.

Sur les 30 sources Biolog où une croissance est expérimentalement observée, le modèle *K-12 MG1655Cbm* réalise 81% de bonnes prédictions. En opposition au seul résultat faux négatif obtenu (le *methyl-pyruvate*), le modèle prédit 14 faux positifs faisant chuter le pourcentage de bonnes prédictions. L'origine possible de ces faux positifs sera abordée dans la prochaine partie.

.4.2 Comparaison globale

La comparaison globale consiste à confronter les données expérimentales aux prédictions des modèles pour 14 souches sur 94 sources de carbone Biolog testées (Table 38). Il existe 45 sources de carbone pour lesquelles une des souches est capable de pousser. Parmi ces sources, 15 permettent la croissance de toutes les souches. A l'opposé, il y a 79 sources de carbone pour lesquelles aux moins une des souches est incapable de croître, et 49 sur lesquelles aucune des souches ne pousse. Ce qui donne 30 sources qui présentent de la variabilité de croissance.

Du côté de la modélisation, il existe 48 milieux minimums pour lesquels au moins un des modèles prédit un flux de biomasse, dont 41 milieux communs à l'ensemble des modèles. Il existe 53 milieux pour lesquels au moins un modèle est incapable de produire de la biomasse, parmi lesquels 47 donnent systématiquement un flux de biomasse nul, peu importe le modèle. En résumé, seulement 6 milieux présentent de la variabilité entre les modèles.

Comme précédemment, on remarque une grande différence entre le nombre de sources de carbone qui donnent toujours une croissance et l'équivalent en modélisation (respectivement 15 et 42). La variabilité des modèles est beaucoup plus restreinte que la variabilité métabolique observée : il y a un facteur 5 entre les observations et les prédictions du nombre de sources de carbones dont le comportement varie en fonction des souches. Cette constatation rejoint les conclusions du chapitre précédent sur la perte de diversité entre les réseaux et les modèles. Si cette diversité est déjà faiblement perceptible dans les réseaux, le manque de précision empêche son utilisation dans les modèles.

Le résultat de l'ensemble des comparaisons est représenté dans la Figure 74. Parmi les milieux, 84 présentent une comparaison positive pour au moins une des souches, et 55 pour l'ensemble des souches. Sur les milieux où la prédiction et l'expérimentation sont systématiquement en accord, 15 correspondent à une croissance et 40 à une absence de croissance. Il reste 39 milieux sur lesquels on observe au moins une inconsistance entre les données *in vitro* et *in silico.* Dans les cas où l'inconsistance concerne l'ensemble des souches, il s'agit d'une prédiction erronée de croissance dans les modèles.

		Table de vérité		Précision	
		Modèle		Modèle	
		Croissance	Pas de croissance	Croissance	Pas de croissance
Biolog	Croissance	373	63	86%	14%
	Pas de Croissance	257	623	29%	71%

Table 43 : Table de vérité et précision.

Sur les 1316 comparaisons, 996 sont concordantes entre l'observation et la prédiction parmi lesquelles 373 concernent la présence de croissance et 626 l'absence de croissance. Au total c'est 76% des prédictions qui sont cohérentes entre l'expérience et la prédiction.

Parmi les 1316 comparaisons observation/prédiction (Table 43), 996 se sont avérées exactes. Ce ratio de 76 %, obtenu pour un ensemble de modèles reconstruits automatiquement est satisfaisant comparé aux 80% de similitude du modèle de référence pour *E. coli* K-12 MG1655.
Sur les 94 environnements, seulement 10 sources de carbone présentent une inconsistance pour l'ensemble des modèles : *4-Aminobutanoate, acétate, citrate, éthanolamine, formate, L-glutamate, ornithine, propionate, putrescine et L-thréonine*. A chaque fois, les modèles prédisent une croissance alors que l'observation montre l'absence de celle-ci. Quatre hypothèses peuvent expliquer ces différences. Tout d'abord la définition du milieu de simulation des CBMs n'est pas conçue pour prendre en compte la notion de micro-aérobie des Biologs. Cependant, l'analyse des voies de dégradation pour ces sources de carbones ne montre pas de dépendance à l'oxygène : cette hypothèse est donc rejetée.
La seconde hypothèse porte sur les GPRs. Si les voies de dégradations utilisent des réactions différentes et des GPRs différentes, on constate que les transporteurs sont tous reliés à la même GPR qui, chez K-12 MG1655, contient les gènes : *ompE* (identifiant MicroScope 1084928) ou *ompN* (1086101) ou *ompC* (1086958) ou *ompF* (1085639). L'annotation de ces gènes indique des transporteurs et des protéines membranaires non spécifiques. Si effectivement la GPR est incertaine et que l'association transporteur/gènes est erronée, d'autres transporteurs existent pour ces composés : par exemple le transport du citrate est dépendant du fer (Hussein et al. 1981). Ceci nous laisse penser que l'erreur de prédiction ne provient pas uniquement des transporteurs qui sont généralement non essentiels dans les voies catabolique notamment dû fait de leur redondance en nombre et de leur non spécificité vis à vis des substrats.
La troisième hypothèse concerne l'un des a priori des CBMs : si une réaction est présente et qu'aucune information n'est disponible sur son activation, alors elle est active par défaut. Sans autres contraintes imposées, le modèle produit donc de la biomasse si les réactions de la voie de dégradation sont présentes. Il manque donc des contraintes supplémentaires pour bloquer certains flux. On sait que dans certaines conditions *E. coli* peut pousser sur *acétate* (Díaz-Guerra et al. 1997; Cozzone 1998) ; or les voies de dégradation et les réactions impliquées sont sous le contrôle de phénomènes de régulations complexes. Les modèles montrent que, sans la régulation et l'inhibition de certaines réactions, les voies de dégradation sont actives. Il ne faut pas en déduire que les modèles reconstruits sont inexploitables ; simplement, et comme cela a été développé dans l'introduction, ils sont incomplets.
Enfin la dernière hypothèse concerne les données expérimentales qui peuvent contenir des erreurs : par exemple O157:H7 Sakai ne pousse pas sur gluconate en mesure de point final alors qu'elle le devrait. La composition des Biologs n'étant pas connue, il peut y avoir des interactions et effets non prévus.

Il existe 30 sources pour lesquelles on observe des comportements différents en fonction des souches. Parmi elles, 5 métabolites peuvent être assimilés par certaines souches et pas par d'autres, et sont complètement absents des modèles (Table 44).

	Croissance	MicroCyc	EcoCyc	Voie de Dégradation
L-alanyl-glycine	7	oui	oui	Non
Raffinose	6	oui	non	chez les eucaryotes
Lactulose	7	oui	non	Non
méthyle pyruvate	13	oui	non	Non

acide glycyl-L-aspartic	1	non	non	Non

Table 44 : Inconsistances dues à l'absence des métabolites.
La première colonne donne le nombre de souches capables d'utiliser la source de carbone. Dans 4 cas, le métabolite est référencé dans la base de données généraliste, mais absent de celle dédiée à *E. coli*. De plus, ils ne participent à aucune voie métabolique chez les bactéries.

Ces inconsistances sont le résultat d'un manque de connaissance à la fois générales, puisqu'aucune voie de dégradation n'est décrite chez les bactéries, et d'un manque de connaissance spécifiques à *E. coli*, puisque dans la base de données dédiée (EcoCyc) les métabolites n'apparaissent pas. Par conséquent, il est impossible de les retrouver dans les modèles.
Pratiquement la moitié des différences (14 sur 30) concernent des sources de carbone assimilables par tous les modèles (Table 45).

Source	2-Oxoglutarate	D-Alanine	L-Alanine	L-Asparagine	L-Aspartate	Dextrine	L-Fucose	D-Glucarate
Croissances	2	4	9	4	4	11	12	8
Source	D-Glucuronate	Glycerol 3-phosphate	L-Rhamnose	D-Serine	L-Serine	Succinate	Sucrose	Uridine
Croissances	13	3	9	7	3	8	6	10

Table 45 : Inconsistances dues à des transporteurs.

Les transporteurs de ces sources de carbone sont associés à la GPR mentionnée précédemment (gènes *ompENCF*). Les conclusions restent identiques : si une erreur expérimentale est possible, il est beaucoup plus probable que la différence soit sur les associations gènes/transporteurs ou sur la régulation des différentes voies.
Un gène de cette GPR est absent dans certaines souches, cependant aucune corrélation n'a pu être établie entre son absence et l'absence d'un transporteur d'un métabolite particulier (e.g. gène *ompE* chez la souche ED1a).
A noter également qu'initialement la *glutamate-pyruvate aminotransferase*, l'une des réactions de la voie de dégradation du *2-Oxoglutarate,* était sans gène associé dans le réseau et le modèle de référence. En analysant si cette réaction pouvait être responsable des incohérences sur cette source de carbone, il est apparu que récemment des gènes ont été associés à cette réaction (S. H. Kim et al. 2010), et qu'ils possèdent des homologues dans les différentes souches (recherche de gènes homologues dans MicroScope) : les incohérences de dégradation sur le *2-Oxoglutarate* ne sont donc pas dûes à cette réaction.
Parmi la dizaine d'incohérences restantes, aucun critère commun n'a été découvert. Toutes les souches, excepté K-12 MG1655, sont capables de dégrader le *N-Acetyl-D-galactosamine.* Ce métabolite est néanmoins présent dans tous les réseaux et modèles, mais reste dans le périplasme où il intervient dans une réaction. Ce composé intervient également dans d'autres réactions mais aucune n'est associée à des voies métaboliques. C'est pourquoi il est impossible de trouver une voie de dégradation fonctionnelle dans les modèles. La comparaison sur le *mélibiose* est incohérente pour sept modèles qui devraient produire de la biomasse. L'erreur peut être corrigée en ajoutant les transporteurs périplasmiques, absents des modèles incriminés. Cette observation rejoint les remarques et les artefacts de reconstruction développés dans le chapitre précédent (partie 4.1).
La voie de dégradation du *4-hydroxyphényl acétate* est présente dans 5 souches : HS, IAI1, O157:H7 EDL933, Sakai, et UMN026. Le flux de biomasse nul pour ces modèles est dû à l'absence du transporteur de cette source de carbone. Les observations montrent une croissance pour HS et IAI1 et une absence de croissance pour les autres : cette hétérogénéité de comportement nous empêche d'ajouter un

transporteur pour le *4-hydroxyphényl acétate* avec une activation de type « *inconnue* ».
La voie de dégradation du *D-galactonolactone* absente du modèle pivot, est de nouveau mise en évidence du fait de l'incapacité de la souche IAI1 à le dégrader. La voie de dégradation comprend deux réactions, une avec gène et l'autre sans. Le gène homologue de la première réaction est présent chez IAI1, il est donc fort probable que la seconde réaction soit absente d'IAI1, ou bien il s'agit d'une erreur expérimentale. On constate aussi pour ce métabolite des résultats cohérents pour les deux souches de O157:H7 : la prédiction des modèles confirme l'absence de croissance observée.
La *dextrine* est un cas particulier puisqu'elle correspond à une classe de métabolite qui possède 5 instances dans nos modèles. Dans la Table 45 nous avons donné les résultats des prédictions de flux de biomasse qui sont identiques pour 4 des instances. La cinquième instance diffère car il n'existe pas de transporteur du métabolite : en rajoutant ce transporteur le résultat obtenu est conforme aux autres instances.
Il existe deux sources de carbone comportant des incohérences de deux sortes différentes : prédiction d'un flux nul et observation de croissance, ou prédiction d'un flux de biomasse et absence de croissance. Dans les deux cas, lorsque la prédiction donne un flux de biomasse non nul, c'est, une fois de plus, un transporteur associé à la GPR générique qui est en cause. La première source est le *D-malate* : par l'absence du gène *yeaU,* les souches 042, APEC O1, CFT073, ED1a, O127:H6 E2348/69 et S88 ne peuvent pas produire la *malate décarboxylase oxydoréductase* pour transformer le *malate* en *pyruvate*. L'observation d'une croissance non nulle dans les Biologs tend à indiquer l'existence d'un isozyme ou d'une voie de dégradation alternative. L'autre source, le *D-sorbitol*, n'est pas dégradée dans le modèle de O127:H6 E2348/69 du fait de l'absence, dans son génome, de l'opéron *srl* qui code un transporteur PTS.
Le dernier exemple d'inconsistance permet de mettre en évidence un *faux positif* issu du processus de reconstruction des réseaux. Une seule souche est capable d'assimiler l'*inositol* : ED1A. Cependant deux modèles produisent de la biomasse : ED1aCbm et *UMN026Cbm* en utilisant la *myo-inositol 2-dehydrogenase*. Si chez ED1a la réaction est associée au gène *iolE* et à un gène « iolE-lke », chez UMN026 cette réaction est liée à un fragment d'ADN d'origine prophagique : à la vue des résultats expérimentaux , c'est un *faux positif* pour UMN026 et un *vrai positif* pour ED1a.
Je terminerais cette comparaison par un autre vrai positif qui montre l'efficacité de la méthode de reconstruction : une seule souche est capable d'utiliser le *D-arabitol*, et seul son modèle est capable de dégrader ce métabolite grâce à une *D-arabinitol 4-dehydrogenase* qui est codée par un locus sans nom chez O127:H6 E2348/69.

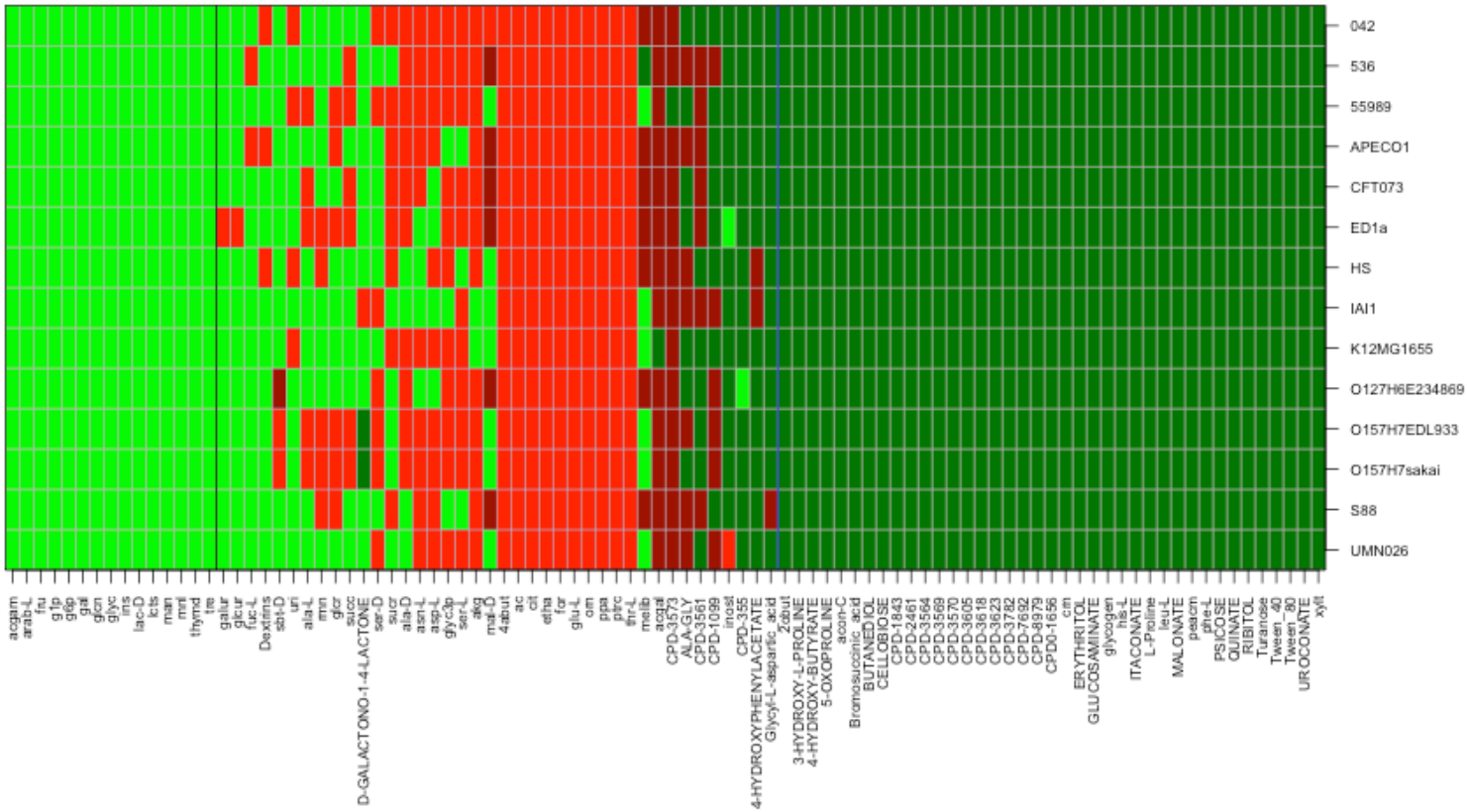


Figure 74 : Comparaison des observations et prédictions.

En vert, les couples (organisme, milieu) pour lesquels l'observation et la prédiction sont cohérentes. La teinte claire indique une croissance, la teinte foncée l'absence de croissance. En rouge, les couples (organisme, milieu) pour lesquels les observations et la prédiction sont en opposition ; la teinte claire indique que le modèle prédit un flux de biomasse non nul, alors que l'observation donne une absence de croissance, la teinte foncée indique que le modèle prédit un flux de biomasse nul alors que l'observation montre une croissance. Il existe 55 milieux pour lesquels les observations et prédictions sont cohérentes pour toutes les souches (à gauche de la barre noire et à droite de la barre bleue) et 10 pour lesquels les deux données sont systématiquement contradictoires, ce qui laisse 29 milieux avec des cohérences ponctuelles.

5 Intégration de données de protéomique

Cette partie est abordée d'un point de vue plus didactique et explique les différentes étapes, hypothèses et réflexions nécessaires à l'intégration de données biologiques en modélisation. Ces réflexions, bien qu'utilisées dans le cadre des CBMs sont, dans les principes, applicables à tous types de modèles et font parties des fondements de la biologie des systèmes. Le développement des méthodes d'intégration de données est très important : si pour les organismes modèles il est possible de trouver une grande variété de données expérimentales, pour des organismes moins étudiés les résultats disponibles ne sont pas toujours les plus adaptés.

.5.1 Problématique

Différents types de données biologiques ont déjà été intégrés dans des modèles à bases de contraintes avec succès. Ces données touchent différents aspects de la biologie : la physiologie avec des mesures de consommation et production de métabolites, de concentration de métabolites ou des mesures de flux ; la génétique avec des données d'essentialité sur des mutants ; la biochimie avec des données de thermodynamique etc. Cependant, il n'existe aucun précédent d'utilisation de concentrations relatives d'enzymes pour limiter l'espace des solutions d'un CBM. Même si en théorie les modèles à bases de contraintes acceptent toutes sortes de contraintes, en pratique certaines sont difficiles voire impossibles à mettre en œuvre. Il faut donc, avant d'intégrer ces données, en estimer la faisabilité.

.5.2 Définition des données et compatibilité avec le modèle

Le premier point est la caractérisation des données disponibles et la compréhension des phénomènes biologiques sous-jacents. On a à disposition un ensemble de valeurs définies par différents critères : organisme, milieu, enzyme, gène, gel, répétition, spot et volume. Même si la démarche est trivial, la première chose à faire consiste à repérer le ou les critères d'intérêts (dans ce cas le volume), et les critères secondaires voir inutiles dans l'analyse. Le volume est calculé à partir d'un spot qui se trouve sur un gel. Un spot contient une protéine particulière. Le gel, support du spot, est réalisé pour un organisme et un milieu donné. En résumé, une observation correspond à un volume d'une protéine enzymatique pour un milieu et un organisme.
Première limite : les modèles ne possèdent pas d'enzyme. Mais grâce au réseau reconstruit et aux références-croisées, il est possible de lier les enzymes aux gènes et aux réactions du modèle.
Les volumes possèdent un certain nombre de propriétés. Tout d'abord ils sont relatifs : le volume de chaque spot a été divisé par le volume médian des spots identifiés dans le gel. Si pour un spot donné le volume est identique sur deux gels différents, la quantité de protéine n'est pas forcément équivalente. Chaque volume est divisé par la masse moléculaire de la protéine. Enfin, ils sont corrigés afin d'éliminer tout artefact tel que l'effet souche, l'effet milieu et les interactions souche/milieu. Les volumes sont donc directement comparables entre eux au sein d'un gel, ou entre deux conditions différentes.
Intuitivement, on suppose qu'il existe un lien entre le volume de l'enzyme et le flux de matière. Cependant la façon d'expliciter cette liaison reste à définir. Pour reprendre un exemple de données déjà cité précédemment, un mutant peut directement être répercuté sur le modèle en bloquant les réactions pour lesquelles le gène est essentiel.
Il est impossible d'introduire directement des volumes dans le modèle, mais il est

possible de passer par une étape intermédiaire : par exemple les concentrations de métabolite sont utilisées par l'intermédiaire de la thermodynamique qui impose un sens à certaines réactions.
On peut envisager deux axes pour intégrer ces données : i) trouver une méthode, basée sur des hypothèses biologiques pour associer directement le volume à une grandeur des modèles ; ii) trouver une interface qui va utiliser les concentrations protéiques en données initiales et produire des contraintes à appliquer sur les flux.

.5.3 Etudes des données

Pour des raisons pratiques j'ai décidé d'utiliser comme identifiant les enzymes plutôt que les spots. Ce choix se justifie par l'unité des enzymes : une enzyme possède le même nom pour tous les organismes, alors que deux spots différents peuvent correspondre à la même enzyme. L'enzyme permet un lien plus simple entre le volume, la souche et le milieu.
Il est difficile de choisir l'un des deux axes avec la seule description des données. C'est pourquoi il est nécessaire de les étudier. Il ne s'agit pas d'effectuer une analyse précise, mais de se familiariser, de voir les tendances, les limites et corrélations qui existent au sein de ces résultats. La connaissance fournie par cette étude servira principalement à définir les hypothèses d'utilisation des données.
La première chose que j'ai regardée, est la répartition des volumes entre les différentes souches et les différents milieux (Table 46).

	LB	Urine	Glucose	Gluconate	Total
K-12 MG1655	663	884	663	663	2873
CFT073	651	651	651	651	2604
O157:H7 Sakai	615	615	615	615	2460
ED1a	636	636	636	636	2544
IAI1	657	657	657	657	2628
Total	3222	3443	3222	3222	13109

Table 46 : Répartition des volumes entre les souches et les milieux
La différence pour K-12 MG1655 s'explique par la présence d'un gel supplémentaire.

Le nombre de volumes est réparti de façon homogène sur toutes les combinaisons de souches/milieux. Ceci nous permet d'éliminer un éventuel biais qui serait lié à la distribution des données. On retrouve environ 215 protéines par gel, et elles touchent la plupart des processus métaboliques (Table 47).

	Nombre de protéines	Processus	Nombre de protéines
Synthèse des acides aminés	38	Métabolisme des nucléotides	22
Réactions anaplérotiques	3	Voie des pentoses phosphate	4
Synthèse des cofacteurs et petites molécules de transport	9	Dégradation des peptides, acides aminés et polyamines	15
Dégradation des Carbohydrates	14	Synthèse des polyamines	1
Détoxification	5	Porines et transporteurs putatifs	11
ATP-proton et transport d'électron	3	Synthèse protéique et processus cellulaire	50
Entner-Doudoroff	2	Régulation	9
Fermentation	5	Métabolisme du sulfure	2
Gluconéogenèse et glyoxylate	4	TCA	10
Glycolyse	12	Toxine production et résistance	3
Métabolisme du fer	7	inconnu	2
Métabolisme des phospholipides et acide-gras	8		

Table 47 : Nombre de protéines impliquées dans les différents processus métaboliques.

Ces deux tableaux montrent que les données sont bien reparties entre les expériences et au sein du métabolisme. La Figure 75 apporte en plus une information importante : la présence d'une diversité dans les valeurs.

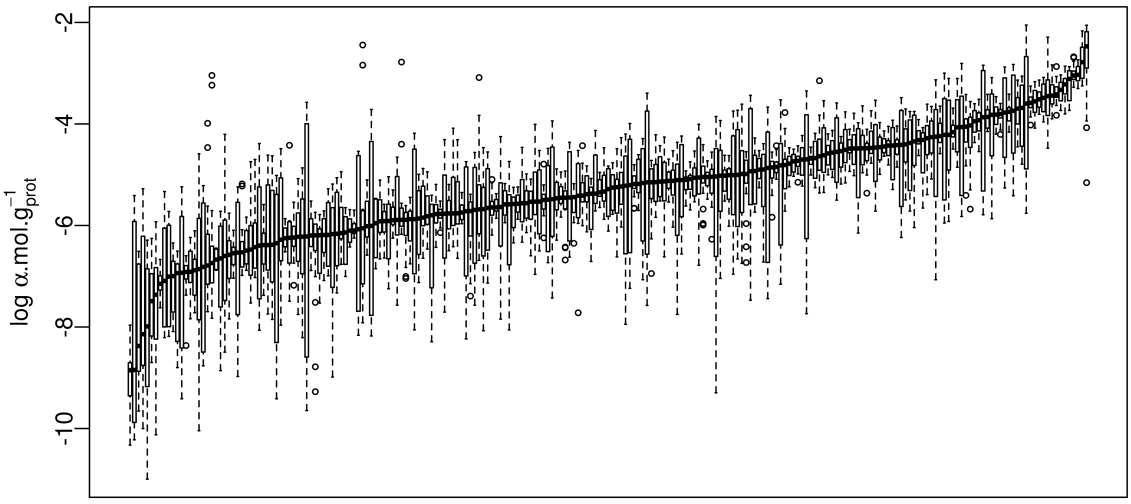


Figure 75 : Diagramme de la concentration massique de chaque protéine sur toutes les combinaisons souche/milieu de culture

Diagramme en « boîte à moustaches » représentant la variabilité du logarithme des concentrations massiques. Les boîtes vont du 1er au 3ème quartile et sont coupées par les médianes. Les segments en pointillés vont jusqu'aux valeurs extrêmes dans la mesure où celles-ci sont situées à moins de 1,5 fois la taille des boîtes par rapport à l'extrémité des boites, dans le cas contraire les valeurs sont représentées par des cercles (Extrait de la thèse de V. Sabarly (Sabarly 2010)).

Je me suis ensuite intéressé à la variation locale en regardant la variation des volumes d'enzyme en fonction des souches et des milieux. Pour chaque enzyme (plus de 230), un graphique est généré (Figure 76). Sur cette figure, on voit nettement un effet milieu : les milieux plus simples ont, en moyenne, un volume beaucoup plus important. Cette séparation n'est pas aussi marquée pour toutes les enzymes. Dans d'autres cas ce sont les milieux complexes qui possèdent les volumes les plus

importants. Il ne s'agit pas ici d'étudier précisément la variation des volumes, mais d'avoir conscience des différentes situations possibles.
Ces observations soulèvent la question de la variabilité dans les modèles : les différences locales au niveau de l'enzyme sont-elles présentes dans les résultats des simulations, avant même l'intégration des données ? Pour cela j'ai regardé les variations et la valeur optimale du flux en fonction du volume de l'enzyme (respectivement Figure 77A et Figure 77B). Là aussi, un graphique est réalisé pour chaque couple souche/milieu, pour estimer si tous les couples se comportent de la même manière. Ces comparaisons montrent que plus le volume augmente plus ce nombre de flux non nuls augmente, mais ils montrent également un certain nombre d'inconsistance entre les prédictions et les observations.
A partir de ces analyses et des différents éléments qu'elles apportent sur le jeu de données, on peut commencer à concevoir la manière de les intégrer.

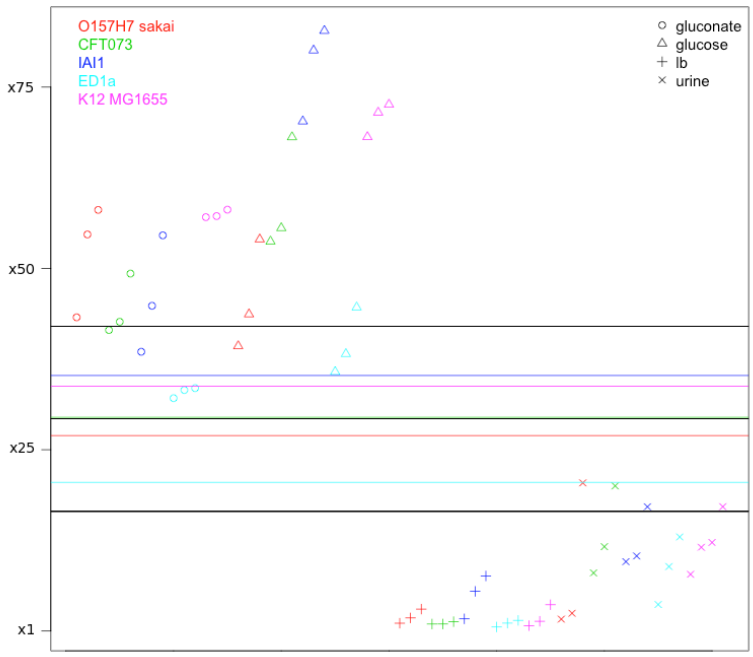


Figure 76 : Variation des volumes pour la D-3-phosphoglycérate déshydrogénase.
En ordonnée, le multiplicateur du volume. En abscisse, les enzymes sont regroupées par combinaisons de souche/milieu. La ligne noire centrale représente la moyenne sur toutes les données, et les 2 autres, la moyenne plus ou moins la moitié de l'écart type. Les lignes de couleurs représentent les moyennes pour chacune des souches.

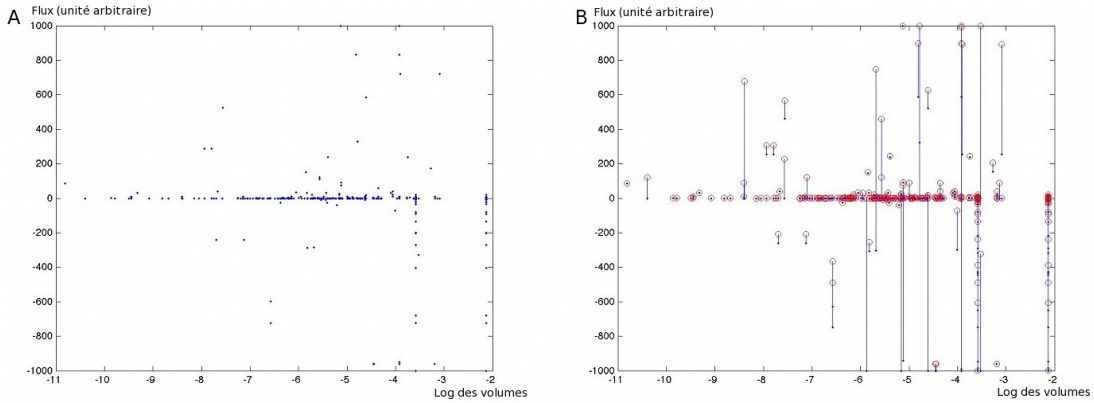


Figure 77 : Comparaison entre le volume d'une enzyme, son flux optimal (A) et sa variabilité (B). La comparaison représentée est celle de K-12 MG1655 sur milieu LB. Les enzymes sont ordonnées par ordre croissant de volume.

.5.4 Hypothèses et limites

L'étude des données de protéomique permet de limiter les hypothèses sur les données. Les observations sont comparables entre elles, reparties équitablement sur l'ensemble des conditions d'expérience et touchent divers aspects du métabolisme. La reproductibilité des résultats lève le doute sur d'éventuelles erreurs expérimentales. Nous pouvons donc maintenant nous focaliser sur les différentes implications des données. L'hypothèse naturelle est : plus le volume de l'enzyme est important, plus le flux de matière qui passe dans la réaction est important. Si cette hypothèse est vraie pour une concentration de substrat en excès, elle devient invalide dans le cas inverse : si le substrat fait défaut, augmenter la quantité d'enzyme permet d'augmenter la probabilité de rencontre entre l'enzyme et le substrat, sans pour autant augmenter le nombre de substrats convertis par unité de temps. Autrement dit l'augmentation de la quantité de protéine peut soit augmenter le flux de matière dans une voie, soit éviter qu'il ne diminue. Et la présence même de l'enzyme n'assure pas obligatoirement la présence d'un flux, du fait des possibilités de régulation et d'inhibition. Néanmoins, il n'est pas exagéré d'associer la présence d'une enzyme à un flux non nul, et de considérer un lien entre la quantité d'enzyme et la quantité de matière qui transite par le flux.

L'absence de valeur pour une protéine dans un gel ne signifie pas forcément l'absence de la protéine dans la combinaison souche/milieu correspondant au gel ; la protéine peut être en trop faible quantité pour être détectée, ou pour une raison inconnue elle n'a pu être identifiée ou conservée dans les expériences. C'est pourquoi l'absence d'une protéine du jeu de données peut ne pas être considérée comme une absence de la protéine dans l'organisme.

Le choix du type d'intégration des données, directe par une relation de proportionnalité ou indirecte en passant par un phénomène biochimique intermédiaire, ne peut se faire qu'après évaluation des méthodes déjà existantes.

L'utilisation de volume (et par extension de concentration) en entrée, et l'obtention de flux en sortie, fait penser aux modèles cinétiques : le modèle CBM de K-12 peut ainsi être couplé avec un modèle cinétique (Figure 78). Ce modèle comprend 29 réactions dont 5 artificielles et une centaine de métabolites. L'intégralité des métabolites et des 24 réactions est associé à leurs homologues dans le modèle CBM. Pour 17 de ces réactions, les enzymes qui les catalysent ont été retrouvées dans les gels. L'utilisation du modèle cinétique présente deux désavantages majeurs qui ont abouti au rejet de cette méthode. Seulement 17 des 239 (7%) enzymes sont prises en compte. De plus, le

couplage des deux modèles demande des ajustements ; un premier test utilisant les données et flux observés dans l'article du modèle cinétique (Chassagnole et al. 2002), a entraîné une utilisation excessive de la voie des *pentoses phosphate* dans le modèle CBM. Bien que le travail d'ajustement des deux modèles soit réalisable, la reconstruction d'un modèle cinétique, qui pourrait prendre en compte au moins ¼ des données est pour le moment impossible. C'est pourquoi l'utilisation du modèle cinétique a été abandonnée.

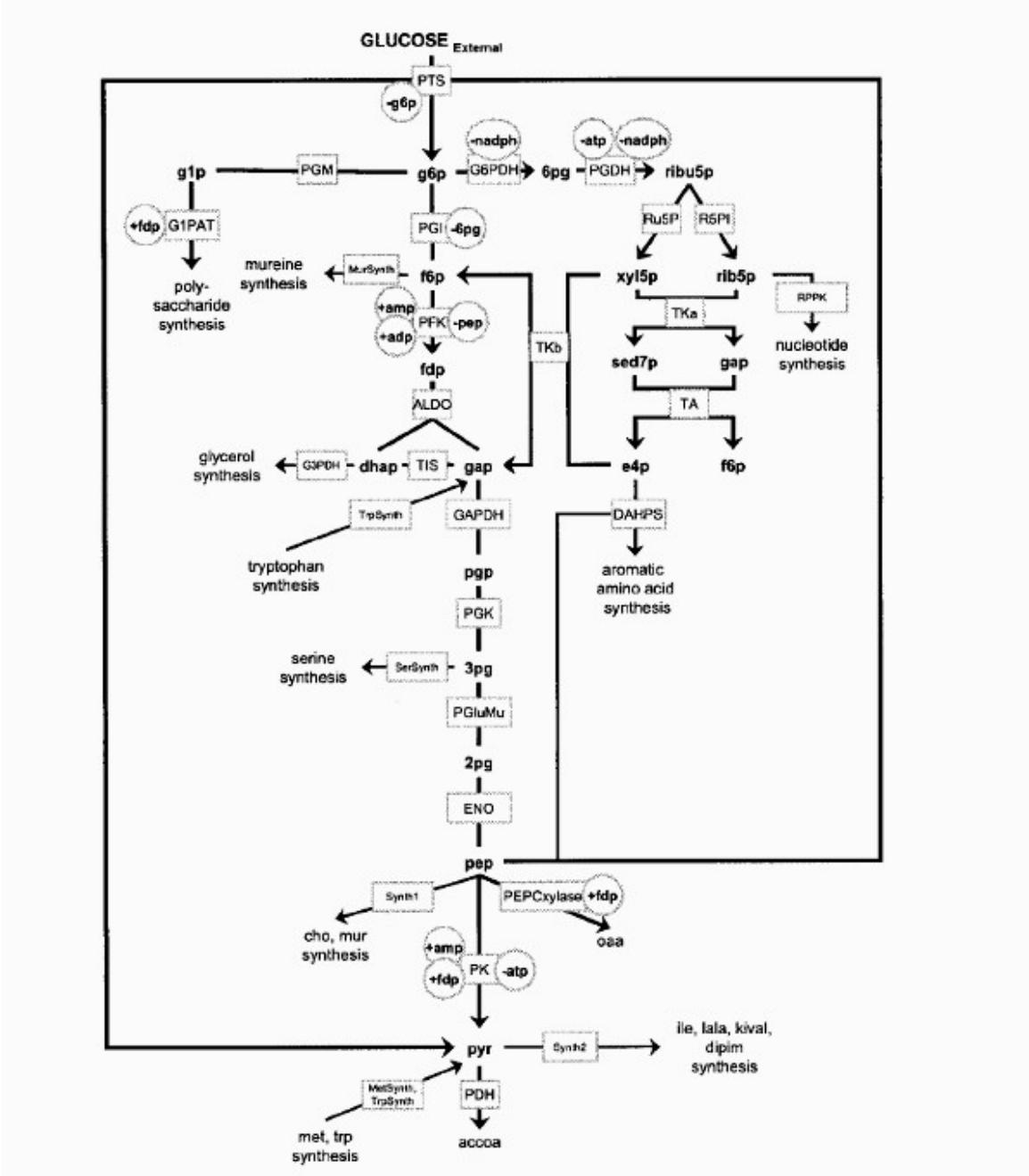


Figure 78 : Schéma du modèle cinétique de la glycolyse et de la voie des pentoses d'*E. coli*
Hormis les réactions artificielles de sortie du modèle, toutes les réactions sont associées à des réactions du modèle CBM. De plus, il existe une information sur le volume d'enzyme pour 17 d'entre elles (d'après (Chassagnole et al. 2002)).

Parmi les différentes façons d'utiliser les données expérimentales, une a retenu mon attention : *iMAT*, pour integrative Metabolic Analysis Tool (Zur et al. 2010). Cette méthode est basée sur l'utilisation de données de transcriptomique et/ou protéomique

comme contraintesde simulations. Le principe de cette méthode est d'utiliser ce type de données expérimentales pour contraindre ou au contraire forcer le passage de matière au travers des réactions. La méthode repose sur l'hypothèse qu'un transcrit ou une protéine reflète directement l'état d'activation du gène. S'ils sont en quantité abondante le gène est fortement exprimé, si au contraire la quantité est faible, le gène est faiblement exprimé ; dans les autres cas, l'expression est considérée comme normale. Ces trois états d'expression du gène (faible, normal, fort) sont répercutés sur la réaction par transitivité sur la GPR, ce qui donne trois ensembles de réactions (faiblement, normalement, fortement utilisée). L'algorithme va ensuite calculer des distributions de flux qui respectent toujours la contrainte de conservation de la matière, mais qui essaie en plus de maximiser le nombre de réactions dans leur catégorie respective (Figure 79, équation 1). Formellement le problème se présente sous la forme suivante :

$$
\begin{aligned}
&(1)\ max(\textstyle\sum_{i\in R_F}(y_i^+ + y_i^-) + \sum_{i\in R_f}(x_i)) \\
&\quad Sujet\ à \\
&(2)\ S.v = 0 \\
&(3)\ v_{min} \leq v \leq v_{max} \\
&(4)\ v_i + y_i^+(v_{min,i} - \epsilon) \geq v_{min,i} : i \in R_F \\
&(5)\ v_i + y_i^-(v_{max,i} - \epsilon) \leq v_{max,i} : i \in R_F \\
&(6)\ (1 - x_i)v_{min,i} \leq v_i \leq (1 - x_i)v_{max,i} : i \in R_f
\end{aligned}
$$

Figure 79 : Formulation du problème, résolus par la méthode *iMAT*.

Où *v* est le vecteur de flux et *S* la matrice stœchiométrique de taille *n.m*, avec *m* le nombre de métabolites et *n* le nombre de réactions. La maximisation du nombre de réactions en accord avec son état d'activation est assurée par (1). La conservation de la matière est assurée par (2). La réversibilité des réactions est prise en compte par (3). Pour chaque réaction du groupe fortement active (R_F) les booléens y^+ et y^- indiquent l'activité de la réaction respectivement dans le sens normal ou inverse, avec comme valeur 1 pour actif et 0 pour inactif. Le booléen *x* indique pour chaque réaction du groupe faiblement activé (R_f), si la réaction est active ($x=0$), ou inactive ($x=1$). (4) Une réaction fortement exprimée est considérée active si son flux est positif, et s'il est supérieur à un seuil ε, ou si son flux est négatif et que sa valeur est inférieure à l'opposé du seuil (5). Une réaction faiblement exprimée est considérée comme inactive si son flux est nul.

En sortie, *iMAT* propose un vecteur d'activité des flux qui comprend 3 niveaux (-1/0/1) correspondant respectivement à : la réaction a lieu dans le sens inverse, la réaction n'a pas lieu et la réaction a lieu dans le sens normal. En parallèle l'algorithme propose un vecteur booléen de confiance sur la prédiction de l'état du flux. En effet, puisque l'hypothèse de lien entre l'expression du gène et l'activité de la réaction peut être mise en défaut par les processus de régulation et puisque le modèle est incomplet, il se peut que dans certains cas il soit impossible d'activer le flux d'une réaction classée comme fortement active, ou au contraire de bloquer le flux d'une réaction qui est normalement faiblement active.

Cette méthode correspond au type d'étude que l'on souhaite réaliser, et j'ai décidé de l'utiliser pour intégrer les données de protéomique. Cependant, elle prend des valeurs discrètes sur trois niveaux (-1, 0,1) en entrée, et l'on dispose de valeurs continues.

.5.5 « *Proof of concept* » et perspectives

La conversion des données continues en données discrètes est facilitée par les travaux précédents i.e, les différentes figures réalisées dans l'étude et la répartition des volumes (partie 5.3). Nous avons vu qu'il existe une grande variabilité des valeurs. On ne peut donc pas utiliser un seuil minimal global en dessous duquel les gènes associés sont sous-exprimés, et un seuil maximal au dessus duquel les gènes sont surexprimés. On considère qu'entre les seuils, les gènes sont normalement exprimés. Puisqu'il est impossible de faire une conversion globale, deux options sont envisageables : prendre en compte la variation entre les souches ou entre les milieux. L'analyse des données a révélé que l'étude en fonction du milieu semble plus appropriée.
Nous allons regarder, pour une souche, les variations des volumes sur les 4 milieux, et à partir d'un seuil minimal et maximal, définir le niveau d'expression des gènes.
Quatre façons de définir ces seuils ont été étudiées : trois se basent sur la moyenne et une sur la médiane. Deux ont rapidement été éliminées : elles étaient de la forme $\mu_E \pm \alpha$ et $\mu_E \pm \alpha\mu_E$ où μ_E est la moyenne de l'enzyme sur les quatre milieux et α un coefficient. L'utilisation d'un coefficient fixe est impossible, car même si les volumes sont normalisés et corrigés, ils varient de plusieurs ordres de grandeur (entre 10^{-5} et 10^{-2}). L'autre méthode basée sur la moyenne fait intervenir l'écart type :

$$\text{(i) } \mu_E \pm \sigma/2.$$

La méthode utilisant la médiane, est elle de la forme

$$\text{(ii) } m_E \pm m_E/2.$$

La méthode (i) a tendance à être très sensible et favorise le classement des gènes en sur/sous exprimés, la méthode (ii) est beaucoup plus modérée et sera utilisée dans la simulation.
Comparé aux méthodes employées dans le chapitre précédent, *iMAT* est extrêmement long et couteux en ressources. La FBA est de l'ordre de la seconde, FVA de la minute avec la version fastFVa et la reconstruction d'un modèle de l'ordre de l'heure sur un ordinateur grand publique. *iMAT* lui est de l'ordre de plusieurs jours, sur un cluster de calcul. Son déploiement, la mise en place de la base de données de faits et la discrétisation de valeurs, m'ont laissé la possibilité d'effectuer un seul calcul. Les résultats trouvés lors de cette exécution sont cohérents. Lorsque l'on regarde les réactions qui sont considérées comme inactives sur milieux complexes, et actives sur milieux simples on trouve 91 réactions dont la plupart sont liées à la synthèse d'acides aminés. A l'inverse on trouve 4 réactions qui doivent être actives sur milieux complexes, mais inactives sur milieux simples, elles concernent le métabolisme de l'urée. Cette simulation a pour vocation de montrer la faisabilité de la méthode ; pour pratiquer une analyse détaillée, il faudrait affiner le procédé de discrétisation des volumes, notamment en étudiant plus en détail les différences entre (i) et (ii) et jouer sur les différents coefficients. Néanmoins la méthode est fonctionnelle et permet l'intégration de données protéomiques sous formes de volumes relatifs.

6 Conclusions et perspectives

Dans ce chapitre, la modélisation est mise en relation avec l'expérimentation, ce qui est l'un des objectifs de la biologie des systèmes. Si un modèle est la représentation de la connaissance et des mécanismes des processus biologiques, alors, d'une part ses prédictions doivent être en accord avec les observations, et d'autre part on doit être

capable d'intégrer des données pour affiner le modèle. En pratique, il y aura toujours des différences entre la prédiction et l'observation, puisque par définition un modèle est *« faux »*. L'intégration est souvent un exercice délicat : les données disponibles sont rarement les plus appropriées.
Le passage des modèles locaux tels que les modèles cinétiques, aux modèles globaux comme les CBMs, demande aussi un passage à une échelle supérieure pour la quantité de données nécessaires. Et la gestion de ces données devient un sujet de préoccupation. La création d'une base de données de faits permet de gérer et stocker les données, mais surtout elle permet de relier les différentes données entre elles et entre les modèles. Par exemple que ce soit les modèles, les Biologs ou les données de protéomiques, ils ont tous en commun l'utilisation d'un milieu. Le schéma de cette base, avec le choix des tables et des champs, est fait de façon à pérenniser l'homogénéisation des données. Si l'on décide d'étudier une souche particulière, on peut avoir accès à son réseau, son modèle, mais aussi aux données expérimentales la concernant. Stocker les données dans une base offre également une souplesse d'exploration de celles-ci par l'intermédiaire des requêtes. Si cette aisance est négligeable pour les Biologs, elle est cruciale pour la protéomique, et a permis de mener rapidement l'étude de la diversité des observations.
L'une des vocations des Biologs est d'estimer la qualité des modèles reconstruits et de mettre en évidence des lacunes soit au niveau du modèle lui-même, soit au niveau des connaissances biologiques. Le modèle reconstruit de la souche K-12 MG1655, offre un nombre de bonnes prédictions comparable, voire supérieur à celui du modèle de référence *iAF1260*. Parmi les autres modèles, celui de IAI1 arrive à un score similaire, alors que les deux souches différent sur certains milieux. Si dans le meilleur des cas les modèles reconstruits sont aussi bons que celui d'origine (80%), en moyenne les modèles ont un score inférieur mais très bon pour des modèles non optimisés, avec 76% de bonnes prédictions. Au delà de ces ratios, cette analyse montre que le processus entier, comprenant la création des réseaux et des modèles, permet de récupérer des voies métaboliques fonctionnelles qui ne sont pas dans l'organisme pivot et qui sont réellement présentes au niveau de l'organisme étudié. Un processus automatique entraîne forcément une perte de qualité par rapport à la référence qui est issue d'une expertise manuelle. Nous avons non seulement limité cette perte mais en plus, nous avons inclus de nouveaux métabolites et de nouvelles réactions.
L'utilisation des Biologs fait partie des analyses classiques lors de la construction d'un CBM. Par contre l'utilisation de volumes protéiques est une première. Pour cela, j'ai tout d'abord identifié les éléments du modèle auxquels ces volumes peuvent être assimilés, en l'occurrence le flux de la réaction. Puis j'ai défini les hypothèses qui permettent d'établir le lien entre le volume et le flux, à savoir un lien de proportionnalité. Enfin, il m'a fallu trouver la façon d'intégrer ces données au sein des modèles à base de contraintes. Pour cela j'ai, d'une part, exploré les données pour en déterminer les limites, les tendances et les distributions et, d'autre part, trouvé une méthode d'intégration. Cette étape a nécessité la sélection de méthodes, l'étude de la faisabilité et des limites de chacune d'elles. Avant de passer aux prédictions, une étape de discrétisation des données a aussi été nécessaire.

Un premier essai a été réalisé : il montre que la méthode est fonctionnelle et que le résultat obtenu est cohérent. Malheureusement, je n'ai pas eu la possibilité d'investiguer plus en détail cette analyse. Notamment pour être totalement

opérationnel, il faut procéder à une étude sur les fonctions de discrétisation possibles et sur leurs ajustements.
La discrétisation, au final, est une perte d'information, ce qui est regrettable. En effet une fois discrétisée il est impossible de savoir si pour deux réactions qui passent du statut « *inactive »* au statut « *active »* entre deux milieux, les variations de volume entre ces milieux sont les mêmes, ou si cette variation est d'ordre 10 pour l'une et 10^3 pour l'autre, auquel cas, on suppose que l'effet de la variation sur la valeur de leur flux respectif n'est pas le même. Il serait donc intéressant d'améliorer *iMAT* afin de pouvoir prendre des valeurs continues à la place des 3 états discrets. Mais, que ce soit d'un point de vue technique ou du point de vue réalisation et exécution, une telle méthode représente un véritable challenge.

Conclusions et perspectives

L'objectif principal des travaux réalisés pendant la thèse était l'exploration de la diversité métabolique de souches bactériennes proches au regard de leur génome.
L'étude de cette diversité par l'intermédiaire des capacités métaboliques, est utilisée depuis longtemps. L'observation des phénotypes de croissance des bactéries sur différents substrats est l'un des critères d'appartenance à une espèce bactérienne. Avec les nouvelles technologies et méthodes en biologie ou en bioinformatique, on peut aller au delà de la simple observation et expliquer pourquoi une souche peut utiliser comme source de carbone le *gluconate* tandis qu'une autre utilise l'*acétate.*
Notre intérêt pour le métabolisme fait suite à deux observations :

(i) Au niveau génomique, la diversité est telle qu'il devient difficile de trouver des signaux clairs et distincts ; le *core* génome des *E. coli* ne représente environ que10% des gènes, parmi lesquels une grande proportion est liée au métabolisme.

(ii) Parallèlement, on sait que les souches ont des capacités métaboliques différentes, capacités qui reposent sur le contenu en gènes.

A partir de (i) et (ii), nous considérons que le métabolisme est un moyen de se focaliser sur la diversité génomique, pour en extraire des liens concrets entre le mode de vie de la bactérie et des éléments génomiques.
Pour arriver à expliciter ces liens, l'ensemble des travaux réalisés dans le cadre de cette thèse repose sur une dualité entre un développement méthodologique d'outils de reconstruction de réseaux et les modèles métaboliques, et une analyse des réseaux et des modèles reconstruits pour étudier les différentes capacités métaboliques.
Ceci nous permet d'offrir à la communauté une stratégie de reconstruction automatique de réseaux et de modèles du métabolisme à l'échelle de la cellule, ainsi qu'un jeu de données sur le métabolisme des *E. coli* unique en son genre.

Conclusions relatives à la méthodologie

Aujourd'hui, les technologies de séquençage et les outils bioinformatiques permettent de séquencer, annoter et comparer automatiquement les génomes bactériens, à l'aide de méthodes dites « haut débit ». Ces comparaisons sont extrêmement fines et peuvent aller de la présence d'un opéron à des différences de nucléotides dans la séquence. La génomique comparative permet d'estimer la diversité biologique des organismes, mais ne permet pas d'expliciter les capacités et les processus biologique mis en cause : il est nécessaire d'utiliser un autre niveau d'étude, tel que les réseaux de régulation ou le métabolisme.
A partir des génomes annotés, il est maintenant possible de construire les réseaux et les modèles métaboliques à l'échelle de la cellule. Cette reconstruction peut suivre deux axes antagonistes : la qualité ou la quantité. Si le premier rend les analyses d'une finesse comparable à celle de la génomique comparative, le second assure de suivre le volume de production de plus en plus important des génomes annotés.
La méthodologie développée durant ces travaux permet d'unir ces deux axes pour obtenir rapidement des représentations du métabolisme de qualité.

Les récentes avancées dans le domaine de la reconstruction métabolique ont donné lieu à de nombreuses méthodes, souvent à l'état de *proof of concept* et plus rarement à des applications concrètes. Plutôt que de créer une n-*ième* méthode, nous nous sommes orientés vers l'utilisation des outils existants. Chacun d'eux possède des atouts et des faiblesses ; nous avons donc cherché des méthodes complémentaires : par exemple, si une méthode est sensible à la qualité des annotations, alors nous allons utiliser en amont une méthode qui homogénéise ces dernières.
La connaissance spécifique d'un organisme est généralement omise des processus de reconstruction automatiques qui reposent sur l'utilisation de bases de données généralistes. Pourtant cette connaissance représente la synthèse de travaux sur de longues années portant sur l'organisme en question ; c'est donc une source d'information différente, mais aussi importante, sinon plus, que celle des bases de données généralistes.

Même si, lors de l'élaboration de la stratégie de reconstruction, nous savions sur quelle espèce nous allions l'appliquer, nous avons volontairement travaillé de façon générique pour rendre ce travail indépendant des données et flexible par rapport aux méthodes existantes. L'apport méthodologique principal est la mise en relation, au sein d'une même stratégie, de données généralistes concernant un grand nombre d'organismes et de données spécifiques d'un organisme appelé *pivot*. S'il existe un réseau (ou modèle) de référence pour un organisme, alors notre stratégie est capable de le réutiliser et de transférer cette synthèse de connaissances sur des organismes phylogénétiquement proches, créant une première ébauche de réseaux (ou modèles) pour ces organismes. Cette ébauche permet d'améliorer le résultat des méthodes d'inférence automatique qui disposent ainsi d'une base solide pour le calcul et l'estimation des parties spécifiques des réseaux (modèles) en reconstruction.

Si la contribution algorithmique n'est pas des plus importantes, la stratégie créée et les synergies des méthodes utilisées apportent une réelle plus-value par rapport à ce qui existe actuellement (Orth et al. 2011; Henry et al. 2010).

La stratégie mise au point peut être appliquée aussi bien aux réseaux qu'aux modèles, à *E. coli* ou toute autre espèce pour laquelle une souche de référence existe. Si la stratégie reste identique, les implications du pivot diffèrent lors de la reconstruction des réseaux et des modèles. Dans le cas des réseaux, le pivot sert à améliorer les résultats des méthodes d'inférence automatique et dans le cas des modèles, il sert de base fonctionnelle à modifier, ce qui permet d'obtenir des modèles capables de produire de la biomasse sur milieu riche. Ces deux utilisations ne sont pas les seules possibles, et le rôle du pivot est à définir lors de la mise en place des processus.
Ces travaux nous permettent d'offrir à la communauté une stratégie de reconstruction efficace et flexible qui peut être utilisée chez différentes espèces. Nous proposons une implémentation fonctionnelle de cette stratégie qui peut être directement utilisée ou adaptée en fonction de méthodes disponibles.
Nous proposons également une centaine de réseaux et une vingtaine de modèles métaboliques d'*E. coli* à l'échelle de la cellule qui, pour la première fois, ne sont pas de simples projections du réseau ou du modèle de référence. Ils bénéficient d'un important travail d'homogénéisation qui permet de passer sans ambigüité du génome au réseau puis au modèle et de réaliser des comparaisons entre les organismes sans aucune difficulté.

Dans son cadre d'utilisation (i.e., les souches d'une même espèce), notre stratégie fournit des résultats bien meilleurs que les processus usuels (chapitre II partie 3.1). Hors de son cadre, elle reste dans le pire des cas au niveau des autres méthodes. L'utilisation de différentes méthodes permet de compenser les faiblesses des unes et des autres. Néanmoins, notre processus reste sensible à la qualité de finition du génome et de l'annotation de celui-ci, même si le processus peut quand même être appliqué à des génomes *« draft »* (non fini). Il est également dépendant des données référencées et des informations disponibles (numéro EC, formule, équation bilan, etc.). Ces limites ne sont pas spécifiques à notre stratégie, mais à toute modélisation et à la biologie des systèmes.
A priori, nous pensions que le choix des méthodes à intégrer aux processus de reconstruction allait constituer le point le plus sensible des travaux. Nous savons maintenant que ce sont les données qui fragilisent la reconstruction et l'automatisation. Il ne s'agit pas ici des connaissances contenues dans les bases, mais de l'exploitation de celles-ci. Si les schémas des bases de données sont proches et cohérents dans leurs concepts, il existe un véritable cloisonnement des ressources entre les différents niveaux d'études (génomes, réseaux métaboliques et modèles métaboliques) et aussi au sein d'un même niveau entre deux bases de données différentes. Il est parfois difficile, voire impossible, de trouver les gènes qui codent l'enzyme d'une réaction, ou pire, d'identifier la même réaction entre deux bases de données. Ces difficultés sont des freins aussi bien pour la reconstruction automatique que pour l'analyse des réseaux et modèles reconstruits. C'est pourquoi nous avons apporté un soin particulier à l'homogénéisation des données et à l'établissement de références croisées.

Conclusions relatives aux analyses

Le développement des processus de reconstruction des réseaux et modèles métaboliques n'est pas une fin en soi, mais une étape indispensable pour réaliser l'étude de la diversité métabolique chez les *E. coli.*

Etant donné la disponibilité tardive d'une partie des génomes, notre étude principale a porté sur 23 souches d'*E. Coli* et 6 *Shigella* bien que nous ayons reconstruit une centaine de réseaux métaboliques. Nous avons estimé les capacités métaboliques des modèles des *E. coli* dont les souches choisies représentent l'ensemble des groupes phylogénétiques de l'espèce, ainsi que les différents pathovars possibles.

Nous ne disposons pas à ce jour de méthode fiable pour estimer le nombre total de réactions d'un organisme. On peut tout de même affirmer que la diversité métabolique est beaucoup plus restreinte que la diversité génomique. Cette diminution de la diversité globale n'est pas synonyme d'absence de diversité dans les différents réseaux ; au contraire cette diversité est bien présente. Le métabolisme est un niveau d'étude qui permet de se concentrer sur une partie de la diversité génomique. Ce « focus » a permis de mettre en évidence certaines réactions et donc certains gènes. Au niveau génomique, ces mêmes gènes sont perdus au sein de la diversité génétique globale.
On distingue les voies ubiquitaires, dont en particulier, des voies de synthèse des métabolites essentiels qui sont fortement conservées dans toutes les souches, des voies de dégradation qui sont beaucoup plus éparses. L'analyse de cette partie moins conservée entre les différents réseaux a mis en évidence un lien fort entre le métabolisme et la phylogénie. Il existe une très forte corrélation entre la distance métabolique et la distance génétique. Pourtant, cette dernière est calculée sur les gènes communs à toutes les souches tandis que la distance métabolique se base sur la différence du contenu en réactions. Ce lien implique des apparitions/disparitions de réactions assez tardives dans l'histoire évolutive de l'espèce. La corrélation entre l'évolution du métabolisme et des génomes est dépendante du mode de vie de l'organisme : pour les *Shigella*, nous avons constaté la disparition de ce lien suite à une levée des contraintes métaboliques imposées par la prototrophie.

Les modèles permettent de quantifier les différences observées dans les réseaux, et bien que les modèles soient très proches les uns des autres, de quantifier les quelques différences de voies métaboliques qui influent sur le flux de biomasse. Une grande partie des réactions du core métabolisme possèdent un flux non nul sur un des milieux de simulation. Cette observation est le reflet des pressions de sélection et de l'obligation de la présence des voies de synthèse des métabolites essentiels (acides aminés, nucléotides etc.) à partir des métabolites ubiquitaires.

L'aspect quantitatif des modèles autorise une approche globale du métabolisme basée sur les flux de matière qui traversent les réactions, et en particulier le flux de biomasse qui représente la capacité du modèle à vivre. Ils permettent aussi une approche focalisée sur une ou quelques réactions pour étudier l'effet de chemins alternatifs. Il ne faut pas oublier que les voies de dégradation décrites dans les bases de données sont définies par l'Homme et que l'union de deux voies métaboliques distinctes peut engendrer, par combinatoire, de nouveaux chemins pour aller d'un

métabolite source à un métabolite produit. Ces voies sous jacentes et non référencées dans la littérature sont un facteur important de la flexibilité et de l'adaptabilité du métabolisme.

Le métabolisme, dans certains cas, est un excellent reflet de l'histoire évolutive des organismes. Les réseaux et modèles métaboliques permettent d'expliciter des liens entre le génome et le phénotype, notamment au travers des contraintes métaboliques indispensables à la vie de l'organisme ; ces liens ne peuvent être déduits à partir d'une étude de génomique comparative seule.
L'utilisation de données expérimentales a validé l'inférence de nouvelles voies métaboliques capables de faire transiter un flux de matière non nul. Ces données ont aussi montré que, dans certains cas, bien que tous les acteurs soient présents, le flux de matière est nul, ce qui laisse supposer des effets de régulation. Le métabolisme est un niveau d'étude intéressant puisqu'il filtre la diversité génomique. L'étude du réseau de régulation est un autre filtre qui apporte lui aussi des liens entre le phénotype et le génotype. C'est pourquoi les CBMs essaient d'intégrer d'autres niveaux d'études au métabolisme pour synthétiser, au sein d'un même objet d'étude, l'intégralité des informations disponibles.

Perspectives

Pour conclure ce travail, je vais évoquer plusieurs pistes qui permettraient de poursuivre ou améliorer ce qui a été réalisé. Mais avant tout, il me semble important de replacer brièvement ce travail dans le paysage de la recherche actuelle.
La reconstruction automatique des réseaux, à l'échelle de la bactérie, a commencé au début des années 2000 (Peter D Karp, S. Paley, et al. 2002). Depuis, même si différentes approches ont vu le jour, la méthode proposée Pathway-Tools, l'une des plus utilisée, reste depuis inchangée.
En parallèle de cette thèse, la première méthode pour reconstruire automatiquement des CBMs de sous processus biologiques était mise au point, et c'est fin 2010 que cette dernière a été propagée à l'échelle de l'organisme (DeJongh et al. 2007; Henry et al. 2010).
En fonction de l'avancée des connaissances, les modèles de référence sont en constante amélioration, et dernièrement le modèle d'*E. coli* a été mis à jour (Orth et al. 2011). Il est accompagné d'une projection du modèle sur différentes souches d'*E. coli*, sans ajout de connaissances spécifiques.
D'autres projets ont débuté pendant le déroulement de cette thèse, tel que le projet Européen Microme (Janvier 2010) qui reprend et développe des axes similaires à mes travaux : reconstruction automatique des réseaux et des modèles métaboliques bactériens à l'échelle de la cellule à partir de génomes séquencés et annotés. Ces réseaux et modèles reposeront sur un travail de caractérisation précis des éléments contenus dans les différentes bases de données généralistes du métabolisme.

Les méthodes destinées à l'utilisation des CBMs se développent également. Certaines ont été publiées pendant mes analyses et m'ont permis d'augmenter le nombre de comparaisons possibles (*fastFVa* (Gudmundsson & Thiele 2010)), ou bien d'intégrer des données de protéomiques (*iMAT* (Zur et al. 2010)). Puisque de plus en plus de CBMs sont disponibles, des bases de données leur étant dédiées ont été, et sont toujours, développées (F Le Fèvre et al. 2009; Schellenberger et al. 2010) et le projet Microme); NemoStudio offre la possibilité d'effectuer des simulations et des comparaisons avec des phénotypes de croissance.
Les travaux réalisés et les méthodes proposées sont non seulement d'actualité mais aussi en parfaite adéquation avec les développements méthodologiques du domaine.

L'utilisation de réseaux et de modèles du métabolisme à l'échelle de la cellule est récente. Par conséquent, nous sommes toujours dans une phase de développement des méthodes de reconstruction et d'analyse. S'il est, pour le moment, difficile de dire quelles seront les méthodes utilisées dans le futur, il est quand même possible d'estimer les limites et les points à améliorer dans les processus et les analyses actuels.
Un premier axe peut être consacré à l'amélioration des réseaux et des modèles. Les réseaux étant incomplets il serait intéressant de combler le manque de connaissance, que ce soit par l'identification de nouvelles réactions et voies métaboliques, ou l'identification des gènes codants des activités enzymatiques orphelines de séquence.
Ces dernières années des projets ont vu le jour pour caractériser ces nouvelles activités enzymatiques, dont certains sont hébergés au Genoscope : analyse de voies de dégradation des acides aminés en anaérobie et une étude de la diversité enzymatique au sein d'une famille de protéines (la famille des Beta-Keto-Acid

Cleaving Enzymes, BKACE). Ces projets sont d'ailleurs couplés avec le projet *CANOE* d'identification de gènes pour les réactions orphelines de séquence: il s'agit d'une méthode bioinformatique basée sur l'étude des contextes métaboliques et génomiques (Adam S. A. et al, en révision). Il existe aussi une autre approche visant à automatiser la recherche de gènes pour des activités enzymatiques orphelines (King et al. 2009).
Dans notre problématique nous pourrions également envisager la curation manuelle d'un des réseaux, de préférence d'une souche éloignée phylogénétiquement de K-12, pour construire un second réseau pivot.
En couplant les données curées qui seront fournies par le projet Microme avec notre stratégie, puis en utilisant une méthode de type *GapFill* (Satish Kumar et al. 2007) pour relier les réactions isolées, la reconstruction des réseaux et modèles du métabolisme à l'échelle de la cellule pourrait devenir systématique après le processus d'annotation d'un nouveau génome.

Au début de mes travaux, les comparaisons entre différents réseaux métaboliques étaient basées d'un côté, sur des aspects topologiques du graphe métabolique, et de l'autre sur la complétion des voies métaboliques. Avec l'augmentation du nombre de réseaux et surtout de leur qualité, d'autres critères et méthodes d'analyses vont apparaître. Nous avons introduit l'utilisation de l'ACM, une technique statistique parmi d'autres, qui a mis en évidence le lien entre le métabolisme et l'évolution. Il est difficile de donner ou même d'estimer les analyses qui verront le jour ; c'est l'objet de l'étude qui dictera une ligne de conduite. Dans notre cas, la création de l'arbre métabolique s'est avérée appropriée, mais pour l'étude des parasites comme les *Shigella* cette méthodologie est inadaptée. La comparaison des modèles est beaucoup plus délicate et, du fait de la nature de l'espace des solutions, elle dépasse le cadre de la bioinformatique et de la biomathématique traditionnelle. Elle nécessite une collaboration entre biologistes, informaticiens et statisticiens pour mettre au point des méthodes d'exploration de l'espace des solutions et ainsi en déduire des ressemblances entre les modèles.
Il est possible de s'intéresser à la flexibilité du métabolisme par l'intermédiaire du couplage des flux (Burgard et al. 2004). Cette méthode consiste à définir la corrélation entre deux flux au sein d'un modèle. Etudier la variabilité du couplage des paires de flux entre deux modèles permet de mettre en évidence les différences de contrainte imposées par la structure du réseau métabolique, et ainsi en déduire des réactions responsables de la flexibilité métabolique.
Les méthodes d'intégration de données vont se développer au fur et à mesure que les technologies expérimentales à haut débit vont apparaître. Aujourd'hui, la transcriptomique est devenue plus facilement accessible et la méthode *iMAT* a vu le jour. Demain, si la fluxomique devient elle aussi une expérience à haut débit, une méthode d'intégration sera mise au point. C'est d'ailleurs l'une des grandes forces des CBMs : ils n'imposent pas de développement expérimental, mais s'adaptent afin de suivre les évolutions des nouvelles technologies pour ajouter de nouveaux types de contraintes et affiner de plus en plus l'espace de solutions des modèles à bases de contraintes. L'utilisation de différents types d'observation comme contraintes dans une même simulation, est une extension des CBMs qui ne saurait tarder. Il ne faut pas oublier que le modèle est une approximation de la réalité et l'utilisation simultanée de plusieurs types de données peut rendre l'espace de solutions nul. Dès lors, il devient indispensable de ne plus travailler dans un cadre déterministe, mais d'autoriser une

certaine souplesse sur les contraintes en passant dans un cadre probabiliste. Une première approche de ce genre est expérimentée dans l'ANR MetaColi.

Un aspect qui n'a pas été évoqué pour le moment est le retour des résultats des CBMs, pour affiner les connaissances. Par exemple, nous avons vu, entre autre, le cas d'une réaction « faux positif » et d'une erreur de séquençage (Chapitre II partie 4.2.2). Nul doute que ce ne sont pas les seules erreurs, et les résultats des simulations peuvent être une nouvelle source de données des méthodes d'identification des réactions orphelines ou même de voies métaboliques. Nous avons bien vu que certains composés assimilables par la bactérie ne figurent pas dans les modèles ; dès lors, ils peuvent alors servir pour trouver des voies de dégradations fonctionnelles.

Enfin, on peut tout simplement appliquer la stratégie et les analyses sur une autre espèce, par exemple *Acinetobacter baylyi* qui est l'organisme modèle du Genoscope pour l'analyse du métabolisme. Un réseau et un modèle de référence sont déjà disponibles, ainsi que les génomes séquencés et annotés de plusieurs souches du même genre.

L'utilisation des CBMs permet aller plus loin que la simple intégration des connaissances et l'étude de la diversité. Ils sont également utilisés dans la biologie de l'ingénierie et la biologie de synthèses. Deux domaines en pleine expansion avec notamment les thématiques d'énergie propre et renouvelable. La création d'une plate-forme de manipulation des modèles, couplée à une base de faits biologiques, pourrait être un outil extrêmement puissant pour la modification, le *« design »* et l'optimisation de voies métaboliques dans un objectif précis. Pousser à l'extrême cette plate-forme pourrait devenir un logiciel de type *conception assistée par ordinateur*, un équivalent pour la biologie de l'ingénierie du célèbre logiciel *CATIA*[7] de Dassault Systèmes.

[7]http://www.3ds.com/

Références

Akerley, B.J. et al., 2002. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2), pp.966-971.

Almaas, E. et al., 2004. Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature*, 427(6977), pp.839-843.

Arakawa, K. et al., 2006. GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics*, 7, p.168.

Arkin, A., Ross, J. & McAdams, H.H., 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, 149(4), pp.1633-1648.

Baba, T. et al., 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2, p.2006.0008.

Bairoch, A., 2000. The ENZYME database in 2000. *Nucleic Acids Research*, 28(1), pp.304-305.

Baker, W. et al., 2000. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 28(1), pp.19-23.

Barabási, Albert-László & Albert, R., 1999. Emergence of Scaling in Random Networks. *Science*, 286(5439), pp.509 -512.

Beard, D.A. et al., 2004. Thermodynamic constraints for biochemical networks. *Journal of Theoretical Biology*, 228(3), pp.327-333.

Becker, S.A. et al., 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols*, 2(3), pp.727-738.

Benson, D.A. et al., 2010. GenBank. *Nucleic Acids Research*, 38(Database issue), pp.D46-51.

de Berardinis, V. et al., 2008. A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Molecular Systems Biology*, 4, p.174.

Berg, J.M., Tymoczko, J.L. & Stryer, L., 2006. *Biochemistry (Biochemistry* Sixth Edition., W. H. Freeman.

Blattner, F.R. et al., 1997. The complete genome sequence of Escherichia coli K-12. *Science (New York, N.Y.)*, 277(5331), pp.1453-1462.

Borodina, I. & Nielsen, J., 2005. From genomes to in silico cells via metabolic

networks. *Current Opinion in Biotechnology*, 16(3), pp.350-355.

Box, G.E.P. & Draper, N.R., 1987. *Empirical Model-Building and Response Surfaces*, Wiley.

Brasch, M.A., Hartley, J.L. & Vidal, M., 2004. ORFeome cloning and systems biology: standardized mass production of the parts from the parts-list. *Genome Research*, 14(10B), pp.2001-2009.

Briggs George Edward & Haldane John Burdon Sanderson, 1925. A Note on the Kinetics of Enzyme Action.

Bumgarner, R.E. & Yeung, K.Y., 2009. Methods for the Inference of Biological Pathways and Networks. In R. Ireton et al., eds. *Computational Systems Biology*. Totowa, NJ: Humana Press, pp. 225-245.

Burgard, A.P. et al., 2004. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research*, 14(2), pp.301-312.

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

Castellani, S.A. & Chalmers, A.J., 1919. *Manual of tropical medicine* 3rd ed., Baillière, Tindall and Cox.

Chang, D.-E. et al., 2004. Carbon nutrition of Escherichia coli in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19), pp.7427-7432.

Chassagnole, C. et al., 2002. Dynamic modeling of the central carbon metabolism of Escherichia coli. *Biotechnology and Bioengineering*, 79(1), pp.53-73.

Chen, L. & Vitkup, D., 2006. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biology*, 7(2), p.R17.

Cherry, J.M. et al., 1998. SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, 26(1), pp.73-79.

Cooke, E.M., Hettiaratchy, I.G. & Buck, A.C., 1972. Fate of ingested Escherichia coli in normal persons. *Journal of Medical Microbiology*, 5(3), pp.361-369.

Covert, M.W. & Palsson, B.Ø., 2002. Transcriptional regulation in constraints-based metabolic models of Escherichia coli. *J Biol Chem*, 277(31), pp.28058–28064.

Cozzone, A.J., 1998. Regulation of acetate metabolism by protein phosphorylation in enteric bacteria. *Annual Review of Microbiology*, 52, pp.127-164.

Croxen, M.A. & Finlay, B.B., 2010. Molecular mechanisms of Escherichia coli pathogenicity. *Nature Reviews. Microbiology*, 8(1), pp.26-38.

DeJongh, M. et al., 2007. Toward the automated generation of genome-scale

metabolic networks in the SEED. *BMC Bioinformatics*, 8, p.139.

Díaz-Guerra, M., Esteban, M. & Martínez, J.L., 1997. Growth of Escherichia coli in acetate as a sole carbon source is inhibited by ankyrin-like repeats present in the 2',5'-linked oligoadenylate-dependent human RNase L enzyme. *FEMS Microbiology Letters*, 149(1), pp.107-113.

Dickinson, J.R. & Schweizer, M., 2004. *Metabolism and Molecular Physiology of Saccharomyces Cerevisiae, 2nd Edition* 2nd ed., CRC Press.

Van Dien, S.J. & Keasling, J.D., 1998. A dynamic model of the Escherichia coli phosphate-starvation response. *Journal of Theoretical Biology*, 190(1), pp.37-49.

Duarte, N.C., Herrgård, M.J. & Palsson, B.Ø., 2004. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, 14(7), pp.1298-1309.

Duchesne, E., 1912. La thèse de médecine oubliée du docteur Ernest Duchesne 1874-1912 = The thesis of medecine forgotten of doctor Ernest Duchesne 1874-1912.

Durot, Maxime, 2009. *Elucidation du métabolisme des microorganismes par la modélisation et l'interprétation des données d'essentialités de gènes : application au métabolisme de la bactérie Acinetobacter baylyi ADP1*, Evry.

Durot, Maxime, Bourguignon, P.-Y. & Schachter, Vincent, 2009. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, 33(1), pp.164-190.

Durot, Maxime et al., 2008. Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Systems Biology*, 2, p.85.

Edelson, E., 2001. *Gregor Mendel: And the Roots of Genetics*, Oxford University Press.

Edwards, J S, Ibarra, R U & Palsson, B O, 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, 19(2), pp.125-130.

Edwards, J S & Palsson, B O, 2000. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10), pp.5528-5533.

Falb, M. et al., 2005. Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis. *Genome Research*, 15(10), pp.1336-1343.

Feist, A.M. et al., 2007. A genome-scale metabolic reconstruction for Escherichia coli

K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, 3, p.121.

Feist, A.M. et al., 2009. Reconstruction of biochemical networks in microorganisms. *Nature Reviews. Microbiology*, 7(2), pp.129-143.

Fell, D A & Small, J.R., 1986. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *The Biochemical Journal*, 238(3), pp.781-786.

Fell, D A, 1992. Metabolic control analysis: a survey of its theoretical and experimental development. *The Biochemical Journal*, 286 (Pt 2), pp.313-330.

Le Fèvre, F et al., 2009. CycSim--an online tool for exploring and experimenting with genome-scale metabolic models. *Bioinformatics (Oxford, England)*, 25(15), pp.1987-1988.

Fields, S. & Johnston, M., 2005. Whither Model Organism Research? *Science*, 307(5717), pp.1885 -1886.

Fischer, E. & Sauer, Uwe, 2003. Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. *European Journal of Biochemistry / FEBS*, 270(5), pp.880-891.

Le Gall, T. et al., 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains. *Molecular Biology and Evolution*, 24(11), pp.2373-2384.

Gay-Lussac, J.L. et al., 1833. *Annales de chimie et de physique*, Masson.

Gevorgyan, A., Poolman, M.G. & Fell, David A, 2008. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics (Oxford, England)*, 24(19), pp.2245-2251.

Ghaemmaghami, S. et al., 2003. Global analysis of protein expression in yeast. *Nature*, 425(6959), pp.737-741.

Gibney, M.J. et al., 2005. Metabolomics in human nutrition: opportunities and challenges. *The American Journal of Clinical Nutrition*, 82(3), pp.497-503.

Gillespie, D.T., 2007. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58, pp.35-55.

Goesmann, A. et al., 2002. PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics (Oxford, England)*, 18(1), pp.124-129.

Gonzalez, O. et al., 2010. Characterization of growth and metabolism of the haloalkaliphile Natronomonas pharaonis. *PLoS Computational Biology*, 6(6), p.e1000799.

Gonze, D., Halloy, J. & Goldbeter, A., 2003. Modèles déterministes et stochastiques pour les rythmes circadiensDeterministic and stochastic models for circadian rhythms. *Pathologie Biologie*, 51(4), pp.227-230.

Gordon, D.M. & Cowling, A., 2003. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. *Microbiology (Reading, England)*, 149(Pt 12), pp.3575-3586.

Green, M. & Karp, P., 2004. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5(1), p.76.

Guarino, N., 1998. Formal Ontology and Information Systems. , p.3--15.

Gudmundsson, S. & Thiele, I., 2010. Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11, p.489.

Guerrier-Takada, C. et al., 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2), pp.849-857.

Güldener, U. et al., 2005. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Research*, 33(Database issue), pp.D364-368.

Hardiman, G., 2004. Microarray platforms--comparisons and contrasts. *Pharmacogenomics*, 5(5), pp.487-502.

Hardy, G.H., 1908. MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science*, 28(706), pp.49 -50.

Heinrich, R. & Rapoport, T.A., 1974. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *European Journal of Biochemistry / FEBS*, 42(1), pp.89-95.

Henriques Normark, B. & Normark, S., 2002. Antibiotic tolerance in pneumococci. *Clinical Microbiology and Infection*, 8(10), pp.613-622.

Henry, C.S. et al., 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech*, 28(9), pp.977-982.

Henry, C.S. et al., 2006. Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophysical Journal*, 90(4), pp.1453-1461.

Heuner, K. & Swanson, M., 2008. *Legionella: molecular microbiology*, Horizon Scientific Press.

Hobman, J.L., Penn, C.W. & Pallen, M.J., 2007. Laboratory strains of Escherichia coli: model citizens or deceitful delinquents growing old disgracefully? *Molecular Microbiology*, 64(4), pp.881-885.

Hornberg, J.J. et al., 2007. Metabolic control analysis to identify optimal drug targets. *Progress in Drug Research. Fortschritte Der Arzneimittelforschung. Progrès Des Recherches Pharmaceutiques*, 64, pp.171, 173-189.

Horton, H.R. et al., 1994. *Principes de biochimie*, De Boeck Université.

Hucka, M. et al., 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*

(Oxford, England), 19(4), pp.524-531.

Huh, W.-K. et al., 2003. Global analysis of protein localization in budding yeast. *Nature*, 425(6959), pp.686-691.

Hussein, S., Hantke, K. & Braun, V., 1981. Citrate-dependent iron transport system in Escherichia coli K-12. *European Journal of Biochemistry / FEBS*, 117(2), pp.431-437.

Ibarra, Rafael U., Edwards, Jeremy S. & Palsson, Bernhard O., 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912), pp.186-189.

Imieliński, M. et al., 2005. Investigating metabolite essentiality through genome-scale analysis of Escherichia coli production capabilities. *Bioinformatics (Oxford, England)*, 21(9), pp.2008-2016.

Jaureguy, F. et al., 2008. Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains. *BMC Genomics*, 9, p.560.

Jensen, R.A., 1976. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology*, 30(1), pp.409-425.

Jerby, L., Shlomi, T. & Ruppin, E., 2010. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular Systems Biology*, 6, p.401.

Kacser, H. & Burns, J.A., 1973. The control of flux. *Symposia of the Society for Experimental Biology*, 27, pp.65-104.

Kanehisa, M. et al., 2007. KEGG for linking genomes to life and the environment. *Nucl. Acids Res.*, 36, p.D480–D484.

Karp, P D et al., 1999. Eco Cyc: encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Research*, 27(1), pp.55-58.

Karp, Peter D, Paley, S. & Romero, P., 2002. The Pathway Tools software. *Bioinformatics (Oxford, England)*, 18 Suppl 1, pp.S225-232.

Karp, Peter D, Riley, Monica, et al., 2002. The MetaCyc Database. *Nucleic Acids Research*, 30(1), pp.59-61.

Karp, Peter D., 2000. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16(3), pp.269 -285.

Kauffman, K.J., Prakash, P. & Edwards, Jeremy S, 2003. Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5), pp.491-496.

Keseler, I.M. et al., 2009. EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Research*, 37(Database issue), pp.D464-70.

Keseler, I.M. et al., 2011. EcoCyc: a comprehensive database of Escherichia coli

biology. *Nucleic Acids Research*, 39(Database issue), pp.D583-590.

Kharchenko, P. et al., 2006. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, 7, p.177.

Kharchenko, P., Vitkup, D. & Church, G.M., 2004. Filling gaps in a metabolic network using expression information. *Bioinformatics (Oxford, England)*, 20 Suppl 1, pp.i178-185.

Kim, S.H., Schneider, Barbara L & Reitzer, L., 2010. Genetics and regulation of the major enzymes of alanine synthesis in Escherichia coli. *Journal of Bacteriology*, 192(20), pp.5304-5311.

King, R.D. et al., 2009. The automation of science. *Science (New York, N.Y.)*, 324(5923), pp.85-89.

Kitagawa, M. et al., 2005. Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 12(5), pp.291-299.

Kitano, Hiroaki, 2002. Systems Biology: A Brief Overview. *Science*, 295(5560), pp.1662 -1664.

Klamt, S. & Stelling, Jörg, 2003. Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 21(2), pp.64-69.

Klamt, S., Saez-Rodriguez, J. & Gilles, Ernst D, 2007. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology*, 1, p.2.

Knoll, A.H., 2003. The geological consequences of evolution. *Geobiology*, 1(1), pp.3-14.

Kobayashi, K. et al., 2003. Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8), pp.4678-4683.

ter Kuile, B.H. & Westerhoff, H V, 2001. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Letters*, 500(3), pp.169-171.

Kümmel, A., Panke, S. & Heinemann, M., 2006. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*, 7, p.512.

Lee, J., Yun, H., et al., 2008. Genome-scale reconstruction and in silico analysis of the Clostridium acetobutylicum ATCC 824 metabolic network. *Applied Microbiology and Biotechnology*, 80(5), pp.849-862.

Lee, J.M., Min Lee, J., et al., 2008. Dynamic analysis of integrated signaling,

metabolic, and regulatory networks. *PLoS Computational Biology*, 4(5), p.e1000086.

Lee, S.Y. et al., 2005. Systems-level analysis of genome-scalein silico metabolic models using MetaFluxNet. *Biotechnology and Bioprocess Engineering*, 10(5), pp.425-431.

Liberati, N.T. et al., 2006. An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8), pp.2833-2838.

Lima-Mendez, G. & van Helden, Jacques, 2009. The powerful law of the power law and other myths in network biology. *Molecular bioSystems*, 5(12), pp.1482-1493.

Liu, T. et al., 2007. Accurate Mass Measurements in Proteomics. *Chemical reviews*, 107(8), pp.3621-3653.

Looman, J., 1976. Biological Equilibrium in Ecosystems 1. A Theory of Biological Equilibrium. *Folia Geobotanica & Phytotaxonomica*, 11(1), pp.1-21.

Macpherson, A.J. et al., 2000. A primitive T cell-independent mechanism of intestinal mucosal IgA responses to commensal bacteria. *Science (New York, N.Y.)*, 288(5474), pp.2222-2226.

Maglott, D. et al., 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(Database issue), pp.D52-57.

Mahadevan, R & Schilling, C H, 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4), pp.264-276.

Majewski, R.A. & Domach, M.M., 1990. Simple constrained-optimization view of acetate overflow in E. coli. *Biotechnology and Bioengineering*, 35(7), pp.732-738.

Markuszewski, M.J. et al., 2005. Human red blood cells targeted metabolome analysis of glycolysis cycle metabolites by capillary electrophoresis using an indirect photometric detection method. *Journal of Pharmaceutical and Biomedical Analysis*, 39(3-4), pp.636-642.

Mauchline, W.S. & Keevil, C.W., 1991. Development of the BIOLOG substrate utilization system for identification of Legionella spp. *Applied and Environmental Microbiology*, 57(11), pp.3345-3349.

Michaelis, L. & Menten, M., 1913. Die kinetik der invertinwirkung. *Biochem. Z*, 49(333-369), p.352.

Milletti, F. et al., 2010. Extending pKa prediction accuracy: high-throughput pKa measurements to understand pKa modulation of new chemical series.

European Journal of Medicinal Chemistry, 45(9), pp.4270-4279.

Mobley, H.L.T., Mendz, G.L. & Hazell, S.L., 2001. *Helicobacter pylori: physiology and genetics*, ASM Press.

Nataro, J.P. & Kaper, J.B., 1998. Diarrheagenic Escherichia coli. *Clinical Microbiology Reviews*, 11(1), pp.142-201.

Neidhardt, F., 1996. *Escherichia coli and Salmonella : cellular and molecular biology* 2nd ed., Washington D.C.: ASM Press.

Neidhardt, F.C. & Umbarger, H.E., 1996. Chemical Composition of Escherichia coli. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Washington D.C.: ASM Press.

Notebaart, R.A. et al., 2006. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics*, 7, p.296.

Notebaart, R.A. et al., 2008. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Computational Biology*, 4(1), p.e26.

Le Novere, N., 2006. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34(90001), p.D689-D691.

Oh, Y.-K. et al., 2007. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *The Journal of Biological Chemistry*, 282(39), pp.28791-28799.

Oliveira, A.P., Nielsen, J. & Förster, J., 2005. Modeling Lactococcus lactis using a genome-scale flux model. *BMC Microbiology*, 5, p.39.

Orth, J.D. et al., 2011. A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. *Molecular Systems Biology*, 7, p.535.

Pál, C., Papp, B. & Lercher, M.J., 2005a. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37(12), pp.1372-1375.

Pál, C., Papp, B. & Lercher, M.J., 2005b. Horizontal gene transfer depends on gene content of the host. *Bioinformatics (Oxford, England)*, 21 Suppl 2, pp.ii222-223.

Papin, J.A. et al., 2002. The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. *Journal of Theoretical Biology*, 215(1), pp.67-82.

Papin, J.A. et al., 2004. Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, 22(8), pp.400-405.

Papoutsakis, E.T., 1984. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and Bioengineering*, 26(2), pp.174-187.

Papoutsakis, E.T. & Meyer, C.L., 1985. Fermentation equations for propionic-acid bacteria and production of assorted oxychemicals from various sugars. *Biotechnology and Bioengineering*, 27(1), pp.67-80.

Patil, K. et al., 2005. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6(1), p.308.

Patterson, S.D. & Aebersold, R.H., 2003. Proteomics: the first decade and beyond. *Nature Genetics*, 33 Suppl, pp.311-323.

Penders, J. et al., 2006. Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics*, 118(2), pp.511-521.

Pfeiffer, T., Soyer, O.S. & Bonhoeffer, S., 2005. The evolution of connectivity in metabolic networks. *PLoS Biol*, 3(7), p.e228.

Pharkya, P. & Maranas, C.D., 2006. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic Engineering*, 8(1), pp.1-13.

Pharkya, P., Burgard, A.P. & Maranas, C.D., 2004. OptStrain: A computational framework for redesign of microbial production systems. *Genome Research*, 14(11), pp.2367-2376.

Picard, B et al., 1999. The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infection and Immunity*, 67(2), pp.546-553.

Pinney, J.W. et al., 2005. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucleic Acids Research*, 33(4), pp.1399-1409.

Price, N.D., Reed, Jennifer L & Palsson, B.Ø., 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews. Microbiology*, 2(11), pp.886-897.

Ravasz, E et al., 2002. Hierarchical organization of modularity in metabolic networks. *Science (New York, N.Y.)*, 297(5586), pp.1551-1555.

Ravasz, Erzsébet, 2009. Detecting hierarchical modularity in biological networks. *Methods in Molecular Biology (Clifton, N.J.)*, 541, pp.145-160.

Reed, Jennifer L et al., 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biology*, 4(9), p.R54.

Reed, Jennifer L & Palsson, B.Ø., 2004. Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Research*, 14(9), pp.1797-1805.

Reed, Jennifer L et al., 2006. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences of the United States of*

America, 103(46), pp.17480-17484.

Ren, Q., Chen, K. & Paulsen, I.T., 2007. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Research*, 35(Database issue), pp.D274-279.

Robertson, D.G., 2005. Metabonomics in toxicology: a review. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, 85(2), pp.809-822.

Romano, A.H. & Conway, T, 1996. Evolution of carbohydrate metabolic pathways. *Research in Microbiology*, 147(6-7), pp.448-455.

Russo, T A & Johnson, J R, 2000. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of Escherichia coli: ExPEC. *The Journal of Infectious Diseases*, 181(5), pp.1753-1754.

Russo, Thomas A & Johnson, James R, 2003. Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem. *Microbes and Infection / Institut Pasteur*, 5(5), pp.449-456.

Sabarly, V., 2010. *Structuration de la diversité métabolique chez Escherichia coli□:Intégration du réseau métabolique, du protéome, des paramètres enzymatiques et des phénotypes de croissances*.

Saito, K. & Matsuda, F., 2010. Metabolomics for Functional Genomics, Systems Biology, and Biotechnology. *Annual Review of Plant Biology*, 61(1), pp.463-489.

Satish Kumar, V., Dasika, M. & Maranas, C., 2007. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(1), p.212.

Sauer, U et al., 1999. Metabolic flux ratio analysis of genetic and environmental modulations of Escherichia coli central carbon metabolism. *Journal of Bacteriology*, 181(21), pp.6679-6688.

Sauer, Uwe, 2004. High-throughput phenomics: experimental methods for mapping fluxomes. *Current Opinion in Biotechnology*, 15(1), pp.58-63.

Savinell, J.M. & Palsson, B O, 1992a. Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *Journal of Theoretical Biology*, 155(2), pp.201-214.

Savinell, J.M. & Palsson, B O, 1992b. Optimal selection of metabolic fluxes for in vivo measurement. II. Application to Escherichia coli and hybridoma cell metabolism. *Journal of Theoretical Biology*, 155(2), pp.215-242.

Schellenberger, J. et al., 2010. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11, p.213.

Schmidt, S. et al., 2003. Metabolites: a helping hand for pathway evolution? *Trends in Biochemical Sciences*, 28(6), pp.336-341.

Schneider, B L, Kiupakis, A.K. & Reitzer, L.J., 1998. Arginine catabolism and the arginine succinyltransferase pathway in Escherichia coli. *Journal of Bacteriology*, 180(16), pp.4278-4286.

Schomburg, I. et al., 2004. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Research*, 32(Database issue), pp.D431-433.

Schuster, S., Kahn, D. & Westerhoff, Hans V., 1993. Modular analysis of the control of complex metabolic pathways. *Biophysical Chemistry*, 48(1), pp.1-17.

Schuster, S., Pfeiffer, T. & Fell, David A., 2008. Is maximization of molar yield in metabolic networks favoured by evolution? *Journal of Theoretical Biology*, 252(3), pp.497-504.

Sears, H.J. & Brownlee, I., 1952. Further observations on the persistence of individual strains of Escherichia coli in the intestinal tract of man. *Journal of Bacteriology*, 63(1), pp.47-57.

Sears, H.J., Brownlee, I. & Uchiyama, J.K., 1950. Persistence of individual strains of Escherichia coli in the intestinal tract of man. *Journal of Bacteriology*, 59(2), pp.293-301.

Segrè, D., Vitkup, D. & Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23), pp.15112-15117.

Shlomi, T., Berkman, O. & Ruppin, E., 2005. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21), pp.7695-7700.

Skurnik, D. et al., 2006. Effect of human vicinity on antimicrobial resistance and integrons in animal faecal Escherichia coli. *The Journal of Antimicrobial Chemotherapy*, 57(6), pp.1215-1219.

Stelling, Jörg, 2004. Mathematical models in microbial systems biology. *Current Opinion in Microbiology*, 7(5), pp.513-518.

Stephanopoulos, G. et al., 1998. *Metabolic engineering: principles and methodologies*, Academic Press.

von Stockar, U. & Liu, J., 1999. Does microbial life always feed on negative entropy? Thermodynamic analysis of microbial growth. *Biochimica Et Biophysica Acta*, 1412(3), pp.191-211.

Suzuki, N. et al., 2006. High-throughput transposon mutagenesis of Corynebacterium glutamicum and construction of a single-gene disruptant mutant library. *Applied and Environmental Microbiology*, 72(5), pp.3750-3755.

Tenaillon, O. et al., 2010. The population genetics of commensal Escherichia coli. *Nature Reviews. Microbiology*, 8(3), pp.207-217.

Terzer, M. & Stelling, Jörg, 2008. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics (Oxford, England)*, 24(19), pp.2229-2235.

Touchon, M. et al., 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genetics*, 5(1), p.e1000344.

Vallenet, D. et al., 2009. MicroScope: a platform for microbial genome annotation and comparative genomics. *Database: The Journal of Biological Databases and Curation*, 2009, p.bap021.

Vallenet, D. et al., 2006. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Research*, 34(1), pp.53-65.

Varma, A, Boesch, B.W. & Palsson, B O, 1993a. Biochemical production capabilities of Escherichia coli. *Biotechnology and Bioengineering*, 42(1), pp.59-73.

Varma, A, Boesch, B.W. & Palsson, B O, 1993b. Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. *Applied and Environmental Microbiology*, 59(8), pp.2465-2473.

Varma, Amit & Palsson, Bernhard O., 1994. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nat Biotech*, 12(10), pp.994-998.

Di Ventura, B. et al., 2006. From in vivo to in silico biology and back. *Nature*, 443(7111), pp.527-533.

VerBerkmoes, N.C. et al., 2009. Systems Biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Micro*, 7(3), pp.196-205.

Verein, N.-M. & Heidelberg, 1877. *Verhandlungen*,

Verwoerd, W.S., 2011. A new computational method to split large biochemical networks into coherent subnets. *BMC Systems Biology*, 5, p.25.

Villas-Bas, S.G. et al., 2007. *Metabolome Analysis*, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Vollaard, E.J. & Clasener, H.A., 1994. Colonization resistance. *Antimicrobial Agents and Chemotherapy*, 38(3), pp.409-414.

Walk, S.T. et al., 2009. Cryptic Lineages of the Genus Escherichia. *Applied and Environmental Microbiology*, 75(20), pp.6534-6544.

Wiback, S.J. et al., 2004. Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *Journal of Theoretical Biology*, 228(4), pp.437-447.

Wong, J., 2003. Dr. Alexander Fleming and the discovery of penicillin. *Primary Care Update for OB/GYNS*, 10(3), pp.124-126.

Wong, P., Gladney, S. & Keasling, J.D., 1997. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnology Progress*, 13(2), pp.132-143.

Yamamoto, N. et al., 2009. Update on the Keio collection of Escherichia coli single-gene deletion mutants. *Molecular Systems Biology*, 5, pp.335-335.

Zur, H., Ruppin, E. & Shlomi, T., 2010. iMAT: an integrative metabolic analysis tool. *Bioinformatics (Oxford, England)*, 26(24), pp.3140-3142.

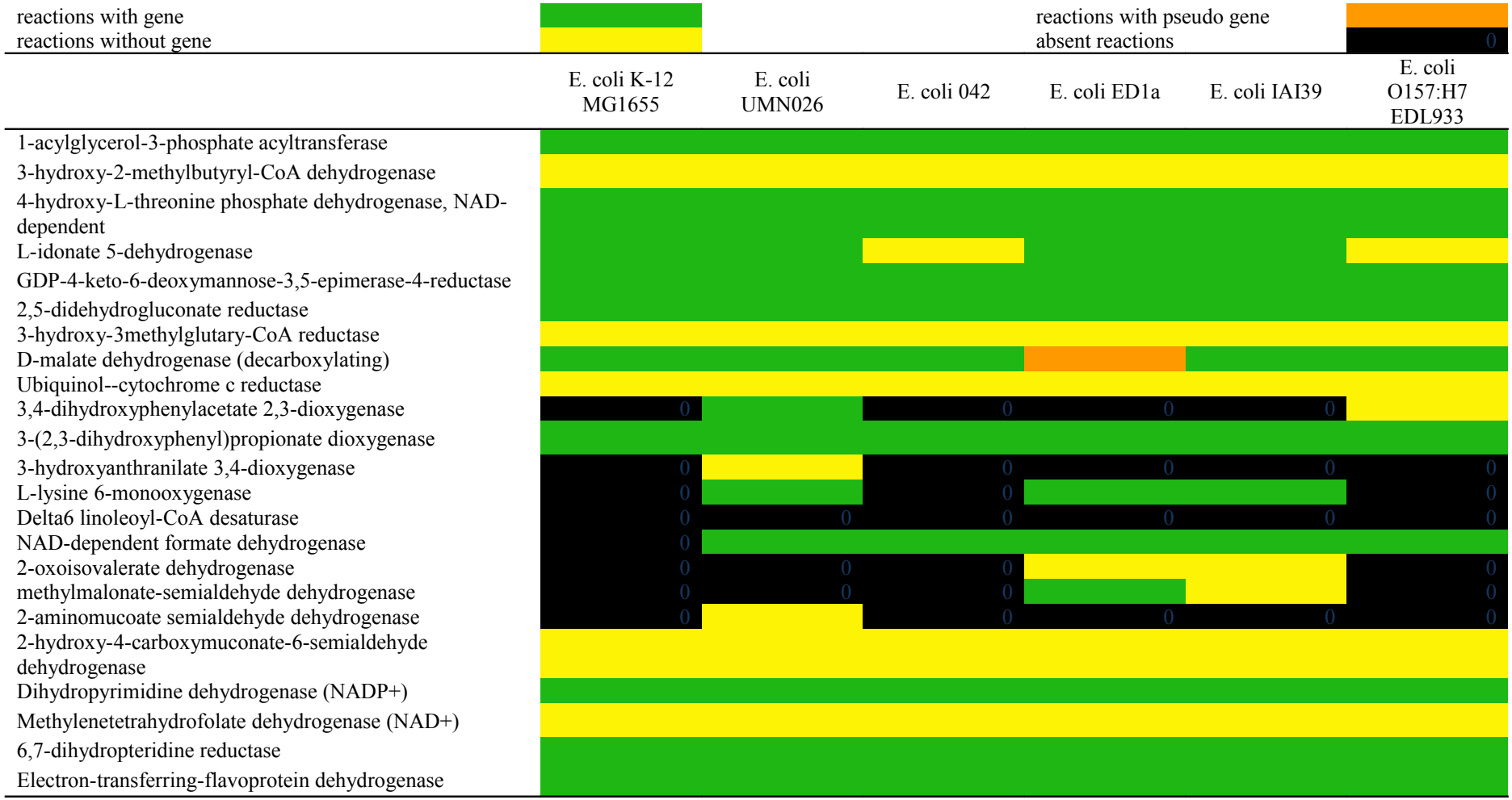
.

Annexes

Les trois premières annexes sont issues des données supplémentaires de l'article
En raison du nombre important de données, seul un extrait des annexes 1 et 2 sont fournis.

Annexe 1 Extrait du tableau des compositions en réactions

La première colonne donne l'identifiant de la réaction dans la base de données EcoCyc, la seconde colonne contient es noms des réactions et les 23 autres les différents réseaux. En vert les réactions associées à un gène, en jaune les réactions sans gène, en orange les réactions associées à des pseudogenes et en noir els réactions absentes du réseau.


reactions with gene
reactions without gene
reactions with pseudo gene
absent reactions
0
E. coli K-12 MG1655
E. coli UMN026
E. coli 042
E. coli ED1a
E. coli IAI39
E. coli O157:H7 EDL933
1-acylglycerol-3-phosphate acyltransferase
3-hydroxy-2-methylbutyryl-CoA dehydrogenase
4-hydroxy-L-threonine phosphate dehydrogenase, NAD-dependent
L-idonate 5-dehydrogenase
GDP-4-keto-6-deoxymannose-3,5-epimerase-4-reductase
2,5-didehydrogluconate reductase
3-hydroxy-3methylglutary-CoA reductase
D-malate dehydrogenase (decarboxylating)
Ubiquinol--cytochrome c reductase
3,4-dihydroxyphenylacetate 2,3-dioxygenase
3-(2,3-dihydroxyphenyl)propionate dioxygenase
3-hydroxyanthranilate 3,4-dioxygenase
L-lysine 6-monooxygenase
Delta6 linoleoyl-CoA desaturase
NAD-dependent formate dehydrogenase
2-oxoisovalerate dehydrogenase
methylmalonate-semialdehyde dehydrogenase
2-aminomucoate semialdehyde dehydrogenase
2-hydroxy-4-carboxymuconate-6-semialdehyde dehydrogenase
Dihydropyrimidine dehydrogenase (NADP+)
Methylenetetrahydrofolate dehydrogenase (NAD+)
6,7-dihydropteridine reductase
Electron-transferring-flavoprotein dehydrogenase

B

Annexe 2 : Extrait de la table bilan du processus de reconstruction des complexes.
En vert les complexes dont tous les gènes sont présents dans l'organisme cible. En orange au moins un gène est absent dans la cible. En rouge il n'existe aucun gène du complexe dans la cible. Le ou les gènes sans homologue d'EcoCyc sont indiqués entre crochets.

Les gènes sont identifiés par le *ECKnumber*, puisqu'il s'agit d'une comparaison directe avec EcoCyc.
Cet extrait est issu d'un tableur trop important pour être intégré au manuscrit, il est disponible à cette adresse :

Ce tableau est composé de trois feuilles :

Homomérique

hétéromérique

comparaisons des hétéromères

Homomérique

Information à propos des complexes homomériques. La colonne ID fait référence à l'identifiant EcoCyc, la colonne Name au nom du complexe, la colonne sous unité contient le monomère ou la sous unité qui compose le complexe, la colonne gène contient l'identifiant du gène associé.
Nous considérons comme homomère la répétition d'un même complexe hétéromériques, c'est pourquoi plusieurs gènes peuvent être associés à un homomère.

Hétéromérique

Information à propos des complexes hétéromériques. La colonne ID fait référence à l'identifiant EcoCyc, la colonne Name au nom du complexe, la colonne sous unité contient les monomère ou les sous unités qui composent le complexe, la colonne gène contient l'identifiant du gène associé.

Comparaison des hétéromères
Comparaisons des hétéromères reconstruits par rapport à ceux d'EcoCyc. En vert les complexes dont tous les gènes sont présents dans l'organisme cible. En orange au moins un gène est absent dans la cible. En rouge il n'existe aucun gène du complexe dans la cible. Le ou les gènes sans homologue d'EcoCyc sont indiqués entre crochets.
Les gènes sont identifiés par le ECKnumber, puisqu'il s'agit d'une comparaison directe avec EcoCyc.

Complete	Complexes for which all the subunit have been reconstructed
AllMissing	Complexes for which none of the subunit have been reconstructed
AtLeastOneMissing	Complexes for which at least one subunit is missing; the K12-MG1655 genes without homologous are given inside the []
orphan	Complexes for which a subunit could not be reconstructed due to missing genes in EcoCyc

	K-12 MG1655	K-12 W3110	042	536	55989	APEC O1	ATCC8739	CFT073
ABC-22-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-23-CPLX	Complete	AtLeastOneMissing [ECK4097]	AtLeastOneMissing [ECK4097]	AtLeastOneMissing [ECK4097]	AtLeastOneMissing [ECK4097]	AtLeastOneMissing [ECK4097]	AtLeastOneMissing [ECK4097]	AtLeastOneMissing [ECK4097]
ABC-24-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-27-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-28-CPLX	Complete	Complete	Complete	AtLeastOneMissing [ECK3744]	Complete	Complete	Complete	Complete
ABC-29-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-3-CPLX	Complete	Complete	Complete	AtLeastOneMissing [ECK2302]	Complete	Complete	Complete	AtLeastOneMissing [ECK2304]
ABC-304-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-32-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-35-CPLX	Complete	Complete	Complete	AtLeastOneMissing [ECK2191]	Complete	Complete	Complete	Complete
ABC-4-CPLX	Complete	Complete	Complete	AtLeastOneMissing [ECK0851]	Complete	Complete	Complete	Complete
ABC-40-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-41-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-42-CPLX	Complete	Complete	AllMissing [ECK4079, ECK4080, ECK4081]	AtLeastOneMissing [ECK4079]	AllMissing [ECK4079, ECK4080, ECK4081]	AtLeastOneMissing [ECK4079]	Complete	Complete
ABC-45-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-46-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-48-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-49-CPLX	Complete	Complete	Complete	AtLeastOneMissing [ECK0820]	Complete	AtLeastOneMissing [ECK0820]	Complete	Complete
ABC-5-CPLX	Complete	Complete	Complete	AtLeastOneMissing [ECK0157]	Complete	Complete	Complete	Complete
ABC-51-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	AtLeastOneMissing [ECK1437]
ABC-55-CPLX	Complete	Complete	AllMissing [ECK1305, ECK1306, ECK1307]	Complete	Complete	Complete	Complete	Complete
ABC-57-CPLX	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete
ABC-58-CPLX	Complete	Complete	Complete	AllMissing [ECK1506, ECK1507, ECK1508, ECK1509]	Complete	AllMissing [ECK1506, ECK1507, ECK1508, ECK1509]	Complete	AllMissing [ECK1506, ECK1507, ECK1508, ECK1509]

Annexe 3 : Composition des arbres de régressions.

La première feuille concerne l'arbre de régression des souches commensales, la deuxième l'arbre des souches pathogènes intestinales et la troisième l'arbre des souches pathogènes extra intestinales. La structure des résultats est identique sur les 3 feuilles : La première colonne contient le nom des réactions, la seconde l'identifiant EcoCyc de la réaction, la troisième le nombre de souches du type de pathogénicité étudié qui possèdent la réaction, la quatrième colonne donne le nom des souches qui possèdent la réaction et la dernière colonne donne le nombre des souches qui ne sont pas de la pathogénicité étudiée et qui possèdent la réaction.

Identifiant MicroCyc	Présent Commensales	8 Commensales	Présent autres	15 InPEc et ExPEc
NADPH-DEHYDROGENASE-RXN	1	SMS-3-5	12	536, APEC O1, IAI39, S88, UMN026, UTI89, 042, LF82, O127H6 E2348/69, O157:H7 EC4115, O157H7 EDL933, O157H7 sakai
PHENDEHYD-RXN, RXN0-2043, AMACETOXID-RXN, PHENYLACETATE--COA-LIGASE-RXN, RXN0-2044, AMINEOXID-RXN, RXN-3641, RXN0-5065, AMINEPHEN-RXN, RXN0-2042	6	ATCC 8739, HS, IAI1, K-12 MG1655, K-12 W3110, SE11	2	55989, E24377A
3.6.3.28-RXN	2	ED1a, SMS-3-5	13	APEC O1, CFT073, IAI39, S88, UMN026, UTI89, 042, 55989, LF82, O127H6 E2348/69, O157H7 EDL933, O157H7 sakai
R262-RXN, HEME-OXYGENASE-DECYCLIZING-RXN	2	ED1a, SMS-3-5	13	536, APEC O1, CFT073, IAI39, S88, UMN026, UTI89, 042, LF82, O127H6 E2348/69, O157:H7 EC4115, O157H7 EDL933, O157H7 sakai
3.1.21.3-RXN	3	ED1a, K-12 MG1655, K-12 W3110	14	536, APEC O1, CFT073, IAI39, S88, UMN026, UTI89, 55989, E24377A, LF82, O127H6 E2348/69, O157:H7 EC4115, O157H7 EDL933, O157H7 sakai

Identifiant MicroCyc	Présent InPEc	8 InPEc	Présent autres	15 commensales et ExPECs
R344-RXN	7	042, E24377A, LF82, O127H6 E2348/69 O157:H7 EC4115, O157H7 EDL933, O157H7 sakai	3	SMS-3-5, IAI39, UMN026
3.6.3.15-RXN	4	O127H6 E2348/69 O157:H7 EC4115, O157H7 EDL933, O157H7 sakai	0	

GALACTONDEHYDRAT-RXN	4	042, 55989, E24377A, LF82	15	ATCC 8739, ED1a, HS, IAI1, K12 MG1655, K12 W3110, SE11, SMS-3-5, 536, APEC O1, CFT073, IAI39, S88, UMN026, UTI89
MALEYLACETOACETATE-ISOMERASE-RXN	4	E24377A, O157:H7 EC4115, O157H7 EDL933, O157H7 sakai	0	
RXN1G-132, RXN-9648, RXN1G-26, RXN-9653, 2.3.1.41-RXN, RXN1G-840, RXN1G-138, RXN-9650, RXN1G-306, RXN-9654, RXN-8391, RXN1G-218, RXN-9651, RXN1G-324, RXN1G-1003, RXN-9632, RXN1G-236, RXN-9652, RXN1G-460	6	042, LF82, O127H6 E2348/69, O157:H7 EC4115, O157H7 EDL933, O157H7 sakai	2	CFT073, UMN026

Identifiant MicroCyc	**Présent ExPEc**	**7 ExPEc**	**Présent autres**	**16 commensales et InPECs**
RXN-10814, 3-SULFINOALANINE-AMINOTRANSFERASE-RXN	4	536, CFT073, IAI39, UTI89	0	
3.2.1.129-RXN	4	APEC O1, IAI39, S88, UTI89	0	
RXN0-4841, RXN0-963	0		12	ATCC 8739, HS, IAI1, K12 MG1655, K12 W3110, SE11, 042, 55989, E24377A, O157:H7 EC4115, O157H7 EDL933, O157H7 sakai
2.7.10.1-RXN	0		12	ATCC 8739, HS, IAI1, K12 MG1655, K12 W3110, SE11, 55989, E24377A, O127H6 E2348/69, O157:H7 EC4115, O157H7 EDL933, O157H7 sakai
3.6.3.38-RXN	7	536, APEC O1, CFT073, IAI39, S88, UMN026, UTI89	4	ED1a, SMS-3-5, 042, LF82
XMPXAN-RXN, RXN-5841	7	536, APEC O1, CFT073, IAI39, S88, UMN026, UTI89	5	ED1a, SMS-3-5, 042, LF82, O127H6 E2348/69

Annexe 4 : Table des 121 organismes utilisés pour la reconstruction des réseaux métaboliques.
Les souches sont ordonnées en fonction de leur phylogroupe et. L'ensemble est composé de 104 *E. coli* : 25 du phylogroupe A, 17 du phylogroupe B1, 30du phylogroupe B2, 16 du phylogroupe D, 7 du phylogroupe E et 2 du phylogroupe F. Deux sont de phylogroupe inconnu et 8 sont des représentant de clades (souches colorée en violet). Les souches sont colorées suivant leur pathogénicité : bleu commensales, rouge pathogène hors de l'intestin et vert pathogène dans l'intestin. Les 7 souches non *E. coli* comportent 6 *Shigella*, 3 *E. albertii*, 2 *E fergusonii* et 3 *Salmonella*.

Nom	Phylogroupe	Pathogenicite	Clade
Escherichia coli ATCC8739	A	Commensal	coli
Escherichia coli HS	A	Commensal	coli
Escherichia coli K-12 MG1655	A	Commensal	coli
Escherichia coli K-12 W3110	A	Commensal	coli
Escherichia coli b175	A	inconnu	coli
Escherichia coli b921	A	inconnu	coli
Escherichia coli e1002	A	inconnu	coli
Escherichia coli e1114	A	inconnu	coli
Escherichia coli e1520	A	inconnu	coli
Escherichia coli e482	A	inconnu	coli
Escherichia coli e560	A	inconnu	coli
Escherichia coli e704	A	inconnu	coli
Escherichia coli h185	A	inconnu	coli
Escherichia coli h288	A	inconnu	coli
Escherichia coli h383	A	inconnu	coli
Escherichia coli h386	A	inconnu	coli
Escherichia coli h454	A	inconnu	coli
Escherichia coli h489	A	inconnu	coli
Escherichia coli h593	A	inconnu	coli
Escherichia coli h617	A	inconnu	coli
Escherichia coli h736	A	inconnu	coli
Escherichia coli r424	A	inconnu	coli
Escherichia coli ta007	A	inconnu	coli
Escherichia coli ta008	A	inconnu	coli
Escherichia coli ta144	A	inconnu	coli
Escherichia coli 55989	B1	InPEc	coli
Escherichia coli E24377A	B1	InPEc	coli
Escherichia coli IAI1	B1	Commensal	coli
Escherichia coli SE11	B1	Commensal	coli
Escherichia coli b088	B1	inconnu	coli
Escherichia coli b574	B1	inconnu	coli
Escherichia coli e1167	B1	inconnu	coli
Escherichia coli h120	B1	inconnu	coli
Escherichia coli h218	B1	inconnu	coli
Escherichia coli h220	B1	inconnu	coli
Escherichia coli h420	B1	inconnu	coli
Escherichia coli h591	B1	inconnu	coli
Escherichia coli m863	CI	inconnu	I

Nom	Phylogroupe	Pathogenicite	Clade
Escherichia coli r529	B1	inconnu	coli
Escherichia coli t408	B1	inconnu	coli
Escherichia coli t426	B1	inconnu	coli
Escherichia coli ta141	B1	inconnu	coli
Escherichia coli ta271	B1	inconnu	coli
Escherichia coli 536	B2	ExPEc	coli
Escherichia coli APEC O1	B2	ExPEc	coli
Escherichia coli CFT073	B2	ExPEc	coli
Escherichia coli ED1a	B2	Commensal	coli
Escherichia coli LF82	B2	InPEc	coli
Escherichia coli O127:H6 E2348/69	B2	InPEc	coli
Escherichia coli S88	B2	ExPEc	coli
Escherichia coli UTI89	B2	ExPEc	coli
Escherichia coli b108	B2	inconnu	coli
Escherichia coli b671	B2	inconnu	coli
Escherichia coli h001	B2	inconnu	coli
Escherichia coli h223	B2	inconnu	coli
Escherichia coli h252	B2	inconnu	coli
Escherichia coli h263	B2	inconnu	coli
Escherichia coli h296	B2	inconnu	coli
Escherichia coli h305	B2	inconnu	coli
Escherichia coli h378	B2	inconnu	coli
Escherichia coli h413	B2	inconnu	coli
Escherichia coli h461	B2	inconnu	coli
Escherichia coli h504	B2	inconnu	coli
Escherichia coli h588	B2	inconnu	coli
Escherichia coli h660	B2	inconnu	coli
Escherichia coli m605	B2	inconnu	coli
Escherichia coli r527	B2	inconnu	coli
Escherichia coli ta014	B2	inconnu	coli
Escherichia coli ta103	B2	inconnu	coli
Escherichia coli ta206	B2	inconnu	coli
Escherichia coli ta435	B2	inconnu	coli
Escherichia coli ta464	B2	inconnu	coli
Escherichia coli h299	B2/F	inconnu	coli
Escherichia coli e1492	CI	inconnu	I
Escherichia coli h442	CI	inconnu	I
Escherichia fergusonii ATCC 35469T	Escherichia fergusonii	inconnu	fergusonii

Nom	Phylogroupe	Pathogenicite	Clade
Escherichia coli tw10509	CI	inconnu	I
Escherichia coli ta004	CIII	inconnu	III
Escherichia coli h605	CIV	inconnu	IV
Escherichia coli e1118	CV	inconnu	V
Escherichia sp. b646	CV	inconnu	V
Escherichia coli 042	D	InPEc	coli
Escherichia coli UMN026	D	ExPEc	coli
Escherichia coli b354	D	inconnu	coli
Escherichia coli b367	D	inconnu	coli
Escherichia coli b706	D	inconnu	coli
Escherichia coli fvec_1302	D	inconnu	coli
Escherichia coli fvec_1412	D	inconnu	coli
Escherichia coli fvec_1465	D	inconnu	coli
Escherichia coli m056	D	inconnu	coli
Escherichia coli m114	D	inconnu	coli
Escherichia coli m646	D	inconnu	coli
Escherichia coli ta024	D	inconnu	coli
Escherichia coli ta143	D	inconnu	coli
Escherichia coli ta249	D	inconnu	coli
Escherichia coli ta255	D	inconnu	coli
Escherichia coli ta280	D	inconnu	coli
Escherichia coli O157:H7 EC4115	E	InPEc	coli
Escherichia coli O157:H7 EDL933	E	InPEc	coli
Escherichia coli O157:H7 sakai	E	InPEc	coli
Escherichia coli b185	E	inconnu	coli
Escherichia coli m718	E	inconnu	coli
Escherichia coli ta054	E	inconnu	coli
Escherichia coli ta447	E	inconnu	coli
Escherichia coli IAI39	F	ExPEc	coli
Escherichia coli SMS-3-5	F	Commensal	coli
Escherichia coli puti459	inconnu	inconnu	inconnu
Escherichia coli e267	inconnu	inconnu	coli
Escherichia albertii b090	Escherichia albertii	inconnu	albertii
Escherichia albertii b156	Escherichia albertii	inconnu	albertii
Escherichia albertii TW07627	Escherichia albertii	inconnu	albertii

Nom	Phylogroupe	Pathogenicite	Clade
Escherichia fergusonii b253	Escherichia fergusonii	inconnu	fergusonii
Shigella boydii	S	Shiga-toxine	coli
Shigella dysenteriae	S	Shiga-toxine	coli
Shigella flexner 5	S	Shiga-toxine	coli
Shigella flexneri 2a	S	Shiga-toxine	coli
Shigella flexneri 2a	S	Shiga-toxine	coli
Shigella sonnei	S	Shiga-toxine	coli
Salmonella enterica subsp. enterica serovar Choleraesuis SC-B67	Salmonella	Enteritidis	Salmonella
Salmonella enterica subsp. enterica serovar Typhi	Salmonella	Typhimurium	Salmonella
Salmonella typhimurium LT2	Salmonella	Typhimurium	Salmonella

Annexe 5 : Intégration des modèles au sein de la plate-forme NemoStudio.

Annexe 5a. Page d'accueil de la plate-forme avec dans le volet droit le menu de sélection des modèles, des différentes options et simulations.

Annexe 5b. Illustration de la page d'information

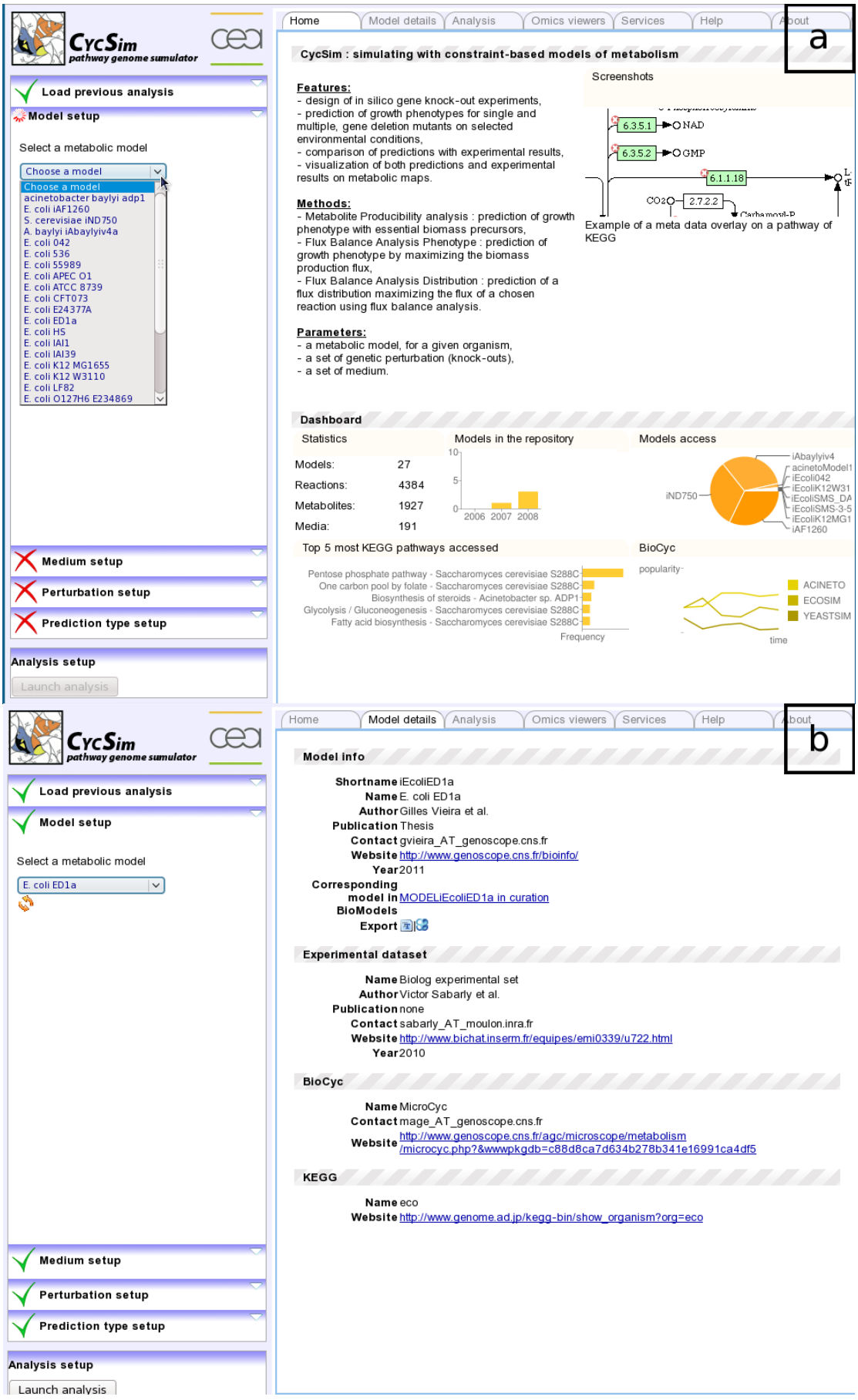
Les informations relatives au modèle : nom du modèle, auteurs, publication, l'année, le status du modèle et l'exportation au format SBML

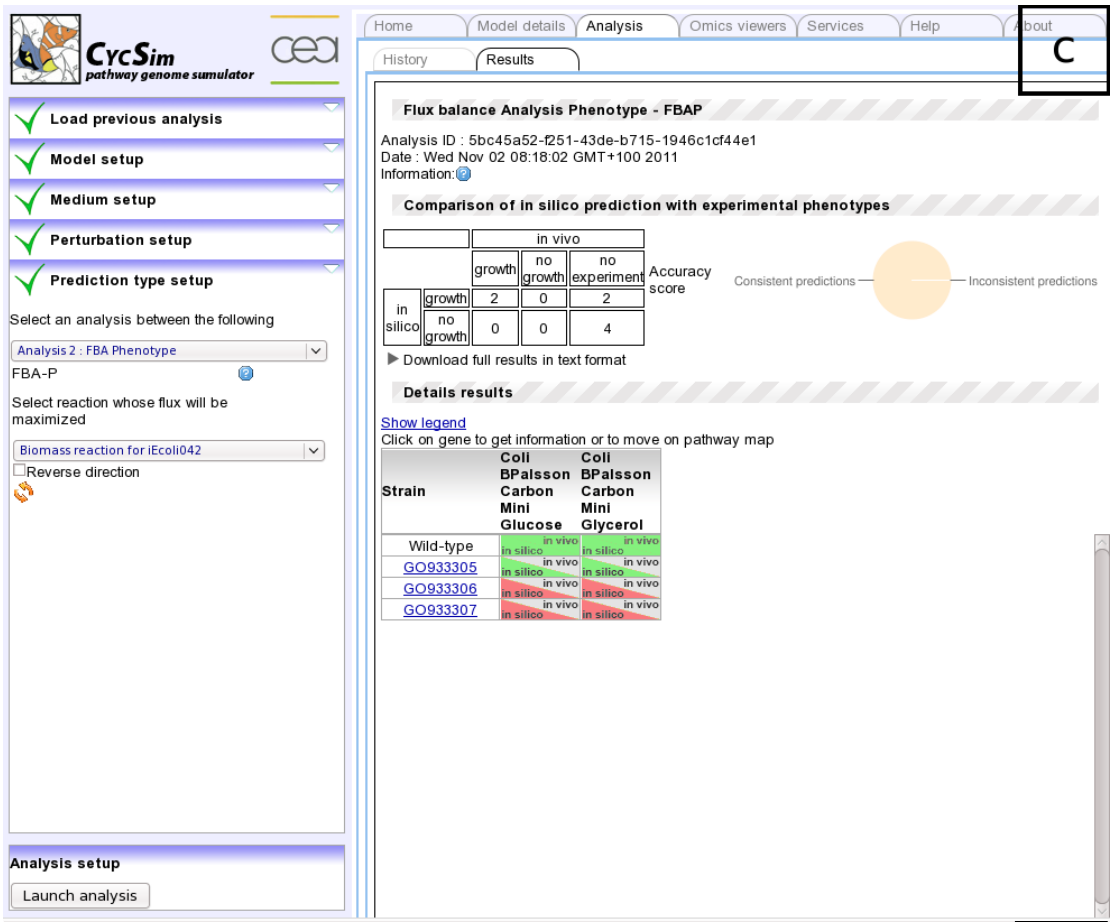
Les informations expérimentales disponibles sur le modèle : nom des données, auteurs, publication, année

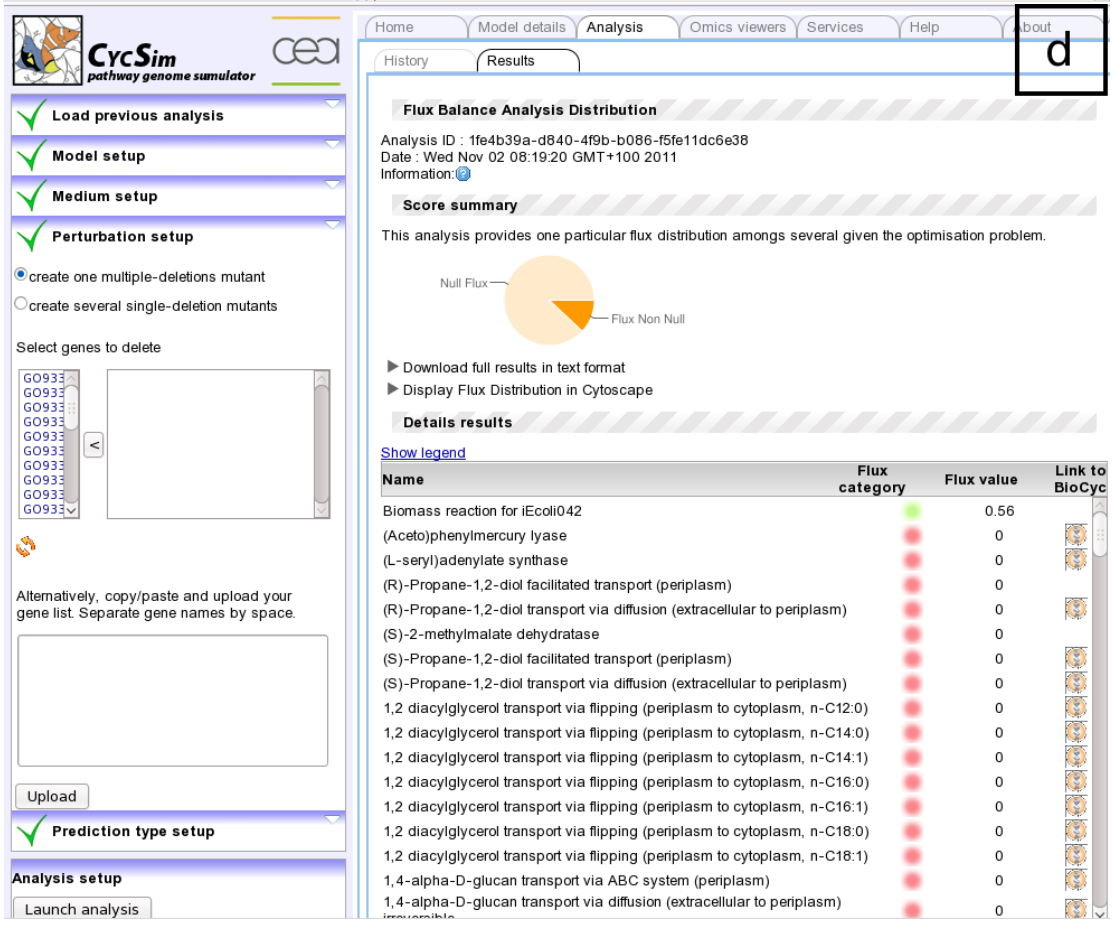
Les bases de données pour lesquelles il existe des références-croisées : nom et URL

Annexe 5c. Ilustration des résultat des comparaisons entre prédictions *in silico* et obersvation *in vitro*.

Annexe 5d. Illustration du vecteur flux résultat de la FBA sur milieu minimum glucose


CycSim
pathway genome simulator
Load previous analysis
Model setup
Select a metabolic model
Choose a model
acinetobacter baylyi adp1
E. coli iAF1260
S. cerevisiae iND750
A. baylyi iAbaylyiv4a
E. coli 042
E. coli 536
E. coli 55989
E. coli APEC O1
E. coli ATCC 8739
E. coli CFT073
E. coli E24377A
E. coli ED1a
E. coli HS
E. coli IAI1
E. coli IAI39
E. coli K12 MG1655
E. coli K12 W3110
E. coli LF82
E. coli O127H6 E234869
Medium setup
Perturbation setup
Prediction type setup
Analysis setup
Launch analysis
Home
Model details
Analysis
Omics viewers
Services
Help
About
a
CycSim : simulating with constraint-based models of metabolism
Features:
- design of in silico gene knock-out experiments,
- prediction of growth phenotypes for single and multiple, gene deletion mutants on selected environmental conditions,
- comparison of predictions with experimental results,
- visualization of both predictions and experimental results on metabolic maps.
Methods:
- Metabolite Producibility analysis : prediction of growth phenotype with essential biomass precursors,
- Flux Balance Analysis Phenotype : prediction of growth phenotype by maximizing the biomass production flux,
- Flux Balance Analysis Distribution : prediction of a flux distribution maximizing the flux of a chosen reaction using flux balance analysis.
Parameters:
- a metabolic model, for a given organism,
- a set of genetic perturbation (knock-outs),
- a set of medium.
Screenshots
6.3.5.1
NAD
6.3.5.2
GMP
6.1.1.18
CO2
2.7.2.2
Example of a meta data overlay on a pathway of KEGG
Dashboard
Statistics
Models: 27
Reactions: 4384
Metabolites: 1927
Media: 191
Models in the repository
Models access
iAbaylyiv4
acinetoModel1
iEcoli042
iEcoliK12W31
iEcoliSMS_DA
iEcoliSMS-3-5
iEcoliK12MG1
iAF1260
iND750
Top 5 most KEGG pathways accessed
Pentose phosphate pathway - Saccharomyces cerevisiae S288C
One carbon pool by folate - Saccharomyces cerevisiae S288C
Biosynthesis of steroids - Acinetobacter sp. ADP1
Glycolysis / Gluconeogenesis - Saccharomyces cerevisiae S288C
Fatty acid biosynthesis - Saccharomyces cerevisiae S288C
Frequency
BioCyc
popularity
ACINETO
ECOSIM
YEASTSIM
time
b
Model info
Shortname iEcoliED1a
Name E. coli ED1a
Author Gilles Vieira et al.
Publication Thesis
Contact gvieira_AT_genoscope.cns.fr
Website http://www.genoscope.cns.fr/bioinfo/
Year 2011
Corresponding model in BioModels MODELiEcoliED1a in curation
Export
Experimental dataset
Name Biolog experimental set
Author Victor Sabarly et al.
Publication none
Contact sabarly_AT_moulon.inra.fr
Website http://www.bichat.inserm.fr/equipes/emi0339/u722.html
Year 2010
BioCyc
Name MicroCyc
Contact mage_AT_genoscope.cns.fr
Website http://www.genoscope.cns.fr/agc/microscope/metabolism/microcyc.php?&wwwpkgdb=c88d8ca7d634b278b341e16991ca4df5
KEGG
Name eco
Website http://www.genome.ad.jp/kegg-bin/show_organism?org=eco


CycSim
pathway genome sumulator
Load previous analysis
Model setup
Medium setup
Perturbation setup
Prediction type setup
Select an analysis between the following
Analysis 2 : FBA Phenotype
FBA-P
Select reaction whose flux will be maximized
Biomass reaction for iEcoli042
Reverse direction
Analysis setup
Launch analysis
Home
Model details
Analysis
Omics viewers
Services
Help
About
History
Results
c
Flux balance Analysis Phenotype - FBAP
Analysis ID : 5bc45a52-f251-43de-b715-1946c1cf44e1
Date : Wed Nov 02 08:18:02 GMT+100 2011
Information:
Comparison of in silico prediction with experimental phenotypes
in vivo
growth
no growth
no experiment
in silico
growth
no growth
2
0
2
0
0
4
Accuracy score
Consistent predictions
Inconsistent predictions
Download full results in text format
Details results
Show legend
Click on gene to get information or to move on pathway map
Strain
Coli BPalsson Carbon Mini Glucose
Coli BPalsson Carbon Mini Glycerol
Wild-type
GO933305
GO933306
GO933307
in vivo
in silico


CycSim
pathway genome sumulator
Load previous analysis
Model setup
Medium setup
Perturbation setup
create one multiple-deletions mutant
create several single-deletion mutants
Select genes to delete
Alternatively, copy/paste and upload your gene list. Separate gene names by space.
Upload
Prediction type setup
Analysis setup
Launch analysis
Home
Model details
Analysis
Omics viewers
Services
Help
About
History
Results
d
Flux Balance Analysis Distribution
Analysis ID : 1fe4b39a-d840-4f9b-b086-f5fe11dc6e38
Date : Wed Nov 02 08:19:20 GMT+100 2011
Information:
Score summary
This analysis provides one particular flux distribution amongs several given the optimisation problem.
Null Flux
Flux Non Null
Download full results in text format
Display Flux Distribution in Cytoscape
Details results
Show legend
Name
Flux category
Flux value
Link to BioCyc
Biomass reaction for iEcoli042 0.56
(Aceto)phenylmercury lyase 0
(L-seryl)adenylate synthase 0
(R)-Propane-1,2-diol facilitated transport (periplasm) 0
(R)-Propane-1,2-diol transport via diffusion (extracellular to periplasm) 0
(S)-2-methylmalate dehydratase 0
(S)-Propane-1,2-diol facilitated transport (periplasm) 0
(S)-Propane-1,2-diol transport via diffusion (extracellular to periplasm) 0
1,2 diacylglycerol transport via flipping (periplasm to cytoplasm, n-C12:0) 0
1,2 diacylglycerol transport via flipping (periplasm to cytoplasm, n-C14:0) 0
1,2 diacylglycerol transport via flipping (periplasm to cytoplasm, n-C14:1) 0
1,2 diacylglycerol transport via flipping (periplasm to cytoplasm, n-C16:0) 0
1,2 diacylglycerol transport via flipping (periplasm to cytoplasm, n-C16:1) 0
1,2 diacylglycerol transport via flipping (periplasm to cytoplasm, n-C18:0) 0
1,2 diacylglycerol transport via flipping (periplasm to cytoplasm, n-C18:1) 0
1,4-alpha-D-glucan transport via ABC system (periplasm) 0
1,4-alpha-D-glucan transport via diffusion (extracellular to periplasm) 0

Annexe 6 : Comparaisons des observations Biologs en mesure sur point final et des prédictions du flux de biomasses par FBA.

13 souches sont comparées sur 94 milieux. Une souche est capable d'utiliser une source de carbone si la moyenne des mesures des points finaux est supérieure à 0.339. Un modèle est capable d'utiliser une source de carbone si le flux de biomasses est supérieur à zéro. 74% des prédictions sont cohérentes avec l'observation. La mesure au point final est de qualité moindre que la mesure sur la courbe de croissance (Annexe 7).

Prédictions

	2obut	3-HYDROXY-L-PROLINE	4-HYDROXY-BUTYRATE	4-HYDROXYPHENYL ACETATE	4abut	5-OXOPROLINE	ac	acgal	acgam
042	-2	-1	-1	-1	0.636152	-1	0.24956	0	1.23297
536	-2	-1	-1	-1	0.636152	-1	0.24956	0	1.23297
55989	-2	-1	-1	-1	0.636152	-1	0.24956	0	1.23297
APEC O1	-2	-1	-1	-1	0.636152	-1	0.24956	0	1.23297
CFT073	-2	-1	-1	-1	0.636152	-1	0.24956	0	1.23297
ED1a	-2	-1	-1	-1	0.636152	-2	0.24956	0	1.23297
HS	-2	-1	-1	-2	0.636152	-1	0.24956	0	1.23297
IAI1	-2	-1	-1	-2	0.636152	-1	0.24956	0	1.23297
O127H6 E2348/69	-2	-1	-1	-1	0.63615	-1	0.24956	0	1.23297
O157H7 EDL933	-2	-1	-1	-2	0.636152	-1	0.24956	0	1.23297
O157H7 sakai	-2	-1	-1	-2	0.636152	-1	0.24956	0	1.23297
S88	-2	-1	-1	-1	0.636152	-2	0.24956	0	1.23297
UMN026	-2	-1	-1	-2	0.636152	-1	0.24956	0	1.23297

Transporteur absent — Metabolite absent du modèle

Flux de biomasse non nul — Flux de biomasse nul

Observations

	2obut	3-HYDROXY-L-PROLINE	4-HYDROXY-BUTYRATE	4-HYDROXYPHENYL ACETATE	4abut	5-OXOPROLINE	ac	acgal	acgam
042	0	0	0	0	0	0	0	1	1
536	0	0	0	1	0	0	0	1	1
55989	0	0	0	0	0	0	0	1	1
APEC O1	0	0	0	0	0	0	0	1	1
CFT073	0	0	0	0	0	0	0	1	1
ED1a	0	0	0	0	0	0	0	1	1
HS	0	0	0	1	0	0	0	1	1
IAI1	0	0	0	0	0	0	0	0	0
O127H6 E2348/69	0	0	0	0	0	0	0	1	1
O157H7 EDL933	0	0	0	0	0	0	0	1	1
O157H7 sakai	0	0	0	0	0	0	0	1	1
S88	0	0	0	0	0	0	0	1	1
UMN026	0	0	0	0	0	0	0	1	1

Flux de biomasse non nul — Flux de biomasse nul

Comparaisons

	2obut	3-HYDROXY-L-PROLINE	4-HYDROXY-BUTYRATE	4-HYDROXYPHENYL ACETATE	4abut	5-OXOPROLINE	ac	acgal	acgam
042	1	1	1	1	0	1	0	0	1
536	1	1	1	0	0	1	0	0	1
55989	1	1	1	1	0	1	0	0	1
APEC O1	1	1	1	1	0	1	0	0	1
CFT073	1	1	1	1	0	1	0	0	1
ED1a	1	1	1	1	0	1	0	0	1
HS	1	1	1	0	0	1	0	0	1
IAI1	1	1	1	1	0	1	0	1	0
O127H6 E2348/69	1	1	1	1	0	1	0	0	1
O157H7 EDL933	1	1	1	1	0	1	0	0	1
O157H7 sakai	1	1	1	1	0	1	0	0	1
S88	1	1	1	1	0	1	0	0	1
UMN026	1	1	1	1	0	1	0	0	1

Similitude prédiction/observation — Incohérence prédiction/simulation

		Table de vérité		précision	
		Modèle		Modèle	
		Croissance	Pas de croissance	Croissance	Pas de croissance
Biolog	**Croissance**	364	95	79%	21%
	Pas de Croissance	200	563	26%	74%
	rappel	65%	14%		
		35%	86%		

	pousse	pousse pas
F-mesure	71%	19%

Cohérent	Incohérent
905	317
74%	26%

O

acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL	CELLOBIOSE
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.6093	0.434192	-1	0.434192	0.792433	0.459324	0.458862	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1

acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL	CELLOBIOSE
0	0	0	0	1	1	0	1	0	0	0
0	0	1	1	1	1	0	0	0	0	0
0	0	0	0	1	1	0	0	1	0	0
0	0	0	1	1	1	0	0	0	0	0
0	0	0	0	0	1	0	1	0	0	0
0	0	0	0	1	1	1	1	0	0	0
0	0	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	0	0
0	0	0	1	0	1	0	0	0	0	0
0	0	0	1	0	1	0	1	0	0	0
0	0	1	1	1	1	1	1	0	0	0
0	0	1	1	1	1	1	1	0	0	0

acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL	CELLOBIOSE
1	0	0	1	1	1	0	1	1	1	1
1	0	1	0	1	1	0	0	1	1	1
1	0	0	1	1	1	0	0	0	1	1
1	0	0	0	1	1	0	0	1	1	1
1	0	0	1	0	1	0	1	1	1	1
1	0	0	1	1	1	1	1	1	1	1
1	0	1	0	1	1	1	1	0	1	1
1	0	0	1	0	0	0	0	1	1	1
1	0	1	0	1	1	1	1	0	1	1
1	0	0	0	0	1	0	0	1	1	1
1	0	0	0	0	1	0	1	1	1	1
1	0	1	0	1	1	1	1	1	1	1
1	0	1	0	1	1	1	1	1	1	1

acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL	CELLOBIOSE
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.6093	0.434192	-1	0.434192	0.792433	0.459324	0.458862	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1
-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1	-1

acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL	CELLOBIOSE
0	0	0	0	1	1	0	1	0	0	0
0	0	1	1	1	1	0	0	0	0	0
0	0	0	0	1	1	0	0	1	0	0
0	0	0	1	1	1	0	0	0	0	0
0	0	0	0	0	1	0	1	0	0	0
0	0	0	0	1	1	1	1	0	0	0
0	0	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	0	0
0	0	0	1	0	1	0	0	0	0	0
0	0	0	1	0	1	0	1	0	0	0
0	0	1	1	1	1	1	1	0	0	0
0	0	1	1	1	1	1	1	0	0	0

acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL	CELLOBIOSE
1	0	0	1	1	1	0	1	1	1	1
1	0	1	0	1	1	0	0	1	1	1
1	0	0	1	1	1	0	0	0	1	1
1	0	0	0	1	1	0	0	1	1	1
1	0	0	1	0	1	0	1	1	1	1
1	0	0	1	1	1	1	1	1	1	1
1	0	1	0	1	1	1	1	0	1	1
1	0	0	1	0	0	0	0	1	1	1
1	0	1	0	1	1	1	1	0	1	1
1	0	0	0	0	1	0	0	1	1	1
1	0	0	0	0	1	0	1	1	1	1
1	0	1	0	1	1	1	1	1	1	1
1	0	1	0	1	1	1	1	1	1	1

cit	CPD-1099	CPD-1843	CPD-2461	CPD-355	CPD-3561	CPD-3564	CPD-3569	CPD-3570	CPD-3573	CPD-3605
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688294	-1	-1	-1	0.872679	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

cit	CPD-1099	CPD-1843	CPD-2461	CPD-355	CPD-3561	CPD-3564	CPD-3569	CPD-3570	CPD-3573	CPD-3605
0	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	1	0	0	0	1	0
0	1	0	0	0	1	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	0	0	0	0	1	0

cit	CPD-1099	CPD-1843	CPD-2461	CPD-355	CPD-3561	CPD-3564	CPD-3569	CPD-3570	CPD-3573	CPD-3605
0	1	1	1	1	1	1	1	1	0	1
0	0	1	1	1	0	1	1	1	0	1
0	0	1	1	1	0	1	1	1	0	1
0	1	1	1	1	1	1	1	1	0	1
0	1	1	1	1	1	1	1	1	0	1
0	1	1	1	1	0	1	1	1	0	1
0	1	1	1	1	1	1	1	1	0	1
0	1	1	1	1	1	1	1	1	1	1
0	0	1	1	1	1	1	1	1	0	1
0	0	1	1	1	1	1	1	1	0	1
0	0	1	1	1	1	1	1	1	0	1
0	1	1	1	1	1	1	1	1	0	1
0	0	1	1	1	1	1	1	1	0	1

CPD-3618	CPD-3623	CPD-3782	CPD-7692	CPD-8979	CPD0-1656	crn	D-GALACTONO-1-4-LACTONE	Dextrins	ERYTHRITOL	etha
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343653
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343653
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343653
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343653
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0.845599	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343597
-1	-1	-1	-1	-1	-1	0	0.845601	-1	-1	0.343597

CPD-3618	CPD-3623	CPD-3782	CPD-7692	CPD-8979	CPD0-1656	crn	D-GALACTONO-1-4-LACTONE	Dextrins	ERYTHRITOL	etha
0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	1	1	0	0
0	0	0	0	0	1	0	1	1	0	0
0	0	0	0	0	1	0	1	1	0	0
0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	1	0	1	0	0	0
0	0	0	0	0	0	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	1	1	0	0
0	0	0	0	0	1	0	0	1	0	0
0	0	0	0	0	1	0	1	1	0	0
0	0	0	0	0	0	0	1	1	0	0
0	0	0	1	0	1	0	1	1	0	0

CPD-3618	CPD-3623	CPD-3782	CPD-7692	CPD-8979	CPD0-1656	crn	D-GALACTONO-1-4-LACTONE	Dextrins	ERYTHRITOL	etha
1	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	1	1	1	0	1	0
1	1	1	1	1	0	1	1	1	1	0
1	1	1	1	1	1	1	1	0	1	0
1	1	1	1	1	1	1	0	1	1	0
1	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	0	1	0	0	1	0
1	1	1	1	1	1	1	1	0	1	0
1	1	1	0	1	0	1	1	0	1	0

for	fru	fuc-L	g1p	g6p	gal	galur	glcn	glcr	glcur	glu-L
0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.0116819	0.971811	0.943244	0.962956	0.993049	0.952926	0.777248	0.882211	0.679604	0.777248	0.688294
0.042029	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042029	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042029	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295
0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725	0.688295

for	fru	fuc-L	g1p	g6p	gal	galur	glcn	glcr	glcur	glu-L
0	1	1	1	1	1	1	1	1	1	0
0	1	0	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	1	0	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	0	1	0
0	1	1	1	1	1	1	0	1	1	0
0	1	1	1	1	1	1	1	0	1	0
0	1	1	1	1	1	1	1	1	1	0

for	fru	fuc-L	g1p	g6p	gal	galur	glcn	glcr	glcur	glu-L
0	1	1	1	1	1	1	1	1	1	0
0	1	0	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	1	0	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	0	1	0
0	1	1	1	1	1	1	0	1	1	0
0	1	1	1	1	1	1	1	0	1	0
0	1	1	1	1	1	1	1	1	1	0

GLUCOSAMINATE	glyc	glyc3p	glycogen	Glycyl-L-aspartic_acid	his-L	inost	ins	ITACONATE	L-Proline	lac-D
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-1	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-2	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-1	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-2	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-2	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0.903174	1.19303	-2	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-1	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-1	-1	0.426301
-1	0.558274	0.597487	-2	-1	0	0	1.19265	-2	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-1	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-1	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0	1.19303	-2	-1	0.426301
-1	0.558276	0.597489	-2	-1	0	0.903174	1.19303	-1	-1	0.426301

GLUCOSAMINATE	glyc	glyc3p	glycogen	Glycyl-L-aspartic_acid	his-L	inost	ins	ITACONATE	L-Proline	lac-D
0	1	1	0	0	0	0	1	0	0	1
0	1	1	0	0	0	0	1	0	0	1
0	1	0	0	0	0	0	1	0	0	1
0	1	1	0	0	0	0	1	0	0	1
0	1	1	0	0	0	0	1	0	0	1
0	1	1	0	0	0	1	1	0	0	1
0	1	1	0	0	0	0	1	0	0	1
0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	1	0	0	1
0	1	0	0	0	0	0	1	0	0	1
0	1	1	0	0	0	0	1	0	0	1
0	1	0	0	0	0	0	1	0	0	1
0	1	1	0	0	0	0	1	0	0	1

GLUCOSAMINATE	glyc	glyc3p	glycogen	Glycyl-L-aspartic_acid	his-L	inost	ins	ITACONATE	L-Proline	lac-D
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	0	1	1	0
1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	0	1	1	1	1

lcts	leu-L	mal	MALONATE	man	melib	mnl	orn	peacm	phe-L	ppa
1.91589	0	-1	-1	0.962959	0	1.03284	0.746316	-1	0	0.426301
1.91589	0	-1	-1	0.962959	0	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	1.91589	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	0	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	0	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	0	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	0	1.03284	0.746316	-1	0	0.426301
1.91589	0	-1	-1	0.962959	1.91589	1.03284	0.774549	-1	0	0.426301
1.91588	0	-1	-1	0.962956	0	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	1.91589	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	1.91589	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	0	1.03284	0.774549	-1	0	0.426301
1.91589	0	-1	-1	0.962959	1.91589	1.03284	0.746316	-1	0	0.426301

1

lcts	leu-L	mal	MALONATE	man	melib	mnl	orn	peacm	phe-L	ppa
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	0	1	0	0	0	0
1	0	1	0	1	1	1	0	0	0	0

lcts	leu-L	mal	MALONATE	man	melib	mnl	orn	peacm	phe-L	ppa
1	1	0	1	1	0	1	0	1	1	0
1	1	0	1	1	0	1	0	1	1	0
1	1	0	1	1	1	1	0	1	1	0
1	1	0	1	1	0	1	0	1	1	0
1	1	0	1	1	0	1	0	1	1	0
1	1	0	1	1	0	1	0	1	1	0
1	1	0	1	1	0	1	0	1	1	0
0	1	1	1	0	0	0	0	1	1	0
1	1	0	1	1	0	1	0	1	1	0
1	1	0	1	1	1	1	0	1	1	0
1	1	0	1	1	1	1	0	1	1	0
1	1	0	1	1	1	1	0	1	1	0
1	1	0	1	1	1	1	0	1	1	0

PSICOSE	ptrc	QUINATE	RIBITOL	rmn	sbt-D	ser-D	ser-L	succ	sucr	thr-L
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943244	0	0.365916	0.365916	0.490148	1.92591	0.588279
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547
-1	0.760581	-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547

PSICOSE	ptrc	QUINATE	RIBITOL	rmn	sbt-D	ser-D	ser-L	succ	sucr	thr-L
0	0	0	0	1	1	1	0	1	1	0
0	0	0	0	1	1	1	1	1	1	0
0	0	0	0	1	1	0	0	1	0	0
0	0	0	0	1	1	1	1	1	1	0
0	0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	1	1	1	0	1	0
0	0	0	0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	1	1	1	0
0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	1	1	1	0	1	1	0
0	0	0	0	0	1	1	1	0	0	0
0	0	0	0	1	1	1	1	1	1	0

PSICOSE	ptrc	QUINATE	RIBITOL	rmn	sbt-D	ser-D	ser-L	succ	sucr	thr-L
1	0	1	1	1	1	1	0	1	1	0
1	0	1	1	1	1	1	1	1	1	0
1	0	1	1	1	1	0	0	1	0	0
1	0	1	1	1	1	1	1	1	1	0
1	0	1	1	1	1	1	0	0	0	0
1	0	1	1	0	1	1	1	0	1	0
1	0	1	1	0	1	1	1	1	0	0
1	0	1	1	0	0	0	0	0	0	0
1	0	1	1	1	0	0	1	1	1	0
1	0	1	1	0	0	0	0	0	1	0
1	0	1	1	1	1	1	0	1	1	0
1	0	1	1	0	1	1	1	0	0	0
1	0	1	1	1	1	1	1	1	1	0

thymd	tre	Turanose	Tween_40	Tween_80	uri	UROCONATE	xylt
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.8667	1.92591	-1	-1	-1	0.856802	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1
0.866702	1.92592	-1	-1	-1	0.856804	-1	-1

thymd	tre	Turanose	Tween_40	Tween_80	uri	UROCONATE	xylt
1	1	0	0	0	0	0	0
1	1	0	0	0	1	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	1	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	1	0	0
1	1	0	0	0	1	0	0
0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	0
1	1	0	0	0	1	0	0
1	1	0	0	0	1	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	1	0	0

thymd	tre	Turanose	Tween_40	Tween_80	uri	UROCONATE	xylt
1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
0	0	1	1	1	0	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1

Annexe 7 : Comparaisons des observations Biologs en mesure sur la courbe de croissance et des prédictions du flux de biomasses par FBA.

13 souches sont comparées sur 94 milieux. Une souche est capable d'utiliser une source de carbone si la moyenne des mesures des points finaux est supérieure à 0.258. Un modèle est capable d'utiliser une source de carbone si le flux de biomasses est supérieur à zéro. 76% des prédictions sont cohérentes avec l'observation

Prédictions	2obut	3-HYDROXY-L-PROLINE	4-HYDROXY-BUTYRATE	4-HYDROXYPHENYLACETATE	4abut	5-OXOPROLINE	ac
042	-2	-1	-1	-1	0.636152	-1	0.24956
536	-2	-1	-1	-1	0.636152	-1	0.24956
55989	-2	-1	-1	-1	0.636152	-1	0.24956
APECO1	-2	-1	-1	-1	0.636152	-1	0.24956
CFT073	-2	-1	-1	-1	0.636152	-1	0.24956
ED1a	-2	-1	-1	-1	0.636152	-2	0.24956
HS	-2	-1	-1	-2	0.636152	-1	0.24956
IAI1	-2	-1	-1	-2	0.636152	-1	0.24956
K12MG1655	-2	-1	-1	-1	0.63615	-1	0.24956
O127H6E234869	-2	-1	-1	-1	0.63615	-1	0.24956
O157H7EDL933	-2	-1	-1	-2	0.636152	-1	0.24956
O157H7sakai	-2	-1	-1	-2	0.636152	-1	0.24956
S88	-2	-1	-1	-1	0.636152	-2	0.24956
UMN026	-2	-1	-1	-2	0.636152	-1	0.24956

Transporteur absent — Metabolite absent du modèle

Flux de biomasse non nul — Flux de biomasse nul

Observations	2obut	3-HYDROXY-L-PROLINE	4-HYDROXY-BUTYRATE	4-HYDROXYPHENYLACETATE	4abut	5-OXOPROLINE	ac
042	0	0	0	0	0	0	0
536	0	0	0	0	0	0	0
55989	0	0	0	0	0	0	0
APECO1	0	0	0	0	0	0	0
CFT073	0	0	0	0	0	0	0
ED1a	0	0	0	0	0	0	0
HS	0	0	0	1	0	0	0
IAI1	0	0	0	1	0	0	0
K12MG1655	0	0	0	0	0	0	0
O127H6E234869	0	0	0	0	0	0	0
O157H7EDL933	0	0	0	0	0	0	0
O157H7sakai	0	0	0	0	0	0	0
S88	0	0	0	0	0	0	0
UMN026	0	0	0	0	0	0	0

Flux de biomasse non nul — Flux de biomasse nul

Comparaisons	2obut	3-HYDROXY-L-PROLINE	4-HYDROXY-BUTYRATE	4-HYDROXYPHENYLACETATE	4abut	5-OXOPROLINE	ac
042	1	1	1	1	0	1	0
536	1	1	1	1	0	1	0
55989	1	1	1	1	0	1	0
APECO1	1	1	1	1	0	1	0
CFT073	1	1	1	1	0	1	0
ED1a	1	1	1	1	0	1	0
HS	1	1	1	0	0	1	0
IAI1	1	1	1	0	0	1	0
K12MG1655	1	1	1	1	0	1	0
O127H6E234869	1	1	1	1	0	1	0
O157H7EDL933	1	1	1	1	0	1	0
O157H7sakai	1	1	1	1	0	1	0
S88	1	1	1	1	0	1	0
UMN026	1	1	1	1	0	1	0

Similitude prédiction/observation — Incohérence prédiction/simulation

	996	320	1316			pousse	pousse pas
	76%	24%			F-mesure	70%	14%
	Table de vérité			précision			
		Modèle		Modèle			
		Croissance	Pas de croissance	Croissance	Pas de croissance		
Biolog	Croissance	373	63	86%	14%	436	

	Pas de Croissance	257	623	29%	71%	880
		59%	9%			
	rappel	41%	91%			
Bpousse/Mpousse		630	686			

acgal	acgam	acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL
0	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.6093	0.434192	-1	0.434192	0.792433	0.459324	0.458862	-1	-1
0.00E+00	1.23297	-2	0.6093	0.434192	-1	0.434192	0.792433	0.459324	0.458862	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1
0.00E+00	1.23297	-2	0.609301	0.434192	-1	0.434192	0.792435	0.459326	0.458864	-1	-1

acgal	acgam	acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL
1	1	0	0	0	0	1	1	0	0	0	0
1	1	0	0	0	1	1	1	0	0	0	0
1	1	0	0	0	0	0	1	0	0	0	0
1	1	0	0	0	1	1	1	0	0	0	0
1	1	0	0	0	0	0	1	0	1	0	0
1	1	0	0	0	0	0	1	1	1	0	0
1	1	0	0	1	1	1	1	1	0	0	0
1	1	0	1	1	1	1	1	1	1	0	0
0	1	0	1	0	0	1	1	0	0	0	0
1	1	0	0	0	0	1	1	1	1	0	0
1	1	0	0	0	1	0	1	0	0	0	0
1	1	0	0	0	0	0	1	0	0	0	0
1	1	0	0	1	1	1	1	0	0	0	0
1	1	0	0	1	1	1	1	0	0	0	0

acgal	acgam	acon-C	akg	ala-D	ALA-GLY	ala-L	arab-L	asn-L	asp-L	Bromosuccinic_acid	BUTANEDIOL
0	1	1	0	0	1	1	1	0	0	1	1
0	1	1	0	0	0	1	1	0	0	1	1
0	1	1	0	0	1	0	1	0	0	1	1
0	1	1	0	0	0	1	1	0	0	1	1
0	1	1	0	0	1	0	1	0	1	1	1
0	1	1	0	0	1	0	1	1	1	1	1
0	1	1	0	1	0	1	1	1	0	1	1
0	1	1	1	1	0	1	1	1	1	1	1
1	1	1	1	0	1	1	1	0	0	1	1
0	1	1	0	0	1	1	1	1	1	1	1
0	1	1	0	0	0	0	1	0	0	1	1
0	1	1	0	0	1	0	1	0	0	1	1
0	1	1	0	1	0	1	1	0	0	1	1
0	1	1	0	1	0	1	1	0	0	1	1

CELLOBIOSE	cit	CPD-1099	CPD-1843	CPD-2461	CPD-355	CPD-3561	CPD-3564	CPD-3569	CPD-3570	CPD-3573
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1

-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688294	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688294	-1	-1	-1	0.872679	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	0.688295	-1	-1	-1	-1	-1	-1	-1	-1	-1

CELLOBIOSE	cit	CPD-1099	CPD-1843	CPD-2461	CPD-355	CPD-3561	CPD-3564	CPD-3569	CPD-3570	CPD-3573
0	0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	1	0	0	0	1
0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	1	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	0	1
0	0	1	0	0	0	0	0	0	0	1

CELLOBIOSE	cit	CPD-1099	CPD-1843	CPD-2461	CPD-355	CPD-3561	CPD-3564	CPD-3569	CPD-3570	CPD-3573
1	0	1	1	1	1	1	1	1	1	0
1	0	0	1	1	1	0	1	1	1	0
1	0	1	1	1	1	0	1	1	1	1
1	0	1	1	1	1	0	1	1	1	0
1	0	1	1	1	1	0	1	1	1	0
1	0	1	1	1	1	0	1	1	1	0
1	0	1	1	1	1	1	1	1	1	0
1	0	0	1	1	1	0	1	1	1	0
1	0	1	1	1	1	1	1	1	1	0
1	0	0	1	1	1	1	1	1	1	0
1	0	0	1	1	1	1	1	1	1	0
1	0	0	1	1	1	1	1	1	1	0
1	0	1	1	1	1	0	1	1	1	0
1	0	0	1	1	1	1	1	1	1	0

CPD-3605	CPD-3618	CPD-3623	CPD-3782	CPD-7692	CPD-8979	CPD0-1656	crn	D-GALACTONO-1-4-LACTONE	Dextrins
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845599	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845599	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.00E+00	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.00E+00	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182
-1	-1	-1	-1	-1	-1	-1	0	0.845601	3.8182

CPD-3605	CPD-3618	CPD-3623	CPD-3782	CPD-7692	CPD-8979	CPD0-1656	crn	D-GALACTONO-1-4-LACTONE	Dextrins
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	1	1

CPD-3605	CPD-3618	CPD-3623	CPD-3782	CPD-7692	CPD-8979	CPD0-1656	crn	D-GALACTONO-1-4-LACTONE	Dextrins
1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

ERYTHRITOL	etha	for	fru	fuc-L	g1p	g6p	gal	galur	glcn	glcr	glcur
-1	0.343597	0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343653	0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343597	0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343653	0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343653	0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343653	0.042145	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343597	0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343597	0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343597	0.0116819	0.962956	0.943244	0.962956	0.993049	0.952926	0.777248	0.882211	0.679604	0.777248
-1	0.343597	0.0116819	0.971811	0.943244	0.962956	0.993049	0.952926	0.777248	0.882211	0.679604	0.777248
-1	0.343597	0.042029	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343597	0.042029	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343597	0.042029	0.971813	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725
-1	0.343597	0.042029	0.962959	0.943246	0.962959	0.993051	0.952928	0.77725	0.882213	0.679606	0.77725

ERYTHRITOL	etha	for	fru	fuc-L	g1p	g6p	gal	galur	glcn	glcr	glcur
0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	1	0	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1	1	0	1
0	0	0	1	0	1	1	1	1	1	0	1
0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	0	1	0	0
0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1	1	1	1

0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1	1	0	1
0	0	0	1	1	1	1	1	1	1	0	1
0	0	0	1	1	1	1	1	1	1	0	1
0	0	0	1	1	1	1	1	1	1	1	1

ERYTHRITOL	etha	for	fru	fuc-L	g1p	g6p	gal	galur	glcn	glcr	glcur
1	0	0	1	1	1	1	1	1	1	1	1
1	0	0	1	0	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	1	1	0	1
1	0	0	1	0	1	1	1	1	1	0	1
1	0	0	1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	0	1	0	0
1	0	0	1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	1	1	0	1
1	0	0	1	1	1	1	1	1	1	0	1
1	0	0	1	1	1	1	1	1	1	0	1
1	0	0	1	1	1	1	1	1	1	1	1

glu-L	GLUCOSAMINATE	glyc	glyc3p	glycogen	Glycyl-L-aspartic_acid	his-L	inost	ins	ITACONATE	L-Proline
0.688295	-1	0.558276	0.597489	-2	-1	0	0	1.19303	-1	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0	1.19E+00	-2	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0	1.19E+00	-1	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0	1.19E+00	-2	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0	1.19E+00	-2	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	9.03E-01	1.19E+00	-2	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0.00E+00	1.19E+00	-1	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0.00E+00	1.19E+00	-1	-1.00E+00
0.688294	-1	0.558274	0.597487	-2	-1	0	0.00E+00	1.19E+00	-1	-1.00E+00
0.688294	-1	0.558274	0.597487	-2	-1	0	0.00E+00	1.19265	-2	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0.00E+00	1.19E+00	-1	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0.00E+00	1.19E+00	-1	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	0.00E+00	1.19E+00	-2	-1.00E+00
0.688295	-1	0.558276	0.597489	-2	-1	0	9.03E-01	1.19E+00	-1	-1.00E+00

glu-L	GLUCOSAMINATE	glyc	glyc3p	glycogen	Glycyl-L-aspartic_acid	his-L	inost	ins	ITACONATE	L-Proline
0	0	1	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	1	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	1	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	1	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0
0	0	1	1	0	1	0	0	1	0	0
0	0	1	0	0	0	0	0	1	0	0

glu-L	GLUCOSAMINATE	glyc	glyc3p	glycogen	Glycyl-L-aspartic_acid	his-L	inost	ins	ITACONATE	L-Proline
0	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	0	1	1	1	1	1
0	1	1	0	1	1	1	0	1	1	1

lac-D	lcts	leu-L	mal-D	MALONATE	man	melib	mnl	orn	peacm	phe-L	ppa	PSICOSE	ptrc
0.426301	1.91589	0	0.3945	-1	0.962959	0	1.03284	0.746316	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0	-1	0.962959	0.00E+00	1.03284	0.774549	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0.3945	-1	0.962959	1.91589	1.03284	0.774549	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0	-1	0.962959	0.00E+00	1.03284	0.774549	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0	-1	0.962959	0.00E+00	1.03284	0.774549	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0	-1	0.962959	0.00E+00	1.03284	0.774549	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0.3945	-1	0.962959	0.00E+00	1.03284	0.746316	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0.3945	-1	0.962959	1.91589	1.03284	7.75E-01	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91588	0	0.3945	-1	0.962956	1.91588	1.03284	7.75E-01	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91588	0	0	-1	0.962956	0.00E+00	1.03284	7.75E-01	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0.3945	-1	0.962959	1.91589	1.03284	7.75E-01	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0.3945	-1	0.962959	1.91589	1.03284	7.75E-01	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0	-1	0.962959	0.00E+00	1.03284	7.75E-01	-1	0.00E+00	0.426301	-1	0.760581
0.426301	1.91589	0	0.3945	-1	0.962959	1.91589	1.03284	0.746316	-1	0.00E+00	0.426301	-1	0.760581

lac-D	lcts	leu-L	mal-D	MALONATE	man	melib	mnl	orn	peacm	phe-L	ppa	PSICOSE	ptrc
1	1	0	0	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	0	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0
1	1	0	1	0	1	1	1	0	0	0	0	0	0

lac-D	lcts	leu-L	mal-D	MALONATE	man	melib	mnl	orn	peacm	phe-L	ppa	PSICOSE	ptrc
1	1	1	0	1	1	0	1	0	1	1	0	1	0

1	1	1	0	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	1	1	1	0	1	1	0	1	0
1	1	1	0	1	1	0	1	0	1	1	0	1	0
1	1	1	0	1	1	0	1	0	1	1	0	1	0
1	1	1	0	1	1	0	1	0	1	1	0	1	0
1	1	1	1	1	1	0	1	0	1	1	0	1	0
1	1	1	1	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	1	1	1	0	1	1	0	1	0
1	1	1	0	1	1	0	1	0	1	1	0	1	0
1	1	1	1	1	1	1	1	0	1	1	0	1	0
1	1	1	1	1	1	1	1	0	1	1	0	1	0
1	1	1	0	1	1	0	1	0	1	1	0	1	0
1	1	1	1	1	1	1	1	0	1	1	0	1	0

QUINATE	RIBITOL	rmn	sbt-D	ser-D	ser-L	succ	sucr	thr-L	thymd	tre	Turanose	Tween_40
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943244	1.03284	0.365916	0.365916	0.490148	1.92591	0.588279	0.8667	1.92591	-1	-1
-1	-1	0.943244	0.00E+00	0.365916	0.365916	0.490148	1.92591	0.588279	0.8667	1.92591	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1
-1	-1	0.943246	1.03284	0.367216	0.367216	0.49015	1.92592	0.588547	0.866702	1.92592	-1	-1

QUINATE	RIBITOL	rmn	sbt-D	ser-D	ser-L	succ	sucr	thr-L	thymd	tre	Turanose	Tween_40
0	0	1	1	0	0	1	0	0	1	1	0	0
0	0	1	1	1	0	0	1	0	1	1	0	0
0	0	1	1	0	0	0	0	0	1	1	0	0
0	0	1	1	1	1	1	0	0	1	1	0	0
0	0	1	1	1	0	0	0	0	1	1	0	0
0	0	0	1	1	0	0	0	0	1	1	0	0
0	0	0	1	1	1	1	0	0	1	1	0	0
0	0	1	1	0	0	1	1	0	1	1	0	0
0	0	1	1	1	0	1	0	0	1	1	0	0
0	0	1	1	0	0	1	1	0	1	1	0	0
0	0	0	0	0	0	0	1	0	1	1	0	0
0	0	0	0	0	0	0	1	0	1	1	0	0
0	0	0	1	1	1	1	0	0	1	1	0	0
0	0	1	1	0	0	1	1	0	1	1	0	0

QUINATE	RIBITOL	rmn	sbt-D	ser-D	ser-L	succ	sucr	thr-L	thymd	tre	Turanose	Tween_40
1	1	1	1	0	0	1	0	0	1	1	1	1
1	1	1	1	1	0	0	1	0	1	1	1	1
1	1	1	1	0	0	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	1	1	1	1
1	1	1	1	1	0	0	0	0	1	1	1	1
1	1	0	1	1	0	0	0	0	1	1	1	1
1	1	0	1	1	1	1	0	0	1	1	1	1
1	1	1	1	0	0	1	1	0	1	1	1	1
1	1	1	1	1	0	1	0	0	1	1	1	1
1	1	1	0	0	0	1	1	0	1	1	1	1
1	1	0	0	0	0	0	1	0	1	1	1	1
1	1	0	0	0	0	0	1	0	1	1	1	1
1	1	0	1	1	1	1	0	0	1	1	1	1

1	1	1	1	0	0	1	1	0	1	1	1	1

Tween_80	uri	UROCONATE	xylt
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856802	-1	-1
-1	0.856802	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1
-1	0.856804	-1	-1

Tween_80	uri	UROCONATE	xylt
0	0	0	0
0	1	0	0
0	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	0	0	0
0	1	0	0
0	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0

Tween_80	uri	UROCONATE	xylt
1	0	1	1
1	1	1	1
1	0	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	0	1	1
1	1	1	1
1	0	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1