Analyse systémique de la symbiose intracellulaire : évolution et organisation du réseau métabolique des endocytobiotes.

Ludovic Cottret

▶ **To cite this version:**

Ludovic Cottret. Analyse systémique de la symbiose intracellulaire : évolution et organisation du réseau métabolique des endocytobiotes.. Sciences du Vivant [q-bio]. Université Claude Bernard - Lyon I, 2009. Français. NNT : . tel-00494581

HAL Id: tel-00494581

https://theses.hal.science/tel-00494581

Submitted on 23 Jun 2010

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

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

N° d'ordre ? ? ?- ? ? ? Année 2009

Thèse

Présentée

devant l'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

et soutenue publiquement le
26 janvier 2009

par

Ludovic Cottret

Analyse systémique de la symbiose intracellulaire : évolution et organisation du réseau métabolique des endocytobiotes

---

Directrice de thèse : Marie-France Sagot
Co-encadrant : Hubert Charles

Jury :	Hubert Charles,	Co-encadrant
	Christine Dillmann,	Rapporteuse
	Frédéric Fleury,	Examinateur
	Arlindo Oliveira,	Rapporteur
	Marie-France Sagot	Directrice

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J. F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur J. LIETO
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. SIMON
Secrétaire Général	**M. G. GAY**

SECTEUR SANTÉ

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Ontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur A. HOAREAU
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur A. BRIGUET
UFR de Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématique	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur M. EGEA
UFR de Chimie Biochimie	Directeur : Mme. le Professeur H. PARROT
UFR STAPS	Directeur : M. le Professeur R. MASSARELLI
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J. C. AUGROS

Table des matières

Introduction

Dans l'imaginaire collectif, les bactéries se rangent parmi les fléaux naturels qui empoisonnent l'humanité. On sait en effet depuis le XIXème siècle que certaines sont à l'origine des épidémies les plus foudroyantes, telles que le choléra ou la peste.

Jusqu'à récemment, l'influence des bactéries sur l'évolution des organismes se limitait à décimer certaines populations et à agir sur l'amélioration des défenses immunitaires. La relation d'une bactérie avec un autre organisme ne pouvait être, par ailleurs, que de courte durée, conduisant à la mort de l'un ou de l'autre.

Toutefois, il est apparu parallèlement que certains organismes étaient capables d'entretenir une relation durable avec l'organisme qui les abrite, leur hôte, et pouvaient même avoir un effet bénéfique sur ce dernier. Dès 1886, le lichénologue Anton de Bary définit ainsi la symbiose comme "*l'association permanente entre plusieurs organismes d'espèces distinctes, au moins pendant une partie de leur cycle de vie*". En outre, on découvre, à la même époque, certaines bactéries symbiotiques, appelées plus tard bactéries endocytobiotes, au sein même des cellules de leur hôte.

Depuis la définition d'A. de Bary, la conception de la symbiose, et particulièrement de la symbiose intracellulaire (endocytobiose), a largement évolué. Aujourd'hui, cette association est considérée comme un phénomène très largement répandu et comme un des facteurs primordiaux dans l'évolution des espèces.

Si l'on considère l'ensemble des interactions durables entre les membres de la symbiose, on remarque que l'effet de l'endocytobiote sur son hôte peut être négatif ou positif. Lorsqu'il est négatif, l'interaction avec l'endocytobiote diminue la valeur sélective de l'hôte, c'est-à-dire sa longévité et le nombre de ses descendants : on parle alors de parasitisme. En revanche, lorsque la valeur sélective de l'hôte est augmentée, on parle de mutualisme. Un continuum est observé dans la nature entre ces deux formes d'interactions.

L'environnement particulier qu'est la cellule de l'hôte, et le type de relations entre les deux partenaires (parasitisme ou mutualisme), ont naturellement des conséquences sur l'évolution de leur métabolisme respectif.

Durant cette thèse, nous nous sommes focalisés sur le métabolisme des bactéries endocytobiotes.

Le métabolisme d'un organisme, même s'il a connu des réductions importantes comme dans le cas des endocytobiotes, est un système dont le nombre d'éléments (réactions ou métabolites) est important. De plus, même si une armée de chercheurs s'attelait à l'analyse individuelle de chaque élément d'un réseau métabolique, son fonctionnement global resterait impossible à prédire. Une fonction métabolique d'un organisme est déterminée par une succession de réactions, appelée voie métabolique, reliées entre elles par les métabolites qu'elles utilisent et produisent. Une réaction d'une telle succession, prise individuellement, serait certainement inutile à la cellule. La même réaction peut également participer à une autre voie métabolique correspondant à une fonction métabolique com-

plètement différente. Nous appréhendons donc ici le métabolisme d'une manière globale (systémique), sous la forme du réseau métabolique, ensemble des réactions biochimiques, des métabolites et de leurs interactions dans un organisme considéré.

Nous nous sommes donnés comme objectif global de mieux comprendre l'influence de la symbiose intracellulaire et du type de relations avec l'hôte (mutualisme ou parasitisme) sur l'évolution et le fonctionnement du métabolisme des endocytobiotes.

Dans un premier temps, nous avons réuni dans une base appelée SymbioCyc les données métaboliques d'une cinquantaine de bactéries parmi lesquelles une trentaine sont intracellulaires. Outre des outils originaux de comparaison de réseaux métaboliques, SymbioCyc dispose également de fonctions de filtre et d'export des données nécessaires pour la modélisation.

Dans un second temps, nous avons confronté les réseaux métaboliques d'une trentaine de bactéries provenant de SymbioCyc afin d'étudier l'influence du mode de vie des bactéries sur la composition et l'organisation de leur réseau métabolique. La comparaison de la composition du réseau métabolique s'est effectuée par l'analyse des ensembles des gènes dédiés au métabolisme, des réactions et des métabolites. Pour comparer l'organisation des réseaux métaboliques, nous avons modélisé le réseau sous la forme d'un graphe, objet mathématique permettant de représenter un réseau comme un ensemble de noeuds reliés entre eux par des arêtes. Des mesures sur ces graphes nous ont permis de distinguer leurs topologies globales mais aussi d'indiquer certaines parties centrales en fonction des bactéries considérées.

Enfin, dans un troisième temps, afin de mieux appréhender les relations entre le métabolisme de l'endocyobiote et celui de l'hôte, nous avons développé une méthode, appelée PITUFO, capable de déterminer les ensembles de métabolites qu'un organisme pourrait capter dans son environnement afin d'assurer certaines fonctions métaboliques. Nous avons appliqué cette méthode sur le métabolisme des acides aminés de *Buchnera aphidicola*, bactérie endocytobiote que l'on trouve au sein des cellules du puceron.

Ce travail s'étend ainsi du traitement des données à l'analyse de celles-ci, en passant par le développement de méthodes originales. Nous espérons que d'avoir réalisé nous-mêmes l'ensemble de ces tâches, malheureusement souvent compartimentées dans les travaux de bio-informatique, nous a permis de garder un oeil critique à la fois sur les données et sur les résultats des méthodes.

Cette thèse s'inscrit dans la biologie des systèmes où l'analyse des interactions entre les éléments d'un système permet de déterminer des propriétés globales du système lui-même. Malheureusement, beaucoup de propriétés globales résultant d'analyses en biologie des systèmes ont une signification biologique très floue

ou même erronée. Nous nous sommes donc attachés à associer des propriétés globales à la nature et à la fonction des éléments eux-mêmes (ici les réactions et les composés). Dans la mesure du possible, ces résultats ont été reliés à la biologie des organismes considérés.

Notre espoir est ainsi que les analyses effectuées fournissent au lecteur une vue à la fois globale et détaillée du métabolisme des bactéries endocytobiotes, et le guident vers une meilleure compréhension de son fonctionnement et de son évolution.

Première partie

Contextes biologique et méthodologique

Chapitre 1

Le métabolisme des endocytobiotes

1.1 Le métabolisme : généralités et définitions

1.1.1 Historique

Le terme *métabolisme* est dérivé du grec *metabolismos* qui signifie "changement" ou "transformation". Dans le Trésor de la Langue Française informatisé, on trouve la définition suivante :

"Ensemble des réactions de synthèse, génératrices de matériaux (anabolisme), et de dégradation, génératrices d'énergie (catabolisme), qui s'effectuent au sein de la matière vivante à partir des constituants chimiques fournis à l'organisme par l'alimentation et sous l'action de catalyseurs spécifiques."

Le concept de métabolisme fut proposé au XIIIème siècle par Ibn al-Nafis (1213-1288), d'origine syrienne et dont les travaux ont été menés au Caire. C'est le premier, ou l'un des premiers, à dépeindre les transformations continues que subit le corps. Il décrit le corps et ses composants comme étant "dans un état continu de dissolution et d'alimentation, ainsi ils subissent inévitablement des changements permanents" (Al-Roubi, 1982). Les premières expériences métaboliques ont cependant été effectuées quatre siècles plus tard par le vénitien Santorio Sanctorius (1561-1636) sur sa propre personne (Eknoyan, 1999). Sanctorius imagina et fit construire une balance métabolique sur laquelle il se livrait à ses diverses occupations quotidiennes (Figure 1.1). Pesant exactement ce qu'il ingurgitait et ce qu'il rejetait dans ses féces et ses urines, il mit en évidence une perte de matière qu'il appela "transpiration invisible".

Jusqu'au XIXème siècle pourtant, ces constatations restèrent sans explication vraiment convaincante du point de vue scientifique. Avant cette période, il était reconnu que les substances organiques ne pouvaient être fabriquées que par les

Figure 1.1. La balance métabolique imaginée par Santorio Sanctorius pour l'étude de la "transpiration invisible". Tiré de *De Statica medecina*, 1690 (Eknoyan, 1999).

êtres vivants. C'est le concept de *vitalisme* selon lequel un principe vital, complètement distinct des forces chimiques gouvernant le monde inanimé, crée la matière organique.

Il faut attendre 1828 avec la synthèse accidentelle d'urée à partir de deux composés inorganiques par l'allemand Friedrich Wöhler (1800-1882), pour montrer qu'il était possible de fabriquer des substances du vivant à partir de substances inanimées (Kinne-Saffran & Kinne, 1999). Cette découverte s'inscrit dans un mouvement essentiellement porté par l'allemand Carl Ludwig (1816-1895) qui considère les lois de la physique et de la chimie comme les seules gouvernant les processus physiologiques (Zimmer, 1996). C'est Wilhelm Kühne (1837–1900), de la même école qui, pour la première fois, utilisa le terme *enzyme* (du grec "zyme" signifiant levain) pour désigner les *ferments* décrits par Louis Pasteur (1822-1895) (Manchester, 1995). Celui-ci, en étudiant la fermentation du sucre en alcool par la levure, parvint à la conclusion que la fermentation était catalysée par des éléments contenus dans les cellules de la levure qu'il appela *ferments*, qu'il considérait comme propres aux organismes vivants.

En 1897, l'allemand Eduard Buchner (1860-1917) compléta ces analyses en mettant en évidence que des extraits de levure étaient capables d'activer la fermentation du sucre en alcool en l'absence de levures vivantes, ce qui lui valut le Prix Nobel en 1907 (Jaenicke, 2007). Il nomma "zymase" l'agent catalysant la fermentation chez la levure. Depuis, les enzymes sont nommées d'après la réaction qu'elles catalysent ou d'après le substrat sur lequel elles agissent, auquel on

ajoute le suffixe -ase. C'est par ces expériences également que l'étude des réactions biochimiques s'est dissociée de l'étude de la cellule, marquant ainsi le début de la biochimie.

Buchner prouva que les enzymes pouvaient fonctionner en dehors de toute matière vivante, il restait à découvrir leur nature biochimique. C'est l'américain James B. Sumner (1887-1955) qui, en 1926, isola et purifia une enzyme, l'uréase, qu'il identifia comme une protéine (Sumner, 1933). Un peu plus tard, les américains John H. Northrop et Wendell M. Stanley parvinrent aux mêmes résultats avec d'autres enzymes (Northrop, 1929; Norrby, 2008). En 1946, le Prix Nobel a été décerné à ces 3 chercheurs pour ces découvertes.

Dans le même temps, l'amélioration des techniques expérimentales permit l'identification de nombreuses voies métaboliques. Le plus célèbre et le plus productif des chercheurs dans ce domaine est sans doute l'allemand Hans A. Krebs, qui donna son nom au cycle de l'acide citrique en 1937 mais fut aussi le découvreur de deux autres cycles : le cycle de l'ornithine en 1931 et le cycle du glyoxylate en 1957 (Kornberg, 2000).

Le lien entre gènes et protéines a d'abord été révélé grâce aux études sur les enzymes. Les américains George W. Beadle (1903-1989) et Edward L. Tatum (1909-1975) montrèrent en effet en 1941 que des mutations génétiques pouvaient affecter directement certaines étapes d'une voie métabolique, proposant ainsi un lien direct entre gènes et enzymes et donnant naissance à ce qu'on peut appeler la génétique biochimique (Horowitz, 1990).

L'accumulation des données métaboliques, enzymatiques en particulier, a necessité l'apport des mathématiques et la création de cadres théoriques afin d'analyser le comportement global de sous-systèmes comme les voies métaboliques. C'est ainsi que les années 1960 et 1970 ont vu la naissance et le développement de la théorie du contrôle métabolique et de la théorie des systèmes biochimiques (voir Section 2.2) qui permettent une analyse numérique globale d'un système métabolique.

Depuis les années 1990, l'explosion du nombre de données générées par les techniques à haut débit a permis de ne plus restreindre l'analyse à une partie du réseau métabolique mais de l'étendre à son ensemble.

L'exploration du réseau métabolique s'inscrit ainsi dans l'étude des systèmes complexes et en particulier de la biologie des systèmes. Un des concepts primordiaux pour comprendre les systèmes complexes est l'émergence. Un système complexe a des propriétés dites "émergentes", c'est-à-dire qui ne peuvent pas s'expliquer par les propriétés individuelles de chacun de ses éléments. Les propriétés d'un système complexe découlent des interactions entre les éléments plutôt que directement des propriétés de ces éléments.

Différents systèmes biologiques peuvent être considérés selon l'échelle à laquelle on se place. Chaque niveau d'intégration peut être expliqué par les relations entre les éléments qui le composent. Ainsi, à l'échelle de la planète, on étudie les

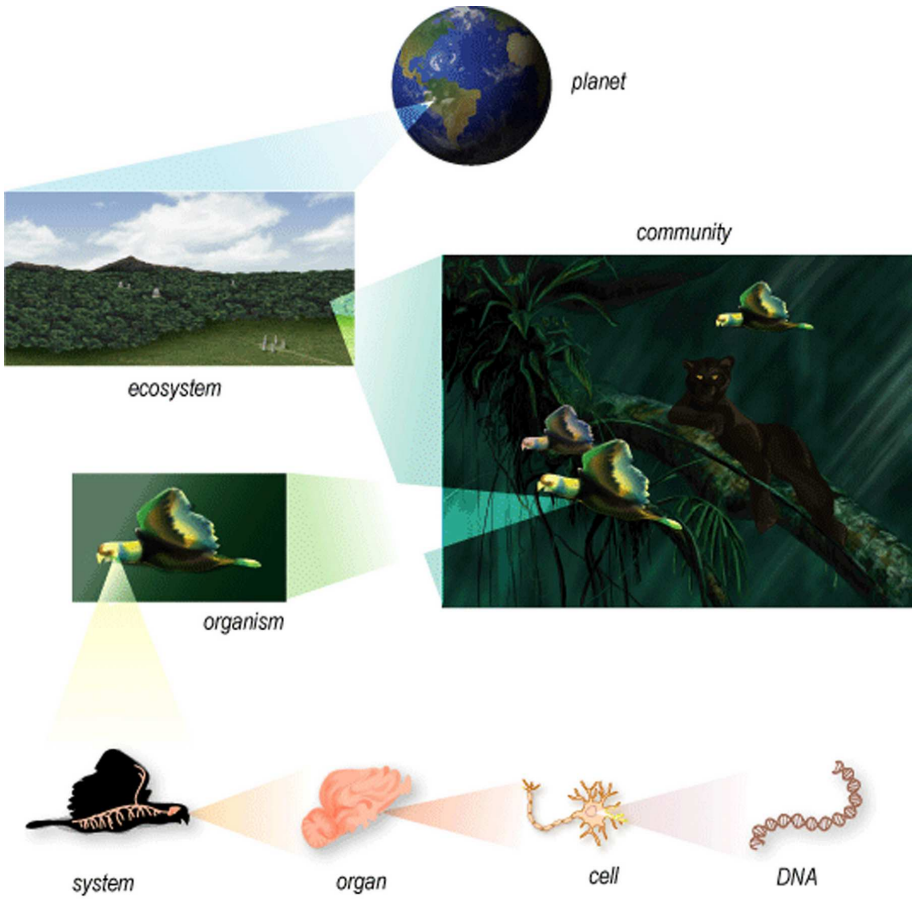

Figure 1.2. Les différents niveaux d'intégration en biologie. Source : The Science Creative Quarterly" (scq.ubc.ca), Jen Philpot)

relations entre écosystèmes, à l'échelle d'un écosystème, on étudie les relations entre populations d'organisme (Figure 1.2). L'étude du réseau métabolique se situe à l'échelle de la cellule dont le fonctionnement global est expliqué par les interactions entre les molécules.

On a l'habitude de considérer trois réseaux d'interactions entre molécules. Cette distinction est basée essentiellement sur les techniques expérimentales associées. Ainsi, les techniques de séquençage du génome et les puces à ADN, entre autres, permettent l'analyse du réseau de régulation de gènes. Un gène peut en effet être régulé positivement ou négativement par un facteur de transcription, souvent un complexe protéique codé par d'autres gènes. D'autres techniques comme les analyses double-hybride permettent d'établir un réseau d'interactions entre protéines. Enfin, certaines protéines ou complexes protéiques ont une activité enzymatique. Chaque enzyme catalyse une ou plusieurs réactions biochimiques qui relient les métabolites entre eux, formant ce que l'on appelle le réseau métabolique.

En réalité, les trois réseaux sont intimement liés (Figure 1.3) et l'analyse de

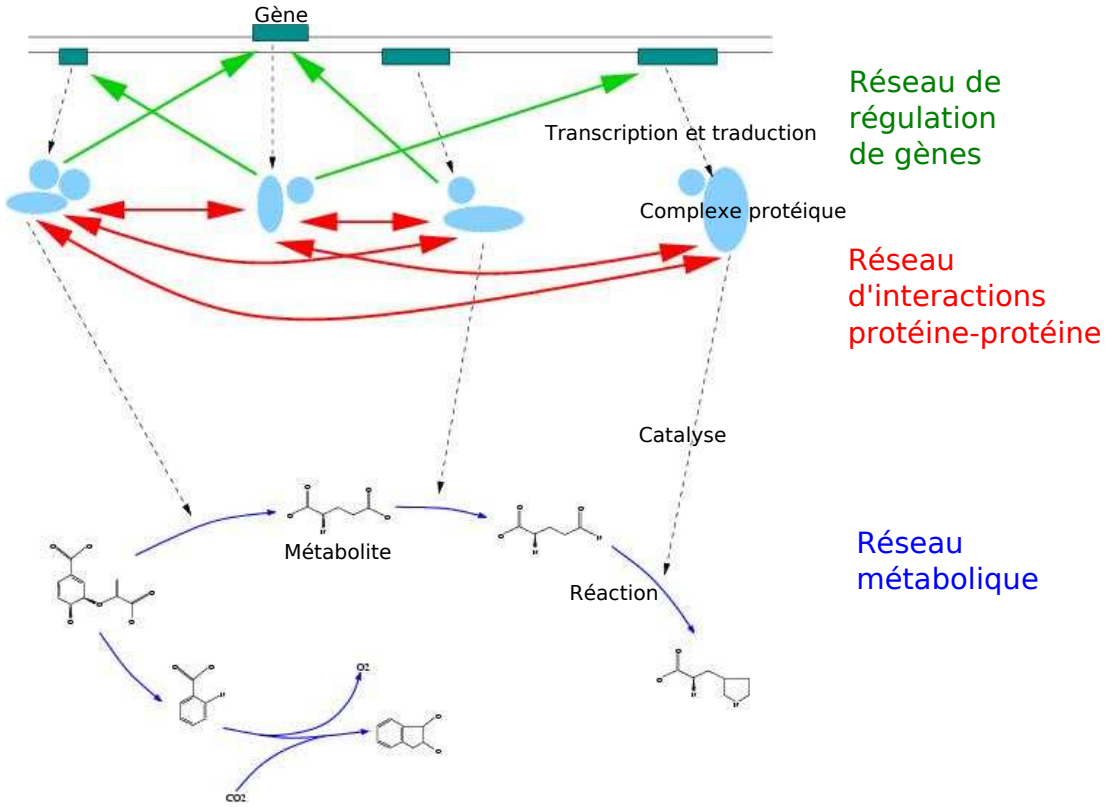

Figure 1.3. Les liens entre réseau de gènes, réseau d'interactions protéine-protéine et réseau métabolique.

chacun d'entre eux devrait se faire à la lumière des deux autres. Afin de mieux comprendre les particularités du réseau métabolique, nous proposons dans la Section suivante une définition de chaque entité qui le compose ou qui influe sur son fonctionnement.

La rapidité des techniques de séquençage actuelles a permis de séquencer à ce jour plus de 500 organismes (essentiellement des bactéries). Comme nous l'avons vu plus haut, il existe un lien entre gènes, enzymes et réactions enzymatiques. L'un des défis actuels est de proposer une reconstruction des réseaux métaboliques des différents organismes séquencés. Nous verrons dans la Section 2.1 les différentes méthodes et stratégies, mais aussi les progrès qu'il reste à faire pour obtenir des reconstructions métaboliques les plus fidèles possible à partir d'informations génomiques. Dans le même temps, les techniques de métabolomique deviennent de plus en plus précises et permettent d'obtenir très vite la liste et les quantités de métabolites présents dans un organisme ou dans un tissu donné. Cependant, la plupart des reconstructions métaboliques actuelles sont effectuées à partir des seules informations génomiques.

Par ailleurs, la quantité et la dispersion des données nécessitent également de faciliter l'échange des données et de mettre à disposition des outils permettant la modélisation. C'est dans cette optique que SymbioCyc a été développé (voir

Section 3).

Des méthodes comme l'analyse de balance des flux et la recherche de modes élémentaires ont été développées pour pallier au manque de données numériques, souvent absentes des reconstructions automatiques (voir Section 2.2). Cependant, ces méthodes ne sont pas tout le temps faciles à mettre en oeuvre. En effet, elles reposent sur l'hypothèse selon laquelle le système est à l'état d'équilibre, ce qui nécessite d'avoir les proportions exactes des métabolites dans chaque réaction mais aussi de connaître les limites du système. La modélisation sous forme de graphes a connu dès les années 2000 un franc succès pour analyser les réseaux biologiques, en particulier métaboliques. En effet, les graphes permettent une modélisation rapide des réseaux biologiques et bénéficient des nombreuses méthodes et mesures associées à la théorie des graphes. Les analyses que nous avons effectuées sur les graphes des composés (Section 4) montrent l'intérêt d'une telle modélisation pour explorer la topologie des réseaux métaboliques.

L'établissement de classes de réseaux et la déduction de propriétés émergentes à partir d'analyses topologiques a fait l'objet de nombreux articles. Cependant, depuis quelques années, la pertinence de telles analyses et surtout de leurs conclusions est sérieusement remis en cause (voir Section 2.2.4). La plupart des analyses sous forme de graphes montre que le passage du réductionnisme à une approche systémique s'est accompagné d'une perte de pertinence des questions et des résultats biologiques. La nature du réseau et surtout les propriétés propres à chaque noeud ont souvent été largement oubliées. Ainsi, l'analyse des graphes métaboliques nécessite aujourd'hui le développement de méthodes adaptées aux caractéristiques propres du réseau métabolique. C'est dans ce cadre que s'inscrit le développement de notre méthode de recherche de précurseurs, PITUFO (voir Section 5).

Par ailleurs, le nombre et la diversité phylogénétique des organismes pour lesquels des reconstructions métaboliques complètes sont disponibles permet d'envisager l'analyse du métabolisme du point de vue évolutif (voir Section 1.1.3). Une question importante est de comprendre comment le réseau métabolique s'est mis en place au cours de l'évolution. En déduisant les fonctions métaboliques essentielles à la vie, beaucoup de chercheurs tentent d'avoir une idée du métabolisme primordial (Wächtershäuser, 1990, 2007; Maden, 1995; Lazcano & Miller, 1999; Morowitz *et al.*, 2000; Caetano-Anollés *et al.*, 2007). Une autre question, à laquelle nous nous intéressons dans la Section 4, est d'étudier la pression du style de vie des organismes sur l'évolution de leur métabolisme.

Un des défis d'ajourd'hui est ainsi de fournir des outils et des méthodes permettant la comparaison de plusieurs réseaux métaboliques. Cette question est encore largement ouverte et le développement de méthodes encore à son balbutiement. Entre autres, la réflexion devra porter sur les moyens de synthétiser l'information contenue dans un réseau métabolique et sur la détermination des variables importantes qui caractériseraient un réseau métabolique.

L'analyse comparative systématique que nous effectuons grâce à SymbioCyc et

à l'analyse des graphes dans la Section 4 est la première effectuée sur un si grand nombre d'organismes et avec un tel niveau de détail. Nous espérons qu'elle fournira quelques éléments supplémentaires non seulement pour mieux comprendre l'évolution des réseaux métaboliques, en particulier ceux des endocytobiotes, mais aussi pour imaginer de nouvelles approches pour la comparaison de réseaux métaboliques.

1.1.2 Définitions

Un **réseau métabolique** peut-être défini comme une collection d'objets et de relations entre eux. La vision la plus simple du réseau métabolique est une liste de réactions biochimiques. Le réseau est alors formé par les chaînes de métabolites reliés par les réactions. Cependant, le métabolisme étant régulé par l'information génétique, il peut être intéressant de prendre en compte également les informations sur les enzymes et les gènes.

La Figure 1.4 présente une vision "d'informaticien" du réseau mettant en relief les relations entre les objets que composent un réseau métabolique. Nous allons maintenant définir succintement chaque entité et les relations qui les relient.

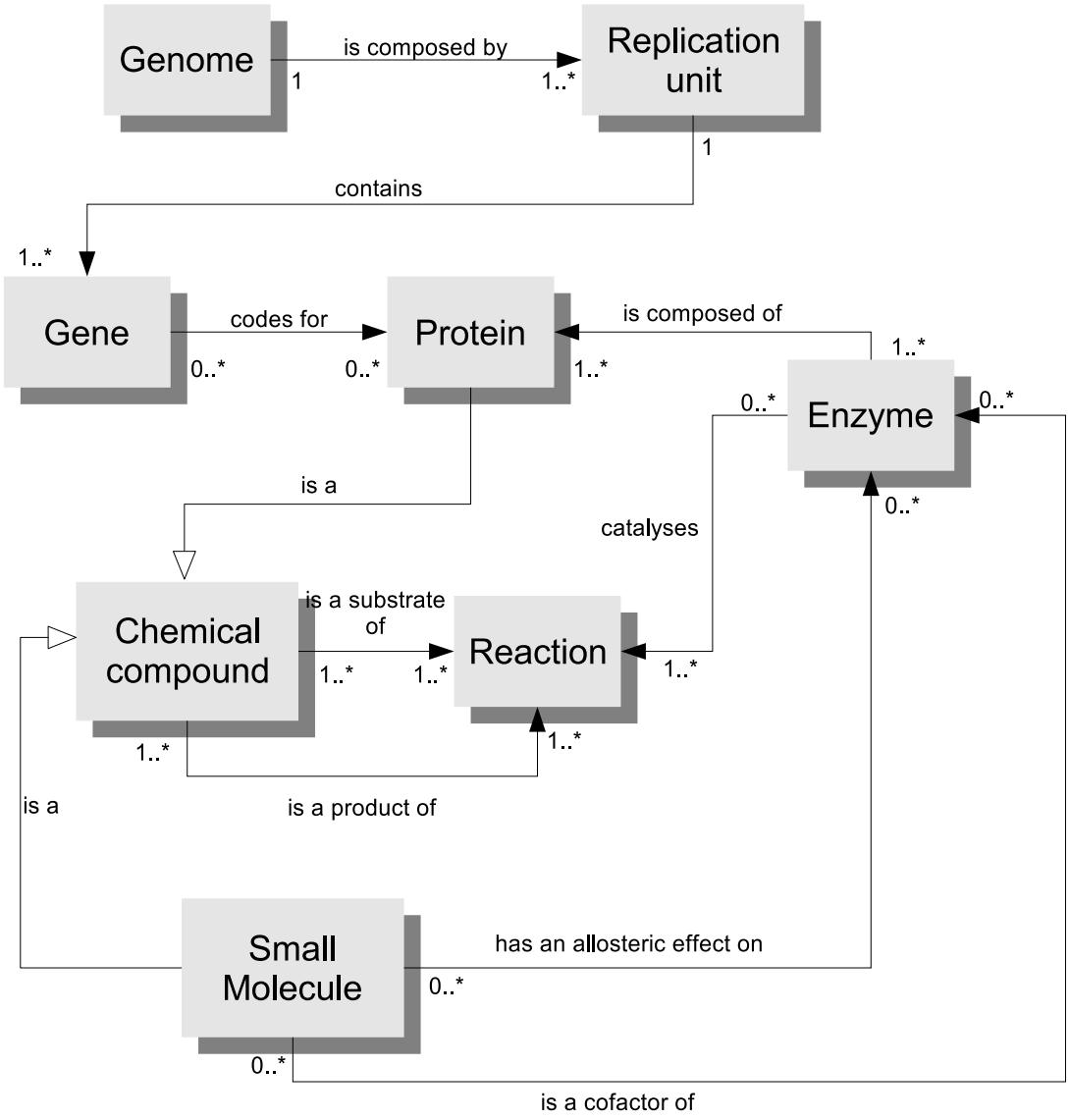

Figure 1.4. Schéma UML (Unified Modeling Language) simplifié des différents objets intervenant dans un réseau métabolique. Les symboles de chaque côté d'une flèche représentant une relation entre les 2 objets indiquent la cardinalité de chaque objet dans la relation : "1" signifie exactement un, "0..*" signifie zéro ou plus, "1..*" signifie un ou plus. Par exemple, les indications de chaque côté de la relation "codes for" entre "Gene" et "Protein" signifient qu'un gène peut produire plusieurs protéines (dans le cas d'épissage alternatif par exemple) ou aucune si le gène n'est pas un gène protéique (ceci explique le "0" du coté de "Protein"), et qu'une protéine peut aussi être produite par plus qu'un gène ou trouvée dans l'environnement (ceci explique le "0" du côté de "Gene").

a. Les composés ou métabolites

Les **composés chimiques**, appelés aussi **métabolites** sont les petites molécules (ensemble d'atomes reliés entre eux) qui sont synthétisées ou dégradées à l'intérieur de l'organisme. Les composés peuvent être importés de l'environnement, auquel cas on les appelle souvent **nutriments**. Ils peuvent également être excrétés. La quantité observée d'un métabolite dépend du tissu ou du compartiment cellulaire dans lequel il est observé. Historiquement, les **composés organiques** sont ceux dont on croyait qu'ils ne pouvaient être synthétisés qu'au sein de la matière vivante par une "force vitale" (voir Section précédente), les **composés inorganiques** représentant tout le reste. Aujourd'hui, la distinction existe encore même si la définition est quelque peu changée. L'atome de carbone étant omniprésent dans la matière vivante, on appelle composés organiques les composés contenant du carbone et synthétisés par la matière vivante. La présence uniforme de carbone dans les métabolites permet de suivre les échanges d'atomes de carbone en les marquant de façon radioactive au sein de l'organisme (Patterson, 1997) ou d'inférer ces échanges de façon automatique (voir Section c.).

On appelle **métabolome** la collection de métabolites contenue dans un organisme. La **métabolomique** identifie et mesure les quantités des différents métabolites dans un échantillon. La **métabonomique** mesure quantitativement les variations dans les concentrations des métabolites en réponse à un *stimulus* ou à une modification génétique. Deux techniques principales permettent de définir le métabolome : la chromatographie couplée à la spectrométrie de masse et la résonnance magnétique nucléaire (Nobeli & Thornton, 2006). Cependant, aucune technique actuelle n'est capable de détecter tous les types de métabolites, il est nécessaire d'utiliser un large éventail de techniques différentes pour obtenir l'ensemble du métabolome (Nobeli & Thornton, 2006). Le traitement automatique des données provenant de ces techniques est d'ailleurs un des défis actuels de la bioinformatique (Nobeli & Thornton, 2006).

b. Les réactions biochimiques

Les **réactions** produisent un ensemble de composés (appelés **produits**) à partir d'un autre ensemble de composés (appelés **substrats** ou **réactifs**). On sait depuis les travaux d'Antoine Lavoisier (1743-1794) qu'une réaction chimique se fait sans variation de masse : "Rien ne se perd, rien ne se crée, tout se transforme". Il n'y a donc pas "apparition d'un produit" mais plutôt transformation d'un substrat en un ou plusieurs produits. Au cours d'une réaction chimique, les métabolites échangent des atomes, ou des groupes d'atomes. Pour décrire ces échanges d'atomes, on écrit une équation chimique qui respecte la conservation des masses. Par exemple, la synthèse de l'acétolactate à partir de deux molécules de pyruvate s'écrit :

$$2C_3H_4O_3 \rightarrow C_5H_8O_4 + CO_2$$

On retrouve la même quantité d'atomes de chaque côté de la réaction. La réaction chimique est une modification des liaisons par déplacements d'électrons : certaines liaisons sont rompues, d'autres sont formées, mais les atomes eux-mêmes sont conservés.

On appelle **stœchiométrie** le calcul des relations quantitatives entre les substrats et les produits d'une réaction chimique. Les **coefficients stœchiométriques** placés devant chaque substrat ou produit dans l'équation d'une réaction chimique indiquent les proportions entre les différents métabolites. Ce sont des nombres sans dimension, l'équation est donc indépendante des quantités de matière des différents métabolites mais permet de les recalculer après la réaction si l'on connaît les quantités initiales.

Les vitesses de réactions sont très variables en fonction des substrats engagés, de la température, de la pression, de la concentration des substrats, du degré de contact entre les réactifs et enfin de la présence d'un catalyseur. Un **catalyseur** est une substance qui augmente la vitesse d'une réaction chimique : il participe à la réaction mais ne fait partie ni des substrats ni des produits, il n'apparaît donc pas dans l'équation de la réaction. Au sein de la cellule, la plupart des réactions, alors appelées **réactions biochimiques**, sont catalysées par des protéines ou des complexes protéiques particuliers, les **enzymes**.

En théorie, une réaction peut toujours fonctionner dans les deux sens. Cependant, dans les conditions physiologiques particulières de la cellule, certaines réactions fonctionnent préférentiellement dans un sens plutôt que dans l'autre. Dans ce cas, les réactions peuvent être considérées comme **irréversibles**.

c. Les enzymes

Les enzymes sont des molécules ou des complexes moléculaires permettant d'accélérer jusqu'à des millions de fois la vitesse d'une ou de plusieurs réactions. Ce sont des catalyseurs, elles agissent à faible concentration et se retrouvent intactes en fin de réaction. Le plus souvent, le constituant de base d'une enzyme est une protéine ou un complexe protéique. Beaucoup d'enzymes sont constituées aussi d'une partie non-protéique appelée **cofacteur**, indispensable à la catalyse (voir Section suivante). La fonction des enzymes est liée à la présence dans leur structure d'un ou plusieurs sites particuliers appelés sites actifs où vont se nicher les substrats ; c'est la proximité des substrats qui va accélérer la réaction (Figure 1.5).

Certains métabolites peuvent accélérer ou ralentir l'activité catalytique de l'enzyme en provoquant un changement dans sa conformation spatiale, ce qui modifie le site de liaison d'au moins un des substrats. Ce phénomène est appelé **allostérie** et l'effet peut être positif (dans ce cas, les composés sont appelés **activateurs**) ou négatif (dans ce cas, les composés sont appelés **inhibiteurs**).

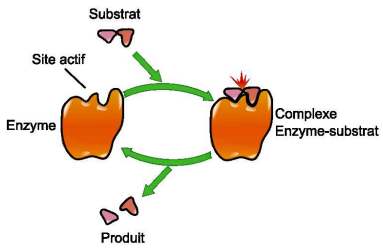

Figure 1.5. L'action d'une enzyme. Source : Wikipedia.

Les noms des enzymes ne sont pas encore standardisés mais contiennent le plus souvent un rappel du substrat ou du produit ainsi que du du type de transformation, suivi du suffixe "-ase" (ex : glucose oxydase). Chaque enzyme est classée par une commission (INTERNATIONAL UNION OF BIOCHEMISTRY AND MOLECULAR BIOLOGY) lui attribuant un code à 4 nombres, appelé **numéro EC** (Tipton, 1994). Pris de gauche à droite, chaque nombre correspond à un niveau de classement toujours plus fin. Les six principaux groupes, correspondant au premier nombre d'un numéro EC, sont :

1. les oxydoréductases,
2. les transférases,
3. les hydrolases,
4. les lyases,
5. les isomérases,
6. les ligases.

Les **oxydoréductases** sont des enzymes catalysant les réactions d'oxydoréduction, c'est-à-dire le transfert de protons et d'électrons.

Les **transférases** catalysent le transfert de fonctions chimiques d'une molécule à une autre de manière spécifique.

Les **hydrolases** sont des enzymes catalysant l'hydrolyse d'une liaison chimique, c'est-à-dire la séparation de molécules d'eau en molécules d'hydrogène et d'ions hydroxide. Ces réaction sont du style : $A-B+H_2O \rightarrow A-OH+B-H$. C'est ce type de réaction qui est utilisé pour dégrader certains polymères (macromolécules constituées de l'enchaînement répété d'un même motif).

Les **lyases** catalysent des cassures de liaisons chimiques autrement que par hydrolyse ou oxydation.

Les **isomérases** catalysent le réarrangement structurel d'isomères (molécules qui ont la même formule mais une forme différente).

Les **ligases** catalysent la jonction de deux molécules.

d. Les cofacteurs

Un cofacteur est une molécule non protéique qui se fixe sur la plupart des enzymes, mis à part les hydrolases. Les cofacteurs favorisent et sont parfois même indispensables à la réaction. On distingue deux types de cofacteurs : les **groupes prosthétiques** qui sont liés de façon permanente à l'enzyme et les **coenzymes** qui sont libérées après la réaction. Parmi les groupes prosthétiques, on trouve essentiellement des ions inorganiques comme Fe^{2+}, Mg^{2+}, Mn^{2+}, ou Zn^{2+}. Ils ne sont pas transformés pendant la réaction et donc n'apparaissent pas dans l'équation.

Les coenzymes sont par contre transformées pendant la réaction et sont donc visibles dans l'équation. On les appelle aussi **cosubstrats**. Les coenzymes sont impliquées dans le transfert de groupes biochimiques et sont continuellement recyclées dans le métabolisme. C'est la raison pour laquelle les coenzymes sont souvent éliminées pour simplifier la modélisation du réseau (voir Section b.). Parmi les coenzymes les plus présentes, on peut citer l'ATP, la coenzyme A, et la nicotinamide adénine dinucléotide (NAD^+).

e. Les voies métaboliques

Une voie métabolique peut être définie comme une séquence de réactions, dans laquelle le produit d'une réaction devient le substrat de l'autre. Certaines voies métaboliques dégradent les nutriments organiques en de simples composés, libérant ainsi de l'énergie et fournissant de nouvelles briques élémentaires pour la synthèse de constituants utiles à la cellule. Ces voies métaboliques forment ce que l'on appelle le **catabolisme**.

D'autres voies métaboliques sont initiées avec des molécules très simples et les convertissent en molécules toujours plus complexes, jusqu'aux protéines et aux acides nucléiques par exemple. C'est ce que l'on appelle l'**anabolisme**. Catabolisme et anabolisme constituent le métabolisme. Les voies métaboliques peuvent être linéaires, branchées ou circulaires.

La mise en évidence et la description de voies métaboliques chez de nombreux organismes est le fruit d'abondantes recherches en biochimie depuis le milieu du XIXème siècle. La généralisation de l'existence d'une voie métabolique à d'autres organismes chez lesquels les mêmes enzymes ont été trouvées est communément effectuée lors de la reconstruction et l'analyse d'un réseau métabolique (voir Section 2.1.3).

1.1.3 Evolution du métabolisme

Même s'ils ont connu plusieurs améliorations et précisions depuis leur énoncé initial (Schmidt *et al.*, 2003), deux modèles principaux tentent aujourd'hui d'expliquer l'évolution du métabolisme.

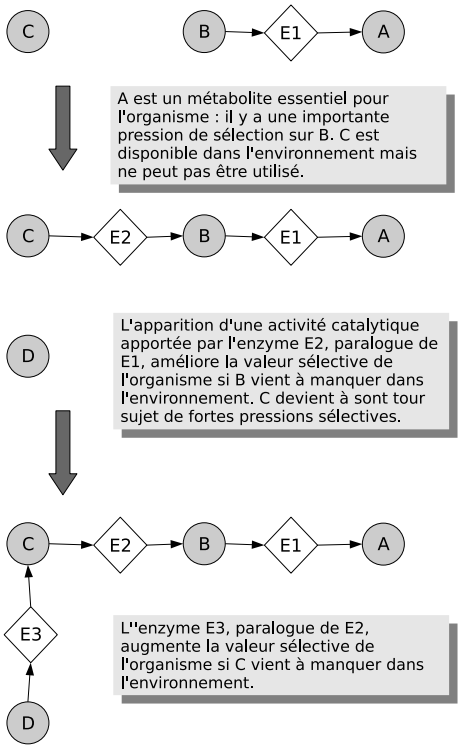

Figure 1.6. Le modèle de l'évolution à rebours. Chaque cercle représente un métabolite et chaque losange une enzyme. Une flèche d'un composé vers une enzyme indique que celle-ci catalyse une réaction prenant le composé comme substrat. Une flèche d'une enzyme vers un composé indique que celui-ci est produit par une réaction catalysée par l'enzyme.

Le premier, proposé par Horowitz (1945), est appelé "évolution à rebours" (en anglais *retrograde evolution*) et s'applique aux voies anaboliques. L'hypothèse est que les organismes qui exploitent des nutriments disponibles dans l'environnement ont un avantage sélectif s'ils deviennent capables de les produire seuls. Si un nutriment vient à manquer, la présence dans un groupe d'organismes d'une enzyme permettant de le synthétiser à partir d'un autre nutriment disponible sera sélectionnée (Figure 1.6).

D'après ce modèle, les enzymes les plus anciennes se trouveraient donc à la fin des voies de synthèse.

Une extension du concept aux voies cataboliques a été proposé par Cordón (1990) dont le modèle prédit la présence des enzymes les plus récentes au début des voies de dégradation.

Le modèle d'Horowitz et l'extension proposée par Cordón reposent tous deux sur l'hypothèse que les métabolites intermédiaires des voies métaboliques ont été disponibles dans l'environnement et transportables au sein de l'organisme, ce qui est en contradiction avec le caractère labile de certains (Jensen, 1976; Rison *et al.*, 2002; Caetano-Anollés *et al.*, 2008).

L'autre modèle, appelé modèle d'évolution "en patchwork" ou "par recrutement", proposé par Jensen (1976), est basé sur le large spectre initial de certaines

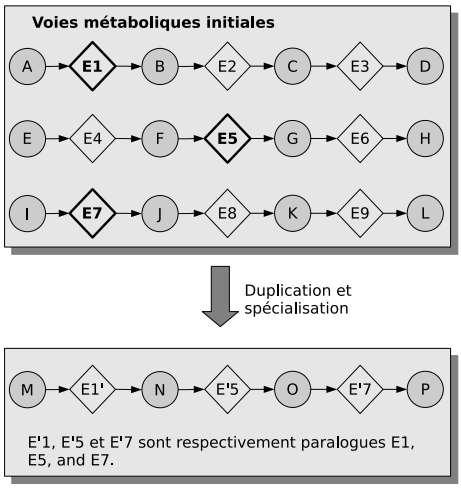

Figure 1.7. Le modèle de l'évolution par recrutement.

enzymes. Le métabolisme initial serait assuré par quelques enzymes très peu spécifiques et non régulées. La duplication des gènes codant pour ces enzymes et la spécialisation de celles-ci permettrait une production plus efficace de certains produits pouvant conférer un avantage sélectif (Figure 1.7). D'après ce modèle, la distribution des enzymes homologues dans le réseau se ferait donc en mosaïque.

D'après plusieurs analyses du réseau métabolique, le métabolisme évoluerait plutôt par recrutement des enzymes que par extension à rebours des voies métaboliques (Rison *et al.*, 2002; Caetano-Anollés *et al.*, 2008).

Cependant, la plupart des études envisagent l'expansion du réseau mais ne s'intéressent pas à la perte de fonctions enzymatiques. L'importance des pertes de fonctions enzymatiques chez les eucaryotes a été étudié (Tanaka *et al.*, 2006) mais à notre connaissance, aucune analyse à l'échelle du réseau n'a été entreprise chez les bactéries intracellulaires. Elle n'est pour l'instant abordée que du point de vue génomique (voir Section suivante).

1.2 Les bactéries endocytobiotes

1.2.1 Définition et découverte de la symbiose

Ce sont des botanistes qui les premiers ont mis en évidence l'importance de la symbiose en biologie. C'est à eux que l'on doit les premières preuves des effets de la symbiose sur la morphologie et la physiologie d'organismes.

L'allemand Albert Frank (1839-1900) utilisa le premier, en 1885, le terme de symbiose (qui signifie "vivre ensemble" en grec) dans un contexte biologique pour décrire l'association entre des champignons et des racines d'arbres : les mycorrhizes (Trappe, 2005). Un an plus tard, un autre allemand, Anton de Bary (1831-1888), découvrit que la structure du lichen émergeait de l'association entre un champignon et une algue. A. de Bary considère déjà les lichens et d'autres exemples de symbiose comme des preuves de l'évolution, venant compléter le gradualisme proposé par Darwin (Sapp, 2004). Même si elle a été sujet de nombreux débats par la suite, la définition de la symbiose selon A. de Bary sera celle que nous adopterons dans ce travail :

"La symbiose est l'association permanente entre deux organismes ou plus d'espèces distinctes, au moins pendant une partie de leur cycle de vie".

La symbiose intracellulaire et le devenir évolutif des endocytobiotes furent envisagés dès 1883 par le botaniste allemand Andreas F. W. Schimper (1856-1901). Ce fut le premier à suggérer un lien évolutif entre des bactéries libres et des organites en notant certaines similitudes entre chloroplastes et cyanobactéries (Sapp, 2004). La théorie selon laquelle l'origine du noyau et des organites cellulaires des eucaryotes serait liée à la symbiose de plusieurs micro-organismes est appelée **symbiogénèse** et a été élaborée par le russe Konstantin Mereschkowski (1855-1921) entre 1905 et 1920.

En 1918, dans son livre *Les symbiotes* très critiqué à l'époque, le français Paul Portier tente de prouver l'aspect fondamental de la symbiose dans le monde vivant. Il propose une origine bactérienne pour la mitochondrie, idée reprise par l'américain Ivan Wallin (1883-1969) qui présenta le phénomène d'héritage de bactéries comme la source de nouveaux gènes et le premier mécanisme de l'origine des espèces (Sapp, 2004).

Il faudra attendre 1967 pour que l'américaine Lynn Margulis (1938-) apporte de nouvelles contributions à la théorie endosymbiotique (Sagan, 1967). La preuve que le matériel génétique du chloroplaste et de la mitochondrie a une origine différente de celle du matériel génétique du noyau vinrent ensuite appuyer ces hypothèses (Chu *et al.*, 2004; Andersson *et al.*, 2003).

Pendant longtemps, la symbiose a été perçue comme un phénomène marginal, une curiosité biologique. Ce n'est que depuis peu que la symbiose est considérée comme un phénomène général ayant un rôle clé dans l'évolution des êtres vivants.

Sapp (2004) explique cette considération tardive par plusieurs facteurs. Premièrement, par sa nature même, l'étude de la symbiose implique la collaboration entre plusieurs disciplines des sciences de la vie qui, surtout à partir du XXème siècle, se sont au contraire hyperspécialisées et se sont confinées dans des instituts aux intérêts différents.

Deuxièmement, il était très difficile de considérer un micro-organisme comme pouvant participer à une association mutualiste à une époque où les microbes étaient considérées comme le principal ennemi de l'homme. Le rôle primordial des microbes dans l'établissement de la vie et dans le fonctionnement des écosystèmes a en effet été longtemps occulté par le pouvoir hautement pathogène de quelques uns d'entre eux. Troisièmement, l'importance de la symbiose héréditaire dans l'évolution des êtres vivants venait à l'encontre de la vision mendelienne de l'hérédité. Quatrièmement, la symbiose en tant que source de changements évolutifs a complètement été mise de côté dans la théorie synthétique de l'évolution construite dans les années 1930-1940, alors que de nombreux cas de symbioses entre organismes pluricellulaires et unicellulaires avaient été mis en évidence. Enfin, la vision de coopération entre organismes a été occultée par la vision classique de conflit, de compétition et de lutte pour la vie énoncée par les premiers évolutionnistes.

C'est la symbiose intracellulaire (ou endocytobiose) des bactéries avec les insectes qui a sans doute été la mieux étudiée, notamment grâce à l'impulsion donnée par le fameux livre "Endosymbiosis of animals with plant microorganisms" (Buchner, 1965). L'endocytobiose est largement répandue chez les insectes, seules quelques familles en semblent dépourvues (Nardon & Charles, 2004). Il a même été mis en évidence chez plusieurs insectes la présence simultanée de plusieurs espèces de bactéries, démontrant ainsi l'existence de relations complexes de complémentarité et de compétition à l'intérieur d'un même individu (Moran *et al.*, 2003a; Wu *et al.*, 2006b; Pérez-Brocal *et al.*, 2006).

1.2.2 Biodiversité de la symbiose

Nous allons définir ici quelques concepts reliés à la symbiose à travers certains exemples, dont la plupart ont été choisis parmi les cas d'endocytobiose.

En général, l'un des partenaires de la symbiose est plus gros que les autres, on l'appelle **hôte**. Les autres partenaires sont appelés **symbiotes**.

Il existe une très grande diversité de cas de symbioses. La distinction se base souvent sur les bénéfices ou les déficits en terme de valeur sélective (fitness) sur l'hôte (Wernegreen, 2005; Gil *et al.*, 2004a). La valeur sélective d'un individu dépend de sa longévité et de sa fertilité (nombre de descendants). Ainsi, on définit les symbiotes comme **mutualistes** quand ils ont un effet positif sur la valeur sélective de leur hôte. Le mutualisme est souvent défini comme une association à bénéfice réciproque. Un exemple connu de mutualisme est celui qui se déroule dans les mycorrhizes où des champignons s'associent aux racines de plantes. Le

champignon se nourrit des glucides synthétisés par la plante lors de la photosynthèse et celle-ci profite de la large surface occupée par les filaments du champignon pour améliorer sa capacité à capter l'eau et les minéraux contenus dans le sol. Mais on peut citer aussi le cas du Bernard-L'hermite et de l'anémone de mer, des bactéries digestives de la vache et de la termite, de l'abeille et des plantes à fleurs... On peut aussi classer les mutualistes selon le degré d'association qu'ils entretiennent avec leur hôte. Ils sont dits **obligatoires** quand la survie des deux partenaires est impossible en dehors de cette association. C'est le cas de l'association entre la bactérie *Buchnera aphidicola* et son hôte, le puceron. Ils sont dits **facultatifs** quand l'hôte et le symbiote peuvent vivre en dehors de l'association.

Au contraire, les **parasites** ont un effet négatif sur la valeur sélective de leur hôte. Parmi ceux-ci, on peut citer par exemple *Plasmodium falciparum*, le protozoaire responsable de la malaria. Le degré du parasitisme est variable selon la durée de l'interaction avec l'hôte : plus la relation est durable, moins les effets du parasite vont être néfastes sur l'hôte. Les organismes les plus nocifs sont plutôt appelés **pathogènes**.

Enfin, les symbiotes **commensalistes** n'ont aucun effet notable sur la valeur sélective de leur hôte. Le terme commensalisme vient du latin *com mensa*, qui signifie "partageant une table". Seul le symbiote tire un bénéfice de l'association (Nair, 2004). Parmi les espèces commensalistes, on peut citer les oiseaux qui nichent dans les trous des arbres.

La distinction peut se faire ensuite sur la localisation de la symbiose (Nardon & Charles, 2004) . Dans une **ectosymbiose** ou **exosymbiose**, les partenaires restent séparés physiquement même s'ils peuvent être très liés comme dans le cas du lichen. Dans le cas d'une **endosymbiose**, les symbiotes appelés **endosymbiotes** sont à l'intérieur de leur hôte mais restent extracellulaires. C'est le cas des bactéries qui composent la flore intestinale humaine. Dans le cas d'une **endocytobiose**, les symbiotes appelés **endocytobiotes** sont intracellulaires. Ils peuvent être temporairement extracellulaires dans le cas de migrations vers les ovaires par exemple. Souvent, les endocytobiotes sont localisés dans des cellules spécialisées appelées **bactériocytes** dans le cas où les endocytobiotes sont des bactéries. Dans les cas d'endocytobiose, on peut considérer également différents niveaux d'intégration de la bactérie. Nardon & Grenier (1993) parlent d'**endocytobiotes intégrés** quand ceux-ci sont toujours présents dans toutes les populations de l'hôte et complètement dépendants de leur hôte pour leur croissance. Ils ne sont pas cultivables *in vitro* et sont obligatoires pour l'hôte. De plus, on observe une congruence très marquée entre les phylogénies de l'hôte et celles des symbiotes, générée par la transmission verticale des symbiotes et la quasi-absence de transferts vers d'autres hôtes (Taylor *et al.*, 2005). C'est le cas par exemple de la bactérie *Buchnera aphidicola* associée au puceron (Baumann *et al.*, 1995). Les **endocytobiotes associés** n'ont pas, par contre, une localisation précise dans l'hôte, n'induisent pas la formation de cellules spécialisées et ne sont pas présents dans toutes les populations. Même dans les populations infectées, elles ne sont

pas forcément présentes dans tous les individus. Ceci indique qu'elles ne sont pas obligatoires pour l'hôte et que leur effet sur la valeur adaptative de l'hôte est faible (Nardon & Grenier, 1993). C'est le cas par exemple des *Wolbachia* parasites de la reproduction.

On peut distinguer les symbioses également selon le mode de transmission du symbiote. Dans le cas d'une **transmission verticale**, la progéniture de l'hôte hérite directement du symbiote. Dans le cas d'une **transmission horizontale**, les symbiotes peuvent passer d'un hôte à l'autre, par relation trophique par exemple. Dans certains cas, la transmission ne semble que verticale, comme pour *Buchnera aphidicola* mais dans d'autres, elle peut être mixte, comme pour les Rickettsies. Ishikawa (2003) considère comme parasites les symbiotes qui sont transmis uniquement de façon horizontale d'un hôte à l'autre.

Les relations entretenues entre symbiotes et hôtes sont très diverses. Certaines stratégies reposent sur une manipulation de la reproduction de l'hôte afin d'assurer une plus grande transmission du symbiote. C'est essentiellement l'oeuvre d'α-protéobactéries du genre *Wolbachia*. La distribution de *Wolbachia* parmi les insectes est estimée à 25-70%. Les *Wolbachia* sont connues pour engendrer, selon les espèces, quatre phénotypes : la mort des mâles (*male killing*), la féminisation, la parthénogénèse (reproduction des femelles sans mâles) et l'incompatibilité cytoplasmique (incapacité des mâles infectés à se reproduire avec une femelle non infectée ou infectée avec une autre souche). Ces 4 types de manipulation ont pour effet la sélection des femelles infectées. En effet, on retrouve les *Wolbachia* dans les oeufs matures mais pas dans le sperme mature : les *Wolbachia* trouvées dans la progéniture proviennent seulement de la femelle. L'effet de mort des mâles a été mis en évidence également chez certaines Rickettsies mais aussi dans des groupes de bactéries beaucoup plus éloignées phylogénétiquement (Hurst & Jiggins, 2000).

D'autres stratégies ont également été sélectionnées et basées sur les échanges trophiques. Les parasites détournent les produits du métabolisme de leur hôte. L'apport des mutualistes à leur hôte peut être nutritionnel : le symbiote fournit des nutriments que son hôte ne peut pas synthétiser. C'est le cas d'un grand nombre de bactéries mutualistes associées aux insectes.

Le caractère mutualiste de certaines bactéries peut s'exprimer également dans la défense de l'hôte contre les parasites. Ainsi, il a été prouvé que la bactérie mutualiste facultative *Hamiltonella defensa* confère une resistance au puceron *Acyrthosiphon pisum* contre le parasitoide *Aphidius ervi* (Oliver *et al.*, 2005). *Photorhabdus luminescens*, une bactérie mutualiste de certains nématodes, contribue à la mort et la digestion des insectes dont se nourrit leur hôte.

Quel que soit le critère utilisé, certains cas de symbioses sont difficiles à classer. La différence de valeur sélective avec ou sans le symbiote est difficile à mesurer. Par exemple, le statut des bactéries que l'on trouve dans la flore intestinale humaine se situe entre le commensalisme et le mutualisme. La localisation des symbiotes n'est pas tout le temps aisée à déterminer. Par exemple, les *Bartonella* ont des stades de vie intracellulaires et d'autres extracellulaires (voir Section 4).

Doit-on les considérer comme de simples endosymbiotes ou comme des endocytobiotes ? Ensuite, il existe très peu de cas de transmission verticale ou horizontale stricte. De plus, nous verrons dans la Section traitant de l'évolution des endocytobiotes qu'il existe un continuum entre le parasitisme et le mutualisme puisque certains parasites peuvent devenir mutualistes au cours de l'évolution.

1.2.3 Histoire des endocytobiotes dans l'évolution

a. Etablissement de la vie intracellulaire

Les ancêtres des endocytobiotes sont des bactéries libres, comme le montrent certaines études phylogénétiques (Gil *et al.*, 2004a). De nombreuses bactéries pathogènes sont capables de pénétrer dans les cellules qu'elles traversent pour coloniser de nouveaux tissus. L'établissement d'une vie intracellulaire durable à l'intérieur de l'hôte implique d'abord une réduction de la pathogénicité de la bactérie. Lorsque une bactérie est capable d'entretenir une relation durable avec l'hôte, on la désigne alors sous le terme de parasite plutôt que sous celui de pathogène.

La défense de l'hôte contre l'intrusion cellulaire est souvent d'isoler la bactérie parasite dans une vacuole, appelée phagosome. Cependant, certains parasites parviennent à quitter le phagosome pour résider dans le cytoplasme. Cette colonisation peut conduire jusqu'à une exploitation complète de la cellule hôte pour la croissance de l'endocytobiote, tel que les Rickettsies qui se reproduisent dans le cytoplasme et le noyau de leurs cellules hôtes.

b. L'influence de la vie intracellulaire sur la réduction du génome

La persistance d'un mode de vie intracellulaire entraîne chez les endocytobiotes une forte réduction de leur génome. Ainsi, parmi les bactéries, on trouve les génomes les plus réduits chez les endocytobiotes.

La réduction du génome apparaît comme une conséquence neutre ou même délétère de l'évolution à long terme sous les conditions imposées par le style de vie des endocytobiotes et ne correspondrait pas une adaptation de l'endocytobiote pour vivre à l'intérieur des cellules de l'hôte (Ochman & Moran, 2001). Chez les endocytobiotes, beaucoup de délétions ne sont pas contre sélectionnées, comme elles le sont chez les bactéries libres. En effet, la stabilité des conditions environnementales et l'absence de compétition inter-spécifiques diminuent la pression de sélection sur les gènes d'adaptation, de résistance et de virulence, dont beaucoup finissent par disparaître. Chez les endocytobiotes les plus intégrés tels que *Buchnera*, certains gènes pourtant généralement considérés comme bénéfiques peuvent aussi disparaître. C'est le cas, par exemple, de gènes codant pour la réparation de l'ADN. La fixations de ces mutations délétères s'expliqueraient par la trans-

mission verticale d'une faible partie de la population d'un hôte femelle vers sa progéniture. La sélection naturelle serait inefficace sur des populations, d'une part fragmentées entre les hôtes, et d'autre part connaissant d'importantes fluctuations numériques (Moran, 1996).

c. Des parasites aux mutualistes

De nombreux indices laissent penser que les ancêtres des bactéries mutualistes étaient parasites. On s'imagine en effet difficilement l'établissement d'une symbiose intracellulaire qui soit directement mutualiste. Il faut voir en effet le mutualisme comme un équilibre plus ou moins stable entre les bénéfices de l'hôte et les bénéfices du symbiote.

La présence d'îlots de pathogénicité (ensemble de gènes à caractère pathogène) conservés chez certaines bactéries mutualistes confirme ainsi le caractère parasite de leurs ancêtres. C'est le cas de *Wigglesworthia glossinidia*, bactérie mutualiste de la mouche *tsé-tsé* (Akman *et al.*, 2002). En outre, des fonctions pathogènes semblent avoir dérivé chez certains mutualistes en fonctions bénéfiques pour l'association entre l'hôte et le symbiote. C'est le cas des *Blochmannia*, mutualistes des fourmis, qui ont gardé l'ensemble de gènes de l'uréase qui code pour des facteurs de virulence chez certains parasites. La conservation de cette fonction permettrait à *Blochmannia* de dégrader l'urée de son hôte en ammoniac utilisé dans la synthèse d'acides aminés (Gil *et al.*, 2003; Degnan *et al.*, 2005). De même, on retrouve dans le génome de *Buchnera aphidicola* une partie des gènes impliqués dans la synthèse des flagelles, cependant insuffisante pour qu'ils gardent leur fonction de mobilité. Ces structures seraient maintenant recyclées dans le transport de métabolites (Maezawa *et al.*, 2006). De même, la surexpression de la protéine GroEL (associée aux conditions de stress de la cellule chez les bactéries libres) chez certains mutualistes d'insectes survient également chez certains parasites. La fonction de cette protéine chez les mutualistes reste incertaine mais elle pourrait aider à la conformation de protéines touchées par les changements d'acides aminés provoquées par les mutations délétères (Fares *et al.*, 2004).

d. Devenir évolutif des endocytobiotes parasites et mutualistes

Le mode de vie intracellulaire, en particulier pour les endocytobiotes les plus intégrés, empêche l'échange de gènes avec d'autres bactéries. La conséquence directe de ces pertes massives de gènes et du manque de transferts horizontaux est que les bactéries endocytobiotes sont incapables de revenir à un style de vie libre. De même, beaucoup de bactéries parasites, respectivement mutualistes, ne peuvent pas devenir mutualistes, respectivement parasites, ce qui expliquerait le caractère uniquement mutualiste ou parasite de groupes phylogénétiques entiers

(Ochman & Moran, 2001).

Chez certains endocytobiotes mutualistes intégrés, ce confinement et la perte irréversible de fonctions métaboliques peut conduire à un remplacement dans sa fonction symbiotique de l'endocytobiote primaire par d'autres endocytobiotes plus récents (Moran *et al.*, 2003b; Pérez-Brocal *et al.*, 2006; Wu *et al.*, 2006b).

Par ailleurs, il est maintenant admis que la réduction et l'intégration du matériel génétique d'endocytobiotes mutualistes peuvent les faire dériver jusqu'à l'état d'organite. Le terme "organite" désigne différentes structures spécialisées contenues dans le cytoplasme des cellules eucaryotes et délimitées par une membrane.

Des analyses phylogénétiques de gènes mitochondriaux ont confirmé que toutes les mitochondries dérivent d'une seule α-protéobactérie (Gray *et al.*, 1999) et que les chloroplastes dériveraient de cyanobactéries. L'apparition des organites dans les cellules eucaryotes est d'ailleurs survenue relativement tôt dans l'histoire de la vie puisqu'elle est estimée à 1,5 milliards d'années (Margulis, 1981). La frontière entre organite et bactérie peut se révéler d'ailleurs très fine. La taille du génome de la bactérie endocytobiote *Carsonella ruddii*, seulement 160 kilobases, pose la question de savoir si on doit la considérer encore comme une bactérie ou comme un organite (Pérez-Brocal *et al.*, 2006). Cependant, d'après Theissen & Martin (2006), la différence majeure entre un organite et un endocytobiote serait que les protéines fonctionelles que l'on trouve dans le cytoplasme des endocytobiotes sont codées par ces derniers et non importées comme dans le cas des organites. Une autre différence est que la plupart des endocytobiotes mutualistes étudiés actuellement sont confinés dans certaines cellules de l'hôte, les bactériocytes, il est donc difficile d'imaginer une possible évolution vers le stade d'organite que l'on trouverait dans toutes les cellules (van der Giezen, 2005).

1.2.4 Evolution du métabolisme des endocytobiotes

La forte réduction du génome des endocytobiotes implique la disparition de voies métaboliques entières. Les voies les plus fréquemment perdues seraient les plus longues et celles qui nécessiteraient le plus d'énergie. La sélection naturelle pourrait également favoriser la perte de voies métaboliques dont le produit final serait devenu présent dans le milieu (Moran, 2007). Chez les espèces parasites, la conservation des voies métaboliques est contrainte par les ressources que la bactérie peut puiser chez son hôte. Chez les espèces mutualistes dont l'association avec l'hôte est d'ordre nutritionnel, la conservation des voies métaboliques chez l'endocytobiote est contrainte également par les métabolites fournis à son hôte. Ainsi, à l'intérieur de l'ordre des Rickettsiales, l'endocytobiote mutualiste *Wolbachia pipientis wBm*, au contraire des *Rickettsiae*, a retenu la capacité de synthétiser la riboflavine et d'autres coenzymes (Foster *et al.*, 2005). La biosynthèse de la riboflavine et celle de l'hème pourraient être les fonctions symbiotiques clés de *Wolbachia pipientis wBm* puisqu'aucune voie de synthèse de ces composés

n'a été détectée à ce jour dans le génome de son hôte.

Chez les endocytobiotes mutualistes les plus intégrés, des voies métaboliques redondantes ont pu être éliminées sans dommage. Grâce à l'existence de gènes orthologues chez *Escherichia coli* pour la quasi totalité des gènes de *Buchnera aphidicola APS*, on a pu ainsi constater que la bactérie a conservé la plupart des voies de biosynthèse des acides aminés essentiels et semble avoir perdu celles des acides aminés non essentiels (que le puceron peut synthétiser) (Shigenobu *et al.*, 2000), il y a donc une réelle complémentarité entre les métabolismes des deux organismes. Une telle complémentarité se retrouve aussi chez des endocytobiotes partageant le même hôte, telles que les bactéries *Baumannia cicadellinicola* et *Sulcia muelleri* (McCutcheon & Moran, 2007).

L'analyse comparative de réseaux métaboliques que nous décrivons dans la Section 4 nous permettra de préciser certains traits de l'évolution du métabolisme des endocytobiotes.

Chapitre 2

La modélisation des réseaux métaboliques

2.1 Reconstruction des réseaux métaboliques à partir des informations génomiques

La reconstruction d'un réseau métabolique consiste à inférer les relations entre gènes, enzymes et réactions qui existent dans un système métabolique donné. Le séquençage complet de nombreux génomes permet maintenant d'avoir une idée générale, quoique souvent imprécise nous le verrons, de l'ensemble des capacités métaboliques d'un organisme.

D'autres types de données peuvent être plus difficiles à obtenir. C'est le cas, par exemple, des effets allostériques (voir Section 1.1.2) qui sont rarement connus. La précision requise dans la définition de chaque lien dépend des questions biologiques que l'on se posera. Pour l'analyse topologique du réseau métabolique, la liste des réactions sera suffisante. Par contre, pour simuler de façon précise le fonctionnement du réseau, les informations cinétiques et les concentrations des composés seront nécessaires. Enfin, pour connaître la relation entre génotype et métabolisme, le lien entre gènes et réactions est indispensable.

Les méthodes de reconstruction d'un réseau métabolique d'un organisme à partir de son génome retournent une liste de réactions considérées comme possibles dans l'organisme. Cependant, bien que le processus ait été largement automatisé ces dernières années, il requiert la plupart du temps l'intervention manuelle d'experts basée sur une recherche intensive dans la littérature ou des preuves expérimentales.

La qualité d'une telle reconstruction dépend non seulement de la qualité de l'annotation du génome utilisé mais également de la position taxonomique de l'organisme considéré. En effet, il existe très peu d'organismes pour lesquels les fonctions associées aux gènes sont bien connues. C'est le cas de quelques organismes modèles comme *Escherichia coli K12*. EcoCyc (Keseler *et al.*, 2005), la

partie de la base de données métaboliques BioCyc (Caspi *et al.*, 2008) dédiée à *E. coli*, offre ainsi un excellent niveau d'expertise avec notamment de nombreuses références expérimentales liées aux données. Ce type de base de données est cependant une exception. La qualité de la reconstruction métabolique d'un organisme dépend ainsi considérablement de sa proximité phylogénétique avec un organisme modèle.

La reconstruction métabolique à partir d'un génome se déroule en trois parties :

1. l'annotation fonctionelle des gènes métaboliques, c'est-à-dire la détermination de l'activité catalytique pour laquelle ils codent,
2. l'établissement de la liste des réactions à partir de la liste des enzymes,
3. l'établissement des voies métaboliques possibles dans l'organisme à partir de la liste de réactions.

2.1.1 Annotation des gènes métaboliques

Les techniques de séquençage à haut débit ont permis de donner accès à de nombreuses séquences de génomes complets. Par exemple, la base de données génomiques de l'EBI (http ://www.ebi.ac.uk/genomes/) contient au moment de la rédaction de ce manuscrit les génomes complets de 54 archées, 680 bactéries et 80 eucaryotes. La plupart de ces génomes ont été annotés en utilisant seulement des méthodes automatiques.

La première étape dans l'annotation du génome consiste à détecter les bornes des gènes sur l'ADN et la seconde à leur assigner une fonction. Plusieurs méthodes complémentaires sont actuellement utilisées pour identifier les limites d'un gène, de la détection de motifs à l'intérieur et autour du gène à la comparaison avec des séquences de gènes déjà connus.

Dans le cas de la reconstruction d'un réseau métabolique, il est particulièrement important d'être capable d'établir les fonctions catalytiques d'une protéine. Les fonctions des protéines correspondant aux gènes détectés sont spécifiées expérimentalement ou de façon automatique. Devant la fréquence de génomes complets nouvellement séquencés, ce sont les méthodes automatiques qui sont de loin le plus souvent employées face à l'impossibilité de fournir des preuves expérimentales pour chaque annotation. Le premier défi de l'annotation automatique des gènes est donc de parvenir à une définition de ce qu'est la "fonction" d'une protéine (Friedberg, 2006).

La définition de fonction biologique est souvent ambiguë et dépend du contexte dans laquelle on l'utilise. Friedberg (2006) donne l'exemple de la kinase. Du point de vue biochimique, la fonction d'une kinase est la phosphorylation du groupe hydroxyle d'un substrat spécifique. Du point de vue physiologique, la même kinase peut intervenir dans une voie de signalisation précise. Enfin, d'un point de vue

médical, une mutation sur cette kinase peut engendrer une maladie. L'annotation automatique de fonctions de protéines doit donc prendre en compte le contexte dans lequel s'incrit la description des protéines et utiliser le vocabulaire approprié.

Dans le cas d'une reconstruction métabolique, le contexte est d'abord biochimique, l'objectif étant de déterminer une ou plusieurs fonctions catalytiques à chaque protéine enzymatique. L'ontologie utilisée est alors communément celle de la classification EC, déjà brièvement décrite dans la Section 1.1.

La manière classique d'annoter une protéine est de la comparer avec l'ensemble des protéines provenant de génomes déjà annotés. L'annotation est basée sur l'hypothèse selon laquelle les enzymes orthologues ont la même activité. Les protéines orthologues sont des protéines homologues (issues d'un ancêtre commun) qui ont divergé après un évènement de spéciation. La difficulté majeure de telles méthodes est de distinguer les protéines orthologues des protéines paralogues, qui sont des protéines homologues provenant d'un évènement de duplication de gène et que l'on considère comme ayant des fonctions différentes. Deux approches existent principalement pour traiter ce problème.

La première est basée sur une comparaison réciproque de toutes les séquences de protéines connues, en utilisant Blast par exemple (Altschul *et al.*, 1990). Deux protéines sont classées dans le même groupe si leurs séquences sont plus proches entre elles que de n'importe quelle autre protéine. Autrement dit, deux protéines $P1$ et $P2$ sont classées dans le même groupe si le meilleur score de comparaison de $P1$ avec toutes les autres séquences correspond à $P2$ et *vice-versa*, ce que l'on désigne classiquement par *Best-Reciprocal-Hits* (BRH). Les méthodes "BRH" sont aujourd'hui communément utilisées pour annoter les nouveaux génomes séquencés car elles sont très rapides et faciles à mettre en place. La première base de données construite avec une méthode similaire, et certainement la plus connue est COG (Clusters of Orthologous Groups) (Tatusov *et al.*, 2000). Le système proposé par KEGG, la classification "Kegg Orthology" (KO) (Kanehisa *et al.*, 2008), est basée sur le même concept et est aujourd'hui largement utilisée dans le contexte d'une reconstruction métabolique. Chaque protéine provenant d'un génome que l'on désire annoter sera ainsi classée dans un groupe d'orthologues, correspondant à une fonction. Dans le cas des gènes métaboliques, cette fonction correspond le plus souvent à un numéro EC. La différence entre orthologues et paralogues est donc ici basée uniquement sur la similarité de séquences. Si deux protéines paralogues coexistent chez un organisme, l'une des deux séquences obtiendra un meilleur score de comparaison pour être classée dans un certain groupe et l'autre sera considérée comme son paralogue. Cependant, ces méthodes ne prennent pas en compte la disparition des paralogues au cours de l'évolution. Imaginons deux protéines paralogues $P1$ et $P1'$, issue de $P1$ après duplication. $P1$ est réellement orthologue des protéines contenues dans un groupe d'orthologie $C1$. Si, pour des raisons évolutives, le gène correspondant à $P1$ disparaît du génome de l'espèce considérée, il se peut très bien que $P1'$ soit classée dans le groupe $C1$ alors qu'une

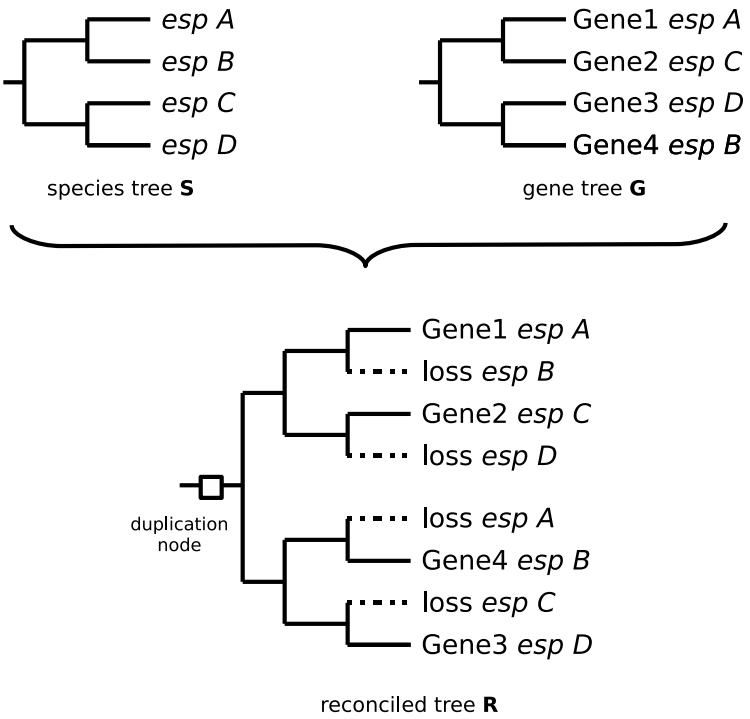

Figure 2.1. Réconciliation de l'arbre des gènes G et de l'arbre des espèces S dans l'arbre de réconciliation R. R est une variation de S, dans lequel un noeud "duplication" a été inséré pour corriger l'incohérence entre les deux arbres G et S (Dufayard *et al.*, 2005).

relation de paralogie et non d'orthologie relie ces protéines.

La seconde approche pour distinguer les paralogues des orthologues essaie de prendre en compte ces possibles disparitions de gènes paralogues. Elle est basée sur la comparaison d'un arbre de gènes, G, avec un arbre des espèces qui fait office de référence, S. La comparaison de ces deux arbres met en relief les pertes de gènes paralogues dans un arbre de "réconciliation" R (Duret *et al.*, 1994) (Figure 2.1). Des méthodes automatiques de réconciliation d'arbres ont été développées pour des annotations à l'échelle du génome mais nécessitent des arbres de gènes et d'espèces exacts. Pour contourner ce problème, Dufayard *et al.* (2005) proposent une méthode autorisant la présence de noeuds non résolus autant dans l'arbre des espèces que dans l'arbre des gènes.

Plusieurs bases de données ont été construites en utilisant une méthode de réconciliation d'arbres, parmi lesquelles on peut citer HOVERGEN (Duret *et al.*, 1994) pour les vertébrés, INVHOGEN (Paulsen & von Haeseler, 2006) pour les invertébrés et TreeFam (Li *et al.*, 2006) pour les animaux.

Cependant, les méthodes de réconciliation d'arbres présentent certaines limitations (Kriventseva *et al.*, 2007) : elles sont basées sur des modèles de duplication et de perte de gènes qui restent mal définis, l'arbre d'espèces utilisé (qui suit le

plus souvent la taxonomie proposée par le NCBI) contient toujours un nombre élevé de noeuds non résolus et la construction de l'arbre de réconciliation est très coûteuse en temps.

Quel que soit le type de méthode choisi, toutes les approches basées sur l'orthologie reposent sur l'hypothèse que les protéines orthologues partagent la même fonction alors que les paralogues ont des fonctions différentes. Cependant, des paralogues issus d'une duplication récente peuvent avoir des fonctions plus proches que des orthologues phylogénétiquement distants. Des protéines orthologues peuvent ainsi présenter des fonctions considérablement différentes, même dans le cas d'une forte similarité de séquences (Friedberg, 2006; Naumoff *et al.*, 2004).

Plutôt que de déterminer la fonction d'une protéine en se basant sur la similarité de séquences entières, d'autres méthodes vont tenter d'identifier dans les séquences protéiques des signatures que l'on peut relier à certaines fonctions. En effet, la fonction d'une protéine est essentiellement déterminée par un ou plusieurs domaines actifs, et non par toute sa séquence. Plusieurs bases de données, comme InterPro (Mulder & Apweiler, 2007) et ProDom (Servant *et al.*, 2002) proposent une classification automatique des domaines homologues qui peut être utilisée pour établir les fonctions d'une protéine. De façon plus spécifique, Claudel-Renard *et al.* (2003) ont développé une méthode appelée PRIAM qui assigne des numéros EC en se fondant sur une classification basée sur une décomposition en modules des enzymes contenues dans la base de données ENZYME (Bairoch, 2000). Un module est défini comme le plus long segment homologue partagé par une collection d'enzymes. Cette décomposition en modules permet de dégager des règles logiques sur la présence simultanée ("et") ou non ("ou") de ces modules dans une collection d'enzymes.

Malheureusement, l'augmentation du nombre et de la diversité des séquences disponibles dans les bases de données multiplie le nombre d'erreurs occasionées par les méthodes basées sur l'homologie. Le principal problème est que les bases de référence sont complétées avec des séquences déjà annotées de façon automatique, ce qui provoque une propagation toujours plus importante des erreurs d'annotation (Friedberg, 2006). En effet, un nouveau gène peut être associé à l'annotation d'une séquence pour laquelle aucune preuve expérimentale n'existe. Par ailleurs, certaines protéines, dites analogues, ont une similarité de séquence très faible mais partagent en fait la même fonction (Galperin *et al.*, 1998).

D'autres méthodes permettent d'inférer la fonction de protéines pour lesquelles l'homologie de séquences n'a rien donné. La première s'appuie sur le phénomène de co-évolution des protéines. L'hypothèse dans ce cas est que les protéines dont les fonctions sont liées évoluent de manière corrélée. Si N génomes sont considérés, on établit pour chaque protéine un profil phylogénétique qui correspond à un vecteur de booléens de longueur N, chaque booléen indicant la

présence ou l'absence d'un orthologue dans un génome donné. Les protéines sont alors classées et leurs fonctions déterminées selon la similarité de leurs profils phylogénétiques (Pellegrini *et al.*, 1999) (Figure 2.2).

Chez les procaryotes, les gènes co-régulés ont tendance à être proches sur le chromosome. La conservation de la co-localisation de gènes au cours de l'évolution est connue sous le nom de synténie (Figure 2.3). L'ordre des gènes dans les groupes de synténie est très souvent conservé au cours de l'évolution. Cette information peut être utilisée pour inférer la fonction d'une protéine dont le gène se trouve dans un tel groupe, même en l'absence de similarité détectable avec d'autres séquences (Rogozin *et al.*, 2002; Vallenet *et al.*, 2006).

La fusion de deux gènes au cours de l'évolution peut être également utilisé pour relier fonctionnellement deux gènes qui seraient homologues (Yanai *et al.*, 2001; Enright & Ouzounis, 2001). Enfin, les profils d'expression obtenus grâce aux puces à ADN, les réseaux protéines-protéines et les informations sur la localisation cellulaire des protéines sont autant de sources pouvant indiquer des relations fonctionnelles entre protéines (Friedberg, 2006). Malheureusement, ces méthodes permettent seulement d'assigner une fonction très générale, le plus souvent au niveau cellulaire, mais ne permettent pas d'assigner précisément une activité enzymatique.

Pour limiter les erreurs dues à l'utilisation d'une seule approche, certains tentent de combiner plusieurs approches de manière efficace. Par exemple, Chua *et al.* (2007) intègrent des données provenant de plusieurs sources (homologie de séquences, interactions protéines-protéines, domaines protéiques, expression et littérature) dans un graphe pondéré. Certaines approches développées pour combler les parties manquantes des voies métaboliques tentent d'intégrer différents types de données pour identifier les gènes correspondant aux enzymes absentes (Green & Karp, 2004; Yamanishi *et al.*, 2007).

Afin d'obtenir une annotation la meilleure possible, une phase d'expertise manuelle est essentielle. Plusieurs plate-formes d'annotation, telles que GenDB (Meyer *et al.*, 2003), MaGe (Vallenet *et al.*, 2006) ou Iogma[1], fournissent de puissantes interfaces graphiques pour aider les experts à nettoyer ou à compléter les annotations générées.

Le partage des tâches devient également une clé dans l'annotation des génomes à venir. En effet, comme il est difficile d'avoir un groupe d'experts pour annoter tous les types de gènes dans un seul génome, Overbeek *et al.* (2005) proposent une approche où tous les gènes impliqués dans un "sous-système" (comme une voie métabolique par exemple) seraient analysés chez plusieurs génomes par un groupe expert de ce sous-système.

De même, il est important que les génomes anciennement annotés puissent profiter des nouvelles données disponibles actuellement. Plusieurs projets fournissent des annotations actualisées pour plusieurs génomes. C'est le cas, par

[1] `http://www.genostar.org`

exemple, de RefSeq (Pruitt *et al.*, 2007), Ensembl (Hubbard *et al.*, 2007) et HAMAP ("High-quality Automated and Manual Annotation of microbial Proteomes") (Lima *et al.*, 2008). Cette dernière base de données intègre des annotations à la fois manuelles et semi-automatiques complétant les annotations des protéomes bactériens. Enfin, le projet Integr8 (Kersey *et al.*, 2005) réunit les informations provenant de GenomeReviews et d'HAMAP.

Finissons cette section sur une note peu optimiste. En effet, malgré l'avènement de nombreuses méthodes automatiques, le partage des tâches et l'échange des connaissances, la fraction de gènes pour lesquels aucune function n'a été attribuée reste très importante (environ 40 %), spécialement chez les eucaryotes, et demeure un problème majeur (Gerlt & Babbitt, 2000; Pouliot & Karp, 2007). Par ailleurs, l'assignation d'un numéro EC, même quand il est correct, peut être assez imprécis. Par exemple, le numéro EC 2.7.1.69 correspond à 60 gènes dans le génome de *Lactobacillus plantarum* (Teusink *et al.*, 2005). De plus, 30 à 40% des activités enzymatiques expérimentalement décrites et possédant un numéro EC restent "orphelines", c'est-à-dire qu'aucun gène ne leur a été assigné (Chen & Vitkup, 2007; Lespinet & Labedan, 2006; Pouliot & Karp, 2007).

Il est donc important de garder à l'esprit cette observation lors de la reconstruction du réseau métabolique d'un organisme à partir de son génome, le réseau métabolique ainsi obtenu ne sera qu'une image partielle et relativement imprécise du réseau réel.

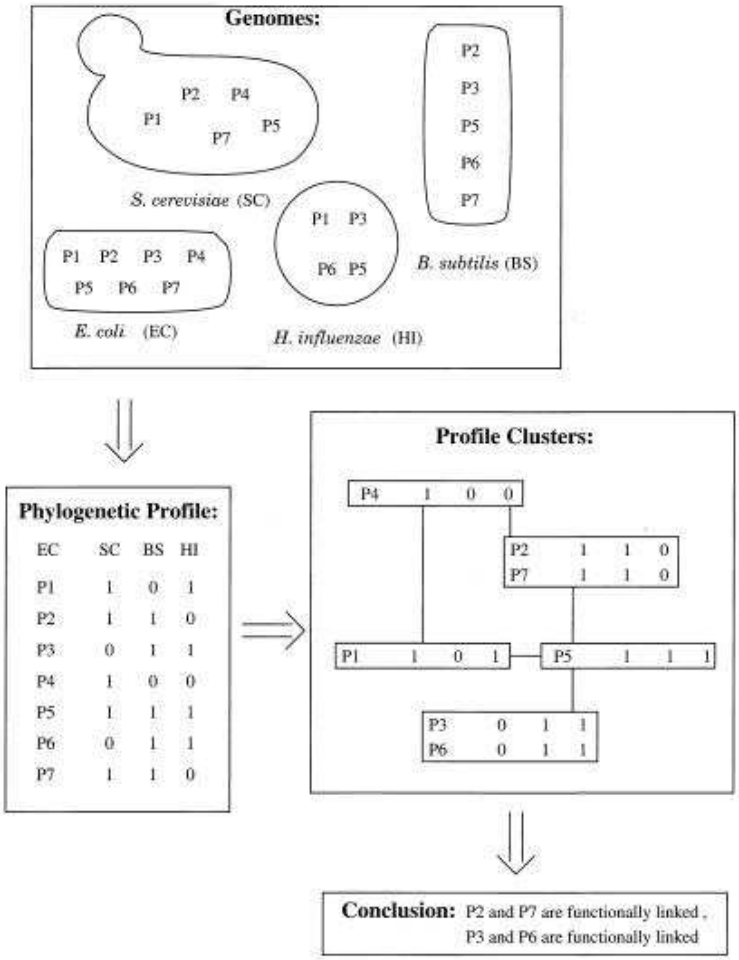

Figure 2.2. Exemple de construction de profils phylogénétiques et d'établissement de liens fonctionnels entre protéines. Dans cet exemple, sept protéines sont considérées à travers quatre génomes d'espèces différentes (*Escherichia coli*, *Saccharomyces cerevisiae*, *Haemophilus influenzae* et *Bacillus subtilis*). Pour chaque protéine d'*E. coli*, est construit un profil quels génomes codent pour des homologues de la protéine. Ensuite, on détermine quelles protéines partagent le même profil. Les protéines avec un profil identique ou similaire sont indiquées comme fonctionnement liées. Exemple tiré de l'article de Pellegrini *et al.* (1999).

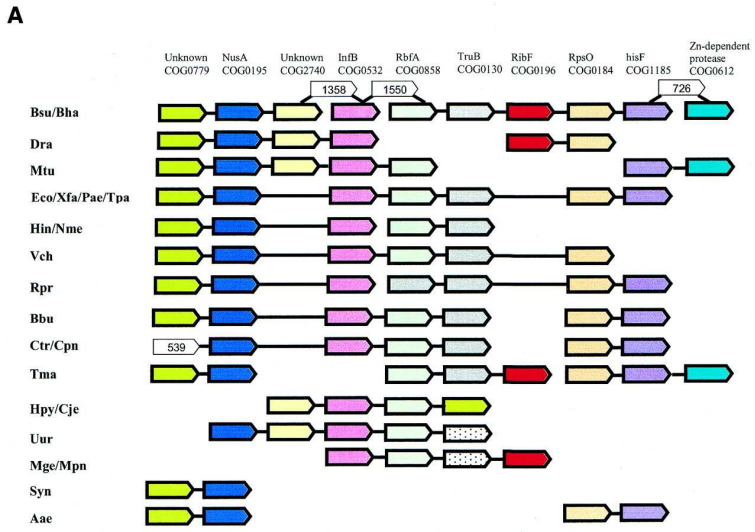

Figure 2.3. Conservation de la synténie à travers diverses espèces de bactéries. Exemple donné par Rogozin *et al.* (2002).

2.1.2 Définition de la liste des réactions métaboliques

Une fois que les fonctions métaboliques des enzymes ont été assignées, les liens entre enzymes et réactions peuvent être définis. La base de données ENZYME décrit chaque enzyme pour lesquelles un numéro EC a été attribué et contient de nombreux liens vers d'autres bases de données métaboliques et la base de données protéiques Uniprot (Wu *et al.*, 2006a). Uniprot réunit les informations protéiques provenant de deux bases : Swiss-Prot (Boutet *et al.*, 2007) qui contient les séquences de protéines décrites expérimentalement, TrEMBL qui contient les séquences de protéines inférées à partir des gènes contenus dans EMBL. La base de données BRENDA fournit des détails supplémentaires, quand ils sont disponibles, sur la spécificité des enzymes pour leurs substrats en fonction des organismes (Schomburg *et al.*, 2004).

Les bases de données BioCyc (Caspi *et al.*, 2008) et KEGG (Kanehisa *et al.*, 2008), certainement celles qui sont les plus largement utilisées actuellement, dispensent des données à la fois génomiques, biochimiques et métaboliques. Chacune de ces deux bases a son outil de reconstruction associé : KAAS (Kegg Automatic Annotation Server) (Moriya *et al.*, 2007) pour KEGG et Pathologic (Karp *et al.*, 2002) pour BioCyc. Ces deux systèmes emploient des stratégies différentes. Dans KEGG, les gènes présents sont identifiés grâce à leur numéro KO qui les relie à une fonction, désignée par un numéro EC dans le cas des gènes métaboliques (voir la section précédente). Une réaction biochimique est considérée comme possible dans un organisme si son génome contient un gène classé par comparaison de séquence dans le groupe KO correspondant au numéro EC de la réaction.

Pathologic n'effectue pas de comparaison de séquences et suppose l'annotation déjà effectuée. Si un numéro EC a été assigné par les annotateurs, la réaction correspondante dans la base de référence utilisée par Pathologic, MetaCyc (Caspi & Karp, 2007), est ajoutée à la liste des réactions possibles dans l'organisme. Si un gène n'a pas de numéro EC assigné, les noms et synonymes des fonctions qu'on lui a assignés sont comparées à la liste des noms d'activités enzymatiques contenues dans MetaCyc. L'efficacité de la stratégie utilisée par Pathologic dépendra ainsi directement de la qualité des annotations génomiques fournies, au contraire de KAAS, qui supporte les génomes sans aucune annotation. En revanche, Pathologic peut bénéficier des annotations supplémentaires ou de corrections effectuées manuellement après l'annotation originale.

Une autre différence entre les deux systèmes est la base de référence utilisée. KAAS utilise les données fournies par la classification KO basée uniquement sur l'homologie de séquences. Pathologic, de son côté, compare les numéros EC ou les noms d'enzymes à la base de référence MetaCyc, qui contient les informations métaboliques de plus de 900 organismes pour lesquelles une preuve expérimentale existe.

2.1.3 Définition des voies métaboliques possibles dans un organisme

Morowitz (1999) considère l'ensemble des voies métaboliques comme "*une vaste généralisation empirique basée sur un siècle et demi de labeur effectué par une armée de biochimistes qui ont travaillé à la caractérisation de toutes les réactions biochimiques ayant lieu dans les cellules vivantes*". En effet, l'inférence des voies métaboliques repose sur la quasi-universalité de celles-ci. Ainsi, au cours du temps, il a été établi expérimentalement chez certains organismes modèles des voies métaboliques qui servent de référence ensuite pour l'établissement des voies métaboliques d'autres organismes. Certaines voies, comme la glycolyse ou le cycle de Krebs, se retrouvent dans la majorité des organismes, même phylogénétiquement éloignés. D'autres, comme les voies reliées à la photosynthèse, ne se trouvent que dans des groupes plus restreints d'organismes. En fonction des organismes ou des environnements, certaines voies peuvent également connaître des variantes.

Dans KEGG, il est possible de repérer sur les cartes métaboliques de référence des gènes ou des réactions à partir de leurs identifiants (voir Section 2.3.1).

Pathologic tente en plus de prédire quelles sont les voies susceptibles de se dérouler dans l'organisme considéré. La prédiction est basée essentiellement sur la proportion de réactions inférées chez l'organisme et se déroulant dans cette voie, sur leur position dans la chaîne de réactions selon que la voie est anabolique ou catabolique, et enfin sur la présence de réactions inférées que l'on trouve seulement dans cette voie (Paley & Karp, 2002).

Cependant, ni KAAS ni Pathologic ne sont capables d'inférer de nouvelles voies métaboliques, c'est-à-dire ne correspondant à aucune autre. Pour ceci, l'outil "Pathway Hunter Tool" (PHT) peut être utilisé (Rahman *et al.*, 2005). À partir d'un ensemble de numéros EC, d'un métabolite source et d'un métabolite destination sélectionnés par l'utilisateur, le PHT calcule toutes les voies métaboliques les plus courtes. Ceci peut aider à proposer des voies alternatives dont la pertinence peut ensuite être testée expérimentalement.

La reconstruction métabolique à partir d'un génome permet très rapidement d'obtenir un aperçu des capacités métaboliques d'un organisme, et, par exemple, d'étudier l'impact de certains évènements génomiques (comme les duplications, les transferts horizontaux, les pertes de gènes) sur un réseau métabolique. Cependant, cette ébauche de reconstruction contient souvent des erreurs ou des imprécisions qui doivent être ensuite nettoyées et complétées manuellement ou à partir de méthodes plus précises décrites ci-dessous.

2.1.4 Raffinements des méthodes de reconstruction métabolique

a. Les gènes et réactions manquants

Après la reconstruction d'un réseau métabolique, certaines réactions apparaissent comme "manquantes". Elles correspondent par exemple à des trous dans les voies métaboliques qui ont été détectées. Ces réactions manquantes peuvent s'expliquer par (Cordwell, 1999; Osterman & Overbeek, 2003) :

- une similarité de séquence faible du gène correspondant avec ceux connus pour coder l'enzyme manquante dans d'autres organismes,
- le fait que les produits de la réaction peuvent être obtenus à partir de voies alternatives ou sont apportés par l'environnement,
- le fait qu'une autre enzyme présente dans l'organisme est capable de catalyser cette réaction.

De multiples approches existent pour compléter ces trous en essayant d'identifier les gènes capables de coder ces fonctions. Ici encore, elles sont essentiellement basées sur des heuristiques. Divers indices génomiques peuvent être combinés pour proposer des gènes candidats à une réaction manquante. L'objectif est ici inverse de l'annotation classique d'un gène comme nous l'avons vu précédemment. Au lieu d'assigner une fonction à un gène inconnu, nous essayons d'assigner une séquence à une fonction. Les méthodes sont donc sensiblement différentes mais les hypothèses biologiques utilisées sont les mêmes. Ainsi, Green & Karp (2004); Gerlt (2003); Kharchenko *et al.* (2006) utilisent entre autres l'hypothèse selon laquelle les gènes codant pour des enzymes intervenant dans la même voie métabolique sont co-localisés sur le génome. Kharchenko *et al.* (2006) proposent d'utiliser en plus les informations de co-expression et de fusion des gènes. L'association de différentes méthodes afin d'inférer les gènes manquants peut être réalisée grâce à une approche supervisée (comme les machines à vecteurs de support (SVM en anglais)) qui nécessite une connaissance partielle du réseau et un ensemble d'apprentissage de qualité (Green & Karp, 2004; Yamanishi *et al.*, 2007).

b. La réversibilité des réactions

La direction d'une réaction dans certaines conditions physiologiques est déterminée par ses propriétés thermodynamiques, les propriétés cinétiques de l'enzyme et la concentration des substrats et des produits. Dans une reconstruction métabolique automatique, la direction des réactions est souvent absente et doit être ajoutée manuellement. La plupart des modélisations prennent en compte les directions des réactions telles qu'elles apparaissent dans les voies métaboliques : si une réaction apparaît toujours dans le même sens quelle que soit la voie métabolique dans laquelle elle intervient, alors on lui assigne cette direction et on la définit comme irréversible.

Cependant, cette manière d'assigner la direction des réactions n'est pas complètement satisfaisante. En effet, chez certains organismes, on peut facilement imaginer qu'une réaction, normalement irréversible d'après les voies métaboliques, puisse se produire dans les deux sens ou même dans le sens contraire, à cause de conditions physiologiques très différentes chez cet organisme et chez ceux pour lesquels la voie métabolique a été définie.

D'autres méthodes tentent de prédire les directions des réactions à partir d'informations contenues dans le réseau métabolique lui-même. Ainsi, Yang *et al.* (2005) montrent comment la direction d'une réaction peut être déterminée en analysant la matrice stœchiométrique d'un réseau métabolique. Les directions possibles des réactions sont calculées grâce à celles imposées aux réactions se situant aux limites du système (comme les réactions de transport) et éventuellement de quelques autres réactions dont on connaît déjà la direction. Cependant, Yang *et al.* ont testé leur méthode sur un réseau contenant seulement 44 réactions. L'algorithme proposé par Kümmel *et al.* (2006) exploite les mesures des énergies de formation (énergies de Gibbs) des métabolites qui, si elles ne sont pas connues exactement, peuvent être estimées à partir de la structure des métabolites, et des concentrations des métabolites si elles sont disponibles. Ensuite, en s'appuyant sur un ensemble d'heuristiques basées sur des règles biochimiques, Kümmel *et al.* identifient les parties du réseau qui sont thermodynamiquement faisables. L'algorithme a été testé sur un réseau métabolique de *Escherichia coli K12* basé sur le génome entier qui compte 920 réactions dont 130 ont été assignées comme irréversibles. Grâce à une méthode similaire, Feist *et al.* (2007) propose une reconstruction métabolique pour *Escherichia coli K12* qui inclut les sens des réactions.

c. Utilisation des métabolites

Certaines méthodes ont été développées pour proposer une liste de réactions possibles à partir d'un ensemble de métabolites. Arita (2000) s'appuie sur 16 types de liens hypothétiques entre métabolites pour inférer les réactions biochimiques possibles à partir d'un ensemble de composés, même si elles ne correspondent à aucun numéro EC. De la même manière, Kotera *et al.* (2004) proposent une méthode capable d'assigner des numéros EC partiels (auxquels il manque le dernier chiffre) à partir d'un ensemble de substrats et de produits.

d. Utilisation d'évidences expérimentales

Des techniques à grand débit ont récemment été développées pour déterminer le métabolome d'un organisme (voir Section 1.1). Les méthodes décrites ci-dessus peuvent ainsi s'appliquer au catalogue de métabolites produit par ces techniques. Breitling *et al.* (2006) utilisent les résultats de spectromètres de masse à ultra-haute résolution et le fait qu'un répertoire limité de transformations intervient

dans les réactions chimiques. Leur méthode est capable d'inférer les transformations chimiques possibles en calculant les différences de masses entre tous les composés et en les comparant à une table de référence donnant la correspondance entre les différences de masses et les transformations chimiques.

D'autres techniques à haut-débit peuvent compléter ou affiner les reconstructions automatiques de réseaux métaboliques. La spectrométrie de masse pour l'identification à grande échelle des protéines dans un organisme permet de confirmer la présence d'enzymes prédites automatiquement (VerBerkmoes *et al.*, 2004; Wagner *et al.*, 2002). De même, l'isolement et la purification d'une enzyme et la définition de ses activités catalytiques précisent sa spécificité et son mode d'action. Les études de phénotype et d'expression de gènes à grande échelle peuvent être également utilisées conjointement avec les résultats de simulations de fonctionnement de réseau dans le but d'affiner un réseau métabolique (Covert *et al.*, 2004). À une plus petite échelle, les informations physiologiques peuvent fournir d'importantes pistes supplémentaires pour compléter l'ensemble des réactions d'un réseau métabolique. Par exemple, dans le réseau métabolique de *Streptomyces coelicor* reconstruit par Borodina *et al.* (2005), 89 % des réactions ont un gène annoté associé tandis que le reste des réactions a été inclus en se basant uniquement sur les connaissances physiologiques de la bactérie.

Enfin, la modélisation du réseau elle-même permet d'affiner la qualité de celui-ci. Lors de la vérification d'un réseau métabolique, les prédictions faites par le modèle sont comparées aux observations expérimentales. En cas de contradiction, le modèle est corrigé et l'opération répétée jusqu'à obtenir un réseau cohérent (Borodina *et al.*, 2005; Duarte *et al.*, 2007). De même, les résultats obtenus par l'analyse du modèle permettent de vérifier ou de tester de nouvelles hypothèses biologiques et ainsi d'améliorer les connaissances du réseau métabolique de l'organisme considéré.

2.2 Modélisation des réseaux métaboliques

2.2.1 Les modèles à base d'équations différentielles

Le modèle par défaut pour représenter une réaction biochimique est basé sur la loi de cinétique de Michaelis-Menten et ses généralisations à plusieurs substrats, aux réactions réversibles et aux différents mécanismes d'inhibition (Garfinkel, 1968). Les paramètres de ces lois cinétiques sont la concentration de l'enzyme, des substrats, des produits et des éventuels activateurs ou inhibiteurs, et une constante de cinétique propre à l'enzyme. Cependant, l'utilisation de ces lois pour modéliser un ensemble de réactions ou un réseau métabolique complet pose certains problèmes (Voit, 2002). D'abord, ces lois n'ont été vérifiées qu'*in vitro* et reposent sur l'hypothèse d'un milieu homogène. Or, la cellule diffère beaucoup

d'un milieu homogène : les métabolites forment des aggrégats et les organites compartimentent la cellule. Ensuite, ces lois ne s'appliquent que sur des enzymes seules alors qu'on sait que les enzymes fonctionnent de manière coordonnée dans la cellule. Enfin, la formulation mathématique de ces lois devient vite extrêmement complexe dès que plusieurs réactions interviennent.

La théorie des systèmes biochimiques (*Biochemical Systems Theory (BST)*) apporte un formalisme mathématique robuste à l'étude dynamique des systèmes métaboliques. Les variables sont classiquement les concentrations des enzymes, des substrats et des modulateurs (activateurs ou inhibiteurs) mais peuvent inclure également les conditions physiologiques telles que le pH ou la température. Une des caractéristiques de la BST est que chaque variation de concentration est approximée par une loi de puissance, approximation observée biologiquement (Voit, 2002).

La BST inclut également la théorie du contrôle métabolique qui relie les propriétés globales d'un système métabolique aux propriétés de ces composants, en particulier les enzymes. Par exemple, le coefficient de contrôle de flux mesure l'effet de la variation de la concentration d'une enzyme sur le flux traversant une voie métabolique (Fell, 1992).

Ces formalismes ont été utilisés avec succès essentiellement dans l'analyse des effets de médicaments. Cependant, ils n'ont été appliqués pour le moment qu'à des sous-parties du réseau, typiquement des voies métaboliques. En effet, ils impliquent une connaissance fine des concentrations des enzymes et des métabolites, ainsi que des paramètres cinétiques des enzymes. Or, ces paramètres ne sont connus que pour des enzymes intervenant dans des voies métaboliques et chez des organismes bien connus. Ces informations n'étant pas disponibles dans les reconstructions métaboliques effectuées à partir des informations génomiques, les formalismes à base d'équations différentielles ne sont pour l'instant pas applicables pour modéliser les réseaux métaboliques à grande échelle.

2.2.2 Les modèles à base de contraintes

En l'absence d'informations détaillées pour l'ensemble du réseau métabolique, on peut se baser sur le fait qu'en réalité les cellules sont soumises à certaines contraintes qui limitent leurs comportements possibles. En imposant ces contraintes dans la modélisation, on peut déterminer ce qui est possible ou non dans une cellule. Les modèles à base de contraintes, ainsi nommés par Palsson (2000), analysent la distribution des flux possibles dans un réseau métabolique. Un flux peut être considéré comme un vecteur v de vitesses, chaque élément du vecteur correspondant à la vitesse relative d'une réaction du système. Si un réseau métabolique contient n réactions, un vecteur flux sera toujours de longueur n. La valeur dans le vecteur flux correspondant à une réaction réversible pourra être positive ou négative selon la direction de la réaction.

La première contrainte appliquée au réseau est la stœchiométrie de chaque

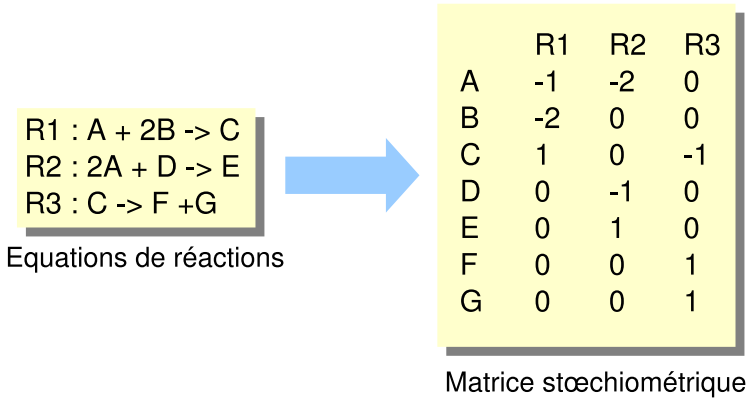

Figure 2.4. Matrice stœchiométrique

réaction. Un réseau métabolique de n réactions et m composés sera donc modélisé sous la forme d'une matrice S d'entiers à n colonnes et m lignes, où chaque valeur correspond au coefficient stœchiométrique d'un métabolite dans une réaction, positif si le métabolite est un produit et négatif si le métabolite est un substrat de la réaction (voir Figure 2.4).

La seconde contrainte toujours présente dans ces modèles est l'hypothèse selon laquelle le réseau est en état d'équilibre dynamique, c'est-à-dire que chaque quantité de métabolite produite est consommée. Cette hypothèse est très importante pour définir les limites du système que l'on étudie. Un flux qui vérifie l'état d'équilibre du système est aussi appelé "mode".

Un flux v sera donc conditionné par l'équation suivante :

$$Sv = 0$$

La troisième contrainte concerne les réactions irréversibles : on contraint les valeurs correspondantes aux réactions irréversibles à être positives ou nulles dans un vecteur flux. Pour toute réaction i irréversible, on aura donc :

$$v_i \geqslant 0$$

D'autres contraintes pourront limiter les flux à travers certaines réactions, selon les capacités des enzymes ou la concentration des métabolites.

La distribution des flux peut être représentée dans un espace à n dimensions, chaque axe représentant le flux à travers une réaction. L'ajout successif de ces contraintes va réduire la distribution des flux possibles à une partie de ce sous-espace (Palsson, 2000).

Deux approches utilisent ce cadre conceptuel : l'analyse de balance des flux (FBA en anglais) et la définition de voies métaboliques.

La FBA a comme fin de trouver un vecteur flux qui optimise une fonction objective. Celle-ci peut être la production maximale d'un composé ou d'un groupe

de composés, comme ceux qui participent à la biomasse de la cellule. Depuis son développement, la FBA a connu de nombreuses applications. Elle est très efficace notamment dans la prédiction de phénotypes (Edwards & Palsson, 2000). Le changement des contraintes sur le flux à travers certaines réactions permet de simuler l'effet de mutations d'un ou plusieurs gènes ou de perturbations dans le système, comme la disparition d'un nutriment dans le milieu. La limitation principale de la FBA est que la solution considérée est optimale du point de vue mathématique mais ne réflète pas forcément la réalité. Plusieurs solutions optimales ou quasi-optimales peuvent en effet être possibles et l'adaptation de la cellule n'atteint pas forcément un état métabolique optimal.

Les modèles à base de contraintes ont été utilisés également pour tenter de définir mathématiquement les voies métaboliques. Plusieurs concepts similaires existent mais le plus connu est certainement celui des modes élémentaires (Schuster & Hilgetag, 1994). Basiquement, un mode élémentaire est un mode (flux) spécial qui ne peut contenir aucun autre mode, ce qui signifie que l'inactivation d'une réaction normalement active dans un mode élémentaire le rend non fonctionnel. Malheureusement, malgré une formalisation élégante, le lien entre voies métaboliques traditionellement définies et les modes élémentaires est difficile à établir. En effet, le nombre de modes élémentaires trouvés dans un réseau explose très vite avec la taille de celui-ci, rendant très difficile leur analyse.

Un mode élémentaire peut être vu comme un ensemble de réactions qui, quand elles sont actives ensembles, accomplissent une fonction donnée. De manière opposée, on peut aussi être intéressé par savoir quels sont les ensembles de réactions suffisants pour empêcher une tâche (indiquée sous la forme d'une liste de réactions cibles). C'est à cette question qu'ont tenté de répondre Klamt & Gilles (2004) en introduisant le concept d'ensembles minimaux de réactions à couper (Minimal Cut Sets ou MCS). Les MCS sont calculés à partir des modes élémentaires et correspondent aux ensembles de réactions dont l'inactivation dans le réseau rend non fonctionnels les modes élémentaires contenant les réactions cibles.

2.2.3 Les réseaux de Petri

Les réseaux de Petri peuvent être considérés comme des graphes bipartis (voir la section suivante) avec deux types de noeuds, les places et les transitions. Les places contiennent des jetons qui passent d'une place à l'autre par les transitions si certaines règles sont respectées. Dans le cas de réseaux de Petri modélisant un réseau métabolique, les places sont les métabolites et les transitions les réactions. Les jetons représentent les coefficients stœchiométriques des réactions. La règle pour activer une transition est que les places "substrats" contiennent toutes au moins un nombre de jetons égal à leur coefficient stœchiométrique dans la réaction. Quand la transition est activée, les jetons sont éliminés des places "substrats" et des jetons sont ajoutés dans les places "produits", suivant les coefficients stœchiométriques de ceux-ci. Dans le cadre des réseaux métaboliques, les

réseaux de Petri ont essentiellement été utilisés pour élucider le fonctionnement de certaines voies métaboliques en mettant en évidence, par exemple, des circuits fonctionnels (Oliveira *et al.*, 2003).

2.2.4 Les graphes métaboliques

La modélisation des réseaux métaboliques sous forme de graphes est rapide à mettre en place et va permettre d'utiliser la batterie de méthodes propres à la théorie des graphes pour analyser des caractéristiques globales du graphe, telles que sa topologie, et ainsi dégager rapidement certains traits métaboliques importants.

a. Les différents types de graphes métaboliques

Un graphe est un objet mathématique composé d'objets appelés noeuds reliés entre eux par des arêtes. Formellement, un graphe G est défini comme un couple (V, E) où V est un ensemble fini de noeuds (*vertices* en anglais) et E un ensemble fini d'arêtes (*edges* en anglais) qui est un sous-ensemble de V^2. Nous reprenons sciemment les notations provenant de l'anglais pour que ne pas égarer le lecteur habitué à ces notations. De même un réseau métabolique sera dans la suite indiqué par la lettre N (*network*).

La modélisation d'un réseau métabolique sous forme de graphe implique de définir quelles entités biologiques vont être associées aux noeuds et quelles relations chimiques ou biologiques vont être associées aux arêtes. Ces choix dépendent directement du type de questions auxquelles on veut répondre avec la modélisation.

La Figure 2.5 représente les différents types de graphes métaboliques.

Dans le graphe des composés, les noeuds correspondent aux métabolites et il y a une arête entre deux métabolites s'il existe une réaction où l'un est substrat et l'autre produit. Dans la Figure 2.5 a, A et C sont reliés par une arête car R1 produit A à partir de C.

Dans le graphe des réactions, les noeuds correspondent aux réactions. Il existe une arête entre deux réactions si l'une produit un métabolite consommé par l'autre. Dans la Figure 2.5 b, il y a une arête entre R1 et R2 car R1 produit D qui est substrat de R2. La réaction R5 n'est reliée à aucune autre réaction car son produit C n'est substrat d'aucune réaction.

Dans le graphe des enzymes, les noeuds correspondent aux enzymes. Il existe une arête entre deux enzymes si l'une catalyse au moins une réaction qui produit le substrat d'au moins une réaction catalysée par l'autre. Dans la Figure 2.5 c, il y a une arête entre E3 et E1 car E3 catalyse R3 qui produit I, lui-même substrat de R4, réaction catalysée par E1.

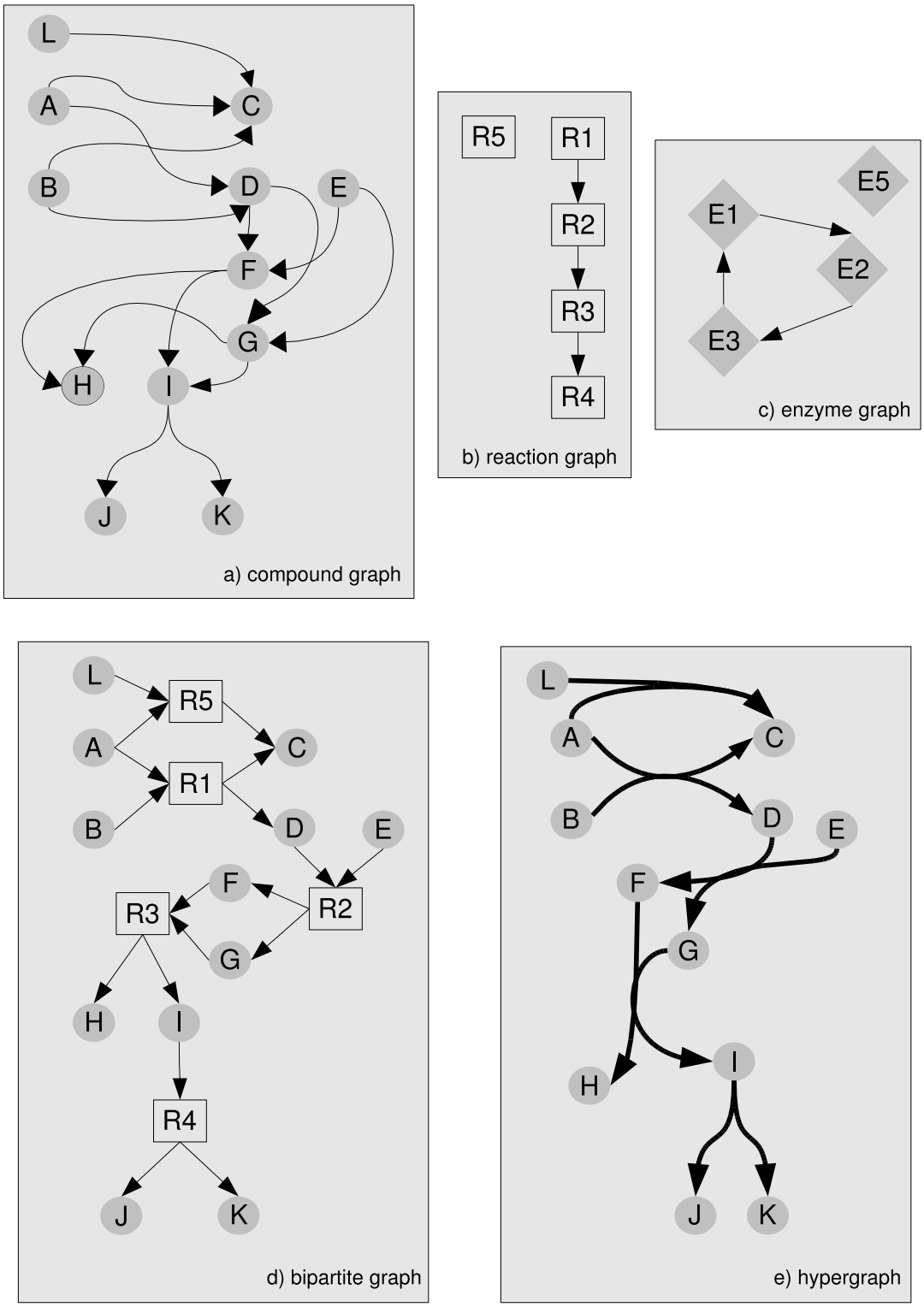

Figure 2.5. Les différents graphes métaboliques représenant le réseau métabolique suivant : $\{R1 : A + B \rightarrow C + D; R2 : D + E \rightarrow F + G; R3 : F + G \rightarrow H + I; R4 : I \rightarrow J + K; R5 : L + A \rightarrow C\}$. L'enzyme E1 catalyse les réactions R1 et R4, E2, E3 et E5 catalysent respectivement R2, R3 et R5.

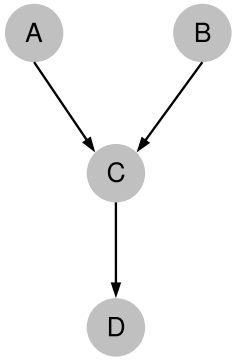

Figure 2.6. Le graphe des composés des réseaux $N1\{R1 : A \rightarrow C; R2 : B \rightarrow C; C \rightarrow D\}$ et $N2\{R1 : A + B \rightarrow C; R2 : C \rightarrow D\}$

On remarque déjà que ces graphes, bien que représentant le même réseau, n'ont pas du tout les mêmes propriétés, soulignant encore l'importance de bien définir quelles relations on veut représenter dans le réseau métabolique. Il arrive souvent dans les articles que les graphes de réactions et les graphes des enzymes soient confondus alors que leur analyse peut mener à des conclusions considérablement différentes. La différence de structure des graphes b et c de la Figure 2.5 en est un bon exemple.

Par ailleurs, la modélisation d'un réseau métabolique en l'un de ces trois graphes entraîne une perte d'information. En effet, dans la Figure 2.5 a, il existe une arête entre A et C laissant penser qu'il suffit de A pour produire C. Or, R1 nécessite à la fois A et B.

D'autres ambigüités apparaissent aussi lors de telles modélisations. Ainsi, dans la Figure 2.6, deux ensembles de réactions distincts sont modélisés de la même manière alors que la signification métabolique des deux réseaux est très différente.

Deux types de formalisations peuvent lever ces ambigüités : les graphes bipartis et les hypergraphes.

Un graphe biparti possède deux types de noeuds, un noeud d'un type donné ne pouvant être relié par une arête qu'à un noeud de type différent. Formellement, un graphe biparti est un graphe dont l'ensemble de noeuds V est divisé en deux sous-ensembles disjoints, V_1 et V_2, tels que chaque arête relie un noeud dans V_1 à un noeud dans V_2. Dans un graphe biparti métabolique, un type de noeuds correspond aux métabolites et l'autre type aux réactions. Dans la Figure 2.5 c, la nécessité de la présence simultanée des 2 métabolites A et B pour produire C est modélisée. De même, les deux réseaux de la Figure 2.6 conduisent maintenant à deux graphes différents (Figure 2.7).

Un hypergraphe est un graphe où les arêtes (alors appelées hyperarêtes) peuvent lier plus que deux noeuds. Formellement, un hypergraphe H est une paire (V, E) où $V = \{v_1, v_2, ..., v_n\}$ est un ensemble de noeuds et $E = \{e_1, e_2, ..., e_m\}$ avec $E_i \subseteq V$, pour $i = 1, ..., m$ est l'ensemble d'hyperarêtes. Dans un hypergraphe

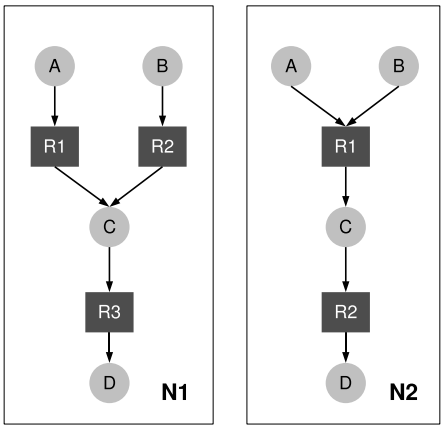

Figure 2.7. Les graphes bipartis des réseaux $N1\{R1 : A \rightarrow C; R2 : B \rightarrow C; C \rightarrow D\}$ et $N2\{R1 : A + B \rightarrow C; R2 : C \rightarrow D\}$

métabolique, les noeuds sont classiquement les métabolites et les hyperarêtes les réactions (Figure 2.5 c).

Tous ces types de graphes peuvent être dirigés ou non. Dans le cas d'un graphe dirigé, chaque arête portera une direction précise. On appelle "arc" un lien entre deux noeuds dans un graphe dirigé. S'il est non dirigé, l'arête ne portera pas de direction et la relation entre les deux noeuds sera réciproque. Dans le cas où toutes ses réactions seraient réversibles, un réseau métabolique est modélisé sous la forme d'un graphe non dirigé. Dans le cas où elles sont toutes irréversibles, le graphe correspondant est dirigé. Si certaines réactions seulement sont irréversibles, on peut utiliser un modèle mixte ou découpler les réactions réversibles en deux réactions irréversibles. Dans beaucoup d'articles, toutes les réactions sont considérées comme réversibles. Pourtant les conditions physiologiques et thermodynamiques font que certaines réactions ont une direction largement favorisée. Cependant, l'assignation des directions dans un réseau métabolique n'est pas immédiate et peu de méthodes rigoureuses permettent de résoudre ce problème actuellement (voir Section 2.1.4).

L'utilisation de graphes non dirigés peut amener à d'autres ambigüités. Par exemple, si les arêtes des graphes bipartis de la Figure 2.7 étaient non dirigées, on ne pourrait plus distinguer de quel côté de la réaction se situerait chaque métabolite, faussant complètement le calcul de chemins entre composés. Dans ce cas, il est nécessaire d'étiqueter les arêtes de façon à pouvoir distinguer les deux côtés de la réaction.

b. Les simplifications des modèles possibles

Dans les modèles de graphes présentés précédemment, tous les composés et toutes les réactions sont considérés comme équivalents. Pourtant, certains composés comme les coenzymes (voir Section 1.1) ont une importance centrale dans le réseau et interviennent dans de nombreuses réactions. On désigne souvent ces

composés comme "ubiquitaires". Considérer ces composés comme les autres peut conduire à des liens artificiels, particulièrement dans les graphes simples. Ainsi, dans un graphe des composés, de nombreux composés seront reliés à l'ATP alors que celui-ci n'intervient que comme cofacteur dans la réaction et ne peut produire la plupart des composés seul. La seconde raison de considérer autrement ce type de composés est d'ordre plus pratique. En effet, ces composés génèrent un nombre d'arêtes très important, ce qui peut augmenter considérablement le temps de calcul de certaines méthodes.

La façon la plus classique de traiter ce problème est de retirer tout simplement les composés qui participent à de nombreuses réactions (Jeong *et al.*, 2000; Ma & Zeng, 2003; Light *et al.*, 2005). Cependant, cette méthode n'est pas satisfaisante sur plusieurs points. D'abord, il est difficile de fixer le seuil de connectivité à partir duquel on retire les métabolites. Ensuite, certains composés comme le pyruvate ou le fructose sont très connectés mais interviennent réellement en tant que substrat ou produit dans le coeur de voies métaboliques importantes. Une autre façon de filtrer les composés ubiquitaires est de retirer du réseau les composés reconnus comme intervenant principalement en tant que cofacteurs. Cependant, même si dans la plupart des réactions leur élimination ne prête pas à conséquence, ils ne devraient pas être éliminés des réactions intervenant dans leur propre synthèse.

Une autre alternative est de retirer les composés dans les réactions où ils interviennent seulement comme substrats ou produits secondaires. La première idée est d'utiliser les informations contenues dans les voies métaboliques et de ne considérer dans le réseau que les composés qui interviennent dans la structure même des voies (Lacroix *et al.*, 2006).

Par ailleurs, certaines transformations de cofacteurs sont bien connues. Il est donc possible d'éliminer les composés intervenant dans ces transformations des réactions où elles ont lieu. Nous reparlerons de ces deux derniers filtres dans la section consacrée à SymbioCyc.

Enfin, il est possible depuis récemment d'utiliser la décomposition des réactions en transformations élémentaires que propose la base de données KEGG (voir Section 2.3.1).

Nous verrons plus tard que l'absence de traitement des composés ubiquitaires peut conduire à des conclusions erronées en ce qui concerne l'analyse des graphes métaboliques.

c. Les mesures classiques

Une fois que le réseau est modélisé sous la forme d'un graphe, les mesures classiques de la théorie des graphes peuvent être utilisées. Soulignons que la plupart de ces mesures s'appliquent aux graphes simples, raison pour laquelle les graphes métaboliques sont des graphes simples dans la plupart des études. Les mesures couramment utilisées pour analyser les graphes métaboliques sont les suivantes : le degré, la distance, la centralité, le diamètre et le coefficient d'agglomération.

Le degré d'un noeud i est le nombre d'arêtes le liant à d'autres noeuds. Si le graphe est dirigé, on parle de degré entrant et de degré sortant. Ainsi, dans un graphe de composés, le degré sortant d'un noeud i correspond au nombre de produits distincts des réactions qui utilisent i en tant que substrat et le degré entrant d'un noeud i correspond au nombre de composés distincts intervenant dans les réactions produisant i.

Dans un graphe de réactions, le degré entrant d'un noeud i correspond au nombre total de réactions qui produisent les substrats de i et le degré sortant d'un noeud i correspond au nombre total de réactions qui utilisent au moins un produit de i en tant que substrat.

La distance entre deux noeuds i et j est la longueur du plus court chemin entre deux noeuds. Autrement dit, c'est le nombre minimal d'arêtes qu'il faut utiliser pour passer d'un noeud à l'autre. Si on considère un graphe des composés, la distance entre deux noeuds représenterait le nombre minimal de réactions utilisées pour produire un métabolite à partir d'un autre.

Le diamètre d'un graphe est la distance maximale entre deux noeuds quelconques.

Le coefficient d'agglomération (*clustering coefficient*) d'un noeud i est la proportion du nombre d'arêtes existantes sur le nombre d'arêtes possibles entre les voisins de i. Le coefficient d'agglomération moyen renseigne sur la tendance des noeuds à former des groupes très connectés autour d'eux.

La centralité d'un noeud i peut se mesurer de deux manières. La première, appelée centralité d'interposition (*betweenness centrality*), mesure la proportion des plus courts chemins passant par i sur le nombre total de plus courts chemins entre toutes les paires de noeuds d'un graphe. Dans le cas d'un graphe de réactions, une telle mesure peut renseigner sur la présence de réactions qui soient des "passages obligés" dans le réseau global. La seconde mesure de centralité, appelée centralité de proximité (*closeness centrality*), mesure la proximité d'un noeud par rapport à tous les autres.

Nous verrons dans la Section 4 comment ces mesures sont utilisées pour décrire et comparer les réseaux métaboliques.

Cependant, l'interprétation de ces mesures nécessite beaucoup de précautions par la nature même du réseau et de ses objets (Lacroix *et al.*, 2008b). Un prétraitement du réseau métabolique, par des filtres comme ceux proposés par SymbioCyc (voir Section 3), est souvent nécessaire afin d'éviter certains artefacts dus à la nature des données.

Par ailleurs, il est important aujourd'hui d'imaginer des mesures propres aux graphes métaboliques, élaborées en gardant à l'esprit le type d'objets que l'on modélise mais aussi la qualité des données disponibles. C'est dans ce cadre que se place le développement de PITUFO, comme nous le verrons plus tard dans (Section 5).

2.3 Exploration et échange des données métaboliques

2.3.1 Les bases de données métaboliques et leurs outils associés

Les deux bases métaboliques les plus utilisées dans le monde de la bioinformatique sont certainement BioCyc (Caspi *et al.*, 2008) et KEGG (Kanehisa *et al.*, 2008) dont on a déjà parlé dans le cadre de la reconstruction d'un réseau métabolique à partir d'un génome. Le grand intérêt de ces deux bases est qu'elles rendent disponibles des informations à la fois génomiques, biochimiques et métaboliques de la plupart des organismes séquencés à ce jour. De plus, elles disposent de nombreux outils associés. Nous allons passer en revue leur points communs et leurs spécificités.

KEGG et BioCyc proposent une exploration visuelle des génomes, et des fiches informatives sur les métabolites, gènes, enzymes, réactions et voies métaboliques.

Comme nous l'avons vu précédemment, les gènes dans KEGG sont classés selon leur numéro KO, classés eux-mêmes dans les voies métaboliques où ils interviennent (Figure 2.8). Dans BioCyc, mis à part dans la partie réservée à *Escherichia coli K12*, EcoCyc, les gènes ne sont classés d'aucune manière.

La représentation des voies métaboliques diffère considérablement selon la base. Dans KEGG, les voies métaboliques sont organisées en cartes métaboliques où toutes les variantes de la voie sont dessinées . À partir d'une liste d'identifiants de gènes ou de numéros EC, il est possible de surligner les réactions correspondantes dans chaque voie métabolique de référence (Figure 2.9). Ces voies sont donc des voies théoriques que l'on ne trouve complète chez aucun organisme.

Au contraire des cartes métaboliques de KEGG, chaque variante de voie métabolique dans BioCyc correspond à une représentation différente (Figure 2.10).

Les deux représentations sont complémentaires. La représentation de KEGG permet de superposer facilement les parties de voies métaboliques possibles chez un organisme par rapport à toutes les variantes possibles. D'un autre côté, la voie métabolique telle qu'elle est représentée dans BioCyc est plus lisible et toutes les informations, du gène à l'enzyme, y sont représentées.

Depuis peu, chaque réaction dans KEGG est décomposée en transformations élémentaires qui correspondent aux transferts d'atomes entre les métabolites participant à la réaction (Kotera *et al.*, 2004; Oh *et al.*, 2007). Cette décomposition peut être très utile pour traiter les données avant la modélisation pour éviter, par exemple, les chemins non réalistes entre composés dans un graphe métabolique.

La comparaison des données est facilitée dans BioCyc. Lorsque l'utilisateur compulse la fiche d'une réaction ou d'une voie métabolique particulière, il est possible de tester leur présence chez les autres organismes présents dans la base.

01100 Metabolism
  01110 Carbohydrate metabolism
  01120 Energy metabolism
  01130 Lipid metabolism
  01140 Nucleotide metabolism
  01150 Amino acid metabolism
    00251 Glutamate metabolism
    .....
    00300 Lysine biosynthesis
      K00003 E1.1.1.3, thrA; homoserine dehydrogenase
      K00928 E2.7.2.4, lysC; aspartate kinase
      K00133 E1.2.1.11, asd; aspartate-semialdehyde dehydrogenase
      K01714 E4.2.1.52, dapA; dihydrodipicolinate synthase
      K00215 E1.3.1.26, dapB; dihydrodipicolinate reductase
      K00674 E2.3.1.117, dapD; 2,3,4,5-tetrahydropyridine-2-carboxylate *N*-succinyltransferase
      K00821 E2.6.1.17; *N*-succinyldiaminopimelate aminotransferase
      K01439 E3.5.1.18, dapE; succinyl-diaminopimelate desuccinylase
      K01778 E5.1.1.7, dapF; diaminopimelate epimerase
      K01586 E4.1.1.20, lysA; diaminopimelate decarboxylase
      .....
    00310 Lysine degradation
    .....
  01160 Metabolism of other amino acids
  .....
01200 Genetic information processing
01300 Environmental information processing
01400 Cellular processes
01500 Human disease

Figure 2.8. Extrait de la classification KO de KEGG.

Lors de la comparaison d'une voie métabolique, l'information sur les gènes et les réactions manquantes apparaît (Figure 2.11).

Une analyse comparative des composés, des protéines, des réactions et des voies est possible également dans BioCyc, fournissant des statistiques détaillées sur ces objets à travers une collection d'organismes. Quoique très complète, cette analyse a pourtant, pour nous, un défaut majeur. En effet, si une voie métabolique a été définie comme présente chez l'organisme considéré, toutes les réactions y participant sont considérées également présentes, même si elles correspondent en fait à des réactions manquantes, pour lesquelles aucun gène ou enzyme n'a été assigné (voir Section 2.1.4).

La différence majeure entre les deux bases de données réside dans leur philosophie. KEGG propose un unique site où les reconstructions sont centralisées et effectuées par la même équipe. L'équipe à l'origine de BioCyc propose un tout autre mode de fonctionnement, fondé sur le partage des tâches. Leur sentiment est qu'aucune équipe n'est capable d'expertiser les annotations d'une diversité aussi importante d'organismes. Des ébauches de reconstruction sont ainsi prêtes à être adoptées par d'autres équipes spécialistes de certains organismes qui prennent en charge le nettoyage et l'amélioration des données. Les organismes dans BioCyc

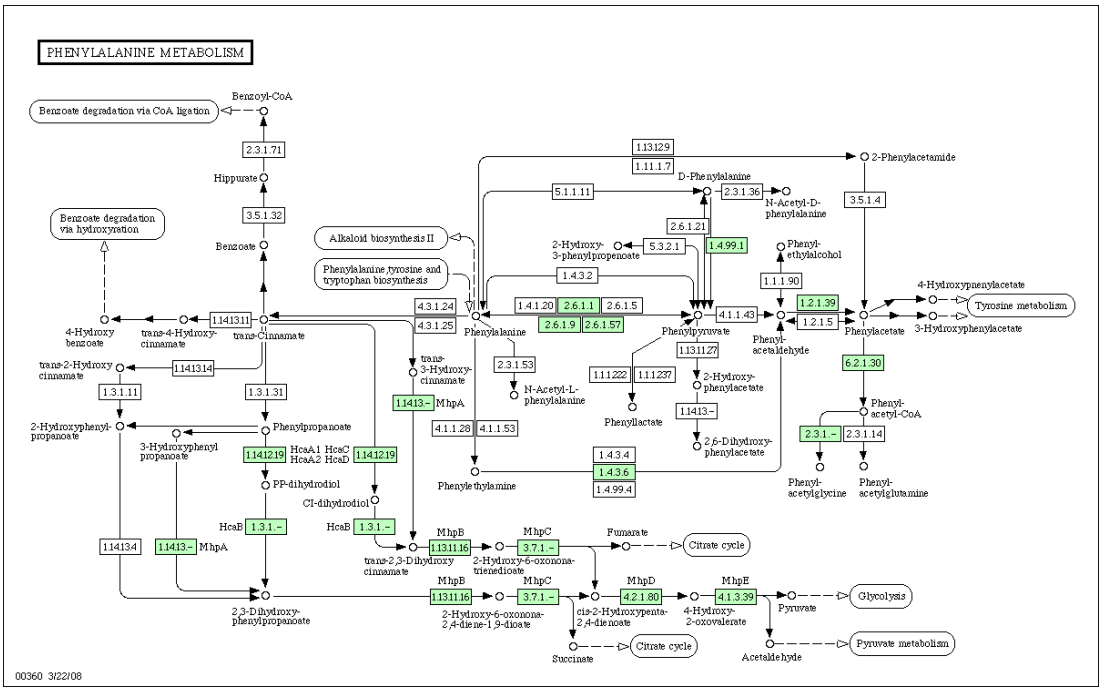

Figure 2.9. Carte métabolique KEGG de la voie de synthèse de la phénylalanine. En vert apparaissent les réactions pour lesquelles un gène de *Escherichia coli K12* a été annoté dans KEGG.

sont classées en 3 niveaux d'expertise. Le premier contient MetaCyc et EcoCyc. MetaCyc contient plus d'un millier de voies métaboliques décrites expérimentalement chez plus de 1500 organismes (Caspi *et al.*, 2008). C'est cette base qui sert notamment de référence pour les reconstructions métaboliques. EcoCyc est dédiée aux génomes de plusieurs souches d'*Escherichia coli* et son expertise est effectuée par plusieurs équipes. C'est certainement la base métabolique la plus expertisée dédiée à une espèce (Karp *et al.*, 2007). Le deuxième niveau d'expertise contient les reconstructions qui ont connu un début d'expertise par d'autres équipes. Les quelques 200 reconstructions effectuées par MaGe (Vallenet *et al.*, 2006) pour le projet MicroScope du génoscope entrent dans cette catégorie. Enfin, le troisième niveau d'expertise contient près de 350 reconstructions métaboliques effectuées automatiquement par l'équipe de BioCyc pour lesquelles aucun nettoyage n'a été effectué. Ces reconstructions sont mises à diposition pour que d'autres équipes les améliorent. Le fait de déléguer l'amélioration de leurs bases de données à d'autres équipes implique l'existence d'un outil rendant capable non seulement la navigation parmi les données, mais aussi l'édition et la création de nouveaux objets. L'outil associé à BioCyc, les pathway-tools (Karp *et al.*, 2002), répond à cette demande. Cet outil possède la même interface de navigation que le site web mais permet de compléter, corriger ou créer de nouveaux objets, notamment de nouvelles voies métaboliques qui seraient propres à l'organisme pris en charge. De plus, les pathway-tools contiennent l'outil Pathologic destiné à la reconstruction

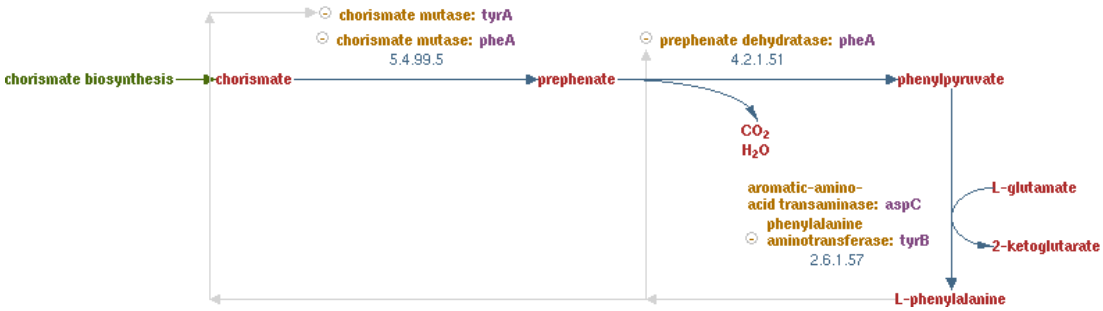

Figure 2.10. Voie de synthèse de la phénylalanine d'*Escherichia coli K12* dans BioCyc.

des réseaux métaboliques à partir d'informations génomiques (voir Section 2.1). Il est possible ainsi d'effectuer une reconstruction métabolique à partir d'annotations génomiques "maison". Par ailleurs, les pathway-tools permettent très facilement de créer une interface web en tout point semblable à celle que propose BioCyc, rendant ainsi navigables les données nouvellement créées. Enfin, des interfaces de programmation en Lisp, Java et Perl sont disponibles et permettent des requêtes complexes et l'automatisation de tâches en vue de modéliser ou d'analyser les données locales créées par les pathway-tools (Krummenacker *et al.*, 2005).

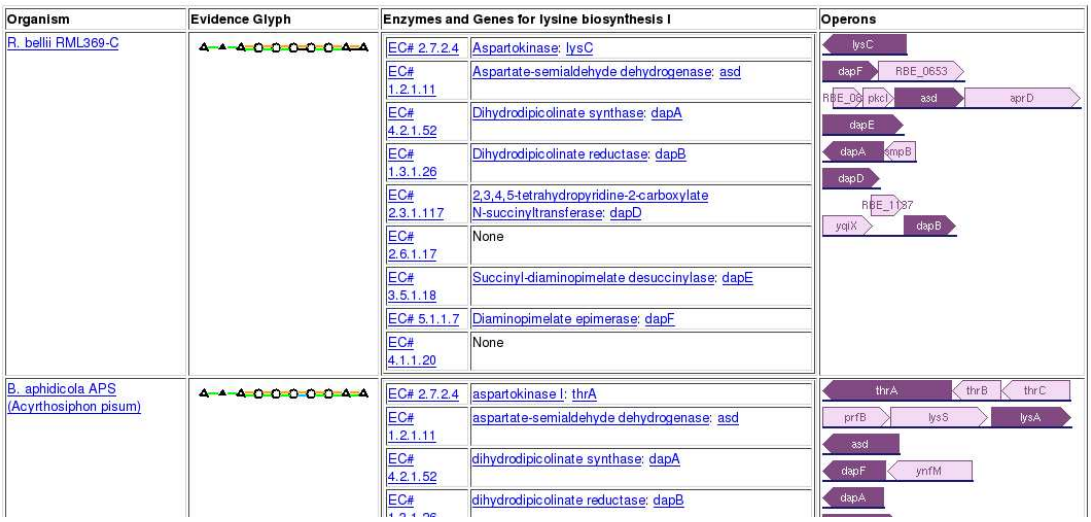

Figure 2.11. Extrait de la comparaison de la voie de synthèse de la lysine chez *Buchnera aphidicola APS* et *Rickettsia bellii* dans BioCyc

2.3.2 Les outils de visualisation des réseaux métaboliques

Nous avons vu que KEGG et BioCyc proposent une visualisation des voies métaboliques. Cependant, dans le but d'effectuer une analyse globale du réseau, il est intéressant de pouvoir le visualiser complètement. Les deux bases de données proposent deux systèmes très semblables pour visualiser l'ensemble du réseau (Okuda *et al.*, 2008; Paley & Karp, 2006) : la carte métabolique est divisée en grands processus divisés eux-mêmes en voies métaboliques. Chacun des deux systèmes propose en outre de colorer sur la carte métabolique des données expérimentales, comme celles provenant des mesures d'expression de gènes.

La différence entre les deux systèmes est la même que celle entre les deux représentations de voies métaboliques. KEGG met en relief sur une carte métabolique globale le réseau métabolique de l'organisme considéré tandis que celui-ci sera seul représenté dans le système de BioCyc (Figures 2.12 et 2.13).

L'organisation spatiale en voies métaboliques dans la représentation graphique implique une duplication des noeuds (composés et réactions), ce qui la rend plus claire. Pourtant, lorsqu'on modélise le réseau sous forme de graphes, on peut vouloir voir l'environnement direct de certains noeuds ou vérifier certains chemins métaboliques. Pour cela, on peut utiliser de nombreux logiciels permettant de visualiser les graphes, comme yEd[2] et Tulip[3]. Ceux-ci proposent de puissants modes de visualisation permettant entre autres de dessiner les noeuds du graphe en fonction de certaines mesures, dont celles exposées dans la Section 2.2.4. Cytoscape propose un bon nombre de ces fonctionnalités mais également d'autres, plus spécifiques des graphes biologiques (Shannon *et al.*, 2003). Depuis sa création, de nombreuses extensions ont été développées par diverses équipes afin d'importer, analyser et dessiner des données de réseaux biologiques, dont les réseaux métaboliques. Si ceux-ci sont dans le format SBML (voir Section suivante), ils sont directement importés et dessinés dans Cytoscape sous la forme d'un graphe biparti (Figure 2.14).

Cependant, dans le cas d'une représentation sous forme de graphes, l'organisation en voies métaboliques que l'on a avec les vues globales de BioCyc et de KEGG est totalement perdue. Récemment, nous avons contribué au développement d'une solution intermédiaire en participant à l'élaboration d'un logiciel permettant de dessiner certaines voies métaboliques correctement tout en n'effectuant pas de duplication de noeuds (Bourqui *et al.*, 2007). Les voies métaboliques correctement dessinées seront celles qui contiendront le plus de noeuds ou celles choisies par l'utilisateur.

[2] `http://www.yworks.com/en/products_yed_about.html`
[3] `http://tulip.labri.fr/`

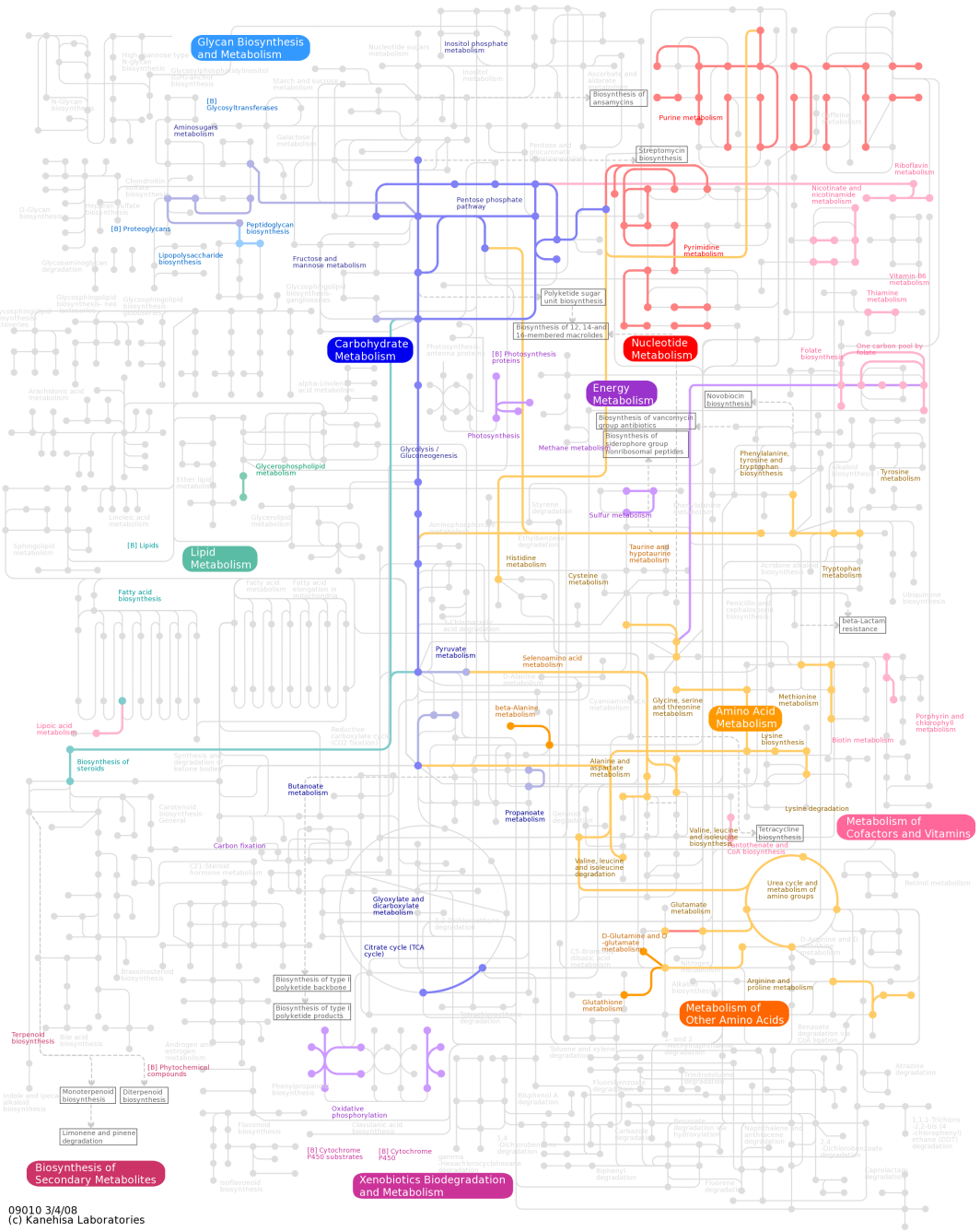

Figure 2.12. Vue globale du métabolisme de *Buchnera aphidicola APS* dans KEGG.

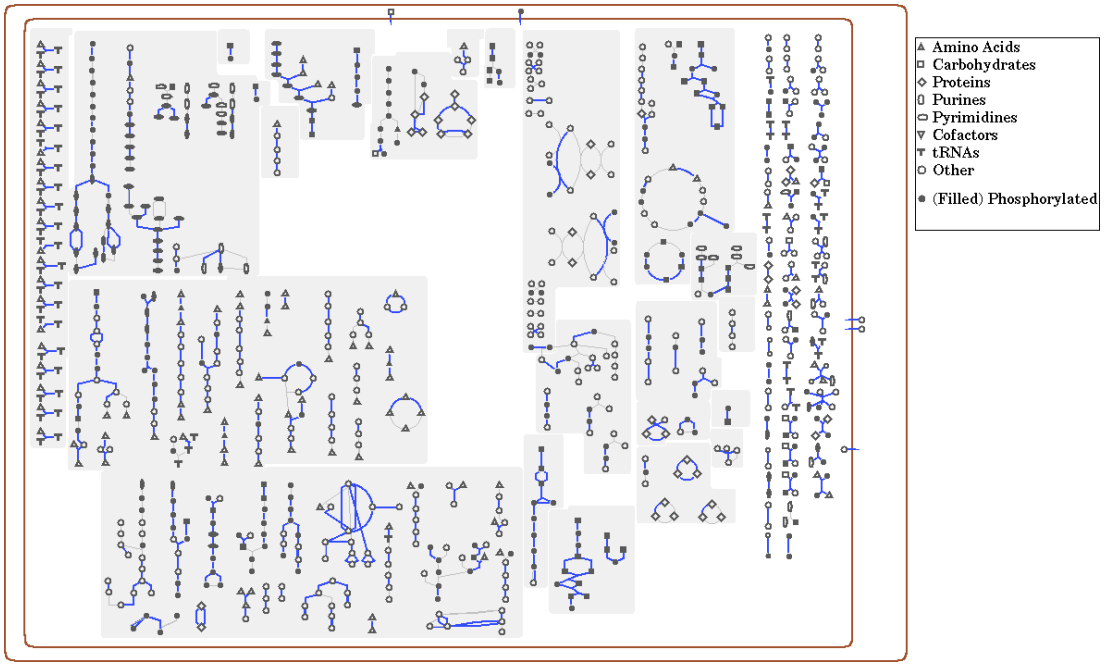

Figure 2.13. Vue globale du métabolisme de *Buchnera aphidicola APS* dans BioCyc.

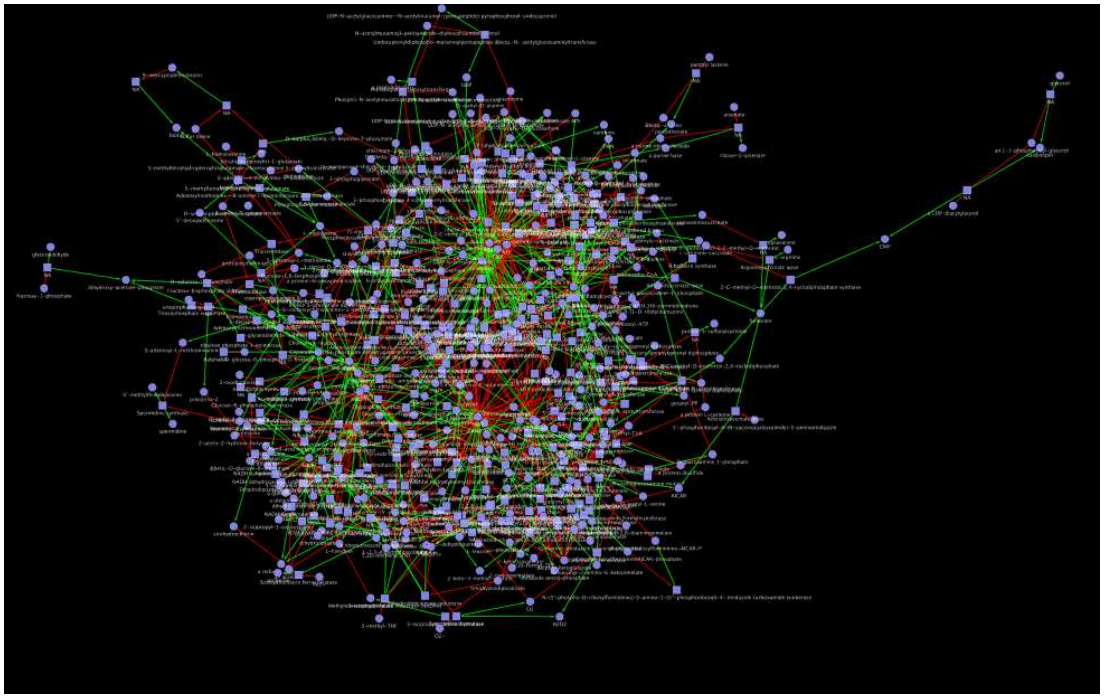

Figure 2.14. Réseau métabolique de *Buchnera aphidicola APS* dessiné par Cytoscape sous la forme d'un graphe biparti. Les noeuds carrés sont les réactions et les noeuds ronds les composés. Une arête rouge représente une relation réaction-substrat et une arête verte une relation réaction-produit.

2.3.3 Les formats d'échange

Afin de partager ou faciliter les analyses des données métaboliques par divers outils, il est important de disposer de formats d'échange communs. Pour les réseaux métaboliques, principalement deux formats sont disponibles : Biopax et SBML (Strömbäck & Lambrix, 2005; Finney A, 2003), tous deux basés sur le format XML.

Le format BioPax est le plus complet puisqu'il est capable de gérer des liens hiérarchiques entre objets et que toutes les informations, des gènes aux composés, peuvent être stockés dans un seul fichier. Cependant, il semble aujourd'hui que sa complexité rend son utilisation rare dans les outils dédiés aux réseaux métaboliques. Notons tout de même que BioCyc permet d'exporter les voies métaboliques sous format BioPax.

D'un autre côté, le format SBML est beaucoup plus simple. Il est composé essentiellement d'une liste de réactions avec la liste des substrats et des produits. Des informations à propos de la concentration des composés et les propriétés cinétiques des réactions peuvent être ajoutées pour permettre les simulations numériques du fonctionnement du réseau. SBML est maintenant utilisé par de nombreux outils (www.sbml.org). BioCyc permet notamment d'exporter ses données sous format SBML. Par ailleurs, les reconstructions métaboliques de qualité effectuées par l'équipe de Palsson (Duarte *et al.*, 2007; Feist *et al.*, 2007; Reed *et al.*, 2003) sont également disponibles en SBML sur la page du groupe (http ://gcrg.ucsd.edu/).

Deuxième partie

Résultats

Chapitre 3

SymBioCyc : une base de données pour la comparaison de réseaux métaboliques bactériens

SymbioCyc est une base de données dédiée aux réseaux métaboliques de bactéries. Elle est particulièrement destinée aux personnes voulant entreprendre la modélisation de ces réseaux. En effet, SymbioCyc met à disposition des filtres sur les données qui peuvent être exportées ensuite sous forme de graphes ou de fichiers SBML. De plus, SymbioCyc propose des statistiques élémentaires permettant d'identifier les grandes caractéristiques de chaque réseau. Enfin, SymbioCyc dispose d'un outil de comparaison original permettant de comparer les données métaboliques entre organismes ou par groupe d'organismes. Ces fonctionnalités en font le premier site internet permettant à la fois l'exploration, le filtrage, la comparaison et la modélisation sous formes de graphes des réseaux métaboliques de plusieurs bactéries et en particulier de bactéries endosymbiotes. Le chercheur dispose ainsi d'un matériel prêt à l'emploi qu'il peut utiliser immédiatement pour l'analyse.

3.1 Les organismes de SymbioCyc

Au moment de la rédaction de ce manuscrit, SymbioCyc contient les informations métaboliques de 50 bactéries. Celles-ci se répartissent phylogénétiquement en quatre grandes classes de bactéries : les Protéobactéries (41), les Firmicutes (6), les Actinobactéries (2) et les Flavobactéries (1) (Figure 3.1).

Les Protéobactéries sont de loin les plus représentées, particulièrement les gamma-protéobactéries (15) et les α-protéobactéries (21). Les β-protéobactéries et les δ-protéobactéries ne comptent qu'un représentant et les ϵ-protéobactéries n'en comptent que trois.

Les bactéries de SymbioCyc ont surtout été choisies pour l'éventail de leur style de vie. Elles peuvent ainsi être classées selon trois critères :

- la relation entrenue avec l'hôte, si elle existe,
- l'existence d'un stade de vie intracellulaire prolongé,
- le mode de transmission principal vers d'autres hôtes : horizontal ou vertical.

Selon ces critères, les bactéries de SymbioCyc peuvent être divisées en sept grands groupes de style de vie (Figure 3.1).

- 12 mutualistes intracellulaires à transmission verticale (**MIV**),
- 8 parasites à stade de vie intracellulaire à transmission verticale (**PIV**),
- 9 parasites à stade de vie intracellulaire à transmission horizontale (**PIH**),
- 7 parasites extracellulaires à transmission horizontale (**PEH**),
- 8 mutualistes extracellulaires à transmission horizontale (**MEH**),
- 3 commensalistes extracellulaires à transmission horizontale (**CEH**),
- 3 libres (**L**).

Nous utiliserons dans la suite les abbréviations entre parenthèses pour désigner chaque grand groupe de style de vie.

Il est évident qu'il existe de nombreux autres critères pour classer les organismes en fonction de leur style de vie. Il n'y a pas à notre connaissance de consensus à ce sujet.

Les bactéries intracellulaires sont au nombre de 29 dans SymbioCyc.

Les bactéries MIV de SymbioCyc sont quasiment toutes des γ-protéobactéries. Seule *Wolbachia pipientis wBm* fait partie des α-protéobactéries. On les retrouve toutes dans des cellules d'insectes à l'exception encore de *Wolbachia pipientis wBm* que l'on trouve dans les cellules du filaire de Malaisie (*Brugia malayi*) qui est un nématode. Le caractère mutualiste s'exprime ici par une association nutritionelle : la cellule de la bactérie fait l'objet d'échanges métaboliques intensifs avec la cellule hôte. *Carsonella ruddii*, les quatre *Buchnera*, les deux *Blochmannia* et *Sulcia muelleri* interviendraient principalement dans la synthèse d'acides aminés essentiels, *Baumannia cicadellinicola* et *Wigglesworthia* dans la synthèse de cofacteurs, *Wolbachia pipientis wBm* dans la synthèse de cofacteurs mais aussi dans celle du hème et de nucléotides (Taylor *et al.*, 2005; Foster *et al.*, 2005). C'est dans ce groupe que l'on retrouve les génomes et les métabolismes les plus réduits (voir Section 1.2.3 p.27).

Les **bactéries PIV** de SymbioCyc sont toutes des α-protéobactéries et plus particulièrement des Rickettsiales. Sept espèces de *Rickettsia* et une espèce de *Wolbachia* sont réprésentées.

Les *Rickettsia* ont pour la plupart comme hôtes habituels des arthropodes. Dans ce cas, la bactérie est transmise par voie ovarienne à la progéniture de l'hôte avec une grande fidélité puisqu'elle est détectée dans les populations de l'hôte à des fréquences considérables (Sakurai *et al.*, 2005). Il a été reconnu que certaines agissent sur la reproduction de leurs hôtes par un effet "male-killing" : elles

tuent les mâles infectés (Hurst & Jiggins, 2000). Même si leur principal mode de transmission se fait par voie ovarienne, il arrive fréquemment que les *Rickettsia* se transmettent horizontalement après que l'hôte se soit nourrit du sang d'un mammifère (Hurst & Jiggins, 2000). Les *Wolbachia* sont de très proches parentes des *Rickettsia*. Elles sont surtout connues pour la diversité des mécanismes de manipulation de la reproduction de leur hôte : incompatibilité cytoplasmique, induction de la parthénogénèse, mort et féminisation des mâles (Werren, 1997).

Les **bactéries PIH** de SymbioCyc sont réparties dans 3 groupes phylogénétiques : les α-protéobactéries, les δ-protéobactéries et les Mycoplasmes. Dans le groupe des α-protéobactéries, les quatre bactéries PIH ne sont représentées que par un seul genre : *Bartonella*. Celles-ci sont des bactéries parasites du sang. Leur cycle de vie est assez complexe. Leur hôte habituel est un mammifère où elles s'installeraient d'abord autour et dans les cellules épithéliales dont elles peuvent causer la prolifération chez des hôtes non habituels ou immunodéprimés. Elles infectent ensuite les érythrocytes où elles se reproduisent. Les arthropodes suceurs de sang jouent le rôle de vecteurs. Les *Bartonella* colonisent l'intestin et persistent à l'extérieur des cellules. L'infection de l'hôte se ferait par les fécès (Birtles, 2005). La seule représentante des PIH dans la division des δ-protéobactéries est *Lawsonnia intracellularis*. Cette bactérie est obligatoirement intracellulaire et envahit les cellules endothéliales principalement du porc où elle est fortement pathogène. Enfin, les bactéries PIH sont représentées dans le groupe des Mycoplasmes par quatre bactéries : trois souches de la même espèce, *Mycoplasma hyopneumoniae*, et la bactérie *Mycoplasma genitalium*. Les Mycoplasmes sont commensaux ou parasites des vertébrés, et peuvent devenir fortement pathogènes. Chez tous les Mycoplasmes, la faculté de fabriquer la paroi cellulaire a été perdue. Elles se fixent sur les cellules des membranes de la trachée, des conduits urogénitaux, des yeux ou des glandes mammaires.
L'absence de paroi rigide pourrait faciliter la fusion des membranes du parasite et des cellules de l'hôte, rendant possible l'échange de composants cytoplasmiques. L'interaction avec les cellules hôtes étant très intime, les Mycoplasmes peuvent être classées parmi les bactéries à stade de vie intracellulaire même si strictement parlant, elles sont considérées comme extracellulaires. Les Mycoplasmes comptent également parmi les génomes bactériens les plus réduits.

Les bactéries extracellulaires sont au nombre de 22 dans SymbioCyc.

Les **bactéries MEH** comptent huit représentantes dans SymbioCyc. Le mutualisme ici s'exprime uniquement par la fixation d'azote fourni à une plante hôte. Six font partie des α-protéobactéries, une des β-protéobactéries et une autre des Actinobactéries. La symbiose se réalise au niveau des racines au sein même d'excroissances de la plante, les nodules, dont la croissance est provoquée par des signaux bactériens.

Les **bactéries PEH** comptent sept représentantes dans SymbioCyc, et se répartissent dans quatre groupes phylogénétiques. *Agrobacterium tumefasciens* est une α-protéobactérie PEH capable de transférer une partie de son ADN aux plantes qu'elle parasite. Cet ADN, intégré ensuite au génome nucléaire de la plante, va produire des hormones de croissance végétale, entraînant ainsi la formation de tumeurs et de nutriments utiles aux parasites.

Campylobacter jejuni est une ϵ-protéobactérie qui colonise les intestins de mammifères et d'oiseaux et peut envahir la couche épithéliale. *Helicobacter pylori*, également une ϵ-protéobactérie, attaque la surface de l'estomac des hommes et peut être très fortement pathogène.

Photorhabdus luminescens est une γ-protéobactérie qui vit dans le corps de nématodes entomopathogènes. Ceux-ci infectent des larves d'insecte et libèrent *Photorhabdus luminescens* dans le sang de leur proie. *Photorhabdus luminescens* produit alors des toxines qui contribuent à accélérer la mort de l'insecte et des enzymes qui dégradent les constituants de l'insecte en nutriments directement assimilables par le nématode. Les bactéries entrent dans la progéniture du nématode lorsqu'elles se développent. *Photorhabdus luminescens* sécrète également des antibiotiques qui protègent le nématode de l'attaque d'autres bactéries. *Photorhabdus luminescens* peut être considérée ainsi comme parasite chez l'insecte et mutualiste chez le nématode.

Salmonella enterica Typhi est responsable de la fièvre typhoïde chez l'homme. Une fois ingérée, la bactérie se multiplie dans l'intestin puis traverse la paroi intestinale et envahit d'autres tissus.

Pseudomonas aeruginosa est une γ-protéobactérie infectant un très large spectre d'hôtes (plantes, humains, arthropodes, ...). Chez l'homme, c'est une bactérie opportuniste qui cause des septicémies ou des pneumonies chez les individus immuno-déprimés.

La bacille du charbon, *Bacillus anthracis*, développe des spores très résistantes qui lui permettent de survivre des dizaines d'années dans le sol. Elle contamine surtout les animaux par l'alimentation et provoque une mort rapide par dissémination sanguine. Lorsqu'elle entre dans l'hôte, elles produit des cellules végétatives qui se multiplient et produisent des facteurs de virulence qui finissent par tuer l'hôte. Après la mort de l'hôte, les bactéries sont relâchées dans l'environnement et sporulent, complétant ainsi leur cycle de vie (Rasko *et al.*, 2005).

Les **bactéries commensalistes** sont au nombre de trois dans SymbioCyc. La γ-protéobactérie *Escherichia coli K12* est communément trouvée dans l'intestin humain.

Wolinella succinogenes est une δ-protéobactérie trouvée dans le rumen des vaches. Elle n'est pas pathogène alors qu'elle possède des gènes de virulence. De même, elle possède des gènes reliés à la fixation d'azote, peut-être transférés horizontalement par des Rhizobiales.

Streptococcus agalactiae est une bacille commensale du tube digestif de l'homme.

C'est aussi une bactérie pathogène opportuniste chez les personnes âgées ou les personnes immunodéficientes.

Les **bactéries libres** sont au nombre de trois dans SymbioCyc. *Rhodobacter sphaeroides* est une α-protéobactérie photosynthétique.

La γ-protéobactérie *Vibrio cholerae* vit dans l'eau et a une grande capacité de survie environnementale. Elle cause cependant le cholera quand l'homme l'ingère.

L'actinomycétale *Corynebacterium glutamicum* est une bactérie du sol, utilisée dans l'industrie pour sa grande capacité à produire du glutamate.

3.2 Annotation des génomes et reconstruction des réseaux métaboliques par MaGe

Les reconstructions métaboliques de SymbioCyc ont toutes été effectuées à partir des annotations génomiques fournies par la plate-forme d'annotation des procaryotes du Génoscope, MaGe (Magnifying Genomes) (Vallenet *et al.*, 2006). Les principales fonctionnalités de MaGe sont la réannotation de génomes bactériens en intégrant le résultat de plusieurs méthodes ainsi qu'une interface web autorisant plusieurs experts à corriger les annotations et une exploration facilitée des données. MaGe propose en effet une annotation complète allant de la détection des bornes des gènes sur le chromosome aux reconstructions métaboliques, intégrant les résultats de plusieurs méthodes automatiques. La détection des limites des gènes se fait avec le logiciel AMIGene d'après un modèle de gène qui repose sur l'usage des codons (Bocs *et al.*, 2003). D'autres outils sont utilisés pour affiner les prédictions : RBSfinder pour la recherche des sites de fixation des ribosomes (Suzek *et al.*, 2001), tRNAscan-SE pour la recherche des gènes codant pour les ARN de transfert (Lowe & Eddy, 1997), la base de données Rfam pour identifier les gènes à ARN non codants (Griffiths-Jones *et al.*, 2005) et le logiciel PETRIN pour identifier les sites de terminaison (d'Aubenton Carafa *et al.*, 1990).

L'annotation fonctionnelle des gènes par MaGe se fait par étapes successives, chaque nouvelle étape correspondant à un niveau de confiance moindre dans les données (Figure 3.2). Une originalité de MaGe est de prendre en compte les informations de synténie à chaque comparaison. Ainsi, le seuil d'acceptation de l'annotation d'un gène fournie par un de ses homologues sera diminué si les deux gènes sont trouvés dans le même groupe de synténie.

Les gènes codant pour les enzymes sont annotés avec PRIAM qui leur assigne un numéro EC (Claudel-Renard *et al.*, 2003) (voir Section 2.1).

Enfin, MaGe fournit sur le site web deux types de reconstructions métaboliques, une provenant de KEGG et l'autres des pathway-tools, combinant ainsi les qualités des deux systèmes (voir Section 2.1). De plus, le site de MaGe est directement connecté au serveur Web du Pathway Hunter Tool (PHT) qui permet

d'identifer des voies métaboliques potentielles qui n'existent pas parmi les voies métaboliques de référence (Rahman *et al.*, 2005).

L'élaboration de SymbioCyc repose entièrement sur la reconstruction métabolique effectuée par les Pathway-Tools. Trente deux organismes proviennent de projets déjà existants tandis que 18 organismes proviennent du projet SmallScope dont nous sommes à l'initiative et qui regroupe des bactéries endosymbiotes.

Le lien de SymbioCyc avec les données MaGe nous permet d'avoir à disposition des reconstructions métaboliques de qualité et mises à jour. De plus, les nombreux outils d'exploration génomiques de MaGe sont disponibles pour chaque organisme de SymbioCyc, ce qui facilite l'analyse des données.

Enfin, les 18 organismes du projet SmallScope pourront dans l'avenir bénéficier des résultats fournis par les analyses que nous pratiquons sur les données de SymbioCyc.

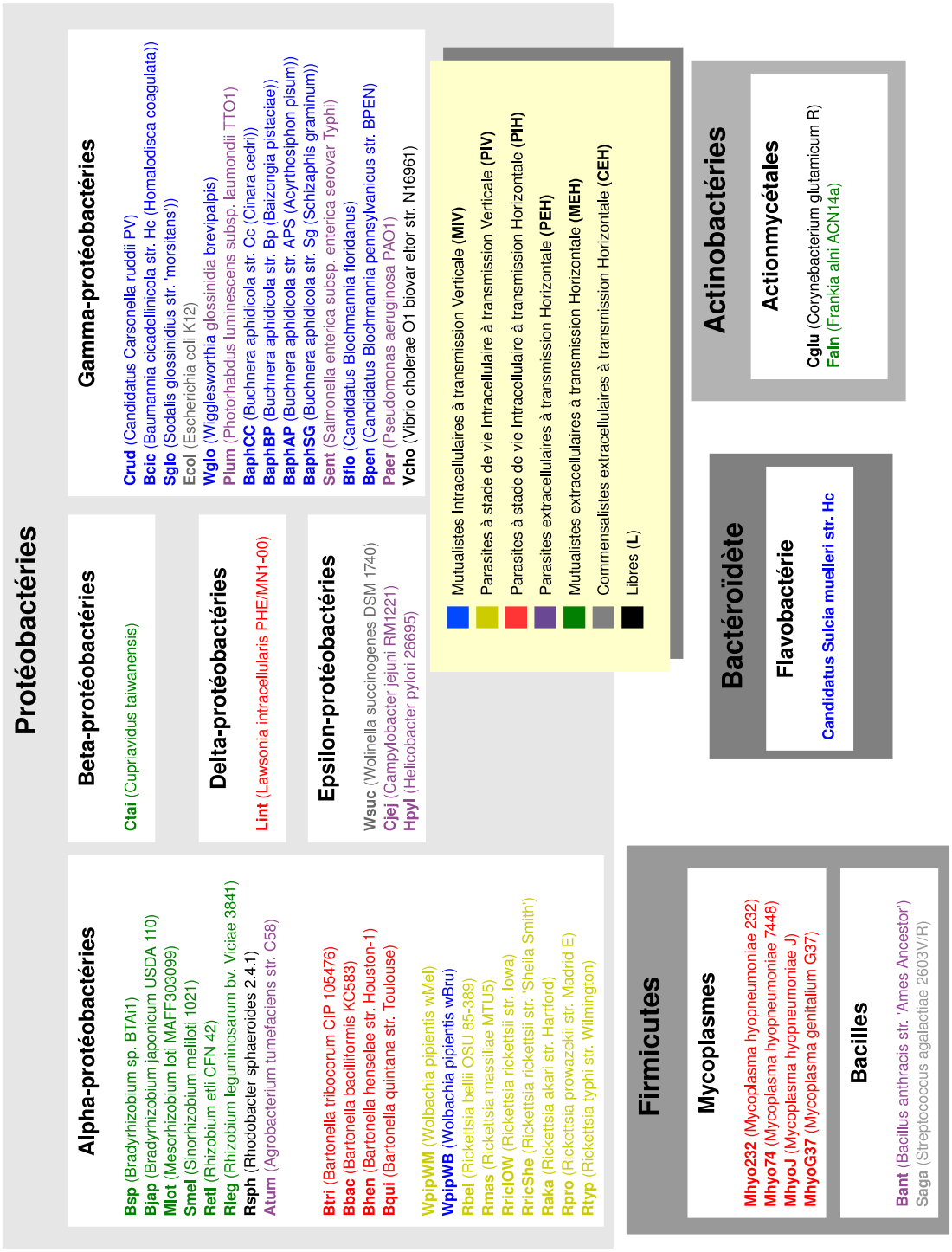

Figure 3.1. Classification générale des 49 organismes de SymbioCyc. Les couleurs correspondent aux styles de vie indiqués dans l'encadré de la figure.

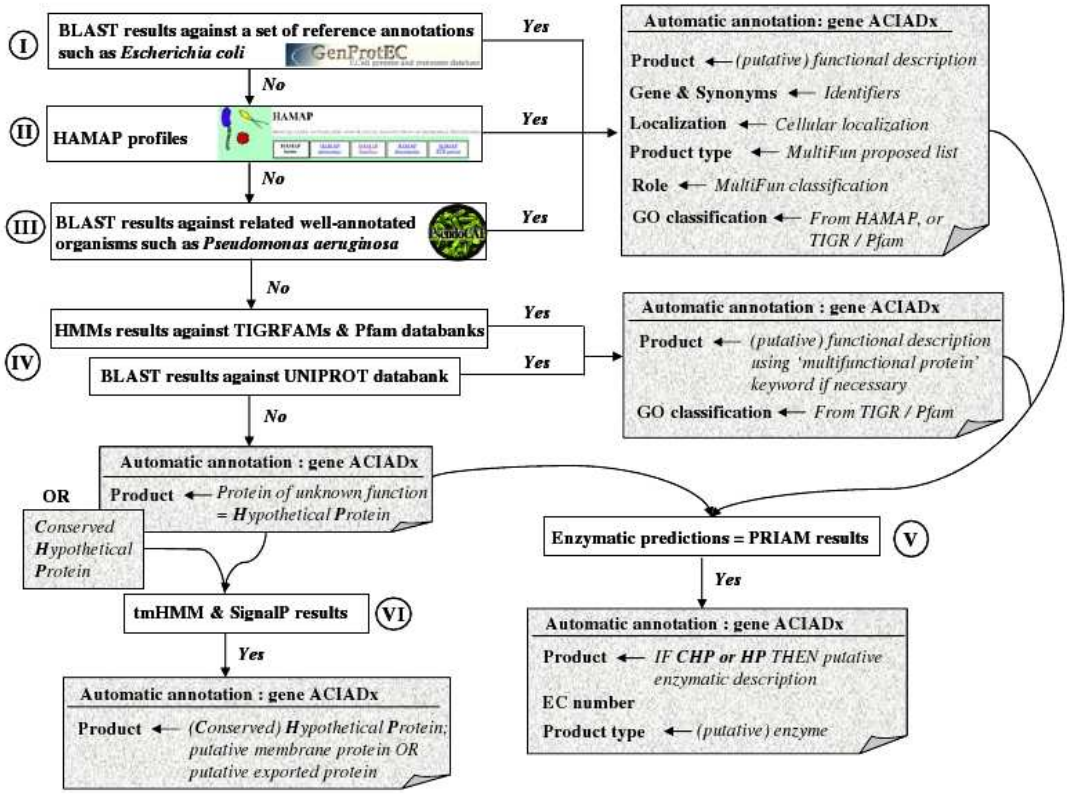

Figure 3.2. Procédure d'annotation automatique de MaGe utilisée pour le génome de *Acinetobacter baylyi* ADP1. Cette figure provient du matériel supplémentaire de l'article écrit par (Vallenet *et al.*, 2006).

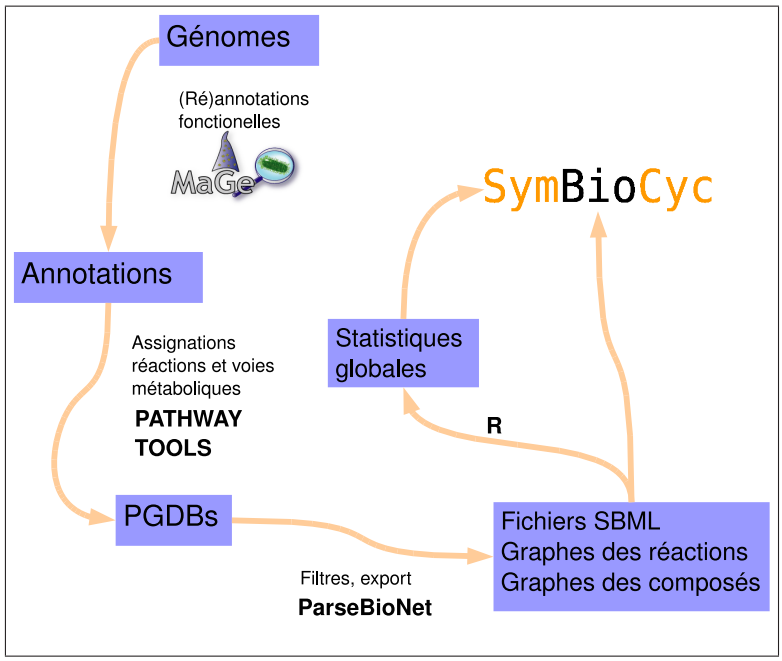

Figure 3.3. Schéma global de la construction de SymbioCyc.

3.3 Construction de SymbioCyc

Les données de SymbioCyc sont calculées à partir des données fournies par les Pathway-Tools. Chaque PGDB (Pathway Genome DataBase) provenant de MaGe est chargée sur une version locale des Pathway-Tools. Nous avons créé une librairie Java permettant d'importer, de filtrer et d'exporter les données des Pathway-Tools. Cette librairie est d'ores et déjà disponible à l'adresse : `http://biomserv.univ-lyon1.fr/baobab/parsebionet/`. Les statistiques générales ont été calculées par de simples routines R. Les données sont exportées sous forme de fichiers plats. Des pages dynamiques développées en PHP permettent ensuite à l'utilisateur d'explorer ces données (Figure 3.3).

Le site est hébergé sur le serveur du pôle bioinformatique lyonnais (PBIL) à l'adresse `http://pbil.univ-lyon1.fr/software/symbiocyc/`.

3.4 Les fonctionnalités de SymbioCyc

SymbioCyc donne un accès direct aux interfaces BioCyc de chaque reconstruction métabolique. Ces interfaces, communes à chaque reconstruction faite par les pathway-tools, permet une exploration à la fois des données génomiques et des données métaboliques (Caspi & Karp, 2007).

SymbioCyc complète les nombreuses fonctionnalités d'une interface BioCyc en s'orientant vers la modélisation et la description des données. De plus, un outil de comparaison original vient s'ajouter aux outils de comparaison déjà existants dans l'interface BioCyc.

SymbioCyc dispose de quatre grandes fonctionnalités :
- le filtre des données,
- la présentation des caractéristiques globales de chaque réseau métabolique,
- l'exportation des données sous forme de graphes et de fichiers SBML,
- la comparaison des réseaux entre organismes et entre groupes d'organismes.

3.4.1 Filtre des données

Avant de commencer une quelconque modélisation d'un réseau métabolique, il convient de filtrer les données pour éviter certains artefacts (cf Section 2.1).

SymbioCyc permet à l'utilisateur d'utiliser quatre filtres différents et surtout de les combiner entre eux en fonction de l'analyse qu'il va effectuer ensuite.

Le premier filtre, communément utilisé dans les analyses des réseaux métaboliques, est de ne prendre en compte que les réactions qui font intervenir des petites molécules, mettant de côté toutes celles qui ne font intervenir que des macromolécules entre elles.

Le second filtre est de retirer toutes les réactions faisant intervenir des composés génériques. Ces composés représentent en réalité des classes de composés qui peuvent intervenir indifféremment dans une réaction donnée. Cependant, pour certaines analyses de graphes notamment, ces composés et ces réactions peuvent conduire à des artefacts en produisant des composantes déconnectées du reste du réseau.

Le troisième filtre prend en compte l'information sur les voies métaboliques contenues dans les Pathway-Tools. En effet, pour chaque voie métabolique dans les pathway-tools, les composés principaux et les composés secondaires sont indiqués. Ce filtre procède de la façon suivante. Chaque réaction est considérée dans chacune des voies métaboliques où elle intervient. Si un composé est indiqué comme secondaire dans chacune de ces voies, il est retiré de la réaction. Comme cette information est contenue dans la description des voies métaboliques, le filtre élimine également toutes les réactions qui n'interviennent dans aucune voie métabolique. Ce filtre a déjà été notamment utilisé par (Lacroix *et al.*, 2006; Bourqui *et al.*, 2007).

Le quatrième filtre considère les cofacteurs intervenant classiquement dans les réactions métaboliques. La liste des cofacteurs s'inspire de celle indiquée par (Handorf *et al.*, 2007) (Tableau 3.1). A chaque fois qu'un couple substrat/produit de cofacteurs apparaît dans une réaction, il est éliminé. Prenons le cas de l'ATP et de l'ADP. La transformation de l'ATP en ADP intervient dans de nombreuses réactions auxquelles elle apporte l'énérgie nécessaire. Ainsi, dans chaque réaction où l'ATP intervient en tant que substrat et l'ADP en tant que produit, ces deux composés seront enlevés de la réaction. Dans la mesure du possible, nous éliminons aussi les sous-produits de cette transformation, par exemple le phosphate dans la transformation précédente.

3.4.2 Propriétés globales des réseaux

SymbioCyc permet de voir les propriétés globales de chaque réseau métabolique (Figure 3.4). A l'écran s'affichent le nombre de voies métaboliques, de composés, de réactions, de numéros EC différents, de gènes métaboliques et d'enzymes impliquées. Ces nombres vont nous renseigner immédiatement sur la taille du réseau et l'étendue de ses capacités métaboliques. A partir de la même interface, il est possible de télécharger un document pdf présentant d'autres statistiques qui vont permettre à l'utilisateur de caractériser plus finement le jeu de données étudié. Par exemple, la fréquence des 30 composés les plus connectés donne une idée de quels sont les composés importants dans le réseau et la distribution de la connectivité des composés renseigne sur la connectivité globale des composés. Il est possible d'appliquer les filtres précédemment décrits et de visualiser aussitôt les propriétés des réseaux filtrés.

3.4.3 Fichiers SBML, graphes de réactions, graphes de composés

Chaque réseau métabolique est téléchargeable sous la forme de fichier SBML, de graphes de composés ou de réactions. Le format SBML est maintenant utilisé par de nombreux outils de modélisation du réseau métabolique (cf Section 2.1).

Les fichiers présents dans SymbioCyc ne contiennent aucune information sur les quantités des composés ou les vitesses de réactions, informations non disponibles dans les reconstructions automatiques. Pour une analyse de balance des flux ou de contrôle métabolique, l'utilisateur doit ajouter manuellement ces informations. Un fichier SBML contient simplement une liste de réactions. Pour chacune de ces réactions, nous avons dans un fichier SBML classique les informations suivantes : la réversibilité, la liste des substrats et des produits et les coefficients stœchiométriques. En nous inspirant du format SBML amélioré utilisé par (Reed *et al.*, 2003), nous avons ajouté les liens gènes-protéines-réaction (GPR), les voies métaboliques dans lesquelles la réaction intervient et son numéro EC (Figure 3.5).

NAD+ + H+ ↔ NADH
NADPH ↔ NADP+ + H+
NAD(P)H ↔ NAD(P) + H+
ADP + Phosphate ↔ ATP
AMP + Diphosphate ↔ ATP
S-Adenosyl-L-homocysteine ↔ S-Adenosyl-L-methionine
CoA ↔ Acetyl-CoA
2-Oxoglutarate ↔ L-Glutamate
UDP ↔ UDP-glucose
Acceptor ↔ Reduced acceptor
CoA ↔ Acyl-CoA
Pyruvate ↔ L-Alanine
Succinate + CO2 ↔ 2-Oxoglutarate + O2
UDP ↔ UDP-N-acetyl-D-glucosamine
GDP + Phosphate ↔ GTP
Oxaloacetate ↔ L-Aspartate
UDP + Phosphate ↔ UDP-D-galactose
Adenosine 3',5'-bisphosphate ↔ 3'-Phosphoadenylyl-sulfate
UDP ↔ UTP
CoA ↔ Malonyl-CoA
CoA ↔ Succinyl-CoA
GDP ↔ GDP-mannose
CoA ↔ Propanoyl-CoA
IDP ↔ ITP
CDP ↔ CTP
CMP ↔ CMP-N-acetylneuraminate
Reduced ferredoxin ↔ Oxidized-ferredoxin
Tetrahydrofolate ↔ 5,10-Methylenetetrahydrofolate
CoA ↔ Palmitoyl-CoA
UDP ↔ UDP-glucuronate
UDP ↔ UDP-N-acetyl-D-galactosamine
dADP + Phosphate ↔ dATP
CoA ↔ p-Coumaroyl-CoA
Dihydrobiopterin ↔ Tetrahydrobiopterin
CoA ↔ Caffeoyl-CoA
Thioredoxin ↔ Oxidized thioredoxin
Ubiquinol ↔ Ubiquinone
Reduced rubredoxin ↔ Oxidized rubredoxin
CoA ↔ S-Benzoate coenzyme A
Reduced-adrenal-ferredoxin ↔ Oxidized adrenal ferredoxin
Coenzyme-F420 ↔ Reduced coenzyme F420
Dithiothreitol ↔ Oxidized dithiothreitol
FAD ↔ FADH2
PQQ ↔ PQQH2
FMN ↔ Reduced FMN
Donor ↔ Oxidized donor
UDP ↔ UDP-L-rhamnose
Deoxynucleoside ↔ Deoxynucleoside 5'-phosphate
CoA ↔ Methylmalonyl-CoA
Reduced-flavoprotein ↔ Oxidized flavoprotein
Protein-histidine ↔ Protein N(pi)-phospho-L-histidine
Electron-transferring flavoprotein ↔ Reduced electron transferring flavoprotein
Oxidized azurin ↔ Reduced azurin
dTDP ↔ dTDP-L-oleandrose
Quinone ↔ Hydroquinone
Ferricytochrome c ↔ Ferrocytochrome c
Protein dithiol ↔ Protein disulfide

Tableau 3.1. Liste des transformations impliquant des cofacteurs, utilisée dans un des filtres de SymbioCyc. Lorsque l'une de ces transformations est trouvée dans une réaction, le filtre élimine de la réaction les métabolites correspondants.

tb

Filters applied on these data : none Apply new filters

Organism	Life style	Pathways	Compounds	Enzymes	Distinct EC numbers	Reactions	Metabolic Genes	Reactions without EC number	PDF file of global characteristics
Agrobacterium tumefaciens C58	Extracellular pathogen of mammals	353	1250	1257	736	1053	1257	79	PDF
Bacillus anthracis Ames Ancestor	Extracellular pathogen of mammals	312	1069	1016	632	892	1016	55	PDF
Bartonella tribocorum CIP 105476	Facultative Intracellular pathogen of mammals	154	603	390	311	457	390	37	PDF
Bartonella henselae Houston-1	Facultative Intracellular pathogen of mammals	158	611	380	316	463	380	40	PDF
Bartonella bacilliformis KC583	Facultative Intracellular pathogen of mammals	138	572	347	285	422	347	40	PDF

Figure 3.4. Visualisation des caractéristiques globales des réseaux métaboliques dans SymbioCyc.

Les fichiers SBML sont directement lisibles par les logiciels de visualisation Cytoscape (Shannon *et al.*, 2003) et MetaViz (Bourqui *et al.*, 2007). Ce dernier logiciel, dont nous avons participé à la conception, prend en compte les informations sur les voies métaboliques contenues dans le fichier SBML pour améliorer la visualisation.

SymbioCyc permet également à l'utilisateur d'enregistrer les graphes correspondant à chaque réseau métabolique. Le graphe peut être dirigé ou non et les noeuds peuvent être, soit les réactions, soit les composés (voir Section 2.2).

Chaque graphe est enregistré sous la forme d'une liste d'arêtes au format sif, directement utilisable par Cytoscape et lisible facilement par de simples routines de programmation (Figure 3.6).

Les filtres décrits précédemment sont applicables, autant sur les fichiers SBML que sur les graphes de réactions et de composés avant leur téléchargement.

3.4.4 Comparaison des réseaux métaboliques

SymbioCyc propose un outil de comparaison de réseaux métaboliques original. Il est possible, en effet, de comparer les ensembles de composés, de réactions et de voies métaboliques entre organismes et entre groupes d'organismes. Cet outil vient compléter les outils de comparaison déjà présents dans les interfaces BioCyc. Ici, les réactions prises en compte sont seulement celles pour lesquelles une enzyme ou un gène a été assigné, au contraire de l'outil de comparaison global proposé par BioCyc qui prend en compte également les autres (voir Section 2.3.1).

```
<reaction id="GLUTATHIONE__45__PEROXIDASE__45__RXN" name="glutathione peroxidase" reversible="true">
  <notes>
    <html:p>GENE_ASSOCIATION: ( G0317145 )</html:p>
    <html:p>PROTEIN_ASSOCIATION: ( G0317145-MONOMER )</html:p>
    <html:p>SUBSYSTEM: glutathione redox reactions I</html:p>
    <html:p>PROTEIN_CLASS: 1.11.1.9</html:p>
  </notes>
  <listOfReactants>
    <speciesReference species="GLUTATHIONE" stoichiometry="2"/>
    <speciesReference species="HYDROGEN__45__PEROXIDE" stoichiometry="1"/>
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="OXIDIZED__45__GLUTATHIONE" stoichiometry="1"/>
    <speciesReference species="WATER" stoichiometry="2"/>
  </listOfProducts>
</reaction>
```

Figure 3.5. Elément "réaction" dans un fichier SBML. Les liens entre gènes, protéines et réactions, les voies métaboliques et le numéro EC ont été ajoutées en note.

La comparaison des voies métaboliques de plusieurs organismes choisis par l'utilisateur permet d'identifier les différences ou les similitudes globales dans les capacités métaboliques des organismes. La comparaison renvoie un tableau avec, pour chaque organisme sélectionné, les voies métaboliques qui lui sont propres et les voies métaboliques partagées entre tous les organismes sélectionnés.

Par rapport au mode de comparaison déjà présent dans l'interface BioCyc, la proportion du nombre de réactions possibles dans l'organisme est indiqué et les voies complètes sont mises en relief (Figure 3.7). Un clic sur une voie métabolique renvoie l'utilisateur sur l'outil de comparaison entre espèces d'une voie métabolique de l'interface BioCyc.

La comparaison des composés de plusieurs organismes renvoie plusieurs tableaux correspondant, pour chaque organisme, aux composés qui lui sont propres, ainsi qu'un tableau correspondant aux composés partagés par tous les organismes choisis. Pour chaque composé, apparaît un lien vers la page correspondante de l'interface BioCyc, son nom, son poids moléculaire, le nombre de réactions dans lequel il intervient en tant que substrat ou produit, et le nombre de voies métaboliques dans lequel il apparaît.

La comparaison des réactions de plusieurs organismes renvoie également plusieurs tableaux, les premiers correspondant aux réactions uniques à un organisme et le dernier aux réactions communes. Pour chaque réaction, apparaît un lien vers l'outil de comparaison d'une réaction de l'interface BioCyc, ses substrats, les enzymes impliquées et les voies métaboliques dans lesquelles on trouve cette réaction.

Le résultat des comparaisons des voies, composés ou réactions entre groupes d'organismes, est présenté de la même manière. Cette fois-ci, l'utilisateur peut préciser jusqu'à six ensembles d'organismes. Un objet (voie, composé ou réaction) sera considéré comme unique à un ensemble s'il apparaît systématiquement dans tous les organismes qu'il contient et s'il est absent de tous les autres organismes de

```
L-XYLULOSE	linkedWith	L-XYLULOSE-5-P
L-XYLULOSE	linkedWith	ADP
CARBON-MONOXIDE	linkedWith	|Co-E-Clth|
CARBON-MONOXIDE	linkedWith	ACETYL-COA
CPD-7224	linkedWith	L-CITRULLINE
CPD-7224	linkedWith	ACET
```

Figure 3.6. Extrait d'une liste d'arêtes dans un graphe des composés sous format sif.

chaque groupe. Ce mode de comparaison original permet de dégager rapidement les grandes différences métaboliques entre groupes d'organismes.

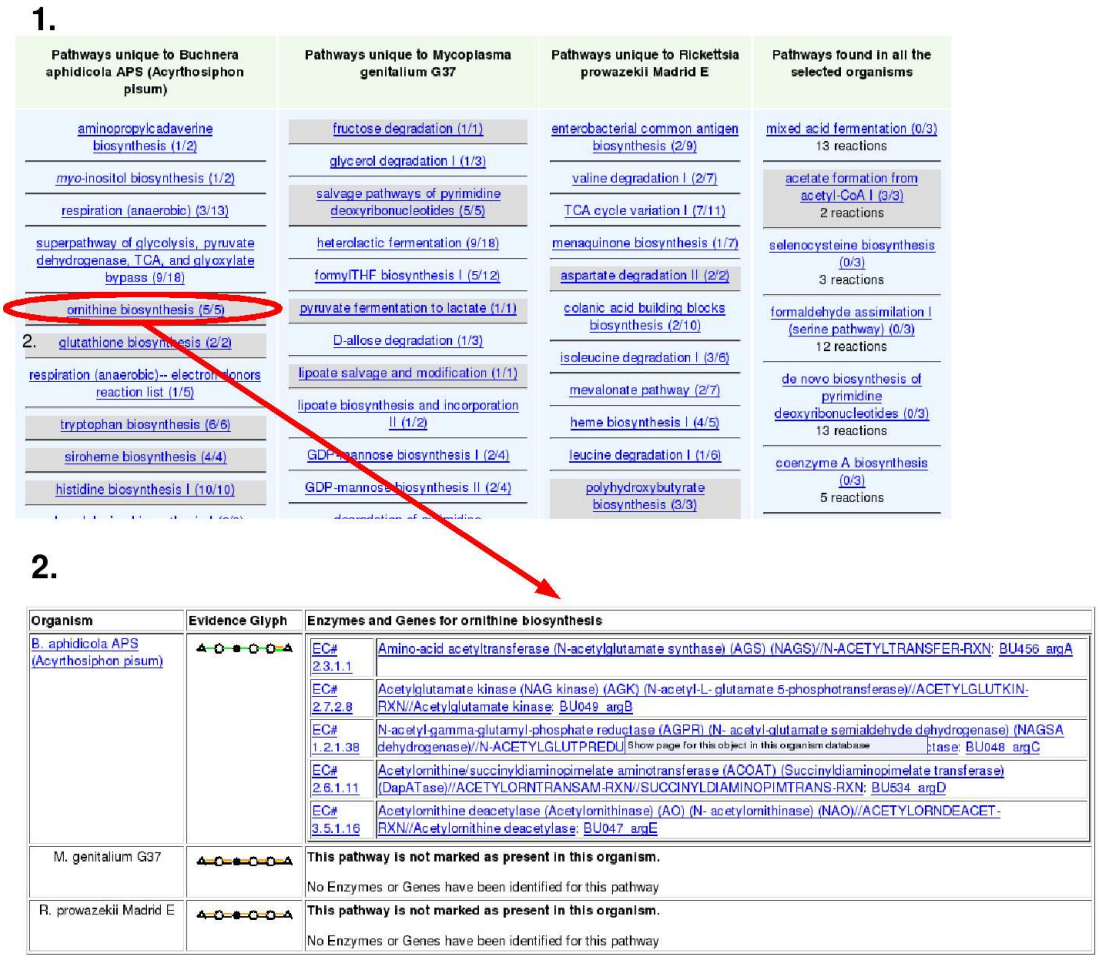

Figure 3.7. Extrait de la comparaison des voies métaboliques de *Rickettsia prowazekii*, *Mycoplasma genitalium* et *Buchnera aphidicola APS* dans SymbioCyc (1.). Un clic sur une voie métabolique renvoie sur le détail de la comparaison dans l'interface de BioCyc (2.)

3.5 Discussion

Les bases de données telles que BioCyc (Caspi & Karp, 2007) et KEGG (Aoki & Kanehisa, 2005) et leurs outils associés facilitent grandement l'exploration des données génomiques et métaboliques de centaines d'organismes. La mise à disposition de réseaux entiers a permis leur modélisation, sous forme de graphes notamment, et leur analyse globale. Cependant, aucun environnement à notre connaissance ne fait de façon satisfaisante le lien entre les données présentes dans les bases et les données utilisées pour la modélisation. SymbioCyc tente de répondre pour la première fois à ce besoin en mettant à disposition pour une cinquantaine de bactéries, les fichiers SBML et les différents types de graphes auxquels plusieurs filtres peuvent être appliqués, économisant ainsi le temps de programmation de l'utilisateur. Le calcul de statistiques globales permet en outre de décrire chaque jeu de données, étape indispensable avant toute modélisation. SymbioCyc également propose un outil de comparaison de réseaux métaboliques original permettant facilement de dégager les principales différences ou ressemblances métaboliques entre organismes et groupes d'organismes. Au-delà de son utilité dans le reste des travaux présentés ici, SymbioCyc est utilisée également dans d'autres travaux auxquels nous avons participés (Bourqui *et al.*, 2007; Picard *et al.*, 2008). Certaines fonctionnalités de SymbioCyc, peuvent facilement être exportées vers d'autres sites. C'est le cas notamment de certaines de ses fonctions de filtre présentes dans MOTUSWeb, projet auquel nous avons également participé (Lacroix *et al.*, 2008a).

Toutefois, plusieurs améliorations sont possibles. Afin de faciliter sa mise à jour et l'ajout de nouveaux organismes, il faudrait d'abord automatiser certaines tâches dans la construction du site. Une limitation de SymbioCyc est que toutes les données sont précalculées, ce qui ne permet pas à un utilisateur d'explorer et de filtrer ses propres données. Nous projetons ainsi de rendre publique une interface graphique dont l'utilisation serait locale pour que l'utilisateur puisse traiter n'importe quel réseau de la même façon que dans SymbioCyc. Ensuite, la visualisation des différents graphes pourrait être facilitée par l'intégration dans SymbioCyc d'un outil comme Cytoscape. Enfin, nous envisageons d'utiliser la visualisation des diagrammes de Venn proposée par la librairie Java Aduna pour améliorer la comparaison des réseaux métaboliques. On peut trouver des exemples de tels diagrammes dans la Section 4.

Utilisée intensivement durant cette thèse et dans quelques travaux auxquels nous sommes associés, nous espérons que SymbioCyc sera bientôt utilisée de façon beaucoup plus large et qu'elle connaîtra très vite de nouveaux développements.

Chapitre 4

Comparaison des réseaux métaboliques des bactéries intracellulaires en fonction de leur style de vie

4.1 Motivation

Le style de vie peut être considéré comme la somme des effets de l'environnement d'un organisme et des relations qu'il établit avec d'autres espèces (Cases *et al.*, 2003). Dans le cas de bactéries intracellulaires, les deux facteurs sont très liés puisque l'environnement correspond directement à la cellule de l'hôte. La vie intracellulaire s'accompagne de pressions de sélection spécifiques, agissant en particulier sur le réseau métabolique de l'hôte et de la bactérie (voir Section 1.2). Certains traits évolutifs, comme la réduction du réseau, sont communs à toutes les bactéries intracellulaires mais d'autres dépendent du type d'interactions entretenues avec l'hôte. Ainsi, chez les bactéries dont l'association avec l'hôte est nutritionnelle, certaines voies métaboliques nécessaires à l'association sont préservées tandis que d'autres, redondantes avec celles de l'hôte, disparaissent. Les reconstructions métaboliques permettent d'obtenir la liste des réactions et des voies métaboliques possibles pour un ensemble d'organismes. La manière la plus classique d'identifier et de comparer les capacités métaboliques d'un groupe d'organismes est de vérifier successivement la présence des voies métaboliques de référence dans chacun des réseaux métaboliques reconstruits.

Non seulement ce processus est coûteux en temps, ce qui limite le nombre d'organismes que l'on peut comparer, mais il limite aussi l'exploration du réseau à quelques voies métaboliques auxquelles on porte un intérêt selon des *a priori* biologiques. L'analyse globale des réseaux métaboliques permet, en gommant la partition en voies métaboliques, une exploration plus large et sans *a priori* des

capacités métaboliques d'un grand nombre d'organismes.

4.2 Les méthodes de comparaison de réseaux métaboliques

4.2.1 La comparaison des ensembles d'entités des réseaux métaboliques

Une manière de comparer les réseaux métaboliques de plusieurs organismes est de confronter les ensembles d'objets qui les composent. A ce jour, la comparaison s'est effectuée sur les ensembles d'enzymes (ou gènes métaboliques) et les ensembles de réactions.

La comparaison des réseaux métaboliques découle souvent de la comparaison des annotations fonctionelles de plusieurs organismes. Pour chaque grand processus métabolique (synthèse des acides aminés, synthèse des lipides, glycolyse...), on recueille les gènes correspondants annotés dans chaque organisme. De cette façon, Zientz *et al.* (2004) répertorient les différents effets de la vie symbiotique intracellulaire sur le métabolisme de trois endocytobiotes mutualistes *Buchnera*, *Blochmannia* et *Wigglesworthia*. De même, Tamas *et al.* (2001) ont montré les différences génomiques mais aussi métaboliques entre une bactérie intracellulaire mutualiste, *Buchnera*, et une bactérie intracellulaire parasite, *Rickettsia*. Ces deux études, quoique riches en informations, ne portent que sur deux ou trois organismes. De plus, l'étude ne se fait que sur quelques voies métaboliques choisies *a priori*. Enfin, une telle approche, basée sur les voies métaboliques de référence, ne permet pas d'appréhender des voies alternatives possibles ou une vision d'ensemble du réseau métabolique.

A l'opposé de ces analyses très précises, certains auteurs ont proposé des études à plus grande échelle pour mettre en évidence des tendances générales à travers de larges collections d'organismes.

Dans l'objectif de comparer le contenu en gènes des organismes séquencés, une centaine à cette époque, van Nimwegen (2003) a examiné la portion de gènes dédiée aux différentes tâches fonctionelles de la cellule. L'auteur indiqua notamment une proportion très conservée des gènes dédiés au métabolisme parmi les génomes analysés.

Dans le même temps, Cases *et al.* (2003) examinèrent la proportion de gènes dédiés à la régulation, au métabolisme et au transport de 60 organismes classés selon leur style de vie. Les auteurs indiquent ainsi une portion plus importante du génome des bactéries intracellulaires pathogènes dédié au métabolisme. Cependant, ce résultat souffre d'un manque de précision, en partie parce que sont classés parmi les intracellulaires pathogènes également les mutualistes tels que

Buchnera aphidicola APS, et d'autre part, par manque de détail sur l'influence de chaque organisme sur les tendances globales observées.

Aguilar *et al.* (2004) ont classé 27 organismes choisis dans les trois grands domaines du vivant (bactéries, archébactéries, eucaryotes) en fonction de leur ensemble d'enzymes, déduits de l'annotation génomique. Celles-ci ont d'abord été classées selon sept grandes classes fonctionnelles. Pour chaque organisme, un profil binaire par classe a été construit, indiquant pour chaque enzyme sa présence (1) ou son absence (0) dans le protéome. A partir de ces profils ont été construits des matrices de distance entre organismes qui ont servi ensuite à construire des arbres. Aguilar *et al.* trouvèrent que la position des organismes dans l'arbre réflétait les pressions de l'environnement et des adaptations spécifiques plutôt que leur classement phylogénétique. Par exemple, les trois bactéries pathogènes obligatoires de leur jeu de données sont groupées ensemble dans les sept arbres malgré leur grand éloignement phylogénétique.

Freilich *et al.* (2005) ont procédé également à une étude comparative des ensembles d'enzymes à travers les trois domaines du vivant. Leur premier objectif était de déterminer si la fraction d'enzymes dans le protéome est constante dans, et entre, chaque grand domaine du vivant. Seulement six bactéries ont montré une proportion d'enzymes significativement élevée et six autres une proportion d'enzymes significativement basse parmi les 85 organismes étudiés, ce qui tend à montrer une relative conservation de la fraction d'enzymes dans le protéome à travers les grands domaines du vivant, confirmant ainsi la tendance indiquée par van Nimwegen (2003).

Parmi les six bactéries ayant une proportion d'enzymes élevée, nous trouvons *Buchnera aphidicola APS*, deux mycoplasmes et une rickettsie. Les auteurs mettent en avant le caractère intracellulaire de ces quatre bactéries pour expliquer ce résultat. La portion élevée du protéome dédiée au métabolisme s'expliquerait par la perte massive des protéines régulatrices chez les bactéries intracellulaires. Cependant, parmi les 85 organismes étudiés, d'autres sont intracellulaires sans montrer une proportion plus élevée que la moyenne d'enzymes dans leur protéome, ce qui invalide l'argument de Freilich *et al.* pour expliquer la proportion d'enzymes élevées chez ces quatre bactéries.

Par ailleurs, en regardant de plus près les résultats de Freilich *et al.* (2005), on remarque que ce résultat n'est vraiment significatif que pour *Buchnera aphidicola APS*. En effet, pour chaque organisme, Freilich et ses collaborateurs ont construit deux ensembles d'enzymes en se référant à la collection d'enzymes de SWISS-PROT. Le premier, appelé "ensemble conservatif", attribue la fonction d'enzyme à une protéine si sa séquence est similaire à plus de 40% avec une enzyme de l'ensemble de référence. Le second, appelé "ensemble permissif", attribue la fonction d'enzyme à des homologues beaucoup plus distants, avec un pourcentage d'identité pouvant être inférieur à 20%. La proportion d'enzymes relativement élevée chez les six bactéries signalées par Freilich *et al.* a été calculée à partir des ensembles d'enzymes permissifs pour, selon les auteurs, éviter le biais

dû aux organismes modèles dans les ensembles conservatifs. Cependant, si l'on prend en compte les ensembles conservatifs, le résultat s'inverse pour les deux mycoplasmes (la proportion d'enzymes devient inférieure à la moyenne) et devient non significative pour la rickettsie. Par contre, la proportion reste significativement élevée pour *Buchnera aphidicola APS* et les deux autres protéobactéries extracellulaires signalées, *Haemophilus influenzae* et *Pasteurella multocida*.

Le second objectif de Freilich *et al.* était de préciser la nature de l'augmentation du nombre d'enzymes dans les différents domaines. Deux modes d'expansion ont été testés. Le premier relie l'élargissement de l'ensemble d'enzymes avec celui du répertoire de réactions, et le second se base sur l'apparition d'une redondance fonctionnelle des enzymes. Les réactions correspondent ici aux numéros EC assignés aux enzymes. Les auteurs calculent ensuite le ratio entre le nombre d'enzymes et le nombre de réactions. Les réactions sont ici déterminées par leur numéro EC. Freilich *et al.* montrent ainsi une dépendance linéaire entre le nombre d'enzymes et le nombre de réactions chez les procaryotes, absente chez les eucaryotes. Cependant, le faible nombre d'eucaryotes dans leur jeu de données pourrait expliquer ce résultat. Par ailleurs, le même calcul indiqua une faible redondance fonctionnelle des enzymes de *Buchnera aphidicola APS* en comparaison avec les autres organismes.

4.2.2 La comparaison des indices mesurés sur les graphes métaboliques

Nous avons déjà indiqué dans la Section 2.2.4, les principaux indices employés dans l'analyse de graphes. Depuis une dizaine d'années, ces mesures ont été utilisées intensivement pour caractériser les réseaux biologiques, en général, et les réseaux métaboliques en particulier.

Le concept de réseau "petit-monde", formalisé par Watts & Strogatz (1998) a été un des premiers à être appliqué sur les réseaux biologiques. Les réseaux "petit-monde" sont caractérisés par un diamètre faible, plus exactement par un diamètre proportionnel au logarithme du nombre de noeuds, ainsi que par un coefficient d'agglomération moyen élevé.

Wagner et Fell sont les premiers à avoir déclaré que le réseau métabolique de *Escherichia coli K12* pouvait être classé parmi les réseaux petit-monde, concluant que ce type de structure permettait un temps de transition très court entre différents états métaboliques (Fell & Wagner, 2000; Wagner & Fell, 2001). Arita (2004) contredit ces affirmations en suggérant que le modèle utilisé par Fell et Wagner n'est pas assez réaliste. Dans le modèle d'Arita, les métabolites eux-mêmes sont modélisés sous formes de graphes où les noeuds sont les atomes et les arêtes les liaisons chimiques. Les chemins dans le graphe sont calculés selon le passage d'atomes de carbone d'un métabolite à un autre, ce qui élimine des chemins entre des composés entre lesquels il n'y a pas d'échange de matière (Arita,

2000). Avec son modèle, Arita montre que le diamètre du réseau est beaucoup plus grand que celui estimé précédemment.

Une autre manière moins coûteuse en temps de calcul d'obtenir des distances plus réalistes entre métabolites est de pondérer les noeuds en fonction de leur degré et de donner ainsi une préférence aux chemins passant par les noeuds les moins connectés, les noeuds les plus connectés intervenant le plus souvent sous forme de cofacteurs et non en tant que source de carbone dans les réactions biochimiques (Croes *et al.*, 2006).

Une autre notion populaire pour décrire les réseaux biologiques est celle des réseaux "sans-échelle" (scale-free). Ce terme est introduit par Barabási & Albert (1999) pour qualifier de nombreux réseaux dont la distribution des degrés suit une loi de puissance, plutôt qu'une loi de Poisson telle qu'attendue dans le cas d'un réseau aléatoire. Ces réseaux sans-échelle incluent ainsi les réseaux génétiques, les réseaux de neurones, les réseaux sociaux et internet. Quelque temps plus tard, les réseaux métaboliques eux-mêmes sont qualifiés de réseaux sans-échelle (Jeong *et al.*, 2000). Une des principales caractéristiques d'un réseau sans-échelle est la présence d'un grand nombre de noeuds peu connectés et d'un nombre faible de noeuds très connectés. Dans le cas de graphes des composés, les premiers sont les composés qui interviennent dans très peu de réactions et les autres sont les composés ubiquitaires dont nous avons parlé précédemment. Les réseaux sans-échelle possèdent également les propriétés des réseaux petit-monde. Barabási & Albert (1999) proposent un modèle d'évolution pour ces réseaux où les noeuds apparaissant *de novo* dans le réseau se lieraient préférentiellement aux noeuds déjà très connectés. Récemment pourtant, ont vu le jour plusieurs articles destinés à mettre en doute la pertinence de ce concept pour les réseaux biologiques.

Le premier argument avancé est la qualité des données utilisées pour créer ces réseaux, particulièrement les réseaux protéine-protéine (Aloy & Russell, 2002; Coulomb *et al.*, 2005) mais on peut étendre facilement ces constatations aux réseaux métaboliques. En effet, les lacunes, imprécisions et erreurs possibles dans les données peuvent avoir un effet considérable sur ce genre de mesures globales. Le deuxième argument est méthodologique et remet en question le fait même que la distribution des degrés des réseaux étudiés suive réellement une loi de puissance (Khanin & Wit, 2006; Stumpf *et al.*, 2005). En effet, le test utilisé dans les études précédentes pour considérer si la distribution suit une loi de puissance est seulement de tracer une ligne dans la représentation log-log de la distribution. Khanin & Wit (2006) montrent ainsi qu'en appliquant des tests plus robustes, plusieurs réseaux ne peuvent plus être considérés comme "sans-échelle". Ils montrent en outre que plusieurs autres types de distributions pourraient correspondre aux propriétés des réseaux biologiques.

Le troisième argument fait intervenir la nature même des noeuds dans un réseau métabolique. En effet, la distribution des degrés comme d'autres mesures globales considèrent tous les noeuds comme équivalents. Cependant, chaque noeud dans un graphe métabolique, que ce noeud représente un métabolite ou une réac-

tion, a des propriétés spécifiques qu'il est important de prendre en compte. C'est le cas des cofacteurs n'intervenant pas dans l'échange d'atomes de carbone et des réactions les utilisant.

Enfin, l'argument final est certainement celui de Keller (2005) : même si les réseaux biologiques étaient reconnus comme étant des réseaux sans-échelle, cette propriété paraît tellement générale quel que soit le réseau qu'on ne peut en dégager une quelconque interprétation biologique.

Au-delà des concepts "petit-monde" et "sans-échelle", un certain nombre d'autres études ont utilisé les mesures sur les graphes afin de décrire les réseaux métaboliques et de les comparer entre eux.

Ma & Zeng (2003) mesurent la distribution des degrés, la longueur des chemins et le diamètre sur 80 graphes métaboliques, mettant ainsi en évidence un diamètre plus élevé chez les eucaryotes et les procaryotes que chez les archébactéries. Par ailleurs, à la différence d'études antérieures, Ma *et al.* proposent quelques règles heuristiques pour déterminer le sens des réactions et éliminent les composés ubiquitaires dans les réactions où ils sont considérés comme secondaires.

Rahman & Schomburg (2006) utilisent une mesure très proche de la centralité d'interposition (betweenness centrality) pour repérer les métabolites et les réactions qui peuvent être considérés comme des "points chauds" du réseau. La comparaison de cette mesure sur deux espèces proches phylogénétiquement dont une seule est pathogène, permet de proposer des cibles thérapeutiques. Afin de comparer les réseaux métaboliques des chloroplastes avec ceux de bactéries photosynthétiques, Wang *et al.* (2006) mesurent leurs connectivité moyenne, coefficient d'aggrégation moyen, longueur moyenne des chemins, diamètre et modularité.

Cette étude révéle un diamètre et une longueur moyenne des chemins plus grands chez le chloroplaste. Les auteurs déclarent utiliser une modélisation en hypergraphe des réseaux métaboliques, ce qui devrait éliminer les artefacts dus aux composés ubiquitaires durant le calcul de chemins entre deux noeuds. Cependant, dans cette analyse, le calcul des chemins ne semble tenir aucun compte de la structure en hypergraphe, les auteurs signalant par ailleurs l'élimination des cofacteurs des réactions où ils interviennent.

Liu *et al.* (2007) proposent de relier l'importance topologique des enzymes dans les graphes d'enzymes avec leur conservation phylogénétique. Ils montrent ainsi un lien entre les profils phylogénétiques des enzymes et le degré et la centralité d'interposition, mais aucun lien avec la centralité de proximité. Cependant, la construction du graphe ne tenait aucun compte de la direction des réactions ni de la présence des composés ubiquitaires, ce qui rend moins pertinents ces résultats. Récemment, Parter *et al.* (2007) ont calculé la modularité de réseaux métaboliques de plusieurs organismes répartis en six classes selon leur environnement (hôte-obligatoire, extrême, aquatique, hôte-facultatif, multiple et terrestre). Un module est ici considéré comme un groupe de noeuds hautement connectés entre eux et peu avec les autres. La mesure de modularité utilisée par Parter

et al. (2007) est définie comme la proportion du nombre d'arêtes à l'intérieur d'un module sur celui entre modules.

Parter *et al.* (2007) indiquent ainsi une modularité différente des graphes de composés selon les environnements des bactéries. Cependant, la construction des graphes de composés est ici critiquable. Les auteurs indiquent que les métabolites hautement connectés sont enlevés du réseau sans en préciser le nombre, ni s'ils sont éliminés de toutes les réactions ou seulement de celles où ils interviennent en tant que cofacteurs. De plus, les réactions sont toutes considérées comme réversibles.

4.3 Objectifs

Notre objectif global est de déterminer les liens entre style de vie d'un coté et composition et topologie du réseau métabolique d'un autre côté. Nous essaierons de déterminer les traits communs et ceux divergents entre les groupes de style de vie et au sein de ces derniers. Nous nous focaliserons essentiellement sur les bactéries intracellulaires, utilisant seulement les bactéries extracellulaires comme élements de comparaison.

Tout d'abord, la comparaison des ensembles de gènes, de composés et de réactions va nous permettre d'aborder les questions suivantes.

- L'adaptation à certains styles de vie s'accompagne-t-elle d'une part différente du génome consacrée au métabolisme ?
- La réduction du génome s'accompagne-t-elle d'une conservation plus importante des enzymes multifonctionnelles ?
- Comment le style de vie se reflète-t-il sur le nombre de métabolites et sur le nombre de réactions intervenant dans le métabolisme des petites molécules ?
- Quelles sont les portions de métabolites et de réactions conservées parmi l'ensemble des bactéries et parmi les différents groupes de style de vie ?
- Comment le style de vie se reflète-t-il sur la redondance des réactions chimiques des différentes bactéries ?
- Comment le style de vie se reflète-t-il sur la nature des métabolites intervenant dans les réseaux métaboliques ?
- Quelles sont les fonctions métaboliques communes et divergentes à travers l'ensemble des bactéries et parmi les différents groupes de bactéries intracellulaires ?

Dans un deuxième temps, nous allons utiliser la modélisation des graphes métaboliques sous forme de graphes des composés pour aborder les questions suivantes :

- Le style de vie se reflète-t-il sur l'accessibilité moyenne des métabolites ?
- Comment le style de vie se reflète-t-il sur l'utilisation et sur la production des métabolites ?

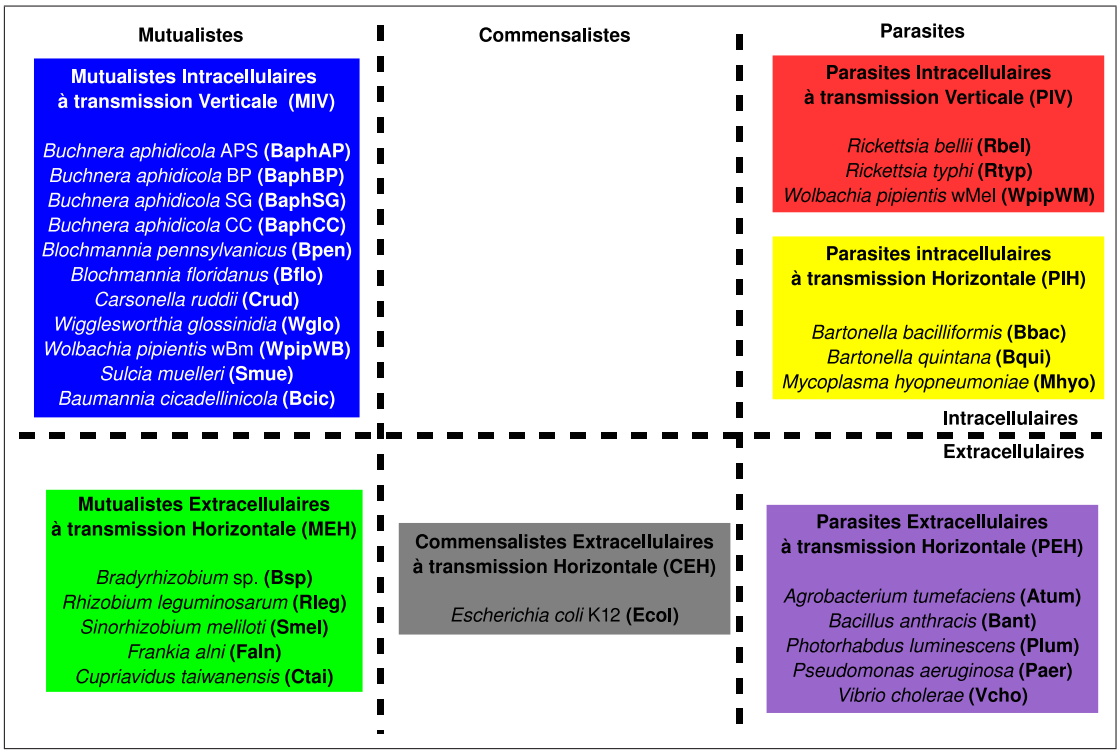

Figure 4.1. Style de vie des bactéries utilisées pour l'analyse comparative. Les abbréviations entre parenthèses seront utilisées dans les illustrations suivantes.

– Comment le style de vie se reflète-t-il sur la centralité de certains métabolites dans le réseau métabolique ?

Les précédentes comparaisons de réseaux métaboliques offrent des analyses très précises sur quelques organismes ou, au contraire, des analyses à grande échelle mais manquant de précision et d'interprétation. Nous essaierons dans notre travail de nous aider des analyses globales afin de repérer les différences ou les similitudes majeures entre les différents réseaux métaboliques, pour ensuite nous focaliser sur certains éléments (métabolites ou réactions) des réseaux. Nous espérons ainsi obtenir une analyse la plus précise possible et proposer une interprétation biologique pertinente.

4.4 Méthodes

Vingt neuf réseaux métaboliques de bactéries ont été sélectionnés dans SymbioCyc (Section 3). Nous n'avons pas analysé l'ensemble des bactéries présentes dans SymbioCyc afin de ne pas trop déséquilibrer les nombres des différents groupes de bactéries. Parmi ces 29 bactéries, 18 sont intracellulaires et 11 sont extracellulaires (Figure 4.1). Ces dernières vont essentiellement nous permettre de distinguer les singularités du métabolisme provenant de la vie intracellulaire.

Nous ne nous sommes intéressés qu'au métabolisme des petites molécules. Les réactions faisant intervenir des macromolécules ne sont donc pas prises en compte. Parmi ces dernières figurent les réactions de synthèse et de modifications de protéines, de modifications de l'ADN, de transcription, etc... Les informations sur les gènes, les réactions et les métabolites ont été extraites des données présentes dans SymbioCyc. Nous nous sommes basés sur les identifiants BioCyc pour faire les comparaisons entre ensembles. Chaque réaction est envisagée dans son ensemble et n'est pas divisée si plusieurs sous-réactions la composent. Ainsi, la décomposition de l'ATP en ADP et en phosphate, présente dans de nombreuses réactions, ne sera pas considérée comme une réaction à part entière. Les intersections entre ensembles de métabolites ou de réactions ont été calculées sous R. La visualisation des intersections a été réalisée avec l'outil de visualisation de diagrammes de Venn d'Aduna[1]. Nous avons utilisé l'outil de comparaison entre groupes d'organismes de SymbioCyc pour explorer et interpréter les résultats. Les représentations graphiques des tableaux de données et les analyses des correspondances multiples ont été réalisées sous R grâce à la librairie ade4 (Dray & Dufour, 2007).

Les graphes des composés proviennent des données de SymbioCyc. Les cofacteurs ont été filtrés et la direction des réactions assignée selon les méthodes décrites dans la Section 3.4.1 p.74. Les mesures sur les graphes ont été réalisées sous R grâce à la librairie igraph (Csardi & Nepusz, 2006).

[1] http ://www.aduna-software.com/technologies/clustermap/overview.view

4.5 Comparaison des réseaux métaboliques par les ensembles d'éléments qui les constituent

4.5.1 Comparaison globale des ensembles de gènes

Le nombre de gènes présents parmi les 29 bactéries est très variable (Figure 4.2), allant de 214 gènes pour *Carsonella ruddii* (Crud) à 7612 gènes pour *Bradyrhizobium* (Bsp). On remarque que ce sont les bactéries extracellulaires qui ont nettement le plus de gènes. Ainsi, la bactérie intracellulaire, *Bartonella henselae* (Bhen), avec le plus de gènes n'en posséde que 1429 alors que la bactérie extracellulaire, *Vibrio cholerae* (Vcho), avec le moins de gènes en possède tout de même 4082. Le nombre de gènes ne semble pas lié à la phylogénie des bactéries : les *Bartonella* (Bhen, Bqui et Bbac) et les Rickettsies (Rbel et Rtyp) ont un nombre de gènes beaucoup plus faible que d'autres α-protéobactéries extracellulaires comme *Bradyrhizobium* (Bsp) et *Rhizobium leguminosarum* (Rleg). On peut faire la même remarque pour les γ-protéobactéries.

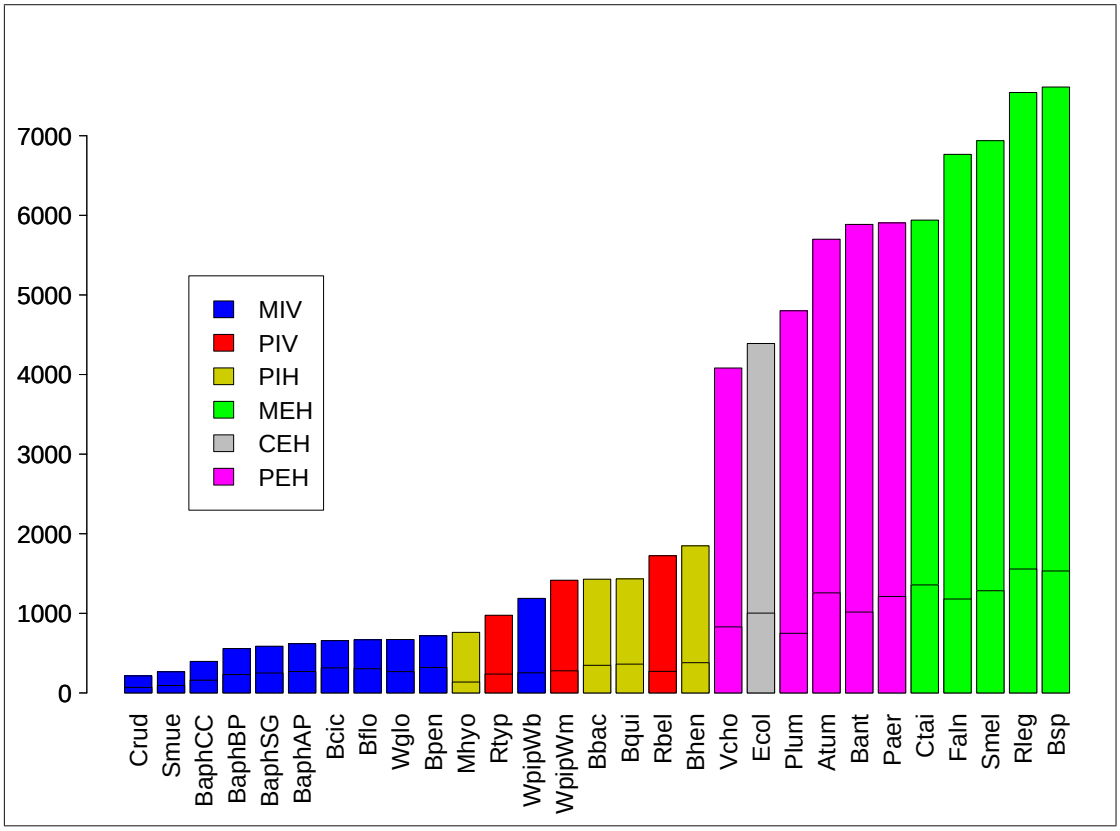

Figure 4.2. Nombre total de gènes et nombre de gènes métaboliques (barres horizontales inférieures). Les abbréviations des noms d'espèces et des styles de vie correspondent à celles données dans la Figure 4.1 p.90.

On retrouve bien ici l'effet de réduction du génome dû à la vie intracellulaire des bactéries décrit dans la Section 1.2. On peut voir également sur la Figure 4.2

que le nombre de gènes métaboliques suit la même tendance que le nombre total de gènes. Cependant, si on calcule la proportion du nombre de gènes métaboliques sur le nombre total de gènes, on observe des différences remarquables selon les organismes (Figure 4.3). Pour 19 de ces bactéries, cette proportion se situe aux alentours de 20 %. Par contre, pour les dix bactéries les plus pauvres en gènes, cette proportion dépasse largement les 30 % et atteint même près de 50 % pour *Baumannia cicadellinicola*. Ces dix bactéries correspondent aux bactéries les plus intégrées (voir définition dans la Section 1.2.2) et dont l'association avec leur hôte est nutritionnelle.

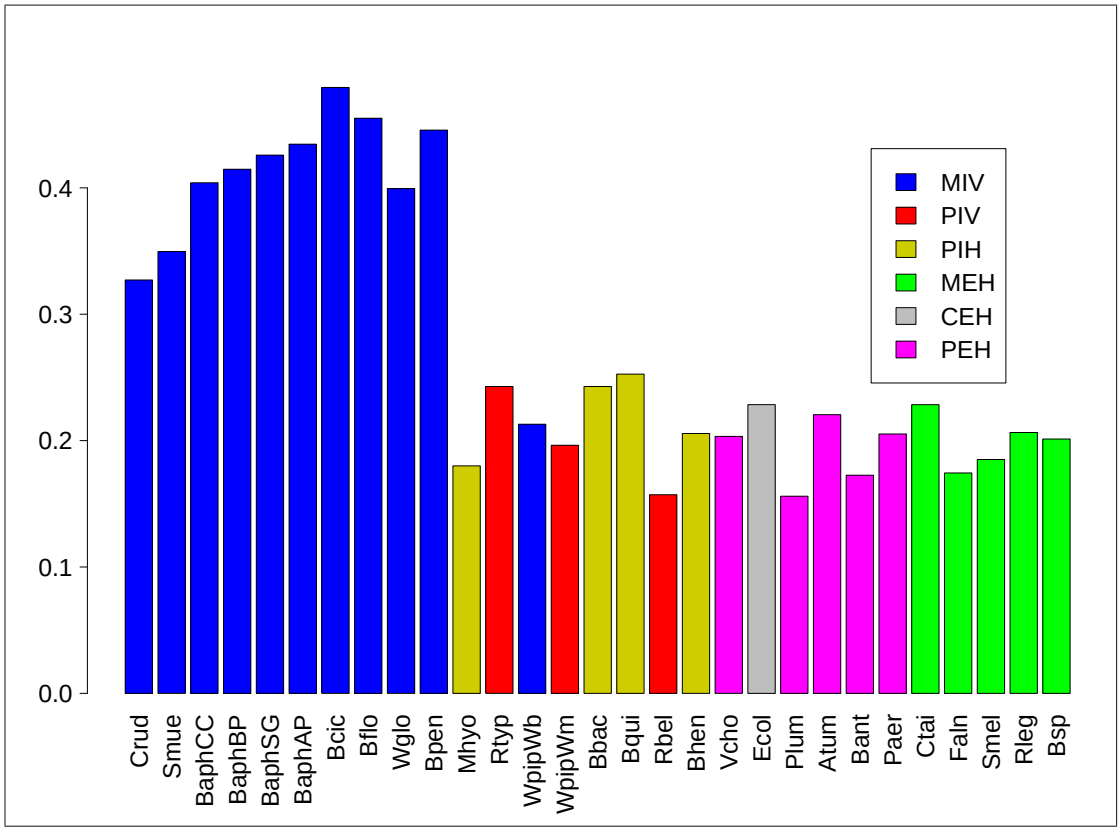

Figure 4.3. Proportion du nombre de gènes métaboliques sur le nombre total de gènes. Les abbréviations des noms d'espèces et des styles de vie correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont classés par nombre de gènes croissant.

Comme nous l'avons vu dans la Section 1.2, ces bactéries ont en effet perdu de nombreux gènes codant pour des fonctions non métaboliques telles que la virulence, l'adaptation aux conditions extrêmes, l'excrétion, etc. Par ailleurs, certaines fonctions métaboliques ont également disparu mais la conservation de celles associées à la symbiose expliquent aussi la grande proportion de gènes métaboliques chez ces bactéries.

On peut remarquer également que la bactérie endocytobiote mutualiste du nématode *Brugia malayi*, *Wolbachia pipientis wBm* (WpipWb), ne fait pas partie

de ces dix bactéries. Le caractère mutualiste et obligatoire de cette bactérie pour son hôte a pourtant été montré (Taylor *et al.*, 2005). Cependant, on peut imaginer que le caractère obligatoire de cette association est plus récent que chez *Buchnera aphidicola*, par exemple. En effet, au contraire de *Buchnera aphidicola* chez les pucerons, *Wolbachia pipientis wBm* a été perdue chez certains nématodes (Taylor *et al.*, 2005; Fenn *et al.*, 2006).

Dans le même sens, on peut voir que la proportion de gènes métaboliques est plus élevée chez *Wolbachia pipientis wBm* que chez *Wolbachia pipientis wMel*, indiquant chez la première une partie plus grande du génome réservée au gènes métaboliques, comme l'indiquent Foster *et al.* (2005). La lignée mutualiste des *Wolbachia* proviendrait en effet d'une lignée parasite (Fenn *et al.*, 2006). Cette plus grande proportion de gènes métaboliques pourrait ainsi marquer une évolution du parasitisme au mutualisme.

En revanche, on s'attendrait à ce que la proportion de gènes métaboliques soit moindre chez les espèces parasites, celles-ci profitant des capacités métaboliques de leur hôte, mais elle n'est pas significativement différente de celles des autres bactéries[2].

[2]Test de Wilcoxon, H_0 : les distributions des proportions de gènes métaboliques chez les parasites et chez les bactéries libres est identique. H_1 : Les deux distributions sont identiques. Niveau du test à $\alpha = 5\%$, P-valeur = 0,42.

4.5.2 Comparaison globale des ensembles de métabolites

a. Comparaison du nombre de métabolites

Comme le nombre de gènes, le nombre de composés intervenant dans le réseau métabolique est très variable selon les espèces (Figure 4.4). Il s'étend de 137 pour *Sulcia muelleri* (Smue) à 1177 pour *Rhizobium leguminosarum* (Rleg). Il faut préciser que ces composés ne correspondent pas nécessairement à ceux que l'organisme produit, certains pouvant intervenir seulement comme substrats. On peut considérer le nombre de métabolites comme une mesure de la diversité métabolique d'un organisme.

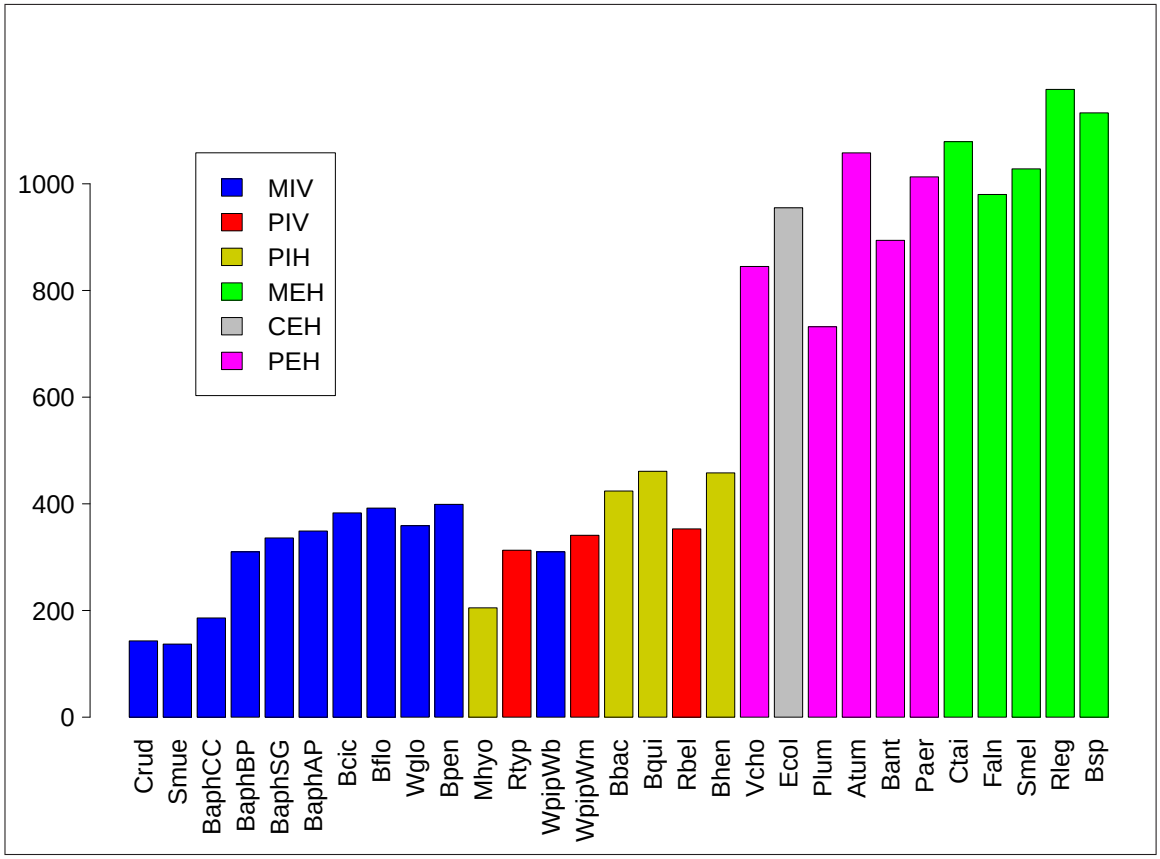

Figure 4.4. Nombre de composés (hors macromolécules) intervenant dans les réseaux métaboliques. Les abbréviations des noms d'espèces et des styles de vie correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont classés par nombre de gènes croissant.

b. Intersections entre les ensembles de composés

Seuls 28 composés, parmi les 1820 identifiés au total, sont communs aux 29 réseaux métaboliques analysés (Figure 4.5). Si l'on retire les classes génériques identifiées parmi ces 28 composés (bas du tableau), ce nombre se réduit à 24.

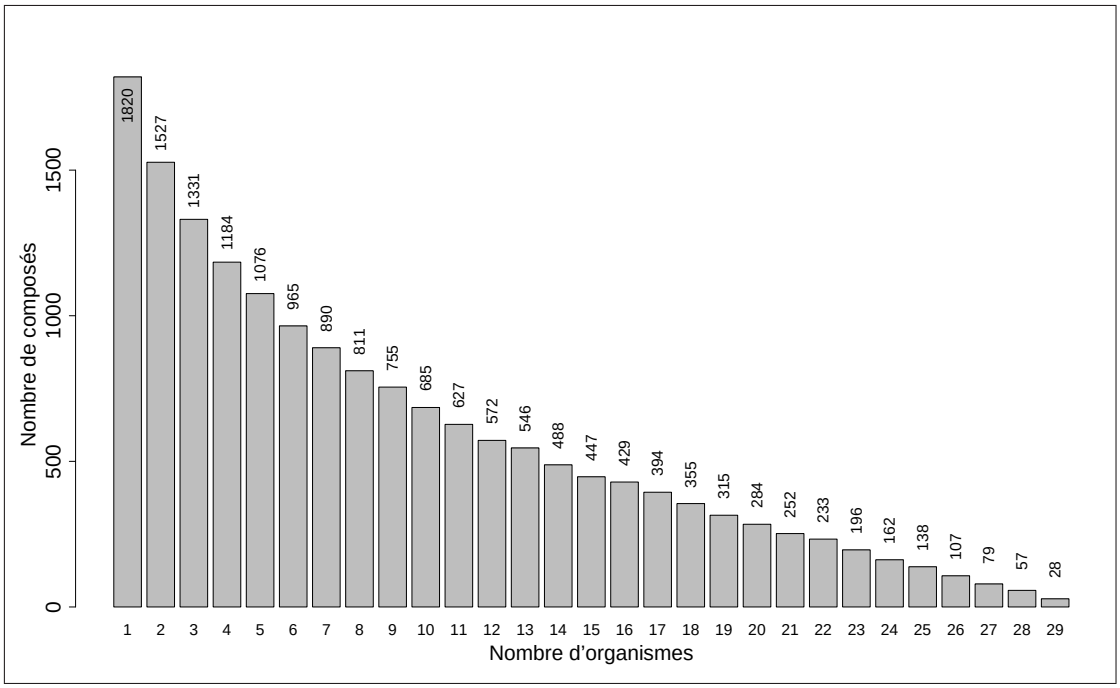

Figure 4.5. Nombre de composés communs (hors macromolécules) entre n organismes

Leurs attributs sont présentés dans le Tableau 4.1. La plupart sont des cofacteurs mais on trouve aussi 7 acides aminés et des métabolites intervenant dans la synthèse des acides nucléiques ou des constituants cellulaires.

Ce nombre, considérablement faible par rapport au nombre de composés trouvés dans un organisme, est directement expliqué par la réduction du métabolisme chez les bactéries intracellulaires. En effet, ce nombre correspond déjà à l'intersection du nombre de composés trouvés en ne considérant que les bactéries intracellulaires alors qu'il est de 393 lorsque l'on ne compare que les 11 bactéries extracellulaires (Tableau 4.2). Ces 393 métabolites correspondent au coeur du métabolisme, c'est-à-dire principalement aux voies complètes de synthèse des acides aminés, des nucléotides, de nombreux cofacteurs, de la glycolyse et du cycle de Krebs.

Dans le cas des bactéries intracellulaires, nous verrons qu'une portion ou la totalité de certaines voies peut disparaître, les fonctions métaboliques correspondantes étant prises en charge par l'hôte. Par ailleurs, l'intersection très faible s'accompagne d'une diversité relativement grande des métabolites chez les bactéries intracellulaires puisque l'on identifie 756 métabolites différents dans l'ensemble de

ces bactéries. L'intersection des ensembles de métabolites représente ainsi moins de 4% de leur union alors que la même proportion représente près de 22% chez les bactéries extracelullaires.

	MIV	PIV	PIH	**Intra**	MEH	PEH	CEH	**Extra**	**Total**
Moyenne	300	335	387	**325**	1079	908	955	**990**	**578**
Union	595	470	541	**756**	1610	1460	955	**1806**	**1820**
Intersection	40	220	132	**28**	626	459	955	**393**	**28**

Tableau 4.2. Taille moyenne, taille des unions et des intersections des lots de composés dans les différents groupes de style de vie. Les abbréviations des styles de vie correspondent à celles données dans la Figure 4.1 p.90. Le groupe des commensalistes (CEH) ne comporte qu'un organisme, *Escherichia coli K12*, l'intersection représente donc le nombre de composés dans cet organisme.

Ces nombres signifient que les réductions du métabolisme ont atteint des fonctions métaboliques différentes selon les bactéries. Cette constatation est particulièrement vérifiée chez les bactéries mutualistes intracellulaires à transmission verticale.

Eau
CO_2
Ion bicarbonate
Proton
ATP
ADP
AMP
NAD
NADH
Phosphate
Diphosphate
Pyruvate
Phospho-énol-pyruvate
D-Ribose-5-Phosphate
Acetyl-Coa
CoEnzyme A
Formate
Cystéine
Glutamate
Glutamine
Méthionine
Sérine
Alanine
Aspartate
Classe générique pour NAD ou NADP
Classe générique pour NADPH ou NADH
Classe générique pour un accepteur
Classe générique pour un donneur

Tableau 4.1. Composés communs (hors macromolécules) intervenant dans les 29 réseaux métaboliques.

c. Représentation des classes de métabolites dans les réseaux métaboliques

Pour préciser de quelle façon diffèrent les ensembles de métabolites, nous avons examiné, dans le réseau de chacun des 29 organismes, la représentation de 6 grandes classes de métabolites : les composés annotés comme cofacteurs, les acides aminés, les vitamines, les acides nucléiques, les sucres, et les lipides (Figure 4.6).

La première constatation est que la réduction du métabolisme chez les bactéries intracellulaires touche les six catégories. Ce trait est moins prononcé en ce qui concerne les lipides puisque les Rickettsies (Rbel et Rtyp) et les *Bartonella* (Bbac, Bqui et Bhen) ont un nombre de lipides intervenant dans leur réseau plus élevé que celui de la bactérie libre *Vibrio cholerae* par exemple. La seconde constatation est le profil extrême de certaines bactéries intracellulaires à l'intérieur même d'un style de vie donné, alors que les profils semblent plus réguliers en ce qui concerne les bactéries extracellulaires. Si l'on considère les vitamines, on observe que la bactérie mutualiste *Sulcia muelleri* (Smue) compte seulement une vitamine (la vitamine K2 ou ménaquinone) intervenant dans son réseau métabolique alors que les autres mutualistes intracellulaires en comptent 16 en moyenne. McCutcheon & Moran (2007) ont émis l'hypothèse que les vitamines pourraient être fournies à *Sulcia muelleri* par l'autre symbiote qui partage le même hôte, *Baumannia cicadellinicola* (Bcic). Les mêmes auteurs émettent l'hypothèse d'un approvisionnement par *Sulcia muelleri* de ménaquinone vers *Baumannia cicadellinicola* et leur hôte. Le nombre de sucres intervenant dans les réseaux métaboliques des bactéries intracellulaires semble dans l'ensemble assez constant, mis à part pour les trois espèces au nombre de gènes le plus faible : *Carsonella ruddii* (Crud), *Sulcia muelleri* (Smue) et *Buchnera aphidicola Cc* (BaphCC). Celles-ci ont en effet un nombre de sucres intervenant dans leur réseau métabolique de, respectivement, 6, 3 et 10 alors que le nombre de sucres chez les 15 autres bactéries intracellulaires s'étend de 22 à 39, avec une moyenne de 32.

Nous retrouvons la même tendance à propos de ces trois bactéries par rapport aux autres intracellulaires en ce qui concerne la représentation des acides nucléiques. Ces trois bactéries, de même que les trois autres *Buchnera* (BaphBP, BaphSG et BaphAP) ont une représentation des lipides considérablement faible si on la compare avec celle des autres bactéries intracellulaires. Aucun lipide n'apparaît dans les réseaux métaboliques de *Sulcia muelleri* et *Buchnera aphidicola Cc* et seulement un lipide apparaît dans ceux de *Carsonella ruddii*, *Buchnera aphidicola Bp*, *Buchnera aphidicola Sg* et *Buchnera aphidicola APS*. Le nombre de lipides dans les réseaux métaboliques des 12 autres bactéries intracellulaires s'étend de 4 à 11, avec une moyenne de 6. Cette observation coïncide avec la dégradation connue chez ces bactéries de la synthèse des membranes cellulaires (Zientz *et al.*, 2004; Pérez-Brocal *et al.*, 2006; McCutcheon & Moran, 2007; Tamames *et al.*, 2007).

Le nombre de composés dans les six classes est très proche pour les trois bactéries parasites intracellulaires à transmission verticale, ce qui peut s'expliquer par leur proximité phylogénétique. Par contre, *Mycoplasma hyopneumoniae* (Mhyo) a un profil souvent différent des *Bartonella* qui partagent son groupe, ce qui pourrait s'expliquer cette fois par leur éloignement phylogénétique. Ainsi, on compte un nombre de vitamines environ trois fois plus faible chez *Mycoplasma hyopneumoniae* (Mhyo) que chez les trois *Bartonella* (Bbac, Bqui et Bhen) que nous avons classées dans le même groupe des parasites intracellulaires à transmission horizontale (PIH).

Cette vue d'ensemble de la nature des métabolites dans les différents réseaux permet d'observer des tendances générales au sein des groupes, mais aussi des différences marquées chez certaines bactéries au sein d'un même groupe. La comparaison des ensembles de réactions va nous permettre de préciser ces différences, en se référant aux styles de vie de chacune des bactéries.

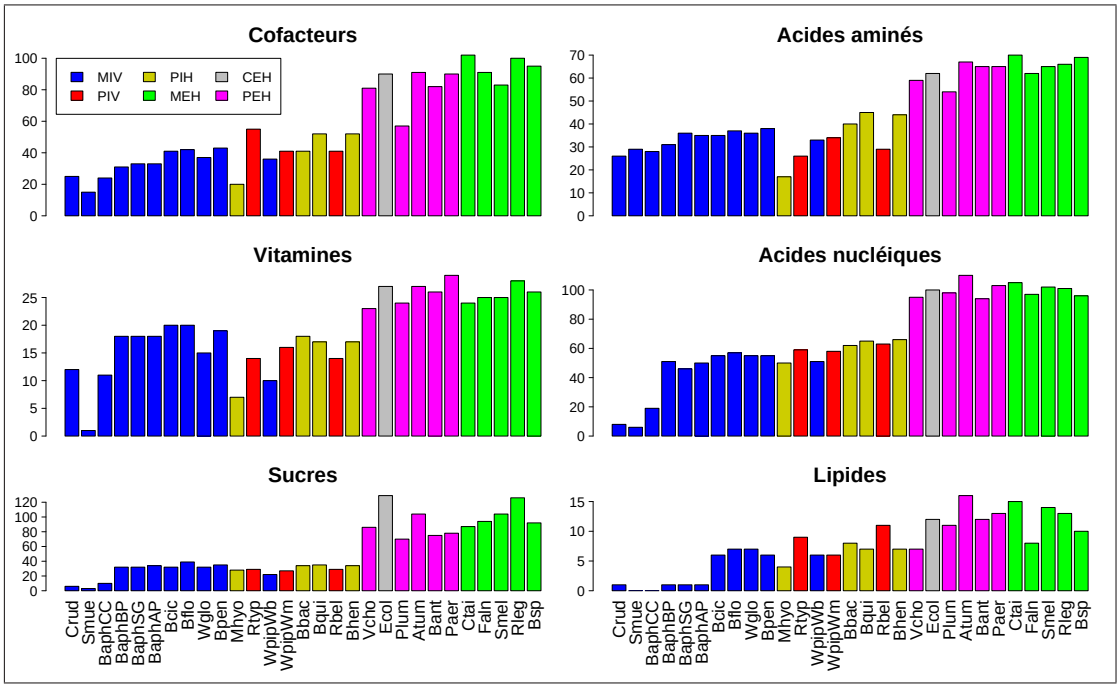

Figure 4.6. Nombre de cofacteurs, d'acides aminés, de vitamines, de sucres, d'acides aminés, d'acides nucléiques, et de lipides dans le réseau métabolique de chacun des 29 organismes, ordonnés selon leur nombre de gènes. Les abbréviations des noms d'espèces et des styles de vie correspondent à celles données dans la Figure 4.1 p.90.

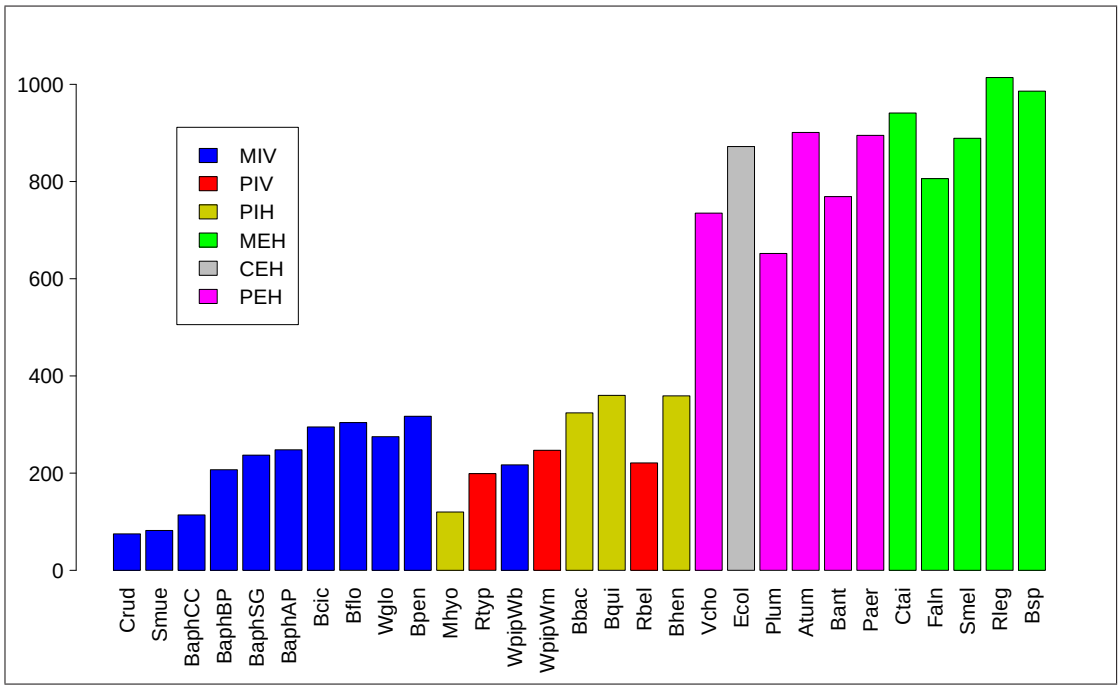

Figure 4.7. Nombre de réactions, par réseau métabolique, n'impliquant que des petites molécules . Les abbréviations des noms d'espèces et des styles de vie correspondent à celles données dans la Figure 4.1 p.90.

4.5.3 Comparaison globale des ensembles de réactions

Si l'on regarde maintenant le nombre de réactions par réseau métabolique (Figure 4.7), on voit qu'il suit globalement la même tendance générale que le nombre de gènes métaboliques et de métabolites. Il fluctue entre 75 réactions pour *Carsonella ruddii* (Crud) et 1014 pour *Rhizobium leguminosarum* (Rleg).

Nous avons testé l'hypothèse selon laquelle la réduction du métabolisme chez les bactéries intracellulaires s'accompagne d'une multi-fonctionnalité des enzymes plus élevée que chez les bactéries extracellulaires.

Dans l'ensemble, en effet, la proportion du nombre de réactions par rapport au nombre de gènes métaboliques est plus importante chez les bactéries intracellulaires[3] (Figure 4.9).

De façon surprenante, aucune réaction du métabolisme des petites molécules n'est commune entre les 29 réseaux métaboliques (Figure 4.8). En regardant de plus près les 28 composés communs trouvés précédemment, on se rend compte en effet qu'ils ne font pas intervenir à chaque fois les mêmes réactions selon les organismes. Rappelons ici qu'une réaction est envisagée dans sa globalité, on ne considère pas chaque sous-réaction qui la compose (voir Section 4.2).

De la même façon que pour les composés, l'intersection nulle entre les lots

[3]Test de Wilcoxon, H_0 : les deux distributions de proportions sont équivalentes. H_1 : les proportions sont significativement plus élevées chez les bactéries intracellulaires. Niveau du test à $\alpha = 5\%$, P-valeur $= 3.373\text{e}^{-5}$.

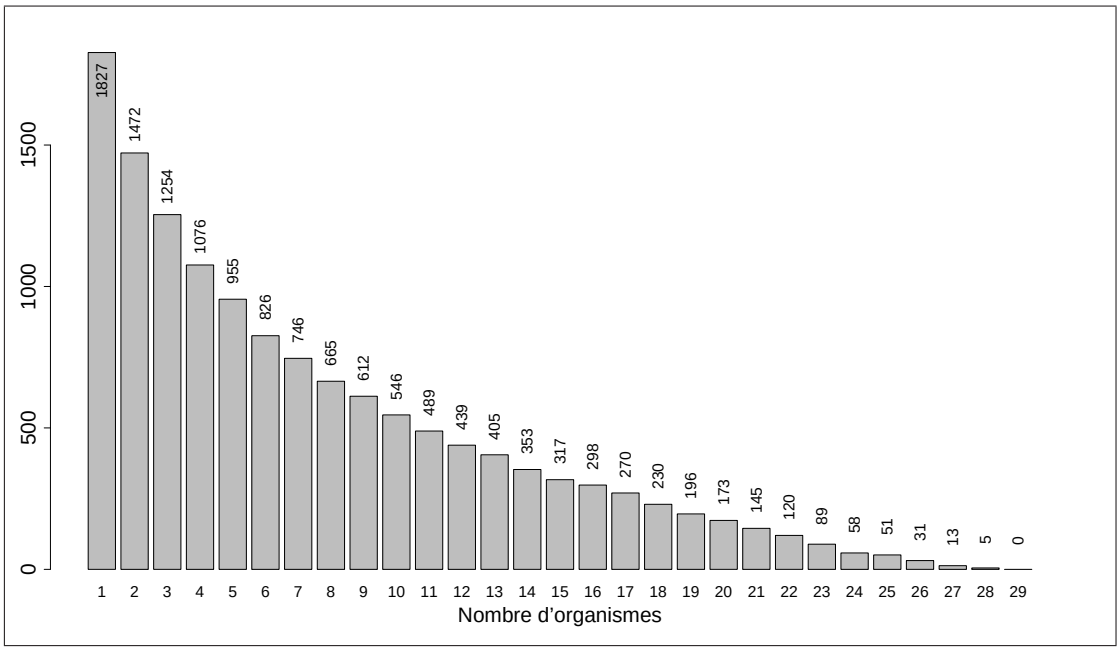

Figure 4.8. Nombre de réactions du métabolisme des petites molécules en commun entre n organismes.

de réactions de notre ensemble de bactéries s'explique par le mode de vie intracellulaire d'une partie de celles-ci, et particulièrement des bactéries mutualistes intracellulaires à transmission verticale (MIV) (Tableau 4.3). En effet, 229 réactions sont partagées par toutes les bactéries extracellulaires, correspondant aux voies métaboliques déjà citées plus haut dans le cas des métabolites, et aucune réaction n'est partagée par toutes les bactéries intracellulaires. On peut voir dans le même tableau que seules trois réactions sont communes aux bactéries MIV.

Ces constatations appuient encore davantage le fait que la réduction du réseau métabolique s'effectue sur des parties du réseau différentes selon les bactéries intracellulaires, et qu'aucune partie du réseau n'est conservée à travers les différents styles de vie des bactéries intracellulaires.

Nous allons maintenant nous focaliser sur le groupe des bactéries intracellulaires et déterminer plus finement quelles fonctions métaboliques sont conservées ou perdues en fonction de leurs styles de vie. D'abord, nous déterminerons les similitudes et les différences métaboliques à l'intérieur des groupes de bactéries intracellulaires que nous avons définis auparavant (Figure 4.1 p.90), puis entre les différents groupes.

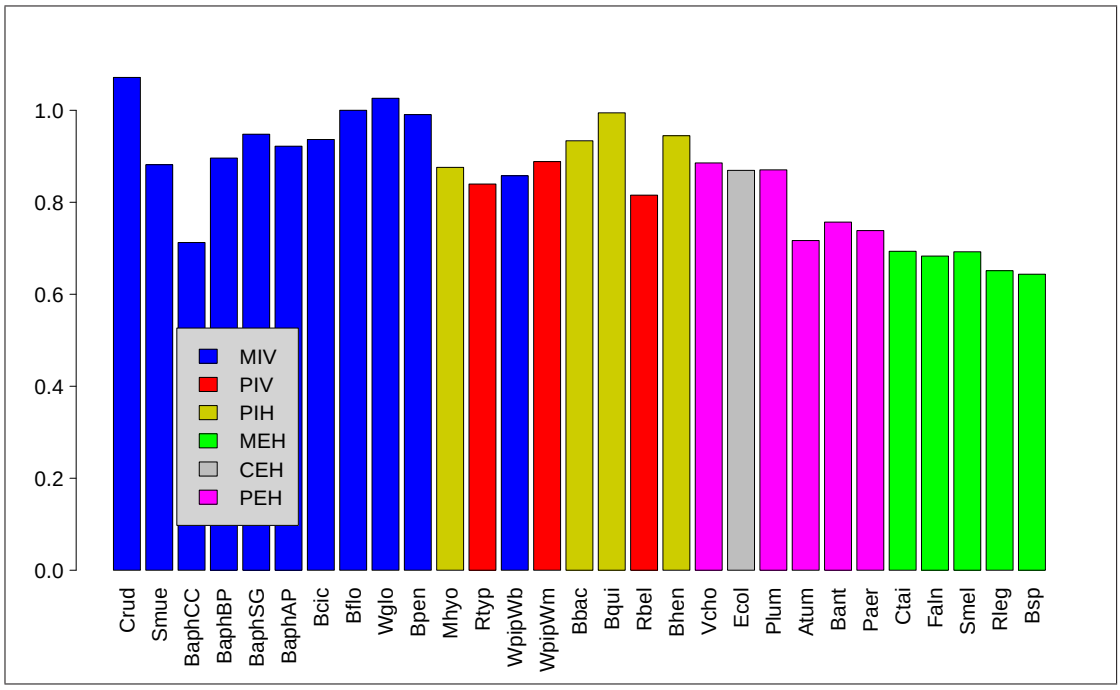

Figure 4.9. Proportion du nombre de réactions sur le nombre de gènes métaboliques. Les abbréviations des noms d'espèces et des styles de vie correspondent à celles données dans la Figure 4.1 p.90.

	MIV	PIV	PIH	**Intra**	MEH	PEH	CEH	**Extra**	**Total**
Moyenne	216	222	291	**233**	927	790	872	**860**	**471**
Union	517	348	431	**674**	1537	1399	872	**1804**	**1827**
Intersection	3	120	57	**0**	437	308	872	**229**	**0**

Tableau 4.3. Taille moyenne, taille des unions et des intersections des lots de réactions dans les différents groupes de style de vie. Le groupe des commensalistes (CEH) ne comporte qu'un organisme, *Escherichia coli K12*, l'intersection représente donc le nombre de composés dans cet organisme.

4.5.4 Comparaison du rapport entre le nombre de métabolites et le nombre de réactions

Le rapport du nombre de métabolites relativement au nombre de réactions (Figure 4.10) a tendance à être plus élevé chez les bactéries intracellulaires[4], signe d'un phénomène général de disparition des voies métaboliques redondantes chez celles-ci. Ce trait est particulièrement marqué chez les bactéries aux plus petits génomes, *Carsonella ruddii* (Crud), *Sulcia muelleri* (Smue) et *Buchnera aphidicola Cc* (BaphCC) mais aussi chez *Mycoplasma hyopneumoniae* (Mhyo) et chez les deux Rickettsies (Rtyp et Rbel).

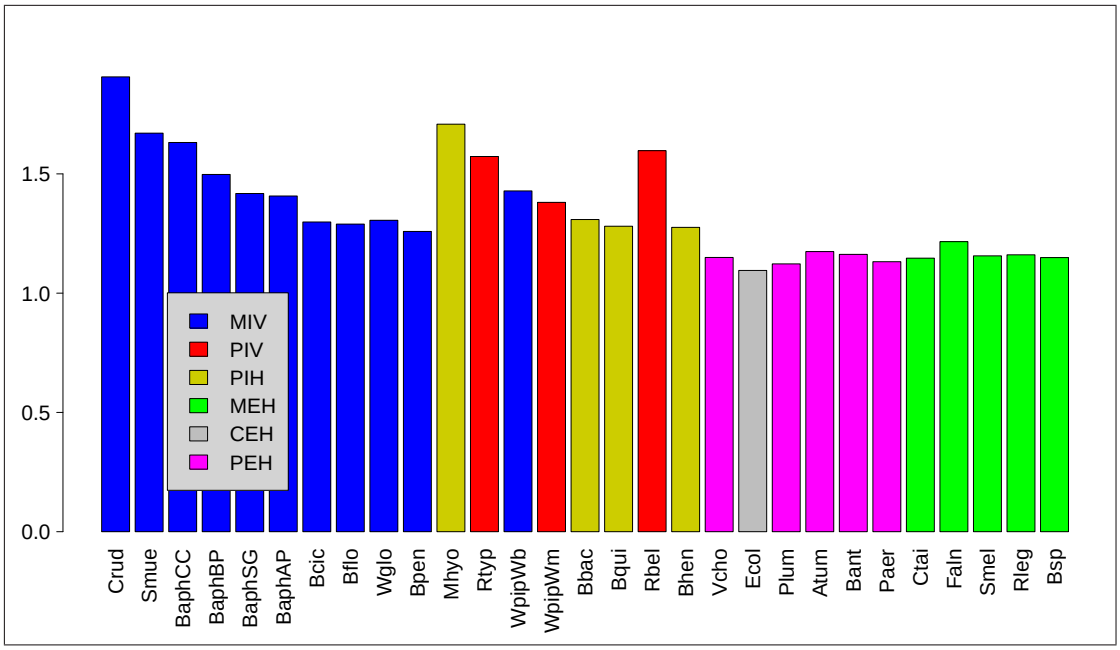

Figure 4.10. Proportion du nombre de métabolites sur le nombre de réactions intervenant dans les réseaux métaboliques des petites molécules

Rappelons que les composés présents dans un réseau métabolique ne représentent pas parfaitement les capacités métaboliques d'un organisme. Certains composés peuvent en effet intervenir seulement en tant que substrat et n'être produits par aucune réaction. Ainsi, afin d'avoir une vision des capacités métaboliques des bactéries, nous allons seulement comparer les ensembles de réactions. Pour nous aider dans l'interprétation de ces comparaisons, nous avons utilisé la vue d'ensemble des réseaux que proposent les pathway-tools et qui permet de localiser une liste de réactions parmi les différentes voies métaboliques (voir Section 2.3.2).

[4]Test de Wilcoxon, H_0 : les distributions des proportions entre les bactéries intracellulaires et extracellulaires sont équivalentes. H_1 : les proportions sont significativement plus élevées chez les bactéries intracellulaires. Niveau du test : 5%, P-valeur = 2,9e^{-8}.

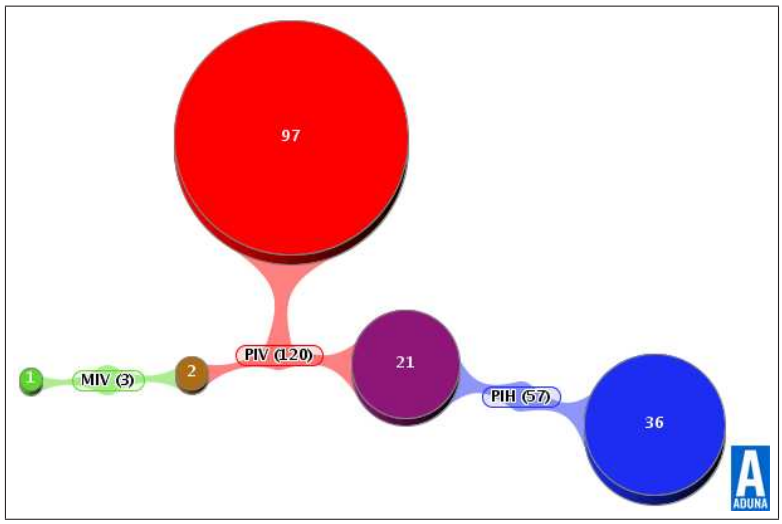

Figure 4.11. Diagramme de Venn des intersections entre les réactions communes aux bactéries MIV (Mutualistes Intracellulaires à transmission Verticale), les réactions communes des bactéries PIV (Parasites Intracellulaires à transmission Verticale) et les réactions communes aux bactéries PIH (Parasites Intracellulaires à transmission Horizontale.). Le nombre inscrit dans chaque cercle indique le nombre de réactions partagées par les ensembles correspondant aux faisceaux liés à ce cercle. Vert : réactions communes aux bactéries MIV ; Bleu : réactions communes aux bactéries PIH ; Rouge : réactions communes aux bactéries PIV.

4.5.5 Comparaison détaillée des ensembles de réactions des trois groupes de bactéries intracellulaires

Si on calcule l'intersection des lots de réactions présentes dans les trois groupes de bactéries intracellulaires, MIV (mutualistes à transmission verticale), PIV (parasites à transmission verticale) et PIH (parasites à transmission horizontale), on peut avoir une idée de ce qui est propre à chaque groupe, même si le nombre plus important de bactéries MIV et la proximité phylogénétique des bactéries PIV a un effet important sur l'intersection des lots de réactions dans ces groupes (Figure 4.11).

Seule une réaction apparaît systématiquement chez les bactéries MIV et pas systématiquement chez les autres : la réaction 6.3.5.5 qui produit du glutamate et du carbamoyl-phosphate à partir de la glutamine. Cette réaction intervient notamment dans la synthèse de l'arginine et celle de l'UMP.

Les 97 réactions qui apparaissent de façon systématique uniquement dans les bactéries PIV interviennent principalement dans la synthèse de cofacteurs, des peptidoglycanes et dans le cycle de Krebs.

Les 36 réactions qui apparaissent de façon systématique dans les bactéries PIH interviennent principalement dans la gluconéogénèse, la synthèse du NAD, et la voie des pentose phosphates.

Enfin, les 21 réactions qui apparaissent de façon systématique seulement dans les bactéries PIV et les bactéries PIH interviennent principalement dans la synthèse des nucléotides.

Afin de dégager plus finement les groupes de bactéries qui se distinguent par

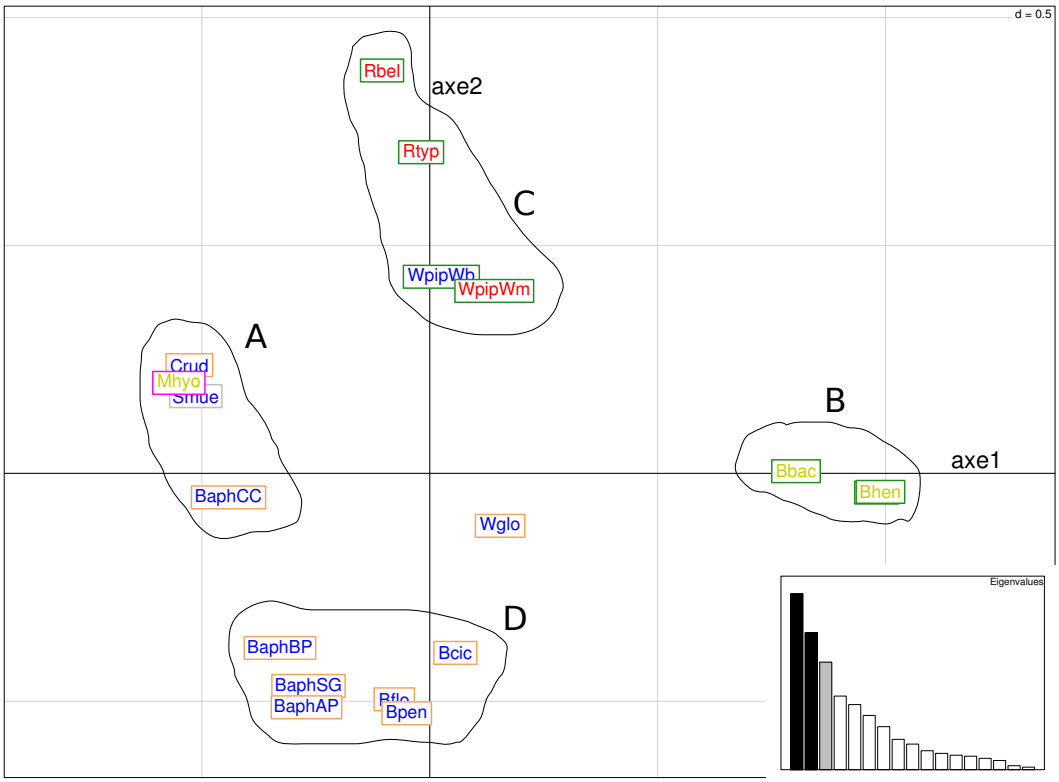

Figure 4.12. Analyse en Correspondances Multiples des bactéries intracellulaires en fonction de leurs ensembles de réactions. Projection sur le plan formé par les deux premiers axes. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. **Cadres**. Vert : α-protéobactéries ; Saumon : γ-protéobactéries ; Gris : Bactéroïdète ; Magenta : Mollicutes. **Textes**. Bleu : Mutualistes ; Rouge : Parasites à transmission verticale ; Jaune : Parasites à transmission horizontale.

leurs ensembles de réactions, nous avons effectué une analyse des correspondances multiples (ACM) (Tenenhaus & Young, 1985) sur le tableau comprenant, en ligne les 18 bactéries intracellulaires, et en colonne les 674 réactions, qui représentent l'union des réactions présentes dans chaque bactérie, divisées en 2 modalités : "présent" et "absent".

La projection sur les deux premiers axes montre tout d'abord que les profils des réactions semblent différents même à l'intérieur d'un même groupe de style de vie (illustré par la couleur des noms des organismes dans la Figure 4.12).

La projection sur le premier axe permet de séparer essentiellement deux groupes : le groupe des *Bartonella* (B) et le groupe formé par *Carsonella ruddii* (Crud), *Sulcia muelleri* (Smue), *Mycoplasma hyopneumoniae* (Mhyo) et *Buchnera aphidicola Cc* (A). Ces 4 dernières correspondent aux organismes aux réseaux métaboliques les plus réduits tandis que les *Bartonella* sont les bactéries intracellulaires de notre jeu de données qui ont le réseau métabolique le plus étendu. Ce premier axe semble donc essentiellement partager les organismes selon la taille de leur réseau métabolique. En effet, si on projette les variables sur cet axe, on note que parmi ceux qui ont une coordonnée positive, la plupart correspondent

à la modalité "présent". La présence de réactions dans le groupe des *Bartonella* absentes chez *Carsonella ruddii*, *Sulcia muelleri*, *Mycoplasma hyopneumoniae* et *Buchnera aphidicola Cc* explique donc en majeure partie la distinction de ces deux groupes.

L'intersection des ensembles de réactions entre les trois *Bartonella* nous montre qu'elles partagent 313 réactions, c'est-à-dire la majeure partie de leur réseau métabolique (données non montrées). En réalisant l'intersection de ces 313 réactions avec celles de *Carsonella ruddii*, *Sulcia muelleri*, *Buchnera aphidicola Cc* et *Mycoplasma hyopneumoniae* (Figure 4.13), on remarque que 198 réactions sont propres aux *Bartonella* et n'apparaissent pas dans les 4 autres organismes. De plus, seules 2 réactions sont partagées par *Carsonella ruddii*, *Sulcia muelleri*, *Buchnera aphidicola Cc* et *Mycoplasma hyopneumoniae*, c'est donc bien l'absence commune plutôt que la présence commune de réactions qui rapproche les 4 bactéries sur la représentation produite par l'ACM.

Les 198 réactions propres aux *Bartonella* interviennent principalement dans la synthèse des enveloppes cellulaires, des nucléotides, de nombreux cofacteurs comme la flavine, la coenzyme A, et l'ubiquinone. Par ailleurs, *Mycoplasma hyopneumoniae* possède 50 réactions non partagées avec *Carsonella ruddii*, *Sulcia muelleri*, *Buchnera aphidicola Cc* et l'intersection des composés des *Bartonella*. Cette différence peut s'expliquer par la distance phylogénétique entre *Mycoplasma hyopneumoniae* et les autres bactéries mais implique aussi que *Mycoplasma hyopneumoniae* possède la capacité de produire des composés que les *Bartonella* ne peuvent pas ou plus produire. C'est ce que nous verrons plus tard en comparant les lots de composés de ces 4 bactéries : *Mycoplasma hyopneumoniae* possède des voies de dégradation des nucléotides que les *Bartonella* ne possèdent pas.

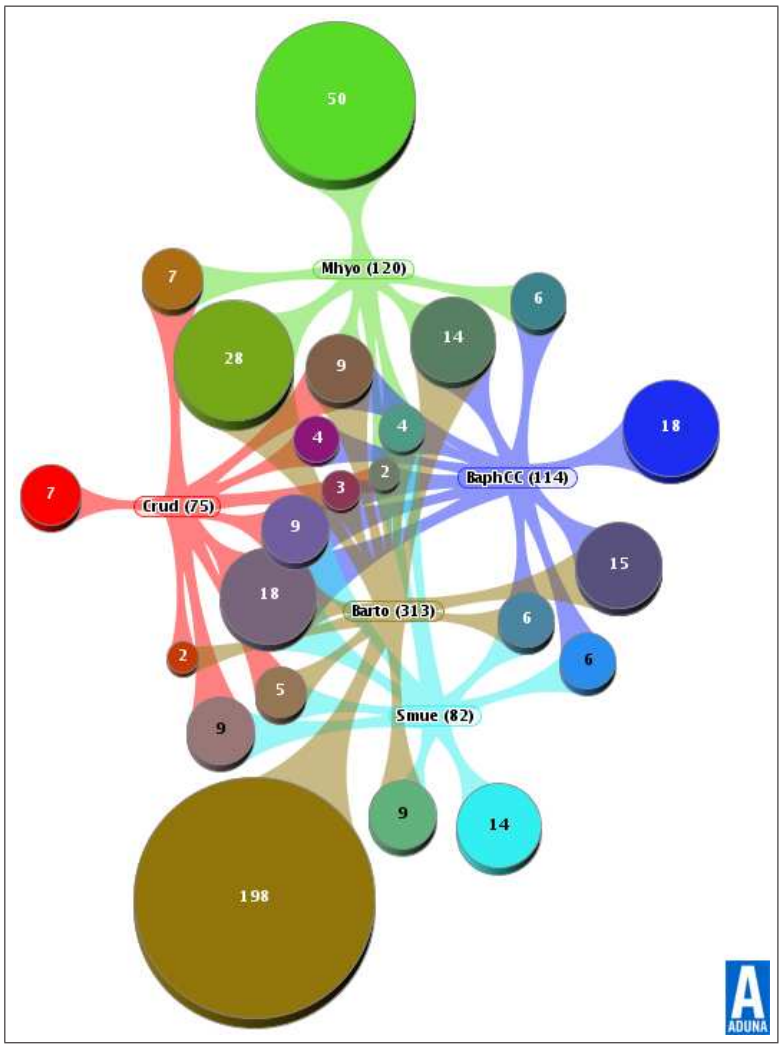

Figure 4.13. Diagramme de Venn des ensembles de réactions de *Carsonella ruddii* (Crud), *Sulcia muelleri* (Smue), *Buchnera aphidicola Cc* (BaphCC), *Mycoplasma hyopneumoniae* (Mhyo) et de l'intersection des ensembles de réactions des trois *Bartonella* (Barto). Le nombre inscrit dans chaque cercle indique le nombre de réactions partagées par les organismes correspondant aux faisceaux liés à ce cercle. Vert : réactions de Mhyo ; Bleu : réactions de BaphCC ; Rose : réactions de Crud ; Marron : réactions communes aux trois *Bartonella*.

Le deuxième axe de l'ACM sépare essentiellement deux groupes notés C et D sur la Figure 4.12. Le groupe C ne contient que des α-protéobactéries, dont trois sont des parasites à transmission verticale et une est classée parmi les mutualistes, tandis que le groupe D ne contient que des γ-protéobactéries mutualistes. Trente-trois réactions ont un pourcentage de corrélation[5] supérieur à 60 % sur l'axe 2. Parmi ces 33 réactions, 20 sont présentes dans toutes les bactéries du groupe D et absentes dans celles du groupe C, et 8 ne sont présentes que dans celles du groupe C et absentes dans l'autre groupe (Figure 4.14). Parmi les 20 premières réactions, 10 participent à la voie de synthèse de l'histidine, c'est-à-dire à la quasi-totalité de la voie. Ces 10 réactions (R6, R7, R8, R10, R14, R15, R16, R17, R19 et R20 sur la Figure 4.14) sont d'ailleurs également absentes des bactéries hors des groupes C et D. L'histidine est un acide aminé essentiel et intervient donc directement dans la fonction symbiotique de *Buchnera* (Douglas, 1998), de *Baumannia cicadellinicola* (McCutcheon & Moran, 2007), et des *Blochmannia* (Zientz *et al.*, 2004).

La voie de référence de synthèse de l'histidine commence avec le ribose-5-phosphate (Nelson & Cox, 2004), produit à partir du glucose dans la voie des pentoses phosphates. On retrouve 2 réactions participant à cette voie (R3 et R13 sur la Figure 4.14) présentes seulement dans le groupe D et complètement absentes des autres bactéries, mises à part *Bartonella bacilliformis* et *Bartonella quintana*. La disparition d'étapes dans la voie des pentoses phosphates semble donc s'accompagner de la disparition de la voie de l'histidine. La disparition complète de la voie de l'histidine et la dégradation des étapes la précédant suggère un transport direct de celle-ci.

C'est la situation inverse dans le cas des bactéries du groupe D. Dans celles-ci, la voie des pentoses phosphates est entièrement préservée. De plus, le début de la voie de l'histidine conduit à la synthèse d'aminoimidazole carboxamide ribonucleotide (AICAR), un des métabolites intermédiaires dans la fabrication des purines. Or, comme le soulignent Zientz *et al.* (2004), les premières étapes de la voie de synthèse des purines sont notamment absentes chez *Buchnera* et *Blochmannia* chez lesquelles la voie débute par AICAR. Chez ces bactéries particulièrement, la conservation de l'ensemble de la voie de synthèse de l'histidine est importante.

La voie de synthèse des purines est par contre entièrement conservée chez *Baumannia cicadellinicola*, qui fait partie aussi du groupe C. Chez cette dernière, les premières étapes de la voie de la synthèse des purines participent également à la synthèse de la thiamine, vitamine qui ferait partie des échanges symbiotiques avec l'hôte et avec le symbiote secondaire, *Sulcia muelleri* (McCutcheon & Moran, 2007).

[5]Chaque modalité est positionnée dans le plan formé par deux axes à la moyenne des coordonnées présentant cette modalité. Le pourcentage de corrélation d'une modalité sur un axe i est le pourcentage de la variance des coordonnées des individus expliquée par cette modalité. Plus le pourcentage de corrélation d'une modalité est élevé, plus celle-ci a participé à séparer les individus sur l'axe i.

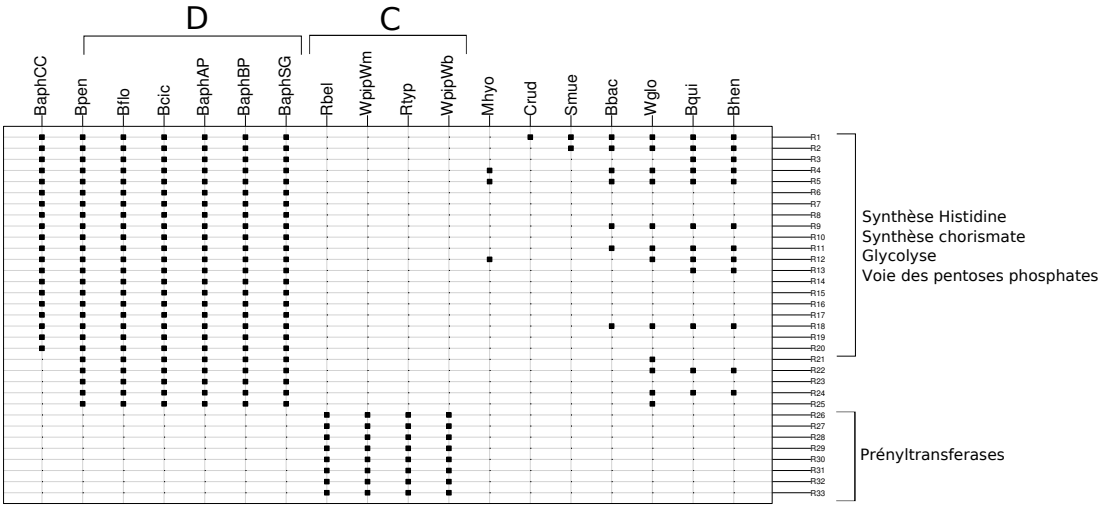

Figure 4.14. Tableau de présence des réactions ayant le pourcentage de corrélation le plus élevé sur l'axe 2. Un carré noir indique que la réaction est présente dans la bactérie correspondante. Pour chaque réaction sont indiquées les processus métaboliques dans lequelles elles interviennent. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les groupes C et D correspondent à ceux entourés sur la Figure 4.12.

Les autres réactions se partagent entre la synthèse du chorismate (qui intervient dans la synthèse des acides aminés) et la glycolyse. Six réactions parmi les huit présentes seulement dans le groupe correspondent en fait à la présence d'un seul gène, le gène *ubiA* qui code pour une prényltransférase capable de catalyser plusieurs réactions.

L'axe 3 de l'ACM (Figure 4.15) est également informatif (il explique 13% de l'inertie totale du nuage des individus). Il sépare les organismes en deux. Cette fois-ci, l'effet de la taille du réseau et de l'origine phylogénétique n'apparaissent plus : les réseaux métaboliques de petite taille sont mêlés aux plus grands et les γ-protéobactéries et α-protéobactéries sont de part et d'autre de l'origine de l'axe.

Neuf réactions ont un pourcentage de corrélation sur l'axe 3 supérieur à 60%. En effet, leur profil de présence dans les bactéries intracellulaires discrimine nettement celles qui ont une coordonnée positive sur l'axe 3 de celles qui ont une coordonnée négative (Figure 4.16). Aucune des neuf réactions n'apparaît dans le groupe B représenté sur la Figure 4.15 tandis qu'elles apparaissent toutes dans le groupe A. Le groupe B est composé des trois bactéries au réseau métabolique le plus réduit (Smue, BaphCC et Crud) et du groupe des parasites intracellulaires, contenant les trois *Bartonella* (Bbac, Bqui et Bhen) et *Mycoplasma hyopneumoniae* (Mhyo). Deux de ces réactions, R4 et R5, sont successives et participent à la synthèse d'un cofacteur, le tétrahydrofolate (THF). Il est intéressant de noter que ces deux réactions apparaissent chez *Wolbachia pipientis wMel* et non chez *Wolbachia pipientis wBm*. Peut être que chez cette dernière le produit de ces réactions est fourni par son hôte. Les réactions R3 et R4, elles aussi successives, participent à la voie de synthèse des tétrapyrolles, amenant à la synthèse

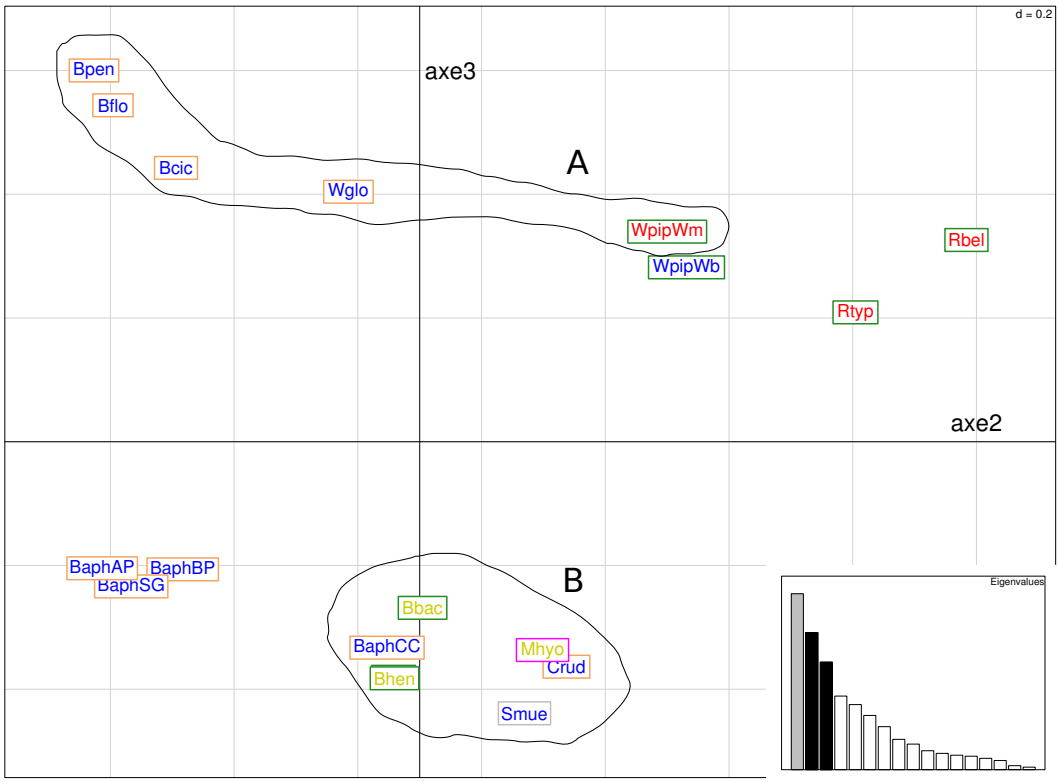

Figure 4.15. Analyse en Correspondances Multiples des bactéries intracellulaires en fonction de leurs ensembles de réactions. Projection sur le plan formé par les axes 2 et 3. Les abbréviations des noms d'espèces et des styles de vie correspondent à celles données dans la Figure 4.1 p.90. **Cadres**. Vert : α-protéobactéries ; Saumon : γ-protéobactéries ; Gris : Bactéroïdète ; Magenta : Mollicutes. **Textes**. Bleu : Mutualistes ; Rouge : Parasites à transmission verticale ; Jaune : Parasites à transmission horizontale.

de certains cofacteurs comme l'hème ou le sirohème. Les réactions R1, R2 et R6 interviennent dans la synthèse d'éléments de la membrane cellulaire. L'absence totale de ces réactions dans le groupe B indique une possible utilisation par les bactéries de ce groupe des produits de ces réactions provenant de leur hôte ou de symbiotes secondaires.

Sur les trois axes, les deux *Wolbachia* sont très proches, ce qui signifie qu'elles ont deux ensembles de réactions métaboliques très proches. La similarité des capacités métaboliques des deux *Wolbachia* a déjà été soulignée par Foster *et al.* (2005). Les deux bactéries partagent la quasi-totalité de leurs réactions (Figure 4.17). Le nombre de réactions est plus important dans le réseau de *Wolbachia pipientis wMel* que dans celui de *Wolbachia pipientis wBm*, ce qui est à l'image de la taille de leur génome. La plus grande réduction du génome de *Wolbachia pipientis wBm* serait un reflet du caractère mutualiste de sa relation avec son hôte (Foster *et al.*, 2005). Les réactions présentes dans *Wolbachia pipientis wMel* et absentes dans *Wolbachia pipientis wBm* participent à la synthèse du folate, du pyridoxal phosphate et à la dégradation de la thréonine, comme indiqué par Foster *et al.* (2005). La majeure différence entre les 2 *Wolbachia* semble être l'utilisation

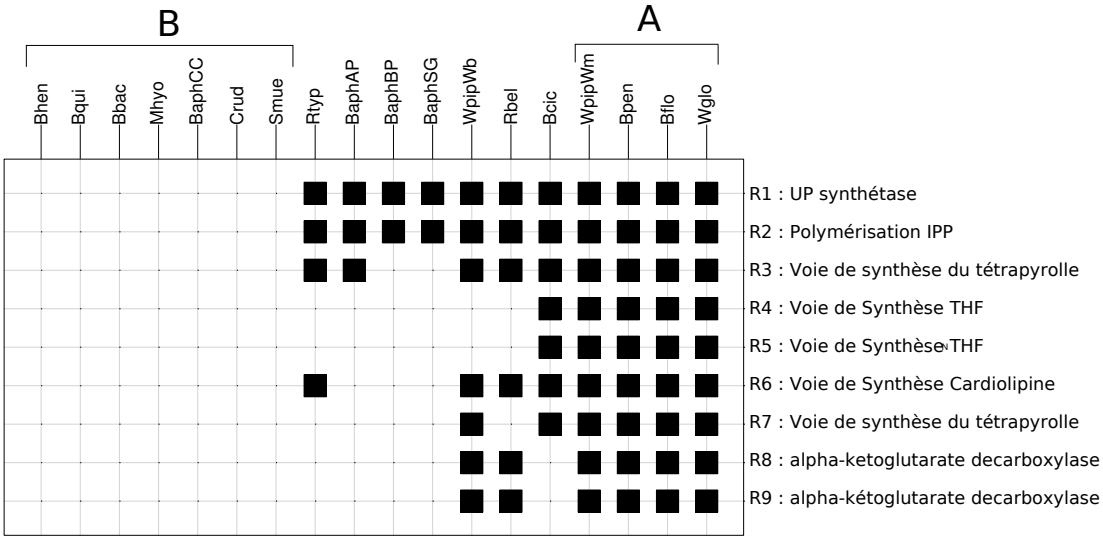

Figure 4.16. Tableau de présence des réactions ayant le pourcentage de corrélation le plus élevé sur l'axe 3. Un carré noir indique que la réaction est présente dans la bactérie correspondante. Pour chaque réaction sont indiquées les processus métaboliques dans lequelles elles interviennent. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les groupes A et B correspondent à ceux entourés sur la Figure 4.15. **IPP** : Isopentenyl diphosphate ; **UP** : undecaprenyl pyrophosphate.

qu'en font leurs hôtes respectifs. En effet, l'hôte de *Wolbachia pipientis wBm*, un nématode, nécessiterait le symbiote pour la synthèse de métabolites, comme la riboflavine ou l'hème, qu'il semble ne pas pouvoir synthétiser (Foster *et al.*, 2005).

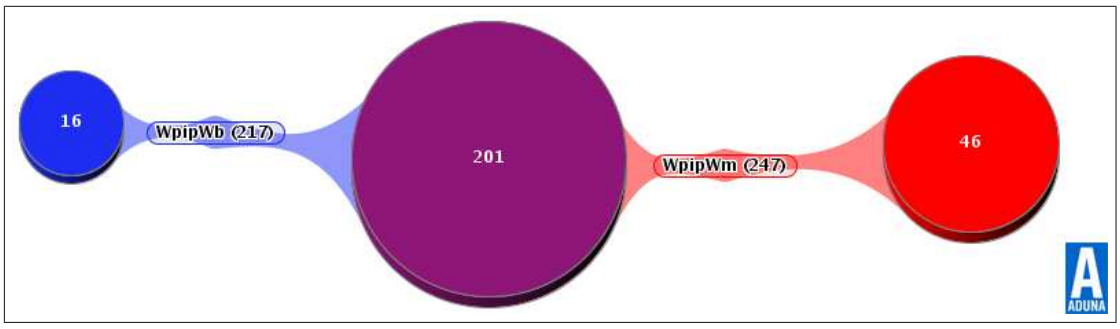

Figure 4.17. Diagramme de Venn des réactions de *Wolbachia pipientis wBm* (WpipWb) et de *Wolbachia pipientis wMel* (WpipWm). Le nombre inscrit dans chaque cercle indique le nombre de réactions partagées par les organismes correspondant aux faisceaux liés à ce cercle. Bleu : ensemble de réactions de WpipWB. Rouge : ensemble de réactions de WpipWM.

Nous allons maintenant comparer les lots de réactions à l'intérieur de chaque groupe de style de vie intracellulaire.

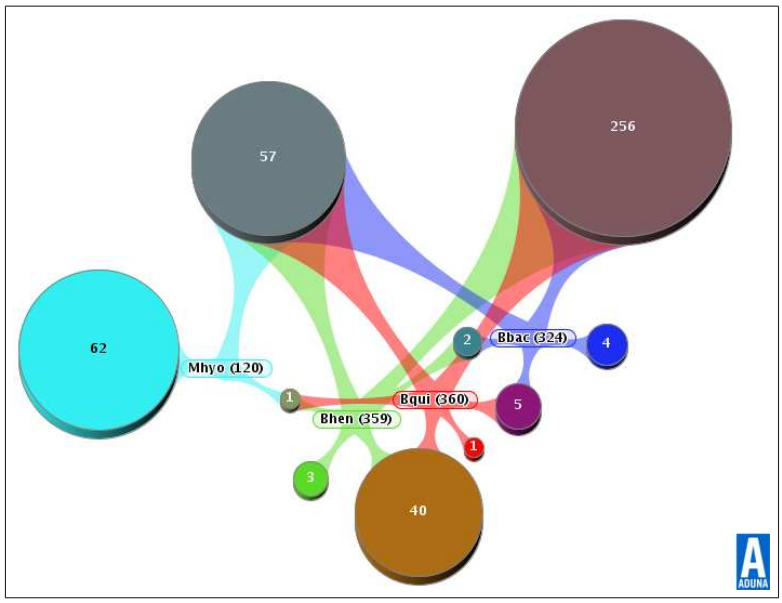

Figure 4.18. Diagramme de Venn des réactions de *Mycoplasma hyopneumoniae* (Mhyo) et des trois *Bartonella* (Bbac, Bqui et Bhen. Le nombre inscrit dans chaque cercle indique le nombre de réactions partagées par les organismes correspondant aux faisceaux liés à ce cercle. Bleu clair : réactions de Mhyo ; Vert : réactions de Bhen ; Bleu foncé : réactions de Bbac, Rouge : réactions de Bqui.

4.5.6 Comparaison détaillée des ensembles de réactions des bactéries parasites à stade de vie intracellulaire à transmission horizontale (PIH)

Nous avons classé *Mycoplasma hyopneumoniae* et les *Bartonella* dans le même grand groupe de style de vie : les parasites à stade de vie intracellulaire à transmission horizontale (PIH). Cependant, leur type de parasitisme et leur habitat sont considérablement différents. En effet, les *Bartonella* peuvent exploiter les érythrocytes de leur hôte habituel sans grand dommage pour celui-ci pendant une période assez longue (Birtles, 2005). *Mycoplasma hyopneumoniae*, de son côté, se fixe sur les cellules épithéliales de la trachée et même si on peut la considérer comme intracellulaire à cause de la fusion de sa membrane avec celles des cellules hôtes, la cellule ne pénètre jamais entièrement dans la cellule hôte. En outre, *Mycoplasma hyopneumoniae* est phylogénétiquement très éloignée des *Bartonella*. L'intersection des ensembles de réactions est ainsi particulièrement faible puisque seulement 57 réactions de *Mycoplasma hyopneumoniae* sont partagées avec les trois *Bartonella* et que 62 sont propres à *Mycoplasma hyopneumoniae* (Figure 4.18).

En regardant de plus près les 62 réactions propres à *Mycoplasma hyopneumoniae* et en les projetant sur les voies métaboliques détectées comme présentes dans cet organisme, nous pouvons voir que beaucoup d'entre elles participent aux voies de dégradation des nucléotides. En effet, *Mycoplasma hyopneumoniae* possède des voies de synthèse *de novo* des nucléotides très dégradées mais sont capables d'in-

terconvertir des nucléotides comme la thymine ou la guanine en d'autres nucléotides (Vasconcelos *et al.*, 2005). Ces voies métaboliques de dégradation ont déjà été mises en évidence expérimentalement chez *Mycoplasma mycoides* (Mitchell & Finch, 1977).

De l'autre côté, les *Bartonella* ne semblent pas posséder ces voies de dégradation mais possèdent les voies complètes de synthèse des nucléotides. Les *Bartonella* ont un stade de vie plus ou moins prolongé dans le sang (Alsmark *et al.*, 2004; Birtles, 2005). Les bactéries doivent réguler leur métabolisme pour survivre dans ce milieu particulier et une étude a montré que la biosynthèse des nucléotides est critique pour la croissance d'*E. coli* lorsqu'elle est cultivée dans le sérum humain (Samant *et al.*, 2008). Le séjour dans le sang pourrait donc provoquer une pression sélective sur la conservation de la synthèse *de novo* des nucléotides chez les *Bartonella*. Dans ce cas, les différences entre les réseaux métaboliques de *Mycoplasma hyopneumoniae* et ceux des *Bartonella* pourraient être dues, outre à un éloignement phylogénétique considérable, aux milieux de vie différents que traversent les bactéries.

Les réactions n'intervenant que chez les *Bartonella* participent principalement, comme nous l'avons vu, à la synthèse des nucléotides mais aussi à la synthèse des acides aminés, de la coenzyme A, de l'ubiquinone, des structures cellulaires, et de la respiration. Ces voies ont disparu chez *Mycoplasma hyopneumoniae*. En effet, cette dernière a subi une réduction de son métabolisme plus important que les *Bartonella*. L'explication peut être une réduction du métabolisme plus récente chez ces dernières. Il se peut également qu'une plus grande partie du réseau soit conservée pour être adaptée au plus grand nombre de milieux différents que traversent les *Bartonella* durant leur cycle de vie (milieu extracellulaire chez le vecteur arthropode, puis érythrocytes chez l'hôte mammifère).

Les réactions communes entre les 4 organismes participent principalement à la glycolyse, à la synthèse du NAD et à la voie des pentoses phosphates, ce qui montre une certaine autonomie dans la dégradation des sucres et en métabolites fournissant de l'énergie.

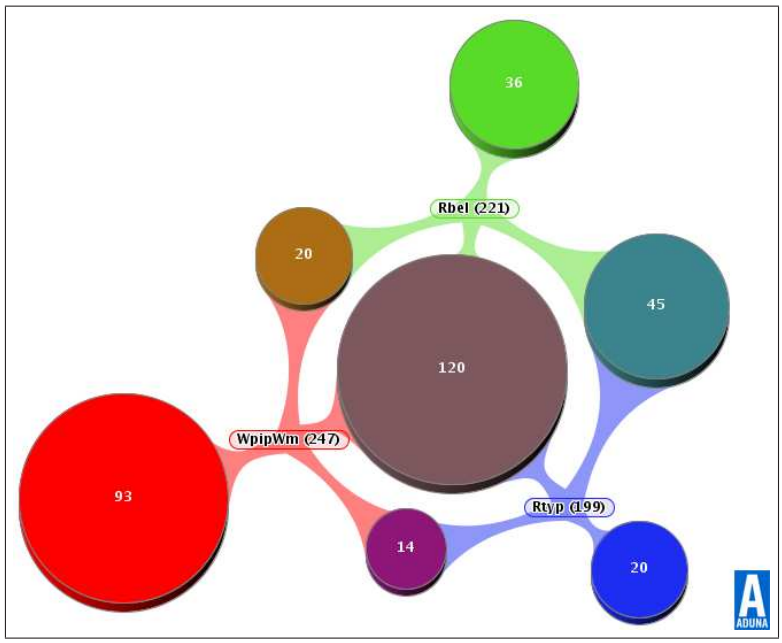

Figure 4.19. Diagramme de Venn des réactions des bactéries PIV. Le nombre inscrit dans chaque cercle indique le nombre de réactions partagées par les organismes correspondant aux faisceaux liés à ce cercle. Bleu : réactions de *Rickettsia typhi* (Rtyp) ; Vert : réactions de *Rickettsia bellii* (Rbel) ; Rouge : réactions de *Wolbachia pipientis wMel* (WpipWm).

4.5.7 Comparaison détaillée des ensembles de réactions des bactéries parasites à stade de vie intracellulaire à transmission verticale (PIV)

Trois bactéries font partie de ce groupe dont deux Rickettsies : *Rickettsia bellii*, *Rickettsia typhi* et *Wolbachia pipientis wMel*. Les *Rickettsia* et les *Wolbachia* sont assez proches phylogénétiquement : elles appartiennent au même ordre des Rickettsiales. Cent vingt réactions sont communes aux trois bactéries et interviennent principalement dans la respiration, la synthèse des pyrimidines, des isoprénoïdes et des peptidoglycanes (Figure 4.19). Cependant, malgré la proximité phylogénétique de ces deux groupes, 93 réactions sont propres à *Wolbachia pipientis wMel* et 45 réactions sont propres à *Rickettsia bellii* et *Rickettsia typhi*.

Parmi les voies métaboliques dans lesquelles les 93 réactions propres à *Wolbachia pipientis wMel* apparaissent, on peut citer principalement celles-ci :

- la synthèse *de novo* des purines,
- la synthèse *de novo* de l'UMP, utilisé ensuite pour la synthèse des pyrimidines,
- la synthèse de la flavine,
- la glycolyse et la gluconéogénèse,
- la synthèse de l'isopentenyl diphosphate (IPP) à partir du méthylérythritol phosphate (MEP),

– la voie non oxydative des pentoses phosphates.

Les Rickettsies sont capables d'importer de l'ATP de leur hôte alors que les *Wolbachia* en sont incapables, ce qui explique la conservation de la glycolyse et des voies de synthèse des purines chez ces dernières. Wu *et al.* (2004) ont signalé également la conservation de la voie de synthèse des pyrimidines et de celle de la flavine.

L'isopentenyl diphosphate (IPP) est un métabolite clé dans la synthèse de l'ubiquinone ou de la métaquinone, requises pour la respiration aérobie. La voie de synthèse de l'IPP est complètement absente chez les *Rickettsia* alors que ces dernières l'utilisent pour la synthèse de quinones, ce qui laisse à penser que les *Rickettsia* importent ce métabolite de leur hôte (Lange *et al.*, 2000).

Les deux Rickettsies ne semblent pas posséder la voie des pentose phosphates. Celle-ci est centrale dans le métabolisme : elle dégrade des hexoses en pentoses et elle est productrice de NADPH. L'absence de cette voie et de la glycolyse montre la dépendance totale des *Rickettsia* envers leur hôte en ce qui concerne la production d'énergie.

En ce qui concerne le métabolisme énergétique et la synthèse de l'IPP, on peut considérer que deux stratégies différentes ont été sélectionnées, si on part de l'hypothèse que l'ancêtre des *Wolbachia* et des *Rickettsia* possédait les deux possibilités : la conservation de la production de l'énergie chez les *Wolbachia* ou l'utilisation de transporteurs chez les *Rickettsia*. On peut formuler l'hypothèse suivante. Si la réduction du génome se fait aléatoirement, on peut imaginer que les *Wolbachia* ont d'abord perdu les gènes codant pour le transport des métabolites tels que l'ATP ou l'IPP et que la pression de sélection s'est donc effectuée ensuite sur les voies de synthèse de ces 2 métabolites. Dans l'autre sens, les *Rickettsia* ont pu perdre une partie des 2 voies métaboliques et la pression sélective s'est ainsi exercée sur les transporteurs.

Les 45 réactions que l'on ne trouve que chez les 2 Rickettsies appartiennent principalement aux voies de synthèse des lipopolysaccharides (LPS) qui interviennent dans la fabrication de la paroi externe. En effet, ces voies sont manquantes chez les *Wolbachia* (Wu *et al.*, 2004). Il est à noter que la perte de ces voies a eu lieu également chez les *Buchnera* (Zientz *et al.*, 2004). La disparition de cette membrane peut être expliquée par la toxicité pour l'hôte de certains éléments qu'elle contient. Zientz *et al.* (2004) suggèrent que la disparition de cette membrane, en tout cas de ses éléments toxiques, pourrait jouer un rôle clé dans l'intégration stable de l'endosymbiote dans son hôte.

Deux éléments semblent donc pouvoir expliquer les divergences métaboliques de *Wolbachia pipientis wMel* et des *Rickettsia* : une réduction différente du réseau métabolique et une intégration plus importante de *Wolbachia* dans son hôte.

4.5.8 Comparaison détaillée des ensembles de réactions des bactéries mutualistes à stade de vie intracellulaire à transmission verticale (MIV)

Les bactéries MIV sont au nombre de 11. Ce sont toutes des γ-protéobactéries sauf *Sulcia muelleri* qui est une flavobactérie et *Wolbachia pipientis wMel* qui est une α-protéobactérie.

Parmi les 517 réactions présentes dans l'ensemble des 11 réseaux, seules trois sont communes à tous. La première, dont le numéro EC est 6.3.5.5, produit le carbamoly-phosphate à partir de bicarbonate et de glutamine et intervient dans la synthèse de l'arginine. La seconde, dont le numéro EC est 5.1.1.7, transforme le L,L-diaminopimélate en méso-diaminopimélate et intervient dans la synthèse de la lysine. La troisième, dont le numéro EC est 3.5.1.88, est une réaction générique qui transforme un peptide formyl-L-méthionyl en peptide méthionyl.

Il est difficile de comparer les 11 espèces à la fois comme nous l'avons fait précédemment avec les autres groupes de style de vie. Afin de repérer les organismes dont les ensembles de réactions se ressemblent et quelles sont les réactions qui diffèrent en fonction des groupes, nous procédons d'abord à une analyse en correspondances multiples (ACM). Notre tableau de données comporte 11 individus (les organismes) et 517 colonnes (l'union des réactions trouvées dans chaque réseau) divisées en 2 modalités : présence ou absence. La Figure 4.20 représente la projection des bactéries sur les deux premiers axes de l'ACM.

La projection des variables sur le premier axe montre qu'une majorité des modalités "présence" se situe à droite de l'origine, ce qui indique que ce sont surtout la présence de réactions chez certaines bactéries et l'absence des mêmes réactions chez d'autres bactéries qui déterminent le premier axe. On compte 76 réactions qui ont un pourcentage de corrélation supérieur à 60 % sur l'axe 1. Parmi ces 76 réactions, 74 sont présentes dans le groupe des quatre bactéries dont les valeurs sont les plus élevées sur le premier axe. Ce groupe contient *Wigglesworthia glossinidia* (Wglo), les deux *Blochmannia* (Bflo et Bpen) et *Baumannia cicadellinicola* (Bcic) (Groupe B, Figure 4.21).

Ces 74 réactions sont complètement absentes des trois bactéries qui ont les valeurs les plus faibles sur le premier axe et qui sont réunies dans le groupe A de la Figure 4.20. Ce groupe contient les trois bactéries au réseau métabolique le plus réduit, *Sulcia muelleri* (Smue), *Carsonella ruddii* (Crud) et *Buchnera aphidicola Cc* (BaphCC). Chez les autres bactéries, c'est-à-dire les trois *Buchnera* réunies dans le groupe C sur la Figure 4.20, et *Wolbachia pipientis wBm*, isolée sur la représentation, une partie seulement des 74 réactions est présente. Comme le premier axe de l'ACM réalisée sur l'ensemble des bactéries intracellulaires (Figure 4.12), le premier axe de cette ACM est déterminé essentiellement par les réactions absentes des réseaux les plus réduits.

Deux réactions, cependant, ont un pourcentage de corrélation supérieur à 60 %

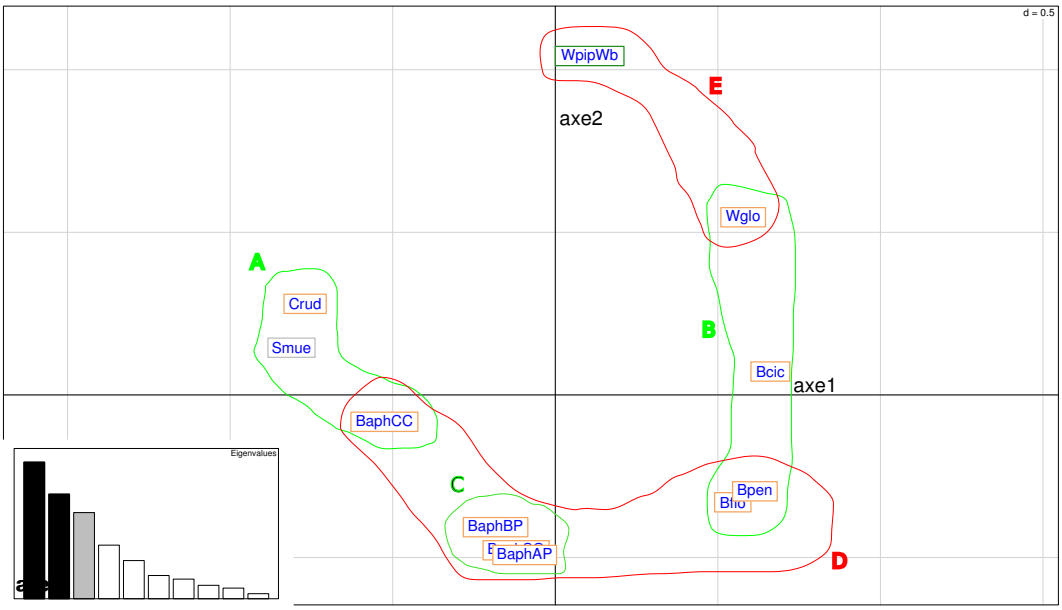

Figure 4.20. Analyse en Correspondances Multiples sur la présence des réactions chez les bactéries MIV. Projection sur les deux premiers axes. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. **Cadres**. Vert : α-protéobactéries ; Saumon : γ-protéobactéries ; Gris : Bactéroïdète. Les ensembles entourés en vert correspondent aux groupes de bactéries qui se distinguent selon le premier axe. Les ensembles entourés en rouge correspondent aux groupes qui se distinguent selon le second axe.

et ne sont absentes que chez les bactéries du groupe B et *Wolbachia pipientis wBm*. Ces deux réactions correspondent à la même activité catalytique et ont le numéro EC 6.3.4.5. L'une de ces deux réactions intervient dans la synthèse de l'arginine, acide aminé essentiel. L'absence de cette réaction chez les *Blochmannia*, dont le rôle nutritionnel serait de fournir leur hôte en acides aminés, est étonnant et avait déjà été soulignée par Zientz *et al.* (2004). Chez *Wigglesworthia glossinidia*, l'arginine pourrait être fournie par l'autre symbiote de la mouche tsé-tsé, *Sodalis glossinidius*, tandis que ce serait *Sulcia muelleri*, l'autre symbiote de la cicadelle, qui pourrait fournir ce composé à *Baumannia cicadellinicola*.

Par ailleurs, 23 réactions sont absentes chez les trois organismes du groupe A et présentes chez toutes les 11 autres bactéries mutualistes. Elles interviennent principalement dans les processus suivants : la synthèse de la riboflavine, des peptidoglycanes, des purines et des pyrimidines. La réduction extrême des réseaux métaboliques semble donc toucher principalement ces voies de synthèse.

Il est intéressant également de regarder les réactions qui ont été conservées dans les trois réseaux métaboliques du groupe A (Figure 4.22). Ving neuf réactions sont communes aux trois organismes et participent essentiellement aux voies de synthèse de la lysine, de l'isoleucine et du chorismate. Ce dernier est un intermédiaire dans la synthèse des acides aminés aromatiques (tryptophane, phénylalanine et tyrosine), les réactions conservées chez ces trois organismes correspondent donc essentiellement à leur fonction symbiotique, la production d'acides

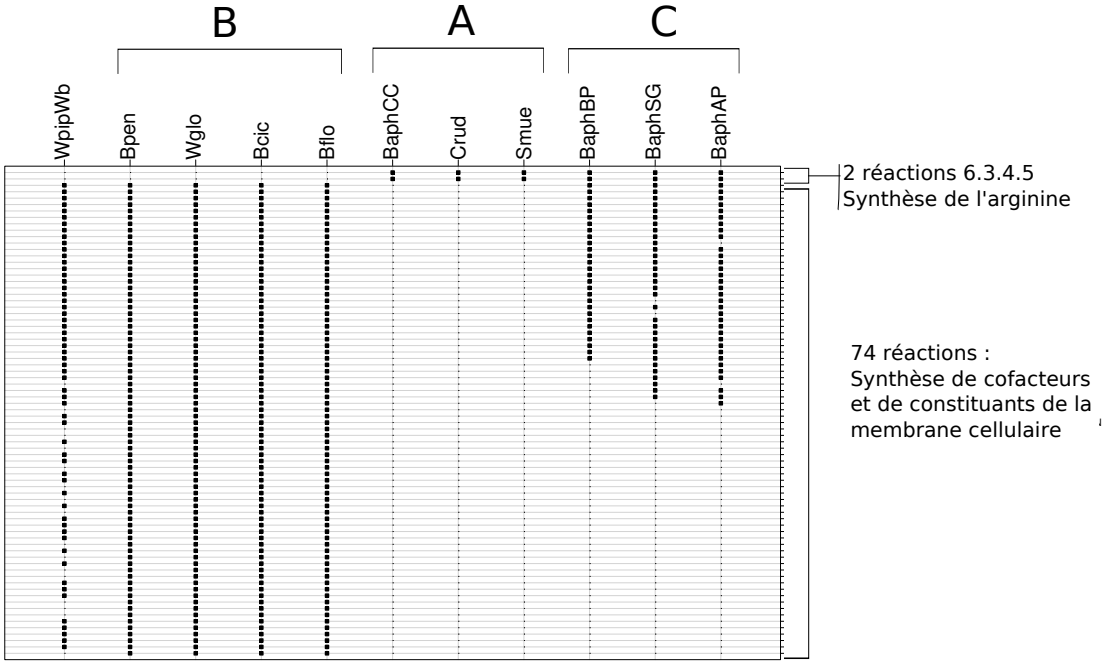

Figure 4.21. Tableau de présence des réactions ayant le pourcentage de corrélation le plus élevé sur l'axe 1 de l'ACM de la Figure 4.20. Un carré noir indique que la réaction est présente dans la bactérie correspondante. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les groupes A, B et C correspondent à ceux entourés sur la Figure 4.20.

aminés (Pérez-Brocal *et al.*, 2006; Tamames *et al.*, 2007; Wu *et al.*, 2006b).

Notons également que malgré son éloignement phylogénétique, seules 23 réactions ne se retrouvent que dans *Sulcia muelleri* et non dans les deux autres. Parmi ces 23 réactions, on retrouve celles intervenant dans les voies de synthèse du tryptophane et de l'ornithine qui ont disparu des deux autres.

Les 53 réactions propres à *Buchnera aphidicola Cc* (Figure 4.22) participent à la voie de synthèse de l'histidine et à la glycolyse, dégradées chez les deux autres bactéries. Dans le cas de *Sulcia muelleri*, c'est l'autre symbiote, *Baumannia cicadellinola* qui partage son hôte, qui produirait l'histidine et les produits de la glycolyse (Wu *et al.*, 2006b; McCutcheon & Moran, 2007).

Enfin, parmi les 16 réactions propres à *Carsonella ruddii*, on note celles intervenant dans les voies de dégradation de la proline en glutamate, ce qui indique une synthèse différente de ce métabolite chez cette bactérie par rapport aux deux autres mutualistes intracellulaires.

La disparition chez ces trois bactéries de voies de synthèse des acides aminés suppose l'intervention d'un symbiote secondaire pour approvisionner non seulement l'hôte mais aussi la bactérie elle-même en ces composés manquants. Un symbiote secondaire a été clairement mis en évidence dans le cas de *Sulcia muelleri* et *Buchnera aphidicola Cc*, mais il reste à identifier dans le cas de *Carsonella ruddii* (Pérez-Brocal *et al.*, 2006; Tamames *et al.*, 2007; Wu *et al.*, 2006b).

Il est intéressant de voir par ailleurs que les deux symbiotes qui partagent le

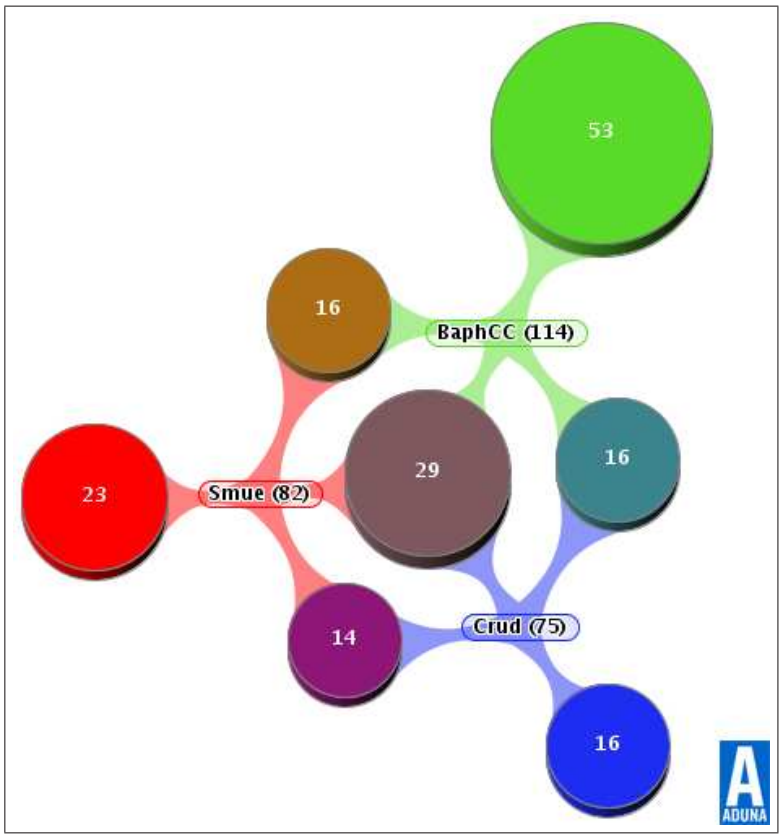

Figure 4.22. Diagramme de Venn des réactions de *Sulcia muelleri* (Smue), *Buchnera aphidicola Cc* (BaphCC) et *Carsonella ruddii* (Crud). Le nombre inscrit dans chaque cercle indique le nombre de réactions partagées par les organismes correspondant aux faisceaux liés à ce cercle. Bleu : réactions de Crud ; Vert : réactions de BaphCC ; Rouge : réactions de Smue.

même hôte, *Baumannia cicadellinicola* et *Sulcia muelleri*, se retrouvent à l'opposé du premier axe de l'ACM. Seules 24 réactions sont communes entre *Baumannia cicadellinicola* et *Sulcia muelleri* et ces réactions correspondent à l'intersection la plus petite entre les réactions de *Baumannia cicadellinicola* ou de *Sulcia muelleri* et celles des autres bactéries MIV (Figure 4.23).

Même si cette faible intersection est expliquée en partie par l'éloignement phylogénétique des deux symbiotes, elle indique également une complémentarité entre les deux réseaux métaboliques.

Les 58 réactions présentes dans *Sulcia muelleri* et non présentes dans *Baumannia cicadellinicola* interviennent quasiment toutes dans la synthèse des acides aminés mais aussi dans la synthèse de la ménaquinone. On retrouve la complémentarité des capacités métaboliques des deux organismes, signalée par McCutcheon & Moran (2007).

Les 74 réactions ayant un pourcentage de corrélation élevé sur l'axe 1 et qui apparaissent toutes dans les bactéries du groupe B (Figure 4.21) participent essentiellement à la synthèse de cofacteurs comme la flavine, le tétrahydrofolate,

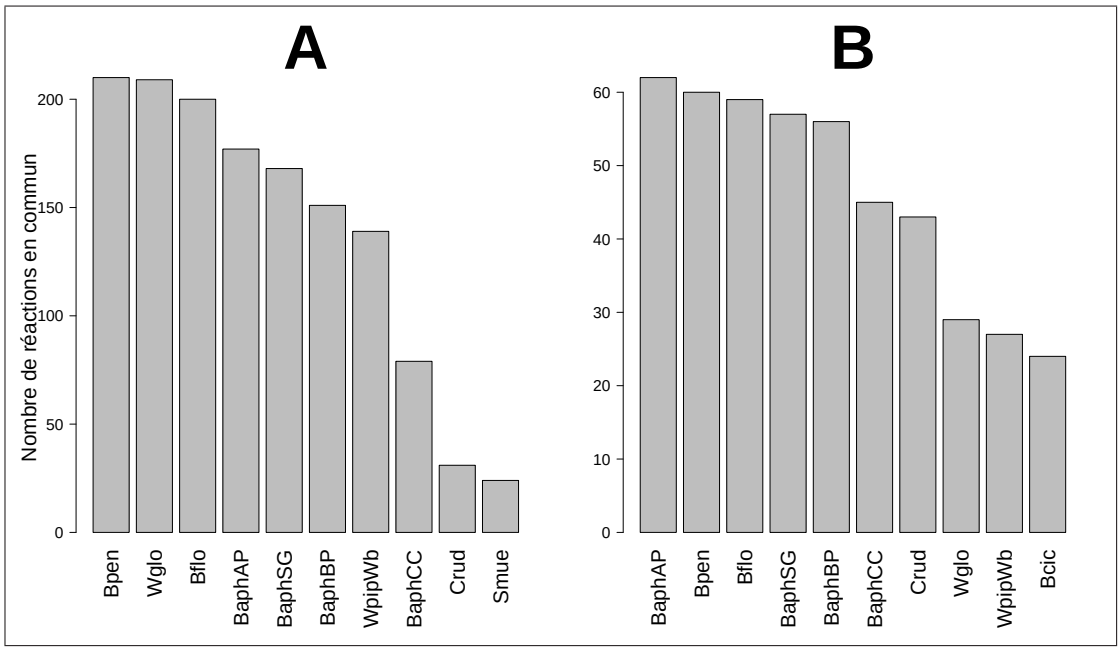

Figure 4.23. A. Nombre de réactions communes entre *Baumannia cicadellinicola* et les autres bactéries MIV. **B.** Nombre de réactions communes entre *Sulcia muelleri* et les autres bactéries MIV.

les tétrapyrolles et le pyridoxal 5'-phosphate, ainsi qu'à la synthèse de constituants de la membrane cellulaire.

L'intersection entre les quatre ensembles de réactions du groupe B de la Figure 4.20, nous montre une grande conservation des réactions entre les quatre bactéries : près de la moitié des réactions de chaque réseau sont partagées par les trois autres (Figure 4.24). L'intersection entre les réactions des 2 *Blochmannia* représente la quasi-totalité de leurs réseaux métaboliques. Les réactions présentes dans *Blochmannia pennsylvanicus* et absentes de *Blochmannia floridanus* interviennent dans la synthèse de la coenzyme A et dans la voie de synthèse des isoprénoïdes, comme le signalent Degnan *et al.* (2005).

La projection sur le second axe de l'ACM (Figure 4.20) permet de séparer essentiellement deux groupes : celui formé par les *Buchnera* (BaphAP, BaphBP, BaphCC et BaphSG) et les *Blochmannia* (Bflo et Bpen), noté D sur la Figure 4.20, et le groupe formé par *Wigglesworthia glossinidia* (Wglo) et *Wolbachia pipientis wBm* (WpipWB), noté E sur la Figure 4.20. Trente huit réactions ont un pourcentage de corrélation supérieur à 60 % sur le deuxième axe. Parmi ces 38 réactions, 15 n'apparaissent que dans le groupe E et sont absentes du groupe D et 23 n'apparaissent que dans le groupe D et sont absentes du groupe E (Figure 4.25).

Ces dernières réactions correspondent essentiellement à la synthèse de l'histidine (11 d'entre elles) et à la synthèse du chorismate (5 d'entre elles). On retrouve aussi des réactions participant à la synthèse de la phénylalanine, de la tyrosine, de la thréonine, et à la voie des pentoses phosphates. La conservation quasi complète

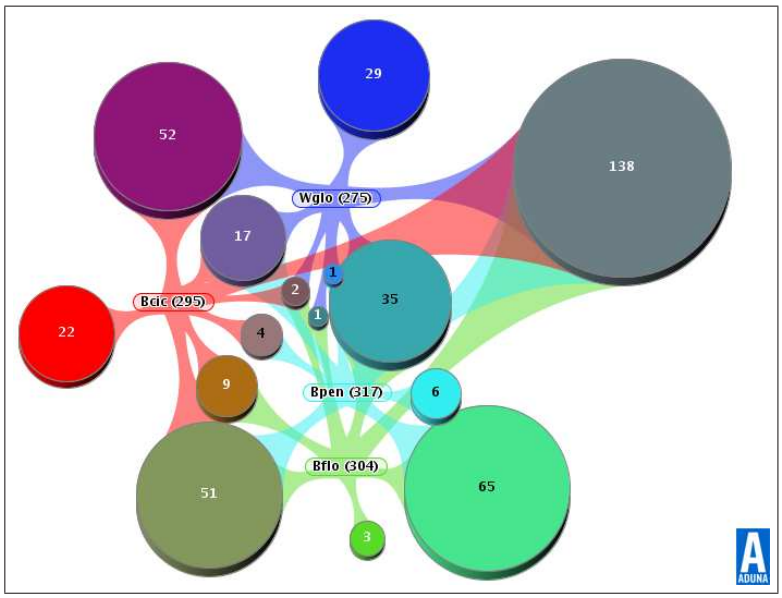

Figure 4.24. Diagramme de Venn des intersections des ensembles de réactions de *Wigglesworthia glossinidia* (Wglo), *Baumannia cicadellinicola* (Bcic), *Blochmannia pennsylvanicus* (Bpen) et *Blochmannia floridanus* (Bflo). Le nombre inscrit dans chaque cercle indique le nombre de réactions partagées par les organismes correspondant aux faisceaux liés à ce cercle. Bleu foncé : réactions de Wglo ; Bleu clair : réactions de Bpen ; Rouge : réactions de Bcic ; Vert : réactions de Bflo.

de la voie de l'histidine chez les bactéries de ce groupe a déjà été commentée lors de la comparaison globale des bactéries intracellulaires.

Le chorismate est un des composés clés dans la synthèse de la phénylalanine, de la tyrosine et du tryptophane. La disparition ou la conservation de la voie du chorismate influe donc logiquement sur la disparition ou la conservation de réactions intervenant dans la synthèse de ces trois acides aminés.

Les 15 réactions qui sont présentes dans les deux bactéries du groupe E et sont absentes du groupe A participent essentiellement à deux processus : la synthèse de l'hème et la dégradation de la proline. La synthèse de l'hème participerait directement à la fonction symbiotique de *Wolbachia pipientis wBm* (Foster *et al.*, 2005) et de *Wigglesworthia glossinidia* (Zientz *et al.*, 2004). La dégradation de la proline, qui compte deux étapes, conduit à la formation de glutamate, acide aminé clé dans la production des autres acides aminés. Parmi les 11 bactéries intracellulaires, cette voie est présente entièrement également chez *Carsonella ruddii*. La conservation chez ces trois bactéries de cette voie et la disparition uniforme de cette voie chez les autres bactéries mutualistes est l'indice d'une synthèse différente des acides aminés. Ainsi, chez les *Buchnera*, il a été montré que le glutamate est fortement présent dans la nourriture de l'hôte et serait transporté activement dans les cellules de *Buchnera* où il serait ensuite utilisé pour la synthèse des autres acides aminés (Sasaki & Ishikawa, 1993). Zientz *et al.* (2004) notent également la présence d'un transporteur de glutamate chez *Blochmannia*.

La projection sur l'axe 3 permet de distinguer essentiellement deux groupes,

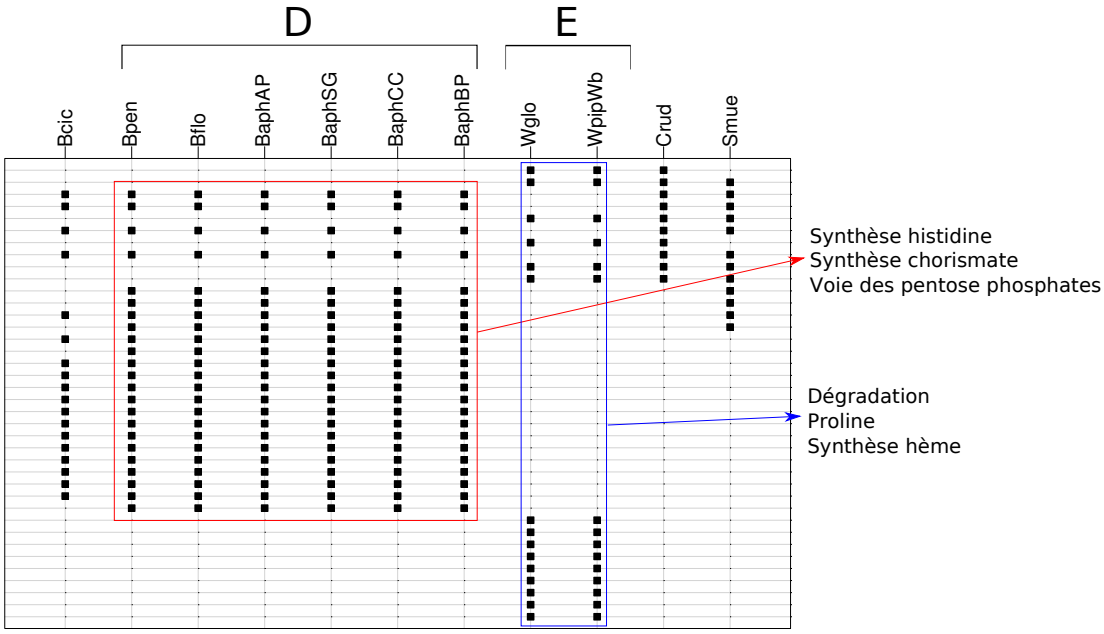

Figure 4.25. Tableau de présence des réactions ayant le pourcentage de corrélation le plus élevé sur l'axe 2 de l'ACM de la Figure 4.20. Un carré noir indique que la réaction est présente dans la bactérie correspondante. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les groupes D, et E correspondent à ceux entourés sur la Figure 4.20.

notés respectivement F et G sur la Figure 4.26.

On compte 21 réactions ayant un pourcentage de corrélation supérieur à 60 % sur le troisième axe de l'ACM. Parmi celles-ci, 15 sont présentes dans toutes les bactéries du groupe F et absentes des bactéries du groupe G, et 6 sont présentes chez toutes les bactéries du groupe G et absentes du groupe F (Figure 4.27).

Les 15 premières réactions participent essentiellement à la synthèse des nucléotides, du glutathion et de la biotine. Le glutathion est un métabolite formé de trois acides aminés : le glutamate, la cystéine et la glycine. Outre le rôle de stockage de ces acides aminés, le glutathion possède un rôle de défense contre les toxines libérées par l'hôte, en particulier les radicaux libres (données MetaCyc). La synthèse du glutathion est complètement absente des *Blochmannia* mais aussi des bactéries aux réseaux métaboliques les plus réduits. On peut voir sa disparition comme l'indice d'une intégration plus importante des bactéries.

La biotine, ou vitamine H, participerait au rôle symbiotique de *Wigglesworthia glossinidia* (Zientz *et al.*, 2004) et de *Baumannia cicadellinicola* (McCutcheon & Moran, 2007). Le rôle de la biotine chez les *Buchnera* est plus flou. Comme l'indiquaient Zientz *et al.* (2004), la synthèse de la biotine est complète chez *Buchnera aphidicola Bp* mais il manque la première étape chez *Buchnera aphidicola Sg* et *Buchnera aphidicola APS*. Elle a par ailleurs complètement disparu chez *Buchnera aphidicola Cc*.

Parmi les six réactions que l'on retrouve dans le groupe G et qui sont absentes du groupe F, on trouve trois réactions dont le numéro EC est 2.6.1.42 et qui

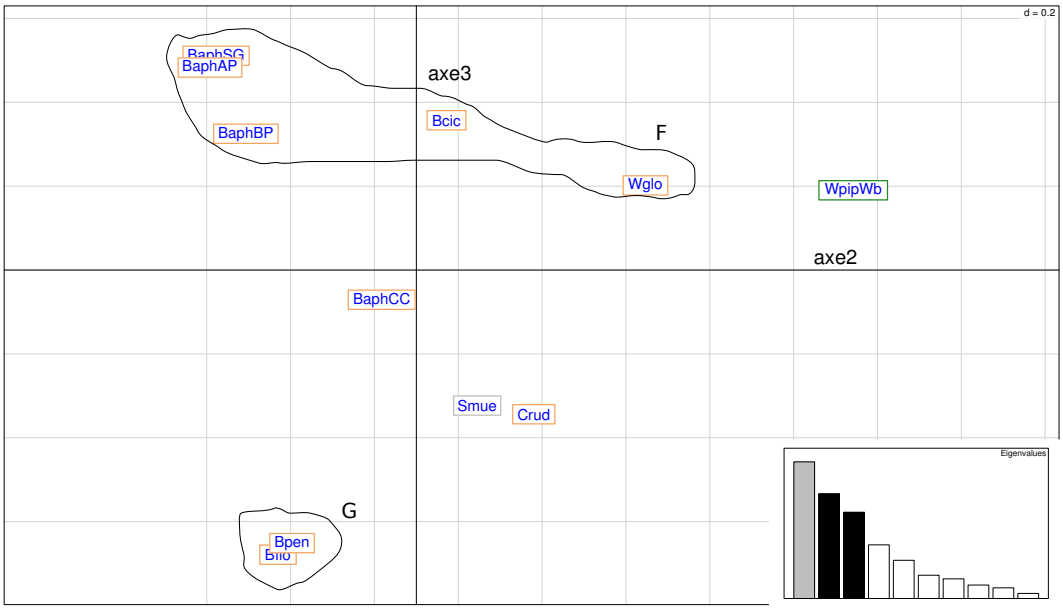

Figure 4.26. Analyse en Correspondances Multiples sur la présence des réactions chez les bactéries MIV. Projection sur les trois premiers axes. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. **Cadres**. Vert : α-protéobactéries ; Saumon : γ-protéobactéries ; Gris : Bactéroïdète. Les ensembles entourés en noir correspondent aux groupes de bactéries qui se distinguent selon le troisième axe.

sont les dernières étapes de la voie de synthèse de la valine, de la leucine et de l'isoleucine. Parmi les autres bactéries, ces trois réactions n'apparaissent que chez *Carsonella ruddii* (Crud) et *Sulcia muelleri* (Smue). Cependant, chez les *Buchnera*, ces trois réactions sont probablement catalysées par une autre enzyme différente de celle que l'on assigne classiquement à cette réaction (Shigenobu *et al.*, 2000). Nous avons d'ailleurs ajouté ces trois réactions à la reconstruction métabolique de *Buchnera aphidicola APS* utilisée pour la recherche de précurseurs par PITUFO (voir Section 5).

Les trois autres réactions que l'on retrouve dans le groupe G et qui sont absentes du groupe F participent, respectivement, à la synthèse de l'isoleucine à partir de la thréonine, à la synthèse de l'ubiquinone et à la synthèse de la ménaquinone. Parmi les autres bactéries mutualistes intracellulaires, ces réactions n'apparaissent que chez *Sulcia muelleri*.

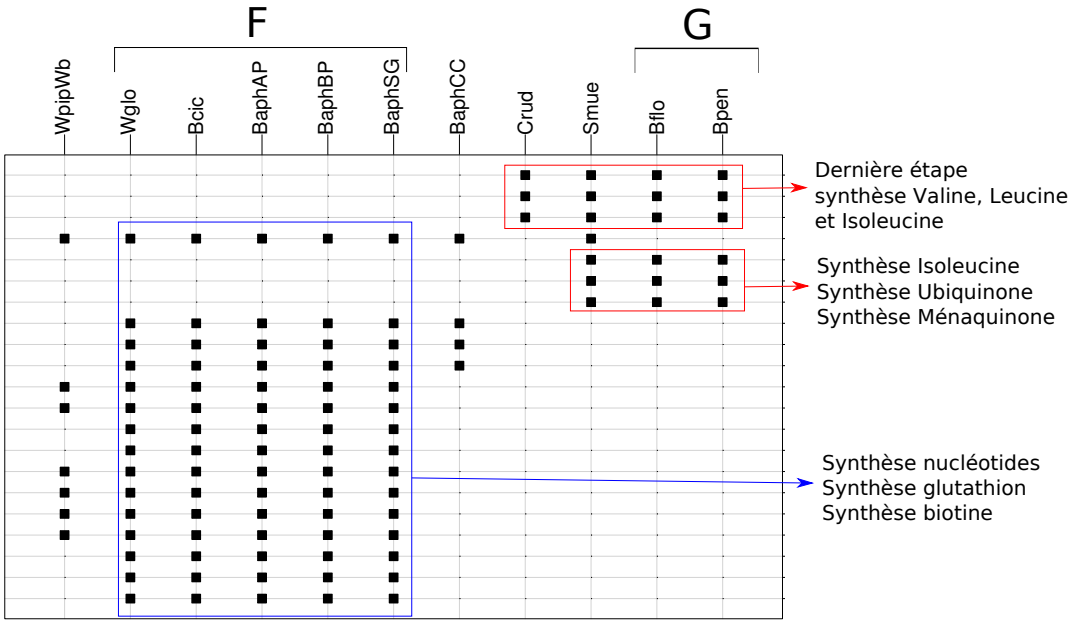

Figure 4.27. Tableau de présence des réactions ayant le pourcentage de corrélation le plus élevé sur l'axe 3 de l'ACM de la Figure 4.26. Un carré noir indique que la réaction est présente dans la bactérie correspondante. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les groupes D, et E correspondent à ceux entourés sur la Figure 4.26.

4.6 Comparaison des graphes de composés

4.6.1 Diamètre des graphes de composés

Rappelons que le diamètre est le plus long des plus courts chemins entre deux noeuds quelconques dans un graphe. Dans un graphe métabolique, le diamètre mesure donc le nombre maximum d'étapes (de réactions) pour produire un métabolite à partir d'un autre. L'idée est de rendre compte d'une certaine "compacité" du réseau.

Les diamètres mesurés sur les graphes de composés varient de 12 pour *Carsonella ruddii* (Crud) à 20 pour *Rickettsia typhi* (Rtyp) et *Rickettsia bellii* (Rbel). La moyenne et la médiane sont de 16. Il ne semble pas y avoir d'effet du nombre de noeuds sur le diamètre si nous prenons l'ensemble des bactéries (Figure 4.28, coefficient de corrélation = -0,27). Ma & Zeng (2003) observaient une augmentation du diamètre avec le nombre de noeuds sur les réseaux de petite taille, c'est-à-dire dont le nombre de noeuds était inférieur à 300. Notre jeu de données ne contient que quatre graphes de composés dont le nombre de noeuds est inférieur à 300. Nous avons donc calculé la corrélation entre le nombre de noeuds et le diamètre des dix graphes de composés ayant un nombre de noeuds inférieur à 350. Le taux de corrélation reste faible puisqu'il n'atteint que 0,53.

Même si on note une plus grande variation du diamètre chez les bactéries intracellulaires, il semble difficile également de relier cette mesure au style de vie des bactéries. On aurait pu penser en effet que la réduction du réseau s'accompagnait d'une réduction du diamètre mais le lien est difficile à établir ici. Mis à part le faible diamètre de *Carsonella ruddii*, les autres valeurs ne diffèrent pas beaucoup de la moyenne. Par ailleurs, il est facile de voir que le diamètre est une mesure très sensible. La différence de diamètre entre des réseaux très similaires, comme ceux des deux *Wolbachia* ou des deux *Blochmannia* en témoigne. En effet, l'ajout d'une seule réaction dans le réseau peut augmenter le diamètre, si cette réaction produit un métabolite qui ne peut être synthétisé qu'à partir d'un seul substrat.

Mentionnons également l'effet important du filtre appliqué sur les cofacteurs et de l'assignation des directions de réactions puisque le diamètre mesuré sur les graphes non filtrés ne varie que de 7 à 10.

4.6.2 Distance moyenne entre deux noeuds dans les graphes de composés

La distance entre deux noeuds est la longueur du plus court chemin entre ces deux noeuds.

Dans un graphe des composés, la distance moyenne entre deux noeuds exprime le nombre de réactions qu'il faut utiliser en moyenne pour produire un métabolite à partir d'un autre. L'hypothèse est ici que la production d'un métabolite en un

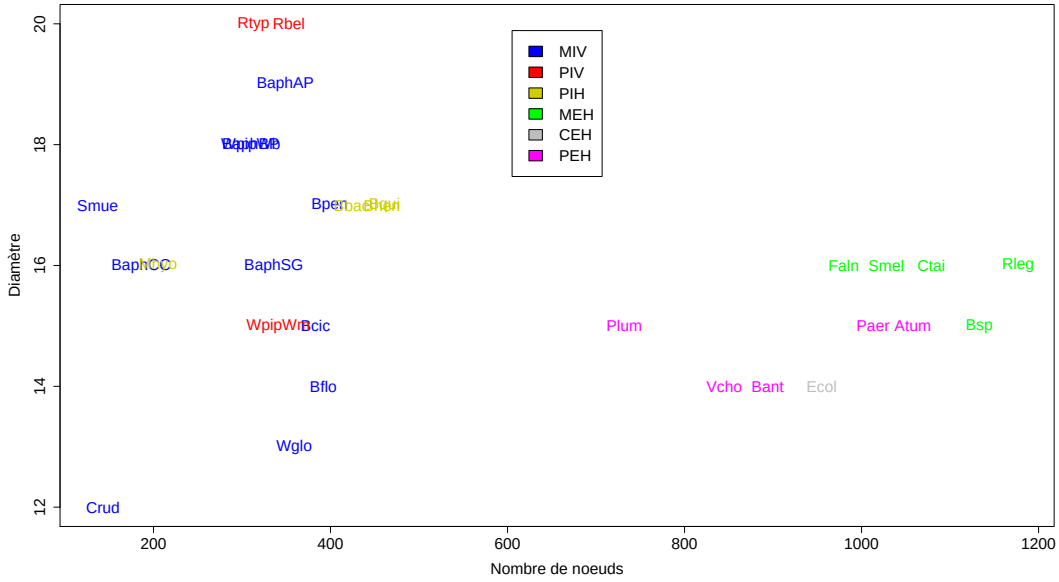

Figure 4.28. Diamètre des graphes de composés dont les cofacteurs ont été filtrés et dont la direction des réactions a été assignée. Les abbréviations des noms d'espèces et des groupes de style de vie correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome.

minimum d'étapes est sélectionnée. La distance moyenne entre deux noeuds ne s'étend que de 4 pour *Carsonella ruddii* (Crud) à 7,4 pour *Rickettsia typhi* (Rtyp). La moyenne est de 5,7 et la médiane de 5,5. Le diamètre et la distance moyenne sont ici corrélés de façon importante (taux de corrélation = 0,8). Comme pour le diamètre, nous observons que la distance moyenne est relativement constante chez les bactéries extracellulaires et plus variée chez les bactéries intracellulaires (Figure 4.29).

Ma & Zeng (2003) notent également un effet du nombre de noeuds sur la moyenne des distances, spécialement dans les réseaux de petite taille, c'est-à-dire ceux dont le nombre de noeuds est inférieur à 300. Comme pour le diamètre, nous ne retrouvons pas cet effet ici. De même, Ma & Zeng (2003) observent des valeurs considérablement plus élevées pour leurs réseaux. Par exemple, les auteurs indiquent une valeur de 8.2 pour la distance moyenne entre deux noeuds dans le graphe des composés d'*Escherichia coli K12* alors que nous trouvons une valeur de 5.3. Les auteurs assignent les directions aux réactions et traitent les cofacteurs d'une manière différente de la nôtre. Nous ne connaissons pas les détails de leur filtre mais le fait que nous trouvons des valeurs plus faibles peut s'expliquer par un nombre plus faible de cofacteurs retirés du réseau, ou un nombre plus faible de réactions assignées comme irréversibles, ce qui tendrait à raccourcir les distances entre les métabolites.

Il est intéressant d'examiner également la distribution des distances entre deux noeuds dans les réseaux métaboliques (Figure 4.30). De façon visuelle et à l'aide

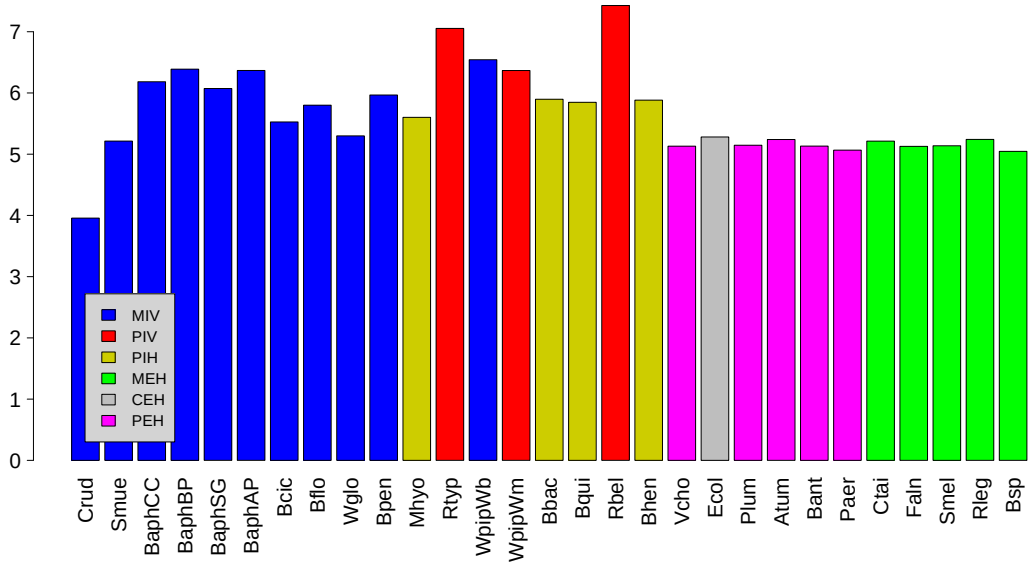

Figure 4.29. Distance moyenne entre deux noeuds des graphes de composés dont les cofacteurs ont été filtrés et dont la direction des réactions a été assignée. Les abbréviations des noms d'espèces et des groupes de style de vie correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome.

des coefficients d'aplatissement et d'assymétrie (notés KU et SK)[6], nous pouvons comparer la forme des distributions.

Sur la Figure 4.30, on peut voir que toutes les distributions s'étalent, à un degré différent selon les bactéries, vers la droite. Ceci signifie que l'ensemble des bactéries présente une minorité de couples de métabolites séparés par un grand nombre d'étapes. Ainsi, la valeur du diamètre correspond à une proportion très faible des distances et ne correspond pas à une tendance moyenne du graphe.

Certaines distributions se rapprochent d'une distribution normale, comme celle de *Wolbachia pipientis wMel* (WpipWM), d'autres présentent un plateau important comme celle de *Rickettsia typhi* (Rtyp) et d'autres encore présentent une distribution beaucoup plus pointue. C'est le cas des bactéries extracellulaires chez lesquelles la forme des distributions est relativement constante. Elle est plus

[6]Le coefficient d'assymétrie (ou Skewness) mesure l'assymétrie d'une distribution. Il est défini par $SK = \frac{\mu_3}{\sigma^3}$ où μ_3 désigne le moment centré d'ordre 3 et σ désigne l'écart type. Un coefficient d'assymétrie positif indique une queue de distribution étalée vers la droite. Un coefficient négatif indique une queue de distribution étalée vers la gauche. Une distribution normale a un coefficient de distribution nul.

Le coefficient d'aplatissement (ou Kurtosis) correspond à une mesure de l'aplatissement, ou au contraire du caractère pointu d'une distribution. Il est défini par $KU = \frac{\mu_4}{\sigma^4} - 3$ où μ_4 désigne le moment centré d'ordre 4 et σ désigne l'écart type. Un coefficient d'aplatissement positif indique une forme pointue de la distribution et un coefficient d'aplatissement négatif indique une forme aplatie de la distribution. Une distribution normale a un coefficient d'aplatissement nul.

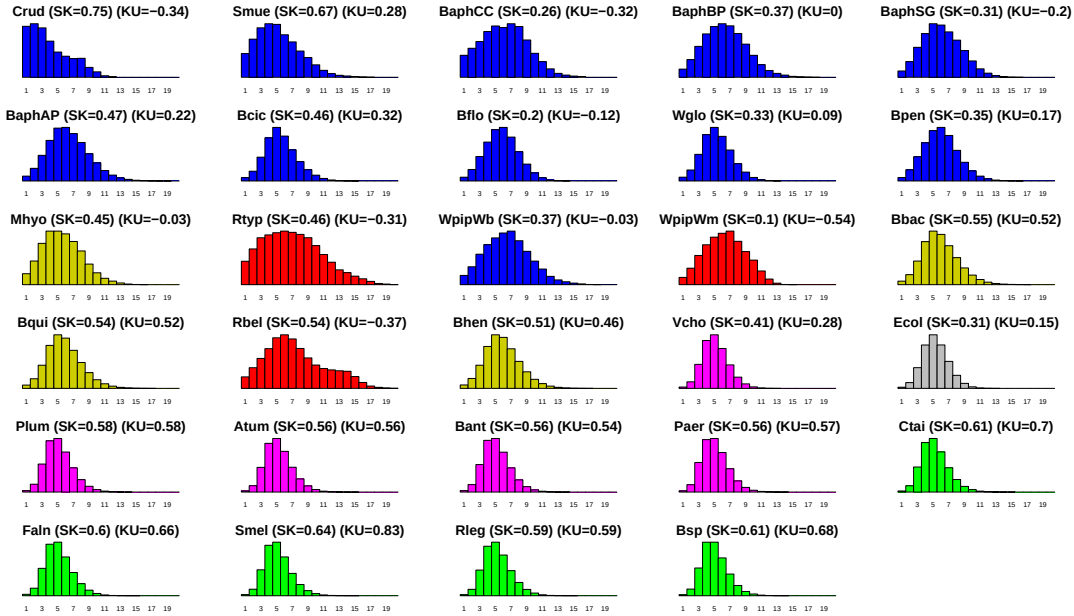

Figure 4.30. Distribution des distances entre deux noeuds des graphes de composés dont les cofacteurs ont été filtrés et dont la direction des réactions a été assignée. On trouve en abcisses la valeur des distances et en ordonnées la proportion de plus courts chemins ayant la longueur correspondante. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome. Bleu : Mutualistes Intracellulaires à transmission Verticale. Rouge : Parasites Intracellulaires à transmission Verticale. Jaune : Parasites Intracellulaires à transmission Horizontale. Magenta : Parasites Extracellulaires à transmission Horizontales. Gris : Commensalistes. Vert : Mutualistes Extracellulaires à transmission Horizontale. SK : Coefficient d'assymétrie. KU : coefficient d'aplatissement.

variée chez les bactéries intracellulaires. Dans le cas de *Carsonella ruddii*, et à un moindre niveau de *Sulcia muelleri* (Smue) et *Buchnera aphidicola Cc* (BaphCC), une grande proportion de couples de métabolites sont situés à très courte distance. En revanche, on retrouve une distribution considérablement étalée chez les 2 Rickettsies et les 2 *Wolbachia*.

Enfin, on peut noter que certains graphes ayant un diamètre et une distance moyenne équivalents ont une distribution de leur distances considérablement différente. C'est le cas par exemple de *Wolbachia pipientis wMel* et *Baumannia cicadellinicola*.

La distribution des distances entre métabolites complète la mesure du diamètre, suggérant ainsi des différences dans la topologie des réseaux métaboliques des bactéries intracellulaires alors qu'elle semble relativement constante chez les bactéries extracellulaires. La mesure du degré et de la centralité d'interposition va permettre de préciser la nature de ces différences.

4.6.3 Connectivité dans les graphes de composés

La connectivité dans un graphe se mesure par la valeur du degré des noeuds. Le degré d'un noeud représente le nombre d'arêtes liées à ce noeud. Dans le cas

d'un graphe dirigé, on peut distinguer le degré entrant d'un noeud, qui compte le nombre d'arêtes dirigées vers ce noeud, et son degré sortant, qui compte le nombre d'arêtes qui partent de ce noeud. Dans le cas d'un graphe des composés, le degré entrant d'un métabolite représente le nombre de métabolites qui sont les substrats des réactions qui le produisent. Le degré sortant d'un métabolite représente le nombre de métabolites produits par les réactions dont il est un des substrats. Les filtres que nous avons appliqués sur les graphes des composés changent évidemment l'allure des distributions. L'assignation des directions permet de distinguer les degrés entrants des degrés sortants, ce qui est impossible lorsque toutes les réactions sont laissées réversibles. Ainsi, dans le graphe des composés dont toutes les réactions sont laissées réversibles, le degré (entrant ou sortant) de l'ATP est 53. Si on assigne les directions des réactions comme expliqué dans la Section 3, on trouve cette fois que l'ATP a un degré entrant de 15 et un degré sortant de 52. L'eau, souvent le métabolite le plus connecté (elle apparaît environ dans un quart des réactions), est complètement supprimé des réseaux par notre filtre. Dans les graphes non filtrés, ce sont les cofacteurs tels que l'ATP ou le NADH qui sont les plus connectés. Ici, ils n'apparaissent que dans les réactions où ils ne participent pas en tant que cofacteur. Ainsi, après le filtrage des cofacteurs dans le graphe des composés de *Buchnera aphidicola APS*, le degré entrant de l'ATP n'est plus que de 2 et le degré sortant seulement de 8.

Malgré les filtres, les distributions des degrés entrants représentées sur la Figure 4.31 semblent correspondre aux distributions décrites lors d'études précédentes de graphes de composés (voir Section 4.2). En effet, on retrouve bien à chaque fois une très forte proportion de noeuds très peu connectés et une faible proportion de noeuds avec un degré entrant important. Par ailleurs, nous remarquons que la queue de distribution est d'autant plus longue que le graphe compte de noeuds. Ceci signifie que malgré le fitre appliqué sur les réseaux, il reste toujours une faible portion de métabolites fortement connectés. On retrouve les mêmes formes de distributions pour les degrés sortants (données non montrées).

Dans un graphe des composés, les métabolites les plus connectés peuvent nous renseigner sur leur importance dans les réseaux métaboliques. Nous avons donc sélectionné les 10 métabolites les plus connectés dans chaque graphe métabolique et considéré leur degré entrant et leur degré sortant. Le nombre de métabolites correspondant à l'union des 10 métabolites aux degrés entrants les plus élevés dans chaque graphe des composés, s'élève à 30 (Figure 4.32). Notons que le même nombre sur les graphes non filtrés s'élève seulement à 16.

Si on effectue les mêmes mesures uniquement sur les 11 bactéries extracellulaires, on obtient une liste de seulement 15 métabolites dont le degré entrant varie très peu selon la bactérie extracellulaire considérée (cadres rouges dans la Figure 4.32). On retrouve les 31 métabolites identifiés avec l'ensemble des bactéries si on effectue la même manipulation seulement sur les bactéries intracellulaires. Les métabolites dont les degrés entrants sont les plus élévés sont donc très conservés chez les bactéries extracellulaires mais très variables chez les bactéries intracellu-

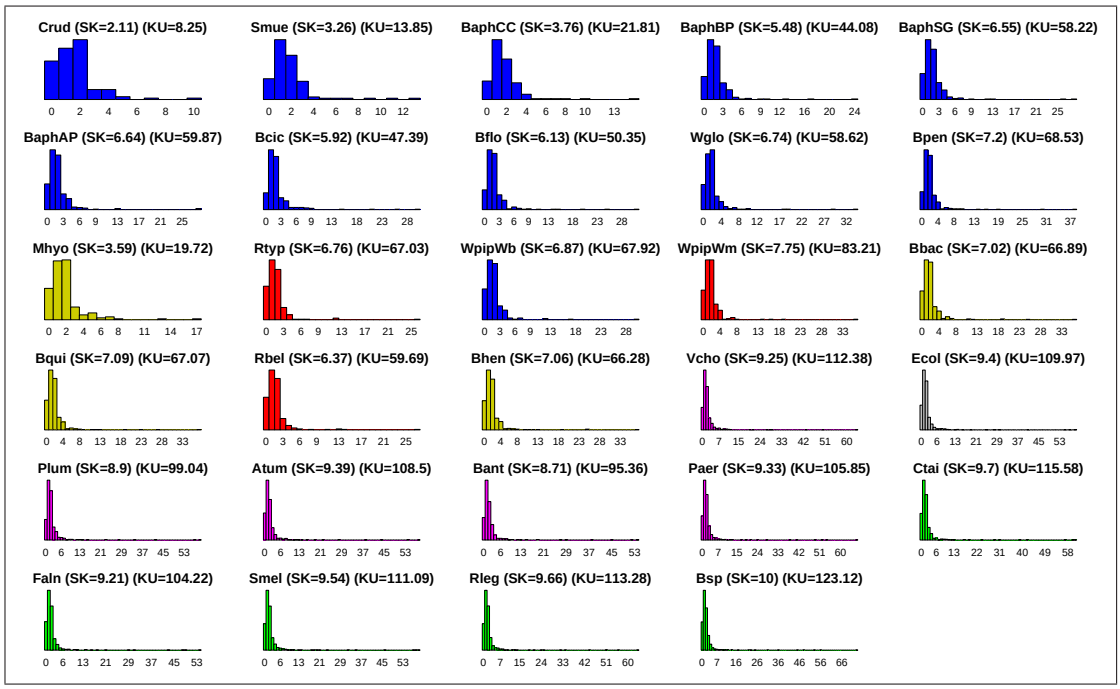

Figure 4.31. Distribution des degrés entrants des graphes de composés dont la direction des réactions a été assignée. On trouve en abcisses la valeur des degrés entrants et en ordonnées la proportion de noeuds ayant le degré entrant correspondant. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome. Bleu : Mutualistes Intracellulaires à transmission Verticale. Rouge : Parasites Intracellulaires à transmission Verticale. Jaune : Parasites Intracellulaires à transmission Horizontale. Magenta : Parasites Extracellulaires à transmission Horizontales. Gris : Commensalistes. Vert : Mutualistes Extracellulaires à transmission Horizontale. SK : Coefficient d'assymétrie. KU : coefficient d'aplatissement.

laires.

La sélection des 10 métabolites aux degrés sortants les plus élevés chez les bactéries extracellulaires retourne une liste de seulement 12 métabolites dont les degrés sortants sont très similaires en fonction de la bactérie extracellulaire considérée (cadres rouges dans la Figure 4.33). La même manipulation pour les bactéries intracellulaires retourne une liste de 35 métabolites qui peuvent avoir des degrés sortants très divers en fonction de la bactérie intracellulaire considérée. Parmi ces 35 métabolites, on retrouve les 12 métabolites précédemment cités (Figure 4.33).

Il est difficile de comparer les graphes métaboliques par la mesure des degrés des composés. En effet, le degré d'un métabolite, surtout quand il est très fréquent dans les réseaux métaboliques, comme le phosphate ou le diphosphate, sera fortement lié à la taille du réseau lui-même. Plutôt que de tenter de comparer l'ensemble des degrés des métabolites les plus connectés, nous allons seulement examiner quelques cas extrêmes. En effet, certains métabolites ayant un degré entrant ou un degré sortant élevé dans certains organismes, ont un degré entrant très faible ou même nul dans d'autres organismes. C'est le cas du diphosphate absent du graphe filtré de *Carsonella ruddii* (Crud) alors que son degré (entrant

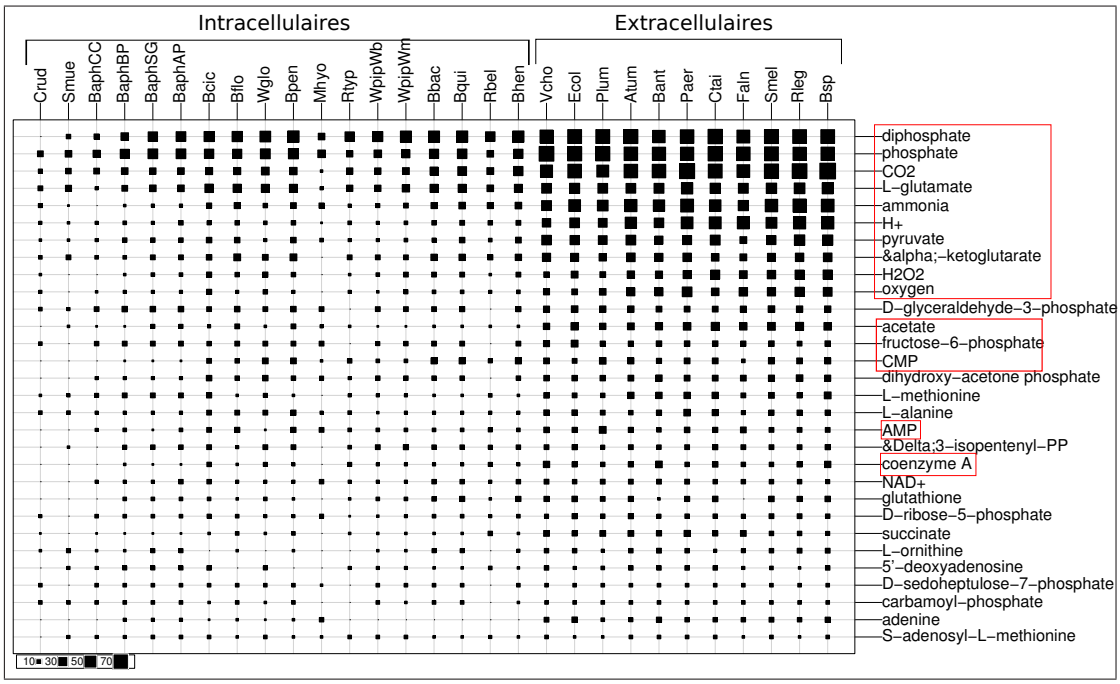

Figure 4.32. Degré entrant des métabolites correspondant à l'union des 10 métabolites aux degrés entrants les plus élevés dans les graphes de composés filtrés. Les composés entourés en rouge correspondent aux 15 métabolites qui sont l'union des 10 métabolites aux degrés entrants les plus élevés dans chacun des graphes de composés filtrés des bactéries extracellulaires. La taille de chaque carré est proportionnelle au degré entrant du métabolite dans la bactérie correspondante. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome. Les métabolites sont ordonnés en fonction de la moyenne de leur degré entrant relatif sur l'ensemble des graphes de composés.

ou sortant) est considérablement élevé chez les autres organismes. Rappelons que ce composé est filtré dans les réactions où il intervient avec l'AMP pour produire de l'ATP ou être produit par ce dernier (Tableau 3.1 p.76) mais il intervient par ailleurs dans de nombreuses voies métaboliques. Chez *Carsonella ruddii*, le diphosphate est même complètement absent des graphes filtrés, ce qui signifie qu'il n'intervient dans le réseau de cette bactérie qu'en tant que sous-produit de la transformation de l'ATP en AMP. De même, l'ammoniac est complètement absent du réseau métabolique de *Buchnera aphidicola Cc* (BaphCC) alors qu'il est produit et utilisé par de nombreuses réactions chez les bactéries libres. On peut voir également qu'il est produit mais non utilisé (même au-delà du réseau des petites molécules) chez *Rickettsia typhi* et *Sulcia muelleri*, ce qui signifie qu'il n'est qu'un sous-produit des réactions qui le produisent. On peut supposer que chez ces organismes, cet ammoniac est diffusé à l'extérieur de leur cellule.

L'α-kétoglutarate, par son rôle dans le cycle de Krebs et/ou dans la synthèse des acides aminés, a un degré entrant ou sortant en général élevé dans les bactéries de notre jeu de données. Une exception est *Mycoplasma hyopneumoniae* (Mhyo) où il est complètement absent, indice d'une dégradation profonde de ces deux processus métaboliques.

Le péroxyde d'hydrogène (H_2O_2) est un composé toxique le plus souvent produit par les réactions utilisant de l'oxygène et dégradé ensuite de nouveau en oxygène et en eau. On le retrouve parmi les métabolites aux degrés entrants les plus élevés mais comme il n'est pas utilisé ensuite, il n'apparaît pas parmi les métabolites aux degrés sortants les plus élevés. Il est complètement absent chez *Buchnera aphidicola Cc* (BaphCC), *Sulcia muelleri* (Smue) et *Mycoplasma hyopneumoniae* (Mhyo). Chez ces deux dernières, son absence coïncide avec celle de l'oxygène dans leur métabolisme.

Le glycéraldéhyde-3-Phosphate (GAP) est un intermédiaire de la glycolyse et de la gluconéogénèse. Il intervient également en tant que sous-produit lors de la synthèse du tryptophane et en tant que substrat lors de la synthèse de la vitamine B1 (thiamine). Alors que son degré est particulièrement élevé dans l'ensemble des bactéries, il est complètement absent des deux Rickettsies (Rtyp et Rbel), ce qui montre une dégradation de ces processus métaboliques, ce que nous avons déjà observé plus haut lors de la comparaison des ensembles de réactions. Le fructose 6-phosphate, lui aussi un intermédiaire de la glycolyse et de la gluconéogénèse est absent des 2 Rickettsies mais aussi de *Sulcia muelleri* (Smue) où ces deux voies sont très fortement dégradées. Le dihydroxy-acétone-phosphate, également un intermédiaire de ces deux processus, n'est pas produit chez *Rickettsia bellii* (Rbel), *Sulcia muelleri* (Smue) et *Carsonella ruddii* (Crud).

Le ribose-5-phosphate, le sedoheptulose et le xylulose-5-phosphate qui participent à la voie des pentoses phosphates ne sont pas non plus produits chez *Sulcia muelleri*. Le premier n'est de même pas produit chez *Rickettsia typhi*, le second chez *Rickettsia bellii*, et le troisième chez aucune des deux Rickettsies.

Les cofacteurs tels que la coenzyme A et le NAD+ ne semblent pas produits chez les trois bactéries aux réseaux métaboliques les plus dégradés (BaphCC, Smue et Crud). La coenzyme A ne semble pas produite non plus dans le réseau métabolique de *Mycoplasma hyopneumoniae* (Mhyo). Ces métabolites sont pourtant utilisés en tant que cofacteurs chez ces bactéries, ce qui indique un approvisionnement possible de la part de l'hôte.

On peut voir également que certains composés participant à la biosynthèse de ceux intervenant dans la relation mutualiste restent parmi les plus connectés. C'est le cas du carbamoyl-phosphate, du chorismate, de l'ornithine, du 2-kéto-isovalérate et du 2-kéto-3-méthylvalérate qui participent à la synthèse d'acides aminés essentiels. Leur degré entrant ou sortant reste ainsi élevé chez les bactéries libres et celles dont la fonction symbiotique est de fournir les acides aminés à leur hôte. Par contre, ces métabolites disparaissent souvent des réseaux métaboliques des bactéries intracellulaires qui n'ont pas ce rôle symbiotique.

Par ailleurs, l'adénine, l'AMP (adénosine monophosphate), le CMP (cytidine monophosphate) et le NAD+ (nicotinamide adénine dinucléotide) sont parmi les 10 métabolites aux degrés les plus élevés chez *Mycoplasma hyopneumoniae*, ce qui suggère une importance particulière des nucléotides dans le métabolisme de *Mycoplasma hyopneumoniae*.

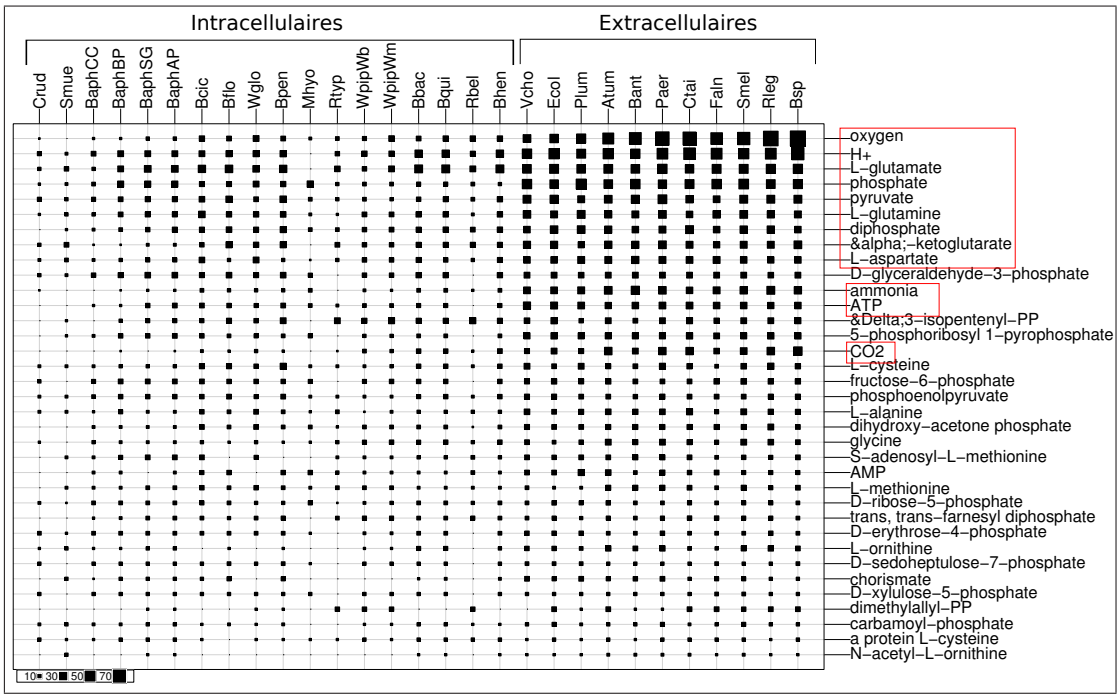

Figure 4.33. Degré sortant des métabolites correspondant à l'union des 10 métabolites les plus connectés dans les graphes de composés filtrés. Les composés entourés en rouge representent les 12 métabolites formant l'union des 10 métabolites aux degrés sortants les plus élevés dans les graphes de composés filtrés des bactéries extracellulaires. La taille de chaque carré est proportionnelle au degré sortant du métabolite dans la bactérie correspondante. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome. Les métabolites sont ordonnés en fonction de la moyenne de leur degré entrant relatif sur l'ensemble des graphes de composés.

Dans l'optique d'étudier la réduction des réseaux métaboliques, il est intéressant également d'avoir une idée plus générale de la proportion de métabolites ayant un degré entrant ou sortant nul. Un métabolite ayant un degré entrant nul n'est pas produit par la bactérie et serait donc puisé dans l'environnement. Un métabolite ayant un degré sortant nul ne produit aucun autre métabolite au sein de la bactérie. Son destin peut donc être de participer à biomasse de la bactérie ou bien de subvenir aux besoins de l'hôte dans le cas d'une association mutualiste.

De façon surprenante, en dehors de quelques bactéries, la proportion de métabolites ayant un degré entrant nul est globalement similaire à la proportion de métabolites ayant un degré sortant nul (Figure 4.34). La plus forte proportion de degrés nuls peut être expliquée par une absorption de nutriments diversifiés et par la libération dans l'environnement de métabolites de natures différentes, ce qui pourrait être le cas des mutualistes extracellulaires de notre jeu de données, ou par des voies métaboliques particulièrement morcelées, ce qui pourrait être le cas de *Carsonella ruddii*.

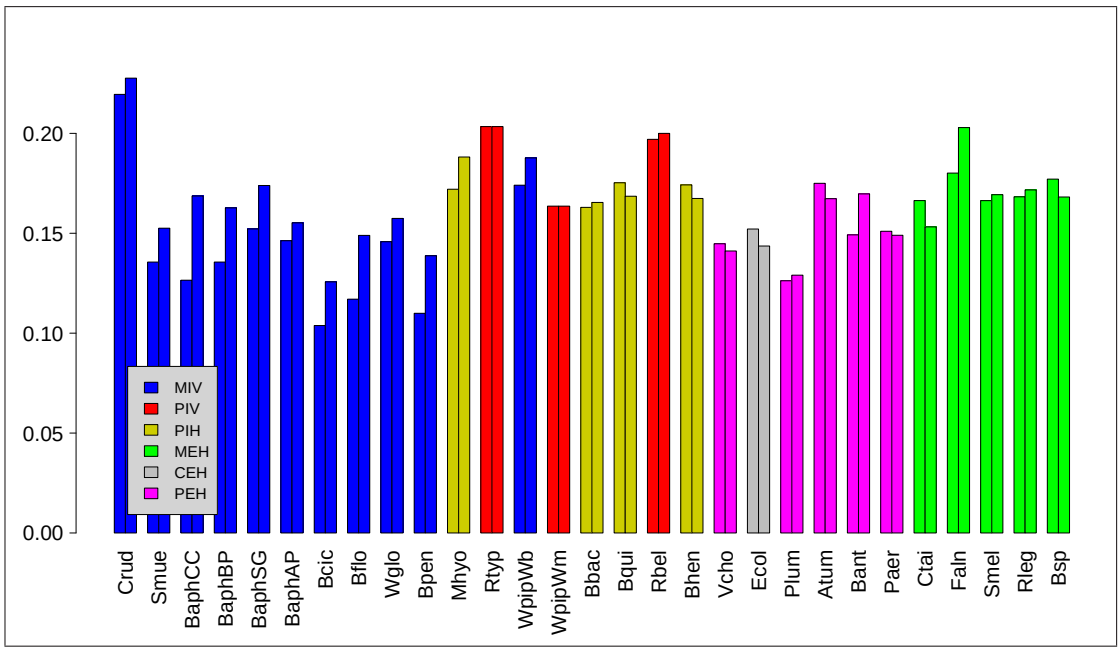

Figure 4.34. Proportion de noeuds ayant un degré entrant nul et un degré sortant nul dans les graphes de composés dont la direction des réactions a été assignée. Pour chaque organisme, la barre de gauche correspond au degré entrant et la barre de droite au degré sortant. Les abbréviations des noms d'espèces et des groupes de style de vie correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome.

4.6.4 La centralité d'interposition des noeuds dans les graphes des composés

Comme nous l'avons indiqué dans la section 2.2 p.44, la centralité d'interposition mesure la proportion des plus courts chemins passant par un noeud i sur le nombre total de plus courts chemins allant d'un noeud u à un noeud v pour tous u et v. Elle peut être normalisée par le nombre de plus courts chemins ne passant pas par i, c'est-à-dire $\frac{(n-1)(n-2)}{2}$ si n est le nombre de noeuds du graphe.

Dans un graphe des composés, la centralité d'interposition élevée d'un noeud peut signifier que de nombreux chemins métaboliques empruntent ce noeud. Par rapport au degré, qui est une mesure locale, l'intérêt de la centralité d'interposition est de prendre en compte également l'ensemble du réseau pour mesurer la centralité d'un métabolite. Par contre, la centralité d'interposition est inutile pour étudier la périphérie du métabolisme, comme nous pouvons le faire avec les degrés entrants ou les degrés sortants dont la valeur est nulle. En effet, un métabolite ayant un degré entrant ou sortant égal à 0 aura une centralité d'interposition nulle puisqu'aucun chemin ne le traversera. Ainsi, la distribution des centralités d'interposition montre une majorité de valeurs nulles, correspondant aux métabolites ayant un degré entrant ou sortant nul.

Ces valeurs nulles mises à part, on observe une majorité de métabolites à la centralité d'interposition faible et une minorité de métabolites très centraux (Figure 4.35). On remarque également que la valeur de centralité la plus haute est très différente selon les bactéries (Figure 4.36). Ainsi, on peut dire qu'il n'existe pas de métabolite central, ni chez *Carsonella ruddii* (Crud) ni chez *Rickettsia typhi* (Rtyp). Cette constatation nous donne déjà un aperçu de la topologie du graphe. Nous allons maintenant regarder plus en détail quels sont les métabolites indiqués comme plus centraux avec cette mesure en fonction des organismes.

L'union des 10 métabolites les plus centraux dans chaque réseau métabolique s'élève à 50 (Figure 4.37). Ce nombre ne s'élève qu'à 15 si on ne considère que les bactéries extracellulaires et à 48 si on ne considère que les bactéries intracellulaires. Seuls le péroxyde d'hydrogène et l'acétate se retrouvent parmi les métabolites les plus centraux des bactéries extracellulaires et ne se retrouvent pas parmi les métabolites les plus centraux des bactéries intracellulaires. Par ailleurs, ces deux métabolites, qu'on retrouvait parmi les métabolites les plus connectés de certaines bactéries intracellulaires, illustrent le fait que certains noeuds peuvent se retrouver parmi les plus connectés sans être parmi les plus centraux.

Les 15 métabolites les plus centraux chez les bactéries extracellulaires ont une valeur de centralité très proche quelle que soit la bactérie extracellulaire considérée. On retrouve d'ailleurs dans ces 15 métabolites la plupart des métabolites les plus connectés chez les même bactéries extracellulaires.

Certains métabolites, pourtant, peuvent présenter une centralité considérablement élevée et être peu connectés. On peut comparer ces métabolites à des

"passerelles" entre plusieurs parties plus connectées du réseau (souvent désignées sous le terme de modules).

Ainsi, l'oxaloacétate est compté parmi les métabolites les plus centraux et n'apparaît pas parmi les métabolites les plus connectés. Ce métabolite est un composé clé dans de nombreuses voies métaboliques telles que le cycle de Krebs, la gluconéogénèse et la synthèse de l'aspartate. Cependant, il n'est produit ou utilisé que par quelques réactions, ce qui explique son degré faible.

De même, le 1,3-diphosphateclycérate ne fait pas partie des métabolites les plus connectés mais apparaît comme l'un des métabolites les plus centraux. Par exemple, chez *Escherichia coli K12*, il n'est produit et consommé que par deux réactions réversibles intervenant dans la glycolyse et dans la gluconéogénèse. Sa centralité est plutôt basse chez les bactéries extracellulaires mais est considérablement élevée chez les bactéries intracellulaires où ces deux voies ont été conservées. La disparition d'autres voies et l'importance de celles-ci augmentent la centralité des métabolites qui la composent.

On remarque également que l'ATP, en dehors des réactions où il intervient en tant que cofacteur, montre une centralité remarquable chez *Buchnera aphidicola APS* (BaphAP) et *Baumannia cicadellinicola* (Bcic). Chez la première, nous verrons en effet avec l'analyse des précurseurs que l'ATP pourrait jouer un rôle central dans la synthèse de certains acides aminés (voir Section 5). On peut noter aussi une centralité élevée de l'UTP chez *Wigglesworthia glossinidia* (Wglo) et chez les trois *Bartonella* (Bqui, Bhen, Bbac).

Le pyridoxal 5-phosphate, une vitamine B6, que l'on ne retrouvait pas parmi les métabolites les plus connectés, figure parmi les métabolites les plus centraux chez les 2 *Blochmannia* (Bflo et Bpen), indiquant potentiellement un rôle central de ce composé dans le métabolisme de ces bactéries. Notons enfin la forte centralité de l'octaprenyl diphoshate et du 3-octaprenyl-4-hydroxybenzoate chez *Rickettsia typhi* (Rtyp), *Wolbachia pipientis wBm* (WpipWB) et *Wolbachia pipientis wMel* (WpipWM). Ces métabolites participent à la synthèse de l'ubiquinone et de la ménaquinone.

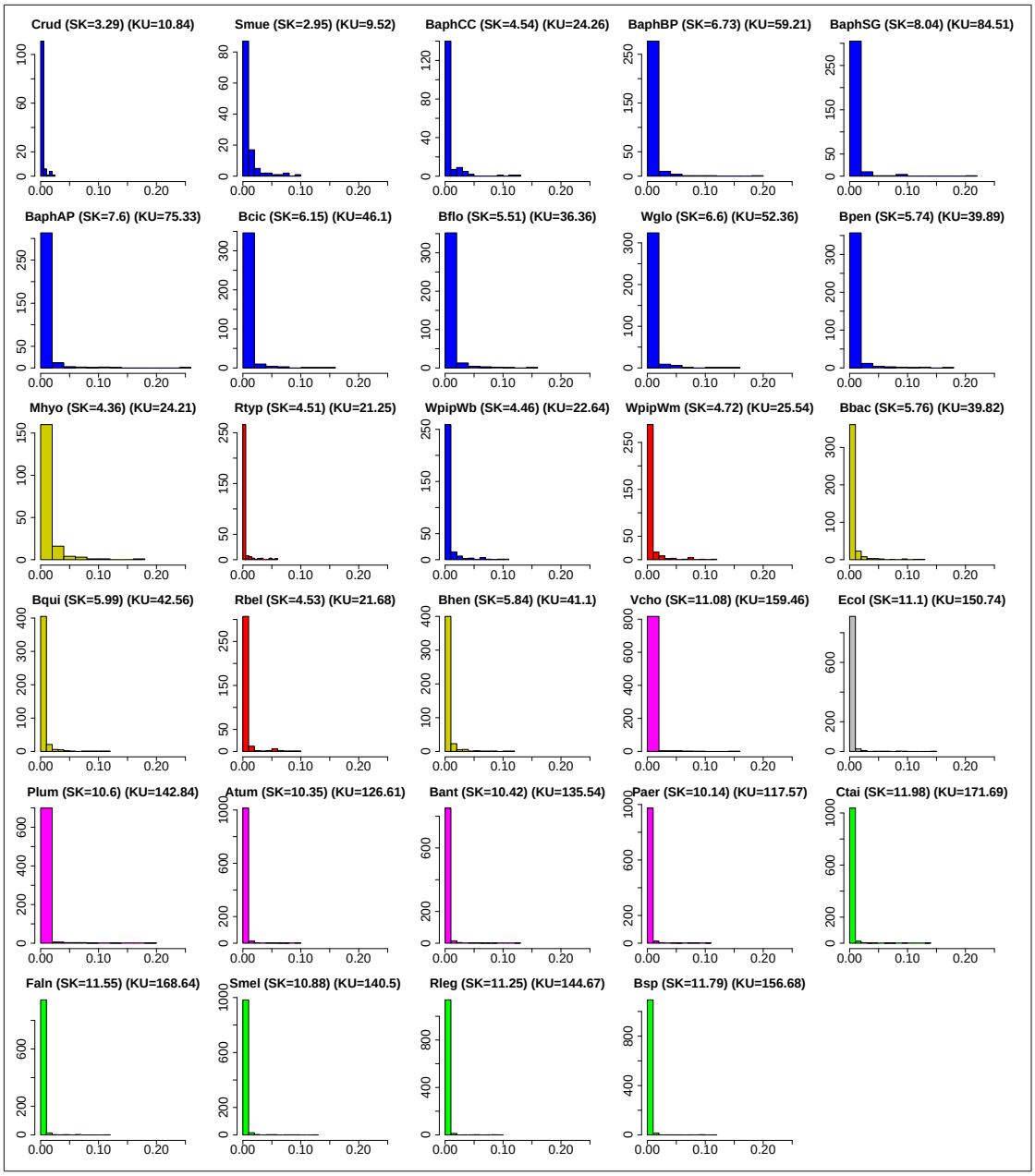

Figure 4.35. Distribution des centralités d'interposition des graphes de composés dont la direction des réactions a été assignée. Il y a une barre par valeur de degré et la hauteur de chaque barre correspond à la proportion de noeuds ayant le degré entrant correspondant. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome. Bleu : Mutualistes Intracellulaires à transmission Verticale. Rouge : Parasites Intracellulaires à transmission Verticale. Jaune : Parasites Intracellulaires à transmission Horizontale. Magenta : Parasites Extracellulaires à transmission Horizontales. Gris : Commensalistes. Vert : Mutualistes Extracellulaires à transmission Horizontale. SK : Coefficient d'assymétrie. KU : coefficient d'aplatissement.

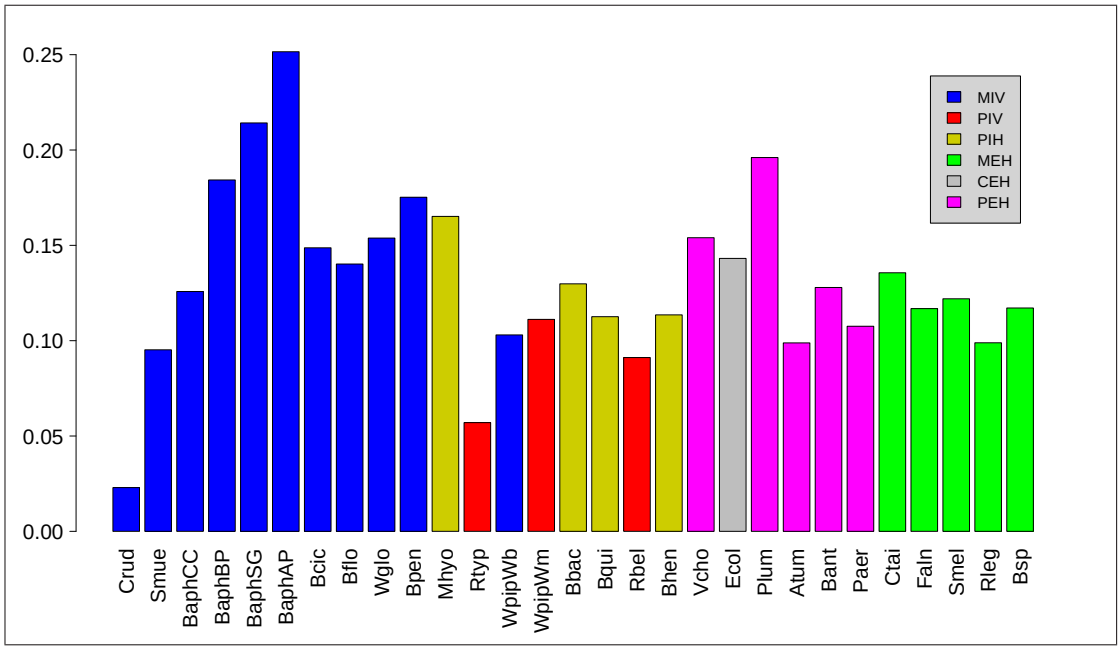

Figure 4.36. Centralité d'interposition maximale dans les graphes de composés dont les cofacteurs ont été filtrés et dont la direction des réactions a été assignée. Les abbréviations des noms d'espèces et des groupes de style de vie correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome.

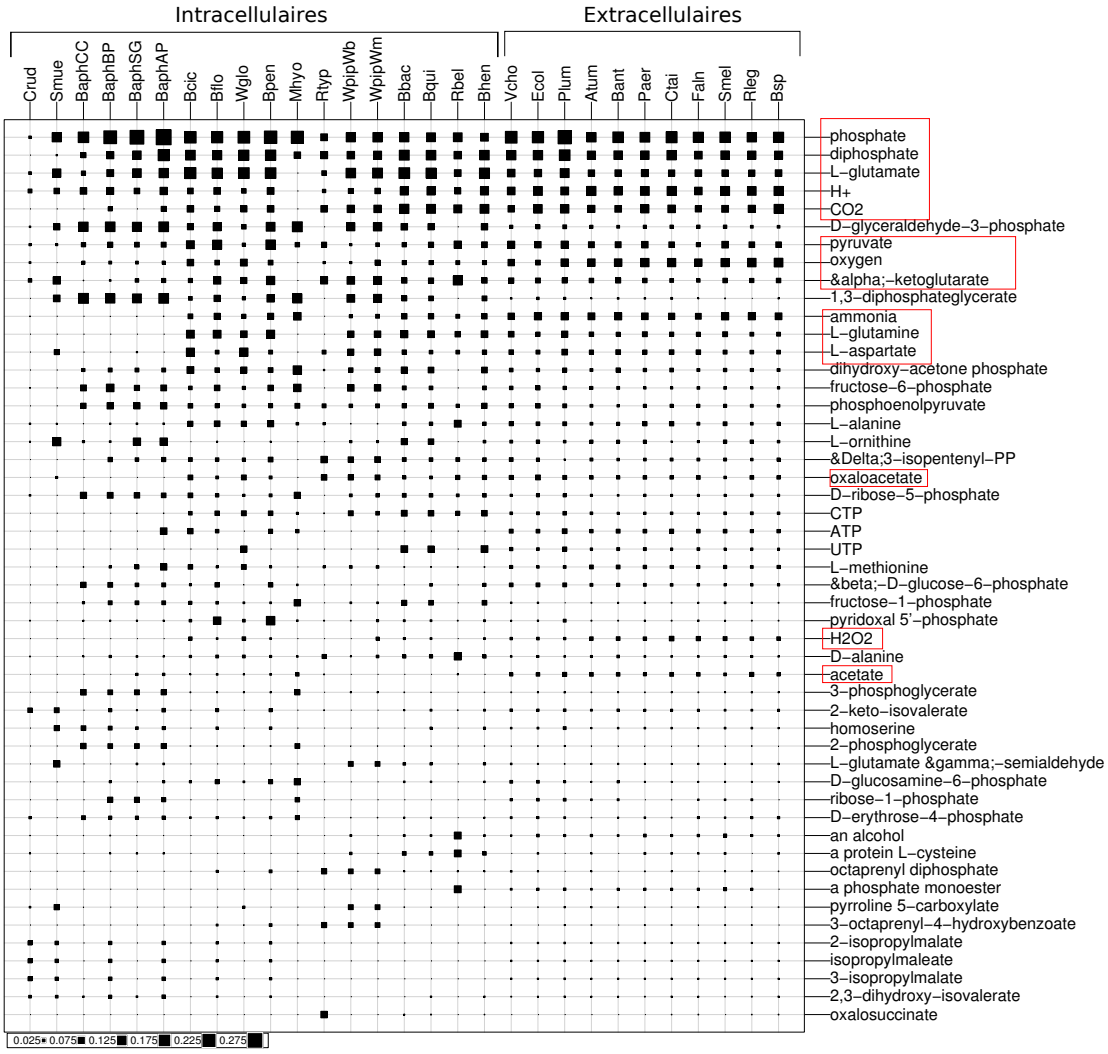

Figure 4.37. Centralité d'interposition des métabolites correspondant à l'union des 10 métabolites les plus centraux selon cette mesure dans les graphes de composés filtrés. Les composés entourés en rouge correspondent aux 15 métabolites correspondant à l'union des 10 métabolites les plus centraux dans les graphes de composés filtrés des bactéries extracellulaires. La taille de chaque carré est proportionnelle avec la centralité du métabolite dans la bactérie correspondante. Les abbréviations des noms d'espèces correspondent à celles données dans la Figure 4.1 p.90. Les organismes sont ordonnés en fonction de la taille de leur génome. Les métabolites sont ordonnés en fonction de la moyenne de leur degré entrant relatif sur l'ensemble des graphes de composés.

4.7 Discussion

Les comparaisons effectuées sur les réseaux métaboliques intracellulaires et extracellulaires nous ont permis de mettre en évidence le contraste entre la conservation d'un coeur métabolique chez les bactéries extracellulaires et l'absence de parties communes dans les réseaux métaboliques extracellulaires. Nous voyons ainsi que le concept de métabolisme minimal ne peut s'appliquer directement pour les endocytobiotes. Celui-ci doit être étudié selon les conditions environnementales de la bactérie comme l'indiquait Koonin (2000), ce qui signifie, dans le cas des endocytobiotes, prendre en compte le métabolisme de l'hôte et d'autres symbiotes potentiels et de considérer le réseau métabolique comme l'union des réseaux métaboliques des différents partenaires. C'est donc aussi le concept d'individu qu'il faut considérer d'une autre manière dans ce cas. Surtout pour ce qui concerne les symbiotes les plus intégrés, l'association avec l'hôte est devenue indissociable. Les fonctions métaboliques (ou autres) des uns s'ajoutent aux capacités métaboliques du second. L'intervention d'autres symbiotes peut encore ajouter à la complémentarité des différentes espèces en jeu où toutes ont un rôle dans la survie des autres. Ce système ne devrait pas être considéré comme un ensemble d'individus à part entière mais comme un "super-individu", que Nardon & Grenier (1993) nomment "symbiocosme".

Par ailleurs, notre étude a permis d'apporter quelques éléments nouveaux dans l'évolution du réseau métabolique des bactéries intracellulaires. Premièrement, nous avons constaté que la proportion de gènes métaboliques, relativement constante sur l'ensemble des bactéries, est presque deux fois plus élevée chez les bactéries mutualistes intracellulaires. Ceci peut s'expliquer à la fois par une perte plus importante de fonctions non métaboliques et une conservation des fonctions métaboliques dédiées aux échanges avec l'hôte chez les bactéries mutualistes.

Nous avons également souligné une proportion significativement plus élevée entre réactions et gènes métaboliques chez les bactéries intracellulaires, suggérant chez celles-ci une utilisation d'enzymes au spectre plus large que chez les bactéries extracellulaires. Ce résultat pourrait être l'indice d'une plus grande conservation des enzymes multifonctionnelles par rapport aux enzymes au spectre moins large chez les bactéries intracellulaires.

En outre, le rapport entre métabolites et réactions est plus élevé chez les bactéries intracellulaires, indiquant une disparition des voies métaboliques redondantes lors de la réduction du réseau métabolique.

Nous avons montré que celle-ci touche des parties différentes du réseau selon les bactéries considérées. La représentation des groupes de métabolites dans les différents réseaux nous révèle des profils très divers dont certains sont mêmes extrêmes, à l'exemple de la disparition totale de la biosynthèse des lipides dans les réseaux de *Buchnera aphidicola Cc* et *Sulcia muelleri*. Lorsque la déletion d'un ensemble de gènes métaboliques est fixée, elle induit à son tour une modification

des pressions de sélection sur d'autres régions du réseau. Lors de notre analyse, nous avons noté de nombreux cas d'une telle répercussion des pressions de sélection. Le maintien de la production de l'ATP dans un cas ou de son transport dans l'autre ont ainsi des conséquences sur l'évolution d'une partie des réseaux métaboliques respectifs de *Wolbachia pipientis wMel* et des Rickettsies, pourtant proches phylogénétiquement. C'est ce que nous avons observé également avec la dégradation de la voie des pentoses phosphates influant sur la conservation de la voie de l'histidine.

La réduction du métabolisme peut conduire à des portions de réseaux similaires même dans le cas d'espèces éloignées phylogénétiquement mais dont le rôle symbiotique est proche. C'est le cas par exemple de la synthèse de certains acides aminés ou cofacteurs chez les espèces où ces métabolites sont importants dans leur relation symbiotique.

A l'inverse, la comparaison des réseaux de *Sulcia muelleri* et de *Baumannia cicadellinicola* met en évidence la complémentarité des deux métabolismes associés dans un même hôte et montre ainsi l'importance que peut avoir la sélection sur les systèmes à plusieurs symbiotes, pouvant aller jusqu'à un partage "optimal" des tâches. Ce phénomène de co-adaptation de plusieurs symbiotes est encore mal connu, mais pourrait commencer par la présence d'un seul endocytobiote intégré dont une partie des fonctions symbiotiques pourrait disparaître et être prise en charge par un symbiote secondaire. C'est l'hypothèse émise pour expliquer la disparition de certaines voies de synthèse des acides aminés chez *Carsonella ruddii* et *Buchnera aphidicola Cc* (Moran *et al.*, 2003b; Pérez-Brocal *et al.*, 2006; Wu *et al.*, 2006b).

L'analyse des graphes des composés nous a montré une diversité importante de la topologie des réseaux métaboliques des bactéries intracellulaires. Cette modélisation nous a également permis d'identifier rapidement les métabolites les plus importants et ceux, qui, au contraire, ont perdu leur importance au cours de la réduction du métabolisme. Nous avons vu qu'en fonction des bactéries intracellulaires et de leur biologie, le réseau s'organise autour de métabolites et de voies différentes.

Ainsi, les éléments que nous apportons ici permettent de proposer trois évènements non exclusifs pouvant expliquer la réduction du métabolisme (Figure 4.38) :

1. les produits finaux des voies anaboliques sont puisés directement dans les ressources de l'hôte sans transformation préalable,
2. les voies anaboliques et cataboliques redondantes ont été éliminées, faisant disparaître du métabolome beaucoup de composés intermédiaires,
3. les voies anaboliques et cataboliques ont été considérablement raccourcies par l'apport d'un composé intermédiaire de la part de l'hôte ou d'un autre symbiote.

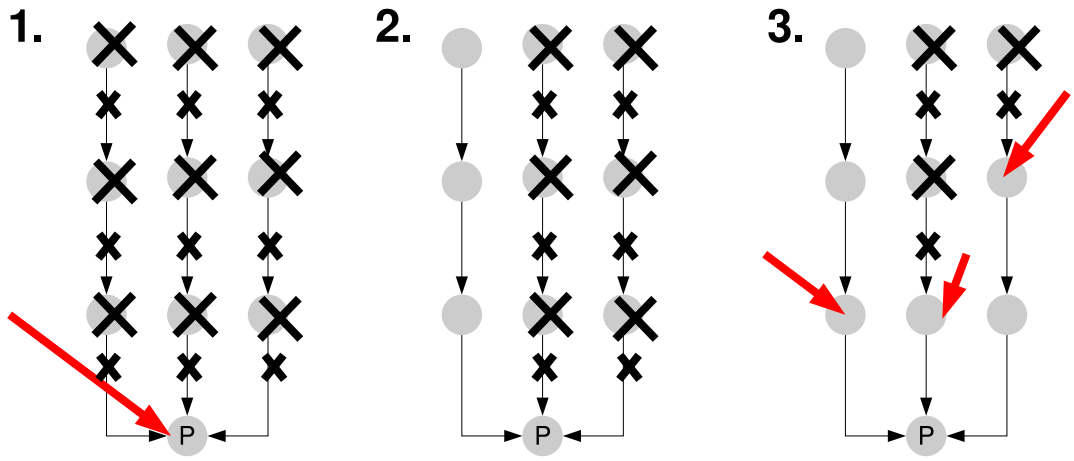

Figure 4.38. Trois hypothèses pour l'évolution des voies métaboliques chez les endocytobiotes. Les ronds symbolisent les métabolites et les flèches noires les réactions métaboliques dans le réseau d'un symbiote. Une croix noire symbolise la disparition d'une réaction ou d'un métabolite dans le réseau. Les flèches rouges indiquent un métabolite importé depuis l'hôte ou un autre symbiote du système.

Ces travaux peuvent être poursuivis de plusieurs manières. La première est d'étendre l'éventail des symbiotes étudiés, autant du point de vue phylogénétique que du point de vue de l'interaction qu'ils entretiennent avec leur hôte. La seconde est de continuer à améliorer encore la modélisation des réseaux métaboliques. La recherche de chemins communs, plus que de noeuds communs, et la mesure de leur centralité serait une des voies intéressantes à continuer. La recherche et l'inférence de motifs dans les réseaux de réactions est une des voies que nous voudrions également explorer (Lacroix *et al.*, 2006). L'alignement de réseaux métaboliques complets est aussi le fruit de réflexions au sein de l'équipe Baobab. De même, d'autres travaux auxquels nous participons permettent de décrire un graphe métabolique en termes de groupes de métabolites connectés de la même manière (Picard *et al.*, 2008). La modélisation sous forme d'hypergraphe ou de graphe bipartite est aussi une des voies pour améliorer la modélisation du réseau métabolique. Le développement d'algorithmes dédiés à des questions métaboliques précises est aussi nécessaire. C'est dans ces deux optiques que s'inscrit le développement de la méthode PITUFO que nous allons maintenant décrire.

Chapitre 5

Analyse fonctionnelle du réseau métabolique de bactéries endocytobiotes : la recherche de précurseurs avec PITUFO

Dans ce chapitre, nous allons présenter PITUFO, la première méthode exacte basée sur la topologie du réseau développée pour trouver tous les ensembles minimaux de métabolites, appelés précurseurs, suffisants pour produire un ensemble de métabolites cibles. Contrairement à d'autres méthodes similaires, notre modèle prend en compte les métabolites "auto-régénérés" impliqués dans des cycles et qui peuvent être utilisés pour générer les métabolites cibles à partir des précurseurs. La méthode a fait l'objet d'un article et d'une présentation acceptés à la conférence WABI 2008 (Cottret *et al.*, 2008).

Cette méthode peut avoir plusieurs applications, comme la vérification de reconstructions métaboliques, ou la définition d'un environnement métabolique minimum pour qu'un organisme puisse assurer certaines fonctions métaboliques. C'est ce dernier aspect qui nous a particulièrement intéressé ici. En effet, cette méthode peut être employée pour proposer des ensembles de précurseurs potentiels qu'un endocytobiote doit importer de son milieu (ici la cellule de l'hôte) pour assurer des fonctions métaboliques vitales ou symbiotiques. Nous avons appliqué et validé notre méthode sur le réseau métabolique de *Buchnera aphidicola APS* en nous intéressant plus particulièrement à la synthèse des acides aminés qui participe à la relation symbiotique entre la bactérie et son hôte, le puceron.

5.1 Contexte

Une fois le réseau métabolique d'un organisme reconstruit (voir Section 2.1), on peut se poser successivement les questions suivantes. Quelles sont les princi-

pales fonctions métaboliques de l'organisme ? Comment ces fonctions sont-elles assurées ?

La première question peut être abordée en inspectant la liste des réactions détectées comme présentes dans le réseau, pour les comparer à celles présentes dans les voies métaboliques connues, tel que nous l'avons fait dans la Section 4. Une fonction métabolique peut être vue comme la production d'un composé ou d'un ensemble de composés.

Il y a plusieurs manières d'aborder la seconde question, et une de celles-ci peut consister à se poser une nouvelle question : quels sont les métabolites que l'organisme doit obtenir de son environnement pour produire un ensemble de composés cibles ? Nous appelerons désormais de tels métabolites des *précurseurs*.

La seule inspection des voies métaboliques pour répondre à cette question pose assurément certains problèmes. Premièrement, les voies métaboliques peuvent connaître des variantes d'un organisme à l'autre. Deuxièmement, ces voies ne forment qu'une sous-partie du réseau et de nombreuses autres voies alternatives peuvent exister. Remonter les séries de réactions d'un produit final jusqu'aux substrats initiaux obtenus de l'environnement devient vite une tâche ardue même lorsqu'on étudie un réseau réduit.

Dans l'objectif de détecter des inconsistances dans la base de données EcoCyc (Keseler *et al.*, 2005), Romero et Karp ont utilisé une approche consistant à explorer le réseau entier pour trouver des ensembles de précurseurs (Romero & Karp, 2001).

Dans une première étape de propagation en avant ("forward propagation"), l'ensemble des métabolites qu'un ensemble de précurseurs peut produire à partir des réactions disponibles dans l'organisme est calculé. L'étape de propagation en avant est un processus itératif prenant en entrée, à chaque itération, l'ensemble des métabolites produits pendant l'itération précédente. Une réaction est "allumée" et peut être utilisée pour continuer la propagation si chacun de ses substrats a été produit par les étapes précédentes. Certains métabolites (appelés "bootstrap compounds"), considérés comme toujours présents dans la cellule, peuvent être utilisés pour allumer des réactions où ils interviennent si les autres substrats font tous partie des composés d'entrée. Ces métabolites "bootstrap" correspondent souvent aux cofacteurs utilisés dans les réactions enzymatiques telles que l'ATP ou la coenzyme A. Lorqu'une réaction est allumée, tous ses produits sont ajoutés à l'ensemble des métabolites d'entrée. Le processus s'arrête lorsque plus aucune réaction ne peut être allumée. Le sous-réseau formé par ce processus sera appelé plus tard "scope" par Handorf *et al.* (2005).

Dans la deuxième étape de la méthode, on vérifie la présence dans le sous-réseau formé par la première étape, de métabolites essentiels à la cellule, tels que les acides aminés ou les constituants de la paroi cellulaire. L'absence d'un de ces composés indique une inconsistance dans le réseau métabolique. Ainsi, pour chaque métabolite essentiel, les auteurs recherchent dans la partie du réseau qui n'a pas été produite par la première étape, les ensembles de précurseurs

qu'il serait nécessaire d'ajouter aux métabolites d'entrée pour produire les métabolites essentiels manquants. Cette étape se fait par un parcours à rebours des métabolites cibles jusqu'à parvenir à des impasses dans le réseau, c'est-à-dire à des métabolites produits par aucune réaction, que nous appelerons métabolites "sources" (E, F, G et H dans la Figure 5.1) et qui définissent ici les précurseurs potentiels. Plusieurs chemins peuvent exister entre les métabolites "sources" et les métabolites cibles, plusieurs ensembles alternatifs de précurseurs existent donc pour un ensemble de métabolites cibles donné.

Cette méthode a permis d'indiquer et de résoudre certaines données manquantes dans Ecocyc mais aussi de proposer des ensembles alternatifs de précurseurs pour quelques métabolites.

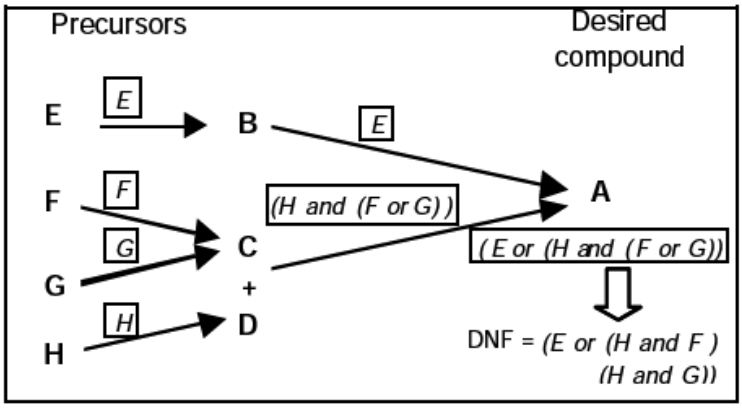

Figure 5.1. Figure tirée de l'article de Romero & Karp (2001). Parcours à rebours du métabolite cible A jusqu'aux précurseurs. L'ensemble des solutions est donné sous la forme d'une combinaison de "et" et de "ou".

Plus récemment, Handorf *et al.* (2007) ont proposé une méthode pour identifier des ensembles minimaux de métabolites requis par un organisme pour produire tous les métabolites d'un autre ensemble.

Ici tous les métabolites du réseau sont considérés comme précurseurs potentiels et placés dans une liste ordonnée. Naturellement, le scope de cette liste contient tous les métabolites cibles. La méthode utilise alors un algorithme gourmand : à partir du haut de la liste, chaque métabolite est successivement retiré et le scope recalculé : si celui-ci ne contient pas tous les métabolites cibles, le métabolite est remis dans la liste. Quand la liste est complètement traversée, un ensemble minimal de précurseurs est obtenu. Pour obtenir toutes les solutions, il serait nécessaire de tester tous les ordres possibles dans la liste, ce qui n'est pas envisageable du point de vue du temps de calcul. Ainsi, pour avoir une approximation de l'ensemble de solutions, plusieurs réarrangements aléatoires de la liste sont utilisés. Pour réduire l'espace des solutions, les auteurs utilisent des heuristiques basées sur des considérations biologiques : les métabolites avec les poids moléculaires les plus élevés sont placés préférentiellement au sommet de la

liste et ceux identifiés comme participant à une réaction de transport préférentiellement au bas de la liste. En utilisant cette méthode, Handorf *et al.* ont prédit des ensembles de nutriments nécessaires à 447 organismes.

Romero et Karp proposent une définition de précurseurs potentiels très restrictive alors que celle proposée par Handorf *et al.* est très large. La méthode que nous avons développée peut prendre en compte n'importe quelle définition de précurseurs potentiels. En effet, l'utilisateur peut définir lui-même ces ensembles. Notre méthode prend en compte également le fait que beaucoup de réactions dans les réseaux métaboliques sont définies comme réversibles à cause du manque d'information sur les propriétés enzymatiques et les concentrations des métabolites.

Dans leur article, Romero et Karp ne fournissent aucun détail sur comment sont traités les cycles, alors que ce point est crucial dans l'analyse des réseaux métaboliques. Dans la méthode décrite par Handorf *et al.*, le processus est itératif et une réaction ne peut être allumée que si tous les métabolites sont déjà produits par les étapes précédentes ou par des métabolites *bootstrap*. La méthode est ainsi incapable de prendre en compte les métabolites qui ne peuvent être atteints par ce processus, typiquement ceux qui participent à des cycles et qui nécessitent leur propre présence pour être produits. C'est le cas par exemple des composés B et D de la Figure 5.2. Nous suggérons dans notre méthode une manière de traiter efficacement ces métabolites dans la recherche des précurseurs. Ceci nous a amenés à introduire le concept, biologiquement fondé, de métabolites "auto-régénérés". Ce concept se rapproche partiellement de celui de métabolites *bootstrap*, indiqués par l'utilisateur, introduit de façon informelle par Romero et Karp. Notre méthode propose un moyen de les utiliser de façon automatique sans définir leur identité *a priori*. Très récemment, Kun *et al.* (2008) ont proposé une méthode pour identifier des métabolites qu'ils appellent réplicateurs autocatalytiques et qui répondent à une définition similaire.

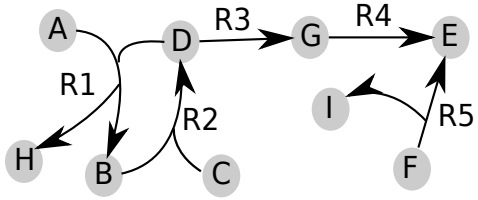

Figure 5.2. Un réseau métabolique G avec comme ensemble de métabolites $\mathcal{C} = \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}, \mathsf{E}, \mathsf{F}, \mathsf{G}, \mathsf{H}, \mathsf{I}\}$, et comme ensemble de réactions $\mathcal{R} = \{\mathsf{R1}, \mathsf{R2}, \mathsf{R3}, \mathsf{R4}, \mathsf{R5}\}$. Le scope de $\{A, C\}$, tel que calculé par Handorf *et al.*, ne contient pas E.

Une autre manière de rechercher les ensembles de précurseurs serait de décomposer le réseau en modes élémentaires (Section 2.2.2), qui représentent tous les plus petits sous-réseaux qui vérifient l'état d'équilibre (chaque quantité de métabolites produite est consommée), et de sélectionner ceux qui contiennent les métabolites cibles et des précurseurs potentiels, quelle que soit leur définition. Au

contraire des deux méthodes décrites précédemment, les métabolites qui nécessitent leur propre présence pour être produits sont pris en compte, à condition que l'état d'équilibre soit vérifié. Le calcul des modes élémentaires est basé sur la matrice stœchiométrique (qui contient tous les coefficients stœchiométriques de chaque composé dans chaque réaction). L'ensemble des modes élémentaires trouvé dans un réseau dépend donc considérablement de cette matrice. Cependant, à l'exception de quelques bases expertisées, de nombreux défauts d'annotation des coefficients stœchiométriques persistent. L'absence de certaines réactions, réellement présentes dans le réseau mais non assignées, peut aussi avoir une grande influence dans le calcul des modes élémentaires. Plus important, il est intéressant dans le cas d'études métaboliques de pouvoir intégrer dans le réseau certaines réactions générales (provenant d'expériences de traçage radioactif par exemple), pour lesquelles la stœchiométrie n'est pas forcément connue.

Dans ce qui va suivre, nous allons donc présenter la première méthode exacte pour détecter tous les ensembles minimaux de précurseurs. Nous proposons une manière efficace de traiter les composés qui nécessitent leur propre présence pour être produits, sans utiliser la matrice stœchiométrique mais seulement la topologie du réseau. Après avoir présenté la complexité du problème de trouver un ensemble minimal de précurseurs et celui de les trouver tous, nous présenterons l'algorithme lui-même et enfin les résultats de l'application de la méthode sur la recherche des précurseurs des acides aminés chez *Buchnera aphidicola APS*.

5.2 Définitions

Nous modélisons le réseau métabolique sous la forme d'un hypergraphe (voir Section 2.2) $G = (\mathcal{C}, \mathcal{R})$ avec $\mathcal{C}$ l'ensemble de noeuds correspondant aux métabolites et $\mathcal{R}$ l'ensemble des hyperarcs correspondant aux réactions. On ajoute un hyperarc r partant d'un ensemble de métabolites $C1$ vers un ensemble de métabolites $C2$ si $C1$ représente les substrats et $C2$ les produits de la réaction. Les réactions réversibles sont considérées comme deux réactions irréversibles.

Une solution du problème de recherche de précurseurs pour un ensemble de métabolites cibles est un ensemble de métabolites dont le stock est infini (par exemple, les métabolites provenant de l'environnement peuvent être, dans une certaine mesure, considérés comme tels) et dont le scope contient tous les métabolites cibles. Cependant, cette façon de considérer la dynamique du réseau n'est pas suffisante. En effet, le réseau peut contenir des cycles où certains métabolites non disponibles initialement pourraient être utilisés comme substrats de réactions parce qu'ils sont capables de se régénérer eux-mêmes une fois que le processus a produit l'un d'entre eux. Nous les appelons métabolites *auto-régénérés*. Ces métabolites ne sont pas considérés comme des précurseurs potentiels, censés provenir de l'environnement, mais ont besoin d'être continuellement régénérés. Cependant,

ils participent à leur propre régénération et à la génération d'autres métabolites. Les métabolites auto-régénérés et ceux dont ils rendent possible la synthèse sont appelés les métabolites *continuellement disponibles*.

Examinons le petit réseau donné en exemple dans la Figure 5.2. Soient A et C les précurseurs potentiels et E le métabolite cible. Les précurseurs potentiels étant considérés comme présents en quantité infinie, les métabolites B et D peuvent s'auto-régénérer. Ainsi, B et D et par extension G sont marqués "continuellement disponibles". Par conséquent, une solution du problème est bien l'ensemble de précurseurs $\{A, C\}$.

Pour chaque réaction r, nous appelons $sub(r)$ l'ensemble des substrats de r et $prod(r)$ l'ensemble des produits de r. Dans la suite, nous notons $\mathcal{P}(\mathcal{S})$ l'ensemble des sous-ensembles d'un ensemble S.

Définition 5.2.1. (Accessibilité) *Etant donné $X \in \mathcal{P}(\mathcal{C})$, un ensemble de métabolites, $Access(X) \in \mathcal{P}(\mathcal{C})$ est l'ensemble de métabolites y pour lequel il existe $r \in R$ avec $sub(r) \subseteq X$ et $prod(r) \ni y$.*

En d'autres termes, $y \in Access(X)$ s'il existe une réaction dans R qui produit y et dont les substrats sont dans X. Notons que X n'est pas nécessairement contenu dans $Access(X)$.

A partir d'un ensemble X de métabolites en quantité infinie, et d'un ensemble Z de métabolites continuellement disponibles, nous souhaitons calculer l'ensemble de métabolites qui peuvent être produits dans le réseau.

Définition 5.2.2. (Fonction d'accessibilité) *Soit $X \in \mathcal{P}(\mathcal{C})$ et $Z \in \mathcal{P}(\mathcal{C})$ deux sous-ensembles de métabolites, la fonction d'accessibilité $f_Z : \mathcal{P}(\mathcal{C}) \longrightarrow \mathcal{P}(\mathcal{C})$ est*

$$f_Z(X) = X \cup Reach(X \cup Z).$$

Par définition, f_Z est monotone, à la fois dans X et dans Z. On définit la fonction $f_Z^k(X) = f_Z(f_Z^{k-1}(X))$, avec $f_Z^1(X) = f_Z(X)$, comme la fonction obtenue en itérant k fois la fonction f_Z.

Définition 5.2.3. (Fonction de scope) *Soit $Z \in \mathcal{P}(\mathcal{C})$ et $X \in \mathcal{P}(\mathcal{C})$ deux ensembles de métabolites. La fonction de scope $f_Z^* : \mathcal{P}(\mathcal{C}) \longrightarrow \mathcal{P}(\mathcal{C})$ est ${f_Z}^*(X) = {f_Z}^k(X)$ pour tout k tel que ${f_Z}^k(X) = {f_Z}^{k+1}(X)$.*

Ainsi, f_Z^* représente ce qui peut être produit à partir de X avec l'aide de Z en utilisant des réactions dans R. Pour définir quand un ensemble de composés X est un ensemble de précurseurs d'une cible T, nous avons donc besoin d'imposer que f_Z^* contient T, et que f_Z^* peut régénérer Z.

Définition 5.2.4. (Ensemble de précurseurs) *Un ensemble de métabolites $X \subseteq \mathcal{P}(\mathcal{C})$ est un* ensemble de précurseurs de $T \subseteq \mathcal{P}(\mathcal{C})$ *s'il existe un ensemble $Z \subseteq \mathcal{P}(\mathcal{C})$ tel que*

$${f_Z}^*(X) \supseteq T \cup Z.$$

On note ici que nous sommes seulement intéressés à savoir si Z existe et non à savoir exactement quels composés il contient. On note évidemment aussi que Z peut être vide.

Un ensemble de précurseurs t est défini en utilisant la modélisation sous forme d'hypergraphe de la façon qui suit.

Définition 5.2.5. Hyperchemin avec un ensemble de métabolites *continuellement disponibles* *Un ensemble $H(X, Z, t) \in \mathcal{P}(\mathcal{R})$ de réactions est un hyperchemin d'un ensemble de métabolites X vers t en utilisant un autre ensemble de métabolites Z s'il satisfait :*

1. *Les réactions dans $H(X, Z, t)$ peuvent être ordonnées $< r_1, r_2, \ldots, r_k >$ de telle façon que :*
   - *i) Pour tout r_i, $Inp(r_i) \subset X \cup Z \cup Out(r_1) \cup \ldots \cup Out(r_{i-1})$;*
   - *ii) $t \in Out(r_k)$;*
   - *iii) Pour tout $s \in Z$, il existe $j(s)$ tel que $s \in Out(r_{j(s)})$,*
2. *Aucun sous-ensemble de $H(X, Z, t)$ ne vérifie ce qui précède.*

Ainsi, s'il y a un hyperchemin $H(X, Z, t)$, alors X est un ensemble de précurseurs de t. Par exemple, dans la Figure 5.2, si E est la cible, et A et C sont des précurseurs potentiels, il y a un hyperchemin H($\{A, C\}$, $\{B\}$, E) = $\{r1, r2, r3, r4\}$ tel que l'ensemble {A,C} est un ensemble de précurseurs de E.

L'inverse est montré dans ce qui suit :

Lemme 5.2.1. *Si X est un ensemble de précurseurs de t, alors il existe un hyperchemin $H(X, Z, t)$ avec $Z \in \mathcal{P}(\mathcal{C})$.*

Preuve. Notons qu'il est suffisant de montrer qu'il existe une liste ordonnée de réactions H qui vérifie la condition 1 de la Définition 5.2.5. Premièrement, nous définissons récursivement une séquence de réactions : en commençant de la cible t, à chaque étape i nous choisissons une réaction r_i qui produit un métabolite qui ne soit pas un précurseur et/ou un substrat qui ne soit pas encore produit par une des réactions précédentes $r_1, \ldots, r_{i-1}$. La séquence obtenue peut contenir des répétitions. En éliminant les répétitions et en inversant la liste obtenue, nous obtenons un ensemble ordonné H qui remplit la condition i.

Dans le but de trouver des réactions qui peuvent être atteintes à partir de X, nous considérons seulement les réactions dans l'ensemble $W = \{r \in \mathcal{R} \mid Inp(r) \subseteq {f_Z}^*(X)\}$, c'est-à-dire les réactions qui prennent comme substrats les composés contenus dans le scope de X.

Soit $N_0 = \{t\}$ et $A_0 = \emptyset$. A l'itération i, $i \geq 1$, nous définissons les ensembles $A_i = \cup_{j=1}^{i} Out(r_j) \setminus X$ et $N_i = \cup_{j=1}^{i} Inp(r_j) \setminus (A_i \cup X)$, c'est-à-dire, A_i est l'ensemble de métabolites produits dans les premières i réactions et N_i l'ensemble de métabolites consommés mais pas encore rendus disponibles dans les premières i réactions. A l'itération i, sélectionnons $c_i \in N_{i-1}$. Puisque $c_i \notin X$, nous savons,

par la définition de W, qu'il existe une réaction $r_i \in W$ avec $c_i \in Out(r_i)$. Nous mettons à jour l'ensemble A_i avec $A_{i-1} \cup (Out(r_i) \setminus X)$ et définissons l'ensemble $S_i = Inp(r_i) \cap A_i$, substrats de r_i qui ont déjà été produits. Nous mettons à jour $N_i = (N_{i-1} \cup Inp(r_i)) \setminus (A_i \cup X)$.

Ce processus est réitéré jusqu'à ce que N_i soit vide, ce qui doit arriver puisque la séquence de A_i est monotone. Soit k la première (et dernière) itération telle que $N_k = \emptyset$ et soit $Z = \cup_{i=1}^{k} S_i$. La séquence $\omega = r_1, \ldots, r_k$ peut avoir des répétitions. Nous définissons $\bar{r}_1, \ldots, \bar{r}_\ell$ comme la sous-séquence qui contient seulement la première occurrence de chaque réaction et définissons $H = \{\bar{r}_1, \ldots, \bar{r}_\ell\}$ comme l'ensemble incluant toutes ces réactions.

Pour conclure la preuve, nous montrons que la séquence inversée $\bar{\omega} =< \bar{r}_\ell, \ldots, \bar{r}_1 >$, satisfait la condition 1 de la Définition 5.2.5. Par construction, $\bar{r_1} = r_1$ produit le métabolite $c_1 = t$, donc $1(ii)$ est satisfaite. De plus, par définition, $Z = \cup_{i=1}^{k} S_i$ et donc $Z \subseteq \cup_{i=1}^{k} A_i$. Ainsi, tout composé $s \in Z$ est le produit d'une réaction $\{\bar{r}_1, \ldots, \bar{r}_\ell\}$ et $1(iii)$ est prouvée.

La preuve de $1(i)$ est plus technique. Pour $\bar{r}_i \in H$, $i \in \{1, \ldots, \ell\}$, nous montrons que $Inp(\bar{r}_i) \setminus (X \cup Z)$ est un sous-ensemble de $Out(\bar{r}_{i+1}) \cup \cdots \cup Out(\bar{r}_\ell)$. Soit y dans $Inp(\bar{r}_i) \setminus (X \cup Z)$ et soit j l'indice de la dernière apparition de $\bar{r}_i$ dans ω. Alors, $y \in Inp(r_j)$. Puisque $y \notin Z$, en particulier $y \notin S_j$, et par définition de S_j, nous avons : $y \notin A_j$. Par ailleurs, nous avons aussi : $y \notin X$ et donc $y \in N_j$. Comme à la fin N_k est vide, alors il doit exister $p \in \{j+1, \ldots, k\}$ tel que $y \in A_p$ mais $y \notin A_{p-1}$. Alors, $y \in Out(r_p)$. Soit $\bar{r}_q = r_p$. Nous avons : $q > i$. En effet, si h est l'indice de la dernière apparition de $\bar{r}_q$, alors $h \geq p > j$. Donc la dernière apparition de $\bar{r}_q$ est après la dernière apparition de $\bar{r}_i$ et donc $q > i$. □

5.3 Algorithme pour énumérer tous les ensembles minimaux de précurseurs

Pour plus de clarté, nous considérerons le cas d'un métabolite cible. La solution pour plusieurs métabolites cibles est calculée simplement en ajoutant une réaction irréversible qui a comme substrats les composés cibles et comme produit un composé fictif, qui sera alors considéré comme l'unique cible.

L'algorithme est composé de deux étapes : la première définit une structure spéciale, appelée un "arbre de remplacement", qui contient une représentation d'au moins un hyperchemin (voir Section précédente) pour chaque ensemble de précurseurs de t. Pour le construire, nous procédons d'une manière analogue à celle que nous avons adoptée dans la preuve du Lemme 5.2.1 ; la principale différence est que dans ce cas-ci, X est inconnu (en réalité l'algorithme cherche tous les X qui sont des ensembles de précurseurs de t). Ainsi, quand l'algorithme parcourt le réseau à rebours en commençant par t, il doit considérer toutes les réactions et non seulement celles dans W. A la fin de la première étape, l'arbre

de remplacement contient une représentation d'au moins un hyperchemin pour chaque ensemble minimal de précurseurs de t mais aussi la représentation d'autres hyperchemins qui ne représentent pas des ensembles minimaux de précurseurs. Dans la seconde étape, l'arbre de remplacement est utilisé pour énumérer tous les ensembles de précurseurs pour le métabolite cible, et éliminer tous les ensembles de métabolites qui ne sont pas des précurseurs. La complexité dans l'espace et le temps est linéaire avec la taille de l'arbre de remplacement qui peut être exponentielle (notons aussi que le nombre de solutions peut être exponentiel avec la taille de P). Pour plus de clarté, les deux étapes sont décrites séparément. Dans l'implémentation, elles sont en réalité exécutées simultanément, améliorant ainsi l'espace et le temps nécessaires.

5.3.1 Construction de l'arbre de remplacement

L'arbre est enraciné et dirigé de la racine vers les feuilles. Les noeuds de l'arbre sont étiquetés, soit par un métabolite, soit par une réaction. Dans le premier cas, nous avons un *noeud métabolite* tandis que dans le second cas, nous avons un *noeud réaction*. Les enfants d'un noeud métabolite sont des noeuds réactions étiquetés par les réactions qui produisent le métabolite en question alors que les enfants d'un noeud réaction sont étiquetés par ses substrats. Dans l'arbre, les noeuds, qu'ils soient métabolite ou réaction, ont seulement un parent alors qu'ils peuvent avoir plusieurs enfants. Par la suite, nous utiliserons les termes *noeud produit* pour le parent d'un noeud réaction et *noeud substrat* pour les enfants d'un noeud réaction.

La construction de l'arbre commence à la racine t. Pour chaque réaction produisant ce métabolite, nous créons un noeud réaction qui a comme parent la racine et comme enfants de nouveaux noeuds métabolites correspondant aux substrats de la réaction considérée. De cette façon, nous obtenons un arbre de profondeur 3 dont les feuilles sont des noeuds métabolites. Ce processus est réitéré pour chaque nouveau noeud métabolite.

Appelons u un noeud métabolite nouvellement créé et c le métabolite correspondant. Le long de n'importe quelle branche de l'arbre, le processus s'arrête quand une de ces 3 conditions est vérifiée :

1. c correspond aussi à un ancêtre de u,
2. c correspond à un enfant d'un noeud réaction ancêtre de u, ou
3. c n'est produit par aucune réaction (nous ne pouvons pas aller plus loin dans le réseau).

Dans le premier cas, nous marquons u comme métabolite continuellement disponible. Un exemple est le noeud métabolite $u1 = H$ qui est un des enfants de $N12$ dans l'arbre de la Figure 5.3. Dans ce cas, ce métabolite noeud est marqué comme continuellement disponible (repéré par une étoile à côté du noeud). En effet, ce métabolite est régénéré par le réseau.

Dans le second cas, c est considéré comme un enfant d'un noeud réaction de u ; donc il n'est pas nécessaire de dupliquer la recherche de précurseurs en continuant après ce métabolite. Un exemple est le noeud métabolite $u2 = I$ dans l'arbre. Puisque I est l'enfant du noeud réaction $N13$ ancêtre de $u2 = I : I$ a déjà été analysé dans la même branche.

Dans le troisième cas, c ne peut être produit par aucun autre noeud réaction. Ainsi, quand le processus s'arrête, toutes les étiquettes des feuilles de l'arbre correspondent à des métabolites produits par aucune réaction ou des métabolites déjà visités dans la même branche.

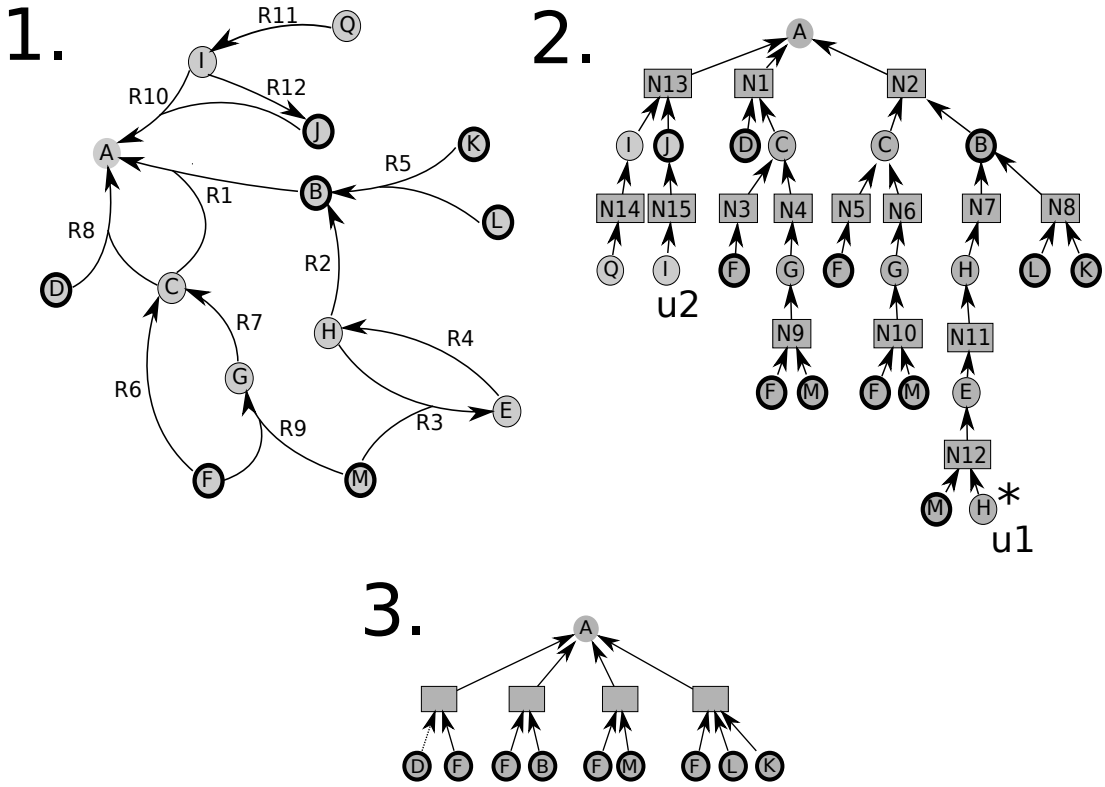

Figure 5.3. Un exemple de réseau métabolique et de l'arbre de remplacement pour le métabolite cible A. **1.** L'hypergraphe métabolique. Les noeuds métabolites entourés en noirs, B, D, F, L, J, K et M sont des précurseurs potentiels. **2.** L'arbre de remplacement avant compression. **3.** L'arbre de remplacement après compression. Chaque ensemble de noeuds substrats d'une réaction correspond à un ensemble minimal de précurseurs.

Cette façon de procéder implique que, pour toute solution X pour le métabolite cible t, il existe un sous-arbre dont l'ensemble de réactions représente un hyperchemin de X jusqu'à t en utilisant des métabolites auto-régénérés (l'ensemble Z du Lemme 5.2.1).

Lemme 5.3.1. *Si X est un ensemble de précurseurs de t, il existe un sous-arbre de l'arbre de remplacement contenant un hyperchemin $H(X, Z, t)$ avec $Z \in \mathcal{P}(\mathcal{C})$.*

5.3.2 Enumération des solutions

Nous utilisons maintenant l'arbre de remplacement pour énumérer tous les ensembles minimaux de précurseurs de t. Ceci est fait en traitant successivement les sous-arbres qui ont un seul noeud réaction r comme racine et dont les noeuds substrats sont tous des feuilles de l'arbre. En fonction de la nature des métabolites (précurseurs potentiels, métabolites marqués ou non), le sous-arbre sera, soit éliminé, soit utilisé pour créer un nouveau sous-arbre qui le remplacera. Ceci aura pour effet une compression progressive de l'arbre d'origine jusqu'à ce qu'il soit composé de la racine (niveau 1), des noeuds réactions produisant la racine (niveau 2) et des ensembles minimaux de précurseurs (niveau 3). L'arbre final préserve les mêmes propriétés que l'arbre initial en ce qui concerne les ensembles minimaux de précurseurs qui produisent la cible t.

L'algorithme de compression commence en considérant un noeud réaction dont les noeuds substrats sont tous des feuilles. Soit r l'étiquette d'un tel noeud, p le noeud produit de r, S l'ensemble des étiquettes des noeuds substrats de r (S est donc un ensemble de métabolites) et r_2 le noeud réaction parent de p. Si r est tel que *1)* au moins un de ses noeuds substrats n'est ni un précurseur potentiel ni marqué comme continuellement disponible, ou *2)* les précurseurs potentiels dans S forment un super ensemble des précurseurs potentiels qui sont les noeuds substrats d'un autre noeud réaction ayant aussi p comme parent, alors le sous-arbre dont r est la racine est simplement éliminé de l'arbre. Si r n'est pas éliminé, le sous-arbre dont r_2 est la racine est dupliqué ; tous les noeuds réactions et métabolites sont dupliqués en conservant la même étiquette. Soit r'_2 la racine de ce nouveau sous-arbre, p' l'enfant de r'_2 qui correspond à p, et r' l'enfant de p' qui correspond à r. Le parent de r_2 devient aussi le parent de r'_2. Nous modifions les sous-arbres dont la racine est r_2 de la façon suivante :

- les noeuds substrats de r sont déconnectés de r et r lui-même est éliminé ;
- nous remplaçons le noeud p' avec l'ensemble des enfants de r'. Les noeuds p' et r' sont éliminés.

Illustrons ceci par un exemple en utilisant l'arbre de remplacement de la Figure 5.3. Faisons l'hypothèse que, à une itération quelconque de l'algorithme de compression, le sous-arbre dont la racine est le noeud réaction $N9$ est considéré. Tous les noeuds substrats de $N9$ sont en effet des feuilles de l'arbre et également des précurseurs potentiels. Le parent de $N9$ n'ayant pas d'autre enfant, $N9$ ne peut pas être éliminé. Puisque le noeud réaction qui est son ancêtre immédiat est $N4$, le sous-arbre dont la racine est $N4$ est dupliqué. Soit $N4'$ l'étiquette de ce sous-arbre dupliqué. $N4'$ devient un enfant du parent de $N4$, étiqueté C. Les noeuds métabolites F et M, substrats de $N9$ deviennent des enfants de $N4'$ et remplacent le noeud substrat de $N4'$, G. $N9$ est retiré de l'arbre. Le parent de $N4$ a maintenant 3 enfants : $N3$, $N4$ et $N4'$. Si $N4$ est le nouveau sous-arbre considéré à la prochaine itération de l'algorithme, l'algorithme l'éliminerait puisque son seul noeud substrat ne correspond pas à un précurseur potentiel et n'a pas

été marqué comme continuellement disponible. Considérons maintenant le sous-arbre dont $N4'$ est la racine. Ses deux noeuds substrats, étiquetés F et M, sont des précurseurs potentiels. Cependant, puisque $\{F, M\}$ est un super ensemble des noeuds substrats de $N3$, l'arbre dont la racine est $N4'$ peut être éliminé. Le processus que nous venons de décrire continue jusqu'à ce que l'arbre compressé final ne contienne que 3 niveaux : la racine étiquetée par le métabolite cible, les noeuds réactions produits par la compression et les noeuds substrats de ces noeuds réactions. La propriété cruciale est que chaque étape de la compression n'élimine aucun ensemble minimal de précurseurs de t ; ceci signifie que les étiquettes des noeuds substrats du niveau 2 correspondent directement à un ensemble minimal de précurseurs de t, comme indiqué dans le lemme suivant.

Lemme 5.3.2. *Si X est ensemble minimal de précurseurs de t, alors il existe un enfant x de la racine de l'arbre compressé final tel que les étiquettes des enfants de x coïncident avec X.*

5.4 Recherche des précurseurs des acides aminés dans le réseau métabolique de *Buchnera aphidicola APS*

5.4.1 Motivation

Des études trophiques ont montré que *Buchnera* approvisionne son hôte, le puceron, notamment en acides aminés essentiels (Douglas, 1998). Ces derniers ne peuvent être synthétisés par le puceron et il ne les trouve pas en quantité suffisante dans la sève dont il se nourrit. Ainsi, la plupart des *Buchnera* ont conservé les dernières étapes des voies de synthèse des acides aminés essentiels. Par contre, les voies de synthèse des acides aminés non essentiels ont été fortement dégradées. *Buchnera* est donc dépendante de l'importation des nutriments provenant de la cellule hôte afin de compléter ces voies.

Les acides aminés sont indispensables à la survie de toute cellule puisque ce sont les briques élémentaires des protéines qui vont constituer les éléments de la cellule et la faire fonctionner. Vingt acides aminés sont communément trouvés dans les protéines. Ce sont tous des acides aminés α : ils sont constitués d'un groupement carboxyl et d'un groupement amine, reliés au même atome de carbone (Figure 5.4). Ils diffèrent les uns des autres par leur chaîne latérale, ou chaîne R, qui varie en composition, taille et charge électrique et qui influence la solubilité des acides aminés dans l'eau. Un acide aminé est dit essentiel pour un organisme quand il ne peut pas être synthétisé par celui-ci. Chez le puceron, il y a 10 acides aminés essentiels (Table 5.1).

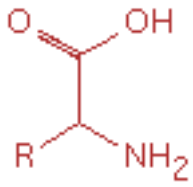

Figure 5.4. Structure générale d'un acide aminé α, mis à part la proline qui contient un cycle entre la chaîne latérale et le groupement amine. Seule la chaîne latérale R diffère entre les acides aminés.

Buchnera, par son petit génome et sa proximité phylogénétique avec la bactérie modèle *Escherichia coli*, bénéficie d'une bonne qualité d'annotation génomique (seuls deux gènes de *Buchnera* n'ont pas d'orthologues identifiés chez *Escherichia coli K12*), qui, ajoutée à son intérêt agronomique, en fait certainement la bactérie endocytobiote la plus étudiée. Par conséquent, les voies de synthèse des acides aminés ont été particulièrement analysées, autant par des études trophiques que génomiques.

Certains résultats bien connus vont ainsi nous permettre de valider notre méthode. Nous proposerons également de nouvelles pistes intéressantes pour la compréhension de la synthèse des acides aminés chez *Buchnera aphidicola APS*.

5.4.2 Acquisition des données et modélisation

Nous avons assigné les réactions présentes dans *Buchnera aphidicola APS* en utilisant les annotations génomiques de la bactérie et en les comparant à celles d'EcoCyc (version 11.5). Le génome de *Buchnera* étant un sous-ensemble de celui d'*E. coli*, on peut raisonnablement penser la même chose à propos de leur liste de réactions. EcoCyc étant intensivement expertisée, nous avons préféré prendre cette base comme support de l'assignation des réactions chez *Buchnera aphidicola APS* plutôt que l'ensemble de la base MetaCyc.

En revanche, nous avons utilisé MetaCyc (version11.5) pour décider du sens préférentiel des réactions. Ainsi, une réaction a été assignée comme irréversible si elle apparaît toujours dans le même sens dans MetaCyc, quelle que soit la voie métabolique considérée.

Nous n'avons considéré que le métabolisme des petites molécules, ce qui signifie que nous ne prenons pas en compte les réactions ne faisant intervenir que des grosses molécules comme les protéines ou les molécules d'ARN par exemple.

Afin de diminuer la taille de l'arbre de remplacement et de rendre plus clairs les résultats, nous appliquons le filtre supprimant tous les couples de cofacteurs connus dans les réactions où ils interviennent (voir Section 3.4.1). De plus, nous retirons l'eau de la liste des composés.

Un certain nombre de corrections manuelles ont été introduites dans le réseau grâce notamment au travail réalisé par Prickett *et al.* (2006) dans la base

BuchneraBase [1]. La réaction 2.6.1.42 permettant de produire la leucine, l'isoleucine ou la valine a été ajoutée. Le gène correspondant est absent du génome de *Buchnera aphidicola APS* mais la réaction pourrait être prise en charge par une autre aminotransférase (Shigenobu *et al.*, 2000). Une autre hypothèse serait que ces réactions se déroulent au sein du puceron. Nous avons ajouté également les réactions de synthèse de la phénylalanine et de la tyrosine (EC : 2.6.1.57). En effet, Shigenobu *et al.* indiquent que la dernière étape de la synthèse de la phénylalanine est catalysée par TyrB chez *E. coli* et que HisC pourrait se substituer à TyrB chez *Buchnera aphidicola APS*.

La voie de synthèse des nucléotides apparaît comme absente chez *Buchnera aphidicola APS*. Pourtant, Zientz *et al.* (2004) indiquent que *Buchnera aphidicola APS* semble pouvoir synthétiser l'ensemble des nucléotides. En regardant de plus près les voies métaboliques inférées par les pathway-tools, nous nous sommes rendus compte qu'une réaction manque particulièrement dans la synthèse des nucléotides : la réaction 2.7.4.6 qui transfère le phosphate d'une molécule d'ATP dans un dinucléoside pour former un trinucléoside. Elle intervient 8 fois dans la synthèse des nucléotides sur divers substrats (CDP, GDP, UDP, dADP, dCDP, dGDP, dUDP, dTDP).

Mushegian & Koonin (1996) indiquent cette réaction comme faisant partie des réactions essentielles à la vie cellulaire. Chez *Escherichia coli*, elle est codée par le gène *ndk*, absent du génome de *Buchnera aphidicola APS*. Pour combler son absence dans le génome de *Buchnera aphidicola*, Gil *et al.* (2004b) proposent le gène *pykA* qui code normalement pour une pyruvate kinase dont le produit pourrait avoir aussi la même activité catalytique que ndk. Nous avons donc assigné à ce gène les réactions correspondantes au numéro EC 2.7.4.6.

Par ailleurs, nous considérons le glucose comme précurseur implicite, il n'apparaît donc pas dans nos résultats. Afin de réduire le temps de calcul, la recherche des précurseurs est effectuée sur la partie du réseau qui ne peut pas être produite par l'injection seule du glucose.

Enfin, sont considérés comme précurseurs potentiels les métabolites sources du réseau, c'est-à-dire les métabolites produits par aucune réaction ou par une seule réaction mais qui est réversible.

Le réseau sur lequel nous appliquons notre méthode comporte ainsi 246 réactions, dont 37 sont restées réversibles, et 365 composés, dont 140 sont définis en tant que précurseurs potentiels.

5.4.3 Résultats

Dans ce qui suit, nous présentons les résultats successivement pour les acides aminés non essentiels, puis pour les acides aminés essentiels. Pour plus de clarté, nous classons aussi les acides aminés en six familles selon le métabolite qui marque

[1] http ://www.york.ac.uk/res/thomas/Buchnerabase/home.cfm

le début des voies chez *Escherichia coli*. Ce groupement en familles, inspiré de celui donné par Nelson & Cox (2004) et exposé dans le Tableau 5.1, nous permet non seulement de rendre les résultats plus clairs, mais aussi d'indiquer les différences entre les voies de synthèse des acides aminés chez *Buchnera aphidicola APS* avec celles chez *Escherichia coli*.

Métabolite origine	**Acides aminés**
α-kétoglutarate	Glutamate, Glutamine†, Proline†, Arginine*
3-Phosphoglycérate	Sérine†, Glycine†, Cystéine†
Oxaloacétate	Aspartate†, Asparagine†, Méthionine*†, Thréonine*†, Lysine*†
Pyruvate	Alanine†, Valine*, Leucine*, Isoleucine*
Phosphoénolpyruvate et érythrose 4-phosphate	Tryptophane*, Phénylalanine*, Tyrosine†
Ribose 5-Phosphate	Histidine*

Tableau 5.1. Groupement des acides aminés par le métabolite à l'origine de leurs voies de synthèse chez *Escherichia coli* et la plupart des organismes libres. $*$ indique que l'acide aminé est essentiel pour le puceron. † indique que le métabolite source n'est plus le même chez *Buchnera aphidicola APS*.

a. Les acides aminés essentiels

- *La méthionine, la thréonine et la lysine*

Chez *Escherichia coli*, les voies de synthèse de la méthionine, de la thréonine et de la lysine commencent par l'oxaloacétate et passent par l'aspartate (Figure 5.5). Chez *Buchnera aphidicola APS*, l'aspartate n'est produit par aucune réaction (voir plus loin), il est donc considéré ici comme un précurseur potentiel.

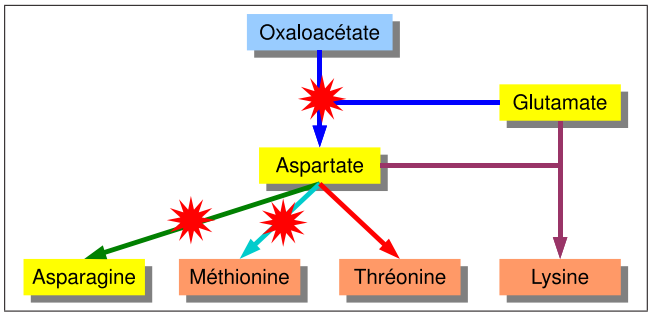

Figure 5.5. Vision simplifiée des voies de synthèse de l'asparagine, de la méthionine, de la lysine et de la thréonine à partir de l'oxaloacétate chez *Escherichia coli*. En jaune, apparaissent les acides aminés non essentiels et en orange, les acides aminés essentiels pour le puceron. Les étoiles rouges indiquent les voies qui n'existent plus chez *Buchnera aphidicola APS*.

Chez *Escherichia coli*, la voie de synthèse de référence de la méthionine à partir de l'aspartate comporte sept étapes. Les trois premières correspondent à

la synthèse de l'homosérine et les trois dernières à la synthèse de la méthionine elle-même. La reconstruction métabolique nous indique que la synthèse de l'homosérine est possible chez *Buchnera aphidicola APS*, mais qu'il semble manquer les trois enzymes capables de catalyser les premières étapes de la synthèse de la méthionine à partir de l'homosérine. Seule subsiste la dernière réaction catalysant la transformation de l'homocystéine en méthionine.

Sept ensembles minimaux de précurseurs ont été trouvés pour la méthionine (Figure 5.6). Il est intéressant de voir que sur ces sept solutions, six proposent la méthionine elle-même en tant que métabolite auto-régénéré. Nous verrons que ces six solutions, plutôt que de correspondre à des cycles régénérant réellement la méthionine, seraient plutôt reliées à deux alternatives différentes, l'une produisant la méthionine, l'autre utilisant cette dernière.

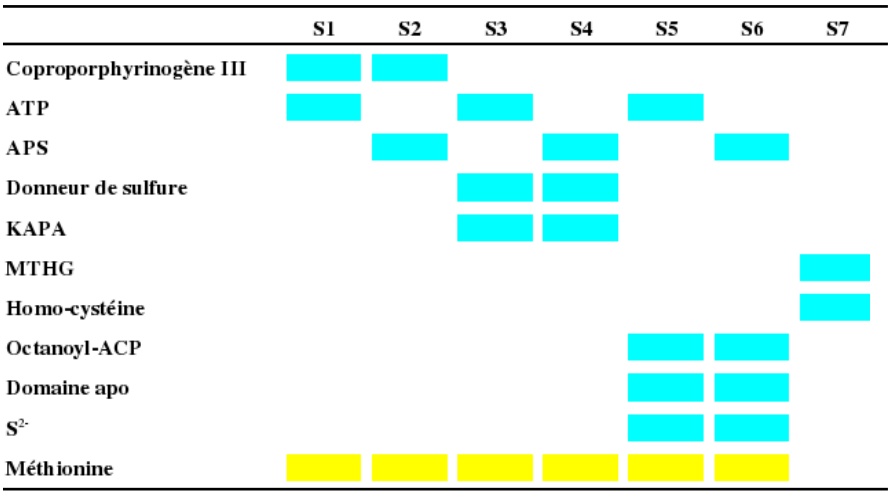

Figure 5.6. Ensembles de précurseurs de la méthionine trouvés par PITUFO. Chaque colonne correspond à une solution. Un carré bleu indique qu'une solution utilise le métabolite correspondant en tant que précurseur. Un carré jaune indique qu'une solution utilise le composé correspondant en tant que métabolite auto-régénéré. **APS** : adénosine 5'-phosphosulfate ; **KAPA** : 7-kéto-8-aminopélargonate ; **MTHG** : 5-méthyltétrahydroptéroyltri-L-glutamate

La solution S7 est la seule n'utilisant pas la méthionine elle-même en tant que métabolite auto-régénéré. Elle correspond à la réaction 2.1.1.14 produisant la méthionine à partir de l'homocystéine et du 5-méthyltétrahydroptéroyltri-L-glutamate (MTHG). Cette réaction correspond à la septième étape de la voie de référence.

L'homocystéine ne semble pas posséder de précurseurs chez *Buchnera aphidicola APS*. Moran *et al.* (2005) ont montré que l'ajout d'homocystéine dans les plantes hôtes du puceron augmentait considérablement la concentration de méthionine dans ce dernier. Les auteurs indiquent que l'homocystéine a pu être transformée en méthionine par la plante elle-même mais aussi par *Buchnera aphidicola APS*. Aucun indice d'un possible transport de l'homocystéine dans la cellule

de *Buchnera* n'a été trouvé dans la littérature, mais il semble probable.

L'autre substrat de la réaction, MTHG, appartient à la classe des tétrahydrofolates polyglutamates. Dans BioCyc, cette classe intervient en tant que produit de réactions de polymérisation du tétrahydrofolate (THF) et du glutamate. Dans l'équation des réactions données par BioCyc, c'est la classe elle-même qui est indiquée. Dans notre modélisation, celle-ci est donc considérée comme un métabolite à part entière alors qu'elle représente en vérité un ensemble de métabolites, présents eux aussi en tant que noeuds ailleurs dans le graphe. Ces classes de métabolites forment ainsi des impasses artificielles dans le réseau. Ici, par exemple, le MTHG ne serait pas un précurseur. Appartenant à la classe des tétrahydrofolates polyglutamates, sa synthèse dépend vraisemblablement de celle du THF. La voie de synthèse de référence du THF semble incomplète chez *Buchnera aphidicola APS* alors que celui-ci intervient en tant que cofacteur dans de nombreuses réactions, notamment avec le 5,10-méthylène-tétrahydrofolate (Met-THF).

Les solutions S1 à S6 utilisent alternativement l'ATP ou l'adénosine 5'-phosphosulfate (APS). Rappelons que l'ATP a été retiré des réactions où il intervenait en tant que cofacteur, c'est-à-dire dans les transformations *ATP* → *ADP+Phosphate* et *ATP* → *AMP+Diphosphate*. Dans les réactions où il n'intervient pas dans ces transformations, il a été conservé. C'est le cas notamment de la réaction irréversible 2.5.1.6 (Figure 5.7), intervenant dans ces six solutions, où l'ATP participe réellement dans l'échange de matière puisque l'on retrouve l'adénine et le ribose de l'ATP dans la S-adénosyl-L-méthionine (SAM).

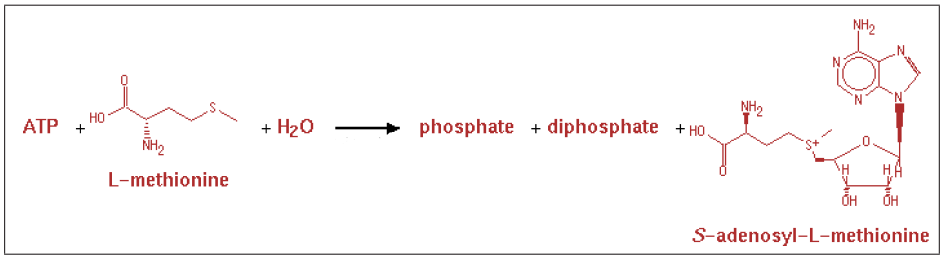

Figure 5.7. Réaction 2.5.1.6. On retrouve l'adénine et le ribose de l'ATP dans la S-adénosyl-L-méthionine.

Après l'application du filtre sur les cofacteurs, l'ATP ne peut être produit que par la réaction 2.7.7.4 qui est réversible (Figure 5.8). Selon notre définition, il est donc considéré comme un précurseur potentiel. Par ailleurs, l'APS qui produit l'ATP par la même réaction n'est produit également que par cette réaction, c'est pourquoi nous retrouvons les deux métabolites alternativement dans les solutions S1 à S6. La réaction 2.7.7.4 est le début de la voie de réduction du sulfate et pourrait intervenir également dans la synthèse de l'histidine et de la cystéine (voir plus loin).

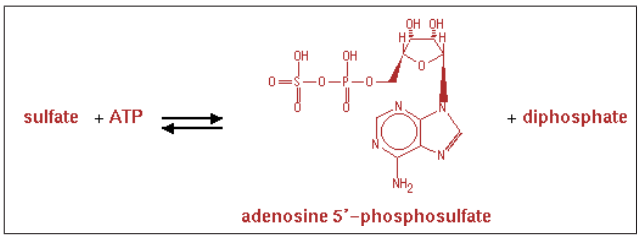

Figure 5.8. Réaction 2.7.7.4. Cette réaction est annotée comme réversible dans notre reconstruction. Après le filtre, c'est la seule qui produise l'ATP et l'APS. Ces deux métabolites sont ainsi considérés comme précurseurs l'un de l'autre.

Les solutions S1 à S6 correspondent à 3 cycles qui régénèrent SAM et la méthionine (Figures 5.9, 5.10 et 5.11). Dans les trois cycles, c'est la réaction 2.5.1.6 qui permet de régénérer SAM, nécessaire à la synthèse de la méthionine (Figure 5.7).

Les solutions S1 et S2 correspondent au cycle de la Figure 5.9. Le coproporphyrinogène (CP) y est indiqué comme précurseur et intervient directement dans l'approvisionnement du cycle. Ce composé intervient dans la synthèse des hèmes. Panek & O'Brian (2002) indiquent d'ailleurs la possibilité d'un approvisionnement de *Buchnera aphidicola APS* en hème ou en un de ces intermédiaires de la part du puceron.

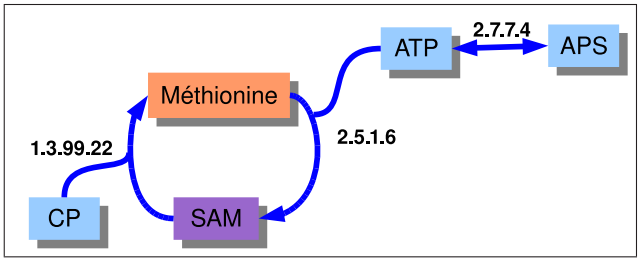

Figure 5.9. Vue synthétique du cycle de la méthionine utilisant le coproporphyrinogène III inféré à partir des résultats de PITUFO (solutions S1 et S2 de la Figure 5.6). Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **CP** : Coproporphyrinogène III ; **SAM** : S-Adénosyl-Méthionine ; **APS** : Adénosine 5'-Phosphosulfate

Les solutions S3 et S4 correspondent au cycle de la Figure 5.10 et indiquent un acide gras, le 7-kéto-8-aminopélargonate (KAPA), en tant que précurseur. Il est le deuxième métabolite intermédiaire dans la voie de référence de synthèse de la biotine. Il manque la première étape chez *Buchnera aphidicola APS*. KAPA produit la déthiobiotine intervenant en même temps que le donneur de sulfure dans le cycle de la Figure 5.10. Aucun indice de transport de KAPA ou de sa synthèse par le puceron n'a été trouvé dans la littérature.

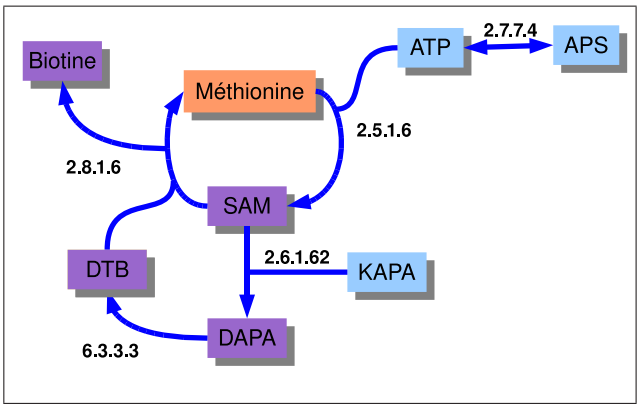

Figure 5.10. Vue synthétique du cycle de la méthionine utilisant KAPA inférré à partir des résultats de PITUFO (solutions S3 et S4 de la Figure 5.6). Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **SAM** : S-Adénosyl-Méthionine ; **APS** : Adénosine 5'-Phosphosulfate ; **KAPA** : 7-kéto-8-aminopélargonate ; **DAPA** : 7,8 Diaminopélargonate ; **DTB** : Déthiobiotine.

Les solutions S5 et S6 correspondent au cycle de la Figure 5.11. Les réactions codées par les gènes *lipA* et *lipB* interviennent dans la voie de synthèse et d'incorportation du lipoate. Le lipoate est un acide gras soufré et est un cofacteur essentiel de complexes enzymatiques, tels que la pyruvate déhydrogénase.

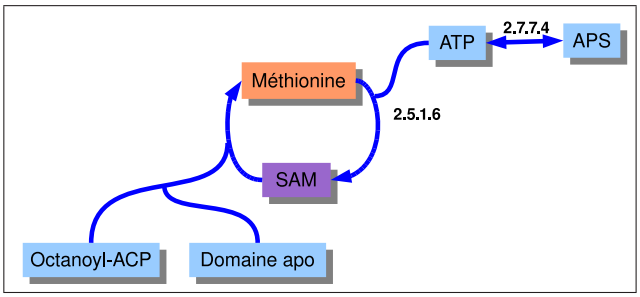

Figure 5.11. Vue synthétique du cycle de la méthionine utilisant la synthèse du lipoate, inférré à partir des résultats de PITUFO (solutions S5 et S6 de la Figure 5.6). Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **SAM** : S-Adénosyl-Méthionine ; **APS** : Adénosine 5'-Phosphosulfate ; **Domaine apo** : Domaine d'une enzyme dépendante du lipoate.

L'un des précurseurs des solutions S5 et S6, l'octanoyl-ACP, est composé d'une molécule d'octanoate (acide gras) lié à une protéine de transport de groupes acyls (**A**cyl **C**arrier **P**rotein). La première réaction de la voie de synthèse de lipoate qui manque chez *Buchnera aphidicola APS* consiste justement à lier ces deux molécules. Dans la deuxième étape, le groupement octanoyl est libéré par l'ACP et fixé sur un domaine précis d'une enzyme dépendante du lipoate. Ce domaine est noté "*domaine apo*" dans les Figures 5.6 et 5.11. La troisième étape transforme la chaîne carbonée en lipoate actif en lui associant deux atomes de souffre. C'est cette troisième étape qui utilise SAM et produit de la méthionine. Aucune information n'a été trouvée dans la littérature sur cette voie possible

de synthèse de la méthionine chez *Buchnera aphidicola APS*. Il est par ailleurs difficile d'imaginer une production intensive de méthionine par cette voie où elle apparaît plutôt comme un sous-produit. Enfin, elle fait intervenir majoritairement des macromolécules et dépasse le cadre de notre étude.

Ces trois cycles faisant intervenir la méthionine et SAM peuvent être considérés selon deux points de vue : le premier correspond à la synthèse de la méthionine, le second à la synthèse de SAM.

Dans le premier cas, pour produire de la méthionine destinée à la synthèse protéique de *Buchnera aphidicola APS* ou à l'approvisionnement du puceron, il faut un apport supplémentaire de SAM. Or, ce dernier ne semble produit qu'à partir de la méthionine chez *Buchnera aphidicola APS*, ce qui suggère un approvisionnement de SAM par le puceron. A notre connaissance, ce transport n'a pas encore été reporté.

Si on considère le second point de vue, celui de la synthèse de SAM, on peut imaginer que la méthionine serait produite essentiellement par la réaction 2.1.1.14 à partir de l'homocystéine. Ensuite, une partie serait utilisée dans la synthèse des protéines de la bactérie, une autre serait fournie au puceron, et enfin la dernière partie serait destinée à régénérer SAM. Une partie de SAM serait utilisée en tant que cofacteur et l'autre pourrait être régénérée en méthionine. Dans l'état actuel de nos connaissances, il n'est pas possible de trancher entre ces deux hypothèses.

L'unique précurseur trouvé pour la thréonine est l'aspartate, ce qui est indiqué également par Shigenobu *et al.* (2000). La voie chez *Buchnera aphidicola APS* est complète par rapport à la voie de référence, notamment chez *Escherichia coli K12*. Elle comporte cinq étapes et contient l'homosérine comme composé intermédiaire. L'aspartate en tant que précurseur de la thréonine est également confirmé par les expériences de radiotraçage de Liadouze *et al.* (1996) : l'introduction de glutamate ou d'aspartate radioactif se traduit par un marquage fort de la lysine dans les acides aminés libres ou protéiques du puceron symbiotique.

Par ailleurs, nous verrons plus tard lors de l'analyse des précurseurs de la glycine que celle-ci pourrait être capable de produire la thréonine.

Les 9 étapes de la voie de synthèse de la lysine à partir de l'aspartate sont bien détectées comme présentes chez *Buchnera aphidicola APS*. La voie commence par l'aspartate et utilise le glutamate. On retrouve bien dans les précurseurs trouvés par notre méthode ceux du glutamate (voir Section suivante) et l'aspartate. Là aussi, ces résultats sont confirmés par les analyses de radiotraçage réalisées par Liadouze *et al.* (1996).

- *La valine, la leucine et l'isoleucine*

Chez *Escherichia coli*, les voies de synthèse de référence de ces trois acides aminés commencent toutes par le pyruvate (Figure 5.12) et sont intimement liées.

Trois étapes des voies de la valine et de l'isoleucine sont catalysées par les mêmes enzymes, et le α-kéto-isovalérate, intermédiaire de la voie de la valine, est le point de départ d'un embranchement conduisant à la synthèse de la leucine. La dernière étape des 3 voies est une transamination avec le glutamate comme donneur de groupement amine.

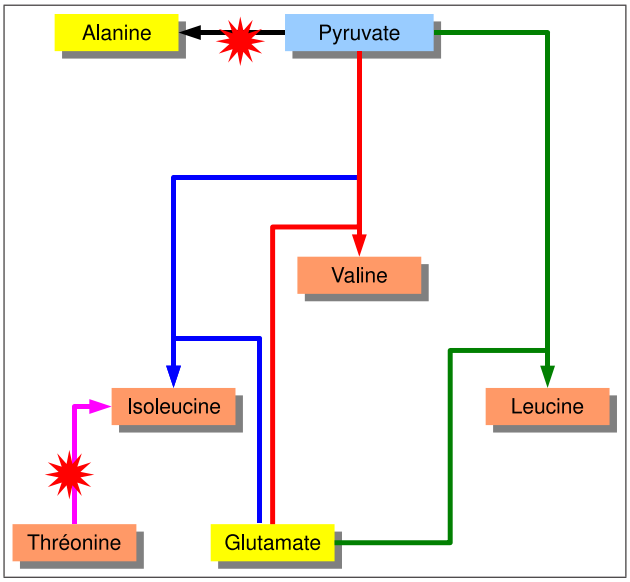

Figure 5.12. Vision simplifiée des voies de référence de synthèse de l'alanine, de la valine, de la leucine et de l'isoleucine à partir du pyruvate chez *Escherichia coli K12*. En jaune apparaissent les acides aminés non essentiels et en orange les acides aminés essentiels. Les étoiles rouges indiquent les voies qui n'existent plus chez *Buchnera aphidicola APS*.

Ces voies sont annotées comme complètes chez *Buchnera aphidicola APS*. Le pyruvate est directement produit par le glucose via la glycolyse. Pour l'isoleucine, la première réaction utilise, outre le pyruvate, le 2-oxobutanoate (OXO) qui n'est produit par aucune réaction et apparaît donc dans les listes de précurseurs. Chez *Escherichia coli* et d'autres organismes, une étape en amont permet de produire OXO à partir de la thréonine mais l'enzyme catalysant cette réaction semble absente chez *Buchnera aphidicola APS*. Nous n'avons trouvé aucun indice en ce qui concerne le transport d'OXO dans les cellules de *Buchnera aphidicola APS*. Cependant, d'après les premières annotations du génome du puceron, la transformation de la thréonine en OXO pourrait se dérouler dans celui-ci.

Le glutamate intervenant dans la dernière étape de ces 3 voies, les ensembles de précurseurs de ces trois acides aminés contiennent également ceux du glutamate. Des expériences de radiotraçage confirment bien le rôle de la bactérie dans la conversion du glutamate en isoleucine chez les pucerons (Liadouze *et al.*, 1996).

Aucun autre ensemble de précurseurs n'a été trouvé qui conduirait à une quelconque voie alternative pour la synthèse de ces acides aminés.

- *Le tryptophane et la phénylalanine*

Chez *Escherichia coli*, le tryptophane, la phénylalanine ainsi que la tyrosine sont produits à partir du phosphoénolpyruvate et de l'érythrose 4-phosphate. Les sept premières étapes sont communes et produisent le chorismate qui marque l'embranchement entre les voies du tryptophane, de la phénylalanine et de la tyrosine (Figure 5.13).

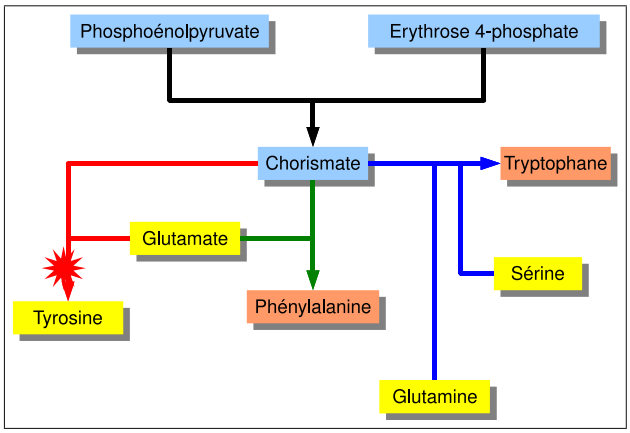

Figure 5.13. Vision simplifiée des voies de référence de synthèse du tryptophane, de la phénylalanine et de la tyrosine à partir du phosphoénolpyruvate et de l'érythrose 4-phosphate. En jaune apparaissent les acides aminés non essentiels et en orange les acides aminés essentiels. L' étoile rouge indique que la voie n'existe plus chez *Buchnera aphidicola APS*.

Chez *Buchnera aphidicola APS*, le chorismate peut être produit à partir du glucose en une douzaine d'étapes empruntant la voie des pentoses phosphates.

La voie du tryptophane apparaît comme complète chez *Buchnera aphidicola APS*. La glutamine et la sérine interviennent dans la voie, les autres métabolites sont produits par le glucose. Les ensembles de précurseurs trouvés pour le tryptophane sont donc le produit de ceux trouvés pour la sérine et pour la glutamine (voir plus loin).

La voie de la phénylalanine apparaît comme complète également chez *Buchnera aphidicola APS*. Parmi les intermédiaires de la voie, seul le glutamate n'est pas produit par le glucose. Les ensembles de précurseurs trouvés pour la phénylalanine correspondent donc à ceux trouvés pour le glutamate.

Pour ces deux métabolites, aucun autre ensemble de précurseurs n'a été trouvé, ce qui semble indiquer l'absence de voies alternatives.

- *L'histidine*

Chez *Escherichia coli*, la voie de synthèse de l'histidine commence par le ribose 5-phosphate et utilise le glutamate et la glutamine (Figure 5.14).

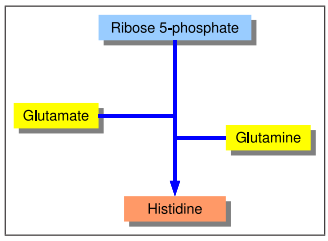

Figure 5.14. Vision simplifiée des voies de référence de synthèse de l'histidine à partir du ribose 5-phosphate. En jaune apparaissent les acides aminés non essentiels et en orange les acides aminés essentiels.

La voie apparaît comme complète chez *Buchnera aphidicola APS*. Le ribose 5-phosphate est synthétisé dans la voie des pentoses phosphate à partir du glucose. Dans la voie, seuls l'ATP, le glutamate et la glutamine ne sont pas produits à partir du glucose. Notre méthode retourne deux ensembles de précurseurs, les deux indiquent la glutamine, et alternativement l'ATP et l'APS en tant que précurseurs. L'ATP et l'APS sont des précurseurs interchangeables car ils ne peuvent être produits que par la même réaction réversible (voir plus haut).

Nous verrons plus tard que la glutamine seule est capable de produire le glutamate, les autres précurseurs du glutamate n'apparaissent donc pas dans les solutions.

On peut noter, comme précédemment pour la méthionine, que l'ATP intervient ici aussi comme fournisseur de matière et non d'énergie en réagissant avec le 5-phosphoribosyl 1-pyrophosphate pour produire le phosphoribosyl-ATP (Figure 5.15).

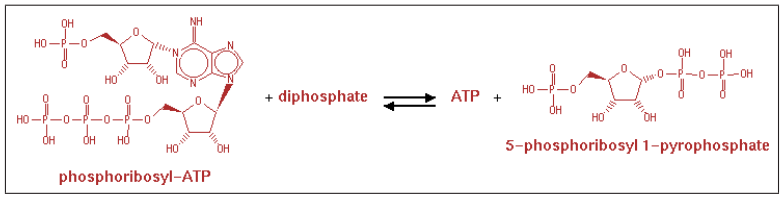

Figure 5.15. Réaction 2.4.2.17 qui utilise l'ATP en tant que fournisseur de matière dans la voie de synthèse de l'histidine.

- *L'arginine*

Chez *Escherichia coli*, la voie de synthèse de l'arginine débute par le α-kétoglutarate et se poursuit par le glutamate et l'ornithine et utilise l'aspartate (Figure 5.16).

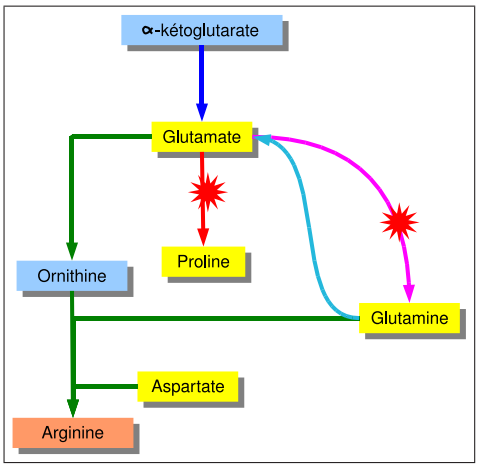

Figure 5.16. Vision simplifiée des voies de référence de synthèse du glutamate, de la glutamine, de la proline et de l'arginine à partir du α-kétoglutarate. En jaune apparaissent les acides aminés non essentiels et en orange les acides aminés essentiels. Les étoiles rouges indiquent les voies qui n'existent plus chez *Buchnera aphidicola APS*.

La voie de synthèse de l'arginine à partir du glutamate apparaît comme complète chez *Buchnera aphidicola APS*. On obtient un seul ensemble de précurseurs : la glutamine, l'ion bicarbonate (HCO_3^-) et l'aspartate. Ces trois métabolites ne sont produits par aucune réaction. De plus, la glutamine permet de produire le glutamate qui intervient dans la voie. Ce résultat est en partie confirmé par les expériences de radiotraçage de Liadouze *et al.* (1996). Ces derniers indiquent également que la proline serait convertie en arginine de façon plus importante dans les pucerons symbiotiques. Nous ne retrouvons pas ce résultat (la proline est même absente du réseau que nous avons reconstruit), ce qui suggère deux hypothèses. La première est qu'une réaction de conversion entre la proline et l'arginine manque dans notre reconstruction. La seconde hypothèse est que cette conversion se déroule dans le bactériocyte mais avec l'aide de métabolites apportés par *Buchnera aphidicola APS*, ce qui expliquerait le resultat de Liadouze *et al.* (1996).

b. Les acides aminés non essentiels

- *Le glutamate, la glutamine et la proline*

Dans les voies de référence, ces trois acides aminés sont produits à partir du α-kétoglutarate (Figure 5.16). Le glutamate et la glutamine ont une fonction très importante dans le cycle de l'azote puisqu'ils jouent un rôle clé dans l'intégration de l'azote dans les métabolites.

Aucune réaction dans notre reconstruction ne produit la glutamine, elle est donc considérée comme un précurseur potentiel. En effet, la glutamine synthétase, qui catalyse chez beaucoup d'organismes la transformation du glutamate en

glutamine est absente de chez *Buchnera aphidicola APS*.

Le glutamate est la source des groupements amines pour la plupart des acides aminés à travers des réactions de transamination. Il manque chez *Buchnera aphidicola APS* la glutamate synthétase qui permet de produire du glutamate à partir de la glutamine et de l'α-kétoglutarate.

Néanmoins, six réactions produisent le glutamate à partir de la glutamine. Elles interviennent dans différentes voies métaboliques : les voies de synthèse du tryptophane, de l'histidine, de l'arginine, du peptidoglycane (intervenant dans la membrane cellulaire), dans la conversion de l'UTP en CTP et dans la formation du NAD^+ à partir du déamino-NAD. Toutes ces réactions sont annotées comme irréversibles.

Ainsi, cinq ensembles de précurseurs sont indiqués pour le glutamate, chacun contenant seulement un acide aminé : la glutamine, l'isoleucine, la leucine, la phénylalanine, et la valine.

Les quatre derniers ensembles de précurseurs indiqués pour le glutamate utilisent le glutamate lui-même en tant que composé auto-régénéré, ce qui signifie que ce dernier est capable de s'auto-régénérer lorsque l'un des quatre acides aminés indiqués est présent, sans qu'il soit lui même le produit d'une voie de biosynthèse.

A partir de chacun de ces quatre acides aminés (isoleucine, leucine, phénylalanine, valine), une seule réaction de transamination produit le glutamate et correspond à la dernière étape de la synthèse de chacun de ces composés. Cette réaction peut être utilisée dans les deux sens, l'un correspondant à la synthèse de l'acide aminé et l'autre à sa dégradation. Chaque réaction utilise le α-kétoglutarate comme substrat. Dans le cas d'une dégradation d'un des acides aminés produisant du glutamate, le α-kétoglutarate serait régénéré par la synthèse d'un des trois autres acides aminés. Cependant, le fait que ces acides aminés soient essentiels pour le puceron rend peu probable leur dégradation et la production de glutamate par ce moyen.

La proline est absente du réseau, ce qui suggère qu'elle est fournie directement par le puceron. En effet, cet acide aminé n'est pas présent en quantité importante dans le phloème des plantes dont le puceron se nourrit mais ce dernier est capable d'interconvertir le glutamate en proline. Le gène a été identifié chez le puceron et cette activité est confirmée par des expériences de radiotraçage (Febvay *et al.*, 1995).

- *La sérine, la glycine et la cystéine*

Chez *Escherichia coli*, la voie de synthèse de référence de la sérine débute par le 3-phosphoglycérate et utilise le glutamate. La sérine est ensuite utilisée pour synthétiser la glycine et la cystéine (Figure 5.17).

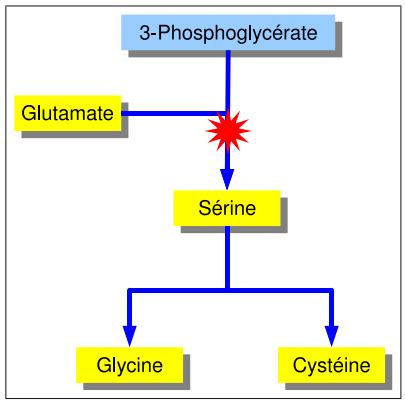

Figure 5.17. Vision simplifiée des voies de référence de synthèse de la sérine, de la glycine et de la cystéine à partir du 3-phosphoglycérate chez *Escherichia coli K12*. En jaune apparaissent les acides aminés non essentiels. Les étoiles rouges indiquent les voies qui n'existent plus chez *Buchnera aphidicola APS*.

La voie de la sérine est incomplète chez *Buchnera aphidicola APS*. Sur les trois étapes qu'elle compte à partir du 3-phosphoglycérate, seule la seconde est présente. *Buchnera aphidicola APS* est donc incapable de synthétiser la sérine par la voie traditionelle. Notre méthode retourne pourtant 9 ensembles de précurseurs (Figure 5.18). A partir de notre reconstruction, il est difficile de déterminer quel composé, de la glycine ou de la sérine, produit l'autre. En effet, la sérine n'est produite que par une seule réaction, la réaction 2.1.2.1 qui est réversible et qui utilise comme substrat la glycine. Rappelons que les métabolites qui ne sont produits que par une seule réaction réversible sont envisagés comme des précurseurs potentiels, la sérine est donc considérée ici comme un précurseur potentiel. Comme la sérine ne peut être produite que par la glycine, les deux composés ont les mêmes ensembles de précurseurs. La réaction 2.1.2.1 apparaît dans la voie de référence de la synthèse de la glycine. Elle apparaît dans l'autre direction dans la voie de polyglutamation du folate où cette fois, la sérine est produite à partir de la glycine. Il est probable que cette réaction se produit préférentiellement dans une direction au sein des cellules de *Buchnera*. Dans le cas où elle fonctionne généralement dans le sens de la production de la glycine, il faut supposer que la sérine est importée par *Buchnera aphidicola APS*. Cependant, aucun indice d'un tel transport n'a été trouvé dans la littérature.

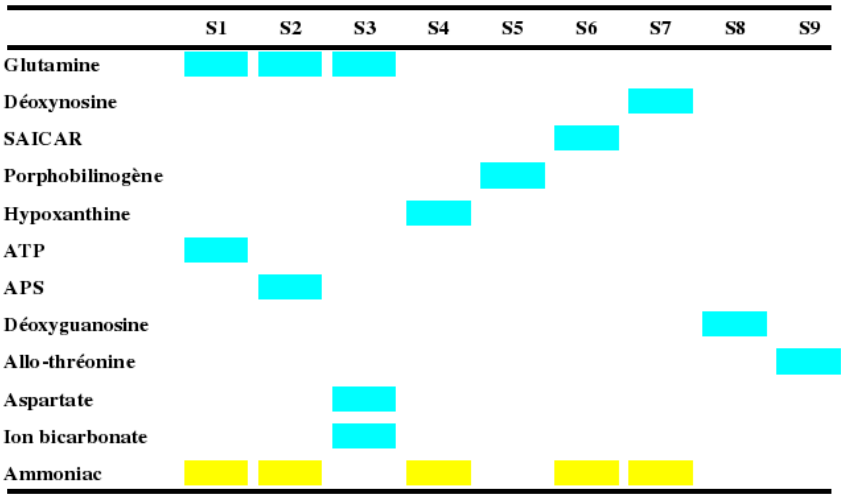

Figure 5.18. Ensembles de précurseurs trouvés par PITUFO pour la sérine et la glycine. Chaque colonne correspond à un ensemble de précurseurs. Un carré bleu signifie que le métabolite est un des précurseurs de la solution et un carré jaune signifie que le métabolite est utilisé en tant que métabolite auto-régénéré.

Un des précurseurs pour la glycine et la sérine trouvés par PITUFO est l'allothréonine. Ce métabolite est décrit dans les pathway-tools comme un mélange probable d'énantiomères de la thréonine. Deux énantiomères d'une molécule ont la même formule chimique mais une configuration spatiale différente : ils sont images l'un de l'autre dans un miroir. La thréonine a ainsi deux formes : L et D, la première étant celle utilisée dans la formation des protéines. Elle n'intervient que dans une seule réaction réversible dans notre reconstruction, catalysée par l'enzyme GlyA et qui produit la glycine à partir de l'allothréonine (voir Figure 5.19). Dans MetaCyc, il est indiqué que la thréonine peut aussi être le substrat direct de cette réaction. La présence de cette réaction indique un lien possible direct entre la glycine et la thréonine chez *Buchnera aphidicola APS*. Cependant, les expériences de radiotraçage de Liadouze *et al.* (1996) montrent que le carbone radioatif de la glycine injectée dans le milieu nutritionel du puceron ne se retrouve pas dans la thréonine extraite des tissus de l'insecte, qu'il soit symbiotique ou aposymbiotique. Par ailleurs, nous n'avons trouvé aucun indice dans la littérature à propos d'une production de glycine à partir de la thréonine.

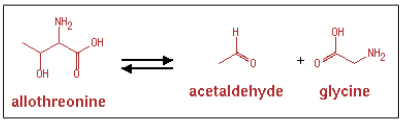

Figure 5.19. Réaction réversible catalysée par l'enzyme GlyA produisant la glycine à partir de l'allothréonine.

Cette solution mise à part, toutes les autres impliquent l'utilisation de l'ammoniac pour synthétiser la glycine à partir de la réaction catalysée par l'enzyme

codée par le gène *lpdA* chez *Buchnera aphidicola APS*. Cette réaction produit la glycine et le tétrahydrofolate (THF) à partir de l'ammoniac, du dioxyde de carbone (CO2) et du 5,10-méthylène-tétrahydrofolate (METHYLENE-THF). Le THF et le METHYLENE-THF forment un couple de cofacteurs, ils ont donc été supprimés de cette réaction, leurs précurseurs n'ont ainsi pas été calculés. Le CO2 est produit par l'injection de glucose. Dans MetaCyc, cette réaction est indiquée comme une voie possible de synthèse de la glycine chez certains eucaryotes. Chez les bactéries, elle semble fonctionner plutôt dans l'autre sens, notamment dans les voies de synthèses des folates. Cependant, rien à notre connaissance ne permet de décider d'un sens préférentiel pour cette réaction.

Si cette réaction se produit chez *Buchnera aphidicola APS* dans le sens de la production de la glycine, l'ammoniac formerait donc un "carrefour" dans la production de la glycine. Selon les ensembles de précurseurs envisagés, l'ammoniac est produit directement (solutions S3, S5 et S8 de la Figure 5.18) ou est considéré comme un métabolite auto-régénéré (solutions S1, S2, S4, S6, et S7). La solution S5 ne comporte que le porphobilinogène (PORPHO). La synthèse d'ammoniac est directe à partir de PORPHO quand les molécules de celui-ci s'assemblent en hydroxyméthylbilane dans la réaction 2.5.1.61 (voir Figure 5.20). Dans les voies de référence, le porphobilinogène intervient dans la synthèse de l'hème, qui semble incomplète chez *Buchnera aphidicola APS*. C'est un candidat intéressant en tant que précurseur puisque son transport dans les cellules de *Buchnera* a déjà été envisagé (Panek & O'Brian, 2002).

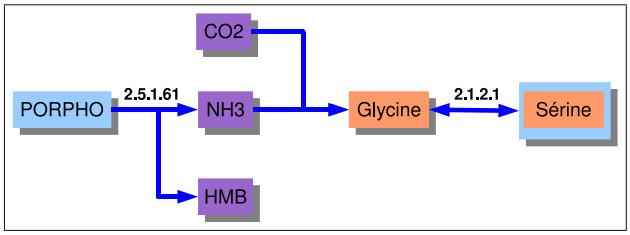

Figure 5.20. Vue synthétique de la voie de synthèse de la glycine et de la sérine à partir du porphobilinogène inférée à partir des résultats de PITUFO (solution S5 de la Figure 5.18). Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **PORPHO** : Porphobilinogène ; **NH3** : ammoniac ; **HMB** : hydroxyméthylbilane ; **CO2** : dioxyde de carbone

La solution S3 qui permet aussi de produire l'ammoniac sans besoin de cycle comprend la glutamine, l'ion bicarbonate et l'aspartate. Ces trois métabolites se trouvent au début de la voie de synthèse *de novo* de l'uridine mono-phosphate (UMP), qui produit en 10 réactions la cytidine diphosphate (CDP) en partant de la glutamine et de l'ion bicarbonate. CDP intervient ensuite comme un des points de départs de la voie de synthèse des pyrimidines déoxyribonucléotides jusqu'au dCTP qui, en se transformant en dUTP par la réaction 3.5.4.13, produit l'ammoniac. Cette solution est particulièrement intéressante puisqu'elle est confirmée par les expériences de radiotraçage de Liadouze *et al.* (1996) montrant un flux de matière de l'aspartate et de la glutamine vers la sérine (sans passer par

la thréonine) chez les pucerons symbiotiques qui n'existe pas chez les pucerons aposymbiotiques.

La solution S8 ne comporte qu'un seul précurseur : la déoxyguanosine. A partir de ce métabolite, l'ammoniac pourrait être synthétisé en 6 étapes (voir Figure 5.21). La déoxyguanosine ne participe qu'à une seule réaction qui est réversible, raison pour laquelle elle est considérée comme précurseur potentiel. La voie passe par une transformation de la guanine en guanosine monophosphate (GMP), qui est phosphorylée en guanosine diphosphate (GDP) puis en guanosine triphosphate (GTP). Ce dernier forme ensuite le 2,5-diamino-6-(ribosylamino)-4-(3H)-pyrimidinone 5'-phosphate (DARP), qui, par la réaction 3.5.4.26, produit de l'ammoniac et du 5-amino-6-(5'-phosphoribosylamino)uracil. Cette réaction est par ailleurs la première étape de la voie de synthèse de la flavine. Aucun indice de transport de déoxyguanosine ni d'ailleurs d'un quelconque nucléotide dans *Buchnera* n'a été trouvé dans la littérature.

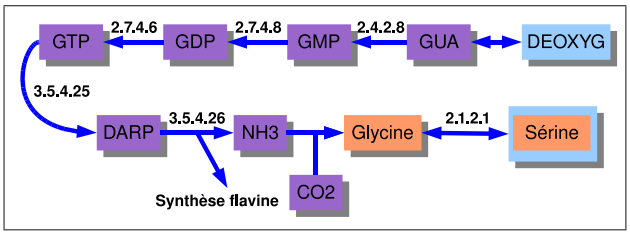

Figure 5.21. Vue synthétique de la voie de synthèse de la sérine à partir de la déoxyguanosine inférée d'après les résultats de PITUFO (solution S8 de la Figure 5.18). La première et la dernière réactions sont réversibles. Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **DEOXYG** : Déoxyguanosine ; **GUA** : Guanine ; **GMP** : Guanosine monophosphate ;**GDP** : Guanosine diphosphate ;**GTP** : Guanosine triphosphate ; **DARP** : 2,5-diamino-6-(ribosylamino)-4-(3H)-pyrimidinone 5'-phosphate ; **NH3** : ammoniac ; **CO2** : dioxyde de carbone.

Cinq solutions (S1, S2, S4, S6 et S7) considèrent l'ammoniac comme métabolite auto-régénéré. Après une inspection du réseau métabolique, nous avons trouvé que le cycle correspondant comporte 5 étapes. Il est approvisionné par l'inosine-5'-phosphate (IMP) et passe par la phosphorylation du GMP en GTP, série de réactions que l'on trouve déjà à partir de la solution S8 décrite plus haut. Il y a trois chemins directs entre ces cinq ensembles de solutions et IMP (Figure 5.22).

Les solutions S4 et S7 proposant respectivement l'hypoxanthine et la déoxynosine sont en fait équivalentes puisque ces deux métabolites ne peuvent être produits que par la même réaction réversible. Les solutions S1 et S2 ont déjà été décrites dans le cas de la synthèse de l'histidine. En effet, la glutamine et l'ATP permettent d'initier la voie de synthèse de l'histidine dont un des intermédiaires est l'aminoimidazole carboxamide ribonucleotide (AICAR). La transformation de l'AICAR en IMP est une des étapes clés dans la synthèse des nucléotides (Zientz *et al.*, 2004). AICAR est produit également dans la voie de référence de la synthèse des purines à partir du 5-phosphoribosyl 1-pyrophosphate (PRPP) mais les premières étapes de cette voie sont absentes dans *Buchnera aphidicola APS* chez

laquelle la voie commence par le 5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole (SAICAR), qui correspond à la solution S6. Aucun indice du transport de SAICAR dans les cellules de *Buchnera* n'a été trouvé dans la littérature.

Cependant, les coefficients stœchiométriques des métabolites participant au cycle approvisionné par l'IMP nous apprennent que le bilan de production de l'ammoniac est nul. Même si le GMP peut être produit également à partir de la déoxyguanosine, la voie jusqu'à l'ammoniac semble coûteuse (2 molécules d'ATP sont consommées).

Par ailleurs, il est très probable que l'ammoniac, en forte quantité dans le phloème des plantes dont se nourrit le puceron, soit diffusé librement à travers les membranes cellulaires de *Buchnera aphidicola APS* (Douglas, 1998). Il est difficilement imaginable que, chez la bactérie, la synthèse de la glycine nécessite la synthèse de l'ammoniac s'il peut être capté directement dans le bactériocyte.

Ainsi, les voies que nous venons de décrire indiquent plutôt une interconnection entre les voies de synthèse des nucléotides, de la flavine et de la glycine. La glycine pourrait ainsi être produite directement à partir de l'ammoniac capté dans le bactériocyte mais pourrait être également produite à partir de l'ammoniac résultant des voies de synthèse des nucléotides et de la flavine.

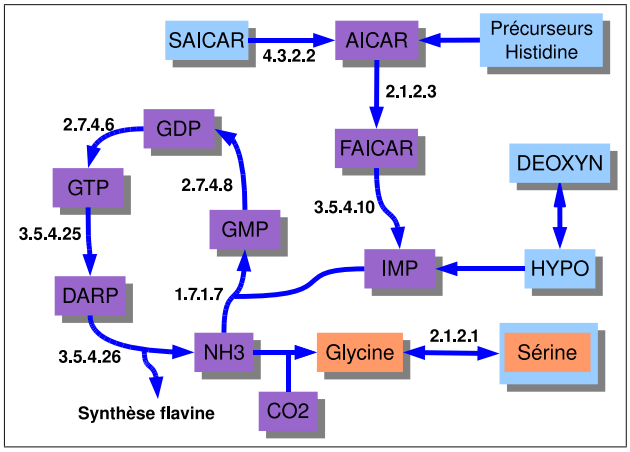

Figure 5.22. Vue synthétique des voies de synthèse de la glycine et de la sérine inférées à partir des résultats de PITUFO (solutions S1, S2, S4, S6, S7 de la Figure 5.18). Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **DEOXYN** : Déoxynosine ; **GMP** : Guanosine monophosphate ;**GDP** : Guanosine diphosphate ;**GTP** : Guanosine triphosphate ; **DARP** : 2,5-diamino-6-(ribosylamino)-4-(3H)-pyrimidinone 5'-phosphate ; **NH3** : ammoniac ; **IMP** : Inosine monophosphate ; **HYPO** : Hypoxanthine ; **FAICAR** : Phosphoribosyl-formamido-carboxamide ; **AICAR** : Aminoimidazole carboxamide ribonucleotide ; **SAICAR** : 5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole ; **CO2** : dioxyde de carbone.

La voie de la cystéine à partir de la sérine apparaît comme complète chez *Buchnera aphidicola APS* (Figure 5.23). Elle utilise le sulfide d'hydrogène provenant de la réduction du sulfate. Celle-ci débute par la réaction réversible 2.7.7.4 qui transforme le sulfate et l'ATP en adénosine 5'phosphosulfate (APS) et en diphosphate (Figure 5.8).

Comme nous l'avons vu auparavant, l'ATP et l'APS ne sont produits que par cette réaction et sont donc considérés comme précurseurs potentiels. Ainsi, les précurseurs trouvés pour la cystéine sont les mêmes que ceux trouvés pour la sérine auxquels on ajoute soit l'ATP et le sulfate, soit l'APS. La synthèse de la cystéine à partir de l'APS a déjà été signalée par Zientz *et al.* (2004). Cependant, aucun indice de transport de l'APS dans les cellules de *Buchnera aphidicola APS* n'a été trouvé dans la littérature.

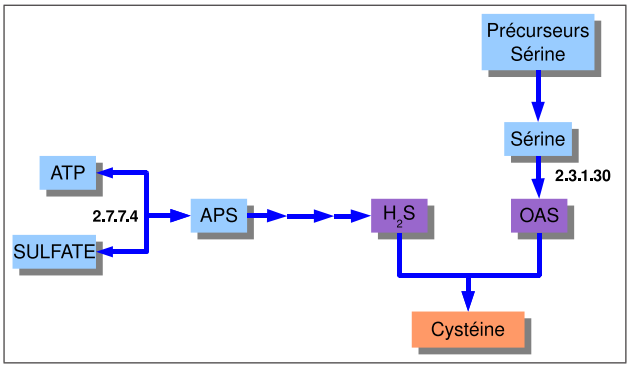

Figure 5.23. Vue synthétique de la voie de synthèse de la cystéine à partir de la sérine et du sulfate inférée par Pathologic et par PITUFO. Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **APS** : Adénosine 5'phosphosulfate ; **H_2S** : Sulfide d'hydrogène ; **OAS** : O-acétyl-L-sérine

- *L'aspartate et l'asparagine*

Chez *Escherichia coli*, les voies de synthèse de l'aspartate et de l'asparagine passent par l'oxaloacétate (Figure 5.5).

Dans notre reconstruction, l'aspartate n'est produit par aucune réaction, il est donc un précurseur potentiel. Il y a d'ailleurs des indices forts dans la littérature sur l'absorption d'aspartate par *Buchnera* (Whitehead & A.E.Douglas, 1993).

L'asparagine est absente du réseau des petites molécules, ce qui suggère qu'elle est fournie directement par le puceron.

- *L'alanine*

Chez *Escherichia coli*, la synthèse de l'alanine passe par le pyruvate comme pour la valine, la leucine et l'isoleucine (Figure 5.12). Chez *Escherichia coli*, trois voies différentes permettent de produire l'alanine. La première se termine par la réaction qui produit l'alanine à partir du pyruvate et du glutamate. La seconde se termine par la réaction qui produit l'alanine à partir du pyruvate et de la valine. La troisième se termine par la réaction qui produit l'alanine à partir d'une molécule de cystéine et d'une cystéine insérée dans un peptide (que nous désignerons par PROT-CYS) (Figure 5.24). Seule cette dernière voie apparaît comme présente dans les voies de synthèse de l'alanine inférées par Pathologic.

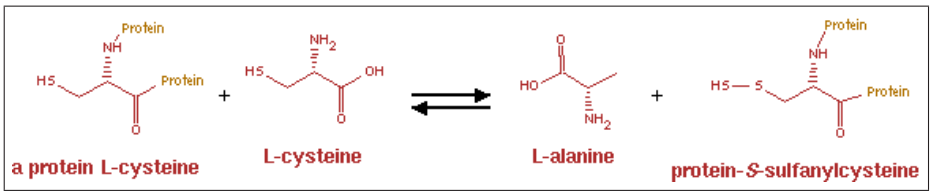

Figure 5.24. Synthèse d'alanine à partir de la cystéine inférée à partir de Pathologic. "a protein L-cysteine" correspond à une molécule de cystéine insérée dans un peptide.

Nous trouvons 32 solutions pour l'alanine. En réalité, une partie correspond aux précurseurs de la la cystéine auxquels on ajoute la PROT-CYS qui n'a pas de précurseur dans notre reconstruction. Les autres solutions contiennent également les précurseurs de la cystéine mais auxquels on ajoute l'uridine (nucléoside). L'uridine n'apparaît que dans une seule réaction irréversible qui produit l'alanine à partir de la cystéine (Figure 5.25). *Buchnera aphidicola APS* ne semble pas pouvoir produire l'uridine mais est capable de produire le nucléotide correspondant, l'uridine 5'-monophosphate (UMP). Les deux molécules diffèrent seulement par la présence d'un groupement phosphate chez UMP. Le passage d'UMP à l'uridine se fait classiquement grâce à une nucléotidase qui catalyse l'hydrolyse d'un nucléotide en nucléoside et phosphate. Aucun indice d'une telle activité enzymatique n'a cependant été trouvé chez *Buchnera aphidicola APS*. Par contre, un gène pouvant coder une telle fonction fait partie de ceux fortement exprimés chez le puceron (Nakabachi *et al.*, 2005). Malheureusement, aucun transporteur de nucléosides n'a pu être mis en évidence à ce jour chez *Buchnera aphidicola APS* (Zientz *et al.*, 2004). D'un autre côté, un gène semble pouvoir coder pour cette fonction chez *Acyrthosiphon pisum* (LOCUS XM_001943584 dans Genbank).

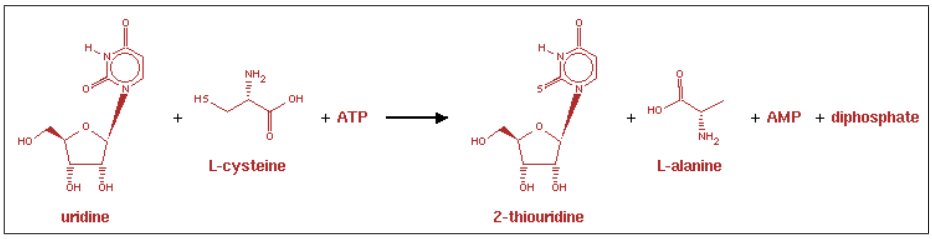

Figure 5.25. Synthèse d'alanine à partir de la cystéine et de l'uridine.

- *La tyrosine*

Chez *Escherichia coli*, la voie de synthèse de la tyrosine passe par le phosphoénolpyruvate et l'érythrose 4-phosphate comme le tryptophane et la phénylalanine. Le chorismate marque l'embranchement entre les trois voies de synthèse (Figure 5.13). La transformation du chorismate se déroule normalement en trois étapes. La première est catalysée par une chorismate mutase dont le gène a été annoté

chez *Buchnera aphidicola APS*. Elle transforme le chorismate en préphénate, utilisée dans la synthèse de la phénylalanine. La dernière étape est catalysée par une tyrosine aminotransférase codée chez *Escherichia coli* par le gène *tyrB*, absent chez *Buchnera aphidicola APS*. D'après les indications données par Shigenobu *et al.* (2000), le gène *hisC* a été associé à cette fonction (voir Section 5.4.2). Par contre, la seconde étape catalysée par une préphénate déhydrogénase codée par le gène *tyrA* apparaît comme absente chez *Buchnera aphidicola APS* (Figure 5.26).

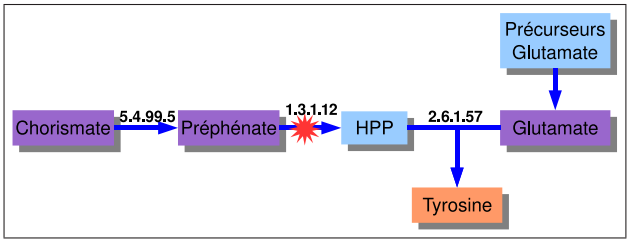

Figure 5.26. Vue synthétique de la voie de synthèse de la tyrosine à partir du chorismate, incomplète chez *Buchnera aphidicola APS* (la réaction 1.3.1.12 n'apparaît pas dans la reconstruction métabolique). Tous les métabolites des réactions ne sont pas représentés. Les métabolites dans les cadres bleus sont les précurseurs indiqués par PITUFO. **HPP** : P-hydroxyphénylpyruvate.

Chez *Buchnera aphidicola APS*, seule la dernière réaction de cette voie peut produire la tyrosine. Cette réaction utilise comme substrats le glutamate et le p-hydroxyphénylpyruvate (HPP). On retrouve dans les précurseurs de la tyrosine le glutamate et ses précurseurs ainsi que HPP qui n'a pas de précurseur et qui n'intervient que dans cette réaction. Aucun indice de synthèse par le puceron et de transport de HPP dans *Buchnera aphidicola APS* n'a été mis en évidence jusqu'à maintenant. Par contre, il a été montré que le puceron est capable de convertir la phénylalanine en tyrosine grâce à une phénylalanine hydroxylase. De plus, Nakabachi *et al.* (2005) ont montré que le gène correspondant est surexprimé chez le puceron. L'hypothèse la plus probable reste donc que la tyrosine est fournie par le puceron à *Buchnera aphidicola APS*.

5.5 Discussion

Avec PITUFO, nous proposons la première définition de précurseurs prenant en compte les cycles de manière explicite. L'algorithme que nous avons développé énumère toutes les solutions minimales. L'intérêt de cette méthode est en outre de s'affranchir totalement des données stœchiométriques du réseau. En effet, notre méthode n'utilise que la topologie du réseau et ne nécessite aucune donnée quantitative. Elle est ainsi plus facile à mettre en place qu'une analyse en balance des flux où la condition d'état d'équilibre nécessite d'avoir une matrice stœchiométrique exacte et des limites du système précises. Enfin, les possibles utilisations de PITUFO vont bien au-delà de la thématique qui nous intéressait

ici, de la définition de milieux nutritionnels à la vérification des reconstructions métaboliques.

Certains points méthodologiques peuvent cependant être améliorés. Le premier est le temps d'exécution de l'algorithme qui empêche une utilisation intensive de la méthode sur de grands réseaux.

D'autres paramètres pourraient être également intégrés dans la méthode. Par exemple, Nikoloski *et al.* (2008) proposent d'ajouter aux paramètres une liste de métabolites qui ne doivent pas apparaître dans le scope des précurseurs potentiels. L'intérêt serait, par exemple, de définir les ensembles de précurseurs d'une cible sans que certains composés toxiques soient produits.

Des extensions de la méthode pourraient aussi aider à l'interprétation des résultats. Par exemple, la reconstitution des chemins entre les ensembles de précurseurs et les cibles en prenant en compte l'intervention des métabolites autorégénérés se fait pour l'instant à la main.

Une autre aide à l'interprétation serait de définir des métabolites clés à partir des solutions. On définirait ces métabolites clés comme ceux intervenant plus d'un certain nombre de fois entre les différentes solutions et les cibles. Il s'agirait ainsi de métabolites "carrefours" dans la synthèse des métabolites cibles, ce qui pourrait enrichir l'analyse des résultats.

L'application sur le réseau de *Buchnera aphidicola APS* a permis de valider notre méthode mais aussi de proposer de nouvelles pistes pour mieux comprendre la synthèse des acides aminés chez *Buchnera aphidicola APS*.

L'ensemble des résultats trouvés par PITUFO sur les 20 acides aminés est résumé dans la Figure 5.27.

Deux acides aminés n'apparaissent pas du tout dans le réseau : la proline et l'asparagine. Ils seraient non seulement fournis directement par le puceron mais ils n'interviendraient pas du tout non plus dans le reste du réseau, mis à part pour se lier aux ARNs de transfert. Ces deux composés sont trouvés en concentrations plus élevées chez les pucerons aposymbiotiques que chez les pucerons symbiotiques (Liadouze *et al.*, 1995), ce qui indiquerait leur utilisation par *Buchnera aphidicola APS*. Aucune preuve d'un transport de proline dans les cellules de *Buchnera* n'a été trouvé dans la littérature. Sasaki & Ishikawa (1995) ont mesuré les concentrations d'asparagine et d'aspartate dans l'hémolymphe du puceron et dans le bactériocyte. Dans l'hémolymphe, l'asparagine a une concentration élevée alors que l'aspartate a une concentration faible. Dans le bactériocyte, la proportion est inversée. Ainsi, les auteurs suggèrent que l'asparagine est transportée dans le bactériocyte puis transformée en aspartate. Par ailleurs, Whitehead & A.E.Douglas (1993) ont montré que l'aspartate est activement absorbé par les *Buchnera*. D'un autre côté, si *Buchnera* ne dispose réellement d'aucun moyen de produire l'asparagine, alors celle-ci doit être prélevée du bactériocyte. Pourtant, aucun indice d'absorption d'asparagine par *Buchnera* n'a été trouvé dans la littérature.

Nous nous trouvons devant le même paradoxe en ce qui concerne la glutamine. Dans notre reconstruction, la glutamine n'intervient que comme substrat

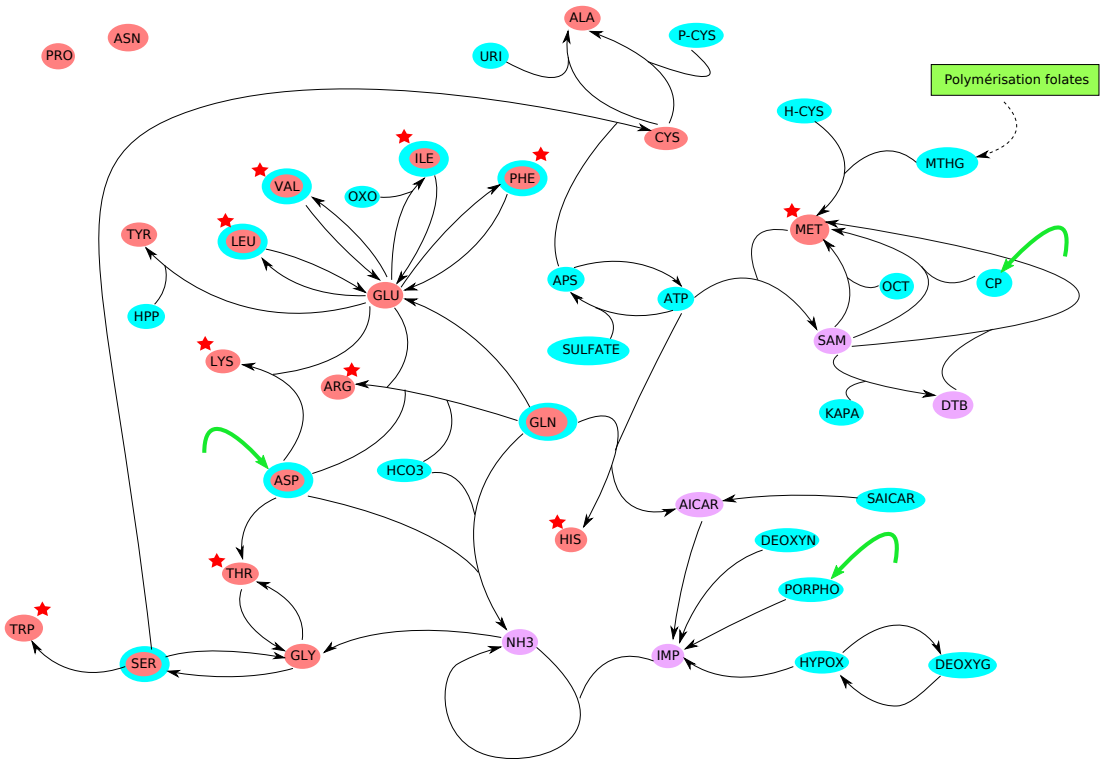

Figure 5.27. Résumé des résultats trouvés par PITUFO pour les acides aminés dans le réseau de *Buchnera aphidicola APS*. En bleu apparaissent les précurseurs potentiels trouvés par PITUFO. En rouge apparaissent les acides aminés. Une étoile rouge indique que l'acide aminé est essentiel pour le puceron. Une flèche verte vers un métabolite indique que le transport de celui-ci a déjà été envisagé chez *Buchnera*. Les flèches indiquent ici une réaction ou des séries de réactions. **AICAR** : Aminoimidazole carboxamide ribonucleotide ; **ALA** : Alanine ; **APS** : Adénosine 5'-Phosphate ; **ARG** : Arginine ; **ASN** : Asparagine ; **ASP** : Aspartate ; **CP** : Coproporphyrinogène ; **CYS** : Cystéine ; **DEOXYG** : Déoxyguanosine ; **DEOXYN** : Déoxynosine ; **DTB** : Déthiobiotine ; **GLN** : Glutamine ; **GLY** : Glycine ; **HCO3** : Ion bicarbonate ; **H-Cys** : Homocystéine ; **HIS** : Histidine ; **HPP** : P-hydroxyphénylpyruvate ; **HYPO** : Hypoxanthine ; **IMP** : Inosine-5'-phosphate ; **KAPA** : 7-kéto-8-aminopélargonate ; **LEU** : Leucine ; **LYS** : Lysine ; **MET** : Méthionine ; **MTHG** : 5-méthyltétrahydroptéroyltri-L-glutamate ; **NH3** : Ammoniac ; **OCT** : Octanoyl-ACP ; **OXO** : 2-oxobutanoate ; **P-CYS** : Cystéine insérée dans un peptide ; **PHE** : Phénylalanine ; **PORPHO** : Porphobilinogène ; **PRO** : Proline ; **SAM** : S-adénosyl-L-méthionine ; **SER** : Sérine ; **THR** : Thréonine ; **TRP** : Tryptophane ; **TYR** : Tyrosine ; **URI** : Uridine ; **VAL** : Valine ; **ILE** : Isoleucine

et jamais comme produit d'une réaction. De même, on trouve un taux élevé de glutamine dans les pucerons aposymbiotiques (Sasaki & Ishikawa, 1995; Liadouze *et al.*, 1995) et une forte activité de synthèse de la glutamine de la part du puceron (Sasaki & Ishikawa, 1993). Pourtant là aussi, aucun indice de transport de la glutamine dans les cellules de *Buchnera* n'a été trouvé. Les expériences de traçage radioactif de Sasaki & Ishikawa (1993) suggèrent que la glutamine serait importée dans le bactériocyte puis hydrolysé en glutamate, qui serait, lui, transporté activement dans les cellules de *Buchnera* et impliqué dans la synthèse des acides aminés. Mais le fait que la glutamine ne semble pas produite au sein de *Buchnera* suggère fortement qu'elle soit elle aussi transportée pour participer à la synthèse des protéines mais également à celle d'autres acides aminés.

La recherche des précurseurs par PITUFO nous montre en effet que seuls

trois acides aminés ne contiennent pas au moins une fois la glutamine parmi leurs précurseurs potentiels : la proline, l'asparagine et la méthionine. En outre, le transport du glutamate "favorisé" par rapport au transport de la glutamine est quelque peu en contradiction avec le fait que le glutamate peut être produit directement à partir de la glutamine au sein des cellules de *Buchnera*. Trouver la glutamine en tant que précurseur potentiel pour la sérine, l'isoleucine et la lysine est en accord avec les résultats des expériences de radiotraçage de Liadouze *et al.* (1996). Ceci semble indiquer que la sérine est bien produite au sein des cellules de *Buchnera* et que la réaction 2.1.2.1 pourrait être utilisée dans le sens de la production de la sérine. Par ailleurs, Liadouze *et al.* (1996) indiquent un flux de carbone entre la glutamine et la thréonine qui serait dû à la présence des symbiotes. Ceci pourrait valider la production de la thréonine à partir de la glycine catalysée par l'enzyme GlyA (voir Figure 5.19).

L'aspartate non plus ne semble produit par aucune réaction dans le réseau de *Buchnera*, suggérant un transport de la part du symbiote, d'ailleurs mis en évidence par Whitehead & A.E.Douglas (1993). Son rôle dans la synthèse globale des acides aminés s'est atténué avec la réduction du métabolisme de *Buchnera*. En effet, il disparaît des voies de synthèse de la méthionine et de l'asparagine. Son rôle en tant que précurseur de la lysine, de la thréonine et de la sérine est confirmé par les expériences de radiotraçage de Liadouze *et al.* (1996). Par ailleurs, les mêmes expériences indiquent un flux de matière entre l'aspartate et l'isoleucine alors que l'aspartate ne figure pas parmi les précurseurs de l'isoleucine trouvés par notre méthode. Ceci pourrait s'expliquer par la transformation au sein du bactériocyte de la thréonine en isoleucine par le biais d'une thréonine déaminase. L'annotation fonctionelle du puceron pourrait valider ou non cette hypothèse.

La production d'isoleucine, de leucine, de phénylalanine et de valine par le biais du glutamate est confirmée par les résultats de radiotraçage de Sasaki & Ishikawa (1995) où le glutamate a été identifié comme donneur d'azote pour ces quatre acides aminés. Notre reconstruction nous indique également l'inverse comme possible : les quatres acides aminés s'ajoutent à la glutamine en tant que précurseurs potentiels du glutamate. Comme aucun autre moyen ne semble exister pour produire ces acides aminés, il est plus que probable que la direction favorisée part du glutamate. Nous avons vu que la glycine pourrait être synthétisée par deux moyens : à partir de la thréonine ou à partir d'ammoniac et de CO2. L'ammoniac est très vraisemblablement importé chez *Buchnera aphidicola APS* mais il est intéressant de voir aussi qu'il est le sous-produit d'autres voies de synthèse, notamment celle de la flavine et des nucléotides.

Le précurseur le plus plausible pour la méthionine semble être l'homocystéine. Cependant, des expériences montrent que des cellules de *Buchnera* isolées produisent de la méthionine à partir de sulfate et que les pucerons symbiotiques produisent cet acide aminé alors que ceux qui sont aposymbiotiques en sont incapables (Douglas, 1988). Le réseau métabolique de *Buchnera* ne permet pas de rendre compte de ces résultats expérimantaux : aucune voie ne semble exister

entre le sulfate et la méthionine. Une possibilité serait que *Buchnera* serait en réalité capable de catalyser la transsulfuration de la cystéine en homocystéine.

Cette catalyse serait pourrait être assurée par une enzyme au spectre plus large qu'habituellement. L'incohérence entre ces résultats expérimentaux et ceux fournis par PITUFO montre également l'intérêt de cette méthode pour relever les inconsistances du réseau.

Grâce aux résultats de PITUFO, nous avons vu également que la méthionine pourrait participer à plusieurs cycles. Ces cycles ne semblent pas fonctionnels en tant que tels pour régénérer la méthionine mais pourraient correspondre à deux fonctionnements alternatifs, selon certaines conditions : la synthèse de la S-adénosyl-méthionine (SAM) à partir de la méthionine, ou la synthèse de la méthionine à partir de SAM. Dans le cas d'une production de méthionine à partir de SAM, le coproporphyrinogène (CP) est indiqué comme un des précurseurs possibles. Comme le porphobilinogène, annoté comme précurseur potentiel de la glycine, le CP est un des intermédiaires dans la synthèse de l'hème. Or Panek & O'Brian (2002) ont déjà envisagé l'approvisionnement de *Buchnera* en hème ou en l'un de ses intermédiaires.

La reconstruction des voies à partir des résultats de PITUFO met en lumière certains métabolites clés dans la synthèse des acides aminés. Nous avons déjà parlé de l'ammoniac dans le cas de la synthèse de la glycine. L'ATP, indiqué ici comme précurseur et source de carbone, apparaît aussi comme central. Sa place apparaît comme essentielle dans la synthèse de la méthionine, de la cystéine, de l'alanine, de l'histidine et potentiellement de la glycine et des acides aminés qui en sont issus.

En résumé, l'image globale de la synthèse des acides aminés chez *Buchnera aphidicola APS* reconstruite à partir des résultats de PITUFO nous confirme l'interdépendance de certains acides aminés mais nous indique aussi quelques métabolites clés dans leur synthèse, comme l'ATP ou l'ammoniac. Ensuite, certains précurseurs potentiels, comme le porphobilinogène ou le coproporphyrinogène seraient intéressants à tester expérimentalement. Par ailleurs, nous pouvons remarquer que la synthèse des acides aminés chez *Buchnera aphidicola APS* pourrait emprunter d'autres voies, traditionellement réservées à la synthèse d'autres métabolites. C'est le cas par exemple ici avec les voies de synthèse des nucléotides et des intermédiaires de l'hème. Chez les bactéries libres comme *Escherichia coli K12*, ces voies sont normalement distinctes et régulées de manière différente. L'altération de la régulation chez *Buchnera aphidicola APS* permettrait une utilisation plus large des voies métaboliques traditionelles (Zientz *et al.*, 2004).

Enfin, cette étude approfondie nous a permis de soulever certains paradoxes entre les résultats trouvés par cette approche, et les conclusions émises après certaines analyses nutritionelles chez le puceron. C'est le cas notamment de l'approvisionnement en glutamine et en asparagine. Ces contradictions entre les deux approches peuvent s'expliquer par deux raisons. La première pourrait être un manque de finesse dans l'annotation génomique et la reconstruction métabolique.

Certains traits propres à *Buchnera aphidicola APS* auraient pu être négligés. On peut imaginer par exemple que certaines enzymes ont perdu de leur spécificité chez *Buchnera aphidicola APS* et sont ainsi capables de catalyser des réactions indiquées comme absentes ici. L'annotation fonctionnelle en cours du puceron et d'autres expérimentations pourront compléter, confirmer ou infirmer certaines de ces hypothèses. Dans les deux cas, notre méthode prouve son utilité, autant pour enrichir les connaissances sur le métabolisme d'un organisme que pour formuler et tester de nouvelles hypothèses.

Conclusions générales et perspectives

Nous avons dans cette thèse abordé l'analyse systémique du métabolisme des endocytobiotes à travers trois pans de la bio-informatique que sont la représentation des données, la bio-analyse et le développement d'algorithmes et d'outils.

Avec SymbioCyc, nous réunissons dans une même interface les outils d'analyse et ceux destinés à la modélisation. Les fonctionnalités de SymbioCyc que sont l'exploration, la comparaison et le traitement des données avant modélisation ont montré leur utilité tout le long de ce projet.
Nous envisageons déjà certaines améliorations à SymbioCyc. La première serait d'automatiser la mise à jour et la synchronisation avec les données MaGe. Ces tâches sont pour le moment effectuées manuellement, ce qui limite le nombre d'organismes présents dans la base.

Par ailleurs, d'autres opérations sur les réseaux métaboliques sont envisageables. Il serait intéressant de pouvoir créer et sauvegarder l'intersection ou l'union de plusieurs réseaux métaboliques. Cette dernière opération serait particulièrement utile dans le cas de l'étude de systèmes symbiotiques complets.

Actuellement, tous les résultats des filtres sont précalculés. Afin de pouvoir les paramétrer plus finement, il faudrait générer les résultats des filtres à la demande. L'utilisateur pourrait ainsi choisir précisément les couples de cofacteurs qu'il voudrait éliminer de son jeu de données.

Enfin, une dernière amélioration serait de pouvoir utiliser les fonctionnalités de SymbioCyc sur d'autres données, propres à l'utilisateur. Nous projetons ainsi de rendre disponibles, à travers une interface utilisateur, toutes les fonctionnalités de parseBioNet, la librairie Java que nous avons développée et qui a permis de construire SymbioCyc.

Avec PITUFO, nous proposons la première définition de précurseurs prenant en compte de manière explicite les cycles. Nous présentons un algorithme permettant de trouver tous les ensembles de précurseurs d'un ensemble de cibles donné. De plus, la modélisation sous la forme d'un hypergraphe prend en compte la structure particulière du réseau métabolique. Outre son utilité pour indiquer les nutriments nécessaires à certaines fonctions métaboliques, on peut employer PITUFO pour la recherche d'incohérences dans les bases de données ou en amont d'une analyse de balance des flux.

Certains perfectionnements de PITUFO sont prévus très prochainement. Le premier est la diminution du temps de calcul. Ensuite, l'analyse des résultats serait grandement facilitée s'il était possible de visualiser de manière automatique les chemins qui vont des précurseurs aux composés cibles. Lorsque plusieurs chemins sont possibles entre les ensembles de précurseurs et les composés cibles, l'indication des composés "carrefours" serait d'une grande aide dans l'interprétation des résultats. On peut imaginer aussi de classer les solutions selon certains critères. Le poids moléculaire des précurseurs pourrait ainsi être pris en compte. La consommation d'énergie (sous la forme de nombre de molécules d'ATP consommées par exemple) dans les voies conduisant des précurseurs aux composés cibles serait une autre manière de classer les résultats. Une interface graphique permet-

trait enfin une diffusion plus large de cet outil.

L'analyse des graphes de composés et les résultats de PITUFO nous montrent que la seule topologie du graphe peut nous renseigner de façon précise sur le fonctionnement du réseau métabolique. Une analyse des flux pourrait certainement nous donner une image plus précise mais requiert une stœchiométrie exacte et des informations souvent absentes des données métaboliques, d'autant plus chez les bactéries endocytobiotes qui sont incultivables. La modélisation sous forme de graphes permet ainsi une analyse rapide et relativement riche de l'organisation des réseaux métaboliques.

Nous avons insisté par ailleurs sur l'importance du traitement des données avant la modélisation sous la forme d'un graphe simple pour éviter les artefacts dus à la nature même du réseau métabolique. La présentation de PITUFO souligne également la nécessité de développer de nouvelles méthodes, spécialement conçues pour l'analyse du métabolisme, en considérant les spécificités du réseau métabolique.

Dans le même ordre d'idée, le développement de méthodes d'alignement de réseaux métaboliques pourrait considérablement enrichir les analyses comparatives comme celle que nous avons effectuée. Au-delà des difficultés méthodologiques, le développement de telles méthodes recquiert un moyen efficace de comparer les réactions entre elles et les composés entre eux pour donner de la flexibilité dans les ressemblances. Ces aspects sont déjà la source de nombreuses réflexions au sein de l'équipe BAOBAB.

Notre étude a porté sur des métabolismes particuliers. En effet, de précédents travaux ont révélé chez les endocytobiotes un métabolisme réduit dont les fonctionnalités dépendent des interactions entretenues avec l'hôte. Nous avons tenté de préciser ces traits selon deux approches différentes, chacune se basant sur une analyse systémique du métabolisme. La première approche repose sur l'analyse comparative de réseaux métaboliques provenant de SymbioCyc. Son objectif est de dégager des propriétés communes ou distinctes des métabolismes des bactéries endocytobiotes, en relation avec le mode de vie intracellulaire et la nature mutualiste ou parasite du symbiote. La seconde approche se focalise sur les relations métaboliques entre les deux partenaires d'une symbiose particulière, celle de la bactérie mutualiste *Buchnera aphidicola APS* et de son hôte, le puceron du pois. Dans cette association, la bactérie approvisionne son hôte en certains acides aminés. La production de ces derniers chez la bactérie nécessite cependant des nutriments apportés par le puceron. Grâce à PITUFO, nous avons prédit les ensembles de métabolites potentiellement fournis par l'insecte et qui rendent possible la production des acides aminées chez la bactérie.

L'analyse comparative des réseaux métaboliques met en évidence la conservation d'une portion considérablement importante du métabolisme parmi les bactéries extracellulaires. En revanche, nous n'avons trouvé aucune réaction commune entre les réseaux métaboliques des 18 bactéries intracellulaires de notre jeu de

données. Un “coeur métabolique” minimal n'existe donc pas chez les endocytobiotes. Ce résultat indique également une réduction du métabolisme qui opère différemment selon les endocytobiotes.

Nos analyses suggèrent que la réduction du réseau métabolique s'accompagne d'une conservation plus importante des enzymes à large spectre. Par ailleurs, le rapport, plus élevé chez les bactéries intracellulaires, entre nombre de métabolites et de réactions, renforce l'idée d'une disparition des voies métaboliques redondantes lors de la réduction du réseau métabolique. Nous avons également montré que les génomes des bactéries mutualistes impliquées dans une symbiose nutritionnelle présentent une proportion plus grande du génome dédiée au métabolisme.

Nous avons montré que la réduction du réseau touchait, selon les bactéries, différentes classes de métabolites. Certaines, comme les lipides, ont même totalement disparu de certains réseaux métaboliques.

Des exemples précis émergeant de notre analyse illustrent la répercussion de la disparition d'une partie du réseau métabolique sur les pressions de sélection s'éxerçant sur d'autres régions du réseau.

Par ailleurs, la très faible intersection entre les réseaux métaboliques de deux bactéries mutualistes associées au même hôte souligne l'importance que peut avoir la sélection sur un système à plusieurs symbiotes.

L'analyse des graphes de composés nous a permis d'affiner l'influence de la réduction du réseau sur son organisation. Celle-ci exhibe des caractéristiques très conservées parmi les bactéries extracellulaires. Ainsi, chez les bactéries extracellulaires, la portion du réseau métabolique conservée l'est non seulement par sa composition mais aussi par son organisation.

En revanche, la topologie du réseau métabolique montre un profil très différent parmi les bactéries intracellulaires. Les mesures de connectivité et de centralité indiquent comment la réduction du réseau peut accentuer, ou au contraire diminuer l'importance d'un groupe de métabolites dans le fonctionnement du réseau. Ainsi, selon la bactérie endocytobiote considérée, le métabolisme se structure autour de régions différentes.

L'analyse comparative nous a permis également de mettre le doigt sur la difficulté à définir certains concepts plus larges. Le premier est le mode de vie. Les notions mêmes de parasitisme et de mutualisme restent dans certains cas difficiles à définir. Le second est la notion d'invidu qui atteint ses limites quand on étudie de tels systèmes.

La comparaison des réseaux que nous avons effectuée pourrait être améliorée sur certains points. Nous pourrions augmenter le nombre de bactéries dans notre jeu de données, pour avoir une étendue phylogénétique plus grande et des modes de vies plus diversifiés. Cependant, cet élargissement empêcherait certainement une analyse fine des points communs et des singularités, comme nous avons pu la faire ici.

Certaines tendances générales mériteraient d'être précisées. Ainsi, il serait intéressant de préciser quelles sont les enzymes retenues préférentiellement au cours de la réduction du réseau métabolique. A quel type d'activité catalytique correspondent-elles ? Leur conservation dépend-elle de leur composition en domaines protéiques ? Une analyse des numéros EC et une décomposition en domaines protéiques permettrait de répondre à ces questions.

Grâce à PITUFO et aux données présentes dans SymbioCyc, nous avons pu effectuer une étude très précise du métabolisme des acides aminés chez *Buchnera aphidicola APS* et de ses liens avec le métabolisme de son hôte, le puceron. Certains de nos résultats concordent avec de précédents résultats expérimentaux mais d'autres lèvent certaines contradictions, utiles pour formuler de nouvelles hypothèses. Nous avons ainsi indiqué certains acides aminés, comme la glutamine ou l'asparagine, comme obligatoirement fournis par le puceron.

Les résultats sont déjà mis à profit pour accompagner l'annotation génomique et la reconstruction métabolique du puceron, auxquelles participent les laboratoires BF2I et BAOBAB. La connaissance du réseau métabolique de l'hôte pourra aussi confirmer, ou infirmer, certaines de nos hypothèses. Elle permettra également de considérer le métabolisme du système symbiotique dans son ensemble.

D'autres réseaux métaboliques d'organismes porteurs d'endocytobiotes seront ainsi bientôt disponibles. Leur analyse nous permettra d'affiner les connaissances sur le métabolisme des systèmes symbiotiques et sur leur évolution.

Références bibliographiques

Références bibliographiques

Aguilar D., Aviles F. X., Querol E. & Sternberg M. J. E. (2004). Analysis of phenetic trees based on metabolic capabilites across the three domains of life. *J Mol Biol*, **340**(3), 491–512. *Cité page 85.*

Akman L., Yamashita A., Watanabe H., Oshima K., Shiba T., Hattori M. & Aksoy S. (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. *Nat Genet*, **32**(3), 402–407. *Cité page 28.*

Al-Roubi A. S. (1982). Ibn al-nafis as a philosopher. In *Symposium on Ibn al-Nafis, Second International Conference on Islamic Medicine : Islamic Medical Organization, Kuwait. Cité page 9.*

Aloy P. & Russell R. B. (2002). Potential artefacts in protein-interaction networks. *FEBS Lett*, **530**(1-3), 253–254. *Cité page 87.*

Alsmark C. M., Frank A. C., Karlberg E. O., Legault B.-A., Ardell D. H., Canbäck B., Eriksson A.-S., Näslund A. K., Handley S. A., Huvet M., Scola B. L., Holmberg M. & Andersson S. G. E. (2004). The louse-borne human pathogen bartonella quintana is a genomic derivative of the zoonotic agent bartonella henselae. *Proc Natl Acad Sci U S A*, **101**(26), 9716–9721. *Cité page 114.*

Altschul S. F., Gish W., Miller W., Myers E. W. & Lipman D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–410. *Cité page 33.*

Andersson S. G. E., Karlberg O., Canbäck B. & Kurland C. G. (2003). On the origin of mitochondria : a genomics perspective. *Philos Trans R Soc Lond B Biol Sci*, **358**(1429), 165–77 ; discussion 177–9. *Cité page 23.*

Aoki K. F. & Kanehisa M. (2005). Using the kegg database resource. *Curr Protoc Bioinformatics*, **Chapter 1**, Unit 1.12. *Cité page 81.*

Arita M. (2000). Metabolic reconstruction using shortest paths. *Simulation Practice and Theory*, **9**, 109–125. *Cité pages 43 et 86.*

Arita M. (2004). The metabolic world of Escherichia coli is not small. *Proc Natl Acad Sci U S A*, **101**(6), 1543–1547. *Cité page 86.*

Bairoch A. (2000). The ENZYME database in 2000. *Nucleic Acids Res*, **28**(1), 304–305. *Cité page 35.*

Barabási A. L. & Albert R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512. *Cité page 87.*

Baumann P., Baumann L., Lai C. Y., Rouhbakhsh D., Moran N. A. & Clark M. A. (1995). Genetics, physiology, and evolutionary relationships of the genus Buchnera : intracellular symbionts of aphids. *Annu Rev Microbiol*, **49**, 55–94. *Cité page 25.*

Birtles R. J. (2005). Bartonellae as elegant hemotropic parasites. *Ann N Y Acad Sci*, **1063**, 270–279. *Cité pages 67, 113 et 114.*

Bocs S., Cruveiller S., Vallenet D., Nuel G. & Médigue C. (2003). Amigene : Annotation of microbial genes. *Nucleic Acids Res*, **31**(13), 3723–3726. *Cité page 69.*

Borodina I., Krabben P. & Nielsen J. (2005). Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res*, **15**(6), 820–829. *Cité page 44.*

Bourqui R., Cottret L., Lacroix V., Auber D., Mary P., Sagot M.-F. & Jourdan F. (2007). Metabolic network visualization eliminating node redundance and preserving metabolic pathways. *BMC Syst Biol*, **1**, 29. *Cité pages 59, 74, 77 et 81.*

Boutet E., Lieberherr D., Tognolli M., Schneider M. & Bairoch A. (2007). Uniprotkb/swiss-prot : The manually annotated section of the uniprot knowledgebase. *Methods Mol Biol*, **406**, 89–112. *Cité page 40.*

Breitling R., Ritchie S., Goodenowe D., Stewart M. L. & Barrett M. P. (2006). Ab initio prediction of metabolic networks using fourier transform mass spectrometry data. *Metabolomics*, **2**(3), 155–164. *Cité page 43.*

Buchner P. (1965). *Endosymbiosis of animals with plant microorganisms*. John Wiley & Sons, Inc., New York, N.Y. *Cité page 24.*

Caetano-Anollés G., Kim H. S. & Mittenthal J. E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci U S A*, **104**(22), 9358–9363. *Cité page 14.*

Caetano-Anollés G., Yafremava L. S., Gee H., Caetano-Anollés D., Kim H. S. & Mittenthal J. E. (2008). The origin and evolution of modern metabolism. *Int J Biochem Cell Biol. Cité pages 21 et 22.*

Cases I., de Lorenzo V. & Ouzounis C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol*, **11**(6), 248–253. *Cité pages 83 et 84.*

Caspi R., Foerster H., Fulcher C. A., Kaipa P., Krummenacker M., Latendresse M., Paley S., Rhee S. Y., Shearer A. G., Tissier C., Walk T. C., Zhang P. & Karp P. D. (2008). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, **36**(Database issue), D623–D631. *Cité pages 32, 40, 54 et 56.*

Caspi R. & Karp P. D. (2007). Using the metacyc pathway database and the biocyc database collection. *Curr Protoc Bioinformatics*, **Chapter 1**, Unit1.17. *Cité pages 40, 74 et 81.*

Chen L. & Vitkup D. (2007). Distribution of orphan metabolic activities. *Trends Biotechnol*, **25**(8), 343–348. *Cité page 37.*

Chu K. H., Qi J., Yu Z.-G. & Anh V. (2004). Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol Biol Evol*, **21**(1), 200–206. *Cité page 23.*

Chua H. N., Sung W.-K. & Wong L. (2007). An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, **23**(24), 3364–3373. *Cité page 36.*

Claudel-Renard C., Chevalet C., Faraut T. & Kahn D. (2003). Enzyme-specific profiles for genome annotation : PRIAM. *Nucleic Acids Res*, **31**(22), 6633–6639. *Cité pages 35 et 69.*

Cordón F. (1990). *Tratado Evolucionista de Biologia.* Aguilar, Madrid, Spain. *Cité page 21.*

Cordwell S. J. (1999). Microbial genomes and "missing" enzymes : redefining biochemical pathways. *Arch Microbiol*, **172**(5), 269–279. *Cité page 42.*

Cottret L., Milreu P. V., Acuña V., Marchetti-Spaccamela A., Martinez F. V., Sagot M.-F. & Stougie. L. (2008). Enumerating precursor sets of target metabolites in a metabolic network. In *Proceedings of the 8th Workshop on Algorithms in Bioinformatics (WABI '08), Springer-Verlag Berlin Heidelberg, Lecture Notes in Computer Science. Cité page 145.*

Coulomb S., Bauer M., Bernard D. & Marsolier-Kergoat M.-C. (2005). Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci*, **272**(1573), 1721–1725. *Cité page 87.*

Covert M. W., Knight E. M., Reed J. L., Herrgard M. J. & Palsson B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**(6987), 92–96. *Cité page 44.*

Croes D., Couche F., Wodak S. J. & van Helden J. (2006). Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol*, **356**(1), 222–236. *Cité page 87.*

Csardi G. & Nepusz T. (2006). The igraph software for complex network research. *InterJournal, Complex Systems*, **1695**. *Cité page 91.*

d'Aubenton Carafa Y., Brody E. & Thermes C. (1990). Prediction of rho-independent escherichia coli transcription terminators. a statistical analysis of their rna stem-loop structures. *J Mol Biol*, **216**(4), 835–858. *Cité page 69.*

Degnan P. H., Lazarus A. B. & Wernegreen J. J. (2005). Genome sequence of blochmannia pennsylvanicus indicates parallel evolutionary trends among bacterial mutualists of insects. *Genome Res*, **15**(8), 1023–1033. *Cité pages 28 et 121.*

Douglas A. E. (1988). Sulphate utilization in an aphid symbiosis. *Insect Biochem.*, **18**, 599–605. *Cité page 180.*

Douglas A. E. (1998). Nutritional interactions in insect-microbial symbioses : aphids and their symbiotic bacteria Buchnera. *Annu Rev Entomol*, **43**, 17–37. *Cité pages 109, 156 et 174.*

Dray S. & Dufour A.-B. (2007). The ade4 package : Implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**(4). *Cité page 91.*

Duarte N. C., Becker S. A., Jamshidi N., Thiele I., Mo M. L., Vo T. D., Srivas R. & Palsson B. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, **104**(6), 1777–1782. *Cité pages 44 et 62.*

Dufayard J.-F., Duret L., Penel S., Gouy M., Rechenmann F. & Perrière G. (2005). Tree pattern matching in phylogenetic trees : automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**(11), 2596–2603. *Cité page 34.*

Duret L., Mouchiroud D. & Gouy M. (1994). HOVERGEN : a database of homologous vertebrate genes. *Nucleic Acids Res*, **22**(12), 2360–2365. *Cité page 34.*

Edwards J. S. & Palsson B. O. (2000). Robustness analysis of the Escherichia coli metabolic network. *Biotechnol Prog*, **16**(6), 927–939. *Cité page 47.*

Eknoyan G. (1999). Santorio sanctorius (1561-1636) - founding father of metabolic balance studies. *Am J Nephrol*, **19**(2), 226–233. *Cité pages 9 et 10.*

Enright A. J. & Ouzounis C. A. (2001). Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*, **2**(9), RESEARCH0034. *Cité page 36.*

Fares M. A., Moya A. & Barrio E. (2004). GroEL and the maintenance of bacterial endosymbiosis. *Trends Genet*, **20**(9), 413–6. *Cité page 28.*

Febvay G., Liadouze I., Guillaud J. & Bonnot G. (1995). Analysis of energetic amino acid metabolism in acyrthosiphon pisum : A multidimensional approach to amino acid metabolism in aphids. *Archives of insect biochemistry and physiology*, **29**, 45–69. *Cité page 169.*

Feist A. M., Henry C. S., Reed J. L., Krummenacker M., Joyce A. R., Karp P. D., Broadbelt L. J., Hatzimanikatis V. & Palsson B. (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. *Cité pages 43 et 62.*

Fell D. A. (1992). Metabolic control analysis : a survey of its theoretical and experimental development. *Biochem J*, **286 (Pt 2)**, 313–330. *Cité page 45.*

Fell D. A. & Wagner A. (2000). The small world of metabolism. *Nat Biotechnol*, **18**(11), 1121–1122. *Cité page 86.*

Fenn K., Conlon C., Jones M., Quail M. A., Holroyd N. E., Parkhill J. & Blaxter M. (2006). Phylogenetic relationships of the wolbachia of nematodes and arthropods. *PLoS Pathog*, **2**(10), e94. *Cité page 94.*

Finney A H. M. (2003). Systems biology markup language : Level 2 and beyond. *Biochem Soc Trans*, p. 1472–3. *Cité page 62.*

Foster J., Ganatra M., Kamal I., Ware J., Makarova K., Ivanova N., Bhattacharyya A., Kapatral V., Kumar S., Posfai J., Vincze T., Ingram J., Moran L., Lapidus A., Omelchenko M., Kyrpides N., Ghedin E., Wang S., Goltsman E., Joukov V., Ostrovskaya O., Tsukerman K., Mazur M., Comb D., Koonin E. & Slatko B. (2005). The wolbachia genome of brugia malayi : endosymbiont evolution within a human pathogenic nematode. *PLoS Biol*, **3**(4), e121. *Cité pages 29, 66, 94, 111, 112 et 122.*

Freilich S., Spriggs R. V., George R. A., Al-Lazikani B., Swindells M. & Thornton J. M. (2005). The complement of enzymatic sets in different species. *J Mol Biol*, **349**(4), 745–763. *Cité page 85.*

Friedberg I. (2006). Automated protein function prediction–the genomic challenge. *Brief Bioinform*, **7**(3), 225–242. *Cité pages 32, 35 et 36.*

Galperin M. Y., Walker D. R. & Koonin E. V. (1998). Analogous enzymes : independent inventions in enzyme evolution. *Genome Res*, **8**(8), 779–790. *Cité page 35.*

Garfinkel D. (1968). The role of computer simulation in biochemistry. *Comput Biomed Res*, **2**(1), i–ii. *Cité page 44.*

Gerlt J. A. (2003). How to find "missing" genes. *Chem Biol*, **10**(12), 1141–1142. *Cité page 42.*

Gerlt J. A. & Babbitt P. C. (2000). Can sequence determine function ? *Genome Biol*, **1**(5), REVIEWS0005. *Cité page 37.*

Gil R., Latorre A. & Moya A. (2004a). Bacterial endosymbionts of insects : insights from comparative genomics. *Environ Microbiol*, **6**(11), 1109–22. *Cité pages 24 et 27.*

Gil R., Silva F. J., Peretó J. & Moya A. (2004b). Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev*, **68**(3), 518–537. *Cité page 158.*

Gil R., Silva F. J., Zientz E., Delmotte F., González-Candelas F., Latorre A., Rausell C., Kamerbeek J., Gadau J., Hölldobler B., van Ham R. C. H. J., Gross R. & Moya A. (2003). The genome sequence of blochmannia floridanus : comparative analysis of reduced genomes. *Proc Natl Acad Sci U S A*, **100**(16), 9388–9393. *Cité page 28.*

Gray M. W., Burger G. & Lang B. F. (1999). Mitochondrial evolution. *Science*, **283**(5407), 1476–1481. *Cité page 29.*

Green M. L. & Karp P. D. (2004). A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76. *Cité pages 36 et 42.*

Griffiths-Jones S., Moxon S., Marshall M., Khanna A., Eddy S. R. & Bateman A. (2005). Rfam : annotating non-coding rnas in complete genomes. *Nucleic Acids Res*, **33**(Database issue), D121–D124. *Cité page 69.*

Handorf T., Christian N., Ebenhöh O. & Kahn D. (2007). An environmental perspective on metabolism. *J Theor Biol. Cité pages 75 et 147.*

Handorf T., Ebenhöh O. & Heinrich R. (2005). Expanding metabolic networks : scopes of compounds, robustness, and evolution. *J Mol Evol*, **61**(4), 498–512. *Cité page 146.*

Horowitz N. H. (1945). On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci U S A*, **31**(6), 153–157. *Cité page 21.*

Horowitz N. H. (1990). George wells beadle (1903-1989). *Genetics*, **124**(1), 1–6. *Cité page 11.*

Hubbard T. J. P., Aken B. L., Beal K., Ballester B., Caccamo M., Chen Y., Clarke L., Coates G., Cunningham F., Cutts T., Down T., Dyer S. C., Fitzgerald S., Fernandez-Banet J., Graf S., Haider S., Hammond M., Herrero J., Holland R., Howe K., Howe K., Johnson N., Kahari A., Keefe D., Kokocinski F., Kulesha E., Lawson D., Longden I., Melsopp C., Megy K., Meidl P., Ouverdin B., Parker A., Prlic A., Rice S., Rios D., Schuster M., Sealy I., Severin J., Slater G., Smedley D., Spudich G., Trevanion S., Vilella A., Vogel J., White S., Wood M., Cox T., Curwen V., Durbin R., Fernandez-Suarez X. M., Flicek P., Kasprzyk A., Proctor G., Searle S., Smith J., Ureta-Vidal A. & Birney E. (2007). Ensembl 2007. *Nucleic Acids Res*, **35**(Database issue), D610–D617. *Cité page 37.*

Hurst G. D. & Jiggins F. M. (2000). Male-killing bacteria in insects : mechanisms, incidence, and implications. *Emerg Infect Dis*, **6**(4), 329–336. *Cité pages 26 et 67.*

Ishikawa H. (2003). *Insect Symbiosis : an introduction*. CRC Press. *Cité page 26.*

Jaenicke L. (2007). Centenary of the award of a nobel prize to eduard buchner, the father of biochemistry in a test tube and thus of experimental molecular bioscience. *Angew Chem Int Ed Engl*, **46**(36), 6776–6782. *Cité page 10.*

Jensen R. A. (1976). Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, **30**, 409–425. *Cité page 21.*

Jeong H., Tombor B., Albert R., Oltvai Z. N. & Barabási A. L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**(6804), 651–654. *Cité pages 52 et 87.*

Kanehisa M., Araki M., Goto S., Hattori M., Hirakawa M., Itoh M., Katayama T., Kawashima S., Okuda S., Tokimatsu T. & Yamanishi Y. (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, **36**(Database issue), D480–D484. *Cité pages 33, 40 et 54.*

Karp P. D., Keseler I. M., Shearer A., Latendresse M., Krummenacker M., Paley S. M., Paulsen I., Collado-Vides J., Gama-Castro S., Peralta-Gil M., Santos-Zavaleta A., Peñaloza-Spínola M. I., Bonavides-Martinez C. & Ingraham J. (2007). Multidimensional annotation of the escherichia coli k-12 genome. *Nucleic Acids Res*, **35**(22), 7577–7590. *Cité page 56.*

Karp P. D., Paley S. & Romero P. (2002). The Pathway Tools software. *Bioinformatics*, **18 Suppl1**, 225–238. *Cité pages 40 et 56.*

Keller E. F. (2005). Revisiting "scale-free" networks. *Bioessays*, **27**(10), 1060–1068. *Cité page 88.*

Kersey P., Bower L., Morris L., Horne A., Petryszak R., Kanz C., Kanapin A., Das U., Michoud K., Phan I., Gattiker A., Kulikova T., Faruque N., Duggan K., Mclaren P., Reimholz B., Duret L., Penel S., Reuter I. & Apweiler R. (2005). Integr8 and genome reviews : integrated views of complete genomes and proteomes. *Nucleic Acids Res*, **33**(Database issue), D297–D302. *Cité page 37.*

Keseler I. M., Collado-Vides J., Gama-Castro S., Ingraham J., Paley S., Paulsen I. T., Peralta-Gil M. & Karp P. D. (2005). Ecocyc : a comprehensive database resource for escherichia coli. *Nucleic Acids Res*, **33**(Database issue), D334–D337. *Cité pages 31 et 146.*

Khanin R. & Wit E. (2006). How scale-free are biological networks. *J Comput Biol*, **13**(3), 810–818. *Cité page 87.*

Kharchenko P., Chen L., Freund Y., Vitkup D. & Church G. M. (2006). Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, **7**, 177. *Cité page 42.*

Kinne-Saffran E. & Kinne R. K. (1999). Vitalism and synthesis of urea. from friedrich wöhler to hans a. krebs. *Am J Nephrol*, **19**(2), 290–294. *Cité page 10.*

Klamt S. & Gilles E. D. (2004). Minimal cut sets in biochemical reaction networks. *Bioinformatics*, **20**(2), 226–234. *Cité page 47.*

Koonin E. V. (2000). How many genes can make a cell : the minimal-gene-set concept. *Annu Rev Genomics Hum Genet*, **1**, 99–116. *Cité page 141.*

Kornberg H. (2000). Krebs and his trinity of cycles. *Nat Rev Mol Cell Biol*, **1**(3), 225–228. *Cité page 11.*

Kotera M., Okuno Y., Hattori M., Goto S. & Kanehisa M. (2004). Computational assignment of the ec numbers for genomic-scale analysis of enzymatic reactions. *J Am Chem Soc*, **126**(50), 16487–16498. *Cité pages 43 et 54.*

Kriventseva E. V., Rahman N., Espinosa O. & Zdobnov E. M. (2007). OrthoDB : the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* *Cité page 34.*

Krummenacker M., Paley S., Mueller L., Yan T. & Karp P. D. (2005). Querying and computing with biocyc databases. *Bioinformatics*, **21**(16), 3454–3455. *Cité page 57.*

Kun A., Papp B. & Szathmáry E. (2008). Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biol*, **9**(3), R51. *Cité page 148.*

Kümmel A., Panke S. & Heinemann M. (2006). Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*, **7**, 512. *Cité page 43.*

Lacroix V., Cottret L., Rogier O., Gomes-Fernandes C., Jourdan F. & Sagot M. (2008a). Motus : a software and a webserver for the search and enumeration of node-labelled connected subgraphs in biological networks. *Bioinformatics*, **Soumis.** *Cité page 81.*

Lacroix V., Cottret L., Thébault P. & Sagot M.-F. (2008b). An introduction to metabolic networks and their structural analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **99**(1). *Cité page 53.*

Lacroix V., Fernandes C. G. & Sagot M.-F. (2006). Motif search in graphs : application to metabolic networks. *IEEE/ACM Trans Comput Biol Bioinform*, **3**(4), 360–368. *Cité pages 52, 74 et 143.*

Lange B. M., Rujan T., Martin W. & Croteau R. (2000). Isoprenoid biosynthesis : the evolution of two ancient and distinct pathways across genomes. *Proc Natl Acad Sci U S A*, **97**(24), 13172–13177. *Cité page 116.*

Lazcano A. & Miller S. L. (1999). On the origin of metabolic pathways. *J Mol Evol*, **49**(4), 424–431. *Cité page 14.*

Lespinet O. & Labedan B. (2006). Puzzling over orphan enzymes. *Cell Mol Life Sci*, **63**(5), 517–523. *Cité page 37.*

Li H., Coghlan A., Ruan J., Coin L. J., Hériché J.-K., Osmotherly L., Li R., Liu T., Zhang Z., Bolund L., Wong G. K.-S., Zheng W., Dehal P., Wang J. & Durbin R. (2006). Treefam : a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, **34**(Database issue), D572–D580. *Cité page 34.*

Liadouze I., Febvay G., Guillaud J. & Bonnot G. (1995). Effect of diet on the free amino acid pools of symbiotic and aposymbiotic pea aphids, acyrthosiphon pisum. *Journal of Insect Physiology*, **41 (1)**, 33–40. *Cité pages 178 et 179.*

Liadouze I., Febvay G., Guillaud J. & Bonnot G. (1996). Metabolic fate of energetic amino acids in the aposymbiotic pea aphid acyrthosiphon pisum (harris) (homoptera : Aphididae). *Symbiosis*, **21**, 115–127. *Cité pages 164, 165, 168, 171, 172 et 180.*

Light S., Kraulis P. & Elofsson A. (2005). Preferential attachment in the evolution of metabolic networks. *BMC Genomics*, **6**, 159. *Cité page 52.*

Lima T., Auchincloss A. H., Coudert E., Keller G., Michoud K., Rivoire C., Bulliard V., de Castro E., Lachaize C., Baratin D., Phan I., Bougueleret L. & Bairoch A. (2008). Hamap : a database of completely sequenced microbial proteome sets and manually curated microbial protein families in uniprotkb/swiss-prot. *Nucleic Acids Res. Cité page 37.*

Liu W., Lin W., Davis A., Jordán F., Yang H. & Hwang M. (2007). A network perspective on the topological importance of enzymes and their phylogenetic conservation. *BMC Bioinformatics*, **8**, 121. *Cité page 88.*

Lowe T. M. & Eddy S. R. (1997). trnascan-se : a program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Res*, **25**(5), 955–964. *Cité page 69.*

Ma H. & Zeng A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, **19**(2), 270–277. *Cité pages 52, 88, 126 et 127.*

Maden B. E. (1995). No soup for starters ? Autotrophy and the origins of metabolism. *Trends Biochem Sci*, **20**(9), 337–341. *Cité page 14.*

Maezawa K., Shigenobu S., Taniguchi H., Kubo T., Aizawa S.-I. & Morioka M. (2006). Hundreds of flagellar basal bodies cover the cell surface of the endosymbiotic bacterium buchnera aphidicola sp. strain aps. *J Bacteriol*, **188**(18), 6539–6543. *Cité page 28.*

Manchester K. L. (1995). Louis pasteur (1822-1895)–chance and the prepared mind. *Trends Biotechnol*, **13**(12), 511–515. *Cité page 10.*

Margulis L. (1981). *Symbiosis in cell evolution.* Freeman, San Francisco. *Cité page 29.*

McCutcheon J. P. & Moran N. A. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proc Natl Acad Sci U S A*, **104**(49), 19392–19397. *Cité pages 30, 99, 109, 119, 120 et 123.*

Meyer F., Goesmann A., McHardy A. C., Bartels D., Bekel T., Clausen J., Kalinowski J., Linke B., Rupp O., Giegerich R. & Pühler A. (2003). Gendb–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*, **31**(8), 2187–2195. *Cité page 36.*

Mitchell A. & Finch L. R. (1977). Pathways of nucleotide biosynthesis in mycoplasma mycoides subsp. mycoides. *J Bacteriol*, **130**(3), 1047–1054. *Cité page 114.*

Moran N. A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A*, **93**(7), 2873–8. *Cité page 28.*

Moran N. A. (2007). Colloquium Papers : Symbiosis as an adaptive process and source of phenotypic complexity. *Proc Natl Acad Sci U S A*, **104 Suppl 1**, 8627–8633. *Cité page 29.*

Moran N. A., Dale C., Dunbar H., Smith W. A. & Ochman H. (2003a). Intracellular symbionts of sharpshooters (insecta : Hemiptera : Cicadellinae) form a distinct clade with a small genome. *Environ Microbiol*, **5**(2), 116–126. *Cité page 24.*

Moran N. A., Dunbar H. E. & Wilcox J. L. (2005). Regulation of transcription in a reduced bacterial genome : nutrient-provisioning genes of the obligate symbiont buchnera aphidicola. *J Bacteriol*, **187**(12), 4229–4237. *Cité page 160.*

Moran N. A., Plague G. R., Sandström J. P. & Wilcox J. L. (2003b). A genomic perspective on nutrient provisioning by bacterial symbionts of insects. *Proc Natl Acad Sci U S A*, **100 Suppl 2**, 14543–8. *Cité pages 29 et 142.*

Moriya Y., Itoh M., Okuda S., Yoshizawa A. C. & Kanehisa M. (2007). Kaas : an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*, **35**(Web Server issue), W182–W185. *Cité page 40.*

Morowitz H. J. (1999). A theory of biochemical organization, metabolic pathways, and evolution. *Complexity*, **4**(6), 39–53. *Cité page 41.*

Morowitz H. J., Kostelnik J. D., Yang J. & Cody G. D. (2000). The origin of intermediary metabolism. *Proc Natl Acad Sci U S A*, **97**(14), 7704–7708. *Cité page 14.*

Mulder N. & Apweiler R. (2007). InterPro and InterProScan : Tools for Protein Sequence Classification and Comparison. *Methods Mol Biol*, **396**, 59–70. *Cité page 35.*

Mushegian A. R. & Koonin E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A*, **93**(19), 10268–10273. *Cité page 158.*

Nair S. (2004). *Marine Microbiology : Facets & Opportunities*, chapter Bacterial Associations : Antagonism to Symbiosis, p. 83–89. National Institute of Oceanography, Goa. *Cité page 25.*

Nakabachi A., Shigenobu S., Sakazume N., Shiraki T., Hayashizaki Y., Carninci P., Ishikawa H., Kudo T. & Fukatsu T. (2005). Transcriptome analysis of the aphid bacteriocyte, the symbiotic host cell that harbors an endocellular mutualistic bacterium, Buchnera. *Proc Natl Acad Sci U S A*, **102**(15), 5477–82. *Cité pages 176 et 177.*

Nardon P. & Charles H. (2004). *Morphological Aspects of Symbiosis. Cité pages 24 et 25.*

Nardon P. & Grenier A.-M. (1993). Symbiose et évolution. *Annales de la Société entomologique de France*, **29**(2), 113–140. *Cité pages 25, 26 et 141.*

Naumoff D. G., Xu Y., Glansdorff N. & Labedan B. (2004). Retrieving sequences of enzymes experimentally characterized but erroneously annotated : the case of the putrescine carbamoyltransferase. *BMC Genomics*, **5**(1), 52. *Cité page 35.*

Nelson D. L. & Cox M. M. (2004). *Lehninger Principles of Biochemistry.* W.H Freeman and Company - New York. *Cité pages 109 et 159.*

Nikoloski Z., Grimbs S., Selbig J. & Ebenhoh O. (2008). Hardness and approximability of the inverse scope problem. In *Proceedings of the 8th Workshop on Algorithms in Bioinformatics (WABI '08), Springer-Verlag Berlin Heidelberg, Lecture Notes in Computer Science. Cité page 178.*

Nobeli I. & Thornton J. M. (2006). A bioinformatician's view of the metabolome. *Bioessays*, **28**(5), 534–545. *Cité page 17.*

Norrby E. (2008). Nobel prizes and the emerging virus concept. *Archives of Virology*, **153**(6), 1109–1123. *Cité page 11.*

Northrop J. H. (1929). Crystalline pepsin. *Science*, **69**(1796), 580. *Cité page 11.*

Ochman H. & Moran N. A. (2001). Genes lost and genes found : evolution of bacterial pathogenesis and symbiosis. *Science*, **292**(5519), 1096–1099. *Cité pages 27 et 29.*

Oh M., Yamada T., Hattori M., Goto S. & Kanehisa M. (2007). Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J Chem Inf Model*, **47**(4), 1702–1712. *Cité page 54.*

Okuda S., Yamada T., Hamajima M., Itoh M., Katayama T., Bork P., Goto S. & Kanehisa M. (2008). Kegg atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, **36**(Web Server issue), W423–W426. *Cité page 59.*

Oliveira J. S., Bailey C. G., Jones-Oliveira J. B., Dixon D. A., Gull D. W. & Chandler M. L. (2003). A computational model for the identification of biochemical pathways in the krebs cycle. *J Comput Biol*, **10**(1), 57–82. *Cité page 48.*

Oliver K. M., Moran N. A. & Hunter M. S. (2005). Variation in resistance to parasitism in aphids is due to symbionts not host genotype. *Proc Natl Acad Sci U S A*, **102**(36), 12795–12800. *Cité page 26.*

Osterman A. & Overbeek R. (2003). Missing genes in metabolic pathways : a comparative genomics approach. *Curr Opin Chem Biol*, **7**(2), 238–251. *Cité page 42.*

Overbeek R., Begley T., Butler R. M., Choudhuri J. V., Chuang H.-Y., Cohoon M., de Crécy-Lagard V., Diaz N., Disz T., Edwards R., Fonstein M., Frank E. D., Gerdes S., Glass E. M., Goesmann A., Hanson A., Iwata-Reuyl D., Jensen R., Jamshidi N., Krause L., Kubal M., Larsen N., Linke B., McHardy A. C., Meyer F., Neuweger H., Olsen G., Olson R., Osterman A., Portnoy V., Pusch G. D., Rodionov D. A., Rückert C., Steiner J., Stevens R., Thiele I., Vassieva O., Ye Y., Zagnitko O. & Vonstein V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, **33**(17), 5691–5702. *Cité page 36.*

Paley S. M. & Karp P. D. (2002). Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics*, **18**(5), 715–724. *Cité page 41.*

Paley S. M. & Karp P. D. (2006). The pathway tools cellular overview diagram and omics viewer. *Nucleic Acids Res*, **34**(13), 3771–3778. *Cité page 59.*

Palsson B. O. (2000). The challenges of *in silico* biology. *Nat. Biotechnol*, **18**, 1147–1150. *Cité pages 45 et 46.*

Panek H. & O'Brian M. R. (2002). A whole genome view of prokaryotic haem biosynthesis. *Microbiology*, **148**(Pt 8), 2273–2282. *Cité pages 162, 172 et 181.*

Parter M., Kashtan N. & Alon U. (2007). Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol*, **7**, 169. *Cité pages 88 et 89.*

Patterson B. W. (1997). Use of stable isotopically labeled tracers for studies of metabolic kinetics : an overview. *Metabolism*, **46**(3), 322–329. *Cité page 17.*

Paulsen I. & von Haeseler A. (2006). INVHOGEN : a database of homologous invertebrate genes. *Nucleic Acids Res*, **34**(Database issue), D349–D353. *Cité page 34.*

Pellegrini M., Marcotte E. M., Thompson M. J., Eisenberg D. & Yeates T. O. (1999). Assigning protein functions by comparative genome analysis : protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**(8), 4285–4288. *Cité pages 36 et 38.*

Picard F., Miele V., Daudin J., Cottret L. & Robin S. (2008). Deciphering the connectivity structure of biological networks using mixnet. *BMC Bioinformatics*, **Soumis**. *Cité pages 81 et 143.*

Pouliot Y. & Karp P. D. (2007). A survey of orphan enzymatic activities. *BMC Bioinformatics*, **8**, 244. *Cité page 37.*

Prickett M. D., Page M., Douglas A. E. & Thomas G. H. (2006). Buchnerabase : a post-genomic resource for buchnera sp. aps. *Bioinformatics*, **22**(5), 641–642. *Cité page 157.*

Pruitt K. D., Tatusova T. & Maglott D. R. (2007). NCBI reference sequences (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**(Database issue), D61–D65. *Cité page 37.*

Pérez-Brocal V., Gil R., Ramos S., Lamelas A., Postigo M., Michelena J. M., Silva F. J., Moya A. & Latorre A. (2006). A small microbial genome : the end of a long symbiotic relationship ? *Science*, **314**(5797), 312–313. *Cité pages 24, 29, 99, 119 et 142.*

Rahman S. A., Advani P., Schunk R., Schrader R. & Schomburg D. (2005). Metabolic pathway analysis web service (pathway hunter tool at cubic). *Bioinformatics*, **21**(7), 1189–1193. *Cité pages 41 et 70.*

Rahman S. A. & Schomburg D. (2006). Observing local and global properties of metabolic pathways : 'load points' and 'choke points' in the metabolic networks. *Bioinformatics*, **22**(14), 1767–1774. *Cité page 88.*

Rasko D. A., Altherr M. R., Han C. S. & Ravel J. (2005). Genomics of the bacillus cereus group of organisms. *FEMS Microbiol Rev*, **29**(2), 303–329. *Cité page 68.*

Reed J. L., Vo T. D., Schilling C. H. & Palsson B. O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol*, **4**(9), R54. *Cité pages 62 et 75.*

Rison S. C. G., Teichmann S. A. & Thornton J. M. (2002). Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J Mol Biol*, **318**(3), 911–932. *Cité pages 21 et 22.*

Rogozin I. B., Makarova K. S., Murvai J., Czabarka E., Wolf Y. I., Tatusov R. L., Szekely L. A. & Koonin E. V. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*, **30**(10), 2212–2223. *Cité pages 36 et 39.*

Romero P. R. & Karp P. (2001). Nutrient-related analysis of pathway/genome databases. *Pac Symp Biocomput*, p. 471–482. *Cité pages 146 et 147.*

Sagan L. (1967). On the origin of mitosing cells. *J Theor Biol*, **14**(3), 255–274. *Cité page 23.*

Sakurai M., Koga R., Tsuchida T., Meng X.-Y. & Fukatsu T. (2005). Rickettsia symbiont in the pea aphid acyrthosiphon pisum : novel cellular tropism, effect on host fitness, and interaction with the essential symbiont buchnera. *Appl Environ Microbiol*, **71**(7), 4069–4075. *Cité page 66.*

Samant S., Lee H., Ghassemi M., Chen J., Cook J. L., Mankin A. S. & Neyfakh A. A. (2008). Nucleotide biosynthesis is critical for growth of bacteria in human blood. *PLoS Pathog*, **4**(2), e37. *Cité page 114.*

Sapp J. (2004). The dynamics of symbiosis : an historical overview. *Can. J. Bot.*, **82**, 1046–1056. *Cité pages 23 et 24.*

Sasaki T. & Ishikawa H. (1993). Nitrogen recycling in the endosymbiotic system of the pea aphid, acyrthosiphon pisum. *Zoological Science*, **10**, 779–785. *Cité pages 122 et 179.*

Sasaki T. & Ishikawa H. (1995). Production of essential amino acids from glutamate by mycetocyte symbionts of the pea aphid, acyrthosiphon pisum. *Journal of Insect Physiology*, **41**, 41–46(6). *Cité pages 178, 179 et 180.*

Schmidt S., Sunyaev S., Bork P. & Dandekar T. (2003). Metabolites : a helping hand for pathway evolution ? *Trends Biochem Sci*, **28**(6), 336–341. *Cité page 20.*

Schomburg I., Chang A., Ebeling C., Gremse M., Heldt C., Huhn G. & Schomburg D. (2004). Brenda, the enzyme database : updates and major new developments. *Nucleic Acids Res*, **32**(Database issue), D431–D433. *Cité page 40.*

Schuster S. & Hilgetag C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J Biol Systems*, **2**, 165–182. *Cité page 47.*

Servant F., Bru C., Carrère S., Courcelle E., Gouzy J., Peyruc D. & Kahn D. (2002). Prodom : automated clustering of homologous domains. *Brief Bioinform*, **3**(3), 246–251. *Cité page 35.*

Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B. & Ideker T. (2003). Cytoscape : a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**(11), 2498–2504. *Cité pages 59 et 77.*

Shigenobu S., Watanabe H., Hattori M., Sakaki Y. & Ishikawa H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature*, **407**(6800), 81–6. *Cité pages 30, 124, 158, 164 et 177.*

Strömbäck L. & Lambrix P. (2005). Representations of molecular pathways : an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics. Cité page 62.*

Stumpf M. P. H., Wiuf C. & May R. M. (2005). Subnets of scale-free networks are not scale-free : sampling properties of networks. *Proc Natl Acad Sci U S A*, **102**(12), 4221–4224. *Cité page 87.*

Sumner J. B. (1933). The chemical nature of enzymes. *Science*, **78**(2024), 335. *Cité page 11.*

Suzek B. E., Ermolaeva M. D., Schreiber M. & Salzberg S. L. (2001). A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**(12), 1123–1130. *Cité page 69.*

Tamames J., Gil R., Latorre A., Peretó J., Silva F. J. & Moya A. (2007). The frontier between cell and organelle : genome analysis of candidatus carsonella ruddii. *BMC Evol Biol*, **7**, 181. *Cité pages 99 et 119.*

Tamas I., Klasson L. M., Sandström J. P. & Andersson S. G. (2001). Mutualists and parasites : how to paint yourself into a (metabolic) corner. *FEBS Lett*, **498**(2-3), 135–9. *Cité page 84.*

Tanaka T., Ikeo K. & Gojobori T. (2006). Evolution of metabolic networks by gain and loss of enzymatic reaction in eukaryotes. *Gene*, **365**, 88–94. *Cité page 22.*

Tatusov R. L., Galperin M. Y., Natale D. A. & Koonin E. V. (2000). The COG database : a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, **28**(1), 33–36. *Cité page 33.*

Taylor M. J., Bandi C. & Hoerauf A. (2005). Wolbachia bacterial endosymbionts of filarial nematodes. *Adv Parasitol*, **60**, 245–284. *Cité pages 25, 66 et 94.*

Tenenhaus M. & Young F. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**(1), 91–119. *Cité page 106.*

Teusink B., van Enckevort F. H. J., Francke C., Wiersma A., Wegkamp A., Smid E. J. & Siezen R. J. (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**(11), 7253–7262. *Cité page 37.*

Theissen U. & Martin W. (2006). The difference between organelles and endosymbionts. *Curr Biol*, **16**(24), R1016–7 ; author reply R1017–8. *Cité page 29.*

Tipton K. F. (1994). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement : corrections and additions. *Eur J Biochem*, **223**(1), 1–5. *Cité page 19.*

Trappe J. M. (2005). A.b. frank and mycorrhizae : the challenge to evolutionary and ecologic theory. *Mycorrhiza*, **15**(4), 277–281. *Cité page 23.*

Vallenet D. *et al.* (2006). MaGe : a microbial genome annotation system supported by synteny results. *Nucleic Acids Res*, **34**(1), 53–65. *Cité pages 36, 56, 69 et 72.*

van der Giezen M. (2005). Endosymbiosis : past and present. *Heredity*, **95**(5), 335–336. *Cité page 29.*

van Nimwegen E. (2003). Scaling laws in the functional content of genomes. *Trends Genet*, **19**(9), 479–484. *Cité pages 84 et 85.*

Vasconcelos A. T. R., Ferreira H. B., Bizarro C. V., Bonatto S. L., Carvalho M. O., Pinto P. M., Almeida D. F., Almeida L. G. P., Almeida R., Alves-Filho L., Assunção E. N., Azevedo V. A. C., Bogo M. R., Brigido M. M., Brocchi M., Burity H. A., Camargo A. A., Camargo S. S., Carepo M. S., Carraro D. M., de Mattos Cascardo J. C., Castro L. A., Cavalcanti G., Chemale G., Collevatti R. G., Cunha C. W., Dallagiovanna B., Dambrós B. P., Dellagostin O. A., Falcão C., Fantinatti-Garboggini F., Felipe M. S. S., Fiorentin L., Franco G. R., Freitas N. S. A., Frías D., Grangeiro T. B., Grisard E. C., Guimarães C. T., Hungria M., Jardim S. N., Krieger M. A., Laurino J. P., Lima L. F. A., Lopes M. I., Loreto E. L. S., Madeira H. M. F., Manfio G. P., Maranhão A. Q., Martinkovics C. T., Medeiros S. R. B., Moreira M. A. M., Neiva M., Ramalho-Neto C. E., Nicolás M. F., Oliveira S. C., Paixão R. F. C., Pedrosa F. O., Pena S. D. J., Pereira M., Pereira-Ferrari L., Piffer I., Pinto L. S., Potrich D. P., Salim A. C. M., Santos F. R., Schmitt R., Schneider M. P. C., Schrank A., Schrank I. S., Schuck A. F., Seuanez H. N., Silva D. W., Silva R., Silva S. C., Soares C. M. A., Souza K. R. L., Souza R. C., Staats C. C., Steffens M. B. R., Teixeira S. M. R., Urmenyi T. P., Vainstein M. H., Zuccherato L. W., Simpson A. J. G. & Zaha A. (2005). Swine and poultry pathogens : the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae. *J Bacteriol*, **187**(16), 5568–5577. *Cité page 114.*

VerBerkmoes N. C., Connelly H. M., Pan C. & Hettich R. L. (2004). Mass spectrometric approaches for characterizing bacterial proteomes. *Expert Rev Proteomics*, **1**(4), 433–447. *Cité page 44.*

Voit E. O. (2002). Metabolic modeling : a tool of drug discovery in the postgenomic era. *Drug Discov Today*, **7**(11), 621–628. *Cité pages 44 et 45.*

Wächtershäuser G. (1990). Evolution of the first metabolic cycles. *Proc Natl Acad Sci USA*, **87**(1), 200–204. *Cité page 14.*

Wächtershäuser G. (2007). On the chemistry and evolution of the pioneer organism. *Chem Biodivers*, **4**(4), 584–602. *Cité page 14.*

Wagner A. & Fell D. A. (2001). The small world inside large metabolic networks. *Proc Biol Sci*, **268**(1478), 1803–1810. *Cité page 86.*

Wagner M. A., Eschenbrenner M., Horn T. A., Kraycer J. A., Mujer C. V., Hagius S., Elzer P. & DelVecchio V. G. (2002). Global analysis of the brucella melitensis proteome : Identification of proteins expressed in laboratory-grown culture. *Proteomics*, **2**(8), 1047–1060. *Cité page 44.*

Wang Z., Zhu X.-G., Chen Y., Li Y., Hou J., Li Y. & Liu L. (2006). Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria. *BMC Genomics*, **7**(1), 100. *Cité page 88.*

Watts D. J. & Strogatz S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**(6684), 440–442. *Cité page 86.*

Wernegreen J. J. (2005). For better or worse : genomic consequences of intracellular mutualism and parasitism. *Curr Opin Genet Dev*, **15**(6), 572–583. *Cité page 24.*

Werren J. H. (1997). Biology of wolbachia. *Annu Rev Entomol*, **42**, 587–609. *Cité page 67.*

Whitehead L. F. & A.E.Douglas (1993). A metabolic study of buchnera, the intracellular bacterial symbionts of the pea aphid acyrthosiphon pisum. *JOURNAL OF GENERAL MICROBIOLOGY*, **139**, 821. *Cité pages 175, 178 et 180.*

Wu C. H., Apweiler R., Bairoch A., Natale D. A., Barker W. C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M. J., Mazumder R., O'Donovan C., Redaschi N. & Suzek B. (2006a). The universal protein resource (uniprot) : an expanding universe of protein information. *Nucleic Acids Res*, **34**(Database issue), D187–D191. *Cité page 40.*

Wu D., Daugherty S. C., Aken S. E. V., Pai G. H., Watkins K. L., Khouri H., Tallon L. J., Zaborsky J. M., Dunbar H. E., Tran P. L., Moran N. A. & Eisen J. A. (2006b). Metabolic Complementarity and Genomics of the Dual Bacterial Symbiosis of Sharpshooters. *PLoS Biol*, **4**(6), e188. *Cité pages 24, 29, 119 et 142.*

Wu M., Sun L. V., Vamathevan J., Riegler M., Deboy R., Brownlie J. C., McGraw E. A., Martin W., Esser C., Ahmadinejad N., Wiegand C., Madupu R., Beanan M. J., Brinkac L. M., Daugherty S. C., Durkin A. S., Kolonay J. F., Nelson W. C., Mohamoud Y., Lee P., Berry K., Young M. B., Utterback T., Weidman J., Nierman W. C., Paulsen I. T., Nelson K. E., Tettelin H., O'Neill S. L. & Eisen J. A. (2004). Phylogenomics of the reproductive parasite wolbachia

pipientis wmel : a streamlined genome overrun by mobile genetic elements. *PLoS Biol*, **2**(3), E69. *Cité page 116.*

Yamanishi Y., Mihara H., Osaki M., Muramatsu H., Esaki N., Sato T., Hizukuri Y., Goto S. & Kanehisa M. (2007). Prediction of missing enzyme genes in a bacterial metabolic network. reconstruction of the lysine-degradation pathway of pseudomonas aeruginosa. *FEBS J*, **274**(9), 2262–2273. *Cité pages 36 et 42.*

Yanai I., Derti A. & DeLisi C. (2001). Genes linked by fusion events are generally of the same functional category : a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U S A*, **98**(14), 7940–7945. *Cité page 36.*

Yang F., Qian H. & Beard D. A. (2005). Ab initio prediction of thermodynamically feasible reaction directions from biochemical network stoichiometry. *Metab Eng*, **7**(4), 251–259. *Cité page 43.*

Zientz E., Dandekar T. & Gross R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiol Mol Biol Rev*, **68**(4), 745–70. *Cité pages 84, 99, 109, 116, 118, 122, 123, 158, 173, 175, 176 et 181.*

Zimmer H. G. (1996). Carl ludwig : the man, his time, his influence. *Pflugers Arch*, **432**(3 Suppl), R9–22. *Cité page 10.*

TITRE en français

Analyse systémique de la symbiose intracellulaire : évolution et organisation du réseau métabolique des endocytobiotes

RÉSUMÉ en français

Les bactéries endocytobiotes vivent de manière durable au sein même des cellules des organismes qui les abritent. L'environnement particulier qu'est la cellule de l'hôte, et le type de relations entre les deux partenaires (parasitisme ou mutualisme), ont naturellement des conséquences sur l'évolution de leur métabolisme respectif.
L'objectif global de cette thèse est de mieux comprendre l'influence du mode de vie sur l'évolution et le fonctionnement du métabolisme des endocytobiotes.
Nous appréhendons le métabolisme d'une manière globale, sous la forme de réseaux métaboliques. Le développement de nouvelles méthodes et d'outils d'exploration du réseau métabolique nous ont permis de réaliser des analyses et des comparaisons de réseaux métaboliques complets avec un niveau de détail élevé.
L'ensemble de ces analyses éclaire ainsi d'un jour nouveau le métabolisme des bactéries endocytobiotes par sa diversité, son évolution et la nature des interactions métaboliques entretenues avec l'hôte.

MOTS-CLEFS en français

métabolisme ; symbiose ; endocytobiose ; réseau ; graphe ; système ;

TITRE en anglais

Systemic analysis of the intracellular symbiosis : evolution and organisation of the metabolic network of endocytobionts.

RÉSUMÉ en anglais

Endocytobiont bacteria live durably inside some cells of their hosts. The peculiar environment, represented by the cell, and the type of relations between the two partners (parasitism or mutualism), have natural consequences on the evolution of the metabolism of both.
The main objective of this PhD was to better understand the links beetween the lifestyle, and the evolution and the functioning of the endocytobiont metabolism.
The metabolism was considered in a global way, and modelled a metabolic network. The development of new methods and original tools to explore and compare the metabolic networks of different symbionts enabled to perform various analyses at a scale and level of detail never realised before.
Such analyses provided new insights into the metabolism of endocytobiont bacteria through their observed diversity, evolution and relation with the host.

MOTS-CLEFS en anglais

metabolism ; symbiosis ; endocytobiosis ; network ; graphe ; system ;

DISCIPLINE : Bioinformatique

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :
Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS
Batiment Gregor Mendel - Université Claude Bernard Lyon1
43, bv du 11 novembre 1918 - 69622 Villeurbanne cedex