Complétion combinatoire pour la reconstruction de réseaux métaboliques, et application au modèle des algues brunes Ectocarpus siliculosus

Sylvain Prigent

ANNÉE 2014

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Sylvain PRIGENT

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
ISTIC

Complétion combinatoire pour la reconstruction de réseaux métaboliques, et application au modèle des algues brunes *Ectocarpus siliculosus*.

Thèse soutenue à Rennes
le 14 novembre 2014

devant le jury composé de :

Alexander BOCKMAYR
Professeur, Freie Universität, Berlin / *Président*

Marie BEURTON-AIMAR
Maître de conférences, Bordeaux 2 / *Rapporteure*

Hubert CHARLES
Professeur, INSA, Lyon / *Rapporteur*

Claudine MÉDIGUE
Directrice de recherche CNRS, CEA-Genoscope / *Rapporteure*

Arnaud MARTIN
Professeur, Université de Rennes 1 / *Examinateur*

Anne SIEGEL
Directrice de recherche CNRS / *Directrice de thèse*

Thierry TONON
Maître de conférences UMPC / *Co-directeur de thèse*

I thought there couldn't be anything as complicated as the universe, until I started reading about the cell.
par Eric de Silva

Remerciements

Je tiens tout d'abord à remercier mes deux directeurs de thèse, Anne Siegel et Thierry Tonon. Merci de m'avoir permis de réaliser ce doctorat avec vous et d'avoir été là pendant ces trois ans. Un grand merci pour votre soutien, qu'il fut professionnel, scientifique ou personnel. Travailler avec vous, en plus d'être très agréable, fut très enrichissant. Pour tout cela je vous remercie.

Merci également à Marie Beurton-Aimar, Claudine Médigue et Hubert Charles d'avoir bien voulu accepter la charge de rapporteur. Je remercie également Alexander Bockmayr et Arnaud Martin pour avoir accepté de juger ce travail.

J'adresse également mes remerciements à tout le "groupe Symbiose", composé des équipes Dyliss, GenScale et de la plate-forme GenOuest. Merci à tous pour cette ambiance exceptionnelle qui règne dans nos couloirs, merci pour l'ensemble des discussions (enrichissantes ou pas, d'ailleurs), pour toutes les activités du midi telles que l'escalade ou la batucada, et pour tout le reste. Espérons que cela reste comme ça encore longtemps, même après ce déménagement imposé. Je remercie également aux Nantais, Damien, Jérémie et Halim, pour les nombreuses discussions tellement intéressantes que nous avons eu, notamment du point de vue de la fonctionnalité des réseaux. Un merci tout particulier à Nicolas et Guillaume qui m'ont accompagné dans le bureau pendant tant de temps. Les deux ans passés avec vous furent tout simplement géniaux, les "pense vite", les courses au retour de la cafet et tout le reste m'ont beaucoup manqué après votre départ. Merci aussi à Malo qui a su les remplacer à merveille durant ces derniers mois. Merci également à Coraline, Nathalie, Valentin et Paulin pour l'ensemble des "cc" qui permettaient à la fois de se distraire, de se réveiller et d'avancer durant les moments de blocage. Au sein de Symbiose, un merci particulier à Gaëlle, Mathilde et Vincent pour l'ensemble du temps passé ensemble depuis ce stage de M2 génial.

Mes remerciement vont également à tout ceux qui m'ont permis de se changer les idées en dehors des heures de travail, qu'il s'agisse d'Adrien, Olivier et Stéphane, de tous les gens du Flood, et de l'ensemble des amis de Rennes et d'ailleurs. Tous les citer serait trop long mais je pense particulièrement à Cécile, Ismahane, Charles, Ludo, Chunky, et tous les autres.

Un grand merci à tout ceux qui m'ont fait découvrir et adorer la vulgarisation scientifique, qu'il s'agisse de l'association Nicomaque avec le festival Sciences en Cour[t]s ou toutes les personnes qui participent à l'aventure bioinfo-fr, par le blog ou le chan IRC si enrichissant.

Un énorme merci à Fréné qui a toujours été là au cours de cette thèse même si ça n'a pas dû être simple tous les jours.

Je ne peux pas terminer ces remerciements sans penser à l'ensemble de ma famille, merci à vous tous pour le soutien sans faille pendant ces trois années.

Et enfin merci à toutes les personnes auxquelles je ne pense pas en ce moment mais qui ont également participé à faire de ces trois années de thèse une expérience inoubliable et s'étant passée dans les meilleures conditions possibles.

Table des matières

Table des matières **1**

Introduction **5**

1 État de l'art **7**
1.1 Biologie des systèmes 7
1.2 Réseaux métaboliques 10
1.2.1 Définition 10
1.2.2 Fonctionnalité d'un réseau métabolique : Analyse en Balance de Flux (ou FBA : Flux Balance Analysis) 12
1.2.3 Optimisation d'une fonction objectif 13
1.2.4 Impact de la structure du réseau 14
1.2.5 Vérification de la fonctionnalité d'un réseau : Flux Variability Analysis . 15
1.3 Pipeline de reconstruction d'un réseau métabolique 15
1.3.1 Construction d'une ébauche métabolique 18
1.3.2 Raffinement de l'ébauche métabolique 22
1.3.3 Évaluation du réseau final 23
1.3.4 Pipelines existants 23
1.4 Complétion de réseaux : problèmes d'optimisation induits 24
1.4.1 Panorama général 24
1.4.2 Optstrain, SMILEY - top-down & MILP 26
1.4.3 GapFind, GapFill - bottom-up & MILP 27
1.4.4 Christian et al - top-down & stochastique 29
1.4.5 Network-expansion - bottom-up & optimisation combinatoire 30
1.4.6 Comparaisons 32
1.5 Reconstruction de réseaux métaboliques pour les algues 34
1.5.1 Les réseaux métaboliques chez les plantes 34
1.5.2 Les réseaux métaboliques chez les algues 34
1.5.3 Pipelines de reconstructions utilisés 35
1.5.4 *Ectocarpus siliculosus* 37
1.6 Principaux apports/résumé 40

2 Complétion combinatoire de réseau métabolique **41**
2.1 Problème d'optimisation 41
2.1.1 Espace de recherche 41
2.1.2 Atteignabilité 42

2.1.3 Problème de complétion . . . 43
2.2 Jeux de test . . . 44
2.2.1 Impact de la taille des bases de complétion : *Ectocarpus siliculosus* & MetaCyc . . . 44
2.2.2 Effet de la taille de la base de données sur la productibilité des cibles . . . 45
2.2.3 Fonctionnalité : E. Coli . . . 46
2.3 Performance . . . 48
2.3.1 Heuristiques de résolutions . . . 48
2.3.2 Nouvelle méthode d'optimisation dans *Network-expansion* . . . 49
2.3.3 Recherche d'optimum . . . 50
2.3.4 Impact de la taille de la base . . . 51
2.4 Impact de la réversibilité sur les performances de *Network-expansion* . . . 54
2.4.1 Réversibilité dans *Network-expansion* . . . 54
2.4.2 Nouvelle modélisation ASP de la réversibilité . . . 55
2.4.3 Impact sur les performances . . . 55
2.4.4 De *Network-expansion* à *Meneco* . . . 57
2.5 Sémantique de productibilité . . . 57
2.5.1 Impact des cycles sur les différents concepts d'accessibilité . . . 57
2.5.2 Comparaison des différentes sémantiques . . . 59
2.5.3 Comparaison des sémantiques qualitatives sur la complétion . . . 61
2.5.4 Fonctionnalité des complétions qualitatives . . . 64
2.6 conclusion . . . 67

3 Pipeline de reconstruction de réseaux à partir de données hétérogènes 69
3.1 Modèle d'application : *Ectocarpus siliculosus* . . . 69
3.2 Construction d'une ébauche du réseau : gestion des données hétérogènes . . . 71
3.2.1 Reconstruction à partir des annotations . . . 71
3.2.2 Reconstruction à partir d'un réseau métabolique de réference . . . 72
3.2.3 Complémentarité des deux réseaux : outils pour leur fusion . . . 72
3.2.4 Conclusion . . . 73
3.3 Complétion . . . 73
3.3.1 Identification de réactions à ajouter . . . 73
3.3.2 Filtrage par analyse sémantique . . . 74
3.3.3 Compartimentation . . . 75
3.4 Curation manuelle du réseau . . . 76
3.4.1 Score pour chaque réaction . . . 76
3.4.2 Identification de faux positifs . . . 76
3.4.3 Ajout de réactions spécifiques . . . 77
3.4.4 Conclusion . . . 77
3.5 Validation fonctionnelle . . . 78
3.5.1 Réseau final . . . 78
3.5.2 Analyse par balance de flux . . . 78
3.6 Le workflow de reconstruction du réseau . . . 80
3.6.1 Description du pipeline . . . 80
3.6.2 Outils intégrés dans le pipeline . . . 81
3.7 Conclusion . . . 84

4 Contribution à l'amélioration des connaissances sur la physiologie des algues brunes 85
4.1 *Ectocarpus siliculosus* : ses spécificités 85
4.2 Analyse de la reconstruction automatique du réseau métabolique 86
4.2.1 La synthèse des alginates 86
4.2.2 Le cycle du mannitol 88
4.3 Étude de voies métaboliques chez *Ectocarpus siliculosus* à l'aide du réseau métabolique 88
4.3.1 Synthèse des acides aminés aromatiques 88
4.3.2 Synthèse du molybdenum 92
4.4 Réannotation des gènes 96
4.5 Conclusion 98

Conclusion et perspectives 99
Conclusion 99
Perspectives 101

Bibliographie 118

Table des figures 119

Annexes 122

A Liste des molécules utilisées lors de l'étude de l'efficacité de *Meneco* 125

B Code ASP original de Network-Expansion 127
B.1 ASP 127
B.2 Network-Expansion 128
B.3 Modélisation 129

C Réannotation de gènes 133

Introduction

Cette thèse a porté sur le développement de méthodes génériques de reconstruction de réseaux métaboliques à l'échelle génomique, c'est-à-dire des cartes complètes de réactions biochimiques qui transforment des molécules en d'autres molécules, pour des organismes eucaryotes. Cette méthode a été appliquée à la reconstruction du premier réseau métabolique du modèle des algues brunes, *Ectocarpus siliculosus*, et a permis de mieux comprendre la biologie et l'histoire évolutive des algues. Ce travail est ainsi naturellement interdisplinaire : pour réaliser la chaine complète d'analyse et d'interprétation, il a été nécessaire de proposer des contributions en informatique, bioinformatique, et biologie.

Classiquement, la reconstruction d'un réseau métabolique s'articule en trois points : la création d'une ébauche métabolique à partir des différentes informations comprises dans le génome, la complétion de cette ébauche métabolique pour être en accord avec les connaissances biologiques et enfin la vérification du résultat obtenu. Ces travaux ont mis en évidence que les différentes approches de reconstruction existantes ne pouvaient pas s'appliquer aux espèces eucaryotes non classiques à cause de deux écueils majeurs.

Tout d'abord les connaissances sur les espèces non classiques sont trop éparses pour reconstruire une ébauche métabolique correcte permettant une complétion manuelle. Pour palier à cette difficulté, nous nous proposons d'étudier un problème d'optimisation combinatoire qui permet de formaliser la question de la complétion automatique d'un réseau. Pour résoudre ce problème combinatoire, l'objectif de ce document sera de tirer profil d'un paradigme de programmation par contraintes relativement récent, la Programmation par Ensemble Réponse (ou Answer Set Programming, ASP). Une modification d'un modèle existant en ASP sera proposée pour améliorer la pertinence biologique de la modélisation et rendre la résolution efficace et applicable à l'échelle des applications visées. Plus généralement, la question de la pertinence biologique du modèle sous-jacent au problème tel que formalisé sous sa forme combinatoire fait l'objet d'une attention toute particulière.

D'autre part, ce travail de thèse s'est heurté à la nécessité de travailler avec des sources de données très différentes pour permettre la reconstruction d'un réseau. L'utilisation de ces sources multiples et hétérogènes nécessite de travailler sur leur unification avant et pendant leur intégration dans le réseau. Il s'avère également nécessaire de traiter *a posteriori* les résultats des méthodes automatiques de reconstruction pour leur donner une signification biologique. Pour donner globalement du sens fonctionnel aux résultats des analyses automatiques, on s'appuiera à la fois sur des approches liées à la représentation des connaissances et aux études de séquences biologiques.

Comme nous le verrons, ces apports méthodologiques ont permis de développer un pipeline de reconstruction de réseaux métaboliques s'appuyant sur un ensemble de données biologiques telles que des données génomiques, transcriptomiques ou de profilage méta-

bolique. Ce pipeline a permis de reconstruire le premier réseau métabolique du modèle des algues brunes, *Ectocarpus siliculosus*. Cependant, sa valeur ajoutée réside aussi dans son exploitation dans le monde de la biologie. Ainsi, nous détaillerons comment l'étude du réseau a permis, en collaboration avec la station biologique de Roscoff, de mieux comprendre la biologie de cette algue, notamment sur deux aspects : une annotation de nouveaux gènes ou une réannotation de gènes, et une meilleur compréhension de l'histoire évolutive et du métabolisme de cette algue.

Une fois cette preuve de concept réalisée pendant la thèse sur cette algue brune, la méthode proposée est actuellement en cours d'application et d'amélioration en utilisant d'autres organismes, ce qui ouvre la voie à différentes perspectives de nature inforamtique, bio-informatique et biologiques.

Chapitre 1

État de l'art. Reconstruction de réseaux métaboliques et applications à des modèles d'algues

Dans ce chapitre de bibliographie, nous allons présenter une introduction aux différentes problématiques informatique et bioinformatique liées à la reconstruction de réseaux métaboliques. Nous nous concentrerons en particulier sur les questions d'optimisation, qu'elle soit combinatoire ou linéaire, et sur les questions liées à la représentation des connaissances sous-jacentes dans cette problématique. En particulier, ces questions seront abordées en ayant à l'esprit des applications sur des organismes eucaryotes non classiques, tels que les algues brunes qui seront notre modèle d'application.

Dans ce chapitre, nous commencerons par une étude globale de la biologie des systèmes dans la partie 1.1 avant de se concentrer sur les réseaux métaboliques et les études possibles à partir de ceux-ci (partie 1.2). Nous présenterons ensuite les différentes méthodes existantes de reconstructions de réseaux métaboliques dans la partie 1.3. Nous nous concentrerons alors sur une étape cruciale de reconstruction de réseaux métaboliques, qui sera particulièrement étudiée dans cette thèse : la complétion des réseaux métaboliques (partie 1.4). Cette thèse s'attachant plus particulièrement au développement du réseau métabolique d'une algue brune, *Ectocarpus siliculosus*, nous verrons enfin les différentes reconstructions de réseaux métaboliques ayant eu lieu jusqu'à présent chez des algues dans la partie 1.5.

1.1 Biologie des systèmes

Depuis les années 1990, les données de biologie moléculaire s'accumulent avec le développement des techniques dites "omiques" [Pal02]. La granularité et la quantité de données produites permet aujourd'hui d'étudier les systèmes biologiques à un niveau beaucoup plus fin qu'au 20ème siècle, au moment où Jacob et Monod décrivaient le fonctionnement de l'opéron lactose et les mécanismes de régulation géniques[JM61].

Pour les espèces les plus étudiées (comme *Escherichia coli*, la drosophile, la souris, l'humain ou encore *Arabidopsis thaliana*) et de plus en plus pour les espèces non classiques, nous possédons actuellement l'ensemble du génome, du transcriptome, du protéome et du

métabolome, un des séquençages de génome le plus connu étant bien évidemment celui du génome humain [LLB+01]. Le domaine de la *biologie des systèmes* s'est naturellement développé pour permettre d'intégrer l'ensemble de ces connaissances par des approches informatiques et mathématiques. L'objectif est de modéliser *in silico* la réponse des différentes espèces à différentes perturbations génétiques ou environnementales en ne limitant plus les études à l'effet d'un seul gène ou d'une protéine, mais à l'étude du comportement d'une cellule dans son ensemble, voire à moyen terme d'un organe ou d'un organisme entier. [Kit01] défini quatre points essentiels à l'étude de systèmes biologiques par la biologie des systèmes : la conception du système étudié, sa structure, sa dynamique et enfin le contrôle de ce système. Si l'étude fine de la dynamique et le contrôle d'un modèle semblent *a priori* les plus intéressants pour analyser la réponse globale d'un système à des modifications internes et/ou externes, les deux premiers points sont tout aussi intéressants à étudier. Ils sont non seulement indispensables dans le processus d'analyse mais intéressants en eux-mêmes. En effet, la reconstruction et l'étude "structurelle" du système étudié sont indispensables à des études plus fines. Mais au delà de ça, la reconstruction même du système à étudier apporte des connaissances sur l'organisme ou les organismes étudiés via l'étape compliquée de représentation des connaissances nécessaire. De même l'étude de la structure du système (la plupart du temps sous forme de réseau) peut apporter de nombreuses nouvelles connaissances.

Désormais les différents composés d'un système biologique (ADN, ARN, protéines, ...) ne seront plus étudiés un à un mais tous ensemble. L'interaction entre ces différentes entités sera la pierre angulaire de la biologie des systèmes. Nous pouvons décomposer ces interactions en trois types de réseaux, comme illustré dans la Figure 1.1 [MCR+11] :

- Les réseaux de signalisation [JLI00] vont permettre de décrire l'ensemble des mécanismes qui permettent à une cellule de répondre rapidement aux modifications de son environnement en transmettant rapidement de l'information au sein de la cellule. Cela se fera notamment par l'activation de différents récepteurs membranaires qui pourront conduire à l'activation ou l'inhibition de gènes.
- Les réseaux de régulation génique [DL05] vont regrouper l'ensemble des interactions relatives à la régulation de l'expression des gènes d'une espèce. Cette régulation pourra avoir en particulier lieu via des facteurs de transcription. La transcription de ces facteurs de transcription sera elle-même régulée par d'autres protéines.
- Les réseaux métaboliques vont regrouper l'ensemble des réactions métaboliques présentes au sein d'un système, c'est-à-dire des transformations chimiques de molécules au sein de l'organisme. Ces réactions sont catalysées par des enzymes qui sont des protéines codées par des gènes.

Dans son article de référence, Kitano [Kit01] introduit quatre points essentiels à l'étude de systèmes biologiques : étudier la structure du système, puis sa dynamique, le contrôler, et enfin concevoir des systèmes minimaux vérifiant des propriétés fixées.

Pour mieux comprendre ces aspects, nous pouvons nous référer à une analogie entre un réseau biologique et un réseau routier.

Ainsi nous pouvons analyser la structure d'un graphe découlant de l'analyse de données biologiques de la même manière que nous pouvons analyser une carte routière. Il est par exemple possible de chercher le chemin le plus court entre deux endroits (molécule ou ville), de regarder les nœuds du réseau qui sont les plus connectés au reste du réseau (qui pourraient correspondre à une molécule ou une ville plus importante que les autres), etc.

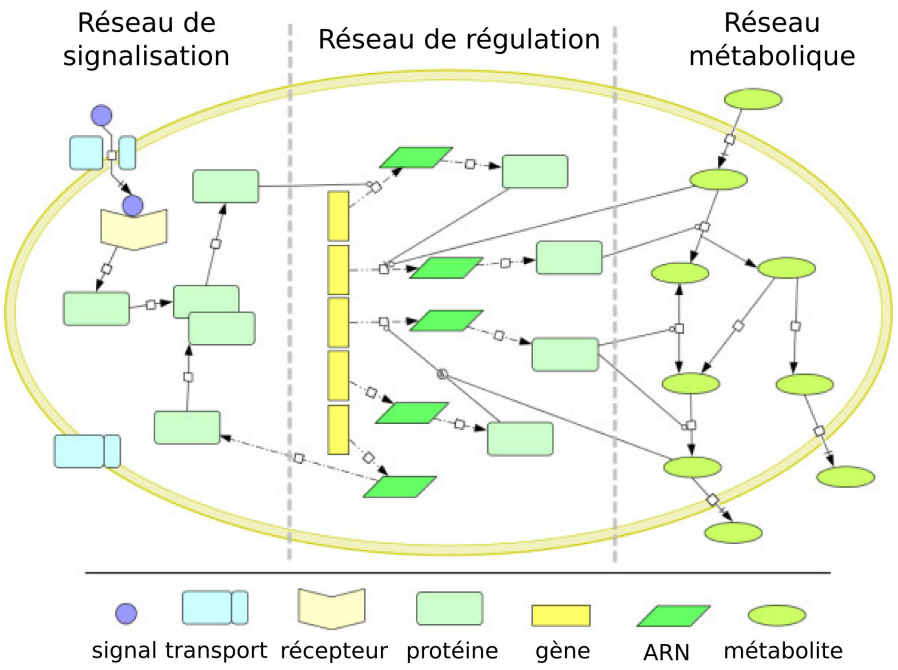


FIGURE 1.1: **Représentation des principaux processus se produisant dans une cellule**. La cellule reçoit des signaux extérieurs qui engendrent des cascades de signalisation à l'intérieur de la cellule. Ces signaux et les cascades induites sont étudiées dans les *réseaux de signalisation*. Ce signal arrive jusqu'au noyau et engendre une régulation de l'expression des gènes. Cette régulation sera étudiée par les *réseaux de régulation géniques*. Les gènes sont ensuite transcrits en ARN messagers qui pourront être traduits en protéines. Certaines de ces protéines (appelées enzymes) pourront catalyser des réactions métaboliques qui transformeront les molécules présentes à l'intérieur de la cellule ou importées de l'extérieur. Ces réactions sont étudiées dans les *réseaux métaboliques*.

Les différents types de réseaux mentionnés ci-dessus ont fait l'objet d'analyses statiques variées, mettant en évidence des structures différentes en fonction de la nature des réseaux, ainsi que des motifs caractéristiques. Nous pouvons par exemple citer les travaux de [JTA+00] qui ont montré que les réseaux métaboliques semblent suivre une structure sans échelle, faisant de ces réseaux des réseaux petit monde. Si une structure globale semble exister dans les réseaux biologiques, il en va de même à un niveau plus précis. Ainsi [Alo07] a distingué et classé de nombreux motifs revenant souvent dans les réseaux biologiques.

Nous pouvons aussi étudier un réseau routier de manière dynamique en ne regardant plus seulement la structure de la carte routière mais les différentes entités qui utilisent ce réseau et les flux qui y circulent. Certaines sont relativement grosses et lentes, et se déplacent habituellement entre des endroits précis prévus à l'avance. D'autres seront plus petites et imprévisibles. L'étude des interactions au cours du temps entre ces différents acteurs permettra de mieux comprendre la réalité d'un réseau routier, comparé à la seule étude de la structure des routes. On pourra d'abord prédire l'importance des flux qui passent sur le réseau routier, identifier les dépendances entre ces flux, y compris éloignés, et l'impact de modifications locales sur l'état global du trafic. On pourra ensuite identifier les points de contrôle de ce trafic pour mieux le réguler ou optimiser son comportement. Des méthodes très variées ont été développées pour étudier la dynamique des réseaux biologiques, méthodes qui peuvent être numériques, qualitatives, ou stochastiques en fonction des informations disponibles sur la réponse du système et de la nature des interactions en jeu. Nous pouvons par exemple citer les travaux de Cornish-Bowden sur les cinétiques enzymatiques [CB79] permettant aux modélisations continues du métabolisme de se développer.

Cette thèse va se concentrer sur la toute première étape de ce processus, la reconstruction de cartes métaboliques, en particulier chez des espèces non classiques pour lesquelles la quantité d'information en notre possession est limitée.

1.2 Réseaux métaboliques

1.2.1 Définition

Un réseau métabolique est défini comme un graphe dirigé biparti :
- un des ensembles de nœuds est constitué par des métabolites au sein de la cellule,
- l'autre ensemble nœuds correspond à des réactions métaboliques, qui tranforment un certain nombre de métabolites en d'autres métabolites, souvent sous l'action d'un catalyseur appelé enzyme. La réaction peut-être unidirectionnelle (irréversible) ou bidirectionnelle (réversible), en fonction des conditions intra-cellulaires.

La figure 1.2 montre une représentation graphique de deux réactions, la première ($R1$) transforme une molécule A en une molécule B et est irréversible. La seconde ($R2$) réalise une transformation conjointe des molécules B et C pour produire D. Cette seconde réaction est réversible.

Ces réactions sont habituellement associées à des gènes via une GPR (Gene Protein Relation). Ainsi, chaque réaction est associée à une ou plusieurs enzymes qui vont catalyser les réactions présentes dans le réseau. Ces enzymes sont elles-mêmes obtenues via la transcription, la traduction, et éventuellement la modification post-transcriptionnelle de un ou plusieurs gènes. De manière générale les enzymes ne sont pas représentées dans les réseaux métaboliques. Le lien est directement fait entre une réaction et un ou plusieurs gènes. Les

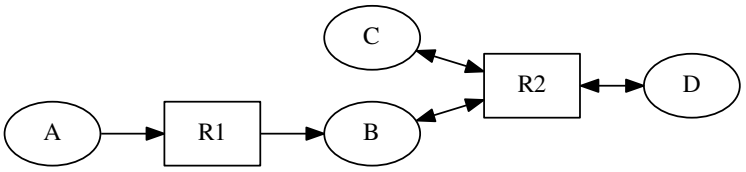


FIGURE 1.2: **Représentation graphique d'un réseau métabolique composé de deux réactions.** La première réaction ($R1$) est irréversible et produit B à partir de A. La seconde ($R2$) transforme $B + C$ en D de manière réversible.

éventuelles modifications post-traductionnelles influençant l'activité des enzymes ne sont pas prises en compte à ce niveau mais pourront être ajoutées directement dans la structure du réseau ou en tant que surcouches de celui-ci.

Durant les cinquante dernières années, les réseaux métaboliques ont été exploités dans de nombreux domaines de la biologie. Nous pouvons en citer deux particulièrement notables. En biologie évolutive, l'étude de réseaux métaboliques, souvent sous la forme de structure de graphe, a permis une meilleure compréhension de la biologie et de l'histoire évolutive des organismes étudiés. Dans ces approches, des alignements de séquences permettent de suggérer une fonction de nature enzymatique pour un ensemble de gènes donnés. La comparaison des voies métaboliques entre différentes espèces et l'identification d'enzymes spécifiques aux espèces étudiées permet de comprendre globalement l'histoire des espèces ciblées d'un point de vue fonctionnel [FMP03].

Inspiré par les biotechnologies, l'étude des réseaux métaboliques a fait l'objet d'études extrêmement nombreuses durant les 20 dernières années, en particulier des dépendances induites par les conservations de masse et les stœchiométries des réactions. Un exemple phare est l'étude de la régulation du métabolisme des globules rouges humains réalisé par [PRP+03], de même que tous les travaux sur le fonctionnement optimal du métabolisme d'*Escherichia Coli*. Ces approches s'appuient sur la décomposition en modes élémentaires de voies métaboliques et des techniques d'optimisation linéaire [SDF99, PVPF04]. On se dirige maintenant vers la biologie synthétique et la production contrôlée de métabolites d'intérêt, par exemple pour l'agroalimentaire ou l'industrie pétrolière [BS05, PW09]. Dans cette direction, de nouvelles questions d'optimisation apparaissent actuellement, telles que le calcul de "capacitance" permettant d'identifier les réactions chimiques à ajouter dans un réseau métabolique pour maximiser la production d'un métabolite donné [LBG+12].

Durant cette thèse, nous nous concentrerons essentiellement sur le premier de ces deux points, en étudiant les points informatiques bloquants pour la reconstruction automatique de réseaux métaboliques pour des organismes non-classiques, et ce qu'on peut en déduire en terme de nouvelles connaissances biologiques sur l'espèce considérée. Les méthodes d'optimisation linéaire utilisées dans les approches de nature biotechnologiques seront plutôt utilisées comme validation des résultats. L'exploitation et le raffinement des réseaux d'espèces non-classiques pour permettre leur étude et leur contrôle sont pour l'instant envi-

sagé dans le cadre de perspectives. Les questions que nous allons développer vont porter sur les questions d'optimisation combinatoire et les questions de représentation des connaissances qui permettent de reconstruire un réseau métabolique.

1.2.2 Fonctionnalité d'un réseau métabolique : Analyse en Balance de Flux (ou FBA : Flux Balance Analysis)

Pour tester et valider les méthodes qui vont être discutées dans cette thèse, nous utiliserons principalement un critère de fonctionnalité quantitative. Dans ce but, nous introduisons ici les différents éléments permettant de modéliser cette fonctionnalité.

La FBA est une technique d'analyse quantitative de la fonctionnalité de réseaux, et en particulier de réseaux métaboliques. Elle va permettre de simuler le métabolisme d'une cellule et de connaître la distribution des flux de matière passant à travers les réactions lorsqu'un modèle est à l'état stable. Le calcul est très rapide même pour de très grands réseaux. Cette technique permet de faire de l'ingénierie, de simuler la réponse des cellules à différentes conditions, des délétions de gènes, etc. La FBA se différencie des méthodes classiques d'étude de réseaux utilisant des systèmes d'équations différentielles par la rapidité de calcul et le peu d'informations nécessaire aux simulations. Ainsi il n'y a pas besoin de connaître les constantes cinétiques des enzymes ou encore la concentration des métabolites internes au système pour pouvoir effectuer des simulations.

Cette approche est basée sur l'hypothèse que le système métabolique est à un état stationnaire, qui implique que des cellules en culture dans des conditions données utilisent l'ensemble des nutriments à leur disposition pour créer de la biomasse. Étant donné qu'il n'y a ni création, ni de perte de matière, les entrées dans le modèle (les nutriments) correspondront parfaitement à la production de biomasse qui est représentée par les métabolites de l'organisme. D'un point de vue mathématique, le réseau sera représenté par une matrice stœchiométrique S formée de n lignes et m colonnes. Chaque ligne représentera un métabolite et chaque colonne une réaction. Les chiffres présents dans cette matrice correspondront aux coefficients stœchiométriques des réactions. Par exemple les réactions :

$$R1 : 2A + 3B \rightarrow 2C + D$$
$$R2 : 3C + 2A \rightarrow 4E$$

correspondent à la matrice stœchiométrique représentée en table 1.1.

TABLE 1.1: Exemple de matrice stœchiométrique

	R1	R2
A	-2	-2
B	-3	0
C	2	-3
D	1	0
E	0	4

Les réseaux métaboliques mettent en jeu un grand nombre de réactions (de l'ordre du millier), et chaque réaction implique en général moins d'une dizaine de métabolites différents. De plus, dans un réseau métabolique, le nombre de métabolites correspond environ

à l'ordre de grandeur de la taille du réseau (de l'ordre du millier de métabolites). À partir de là, il ressort immédiatement que la matrice stœchiométrique est une matrice creuse, contenant moins d'une dizaine de valeurs non nulles par colonne.

Une des hypothèses fondamentale de la FBA consiste à considérer que lors de l'étude du système, celui-ci se situe à "un état stationnaire". D'un point de vue mathématique, cela se représente par une contrainte simple :

$$S.v = 0,$$

avec S la matrice stœchiométrique, v un vecteur de flux dont la taille fait le nombre de réactions dans le modèle, et v_i le flux passant dans une réaction i. Ce flux correspond à la quantité de matière qui passe à travers chaque réaction pour que le système soit à l'état stable. Il est bien évidemment possible d'avoir plusieurs vecteurs v qui satisfont cette contrainte.

Une fois cette contrainte émise, on va pouvoir travailler sur le reste du système. Pour cela, on aura un autre type de contrainte : les bornes de flux pour chaque réaction. En effet, chaque réaction ne pourra avoir qu'une quantité de flux comprise entre une borne inférieure et une borne supérieure qui passera à travers elle. Typiquement, une réaction irréversible aura par exemple toujours une borne inférieure qui sera positive ou nulle et une bonne supérieure positive. D'autre part, une réaction réversible aura forcément une borne inférieure négative et une borne supérieure positive.

1.2.3 Optimisation d'une fonction objectif

Une fois ces deux contraintes posées (état stable et bornes sur les flux des réactions), reste à savoir ce que l'on souhaite modéliser. Habituellement, le choix se porte sur une maximisation de la biomasse. Le concept de "biomasse" représente le poids sec d'une cellule, c'est à dire la somme des quantités de chacun de ses composants les plus importants. Elle se calcule sous la forme d'une combinaison linéaire $Z = c^T v$ dépendant des flux traversant le réseau. Le vecteur Z peut être estimé par des techniques de biologie permettant de mesurer la concentration des composés au sein d'une cellule, ou il peut être estimé en se basant sur des biomasses précédemment définies chez des espèces proches physiologiquement. Lors de l'étude, cette fonction de biomasse sera maximisée. Au final le problème peut être représenté de la manière suivante (pour une matrice S de taille NxM) :

$$\begin{gathered}\text{Maximiser } Z = c^T v \\ \text{Avec : } \sum_{j=1}^{M} S_{ij} v_j = 0, \forall i \in 1..N \\ v_j^{\min} \leq v_j \leq v_j^{\max}, \forall j \in 1..N\end{gathered}$$

Un vecteur de flux v qui permettra d'obtenir l'ensemble des métabolites de la fonction objectif dans les proportions voulues sera produit par les algorithmes de résolution de ce problème. Le vecteur c est alors un vecteur de poids qui représente la proportion dans laquelle chaque réaction de v participe à l'objectif. Étant donné qu'habituellement une seule réaction (la réaction de biomasse) est comprise dans l'objectif, le vecteur c est entièrement

constitué de 0 et de un seul 1 à la position d'intérêt. Il est cependant possible d'optimiser une combinaison de plusieurs réactions

Différents outils existent pour réaliser des études de FBA sur des réseaux métaboliques. Le plus connu et le plus utilisé est la Cobra Toolbox [BFM$^+$07] pour Matlab. Si l'outil est disponible librement, il est à noté qu'il tourne sur une plateforme (Matlab) qui n'est ni libre ni gratuite, bien que présente dans la plupart des laboratoires d'informatique et de bioinformatique. En revanche, la plupart des laboratoires de biologie ne sont pas équipés de cette plateforme, ce qui limite son utilisation dans ces laboratoires. On peut en revanche signaler qu'une implémentation en Python de la Cobra Toolbox existe, appelée CobraPy [ELPH13], elle aussi totalement libre et gratuite et ne nécessitant pas l'installation préalable de Matlab ou de solveur payants.

Les résultats prédits par les analyses basées sur ces concepts sont nombreux et variés. Nous pouvons par exemple citer la prédiction du comportement d'un système à l'inactivation d'un gène [EP00, BES$^+$01], la prédiction de croissance cellulaire [EIP01, IEP02], la modification de la performance d'un réseau métabolique suite à une déletion ou une addition de gènes [BVM01] ou encore la suggestion d'inactivation de gènes afin d'augmenter la production de métabolites [BPM03, PBM04].

En ce qui concerne le sujet de cette thèse (reconstruction des réseaux métaboliques), ces méthodes sont utilisées principalement pour vérifier la fonctionnalité des réseaux reconstruits, via la définition d'une fonction de biomasse ou de cibles métaboliques prédéfinies.

1.2.4 Impact de la structure du réseau

L'analyse de réseaux par balance des flux se heurte aujourd'hui à un problème : l'utilisation de la stœchiométrie des réactions. En effet, si celle-ci est déterminante dans l'analyse, elle n'est pas toujours connue précisément et peut même, parfois, ne pas être déterminable. Nous pouvons par exemple citer les réactions de production de certaines molécules telles que les acides gras qui, en s'allongeant, produisent toujours la même molécule ($fatty\ acid + molecule \rightarrow same\ fatty\ acid$). Ce genre de réactions seront souvent retirées du système et pas prise en compte. De plus une des hypothèses de base de la FBA consiste à avoir un système à l'état stable. Or on ne peut considérer qu'un système biologique est à l'état stable uniquement quand celui-ci n'est pas perturbé. Ainsi l'utilisation de la FBA pour étudier des modifications de l'environnement, de conditions de stress ou de systèmes de régulation peut se trouver limitée [RC09].

Dans le même esprit, on note une certaine instabilité des tests de fonctionnalité. Les implémentations et les vérifications inclues dans les suites logicielles liées aux réseaux métaboliques ne sont pas toujours explicites et peuvent procéder à des transformations cachées qui peuvent avoir un impact important sur la fonctionnalité des réseaux. Ainsi, le réseau EcoCyc [KCVGC$^+$05] présent dans la banque MetaCyc est présenté comme fonctionnel par le logiciel Pathway Tools, qui inclut sa propre vérification de production de biomasse par FBA [LKTK12]. Or, si nous récupérons le réseau directement depuis Pathway Tools pour tenter d'appliquer des techniques de FBA sur ce modèle, en utilisant la même fonction de biomasse, le réseau devient étrangement non fonctionnel. Il semblerait que Pathway Tools procède à des opérations de dégénéricisation de réactions qui transforment complètement le résultat des analyses.

De nombreuses méthodes ont été dérivées du cadre du FBA pour étudier les réseaux métaboliques. On se référera par exemple à la figure 1.3 tirée de [PRP04] pour un panorama de ces méthodes.

Parmi ces méthodes, nous voudrions insister sur une approche particulièrement utile pour vérifier la fonctionnalité des réseaux et le rôle de certaines réactions en leur sein, l'analyse de variabilité de flux (figure 1.3, 11).

1.2.5 Vérification de la fonctionnalité d'un réseau : Flux Variability Analysis

Lors de l'utilisation de techniques de FBA, nous obtenons habituellement une solution qui optimise une fonction objective, comme par exemple une fonction de biomasse. Cependant, il peut exister des optimaux alternatifs à cette solution. L'analyse de variabilité de flux (ou Flux Variability Analysis, FVA) [GT10] va permettre d'étudier ces optimaux alternatifs. Nous pourrons ainsi obtenir les flux minimaux et maximaux possibles pour chaque réaction tout en maintenant la fonctionnalité du réseau. Cette fonctionnalité globale pourra être soit complète (on garde le même optimum), soit suboptimale si on ne garde qu'une partie de l'optimum.

L'étude d'un réseau par FBA (décrite précisément au-dessus) nous permet d'obtenir la valeur de l'optimum souhaité. Une fois cet optimum fixé, la FVA nous permettra de connaitre les plages de flux pour chaque réaction qui permettent d'obtenir cet optimum. On pourra ainsi définir différents types de réactions :

- Les réactions obligatoires : la plage de flux passant à travers elles sera toujours positive
- Les réactions bloquées : les flux seront nuls dans tous les scénarios possibles
- Les réactions "accessoires" : les flux pourront être nuls ou non

Les études de FVA peuvent se faire soit en conservant l'optimum inchangé, soit en prenant des sous-optimums. D'un point de vue mathématique, on aura donc pour les plages de flux correspondant parfaitement à l'objectif Z :

$$\begin{gathered}\text{Maximiser ou minimiser } v_j \\ \text{Avec : } Sv = 0 \\ v^{min} \leq v \leq v^{max} \\ Z^t v = Z_{obj}\end{gathered}$$

Pour obtenir les plages de flux correspondant à des solutions sous-optimales, il suffit de remplacer $Z^t v = Z_{obj}$ par $Z^t v \geq Z_{obj}.\gamma$ avec $\gamma \in [0;1]$ (plus gamma est proche de 1, plus on est proche de l'optimum).

Différentes implémentations de la FVA existent, en utilisant des plateformes telles que Matlab ou fonctionnant par elles-mêmes. Nous pouvons notamment citer fastFVA [GT10] compilé en tant que programme Matlab et qui peut utiliser les solveurs GLPK ou CPLEX.

1.3 Pipeline de reconstruction d'un réseau métabolique

Le développement des techniques de séquençage haut débit et leur coût de plus en plus faible, couplé au développement des autres méthodes "omiques" permettant d'obtenir de

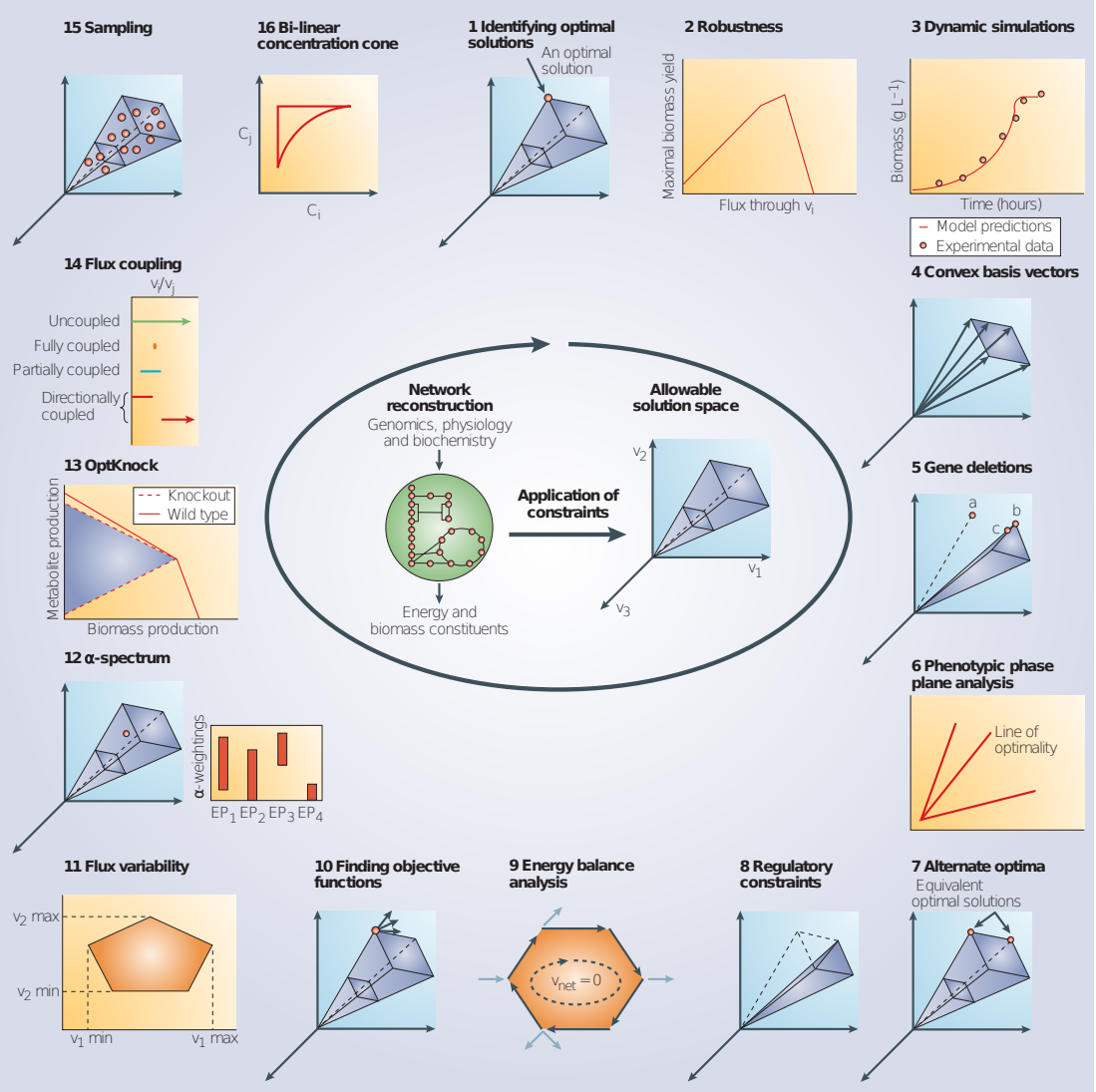


FIGURE 1.3: **Les méthodes classiques à base de contraintes pour l'analyse de réseaux métaboliques [PRP04].** La construction de l'espace des solutions est représentée au centre, les méthodes d'analyse de cet espace des solutions sont représentées autour.

grandes quantités de données, nous permet aujourd'hui d'avoir de grandes quantités d'informations. Ces informations, combinées entre elles, sont aujourd'hui suffisantes pour reconstruire des réseaux métaboliques. Ces réseaux peuvent être de taille et de qualité très variables.

Dans un récent article [MNP14] les réseaux métaboliques existant en février 2013 ont été étudiés, notamment du point de vue de la couverture phylogénétique des espèces modélisées. La figure 1.4 présente un arbre de l'ensemble de ces reconstructions. Il en ressort qu'un très grande majorité des réseaux métaboliques à l'échelle génomique appartiennent au règne bactérien, et notamment aux protéo-bactéries. À l'inverse les eucaryotes et les archaea-bactéries sont sous représentées, avec de nombreux phylum totalement non représentés.

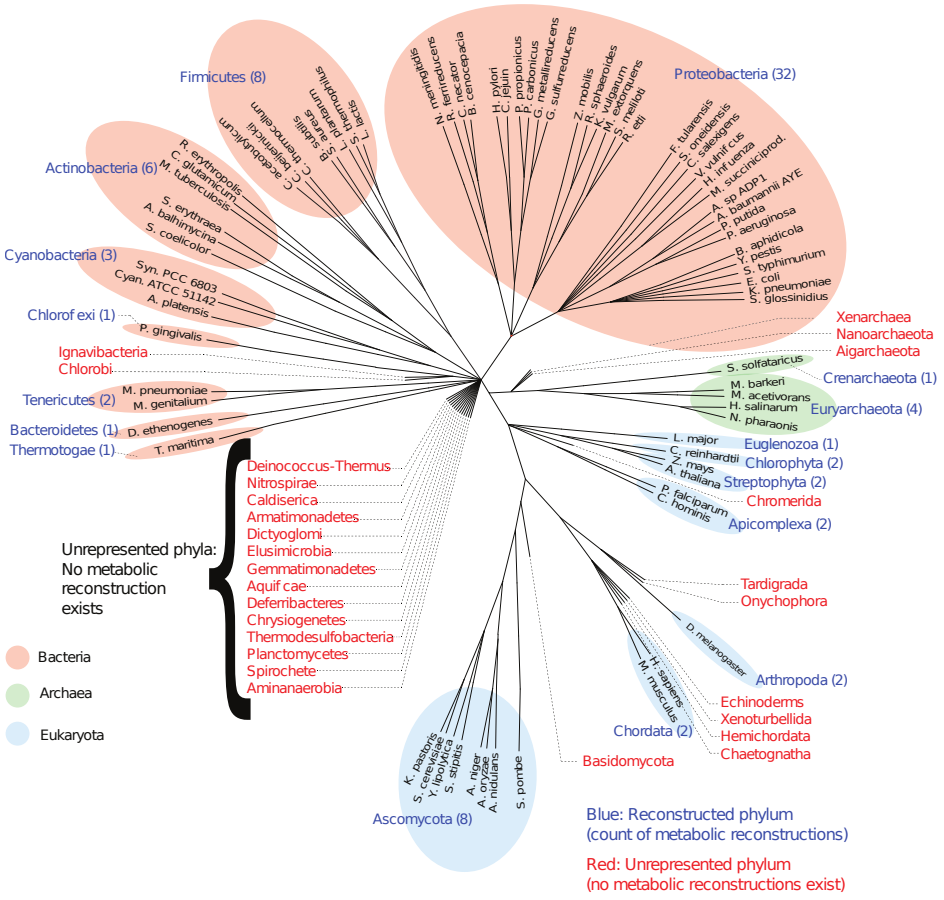


FIGURE 1.4: **Couverture phylogénétique des reconstructions de réseaux métaboliques en février 2013[MNP14].**

Pour tenter d' homogénéiser les reconstructions, Thiele et Palsson [TP10] ont décrit une méthodologie globale de reconstruction de réseaux métaboliques. La figure 1.5 tirée de cet article résume l'ensemble de ce processus, qui sera décrit plus précisément par la suite. Une analyse proposée par [HR14] reprend cette méthodologie en étudiant des outils de reconstruction de réseaux métaboliques bactériens existants et réalisant plus ou moins automati-

quement certaines des étapes de reconstruction.

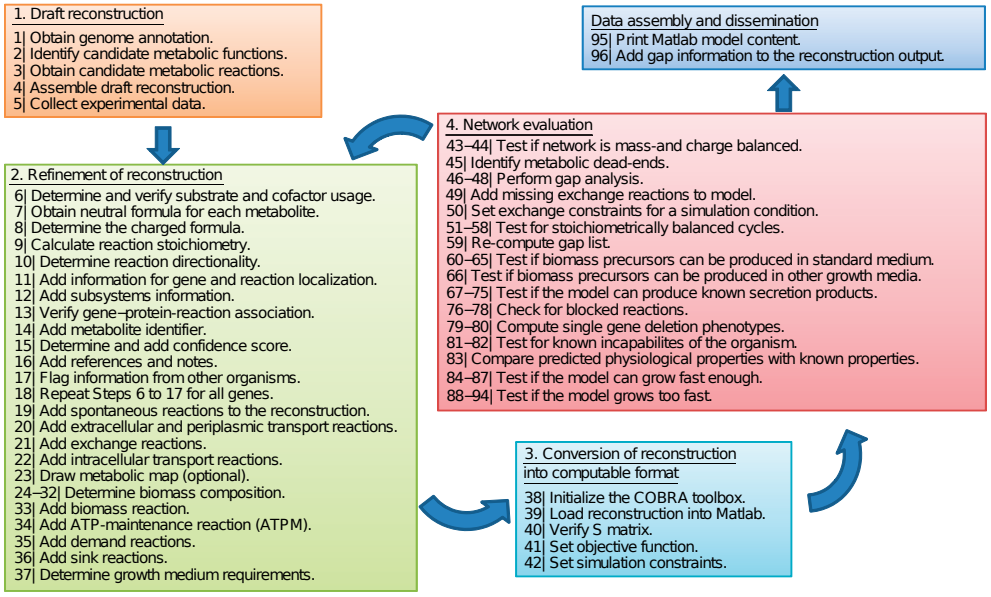


FIGURE 1.5: **Vue d'ensemble de la procédure de reconstruction de réseaux métaboliques [TP10].** Les étapes 2 à 4 seront itérées jusqu'à ce que les prédictions réalisées par le réseaux métaboliques soient en accord avec les observations biologiques.

L'analyse de cette référence permet de dégager quatre étapes majeures pour ce type de reconstruction :

1. La construction d'une ébauche métabolique
2. Une amélioration de cette ébauche
3. La conversion du réseau en un format analysable par des méthodes automatiques
4. L'évaluation de la qualité du réseau final

Ces quatre étapes, si elles ont été bien exécutées et ont conduit à un réseau de bonne qualité, peuvent être suivies de simulations et de prédictions. Nous allons maintenant décrire plus en détail chacune de ces étapes.

1.3.1 Construction d'une ébauche métabolique

La première étape de la reconstruction d'un réseau métabolique consiste en la construction d'une *ébauche* ou *esquisse* (en anglais, *draft*) de réseau métabolique. Cette ébauche correspond à un réseau de réactions métaboliques brut, qui est le plus souvent obtenu automatiquement par extraction d'informations dans les annotations d'un génome ou par l'étude de réseaux métaboliques d'espèces cousines. En particulier, certaines réactions que l'on sait être présentes, i.e. fonctionnelles, chez l'espèce étudiée pourront être absentes ou "erronées" dans cette ébauche du fait d'annotations manquantes (ou de mauvaises annotations) des gènes.

La réalisation de cette étape s'appuie principalement sur une annotation fonctionnelle précise du génome étudié, ou encore du transcriptome ou du protéome. Cette étape vise à proposer une fonction potentielle à chaque gène d'intérêt présent dans le génome.

Cela se fait généralement en utilisant différentes informations comme :

- Le numéro E.C.
- Les termes GO
- Des mots clefs
- Des noms génériques de réactions

Une fois que le génome est suffisamment annoté, il est, en théorie, possible de connaître précisément les réactions métaboliques associées à chaque annotation et donc à chaque gène connu. Les bases de données de réactions métaboliques telles que KEGG [KGS+14] ou MetaCyc [CAB+14] contiennent des informations détaillées sur ces réactions. Ces bases de données ont, en outre, le gros intérêt de posséder des identifiants internes cohérents à travers les différentes réactions. Ainsi, une molécule donnée (comme le β-D-glucose par exemple) possédera toujours le même identifiant ("GLC" pour MetaCyc, "C00221" pour KEGG) dans toutes les réactions produisant ou consommant cette molécule. Cette cohérence des identifiants permettra de construire des réseaux métaboliques où les différentes réactions seront interconnectées entre elles grâce à ces métabolites.

La qualité d'une ébauche métabolique, en particulier pour les espèces non modèles, va ainsi dépendre directement de l'annotation du génome. La qualité d'une annotation va, elle, dépendre de la façon dont elle a été réalisée. En effet, il est possible de distinguer deux classes d'annotations fonctionnelles : celles réalisées manuellement (ou validées manuellement) et celles ne reposant que sur une découverte automatique de la fonction des gènes. Les annotations manuelles, bien que moins nombreuses et plus longues à réaliser que les annotations automatiques, sont normalement de bien meilleure qualité. Ainsi plus un génome sera annoté manuellement, meilleure devrait être l'ébauche métabolique.

Différentes méthodes se basent en parallèle sur les connaissances sur le métabolisme d'espèces cousines déjà connues. Dans ces cas là, on construit des modèles de références en extrayant de l'information sur le métabolisme d'une sélection d'espèces. Différents scores d'alignement permettent, soit à priori, soit à posteriori, de vérifier si les différentes réactions sélectionnées doivent être intégrées au réseau de l'espèce ciblée. Cela a par exemple été utilisé lors de la reconstruction d'un réseau compartimenté et tenant compte de la spécificité des tissus chez *Arabidopsis thaliana* [MOMM+12].

Cependant, comme noté par Monk, Nogales et Palsson dans [MNP14], l'utilisation de réseaux métaboliques et d'annotations pour des espèces voisines perd de son efficacité au fur et à mesure que s'accroît la distance phylogénétique entre l'espèce cible et les espèces sur lesquelles les réseaux métaboliques dont on dispose sont de bonne qualité. Ce problème est d'autant plus important chez certains eucaryotes pour lesquels la couverture phylogénétique est faible, rendant la reconstruction de réseaux métaboliques chez ces phylum d'autant plus important (figure 1.6).

Lors de la reconstruction de réseaux métaboliques, nous sommes confrontés à des problèmes importants concernant l'hétérogénéité des données. Un gros travail d'intégration de ces données précède obligatoirement la reconstruction de réseaux.

Une fois la liste des réactions métaboliques à inclure dans un réseau établie, il convient de transformer cette liste en un fichier analysable par la plupart des méthodes informatiques. Le format préconisé pour les réseaux métaboliques est le format SBML (Systems

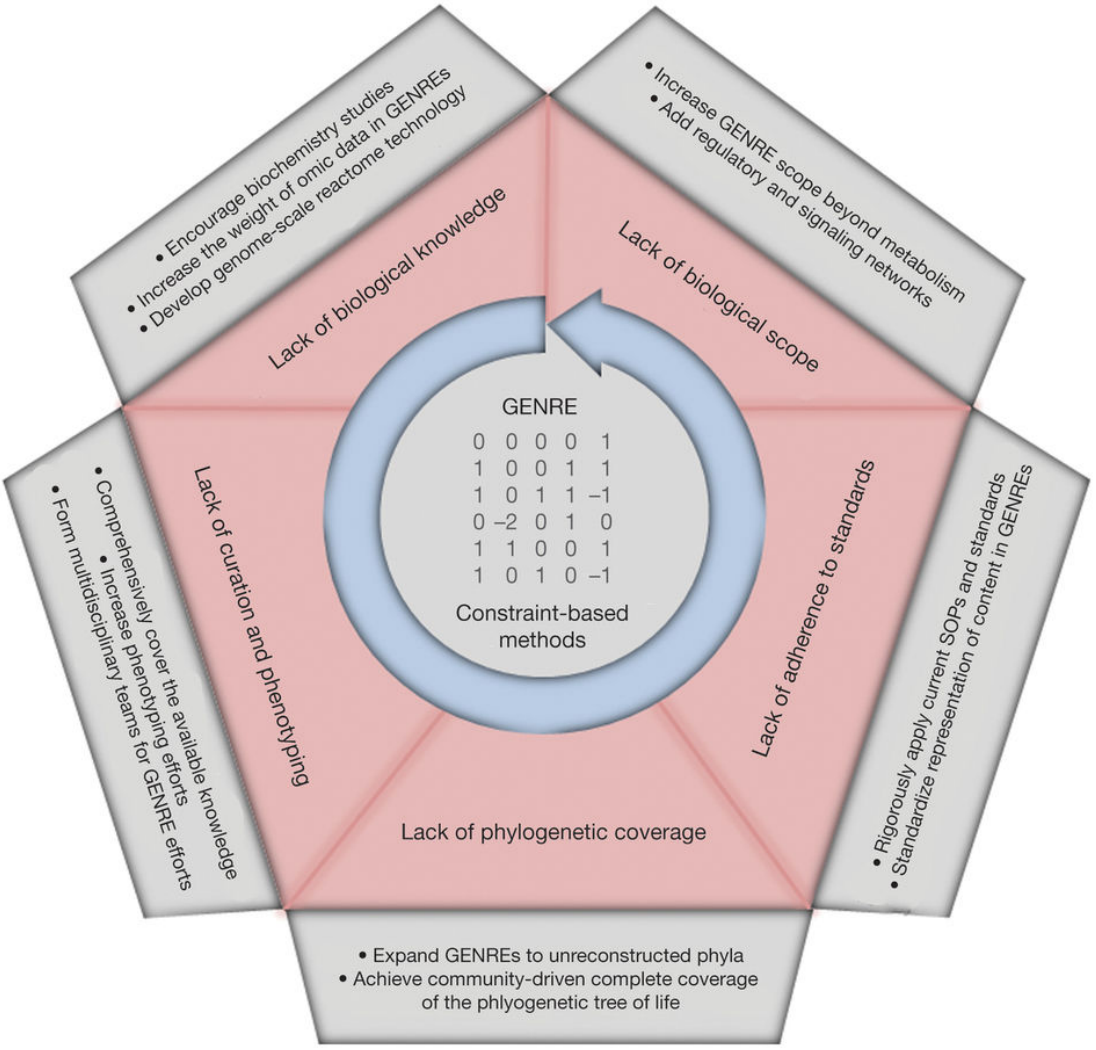


FIGURE 1.6: **Limitations actuelles au développement rapide de la reconstruction de réseaux métaboliques.** Les éventuels points à améliorer sont indiqués en gris, les méthodes à base de contraintes étant au centre des améliorations [MNP14].

Biology Markup Language). Cette norme SBML n'est malheureusement pas suivie à la lettre par grand nombre d'outils ce qui rend l'étude des réseaux métaboliques compliquée d'un point de vue "formatage des données". De plus de nombreuses versions de la norme SBML existent et ne sont pas forcément compatibles entre elles, rendant d'autant plus difficile l'utilisation d'outils plus ou moins récents. Dans le cas des espèces non-classiques, nous avons été particulièrement confronté à la modélisation des réactions réversibles dans différentes études de cette thèse. Les discussions sur cette question seront détaillées dans la section 2.2.3.

Par ailleurs, si les ébauches métaboliques sont créées par orthologie directe entre des enzymes connues chez des espèces modèles et un nouveau génome, il faut être capable de faire le lien entre une réaction associée à l'enzyme connue (pouvant provenir de KEGG, The SEED, ou les réseaux utilisés dans CobraToolBox) et une réaction dans les formats ciblés (MetaCyc par exemple). Ces liens entre bases de données de réactions sont extrêmement problématiques et sources d'imprécisions. Il existe cependant des méthodes pour établir du lien entre elles. Par exemple la base de données MetaCyc donne parfois des liens vers les réactions KEGG correspondantes. De plus, les numéros E.C. (lorsqu'ils sont précis) permettent de relier entre elles les différentes réactions, un numéro E.C. précis ne pouvant en théorie correspondre qu'à une seule réaction. Dans cette direction, différentes approches proposent de réconcilier les réactions métaboliques en fonction de différents critères et annotations, mais aucune n'est exhaustive. Différents outils existent déjà pour réaliser de telles réconciliations entre les différentes bases de données. Nous pouvons par exemple citer BridgeDb [vIPK+10], MetaMerge [CSH+12] ou encore MNXref [BBM+14]. L'outil MetaMerge, par exemple, devrait permettre de fusionner des réseaux métaboliques créés pour une même espèce à partir de différentes méthodes et bases de données. D'autre part MNXref est un outil permettant de faire le lien entre les différentes bases de données de réactions métaboliques telles que MetaCyc, KEGG et beaucoup d'autres. Tout ces outils se basent sur différentes informations pour réaliser l'homogénéisation entre les différentes bases de données telles que la nomenclature chimique des réactifs des réactions, des références croisées entre différentes bases de données, des noms ou synonymes, etc.

Outre la nature des données rencontrées, leur qualité peut également être très hétérogène. Les annotations, par exemple, sont à considérer avec précaution, notamment car certaines d'entre elles ne sont pas complètes (ex. : absence du numéro E.C.) ou parce que certains gènes ne sont pas annotés du tout. Ces situations représentent une source importante de faux négatifs et de faux positifs lors de la création de l'ébauche métabolique. Inversement, un gène entièrement annoté manuellement et possédant l'ensemble des éléments nécessaires pour une recherche automatique de la réaction associée sera extrêmement fiable.

Il faut également faire attention au taux élevé de faux positifs générés par les recherches d'homologues à partir de profils HMMs (Hidden Markov Models, ou modèles de Markov cachés). Certains des profils sont en effet très génériques du fait de l'existence de certaines enzymes chez des organismes présents à travers l'ensemble de l'arbre du vivant, ce qui engendre parfois des signatures floues et donc des faux positifs lors de la recherche d'homologues. L'absence de cohérence entre les différents outils d'alignements [RSS01, LSR03] peut-être source d'erreur et nécessite de définir des scores de réconciliations tels que décrits, par exemple, dans [LDNS12].

1.3.2 Raffinement de l'ébauche métabolique

Le résultat de l'étape précédente est un ensemble de réactions qui sont reliées entre elles par les métabolites qu'elles partagent. Ces réactions ne correspondent pas forcément à l'ensemble de celles ayant lieu dans l'espèce étudiée du fait des erreurs et des manques dans l'annotation. Il faut donc une seconde étape pour améliorer la qualité du réseau, qui consiste à rajouter les réactions que l'on aurait pu manquer lors de la création de l'ébauche métabolique ainsi que les gènes qui pourraient y être associés.

Pour réaliser cette étape, des connaissances biologiques qui ne seraient pas expliquées par le réseau actuel sont nécessaires. Pour cela, il est possible d'utiliser l'ensemble des métabolites dont la présence dans l'organisme d'intérêt a été prouvé expérimentalement, notamment par du profilage métabolique. En effet, il apparaît évident que le réseau métabolique correspondant à cet organisme doit contenir les voies complètes nécessaire à la production de ces molécules. Nous nous concentrerons cependant ici sur l'exploitation de profils métaboliques, puisqu'il s'agit de la base de la plupart des méthodes de reconstruction de réseaux métaboliques.

Pour étudier leur productibilité, il faut connaître le milieu de croissance dans lequel l'organisme d'intérêt est cultivé. En effet, les cellules d'un organisme vont récupérer, dans ce milieu, tous les nutriments nécessaires à la croissance, et donc à la production de biomasse. Bien évidemment, plus l'information sur ce milieu de culture est grande, meilleure sera la reconstruction. Le cas parfait correspond à un organisme pouvant croître à différentes vitesses dans différentes conditions, et pour lequel la vitesse de consommation des différentes molécules du milieu de culture est connue. Si l'organisme n'est pas cultivable en laboratoire, il faudra estimer les conditions nécessaires à la croissance à partir de la bibliographie ou de toute autre source d'informations.

Une fois le milieu de culture parfaitement décrit, nous allons pouvoir étudier par différentes méthodes si ce milieu de culture, associé à l'ébauche métabolique, permet de prédire la production des molécules identifiées expérimentalement. Cette prédiction peut se faire par différents types de simulations, que l'on peut classer globalement en deux groupes : les simulations topologiques et les simulations numériques. Les simulations topologiques consistent en une recherche de chemins dans le graphe métabolique partant d'un set de nœuds "molécules" correspondant au milieu de croissance de l'organisme et allant jusqu'à l'ensemble des nœuds correspondant aux métabolites identifiés comme productibles par l'espèce étudiée [RK01]. Les simulations numériques consistent le plus souvent en une analyse d'un système de contraintes linéaires dictées par la structure et la stœchiométrie du réseau métabolique, comme c'est le cas avec la FBA [FS86]. Elles nécessitent plus de connaissances biologiques qu'une recherche topologique, notamment sur la quantité de molécules présentes dans la biomasse et la stœchiométrie précise des réactions, afin de permettre une optimisation des flux de molécules passant à travers les réactions pour que le résultat théorique corresponde aux connaissances biologiques. Un des apports de cette thèse consistera précisément à étudier l'impact de ces différentes modélisation du concept de productibilité sur la qualité de la reconstruction (chapitre 2).

Si toutes les molécules sont productibles à partir de l'ébauche métabolique et de la composition du milieu de culture, c'est un signe que la reconstruction initiale à partir des annotations a été particulièrement efficace. Étant donné les erreurs d'annotations, de séquençage et les spécificités de chaque espèce, ce cas est extrêmement rare. Pour produire les mé-

tabolites identifiés expérimentalement, il a toujours été nécessaire de rajouter des réactions au réseau. Cette addition peut être manuelle [FMR+12] quand des experts de l'organisme connaissent parfaitement le métabolisme de celui-ci. Une complétion manuelle peut cependant être extrêmement longue et risque de n'apporter au réseau que des informations déjà connues par les biologistes.

La recherche des réactions à ajouter dans le réseau peut également être réalisée automatiquement. De très nombreuses méthodes existent pour effectuer cette étape mais reposent presque toutes sur le même principe de parcimonie, c'est-à-dire d'inclure le minimum possible de réactions dans le réseau. Pour choisir ces réactions, la plupart des méthodes ne donnent pas une solution unique mais un ensemble de solutions possibles. Ces solutions sont ensuite classifiées selon des scores généralement basés sur des données d'orthologie ou de phylogénie. Les approches de complétion se ramènent à des questions d'optimisation que nous détaillerons dans le chapitre 2.

1.3.3 Évaluation du réseau final

Après cette étape, le réseau métabolique obtenu doit être de qualité suffisante pour pouvoir réaliser différentes prédictions.

La complétion du réseau métabolique va produire une liste de réactions susceptibles d'être ajoutées au réseau. Ces réactions étant en très grande majorité catalysées par des enzymes, il convient de rechercher si les enzymes en question sont présentes chez l'espèce étudiée. Pour cela, il faut rechercher d'éventuels gènes qui pourraient coder pour ces enzymes dans le génome. Cette recherche peut se faire en utilisant différentes techniques telles que la recherche d'orthologues [LSR03], la recherche de signatures de familles de protéines [FBC+14], etc.

L'évaluation du réseau final se fait habituellement au niveau fonctionnel en utilisant des méthodes de Flux Balance Analysis (FBA). Si des données précises de quantification de certaines molécules sont disponibles, il est assez aisé de créer une fonction objectif et de tester la croissance virtuelle de l'organisme en utilisant des algorithmes d'optimisation linéaire. Si ces données ne sont pas disponibles, il est toujours possible de reconstruire une fonction objective "virtuelle" en se basant sur celles existantes chez d'autres espèces, si possible proches au niveau phylogénétique et/ou au niveau physiologique.

Si différentes données de croissance obtenues au cours de plusieurs conditions de culture sont disponibles, la FBA pourra être appliquée pour chaque condition et les résultats pourront être comparés avec la croissance mesurée expérimentalement. Si les résultats sont corrélés entre les analyses de simulation et les observations biologiques, cela suffit habituellement à valider la qualité globale d'un réseau métabolique.

1.3.4 Pipelines existants

Plusieurs logiciels ou groupes de logiciels permettent de réaliser une ou plusieurs des étapes de reconstruction citées précédemment. Comme indiqué par Hamilton et Reed en 2014 [HR14], la reconstruction automatique d'une ébauche est maintenant proposée de manière automatique par la plupart des plate-formes logicielles. La plupart de ces méthodes nécessitent tout de même une part de curation manuelle, souvent guidée par différentes approches. En revanche, l'ensemble des analyses relatives à la fonctionnalité du système sont

généralement laissées à la charge de l'utilisateur, avec un accompagnement limité, et peu d'automatisation. Il faut cependant noter que les principaux pipelines ici mentionnés sont plutôt des canevas sur lesquels se positionnent différents outils qui permettent d'améliorer l'efficacité et/ou l'automatisation des méthodes dans des contextes applicatifs particulier.

Dans leur article, Hamilton et Reed [HR14] étudient quatre outils de reconstruction globale de réseaux métaboliques microbiens : Subliminal [SSM+11b], Model SEED [HDB+10], Raven [ALS+13] et Pathway Tools [KPK+10]. La figure 1.7 résume les points positifs et négatifs de ces outils. Il est également possible d'étendre cette étude aux organismes non microbiens en citant MetaNetter [JBBG08] qui permet d'inférer des réseaux métaboliques à partir de données de métabolomique, en se basant sur la différence de masse entre les différents métabolites pour inférer des réactions (comme la disparition d'une molécule de CO_2 par exemple). Cet outil et ceux associés mettent particulièrement en avant la visualisation des réseaux métaboliques après reconstruction. Une autre plateforme de reconstruction automatique de réseaux métaboliques microbien existe, MicroScope [VBC+13]. Cette plateforme permet de reconstruire et d'étudier un réseau métabolique bactérien à partir d'un génome brut, en utilisant notamment la structure particulière de ces génomes et en accompagnant l'utilisateur tout le long du processus.

L'analyse de l'ensemble de ces plate-formes fait ressortir un manque dans la reconstruction automatique ou semi-automatique de réseaux métaboliques eucaryotes à partir de données génétiques.

1.4 Complétion de réseaux : problèmes d'optimisation induits

Les différentes problématiques informatiques relatives à la reconstruction d'un réseau métabolique à partir de données brutes, relèvent finalement de trois domaines différents : il y a d'abord différentes questions autour de la représentation des connaissances pour homogénéiser et réconcilier les données provenant de différentes sources, il y a ensuite des questions complexes en optimisation entière, relative à la fonctionnalité des réseaux ou des réactions. Enfin, la question de la complétion des réseaux métaboliques a été abordée avec différentes méthodes et algorithmes. Dans cette partie, nous allons détailler les principales approches existantes concernant ce troisième point et discuter leurs limites.

1.4.1 Panorama général

La principale difficulté dans cette étude bibliographique réside dans le fait que les différentes approches ne résolvent pas toujours le même problème. Dans tous les cas, il s'agit de construire un réseau métabolique consistant, supporté par des observations biologiques en partant d'un réseau brut incomplet et de connaissances additionnelles. Souvent, la consistance avec les observations biologiques réside dans le fait que, lorsque l'on connaît la composition du milieu de culture d'un organisme et que l'on a des données de métabolomique, le réseau doit être en mesure de trouver un moyen de produire les métabolites identifiés à partir des nutriments du milieu de culture. Cependant, la notion de productibilité peut varier en fonction des approches (quantitative, qualitative, globale ou probabiliste...). La consistance peut aussi être mesurée à l'aide de données d'expression, de protéomique, ou à l'aide de critères topologiques tel que réalisé lors de la création d'un réseau métabolqie compartimentalisé d'*Arabidopsis thaliana* [MOMM+12].

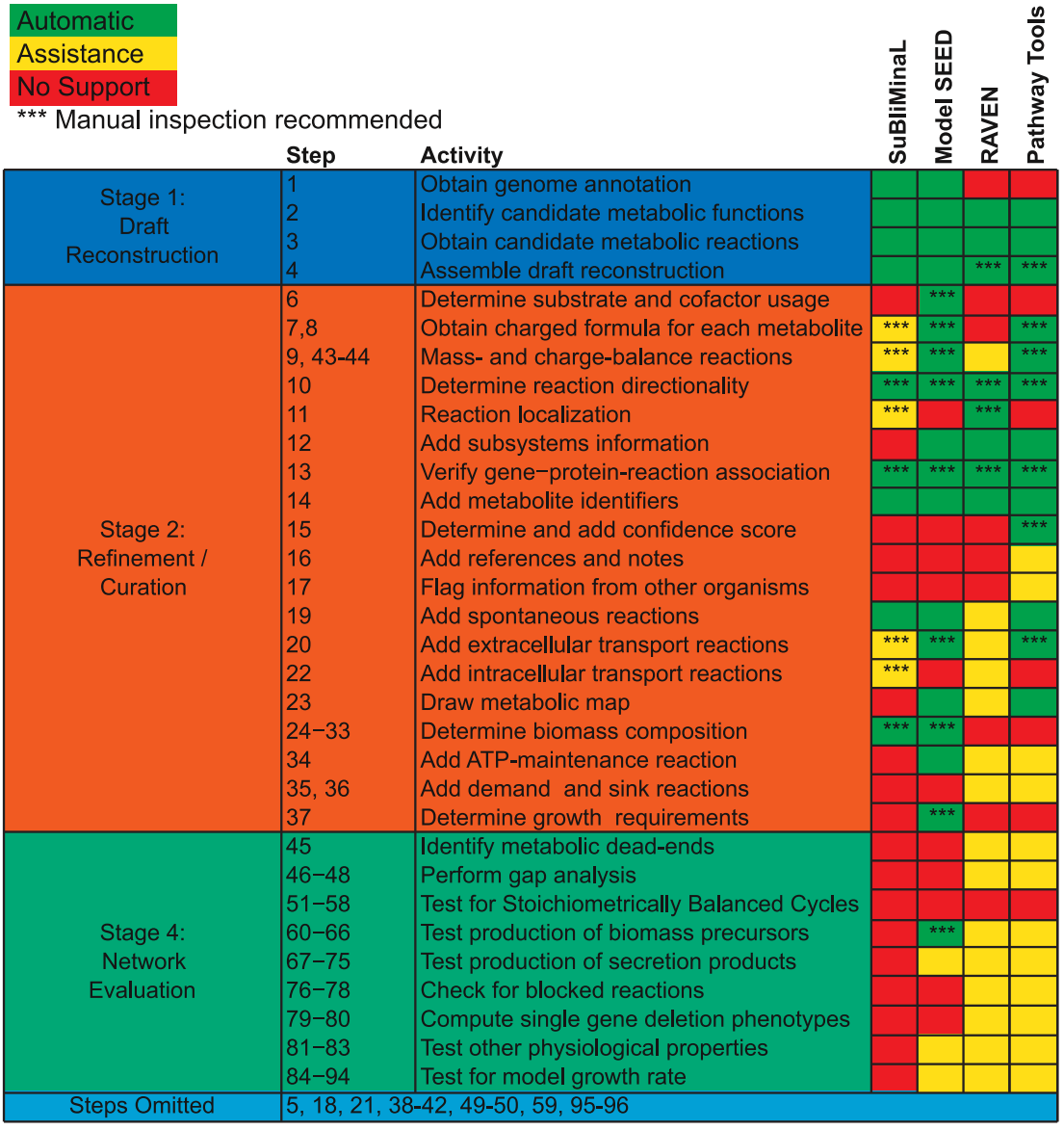


FIGURE 1.7: **Support apporté par les différents outils de reconstruction de réseaux métaboliques, de la création de l'ébauche jusqu'à l'évaluation de la qualité du réseau [HR14].** Les quatre étapes principales et certains des 96 points relevés par [TP10] sont étudiés.

Si on prend du recul, les méthodes de complétion de réseaux métaboliques se basent habituellement sur deux stratégies différentes :

- une approche descendante (dite top-down), qui consistera à ajouter toutes les informations disponibles à un instant donné, qu'elles soient intéressantes ou non, avant de supprimer petit à petit les données les moins relevantes. On citera ici la méthode OptStrain [PBM04] et sa dérivation SMILEY [RPC+06] qui, à l'aide de programmation MILP, a été utilisée pour reconstruire par exemple un réseau métabolique d'une souche d'*Escherichia coli* [HJFP06] et une approche stochastique [CMK+09] utilisée pour reconstruire un réseau métabolique de *Chlamydomonas reinhardtii* [MWK+08] par exemple ,
- une approche ascendante (dite bottom-up), qui partira uniquement de l'ébauche métabolique avant de rajouter progressivement des informations (et donc des réactions) afin d'obtenir au final un réseau de bonne qualité. C'est en particulier l'approche choisie dans la méthode GapFind [SKDM07] et la méthode combinatoire Network-expansion [ST09].

Les algorithmes sous-jacents aux différents méthodes ne sont pas toujours explicités, et leur complexité peu étudiée. Cependant, Nikolski et ses collaborateurs ont montré dans un cadre assez générique que l'extraction d'un sous-réseau satisfaisant un score minimal dans un réseau métabolique est NP-difficile, en se ramenant à un problème de Weighted Tree Packing [NGMS08]. On peut aussi noter l'ensemble des travaux de Acuna et Sagot autour de la complexité de différents problèmes relatifs à l'identification de sources dans un réseau métabolique [CMA+08].

1.4.2 Optstrain, SMILEY - top-down & MILP

OptStrain [PBM04] est une méthode impliquant quatre étapes afin de rajouter des réactions provenant d'une base de données de réactions métaboliques à l'intérieur d'un réseau pré-existant.

- La première étape consiste à télécharger et nettoyer la base de données de réactions métaboliques. Les erreurs de stœchiométrie relativement importantes pouvant impacter grandement les résultats de l'approche, toutes les réactions mal équilibrées ou impliquant au moins un élément donc la composition n'est pas précise sont supprimées de la base de données,
- La seconde étape consiste à ajouter l'ensemble de la base de données au réseau pré-existant. Cela permet d'obtenir un "méta-réseau" grâce auquel il est possible d'étudier, à partir d'un ensemble de substrats, le rendement maximum possible pour la production d'une cible donnée. Cette étape ne prend pas du tout en compte l'origine des réactions (base de données ou organisme étudié) utilisées pour maximiser la production de la cible. Cette maximisation correspond à un problème de programmation linéaire,
- Une fois le rendement maximal identifié, la troisième étape cherche à minimiser le nombre de réactions provenant de l'extérieur en maintenant ce rendement. Ce problème est modélisé et résolu par programmation linéaire mixte en nombres entiers (MILP),
- Enfin la dernière étape va consister à faire en sorte de ne garder dans le réseau que les réactions nécessaires pour maximiser le rendement de production de la cible étu-

diée. L'ajout des réactions n'est pas automatique pour autant mais utilise l'algorithme OptKnock [BPM03, PBM03] pour éviter de découpler la production de biomasse et celle du métabolite d'intérêt qui risquerait de conduire à une mort de l'organisme d'intérêt.

Aucune étude particulière ne semble avoir été effectuée pour tester la rapidité ou la qualité de cette méthode. Il est uniquement indiqué que les optimisations prennent quelques heures et que certaines prédictions réalisées sur une espèce bactérienne donnée semblent cohérentes avec des résultats biologiques. Cette méthode a été très utilisée par la communauté. Nous pouvons notamment citer son utilisation pour étudier la production de succinate par *Mannheimia succiniciproducens* [LLK05] ou encore la production de lactate par *Escherichia coli* [HJFP06].

Une extension de cette approche est la méthode SMILEY [RPC+06], qui donne la possibilité d'inclure des réactions de transport permettant d'excréter n'importe quel produit. Cette approche a en particulier été utilisée pour la reconstruction de réseaux métaboliques de *Ostreoccocus* [KYW+12].

1.4.3 GapFind, GapFill - bottom-up & MILP

En 2007, Kumar et ses collaborateurs [SKDM07] ont développé une procédure d'optimisation linéaire pour identifier des trous dans les réseaux métaboliques, et proposer des solutions afin de combler ces trous. Pour cela, ils commencent par identifier les métabolites ne pouvant pas être produits dans un réseau métabolique avant de restaurer la connectivité du réseau en modifiant celui-ci ou en rajoutant des réactions à l'intérieur du modèle.

GapFind va ainsi permettre d'identifier des trous dans les réseaux métaboliques. Ceux-ci se manifestent par le fait qu'il existe des métabolites qui ne sont pas produits ou consommés par le réseau. À l'état stable, aucun flux ne pourra alors passer à travers ces métabolites ce qui indique une erreur dans le réseau. De plus, si un ou plusieurs métabolites sont bloqués, cela entraînera très souvent une cascade de réactions bloquées. Par exemple, un métabolite non consommé bloquera les flux dans toutes les réactions en amont de celui-ci à l'état stable.

Les métabolites non produits peuvent être identifiés en scannant la matrice stœchiométrique du réseau. En effet, il suffit de chercher toutes les colonnes contenant des valeurs non positives (pour les réactions irréversibles) ou non nulles (pour les réactions réversibles). Pour les métabolites non consommés, on applique la même méthode que pour les réactions réversibles et on cherche des valeurs non négatives pour les réactions irréversibles. Cela permet d'identifier les "racines" des voies métaboliques bloquées par un métabolite non produit ou non consommé. Pour trouver les métabolites bloqués respectivement en amont et en aval de ceux-ci, une simple inspection de la matrice stœchiométrique ne suffit pas. Les auteurs ont alors mis en place un ensemble de contraintes permettant, pour chaque métabolite, de connaître si celui-ci est bien produit et consommé. Les contraintes sont telles que :

- Pour chaque réaction irréversible produisant un métabolite i, la réaction produit une quantité ϵ de i supérieure à zéro,
- Pour chaque réaction réversible qui inclut un métabolite i, le métabolite est considéré comme produit uniquement si la réaction produit une quantité ϵ de i supérieure à zéro,

- Un métabolite i doit posséder au moins une voie de production pour être considéré comme productible, c'est-à-dire qu'il doit être le produit d'au moins une réaction active,
- Les flux passant dans chaque réaction doivent être compris entre des bornes supérieures et inférieures imposées par le modèle,
- Étant à l'état stable, les cellules continuent toujours à se diviser, il faut donc produire un minimum de métabolites. Ils fixent dont $\sum_{j \in M} S_{ij} v_j \geq 0 \ \forall i = 1...N$,
- Enfin, la fonction objective va maximiser le nombre de métabolites productibles.

L'ensemble de ces contraintes fait en sorte de maximiser le nombre de métabolites considérés comme productibles. S'il reste des métabolites non productibles, c'est la preuve d'un manque dans le réseau métabolique.

Une fois les trous identifiés, il convient de les combler. Pour cela les auteurs ont développé GapFill afin de rechercher si inverser le sens des réactions, rajouter des réactions provenant de bases de données, ou rajouter des réactions de transport permettent de combler les trous du réseau et rendre l'ensemble des métabolites productibles et consommés. Il s'agit là encore d'un problème d'optimisation linéaire en nombres entiers ayant comme objectif de minimiser le nombre de réactions provenant de bases de données ajoutées dans le réseau initial. Les différentes contraintes à respecter sont :

- Pour chaque réaction réversible qui inclut un métabolite i, le métabolite est considéré comme produit uniquement si la réaction produit une quantité ϵ de i supérieure à zéro,
- Les additions de réactions doivent conduire à une production minimale des métabolites jusqu'à présent non productibles,
- Les flux passant dans chaque réaction doivent être compris entre des bornes supérieures et inférieures imposées par le modèle,
- Les réactions ajoutées au réseau métaboliques doivent toutes avoir des flux non nuls passant à travers elles.

L'ensemble du processus est réalisé pour trouver le minimum de réactions à ajouter pour pouvoir produire et/ou consommer chaque métabolite cible du réseau.

L'ensemble des contraintes définissant GapFind et GapFill est extrêmement intéressant et produit des résultats biologiquement valides. Ainsi, GapFind/GapFill ont été utilisés intensivement par l'équipe ayant créé la méthode. En dehors de cette équipe, nous pouvons citer son utilisation pour la reconstruction du réseau métabolique de *Gordonia alkanivorans* [AKRI13], du pathogène *Salmonella Typhimurium LT2* [THS+11] et d'une souche de *Clostridium beijerinckii* produisant du butanol [MER+11].

On peut cependant apporter quelques critiques à cette méthode. La première est que, si l'encodage des contraintes est disponible librement, celui-ci s'applique uniquement au solveur GAMS qui, lui, est payant et relativement cher. Cela empêche toute comparaison neutre entre cet outil et les outils existants. De plus, à ma connaissance, aucun test n'a été effectué pour connaître le comportement de la méthode face à un accroissement de la taille des jeux de données. Il est donc difficile de savoir si ce logiciel fonctionne dans un contexte où la taille des bases de données et la complexité des réseaux augmentent sans cesse.

Il est également possible de se poser la question de la conséquence que pourrait avoir l'introduction d'un faux positif au niveau des réactions lors de la première reconstruction du réseau. Une seule réaction consommant deux molécules et en produisant deux autres, si

celle-ci est ajoutée "loin" de toutes les voies métaboliques existantes, amènerait la reconstruction d'une longue voie métabolique pour raccorder l'ensemble des quatre métabolites à l'ensemble du réseau. De même, si la technique initiale semble limitée à la recherche d'une solution unique, des techniques d'optimisation linéaire pourraient être appliquées pour obtenir l'ensemble des solutions existantes, mais là encore nous pouvons nous interroger sur l'utilisabilité en pratique de ces méthodes.

Enfin on peut s'interroger sur la contrainte $\sum_{j\in M} S_{ij} v_j \geq 0 \; \forall i = 1...N$ à l'état stable. Cette contrainte indique que le réseau (étant à l'état stable) peut produire une quantité illimitée de tous les métabolites, ce qui relâche énormément les contraintes de modélisation classiques où, habituellement, nous avons $\sum_{j\in M} S_{ij} v_j = 0$ qui force le système à ne pas créer de matière.

1.4.4 Christian et al - top-down & stochastique

Nils Christian et ses collaborateurs proposent une méthode stochastique de complétion de réseaux métaboliques[CMK+09]. Cette méthode est basée sur une approche "top-down", commençant par une complétion très large qui sera ensuite précisée au fur et à mesure des différentes itérations.

Le cœur de la méthode consiste à identifier des groupes minimaux de réactions (appelées "extensions") permettant de compléter le réseau métabolique pour produire des métabolites sélectionnés à partir des nutriments du milieu de culture.

- L'algorithme consiste, dans un premier temps, à ajouter l'ensemble des réactions contenues dans une base de données de réactions métaboliques au réseau incomplet,
- Une fois cette addition réalisée, il convient de regarder l'ensemble des métabolites d'intérêt productibles à partir du réseau incomplet auquel on a ajouté la base de données. Ces métabolites d'intérêt formeront alors la base de travail autour de laquelle la méthode fonctionnera,
- Lors d'un processus itératif, à chaque pas de temps, une réaction (provenant de la base de données) tirée aléatoirement sera enlevée du réseau. Si le réseau garde sa fonctionnalité (c'est à dire si les métabolites d'intérêts sont toujours productibles à partir des nutriments), cette réaction est définitivement retirée de l'extension. Si le fait de retirer cette réaction supprime la fonctionnalité du réseau, celle-ci semble être indispensable à cette fonctionnalité et sera donc remise dans l'extension et ne sera plus testée avant la fin du déroulement de la méthode. La fonctionnalité utilisée par les auteurs est topologique mais, hormis les temps de calcul, rien ne semble s'opposer à une étude quantitative de la fonctionnalité,
- Puisque les extensions obtenues vont dépendre énormément de l'ordre dans lequel les réactions sont tirées, les auteurs recommandent de réaliser un grand nombre de tirages aléatoires de l'ordre de test des réactions,
- Le résultat de cette méthode correspondra à un ensemble d'extensions possibles. Les réactions présentes dans les extensions vont être analysées d'un point de vue statistique pour classer les réactions par ordre d'importance afin de pouvoir les étudier manuellement par la suite.

Finalement, cette approche consiste donc à explorer l'espace de toutes les complétions fonctionnelles qui sont minimales au sens ensembliste du terme (si on enlève une réaction, la fonctionnalité globale du système n'est plus vérifiée). On peut cependant regretter qu'au-

cune estimation de la taille de cette espace ne soit proposée, ce qui ne permet pas de décider si le nombre d'itérations est suffisant pour l'explorer de manière fiable.

De plus cette méthode ne propose pas une "bonne" solution. En effet, rien n'assure qu'ajouter l'ensemble des réactions revenant souvent dans les simulations suffit au final à compléter le réseau pour le rendre fonctionnel. Imaginons par exemple un cas où dix réactions différentes permettraient de compléter le même "trou" dans un réseau. D'après l'aléa créé par la méthode, chaque réaction reviendrait dans 1/10ème des simulations, ces réactions prises une par une auraient alors un score final faible et ne seraient probablement pas inclues automatiquement dans le réseau final. Celui-ci perdrait ainsi sa fonctionnalité, ce qui est pourtant la raison initiale du développement de cette méthode.

Pour essayer de limiter ce biais, les auteurs associent un "score biologique" aux réactions et biaisent plus ou moins fortement le tirage aléatoire de l'ordre dans lequel les réactions sont étudiées. Plus le score sera fort, moins une réaction aura de chance d'être tirée "tôt". Ce score peut se baser sur différentes informations telles que des scores de similarité de séquence entre des enzymes connues pour catalyser les réactions présentes dans les bases de données et les protéines produites par le génome de l'espèce d'intérêt. Puisque plus une réaction est testée tôt, plus elle a de chances de ne pas être présente dans l'extension au final, ces classifications des listes de réactions permet d'améliorer grandement les capacités de prédiction de la méthode.

Cette approche a notamment été utilisée pour améliorer les réseaux métaboliques des algues vertes *Chlamydomonas reinhardtii* [MWK+08] et *Ostreococcus* [KYW+11]. Il est également possible de citer son utilisation pour étudier les relations métaboliques existantes entre des plantes et des phytopathogènes [DCS+13]. On peut regretter que les implémentations des algorithmes ne soient pas mises à disposition. Une ré-implémentation de l'algorithme au laboratoire (durant un stage) a cependant suggéré qu'il ne passait pas à l'échelle quand la taille de la base de données de référence augmente.

1.4.5 Network-expansion - bottom-up & optimisation combinatoire

En 2009, Torsten Schaub et Sven Thiele proposent une méthode [ST09] permettant de compléter des réseaux métaboliques d'un point de vue qualitatif en s'affranchissant de la nécessité d'obtention de données cinétiques. Par rapport aux approches précédentes, l'idée est de vérifier la fonctionnalité du réseau d'un point de vue qualitatif (accessibilité des métabolites) sans prendre en compte les contraintes induites par la stoechiométrie du système qui pourront être vérifiées dans un second temps. En reformulant ce problème de manière combinatoire, Schaub et Thiele proposent d'utiliser des méthodes d'optimisation récentes pour énumérer intégralement l'espace des solutions du problème.

Comme pour les précédentes approches, les auteurs sont partis du constat que la grande majorité des réseaux métaboliques reconstruits automatiquement sont incomplets. De plus, les bases de données de réactions métaboliques sont de plus en plus exhaustives et leur utilisation n'en devient que plus pertinente. Enfin, il y a à disposition des données biologiques pour la plupart des organismes biologiques pour lesquels un réseau métabolique est reconstruit ou en cours de reconstruction, et qui sont utilisables pour l'étape de complétion.

L'idée générale des auteurs est qu'une réaction ne peut avoir lieu uniquement que si les réactants de celle-ci sont présents, soit dans un milieu de culture donné, soit comme produit d'une autre réaction. Ainsi, en partant des métabolites présents dans un milieu de

culture (les métabolites graines) il devient possible de déterminer le "scope" des graines, c'est à dire l'ensemble des métabolites productibles à partir des graines et des réactions présentes dans un modèle.

L'approche s'appuie ainsi sur des données expérimentales comme la présence de certains métabolites dans une cellule (métabolites cibles). La méthode regarde si ces cibles font partie du scope des graines. Si c'est le cas, le réseau métabolique peut être considéré comme suffisamment complet pour expliquer l'existence de ces cibles. Si ce n'est pas le cas, il conviendra d'ajouter des réactions dans le réseau pour permettre aux cibles de faire partie du scope des graines. Ajouter l'ensemble des réactions provenant d'une base de données de réactions comme MetaCyc permet (généralement) de produire l'ensemble des cibles. Cependant, toute capacité d'étude biologique ultérieure du réseau sera perdue du fait de la trop grande quantité de faux-positifs. Les auteurs proposent donc d'utiliser un principe de parcimonie pour minimiser le nombre total de réactions à ajouter au réseau, et ainsi minimiser les modifications faites à un réseau que l'on considère initialement de bonne qualité. Ce problème est par nature extrêmement combinatoire et donne de très nombreuses solutions.

Contrairement aux approches précédentes, cette modélisation n'a pas fait l'objet d'une résolution algorithmique ou en programmation entière, mais d'une résolution déclarative. En effet, les auteurs ont modélisé les concepts de scope et d'accessibilité dans un langage déclaratif puis utilisé des techniques d'optimisation combinatoire particulièrement efficaces pour résoudre ce problème. Plus précisément, la modélisation et la résolution du problème sont basées sur des technologies apparues relativement récemment, dite de programmation par ensemble réponse (ou Answer Set Programming, ASP) [Bar03, GKKS12], qui sont connues pour être particulièrement efficaces pour résoudre des problèmes NP-dur. La programmation par ensembles réponses propose un cadre déclaratif pour modéliser des problèmes combinatoires. Le caractère déclaratif et la haute performance des solveurs ASP actuels permettent de se concentrer sur les problèmes de modélisation plutôt que de rechercher des façons intelligentes de traiter ces problèmes. L'idée de base d'ASP est d'exprimer un problème sous forme logique de manière à ce que les modèles sortant de cette représentation donnent les solutions au problème initial. Les problèmes sont exprimés en tant que programmes logiques et les modèles résultants sont appelés "ensembles réponses" (ou Answer Sets). Ces ensembles-réponses sont identifiées à l'aide de technologies inspirées par les approches SAT et les bases de données. Il est bien évidemment possible de déterminer, à l'aide d'ASP, si un programme possède un ensemble réponse, mais d'autres modes de raisonnement sont nécessaires pour couvrir l'ensemble des problèmes que l'on peut rencontrer en pratique. Ainsi il sera possible d'identifier aisément l'intersection ou l'union de l'ensemble des ensembles réponses. De la même manière, lister l'intégralité des ensembles réponses pourra être possible.

Dans leur article, Schaub et Thiele proposent une implémentation de cette approche en ASP, en utilisant la suite Potassco [GKK+11]. L'implémentation est détaillée en annexe B. Ils ont ainsi montré sur différents exemples tirés de dégradations d'un réseau métabolique d'*Escherichia coli* que leur implémentation en ASP permet de compléter un réseau en un temps raisonnable tant que les dégradations de réseaux et la taille de la base de données utilisée pour compléter n'est pas trop grande. En revanche dès que la taille de la base de données de référence devient trop grande l'outil n'arrive plus à calculer de complétions, laissant de la place à de grandes améliorations de la méthode.

TABLE 1.2: Description générale d'outils de complétion de réseaux métaboliques.

Nom	Approche	Méthode	Disponibilité	principe
OptStrain [PBM04]	top-down	MILP	Non disponible	*Nombre minimal de réactions à ajouter pour maintenir le rendement optimal du réseau initial associé à une banque de réaction.*
SMILEY [RPC+06]	top-down	MILP	Code Matlab	*Comme OptStrain avec prise en compte de réactions de transport*
GapFill [SKDM07]	bottom-up	MILP	Code GAMS	*Nombre minimal de réactions à ajouter pour produire des métabolites internes (préalablement identifiés) empêchant la formation de biomasse.*
Christian [CMK+09]	top-down	stochastique	Non disponible	*Exploration non exhaustive de l'ensemble des sous-ensembles de réactions minimaux au sens ensembliste qui restaurent la fonctionnalité globale du système.*
Network Expansion [ST09]	bottom-up	recherche combinatoire	ASP	*Exploration exhaustive de l'ensemble des plus petits sous-ensembles qui restaurent la fonctionnalité du système au sens topologique.*

L'intérêt principal de cette approche réside dans le fait que les auteurs effectuent une énumération exhaustive de l'ensemble des solutions qui permettent la production de l'ensemble des métabolites cibles plutôt que ces cibles une à une. Cependant, le prix à payer pour cela est de relaxer la contrainte de fonctionnalité puisqu'ils ne font plus qu'une vérification topologique de la productibilité des cibles sans prendre en compte les contraintes induites par la stœchiométrie et les équilibres internes du système.

1.4.6 Comparaisons

Finalement, les différentes propriétés des méthodes étudiées ci-dessus sont être synthétisées dans la table 1.2. On notera en particulier des difficultés liées à l'accessibilité des méthodes, leur utilisation générique, et leur comparaison.

En détaillant plus précisément les différentes approches, on constate finalement qu'elles ne résolvent pas toutes le même problème d'optimisation. Pour résumer, nous pouvons dire que OptStrain et SMILEY vont chercher à obtenir une fonctionnalité maximale des réseaux pendant que GapFill cherchera à produire chaque métabolite cible individuellement plutôt

TABLE 1.3: Description fonctionnelle d'outils de complétion de réseaux métaboliques.

Nom	Production	Fonctionnalité	Espace de solution	Minimalité
OptStrain et SMILEY	biomasse	quantitative	une seule solution	Taille des solutions
GapFill	cibles intermédiaires	quantitative	une seule solution	Taille des solutions
Christian	ensemble de cibles	topologique	échantillonnage	Ensembliste
Network-expansion	ensemble de cibles	topologique	énumération globale et union	Taille des solutions

que d'essayer d'optimiser la biomasse dans son ensemble. Dans les deux cas il est possible de trouver une solution unique en suivant la méthode présentée même s'il semble possible d'obtenir l'ensemble des solutions en utilisant des techniques propres aux MILP tel que l'"integer cut", mais aucune donnée sur le fait que cette technique serait utilisable en pratique d'un point de vue du temps de calcul n'est disponible à notre connaissance. La méthode développée par Nils Christian permet, elle, une exploration plus complète de l'espace des solutions via un échantillonnage de solutions minimales d'un point de vue ensembliste. Enfin Network-expansion va arriver à être exhaustif dans l'énumération des solutions tout en cherchant à produire l'ensemble des molécules présents dans la biomasse. Ces deux dernières méthodes relâchent quelque peu la contrainte de fonctionnalité étant donné que la fonctionnalité ne sera plus étudiée que d'un point de vue topologique.

D'un point de vue biologique, le fait de choisir une méthode de complétion globale, qui prend en compte l'ensemble des métabolites que l'on sait présents chez une espèce, sans pour autant avoir besoin de connaître la concentration précise de ceux-ci, permet d'avoir une complétion beaucoup plus large du réseau avec des informations incomplètes. Ceci est également un gros apport comparé aux méthodes de complétions manuelles qui n'utilisent que les connaissances pré-existantes chez une espèce pour créer le réseau, au risque de n'apporter aucune nouvelle connaissance au modèle ainsi créé.

Le fait d'avoir une énumération exhaustive des solutions (dont on peut faire l'union ou l'intersection) permet d'éviter de faire un choix entre les différents modèles proposés, ce choix se faisant habituellement par la création de scores impliquant la similarité de séquence et/ou des informations provenant de la phylogénie. Chez des espèces trop lointaines des espèces modèles, ces deux informations risquent de biaiser une éventuelle reconstruction du fait d'une trop grande divergence des séquences avec les séquences connues. D'éventuels transferts horizontaux de gènes risquent également de poser problèmes, les gènes impliqués pouvant être très éloignés phylogénétiquement entre deux espèces étudiées.

Cependant, il faut bien noter que l'approche de Network-expansion, qui a les deux caractéristiques précédentes, nécessite de relaxer le critère de fonctionnalité du système puisque les équilibres stœchiométriques internes ne sont plus pris en compte. Un des objectifs de cette thèse fut d'utiliser et d'améliorer les méthodes combinatoires de Network-expansion pour permettre à cette approche de passer à l'échelle sur des réseaux eucaryotes pour des

espèces non-classiques. Nous en avons profité pour tester l'impact de la modélisation combinatoire de la productibilité au lieu de la fonctionnalité issue de la stœchiométrie. Ces questions seront abordées au chapitre 2.

1.5 Reconstruction de réseaux métaboliques pour les algues

L'objectif de la thèse est de développer des méthodes informatiques pour permettre la reconstruction de réseaux métaboliques pour des espèces non classiques telles que celles étudiées dans le projet Investissement d'avenir Idealg (macro-algues). Dans cette section, nous proposons une revue bibliographique des enjeux spécifiques à la reconstruction de réseaux pour les algues.

1.5.1 Les réseaux métaboliques chez les plantes

De par son statut d'organisme modèle en génétique, *Arabidopsis thaliana* a profité très rapidement des développements de la reconstruction de réseaux métaboliques. D'autres plantes telles que le maïs ont également reçu l'attention de la communauté scientifique de par leur intérêt dans l'agroalimentaire. Les plantes étant souvent des organismes possédant plusieurs tissus et organites, leur modélisation s'en retrouve d'autant plus compliquée et leur analyse d'autant plus intéressantes. Des revues complètes et récentes de ces reconstructions [SHH12] et des perspectives [dODN13] existent, nous allons ici discuter des spécificités de la reconstruction de réseaux chez ces organismes photosynthétiques plutôt que de rentrer en détail dans leur processus de reconstruction.

En effet, les plantes sont des organismes complexes possédant un génome souvent très grand caractérisé par de nombreux gènes enzymatiques ou de transports non annotés. La chronologie de création des réseaux métaboliques des plantes est un bon exemple de la difficulté de reconstruction de tels réseaux. Ainsi en 2009, Poolman et ses collaborateurs [PMSF09] publiaient un modèle de cellules en suspension d'*Arabidopsis thaliana* comprenant 1.336 réactions, suivi de près par la création d'AraCyc [dQP$^+$10a] (1.567 réactions) qui se concentrait sur le métabolisme primaire de cette plante au niveau des feuilles, avant de passer à l'ensemble des tissus d'une plante (le maïs) avec [SSM11a] (1.588 réactions). Aujourd'hui, un réseau compartimenté et multi-tissus d'*A. thaliana* existe permettant des analyses fines de son métabolisme [MOMM$^+$12]. Celui-ci est formé de différents sous-réseaux selon l'organe étudié, le réseau du fruit, par exemple, est formé de 1.143 réactions.

Les reconstructions de ces réseaux ont été grandement facilitées par le fait qu'*Arabidopsis thaliana* est une espèce modèle en génétique depuis de nombreuses années [Fin98]. Ainsi de très nombreuses données "omiques" existent sur cette plante et de nombreuses techniques telles que l'inactivation ciblées de gène sont possibles. Tout cela permet également de valider fonctionnellement les réseaux reconstruits.

1.5.2 Les réseaux métaboliques chez les algues

Les algues sont actuellement des organismes étudiés par les biologistes du fait de leur grande diversité, leur métabolisme chimérique et leur caractère évolutif très particulier. L'étude des algues est également en plein essor au niveau applicatif en biotechnologie avec l'arrivée de plus en plus de données omiques.

TABLE 1.4: Comparaison des différentes méthodes de reconstruction de réseaux métaboliques chez des algues. Plus de détails sont donnés en 1.5

Espèce	Nom du réseau	Source de données	Reconstruction initiale	Complétion	FBA
Chlamydomonas reinhardtii	AlgaGem	KEGG	mapping sur KEGG	manuel, composé par composé	Oui
Chlamydomonas reinhardtii	*i*RC1080	réseau pré-existant & bibliographie	Manuel	Effectué mais pas d'informations	Oui
Ostreococcus		KEGG	mapping sur KEGG	Top-down et Bottom-up, pas de solution unique	Oui
Phaeodactylum tricornutum	DiatomCyc	KEGG et Metacyc	Pathway tools	Manuel pour les voies métaboliques connues	Non

Cependant, les connaissances globales sur le métabolisme des algues est finalement assez limité. Nous pouvons étudier quatre réseaux métaboliques ayant été reconstruits chez des algues, pour trois espèces différentes (chronologiquement) :

- *Chlamydomonas reinhardtii* (AlgaGem et *i*RC1080)
- *Ostreococcus* (pas de nom)
- *Phaelodactylum tricornutum* (DiatomCyc)

Pour ce faire, les étapes classiques de reconstruction de réseaux métaboliques décrites en 1.3 ont été suivies, cependant chaque étape a été réalisée avec des outils et des méthodes différentes. On notera en particulier qu'aucun de ces réseaux n'a fait l'objet d'une étude fonctionnelle équivalente à celle opérée sur les plantes plus classiques citées plus haut.

Le tableau 1.4 résume l'ensemble de ces informations de manière synthétique. Le tableau 1.5 donne des informations sur les réseaux obtenus.

1.5.3 Pipelines de reconstructions utilisés

Nous allons décrire plus précisément chaque étape et les méthodes utilisées pour chaque réseau en insistant sur les spécificités de chacune.

Pour la reconstruction des ébauches métaboliques, concernant *Chlamydomonas reinhardtii*, le réseau AlgaGEM [GdODQPN11] a été reconstruit d'après le génome de *C. reinhardtii*, ainsi que l'ensemble des réactions disponibles dans KEGG. L'ébauche du second réseau de cette espèce (*i*RC1080) [CGM+11] a été reconstruite en se basant sur un précédent réseau reconstruit manuellement en 2009 (*i*AM303) [MGH+09] et contenant 259 réactions impliquées dans des voies métaboliques d'intérêt pour cette microalgue. Le réseau de 2011 reprend le premier en rajoutant des réactions provenant de plus de 250 publications différentes.

TABLE 1.5: Description des réseaux obtenus

Espèce	Nom du réseau	Nombre de gènes	Nombre de réactions	Nombre de métabolites	Nombre de métabolites dans la fonction de biomasse
Chlamydomonas reinhardtii	AlgaGem	2 249	1 725	1 862	39
Chlamydomonas reinhardtii	*i*RC1080	1 080	2 190	1 068	172
Ostreococcus tauri			871	1 014	48
Ostreococcus lucimarinus			964	1 100	48
Phaeodactylum tricornutum	DiatomCyc	1 069	1 719	1 073	-

Pour la reconstruction de l'ébauche du réseau métabolique des pico-algues vertes $Ostreococcus$ [KYW+12], des annotations de gènes provenant de KEGG ont été récupérées et traitées par Matlab avant que la relation gène-réaction soit créée grâce aux données d'orthologie contenues dans KEGG Orthology (KO).

Enfin, la création de l'ébauche métabolique de DiatomCyc (pour $Phaelodactylum\ tricornutum$) a été faite en utilisant la suite logicielle Pathway Tools à partir d'annotations provenant de KEGG et de MetaCyc.

Tous les réseaux mentionnés ont fait l'objet d'une complétion des informations manquantes lors de la première reconstruction suite à des erreurs d'annotations, des données manquantes ou hétérogènes, etc.

Lors de la reconstruction d'AlgaGem, la complétion a été réalisée en reconstruisant manuellement les voies métaboliques intéressantes qui mènent aux métabolites intégrés dans la fonction de biomasse. Pour chaque métabolite présent dans cette biomasse, Dal'Molin $et\ al.$ ont regardé si ce métabolite était productible d'un point de vue "fonctionnel" en utilisant de la FBA. Si ce n'était pas le cas ils ont regardé manuellement quelles réactions sont à ajouter afin de rendre ce métabolite productible.

Une complétion du réseau iRC1080 a été réalisée, mais aucune information autre que "Network gap-filling was performed to make pathways functionnal and account for dead-end metabolites" n'est disponible dans l'article.

La complétion du réseau $Ostreococcus$ a été réalisé automatiquement, de deux manières différentes, en utilisant comme base de données le réseau meta-plant correspondant à l'union des 17 réseaux métaboliques de plantes contenus dans KEGG en 2012. La première complétion a été réalisée en utilisant l'approche "bottom-up" SMILEY [RPC+06] (voir section 1.4.2) et l'approche top-down de Christian et al [CMK+09] (voir section 1.4.4). Ces deux approches ne possédant pas de solution unique, un score prenant en compte la distance phylogénétique des espèces dont proviennent les réactions ajoutées et la similarité de séquence entre les enzymes connues et celles du génome d'intérêt a été utilisé. Si l'approche

"bottom-up" donne un nombre plus faible de réactions ajoutées, l'approche "top-down" donne de meilleurs scores globaux.

L'étape de complétion du réseau DiatomCyc a été réalisée entièrement à la main et en se concentrant uniquement sur les voies métaboliques connues et considérées comme intéressantes pour cette espèce, comme la biosynthèse des acides gras, des sucres ou encore des isoprénoïdes. Le réseau final représente le métabolisme central de $P.\ tricornutum$.

La validation des réseaux considérés a également été plus ou moins soignée. De la FBA a été réalisée sur le réseau AlgaGem. Le critère ayant servi de base à la complétion manuelle étant la production de biomasse par le réseau via les techniques de FBA, il est normal qu'à la fin de la reconstruction la FBA fonctionne toujours. Pour le réseau iRC1080 une fonction objectif a également été créée et testée dans différentes conditions. Une attention toute particulière a été portée sur la validation des voies métaboliques impliquées dans la photosynthèse.

Concernant $Ostreococcus$, là encore une fonction de biomasse a été créée en se basant sur la littérature. Très peu d'informations sont disponibles sur les expérimentations réalisées dans le cadre de l'étude du réseau par FBA.

Enfin le réseau de $Phaelodactylum\ tricornutum$ n'a reçu aucune validation fonctionnelle. Seules certaines voies métaboliques ont été étudiées manuellement d'un point de vue topologique. Une étude des données transcriptomiques disponibles pour cette espèce a toutefois permis de valider la prédiction de certaines de ces voies métaboliques.

1.5.4 *Ectocarpus siliculosus*

Le modèle d'application de cette thèse est $Ectocarpus\ siliculosus$ (représentée en figure 1.8), un eucaryote appartenant au groupe des algues brunes. Les algues brunes sont des organismes photosynthétiques multicellulaires et vivant principalement dans l'eau de mer qui appartiennent au groupe des straménopiles (aussi appelés hétérocontes). Les straménopiles sont distants phylogénétiquement des plantes vertes et des opisthocontes (animaux et champignons) [CPC11], comme le montre la figure 1.9. L'ancêtre commun à ces trois lignées date de la radiation initiale des eucaryotes il y a plus d'un milliard d'années [YHC$^+$04]. Depuis cette divergence, les algues brunes ont développé des caractéristiques spécifiques, qui leur ont notamment permis de s'adapter aux environnements extrêmes de la zone de balancement des marées (ou zone intertidale) et aux stress abiotiques induits. Ces organismes ont acquis leurs plastides lors d'une endosymbiose secondaire [Kee04] impliquant une capture d'une algue rouge, impactant grandement la structure interne de leurs cellules et sur la composition de leur génome du fait de transferts de gènes depuis l'endosymbiote. Les algues brunes sont des piliers des écosystèmes marins et costaux [GKD$^+$07]. Elles sont également très importantes pour l'industrie de l'aquaculture en pleine expansion, et représentent une ressource durable de composants importants pour des applications biotechnologiques [WLW$^+$12, ENFB$^+$14, WQJ13].

Parmi les algues brunes, $Ectocarpus\ siliculosus$ a été choisie comme modèle génomique [CSR$^+$10, BNLC13] et génétique [HCP$^+$10], entraînant l'arrivée d'une nouvelle vague d'études sur cet organisme. Si les recherches sur $Ectocarpus\ siliculosus$ se concentrent classiquement sur la biologie du développement [CGA$^+$11, LBBLP$^+$11], les plus récentes couvrent aujourd'hui des champs beaucoup plus larges tels que son évolution [PVC$^+$10, DPR$^+$11], son métabolisme [MTS$^+$10b, MTS$^+$10a, GDR$^+$10, MCDL$^+$13] ou son adaptation et

FIGURE 1.8: **Photo de sporophyte d'*Ectocarpus siliculosus* en culture.**

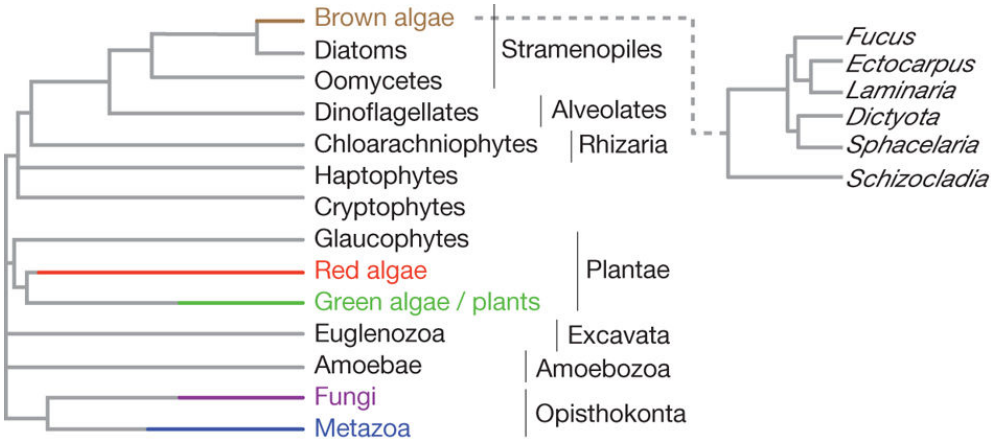


FIGURE 1.9: **Représentation simplifiée de l'arbre phylogénétique des eucaryotes [CPC11]. Les cinq groupes majeures ayant développé une multicellularité complexe sont représenté en couleur. La longueur des barres de couleur indique le temps relatif approché de développement de la multicellularité dans chaque lignée.**

acclimatation aux changements environnementaux [DSP+09, DGR+11, DGG+12, RUR+10]. Toutes ces études suggèrent que les algues brunes possèdent un métabolisme chimérique façonné par des endosymbioses et des transferts latéraux de gènes. Ces études suggèrent également une importance du mannitol dans le stockage de l'énergie et la défense face au stress. Nous reviendrons sur cette molécule plus particulièrement dans le chapitre 4.

Malgré la disponibilité de jeux de données transcriptomiques, protéomiques et de profilages métaboliques [TEP+11], aucun réseau métabolique à l'échelle génomique n'avait été proposé avant cette thèse. Un tel réseau métabolique visait à servir de modèle pour le métabolisme de l'ensemble des algues brunes.

Cependant, l'application des méthodes utilisées pour les réseaux métaboliques d'algues s'avère délicate sur différents points.

D'abord, la reconstruction de la plupart des ébauches métaboliques mentionnées ci-dessus a été réalisée en utilisant des informations pré-existantes dans KEGG. Cette base de données de référence ne contenant aucune information sur le génome ou les réactions présentes chez *Ectocarpus siliculosus*, il est nécessaire d'utiliser uniquement les annotations (manuelles ou automatiques) existant chez cette espèce et disponibles via un portail web (`http://bioinformatics.psb.ugent.be/orcae/overview/Ectsi`). KEGG ayant changé de statut en juillet 2011 et l'accès étant maintenant payant, il fut nécessaire d'utiliser l'autre référence parmi les bases de données de réactions métaboliques, MetaCyc et le logiciel qui lui est associé, Pathway Tools.

De plus, l'analyse bibliographique a mis en évidence le besoin de creuser les approches de complétions qui devraient être utilisées. En effet, la plupart des réseaux algues ont été complétés manuellement en utilisant les connaissances bibliographiques pré-existantes. Or l'utilisation de complétion manuelle est particulièrement compliquée lorsqu'on travaille sur des réseaux métaboliques chez des espèces non classiques. En effet, ceux-ci évoluent en permanence et réaliser une étude bibliographique à chaque modification minime du réseau devient vite irréalisable.

À l'inverse, utiliser une méthode automatique non biaisée permet d'obtenir de nouvelles connaissances. Cependant, la reconstruction des réseaux d'*Ostreococcus* met en jeux deux techniques différentes de complétion : SMILEY [RPC+06] et une approche stochastique [CMK+09]. La première approche, décrite plus précisément en 1.4.2, nécessite d'avoir un grand nombre d'expérimentation sur l'espèce étudiée. En effet, cette méthode utilise les différences de croissance sur différents substrats pour que les réactions présentes dans le modèle amènent à des simulations en adéquation avec les observations numériquement parlant. On peut également utiliser des données de flux connues à travers certaines réactions. Dans le cas d'*Ectocarpus siliculosus* nous ne possédons pas de telles données mais uniquement des données de croissance en milieu eau de mer. L'application de la méthode stochastique ([CMK+09]) (là encore décrite plus précisément en 1.4.4) aurait été possible en théorie. Cependant en pratique, nous nous sommes rendu compte que cette méthode était non seulement très longue mais surtout donnait un grand élevé de solutions différentes. Celles-ci n'étant pas toutes équivalentes, pour les discriminer il faut introduire des scores pour chaque réaction. Ces scores sont habituellement définis en fonction de données d'orthologie et de données phylogénétiques. En présence d'une espèce éloignée phylogénétiquement des espèces classiquement étudiées, ces scores peuvent rapidement perdre de leur pertinence. De plus, dans le cas d'*Ectocarpus siliculosus*, cette espèce résulte de deux

endosymbioses, et les transferts horizontaux de gènes, non négligeables, risquent de ne pas être pris en compte par les données phylogénétiques.

Cette analyse a mis en évidence que la reconstruction du réseau métabolique pour *Ectocarpus siliculosus* ne pourrait pas être réalisée en utilisant les pipelines utilisés pour les autres réseaux algues existant, en particulier à cause de l'hétérogénéité des sources de données permettant de reconstruire l'ébauche et de la nature des données fonctionnelles (profils métaboliques) connues sur *Ectocarpus siliculosus* pour le compléter.

1.6 Principaux apports/résumé

Durant cette thèse, nous nous sommes appliqués à proposer un processus de reconstruction de réseaux métaboliques appliqués aux espèces non classiques. Nous avons appliqué l'ensemble de ce processus à *Ectocarpus siliculosus* en vue d'obtenir un réseau métabolique modèle pour les algues brunes. Ce processus global s'articule autour d'un point particulier, l'amélioration d'une technique existante de complétion de réseaux métabolique résolvant un problème combinatoire difficile. Cette amélioration sera présentée dans le chapitre 2. Nous sommes ainsi passés d'une méthode expérimentale fonctionnant sur de petits jeux de données à une méthode rapide, précise et permettant d'obtenir l'ensemble des complétions minimales possibles pour un réseau métabolique donné.

Une fois ce problème combinatoire résolu, nous nous sommes appliqués à intégrer cet outil dans un processus global de reconstruction de réseaux métaboliques que nous avons appliqué à *Ectocarpus siliculosus* . Ce pipeline décrit dans le chapitre 3 utilise des données génomiques (annotations et séquences) pour reconstruire une ébauche métabolique. Cette ébauche sera ensuite améliorée par l'utilisation de données de profilage métabolique permettant de faire des hypothèses sur les réactions manquantes dans le réseau. Enfin nous proposons des gènes pouvant coder pour les enzymes pour lesquelles nous émettons l'hypothèse qu'elles sont présentes chez l'organisme malgré l'absence d'annotation équivalentes au niveau du génome.

La reconstruction du réseau métabolique d'*Ectocarpus siliculosus* et sa curation manuelle ont permis d'obtenir de nouvelles indications sur le métabolisme de cette algue. Ces hypothèses, qui concernent par exemple la synthèse des acides aminés aromatiques ou encore du molybdenum seront présentées dans le chapitre 4.

Chapitre 2

Complétion combinatoire de réseau métabolique

Durant ce chapitre nous allons étudier plus précisément le problème de complétion d'ébauches de réseaux métaboliques tel que défini par Schaub and Thiele [ST09]. Après avoir décrit le problème d'optimisation auquel nous faisons face dans la partie 2.1, nous étudierons différentes extensions de ce problème pour en améliorer les performances : nouvelle modélisation de la réversibilité (partie 2.4), impact des heuristiques de résolution (partie 2.3), impact de la sémantique de productibilité (partie 2.5). Dans une dernière section (partie 2.5.4), nous étudierons l'efficacité de l'approche de complétion combinatoire revisitée en terme de fonctionnalité quantitative.

Une partie des résultats de ce chapitre a été présentée lors de la conférence LPNMR 2013 et avant d'être publiée dans les actes de cette même conférence [CEG+13].

2.1 Problème d'optimisation

2.1.1 Espace de recherche

Un *réseau métabolique* est représenté par un graphe dirigé bipartite $G = (R \cup M, E)$ où R et M sont des nœuds représentant respectivement les réactions et les métabolites.

Les *réactants* d'une réaction correspondent à l'ensemble des nœuds reliés par des arêtes entrant dans un nœud R, soit $reac(r) = \{m \in M | (m, r) \in E\}$.

Les *produits* correspondent aux arêtes sortantes : $prod(r) = \{m \in M | (r, m) \in E\}$.

Les entrées du problème d'optimisation consistent en l'introduction d'un ensemble de métabolites M et deux ensembles de réactions reliant des métabolites de l'ensemble M :

– Une ébauche de réseau métabolique formée du groupe de réactions : R_{draft},

– Une base de données de réactions métaboliques formée de réactions : R_{db}.

Parmi les métabolites M on définit aussi deux sous-ensembles de métabolites :

– Un ensemble de *métabolites graines* : $M_{seed} \subset M$,

– Un ensemble de *métabolites cibles* : $M_{target} \subset M$.

L'*espace de recherche* va être constitué de l'ensemble des sous-ensembles de $R = R_{draft} \cup R_{db}$.

2.1.2 Atteignabilité

Le problème de complétion est particulièrement dépendant du concept d'atteignabilité.

La définition la plus simple d'atteignabilité est purement topologique et consiste à propager les dépendances du graphe. On dit qu'une réaction $r \in R$ est *topologiquement atteignable* depuis M' si l'ensemble de ses réactants est dans le scope de M', c'est-à-dire si $reac(r) \subseteq M'$. De manière similaire, on dit qu'un métabolite est atteignable depuis M' si $m \in M'$ ou si $m \in prod(r)$ pour au moins une réaction $r \in R$ atteignable depuis M'. L'ensemble des métabolites atteignables depuis M_{seed}, noté $scope_{FWD}(M_{seed})$ (ou $Fwf(M_{seed})$ dans [CMA+08]), est l'ensemble des métabolites produits à partir des graines en utilisant les réactions du réseau. Formellement, $scope_{FWD}(M_{seed})$ est le résultat de l'itération $M_{i+1} = M_i \cup prod(reac(M_i))$ partant de M_{seed} jusqu'à atteindre un point fixe. L'ensemble des métabolites atteignables depuis M_{seed}, noté $scope_{FWD}(M_{seed})$, pourra être calculé en temps polynomial [RK01, ST09].

Une deuxième définition plus fine d'atteignabilité a été introduite par Cottret, avant d'être utilisée par Acuna et Sagot, dans [CMA+08, AMC+12] et cherche à introduire les effets des cycles dans la production des cibles. Il s'agit de définir l'ensemble des précurseurs d'une manière plus élaborée. Au lieu de commencer la propagation depuis un ensemble de sources, les auteurs autorisent l'inclusion de métabolites internes, sous la condition que ces métabolites internes soient produits par une réaction dans une étape ultérieure de la propagation. On définit ainsi $Fwd_Z(M)$, l'ensemble des cibles de M avec Z comme métabolite interne, comme le résultat de l'itération $M_{i+1} = M_i \cup prod(reac(M_i \cup Z))$ partant de $M_0 = M_{seed}$ jusqu'à atteindre un point fixe. Ainsi, un ensemble $S \subset M_{seed}$ est précurseur d'un ensemble T s'il existe Z tel que T et Z sont inclus dans l'ensemble $Fwd_Z(S)$. Cela permet d'identifier les éléments assurant le maintien de différents cycles sans provenir de sources externes. À partir de ce concept, il est possible de définir un nouvel ensemble $scope_{InternalSupply}(M_{seed})$ de cibles atteignables à partir d'une ensemble de sources M_{seed}. Il va s'agir de tous les métabolites qui admettent M_{seed} comme ensemble de précurseurs. L'ensemble $scope_{InternalSupply}(M_{seed})$ pourra être calculé en temps polynomial [AMC+12].

Enfin, lorsque la stœchiométrie du réseau est disponible, on peut aussi considérer que l'ensemble des métabolites atteignables à partir d'un ensemble de sources correspond à l'ensemble des produits de réactions non bloquées, c'est-à-dire pouvant posséder un flux non-nul en FVA. L'ensemble $scope_{FBA}(M_{seed})$ sera alors défini par des contraintes linéaires et pourra être, la plupart du temps, calculé en temps polynomial [MB13].

Dans leur article, Acuna *et al.* [AMC+12] proposent un exemple permettant de bien comprendre leur définition de l'atteignabilité. Cet exemple est présenté en figure 2.1.

Dans cet exemple, la cible t sera atteignable aussi bien lorsque l'on utilise la sémantique topologique simple que celle avec recyclage interne.

Pour la sémantique topologique simple, pour pouvoir produire t, il va falloir que h appartienne au scope des sources. Cette molécule est productible à partir de deux réactions, r_3 et r_5, cette dernière nécessitant la présence de e dans le scope. Or la molécule e pourra être produite directement à partie de deux sources, b et c. La cible t sera donc productible et ce malgré le fait que g et f ne fassent pas partie du scope des sources.

Pour la sémantique topologique avec recyclage interne, si l'on suit la propagation, avec $M_0 = \{a, b, c\}$ nous aurons alors $M_1 = \{a, b, c, e\}$ de manière classique puis $M_2 = \{a, b, c, e, h\}$ et enfin $M_3 = \{a, b, c, e, h, i, t\}$ qui sera un point fixe. Nous aurons donc $Fwd(\{a, b, c\}) =$

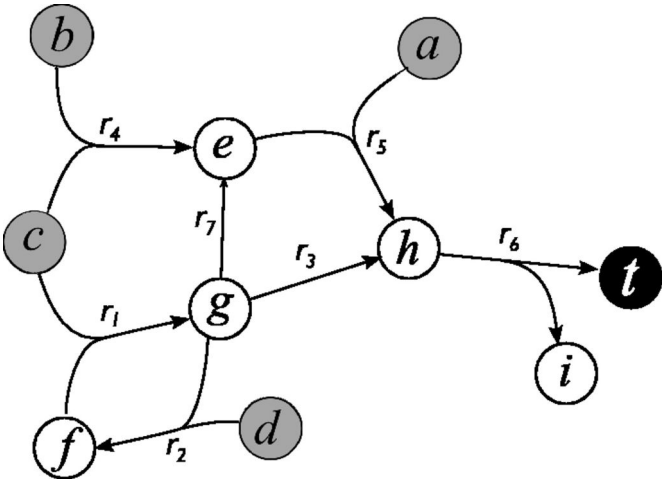


FIGURE 2.1: **Représentation schématique d'un réseau métabolique[AMC+12].** Les nœuds représentent les métabolites et les hyper-arcs représentent les réactions. Les nœuds gris correspondent aux sources et le nœud noir la cible.

$\{a,b,c,e,h,i,t\}$. En revanche, avec l'ensemble des sources présentes dans cet exemple nous aurons $M_0 = \{a,b,c,d\}$ puis $M_1 = \{a,b,c,d,e,f,g,h\}$ étant donné que les métabolites f et g sont des métabolites internes à un cycle pouvant être produites plus tard par ce même cycle. Nous aurons enfin $M_2 = \{a,b,c,d,e,f,g,h,i,t\}$ qui sera le point fixe. Une fois ce point fixe atteint, nous pouvons conclure que $Fwd_Z(\{a,b,c,d\}) = \{a,b,c,d,e,f,g,h,i,t\}$. Il est intéressant de remarquer que $Fwd_Z(\{a,b,c,d\}) = Fwd_Z(\{c,d\})$ si l'on ne prend pas en compte les sources a, b et c qui ne peuvent être produites par aucune réaction.

2.1.3 Problème de complétion

Trouver une complétion de réseaux métaboliques correspondra à trouver un sous-ensemble de réactions $R_{completion} \subset R_{db}$ de telle manière que le scope de M_{seed} associé aux réactions $R_{draft} \cup R_{completion}$ contienne l'ensemble des métabolites M_{target}.

Plus formellement, une *complétion* de $(R_{draft} \cup M_{draft}, E_{draft})$ à partir de $(R_{db} \cup M_{db}, E_{db})$ par rapport à M_{seed} et M_{target} est un ensemble de réaction $R_{completion} \subseteq R_{db} \setminus R_{draft}$ tel que M_{target} est atteignable à partir de M_{seed} dans le réseau $((R_{draft} \cup R_{completion}) \cup (M_{draft} \cup M_{completion}), E_{draft} \cup E_{completion})$, où $M_{completion}$ et $E_{completion}$ sont les projections de M_{db} et E_{db} sur le réseau complété. En particulier, cette définition dépend de la notion d'accessibilité utilisée : $scope_{FWD}$, $scope_{InternalSupply}$, $scope_{FBA}$.

Lors de la complétion, nous allons minimiser le nombre de réactions ajoutées aux réseau métabolique initial. Pour cela nous définissons un score S correspondant au nombre de réactions de R' à ajouter pour obtenir une solution. Un premier problème d'optimisation consistera donc à minimiser S.

Une fois la taille de l'ensemble minimal de réactions à ajouter déterminée, nous allons chercher exhaustivement l'intégralité des ensembles de réactions de cette taille permettant de compléter le réseau. Les différents problèmes de complétion que nous allons étudier consistent donc à calculer :

– Une complétion minimale en taille à un réseau métabolique,
– Toutes les complétions minimale en taille,
– L'union ou l'intersection des complétions minimales en taille.

Dans les applications, nous utiliserons plutôt l'union de l'ensemble des solutions obtenues pour ne pas avoir à discriminer des complétions sans connaissance biologique préalable.

Les variantes de ce problèmes dépendent bien évidemment de la notion d'accessibilité. Une variante, dont la complexité est bien plus élevée, consiste à recherche des complétions minimales au sens ensembliste.

Comme indiqué dans l'état de l'art, Schaub et Thiele [ST09] ont proposé une modélisation en programmation par ensemble-réponse du problème d'optimisation correspondant à la notion d'accessibilité topologique $scope_{FWD}$. Ce programme ASP est décrit en annexe. Dans [ST09], les auteurs ont montré que ce programme permettait de répondre à la question de la complétion de réseaux métaboliques. Leurs expériences ont été réalisés sur des réseaux bactériens (*E. coli*) ayant été dégradés et dont la complétion est réalisée à partir de sous ensembles de taille croissante de la base de données de réactions métaboliques, MetaCyc. Ce programme montrait alors des performances acceptables pour une petite taille de base de données mais ces performances se dégradaient dès que la taille de celle-ci approchait la taille réelle.

2.2 Jeux de test

De manière à étudier les performances de *Network-expansion* vis-à-vis de différents critères, nous avons construits plusieurs jeux de donnés de test.

2.2.1 Impact de la taille des bases de complétion : *Ectocarpus siliculosus* & MetaCyc

Afin de mesurer l'impact de la taille de la base de complétion, nous avons créé un jeu de test se rapprochant le plus possible de la réalité. Une des applications étant la complétion de réseaux métaboliques de nouvelles espèces d'intérêt, nous avons décidé de faire ce benchmark en se basant sur un réseau "brut" d'*Ectocarpus siliculosus*. Ce réseau (à l'inverse de celui présenté dans le chapitre 3) a été créé à partir de la fusion entre un réseau créé depuis d'anciennes annotations du génome d'*Ectocarpus siliculosus* et une toute première version de l'ébauche créée à partir de données d'orthologie. La méthodologie de création de ce réseau sera plus détaillée dans le chapitre 3, le but ici étant de démontrer la faisabilité de la complétion d'un point de vue informatique et non la pertinence biologique de la création de l'ébauche métabolique.

L'ébauche métabolique d'*Ectocarpus siliculosus*, non complétée, contenait à l'époque 1210 réactions et 1454 métabolites. D'après les informations biologiques que nous possédons, nous pouvons recenser 44 métabolites graine, qui correspondent aux constituants du milieu de croissance de l'algue, et 48 métabolites cibles, qui correspondent à des molécules identifiées comme étant productibles par l'algue. En utilisant la sémantique topologique simple ou avec recyclage interne, nous pouvons constater que le réseau initial n'était pas capable de produire 25 des cibles.

La taille de la base de donnée étant le facteur majeur de la complexité, nous avons décidé de jouer principalement sur celle-ci pour la construction du jeu de test. Nous avons donc créé des sous-ensembles de la base de données MetaCyc (version 17.0) de taille comprise entre 10000 et 5000 réactions, en enlevant 1000 réactions à chaque fois. Pour chaque taille de base de données, 10 réplicats ont été réalisés. Nous avons veillé à ce que chaque sous-ensemble contienne à peu près la même proportion de réactions réversibles que la base de données initiale, c'est à dire 42%.

2.2.2 Effet de la taille de la base de données sur la productibilité des cibles

Étant donné que nous prenons des sous-parties de la base de données MetaCyc dans cette étude, les métabolites cibles productibles à partir des graines vont varier selon les réactions présentes dans ces sous-ensembles. De manière intéressante, les deux sémantiques étudiées (topologique simple et avec recyclage interne) permettent dans tous les cas étudiés de produire les mêmes métabolites à partir du réseau initial auquel a été ajouté toutes les réactions présentes dans les sous-ensembles des bases de données. Cela semble indiquer que les cycles ne sont jamais bloquant pour produire des métabolites d'un point de vue topologique, du moins dans les exemples étudiés ici. La figure 2.2 présente la distribution du nombre de cibles productibles selon les sous-ensembles de la base de données utilisés.

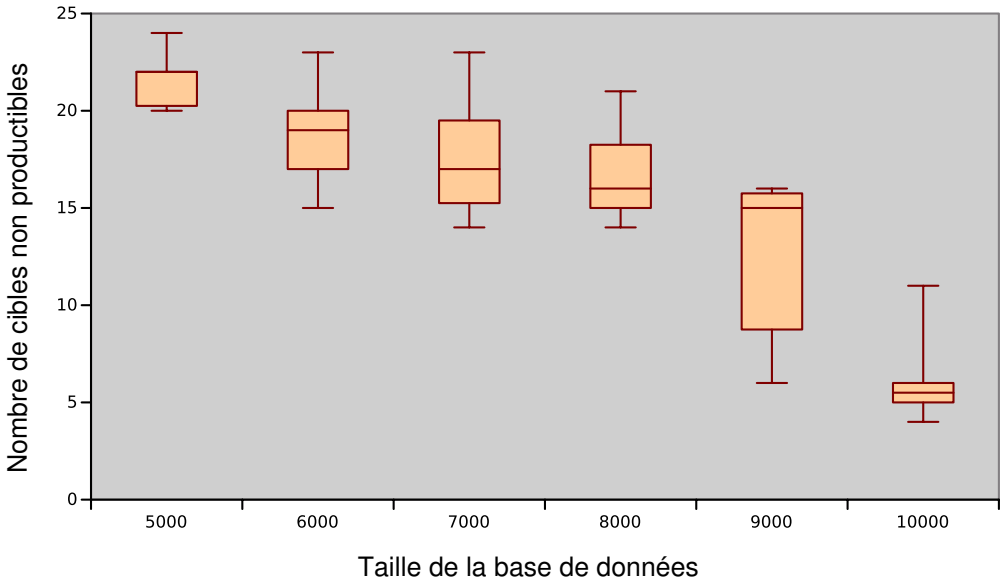


FIGURE 2.2: **Représentation de la répartition statistique du nombre de métabolites non productibles en fonction de la taille de la base de données utilisée.** Les boites correspondent à la taille du premier et du troisième quartile, la ligne centrale correspond à la moyenne et les barres correspondent aux valeurs extrêmes.

Le réseau initial est capable de produire 23 cibles sur les 48 identifiées. Il reste donc 25 cibles à reconstruire à l'aide des bases de données. La figure 2.2 nous montre que, globalement, plus la taille des sous-ensemble de MetaCyc considérés augmente, plus il est possible d'expliquer la productibilité des cibles. Ce point particulier sera à retenir pour la suite, la

taille des complétions proposées ainsi que le temps de calcul pour les identifier sera évidemment influencé par le nombre de voies métaboliques à compléter.

2.2.3 Fonctionnalité : E. Coli

Pour tester la différence entre la sémantique de productibilité topologique et une sémantique quantitative, nous avons exploité un réseau de référence pour sa fonctionnalité et la qualité de l'annotation de ses cycles. Il s'agit d'un réseau d'*Escherichia coli*, Ec_iJR904 issu de la banque de référence gérée par H. Pallson [RVSP03].

Ce réseau est composé de 1.074 réactions (dont 143 réactions à la frontière du système pour l'import et l'export de molécules) et 905 composés métaboliques différents. Une fonction de biomasse formée de 49 métabolites différents a été utilisée dans le cadre de cette étude. Ce réseau existe sous la forme d'un fichier sbml, mais celui-ci n'étant pas totalement cohérent intrinsèquement (voir par la suite) nous avons finalement choisi de partir de la matrice stœchiométrique associée pour reconstruire des fichiers sbml. On peut regretter que les identifiants (des métabolites et des réactions) de ce réseau ne soient pas normalisées vis-à-vis des bases de données de réactions de référence (KEGG ou MetaCyc). En particulier, cela nous a empêcher d'utiliser MetaCyc comme base de référence pour tester la complétion sur des versions dégradées de ce réseau

L'idée générale est de regarder dans ce réseau, pour des fonctions objectives différentes, les types de réactions qui existent (réactions obligatoires, bloquées et alternatives pour la production de biomasse) en utilisant les techniques de FBA (pour tester la fonctionnalité globale du réseau) et de FVA (pour trouver les classes de réactions). Une fois ces réactions classées, nous dégradons chaque réseau en enlevant une proportion équivalente de chaque classe de réactions. Une fois la dégradation accomplie, nous utilisons une approche de complétion topologique simple pour compléter le réseau afin de rendre productible d'un point de vue topologique l'ensemble des métabolites compris dans la biomasse.

Les jeux de tests qui ont été extraits de ce réseaux ont consisté à les dégrader de manière à casser la production de la biomasse. À partir de la fonction de biomasse associée au réseau, 90 biomasses intermédiaires ont été créés aléatoirement en enlevant 9 à 26 composants à cette biomasse.

Le déroulement globale de la création du jeu de test a été celui-ci :

- Identification de la classe de chaque réaction en rapport avec la fonction de biomasse étudiée (Obligatoire, bloquée ou alternative) à l'aide de techniques de FVA.
- Dégradation du réseau selon différentes contraintes :
  - Enlever entre 10 et 40% des réactions du réseau par palier de 10%,
  - Rompre la production de biomasse en FBA,
  - Enlever une proportion égale de réactions obligatoires, bloquées et alternatives,
  - Réaliser 10 réplicats aléatoires pour chaque dégradation.

Au final nous obtenons 9.600 réseaux dégradés : 4 pourcentages de dégradation * 10 réplicats * 90 fonctions de biomasse. Tout ces réseaux sont non fonctionnels vis-à-vis de la biomasse qui leur est associée.

La complétion effectuée par la suite correspond à compléter ces réseaux dégradés en utilisant comme base de référence le réseau initial non dégradé, le but étant d'identifier à quelle classe appartiennent les réactions ainsi ajoutées.

Une difficulté pour la préparation de ces jeux de données a consisté à identifier correctement les informations relatives à la réversibilité des réactions ou de la direction des réactions des fichiers sources. Dans un fichier sbml il y a en effet deux manières de spécifier la réversibilité et/ou la direction d'une réaction.

Ainsi, pour la réversibilité, il est possible de spécifier un argument "reversible" à une réaction de la manière suivante :

Listing 1: Une réaction définie comme réversible en format SBML

```
<reaction id="XXX" name="YYYY" reversible="true">
```

ou

Listing 2: Une réaction définie comme irréversible en format SBML

```
<reaction id="XXX" name="YYYY" reversible="false">
```

Cependant, il est également possible de spécifier l'espace des possibles pour les flux passant à travers cette réaction en fixant une upper-bound (borne supérieure) et une lower-bound (borne inférieure) pour ces flux. Typiquement pour la réaction 3 il est indiqué qu'il s'agit d'une réaction réversible alors que les flux ne peuvent aller que dans une seule direction.

Il existe également deux manières d'indiquer la direction des réactions, là encore en prenant en compte les valeurs des flux ou non. Ainsi, dans les spécifications de SBML, il est possible de désigner une molécule comme réactant d'une réaction ou comme produit de celle-ci. Une réaction irréversible (peu importe la manière dont l'irréversibilité est introduite) ayant pour réactant une molécule A et une molécule B, et ayant pour produit une molécule C correspondrait donc à : $A + B \rightarrow C$.

Cependant, là encore, il faut prendre en compte les flux. Pour certaines réactions, la borne supérieure du flux peut être nulle et la borne inférieure peut être négative. Cela indique que la réaction ne doit pas être prise dans le sens initial mais dans le sens inverse, où les réactants deviennent les produits, et les produits deviennent les réactants. C'est le cas notamment de la réaction 3 (ayant pour identifiant "R_EX_o2_LPAREN_e_RPAREN_") qui est définie comme faisant rentrer du dioxygène dans une cellule, mais, si l'on se fie aux flux, qui exporte du dioxygène de l'intérieur vers l'extérieur de la cellule.

Listing 3: Une réaction irréversible spécifiée comme réversible mais dont le sens des flux ne correspond pas à la définition

```
<reaction id="R_EX_o2_LPAREN_e_RPAREN_" name="O2 exchange" reversible="
    true">
  <notes>
    <html:p>Abbreviation: R_EX_o2_LPAREN_e_RPAREN_</html:p>
    <html:p>Synonyms: _0</html:p>
    <html:p>Equation: [e] : o2 &lt;==&gt;</html:p>
    <html:p>Confidence Level: </html:p>
    <html:p>GENE ASSOCIATION: </html:p>
  </notes>
  <listOfReactants>
    <speciesReference species="M_o2_e" stoichiometry="1"/>
```

```
  </listOfReactants>
  <listOfProducts>
    <speciesReference species="M_o2_b" stoichiometry="1"/>
  </listOfProducts>
  <kineticLaw>
    <math xmlns="http://www.w3.org/1998/Math/MathML">
      <ci>FLUX_VALUE</ci>
    </math>
    <listOfParameters>
      <parameter id="LOWER_BOUND" value="-20" units="mmol_per_gDW_per_hr"
          />
      <parameter id="UPPER_BOUND" value="0" units="mmol_per_gDW_per_hr"/>
      <parameter id="OBJECTIVE_COEFFICIENT" value="0" />
      <parameter id="FLUX_VALUE" value="0" units="mmol_per_gDW_per_hr"/>
    </listOfParameters>
  </kineticLaw>
</reaction>
```

Ces différentes formalisations indiquent que les réseaux présentés sous leurs formats SBML et matriciels peuvent en fait avoir des comportements et des structures très différentes. Pour homogénéiser les analyses et les comparaisons entre approches quantitatives et qualitatives, il a été décidé de ne pas prendre en compte les informations présentes dans les fichiers SBML mais de reconstruire des fichiers SBML cohérents à partir de matrices stœchiométriques et de vecteurs de flux.

2.3 Performance

Comme indiqué dans l'état de l'art, Schaub et Thiele [ST09] ont proposé une modélisation en programmation par ensemble-réponse du problème d'optimisation correspondant à la notion d'accessibilité topologique $scope_{FWD}$. Ce programme ASP, appelé Network-expansion, est décrit en annexe. Dans [ST09], les auteurs ont montré que ce programme permettait de répondre à la question de la complétion de réseaux métaboliques. Leurs expériences ont été réalisés sur des réseaux bactériens (*E. coli*) ayant été dégradés et dont la complétion est réalisée à partir de sous ensembles de taille croissante de la base de données de réactions métaboliques, MetaCyc. Ce programme montrait alors des performances acceptables pour une petite taille de base de données mais ces performances se dégradaient dès que la taille de celle-ci approchait la taille réelle.

2.3.1 Heuristiques de résolutions

En 2012, une nouvelle méthode de recherche pour les solveurs ASP a été développée par l'équipe qui maintient la collection Potassco [GKK+11] et qui développe notamment le solveur ASP Clasp. Cette méthode de recherche utilise une optimisation basée sur l'insatisfiabilité de contraintes [AKMS12].

Jusque avant la version 3, le solveur Clasp [GKS12] utilisait uniquement des méthodes d'optimisation très efficaces basées sur des algorithmes de séparation et d'évaluation (ou

branch and bound). Ces techniques top-down ont montré leur efficacité pour résoudre de nombreux problèmes d'optimisation en travaillant énormément sur les bornes supérieures et inférieures d'un problème. La stratégie globale consistera à descendre plus ou moins rapidement une borne supérieure jusqu'à atteindre une configuration insatisfiable qui marquera une borne inférieure. Cette méthode d'optimisation a montré sa forte efficacité sur certains types de problèmes. En parallèle, une nouvelle méthode a émergé du domaine des solveurs SAT pour résoudre les problèmes MaxSAT [CKS01]. Cette approche se base sur l'identification et la relaxation de cœurs insatisfiables et a montré son efficacité notamment lors de compétitions comme la "2008 MaxSAT evaluation". Le solveur Clasp étant peu efficace pour certains problèmes résolus efficacement par cette approche les développeurs de ce solveur se sont inspirés de ces nouvelles techniques pour développer une heuristique efficace combinant les caractéristiques d'ASP et les techniques développées dans ce cadre, en étendant l'algorithme à la résolution de problèmes d'optimisation pondérée [AKMS12].

La base de ce travail consiste en l'identification de cœurs insatisfiables. Un cœur insatisfiable est un sous-ensemble de clauses du problème initial dont la conjonction est insatisfiable. La méthode essaie donc de résoudre le problème une première fois. Si elle y arrive directement, le problème est satisfiable, la solution est trouvée. Sinon on identifie ainsi un ensemble insatisfiable de contraintes. Le solveur va alors relaxer des contraintes afin que l'ensemble soit satisfiable avant d'identifier un autre ensemble insatisfiable, les relaxer à leur tour et cela jusqu'à ce qu'il n'y ai plus de contraintes insatisfiables dans le problème.

Dans le cadre de notre problème d'optimisation, nous cherchons à minimiser le nombre de réactions ajoutées dans un modèle à partir d'une base de données de réactions métabolique très grande. Nous avons donc un problème de minimisation qui sera traditionnellement traité par une approche "top-down" en prenant un modèle stable résout l'ensemble des contraintes, peu importe sa taille, avant de réduire la taille du modèle petit à petit. Étant donné que la taille de la base de données est très grande par rapport à la taille de la solution recherchée (de l'ordre de la dizaine de milliers pour la base de données contre des solutions de taille inférieure à 100), cette minimisation pourra être longue. À l'inverse l'heuristique se basant sur les cœurs insatifiables va traiter le problème par une approche "bottom-up" en cherchant des solutions de très petite taille puis en augmentant progressivement la taille de ces solutions jusqu'à trouver un modèle stable satisfaisant l'ensemble des contraintes qui sera donc de taille minimale.

Cette méthode ayant été validée par les chercheurs de Potsdam sous la forme d'un solveur à part entière (Unclasp),elle a ensuite été intégrée au solveur Clasp (à partir de Clasp 3.0 en février 2014) sous la forme d'une paramètre de recherche : `-opt-strategy=usc`.

Par soucis de lisibilité, nous désignerons dans la suite par :

- *Clasp* : la recherche par séparation et évaluation (Clasp version 2, ou version 3 avec le paramètre `-opt-strategy=bb`),
- *Unclasp* : la recherche MaxSat (Unclasp version 0.1 ou Clasp version 3 avec le paramètre `-opt-strategy=usc`).

2.3.2 Nouvelle méthode d'optimisation dans *Network-expansion*

L'apparition de nouvelles techniques de recherche a permis de faire évoluer *Network-expansion* qui réalisait la complétion topologique. Clasp a été configuré de manière à pouvoir trouver au plus vite une solution optimale en autorisant des redémarrages de la re-

cherche afin de ne pas rester bloqué dans des optimas locaux. Cette solution semble être la meilleure pour ce solveur après utilisation quotidienne de celui-ci pendant plusieurs mois.

Unclasp n'étant pas capable de réaliser l'énumération des solutions ainsi que l'intersection des solutions, seul Clasp a été utilisé dans ce cas afin d'observer l'influence de l'inclusion de la réversibilité dans la modélisation.

2.3.3 Recherche d'optimum

Nous avons utilisé les jeux de tests pour tester l'impact des deux méthodes de résolution Clasp ou Unclasp sur les performances de la complétion. Les critères testés ont été les suivants :

– Calcul des optimaux de complétion,
– Énumération de toutes les solutions possibles,
– Calcul de l'intersection de toutes les solutions possibles.

La figure 2.3 donne les résultats de ce test pour le temps de calcul des optimaux en fonction du solveur.

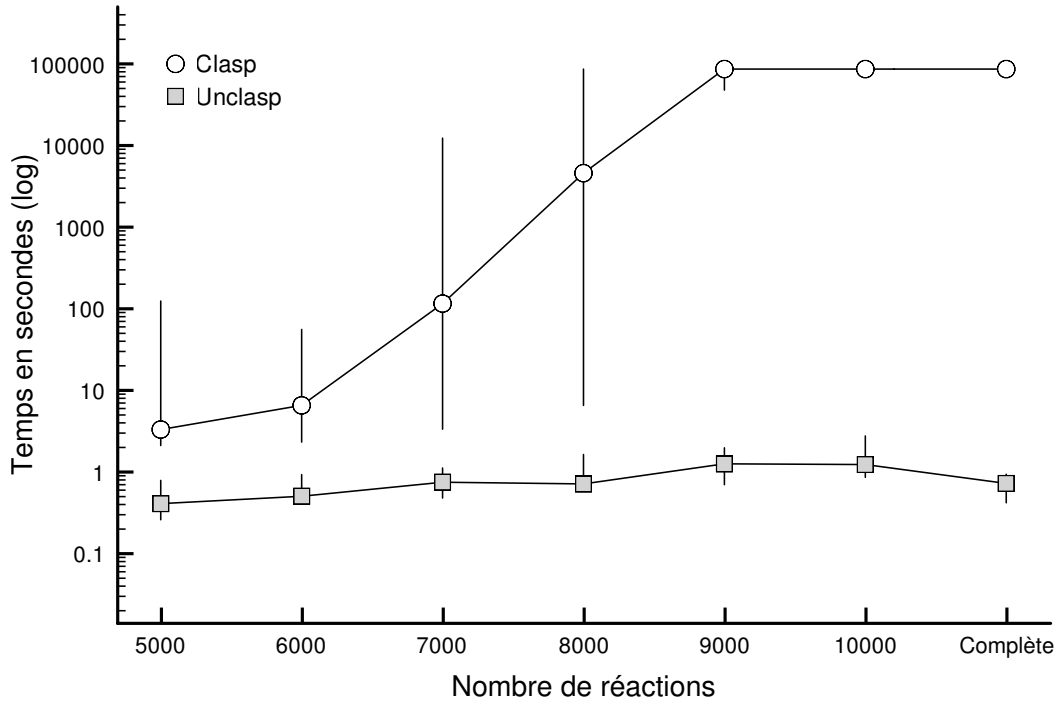


FIGURE 2.3: **Temps de calcul de Clasp et Unclasp pour la recherche de la taille minimale de complétion.** Les cercles transparents correspondent à la médiane des expérimentations avec Clasp. Les carrés grisés correspondent à la médiane des expérimentations avec Unclasp. Pour chaque expérience, les valeurs minimale et maximale sont reportées avec des lignes verticales.

Comme nous pouvons le voir, le changement de méthode de recherche entraîne un gain important dans le temps de calcul des optimaux. Étant donné que nous avons un timeout à 100.000 secondes, nous remarquons que, même avec une taille de base de données de

9000 réactions, il était souvent impossible de trouver des solutions au problème étudié. Les résultats de la complétion ainsi que le détail des timeout avec Clasp sont donnés dans le tableau 2.1.

TABLE 2.1: **Tailles et nombre de solutions optimales pour chaque sous-ensemble de réactions.** Les time-outs de Clasp sont également reportés (100.000 secondes).

Nombre de réactions	5000	6000	7000	8000	9000	10000	Complet
Taille de l'optimal	[6,14]	[7,22]	[7,29]	[9,29]	[16,47]	[33,50]	52
Nombre de solutions	[4,32]	[6,324]	[6,1728]	[16,3456]	[80,1150]	[180,22800]	2600
Time-out de Clasp	0	0	1	3	9	10	10

2.3.4 Impact de la taille de la base

Une fois la recherche d'optima réalisée, nous pouvons utiliser cet optimum pour rechercher l'ensemble des solutions de taille minimale permettant de compléter le réseau métabolique. Unclasp n'étant pas capable de faire cette énumération, nous avons utilisé uniquement Clasp pour cela. Les résultats sont présentés en figure 2.4. Ces résultats nous montrent que le nombre de solutions différentes augmente jusqu'à atteindre un plateau autour d'une taille de base de données de 9.000 réactions, tandis que le temps de calcul augmente exponentiellement avec la taille de la base de données. Cela semble montrer que 9000 réactions sont suffisantes pour compléter le réseau, les autres réactions augmentant l'espace de solutions, et donc le temps de calcul. Malgré cette croissance exponentielle, le calcul de l'ensemble des solutions reste réalisable en un temps suffisamment court pour être utilisé en routine. Cela est principalement dû au fait que l'utilisation d'Unclasp nous permet d'avoir la valeur de l'optimum en un temps très court.

Ceci montre tout d'abord que l'utilisation des méthodes de recherche d'Unclasp permet de calculer le nombre minimum de réactions à ajouter lors de l'étape de complétion du réseau métabolique en quelques secondes malgré un espace de solution très grand. Cette première étape est indispensable à toute autre étude de complétion. Il est à noter que la taille de la base de données ne semble pas impacter énormément la recherche d'optimas avec Unclasp. Ce solveur est en effet particulièrement adapté à la recherche d'optima très petits (ici de taille 50 environ) face à un espace de solutions très grand (plus de 10000 réactions).

Enfin, nous avons déterminé la valeur de l'intersection des solutions ainsi que le temps nécessaire au calcul de cette intersection. L'intersection des solutions est calculée en utilisant l'option `-enum-mode=cautious` de Clasp. La figure 2.5 présente les résultats obtenus. On remarque encore une fois que le temps de calcul est exponentiel par rapport au nombre de réactions dans la base de données. Par contre, la taille de l'intersection grandit beaucoup moins vite alors que le nombre de solutions optimales augmente énormément. Cela est cohérent avec les observations réalisées au cours de la thèse selon lesquelles le nombre total de solutions possibles n'impacte que très peu sur la taille de l'union et/ou de l'intersection de ces solutions. Cela semble indiquer une très forte redondance dans l'ensemble des solutions étudiées.

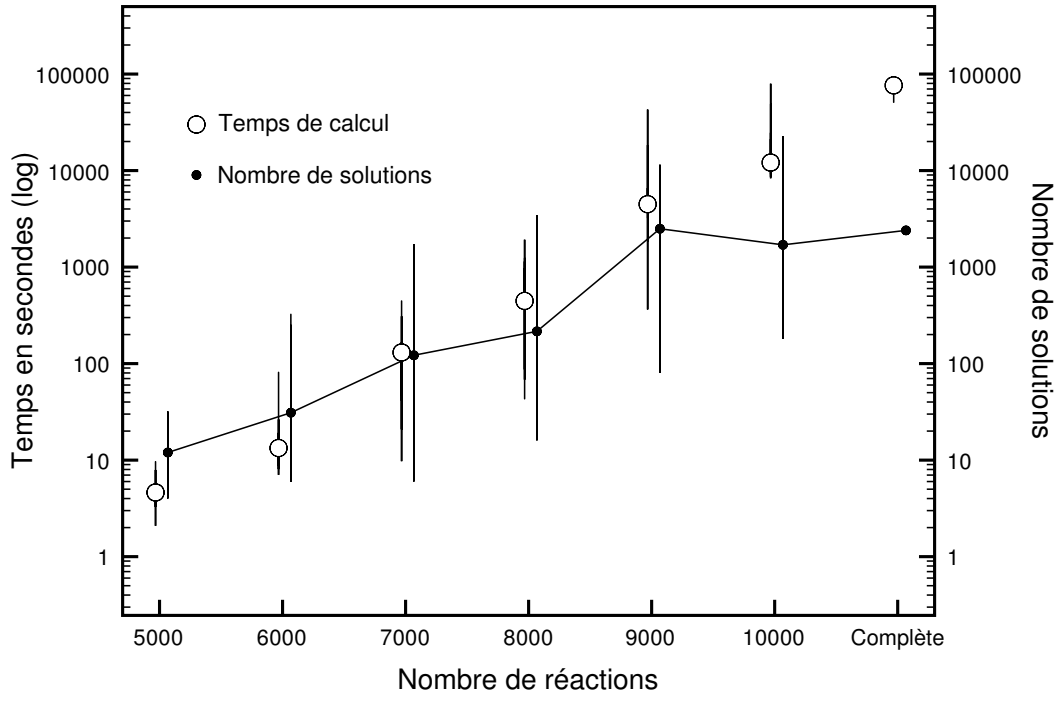


FIGURE 2.4: **Temps de calcul de Clasp pour énumérer l'ensemble des solutions** Les cercles transparents correspondent aux médianes des durées des expérimentations. Les points noirs correspondent aux médianes du nombre de solutions. Les lignes verticales correspondent aux valeurs maximales et minimales.

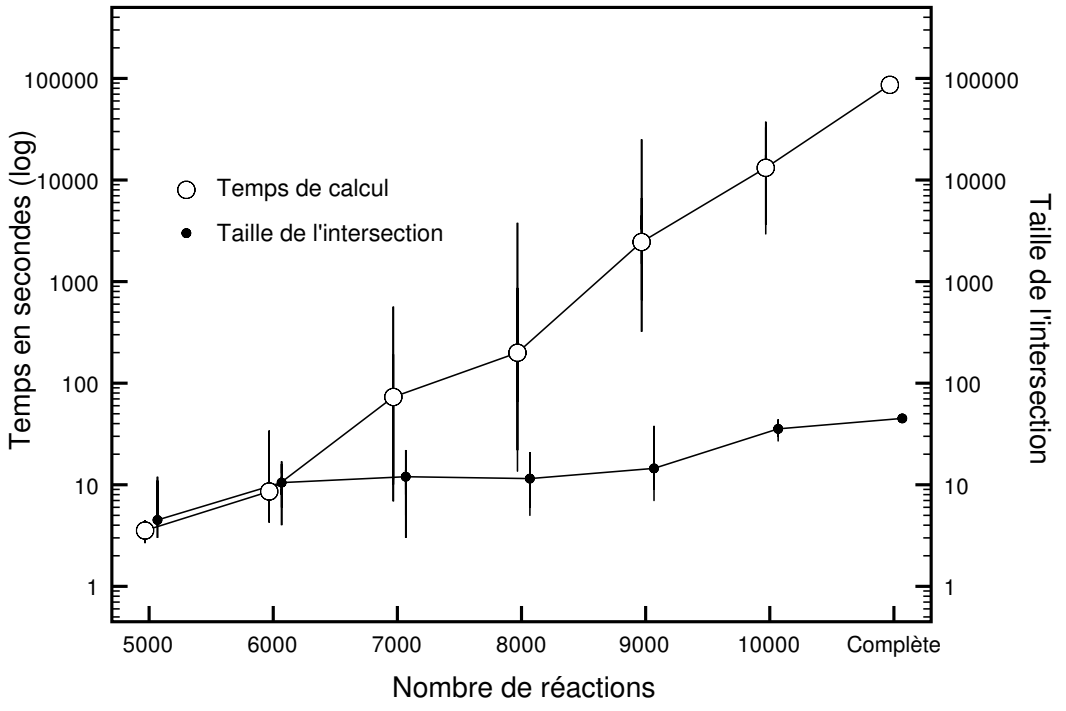


FIGURE 2.5: **Temps de calcul de Clasp pour déterminer les intersections des solutions optimales.** Les cercles transparents correspondent aux médianes des résultats de durée des expérimentations. Les points noirs correspondent à la taille médiane des intersections. Les lignes verticales correspondent aux valeurs maximales et minimales.

Suite aux différentes observations, nous pouvons dire que l'utilisation de la combinaison des méthodes de recherche d'Unclasp et Clasp est indispensable pour permettre la reconstruction de réseaux métaboliques avec ce programme ASP. La méthode de recherche d'Unclasp est la seule capable (en un temps raisonnable) de calculer le nombre minimal de réactions à ajouter au réseau métabolique, mais n'est pas capable, ni d'énumérer l'ensemble des solutions optimales, ni de calculer leur intersection ou leur union. En revanche, une fois que l'optimum recherché est connu, Clasp permet de réaliser ces trois tâches en un temps suffisamment court pour être utilisé en routine. Maintenant que la méthode d'optimisation présente dans Unclasp a été ajoutée dans Clasp (version 3), la combinaison des deux outils en un seul simplifie cette étape de complétion.

2.4 Impact du codage de la réversibilité sur les performances de *Network-expansion*

2.4.1 Réversibilité dans *Network-expansion*

Au début du projet, *Network-expansion* ne prenait pas du tout en compte les réactions réversibles. Si une réaction dans une base de données ou dans un modèle était considérée comme réversible, cette réaction était dédoublée pour créer deux réactions différentes, une allant dans un sens et l'autre dans le sens inverse. Cette modélisation initiale peut poser différents problèmes, en terme de fonctionnalité comme de performance.

Tout d'abord, créer deux réactions en sens inverse l'une de l'autre pour chaque réaction réversible est incompatible avec la réalité biologique que l'on souhaite modéliser lors de la création d'un réseau métabolique et pour sa complétion avec *Network-expansion*. En effet, la réversibilité d'une réaction est une donnée importante en biologie et deux réactions possédant les mêmes réactants et mêmes produits mais ne partageant pas la même réversibilité pourront être catalysées par des enzymes différentes. Un tel exemple est présenté en figure 2.6.

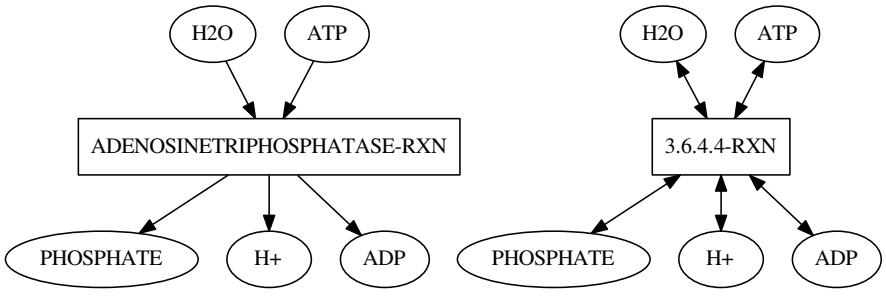


FIGURE 2.6: Exemple de deux réactions impliquant les mêmes molécules et provenant de la même base de données (MetaCyc), l'une étant réversible et l'autre irréversible.

Ainsi, ajouter la réaction ayant pour identifiant "ADENOSINETRIPHOSPHATASE-RXN"

OU "3.6.4.4-RXN" n'aura pas les mêmes implications en matière de recherche biologique.

De plus, étant donné que nous essayons d'ajouter le nombre minimal de réactions pour compléter le réseau, si nous avons besoin qu'une réaction réversible fonctionne dans les deux sens, nous devrions ajouter deux réactions différentes plutôt qu'une seule dans le réseau, au risque de biaiser le résultat final en terme de cardinalité.

D'autre part, cette dissociation d'une réaction réversible en deux réactions irréversibles n'était pas automatique lors de la complétion, et devait se faire manuellement en amont. L'objectif de cette thèse étant notamment de produire un outil utilisable aisément par toute personne travaillant sur des réseaux métaboliques, cette étape risquait d'être oubliée, entraînant ainsi une totale incohérence au niveau des résultats, seul un des sens des réactions réversibles étant pris en compte aléatoirement.

Enfin, comme montré précédemment, la complétion d'un réseau métabolique est un problème hautement combinatoire. Hors, les bases de données contiennent énormément de réactions réversibles (autour de 40% des réactions dans MetaCyc 17.0). L'espace des solutions de complétion des réseaux métaboliques est très grand. Dédoubler cet espace des solutions pour chaque réaction réversible agrandit encore plus l'espace des solutions, avec un risque important de ralentir la recherche de l'optimum.

2.4.2 Nouvelle modélisation ASP de la réversibilité

À l'inverse, conserver les réactions réversibles telles quelles en changeant l'encodage, outre le fait d'outrepasser l'ensemble des problèmes cités précédemment, possède quelques avantages vis-à-vis du solveur de contraintes. Ainsi, si l'on garde les réactions comme réversibles, l'espace des solutions ne grandit plus, et ajouter une réaction réversible permet de s'affranchir plus tard de l'étude de l'autre sens de réaction.

En ASP, cela a consisté à créer un fait "reversible/1" pour chaque réaction réversible du modèle et de la base de données de réactions, et d'utiliser ce fait pour changer la définition des scopes utilisant des réactions réversibles. Ainsi, par exemple, le scope d'une molécule M pourra désormais être calculé de la manière présentée en listing 4. Ceci est à comparer avec la précédente version de la définition d'un scope avec le code ASP initial présenté en annexe B. Il était alors nécessaire de créer une seconde réaction en inversant les réactants et les produits.

Listing 4:

```
scope(M) :- seed(M).
scope(M) :- product(M,R), reaction(R), scope(M') : reactant(M',R).
scope(M) :- reactant(M,R), reversible(R), scope(M') : product(M',R).
```

2.4.3 Impact sur les performances

La figure 2.7 donne les résultats de ce test pour le temps de calcul des optimaux en fonction de la réversibilité pour l'utilisation de Clasp et d'Unclasp.

Nous pouvons remarquer que l'introduction de la réversibilité, en plus d'être plus proche de la réalité biologique, a permis d'accélérer grandement la recherche d'optimaux avec Un-

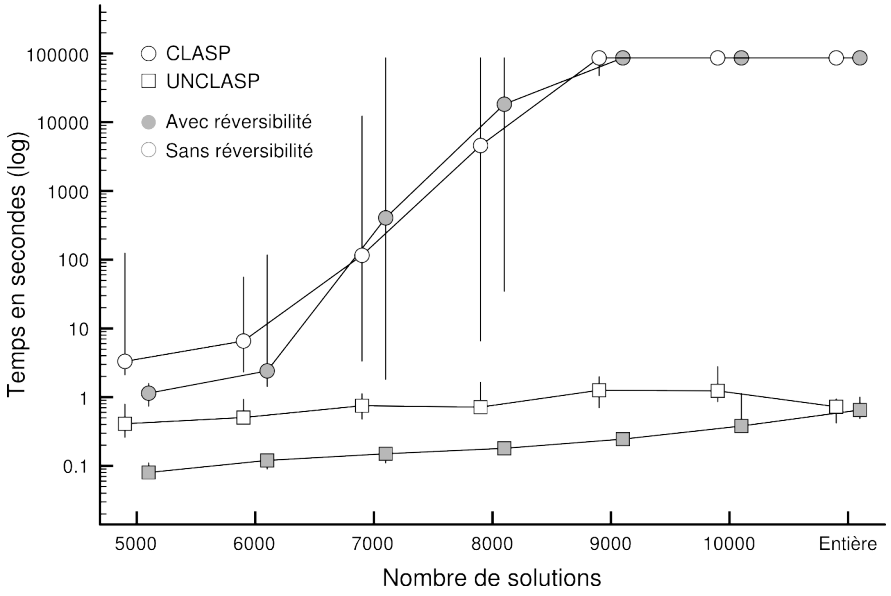


FIGURE 2.7: **Temps de calcul de Clasp et Unclasp pour la recherche de la taille minimale de complétion.** Les cercles correspondent à la médiane des expérimentations avec Clasp. Les carrés correspondent à la médiane des expérimentations avec Unclasp. Les ronds et carrés grisés correspondent à la modélisation avec prise en compte de la réversibilité, les ronds et carrés vides correspondent à la même modélisation sans cette prise en compte. Pour chaque expérience, les valeurs minimale et maximale sont reportées avec des lignes verticales.

clasp. Cette accélération va d'un facteur 2 à 11. Il est intéressant de remarquer l'absence d'accélération significative avec Clasp.

Une des explications possible pourrait se trouver dans les techniques d'optimisation différentes entre les deux solveurs. En effet, Clasp travaille beaucoup sur les bornes supérieures et inférieures de la taille des solutions, avec donc un effet limité de la taille de l'espace de recherche si Clasp arrive rapidement à réduire ces bornes. En revanche Unclasp recherche à prouver l'insatisfiabilité de cœurs contraints, réduire l'espace de recherche permettra donc d'accélérer cette preuve d'insatisfiabilité. Ceci n'est pour l'instant qu'une hypothèse et il serait intéressant d'étudier ce cas plus en détail.

Le nouvel encodage de la réversibilité des réactions permet donc de mieux prendre en compte la réalité biologique des réactions étudiées et d'accélérer la recherche d'optima.

2.4.4 De *Network-expansion* à *Meneco*

Après changement des méthodes de recherche avec l'introduction d'Unclasp pour la recherche d'optimum et Clasp pour les énumérations, et recodage de la réversibilité, $Network\text{-}expansion$ a été inclus dans un package Python autonome et a pris le nom de $Meneco$ (pour Metabolic network completion : `https://pypi.python.org/pypi/meneco/`), développé par Sven Thiele.

2.5 Sémantique de productibilité

Nous l'avons vu précédemment, le problème de complétion peut être considéré comme un problème d'optimisation où l'on rajoute le minimum de réactions pour produire un ensemble de métabolites à partir d'un autre ensemble de métabolites. Par contre, la notion d'accessibilité ou de productibilité peut être définie de manières différentes.

Nous avons déjà vu qu'il existe différentes manières de considérer un métabolite comme productible ou non. Nous allons ici étudier plus finement ces concepts pour mieux les comparer, en particulier vis-à-vis de leur utilisation pour compléter un réseau métabolique.

2.5.1 Impact des cycles sur les différents concepts d'accessibilité

Accessibilité quantitative (FBA) La recherche de productibilité en FBA est une recherche de nature numérique. Pour savoir si un ensemble de métabolites est productible, nous mettons l'ensemble de ces métabolites (avec la concentration associée si celle-ci est disponible) dans une fonction objectif globale. Les métabolites graines, à partir desquels les métabolites de la fonction objectif devront être produits, seront considérés comme des métabolites à la frontière du système. Aucune contrainte n'est fixée quand à leur conservation dans le système. Des bornes pourront être placées pour mettre des contraintes numériques à leur import dans le système en fixant les flux maximaux et minimaux passant à travers les réactions d'import. La stœchiométrie des réactions revêt ici une importance toute particulière, comme le montre la figure 2.8.

La partie (a) représente un cycle formé de quatre réactions métaboliques. Nous avons une entrée à ce cycle (métabolite A) et une sortie (métabolite F). En FBA ce cycle sera fonctionnel car le métabolite A permettra d'alimenter le cycle en matière et la sortie de F ne consommera pas de matière.

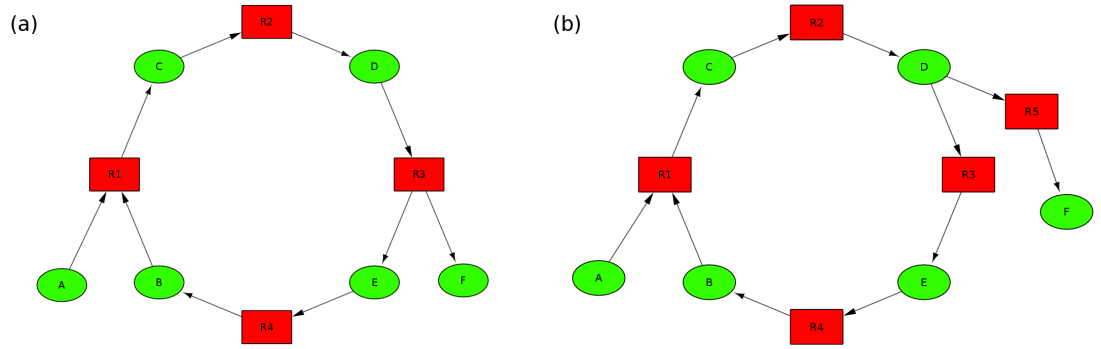


FIGURE 2.8: **Deux exemples de réseaux métaboliques contenant des cycles.** Les métabolites sont représentés par des ovales verts, les réactions sont représentées par des rectangles rouges. La direction des réactions est représentée par la direction des arcs. La stœchiométrie des réactions n'est pas représentée.

La partie (b) représente également un cycle formé de quatre réactions métaboliques ainsi que d'une réaction de sortie. Ici l'amorçage du cycle par le métabolite A fonctionne toujours. En revanche, pour que le cycle puisse avoir lieu dans son intégralité et continue de produire le métabolite F, il faudra que la réaction R2 produise au moins deux molécules de D. Une d'entre elles sera transformée en E pour permettre au cycle de continuer. L'autre sera transformée en F qui pourra être utilisé par le reste du réseau métabolique. Si ce n'est pas le cas, et que la réaction R2 produit moins de deux molécules de D, le cycle ne pourra pas être maintenu et le métabolite F ne pourra pas s'accumuler.

Accessibilité topologique (*Meneco*) La recherche de productibilité par *Meneco* se fera uniquement de manière topologique. Nous allons ici chercher un chemin à travers le graphe bipartite des réactions métaboliques. Pour qu'un métabolite soit productible, il faut que l'ensemble des réactants d'au moins une réaction qui le produit soit présent. Cette définition de la productibilité a l'avantage de ne pas prendre en compte la stœchiométrie des réactions qui n'est pas toujours exacte dans les bases de données, notamment pour ce qui concerne les cofacteurs. Par contre, cette définition a l'inconvénient de créer des problèmes lors de la présence de cycles. Ainsi, si les exemples présentés en figure 2.8 conduisent à une production du métabolite F en FBA, ce n'est pas le cas lors de l'utilisation de la sémantique de *Meneco*.

En effet, dans l'exemple 2.8(a), le métabolite A ne pourra pas produire B car cela nécessite la présence préalable de E, qui n'est pas présent ici. Il en va de même pour l'exemple 2.8 (b). En revanche si une autre molécule du cycle est produite à un autre endroit du réseau, cela ne pose plus de problème pour ce cycle. Par exemple, si une réaction du réseau produit la molécule E, celle-ci permettra d'alimenter la réaction R4 pour produire B rendant la réaction R1 possible. Dans ce cas, les exemples (a) et (b) permettront de produire F. Cela ne serait pas forcément en contradiction avec la FBA, dans le cas (b) et une production d'une seule molécule de D par R2, étant donné que nous avons une alimentation extérieure du cycle.

Accessibilité avec recyclages internes (Pitufo) Pour tenter de résoudre le problème des cycles, nous pourrions utiliser la méthode développée dans le cadre de recherche de précurseurs par [CMA+08]. Un précurseur est un ensemble de métabolites minimal suffisant pour produire un ensemble de métabolites cibles. Pour cela les auteurs définissent un type de molécules "continuellement disponibles" qui, lorsqu'ils sont présents dans un cycle et produits par ce même cycle, pourront être produits en continu. Ainsi, dès qu'un cycle simple est obtenu, l'ensemble des réactions de ce cycle est réalisable et l'ensemble des produits de ce cycle est réellement produit. Dans les exemples précédents, dès que la molécule A est présente dans le réseau, l'ensemble des réactions est dès lors atteignable et l'ensemble des molécules impliquées est productible, peu importe la stœchiométrie des réactions. Si la méthode Pitufo ne s'applique pas directement à complétion de réseaux métaboliques, la sémantique de productibilité utilisée est intéressante à étudier dans ce cas.

2.5.2 Comparaison des différentes sémantiques

Nous l'avons vu, la productibilité semble dépendre grandement de la sémantique utilisée. Pour clarifier ce point, nous allons ici comparer la productibilité de molécules dans cinq exemples qui semblent correspondre à l'ensemble des cas possibles que l'on peut rencontrer. Les métabolites sont représentés par des ronds et les réactions par des rectangles. La stœchiométrie des réactions est indiquée sur les arêtes. Les molécules S correspondront aux graines et les molécules T aux cibles. L'ensemble de ces cas a été préparé avec l'aide de Sven Thiele et sont présentés en figure 2.9. Le résumé de la productibilité de ces différents exemples est présenté dans la table 2.2. Dans un article de 2009 [dFSKF09], De Figueiredo et ses collaborateurs présentent un cas réel de recherche de productibilité de métabolites, le cycle de Krebs associé à la glycolyse/néoglucogénèse chez l'humain. Ils montrent dans leur article que les méthodes basées sur une étude uniquement topologique de réseaux métaboliques résultent habituellement en une sur-approximation des molécules considérées comme productibles ces méthodes. Le réseau présenté en exemple dans cet article peut être décomposé en les différents exemples suivants, montrant que la sémantique topologique simple ne sur-approxime pas les résultats mais, à l'inverse, prédit certains métabolites comme étant non productibles alors que les méthodes numériques les considèrent comme productibles du fait de l'auto-alimentation du cycle.

TABLE 2.2: **Résultat de productibilité des métabolites T à partir de S pour les exemples présentés en figure 2.9.**

Cas	**(a)**	**(b)**	**(c)**	**(d)**	**(e)**
FBA	Productible	Non productible	Productible	Non productible	Productible
Meneco	Productible	Non productible	Non productible	Non productible	Non productible
Pitufo	Productible	Non productible	Productible	Productible	Productible

Dans l'exemple 2.9 (a), le cas est simple. Nous avons une réaction qui transforme une molécule graine en une molécule cible. Pour toutes les sémantiques utilisées, la cible sera toujours productible.

Dans l'exemple 2.9 (b) pour pouvoir produire T, il faut être en présence de deux molécules, S et C_1. S étant une molécule graine, celle-ci est présente. En revanche C_1 n'est pro-

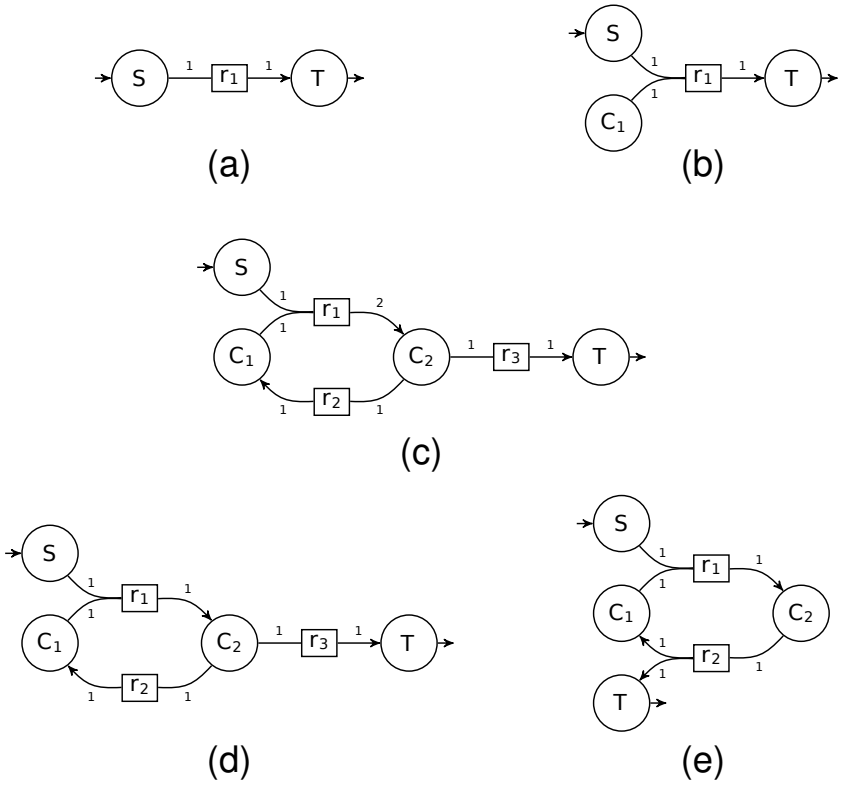


FIGURE 2.9: **Exemples de réseaux métaboliques.** Les cercles représentent les métabolites, les rectangles représentent les réactions. Les chiffres sur les arcs représentent la stœchiométrie des réactions. Les métabolites S correspondent aux sources et les métabolites T correspondent aux cibles.

duite par aucune réaction et ne pourra donc pas être produite, impliquant une impossibilité de produire T que ce soit en FBA, par $Meneco$ ou en utilisant la sémantique Acuna - Sagot.

Nous rencontrons, dans le cas 2.9 (c), notre premier cycle et la première différence entre $Meneco$ et la FBA. Dans le cas présent, en FBA, la réaction r_1 va produire deux molécules de C_2. La première va pouvoir être utilisée pour alimenter le cycle en C_1 via la réaction r_2 et la seconde pourra être utilisée par la réaction r_3 pour produire T. À l'état stable et avec uniquement une entrée de S dans le système, le métabolite cible pourra donc être produit. En revanche en utilisant $Meneco$, pour produire C_2, il faudra que S et C_1 soient présents. Or C_1 ne peut être produit qu'à partir de C_2, le cycle ne pourra donc jamais être amorcé et T ne pourra jamais être produit. Une production de C_1 à un autre endroit du réseau rendrait en revanche le cycle actif et permettrait de produire T. En utilisant Acuna - Sagot, les métabolites C_1 et C_2 sont "continuellement disponibles" et permettent donc aux réactions r_1 et r_2 de fonctionner, ce qui alimente la réaction r_3 en substrat.

La différence entre le réseau 2.9 (d) et le précédent consiste en la production d'un seul métabolite C_2 par la réaction r_1. $Meneco$ ne prenant pas en compte la stœchiométrie des réactions, T ne sera pas produit à partir de S ici non plus. En revanche nous avons une différence en utilisant la FBA, étant donné qu'une seule molécule de C_2 est produite, le métabolite C_2 ne pourra plus, à la fois, entretenir le cycle et être utilisé par la réaction r_3. T ne sera donc pas non plus productible en FBA. D'après Acuna - Sagot, les métabolites C_1 et C_2 sont disponibles et permettent donc d'alimenter r_3 malgré la stœchiométrie qui devrait interdire cela.

Enfin, dans l'exemple 2.9 (d), encore une fois la FBA donnera T comme productible et $Meneco$ considérera T comme non productible. En effet en FBA, à l'état stable, le cycle se maintiendra de lui même par une entrée de S et la réaction R_2 qui produit $C_1 + T$ ne "consommera" pas de matière à proprement parlé. En revanche, pour que T soit productible par $Meneco$, il faudra comme dans l'exemple précédent que C_2 soit productible or celui-ci étant produit par $C_1 + S$, ce ne sera pas le cas car C_1 nécessite C_2 pour être produit. Encore une fois, avec Acuna - Sagot, C_1 et C_2 sont "continuellement disponibles" entrainant une productibilité de T par r_2.

Cette étude des différents exemples d'analyses qualitatives indique que la sémantique de productibilité utilisée par $Meneco$ peut être vue comme une sur-approximation de la sémantique quantitative. Une complétion basée sur ce critère sera globalement plus restrictive étant donné qu'elle empêche d'alimenter des cycles par des métabolites internes. On risque donc de créer des faux négatifs lors de l'identification des métabolites productibles.

Au contraire, la sémantique autorisant les alimentations de cycles par leurs métabolites internes risque ici de produire des faux positifs. En effet en ne prenant pas en compte la stœchiométrie des réactions, il y a de grandes chances de rendre productibles des métabolites qui ne le sont pas réellement.

2.5.3 Comparaison des sémantiques qualitatives sur la complétion

Une implémentation en ASP de la complétion de réseaux métaboliques utilisant le critère de productibilité topologique avec recyclage interne a été proposée par Sven Thiele. Les solveurs ASP ne pouvant, pour le moment, utiliser que des valeurs entières, il n'a pas été possible de tester une complétion s'appuyant sur une sémantique de productibilité par FBA. Cependant des efforts sont actuellement réalisés pour intégrer la programmation MILP dans

ces solveurs, ouvrant la voie à de futures recherches sur le sujet.

Pour mieux comprendre l'impact de la sémantique de productibilité, nous avons d'abord considéré un critère de performance vis-à-vis de la taille de la base de référence. Les tests ont donc été réalisés avec le jeu décrit dans la section 2.2, c'est-à-dire le même jeu de données que celui utilisé pour tester l'amélioration de l'efficacité dû au changement de méthode de recherche (Unclasp VS Clasp). Le programme ASP n'ayant pas été particulièrement optimisé, nous n'allons pas comparer les temps de calcul mais uniquement les résultats obtenus, et notamment le nombre de réactions à ajouter pour rendre tous les métabolites cible productibles ainsi que l'union de ces complétions. La figure 2.10 présente la taille médiane des optimaux et de l'union des solutions proposées par les complétions utilisant la sémantique de *Meneco* et la sémantique topologique avec recyclage interne. La taille médiane de l'union utilisant cette dernière sémantique, pour une base de données de taille 10.000, n'est pas disponible, les calculs n'ayant pas terminés après deux semaines. D'autre part la répartition statistique des unions des complétions est présentée en figure 2.11.

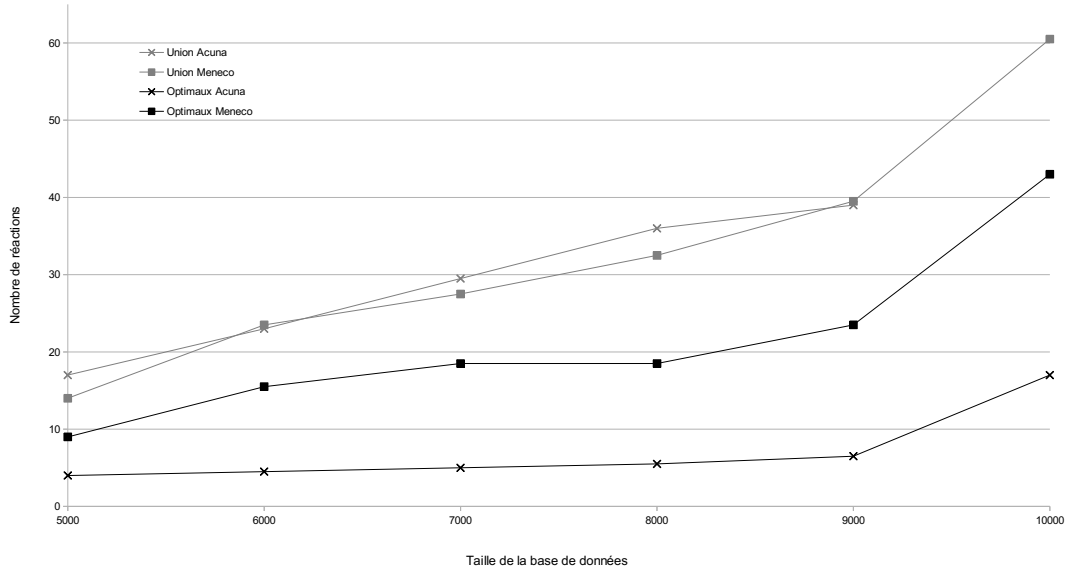


FIGURE 2.10: **Tailles des solutions et de l'union des solutions selon la sémantique de productibilité utilisée.** Les croix correspondent aux expériences utilisant la productibilité d'Acuna-Sagot, les carrés celle de *Meneco*. Les courbes noires correspondent à la taille des optimaux, les courbes grises correspondent à la taille des unions. Il n'existe pas de valeur pour l'union avec Acuna pour une base de données de 10.000 réactions, les calculs étant trop long.

Comme nous pouvions nous y attendre, le changement du critère de complétion en prenant en compte les cycles réduit énormément la taille de chacune des complétions. En revanche nous pouvons remarquer que la taille de l'union des complétions est beaucoup plus variable et reste dans le même ordre de grandeur si l'on prend les critères de productibilité de *Meneco* ou de Pitufo. Cela pourrait être dû au fait que, avec ce critère de productibilité, nous avons plus d'endroits possibles où réaliser la complétion. Pour étudier la composition de ces unions, et comparer d'un point de vue qualitatif les deux complétions,

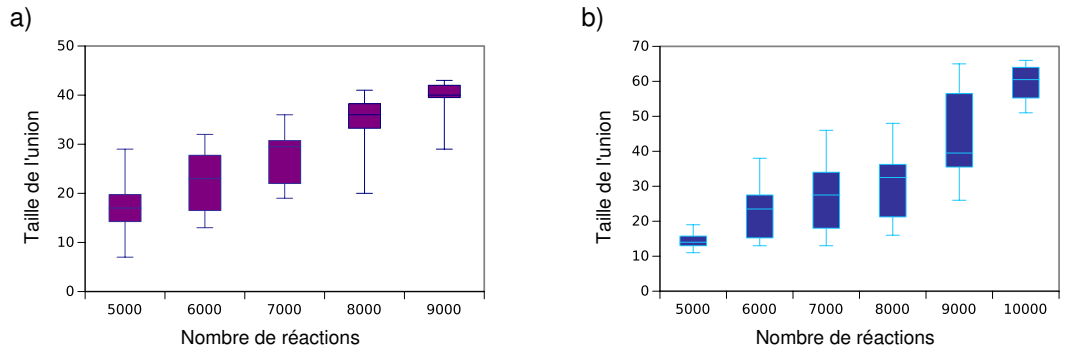


FIGURE 2.11: **Représentation de la répartition statistique des différentes complétions en fonction de la taille de la base de données utilisée.** La partie a) correspond à la complétion utilisant la sémantique proposée par Acuna et Sagot, la partie droite correspond à la complétion utilisée par *Meneco*. Les boites correspondent à la taille du premier et du troisième quartile, la ligne centrale correspond à la moyenne et les barres correspondent aux valeurs extrêmes.

nous avons donc regardé la composition de l'intersection des complétions. Ces intersections sont en accord avec cette hypothèse, celles-ci étant souvent de taille nulle et toujours de taille inférieure à deux. Il y a donc extrêmement peu de redondance lorsqu'on complète un réseau avec une sémantique autorisant les cycles, au contraire d'une sémantique topologique simple où les complétions se recoupent énormément (cf figure 2.5).

Les complétions réalisées en utilisant la sémantique de productibilité topologique avec recyclage semblent donc être des complétions petites en taille et distribuées à plusieurs endroit du réseau. Cela est supporté par le fait que les complétions minimales sont de petite taille par rapport à celles proposées par la sémantique de $Meneco$ (2.10) , alors que les unions de complétions sont de taille similaire pour les deux sémantiques. De plus, le fait que l'intersection des complétions réalisées avec cette sémantique soit très petite voire nulle indique une complétion répartie en plusieurs endroits du réseau. À l'inverse les complétions réalisées par $Meneco$ semblent beaucoup plus "stables" avec des optimaux de plus grande taille, un cœur de réactions représenté par les intersections de grande taille également et une union petite comparée à la taille des optimaux (2.10).

D'un point de vue biologique, il semble plus logique d'utiliser une méthode complétant le réseau métabolique à un endroit précis avec différentes possibilités pour finir cette complétion (comme cela est le cas avec $Meneco$), plutôt que de compléter un réseau à beaucoup d'endroits différents (comme cela semble être le cas avec la sémantique topologique avec recyclage).

Enfin du point de vue de l'efficacité informatique, ce dernier critère de productibilité s'avère beaucoup plus long à calculer, étant donné qu'il y a beaucoup plus d'endroits différents à compléter. Concrètement, pour des bases de données à plus de 10000 réactions, les calculs de complétions selon la sémantique de Acuna-Sagot ne terminent pas en un temps raisonnable (plusieurs jours de calculs nécessaires).

Dans la suite du document, nous nous sommes donc concentrés sur l'utilisation de la sémantique de $Meneco$. Cependant, il faut garder à l'esprit que, une fois résolu la question

des performances pour le calcul des complétions avec la sémantique topologique avec recyclage, il sera intéressant d'en utiliser les résultats pour proposer des réactions complémentaires pour la complétion des effets de cycles. Ce point sera rediscuté en perspectives.

2.5.4 Fonctionnalité des complétions qualitatives

Il reste maintenant à étudier les pertes de fonctionnalité induites par le fait d'utiliser une méthode qualitative. En effet, si la complétion basée sur un point de vue topologique nous permet d'obtenir des résultats précis et rapides, il n'est pas évident que ceux-ci soient corrects d'un point de vue fonctionnel et numérique.

Il a donc été choisi de comparer les résultats de complétion obtenus d'un point de vue "topologique" à des méthodes numériques basées sur l'analyse en balance de flux (ou Flux Balance Analysis, FBA) et l'analyse par variabilité de flux (ou Flux Variability Analysis, FVA). Les jeux de test s'appuyant sur un réseau d'*E. Coli* sont décrits en partie 2.2.3.

Une fois les réseaux métaboliques dégradés, la phase de reconstruction est réalisée en utilisant *Meneco*. La base de données de réactions utilisée correspond au réseau non dégradé. Les graines correspondent aux métabolites pouvant être importés dans le modèle et les cibles correspondent aux métabolites de la fonction de biomasse qui peuvent être produits topologiquement à partir des graines. La complétion est réalisée avec *Meneco* de manière à obtenir l'union des solutions optimales.

L'analyse de la qualité de la complétion se basera sur les différentes classes de réactions définies en FVA : les réactions obligatoires, bloquées et alternatives. En effet, la complétion devant être minimale mais rester fonctionnelle, une complétion "parfaite" correspondrait à un ajout de l'ensemble des réactions obligatoires, aucune réaction bloquée et une faible proportion de réactions alternatives.

La figure 2.12 présente les résultats globaux de cette analyse.

Nous remarquons que, globalement, les reconstructions ont bien fonctionné. À 10% de dégradations, nous ajouterons l'ensemble des réactions obligatoires dans le réseau dans 88% des cas . À 20% de dégradation, nous ajouterons toujours l'ensemble des réactions obligatoires ayant été retirées du réseau dans 87% des cas. À 40% de dégradation ce pourcentage descend à 65% et 57% à 40% de dégradation. Ainsi, assez logiquement, plus le réseau a été détérioré, plus il est compliqué de le compléter d'un point de vue fonctionnel.

Il en va de même pour les réactions bloquées. Dans ce cas plus le réseau sera détérioré plus le nombre de réactions bloquées que l'on rajoutera lors de la complétion sera élevé.

Nous l'avons mentionné précédemment, une complétion "parfaite" correspondrait à une complétion contenant 100% des réactions obligatoires ayant été retirées et aucun ajout de réaction bloquée n'aurait été effectuée. C'est le cas dans 47% des complétions réalisées avec une dégradation à 10%.

Concernant les réactions alternatives, nous remarquons que la proportion ajoutée augmente légèrement avec la dégradation, tout en restant très faible. Ce faible nombre valide le choix de rechercher des complétions minimales en taille que nous avions fait. Le fait que la proportion augmente avec le pourcentage de dégradation s'explique probablement par le fait que nous considérons l'union des solutions minimales. En effet, une forte dégradation du réseau revient à retirer des réactions un peu partout dans le réseau. *Meneco* risque alors de trouver un grand nombre d'endroits à compléter de manière minimale ce qui agrandira la taille de l'union des complétions minimales.

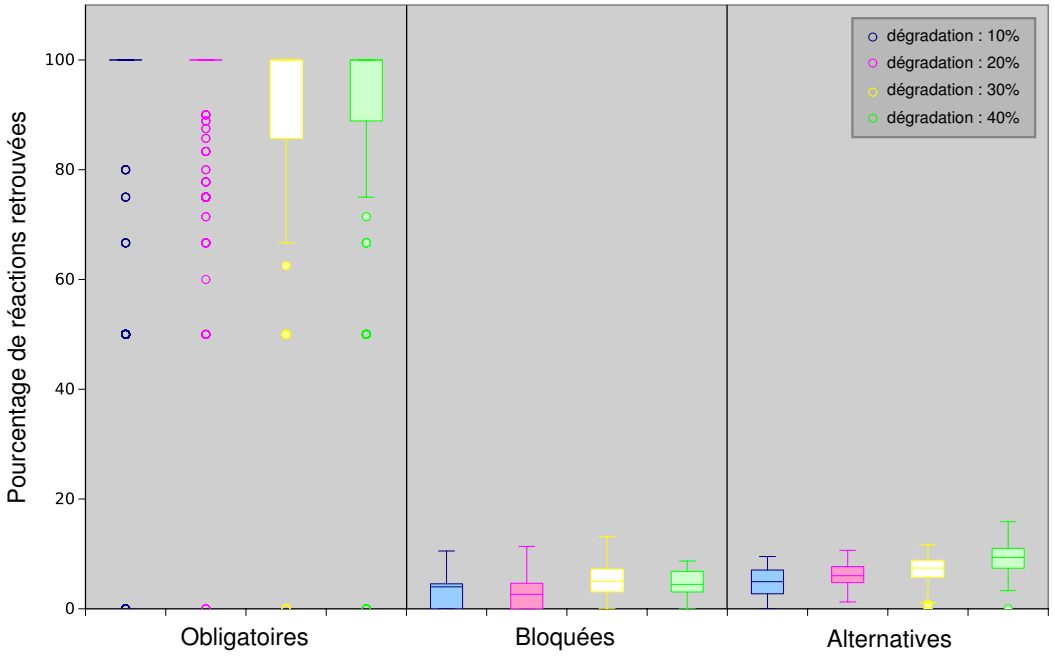


FIGURE 2.12: **Distribution des pourcentages de réactions rajoutées aux réseaux dégradés selon les différentes classes de réactions et le pourcentage de dégradation.** Les boites représentent l'espace interquartile, les moustaches représentent 1.5 espace interquartile. Les ronds représentent les valeurs extrêmes sûres (vides) ou potentielles (pleins).

Nous avons également étudié, pour chacune des 90 fonctions objectives définies, ce qu'il se passe si nous essayons de compléter un réseau vide à l'aide du réseau initial d'*E. coli*. Cela correspondrait à une dégradation de 100% du réseau. Nous allons ainsi identifier, dans le réseau initial, uniquement les réactions nécessaires à la production de biomasse.

De manière globale sur l'ensemble des fonctions objectives étudiées, nous retrouvons ainsi 94% des réactions obligatoires, 5% de réactions bloquées et 14% de réactions alternatives. La figure 2.13 représente la distribution statistique sur l'ensemble des 90 fonctions objectives.

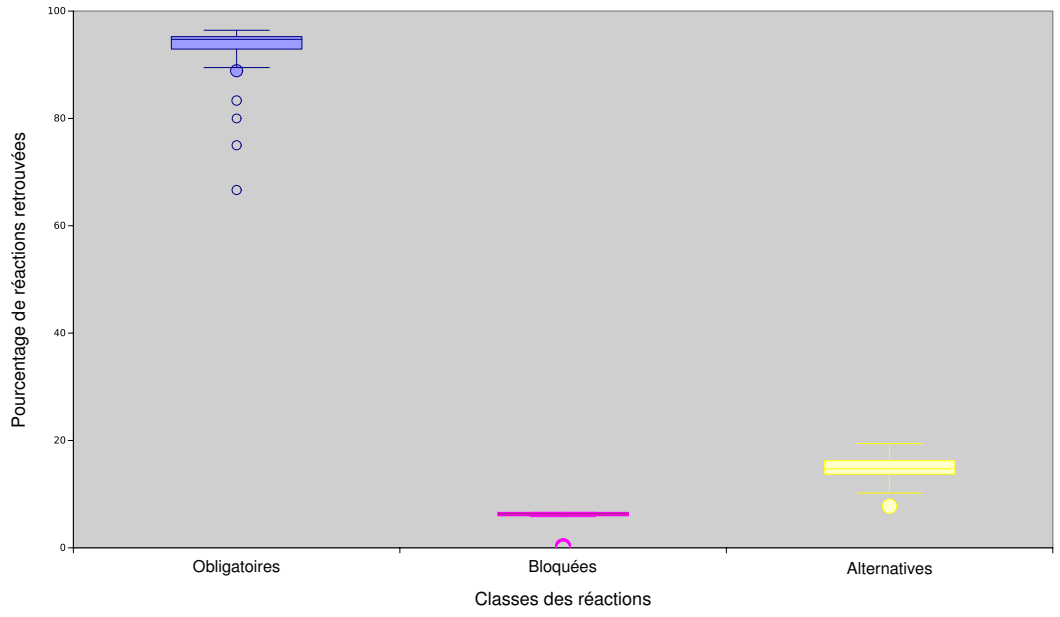


FIGURE 2.13: **Distribution des pourcentages de réactions rajoutées au réseau vide selon les différentes classes de réactions.** Les boites représentent l'espace interquartile, les moustaches représentent 1.5 espace interquartile. Les ronds représentent les valeurs extrêmes sûres (vides) ou potentielles (pleins). Les réactions essentielles à la production de biomasse (au sens de la FVA) sont représentées en bleu, les réactions bloquées qui ne possédent jamais aucun flux non nul en FVA sont représentées en violet, les réactions alternatives pouvant posséder un flux nul ou non nul lors d'optimisation de la fonction objective sont représentées en jaune.

Il est intéressant de remarquer que dans la majorité des réseaux ainsi reconstruits, une réaction obligatoire manque. Il s'agit de la réaction irréversible appelée "inorganic diphosphatase" qui correspond à la réaction $H_2O + diphosphate \rightarrow H^+ + phosphate$. Cette réaction peut être importante d'un point de vue quantitatif pour produire le phosphate qui sera consommé dans énormément de réactions. En revanche d'un point de vue topologique simple cette réaction n'est pas obligatoire, d'autres réactions existant déjà permettant de produire du phosphate.

2.6 conclusion

Au cours de ce chapitre nous avons étudié la pertinence de la complétion d'une ébauche métabolique d'un point de vue combinatoire vis-à-vis d'une complétion quantitative. Nous avons montré la grande efficacité en terme de temps de calcul de la résolution du problème combinatoire à l'aide d'une approche déclarative (programmation par ensembles-réponses - ASP), notamment en utilisant un nouveau solveur ASP qui a permis d'accélérer grandement les calculs, tout en changeant la modélisation existante précédemment pour mieux prendre en compte la réalité biologique du modèle.

Nous avons également montré que cette complétion topologique d'un réseau métabolique, si elle relâche des contraintes par rapport à une étude quantitative prenant en compte la stœchiométrie des réactions, donne de très bons résultats en pratique. Cette complétion topologique pourrait être améliorée dans la suite en prenant en compte une sémantique topologique avec recyclage interne qui semble proposer des résultats complémentaires que sémantique topologique simple. Pour arriver à intégrer cette sémantique il faudrait tout d'abord résoudre le problème combinatoire associé de manière suffisamment efficace pour pouvoir utiliser des bases de données de réactions de grande taille.

Ce problème combinatoire de complétion de réseaux métaboliques étant résolu, nous nous sommes attelés à l'intégrer dans un pipeline global de reconstruction de réseaux métaboliques qui sera présenté dans le chapitre suivant.

Chapitre 3

Pipeline de reconstruction de réseaux à partir de données hétérogènes

Dans ce chapitre nous allons étudier plus précisément le processus global de reconstruction de réseau métabolique que nous proposons. Ce processus est décomposé en différentes étapes, chacune possédant des spécificités propres que nous expliciterons, notamment dans le cas de la reconstruction de réseaux métaboliques chez des espèces non classiques. Cette thèse se concentrant plus particulièrement sur la reconstruction du réseau métabolique d'*Ectocarpus siliculosus*, nous allons prendre cet organisme comme exemple d'application durant tout ce chapitre.

Ces travaux ont été décrits dans la publication [PCD+14] et présentés sous forme de poster à ECCB 2014. L'application de ces méthodes est également en cours sur d'autres espèces : bactéries associées à l'espèce *Ectocarpus* d'eau douce [DET14], micro-algues vertes, protistes, etc.

3.1 Modèle d'application : *Ectocarpus siliculosus*

Le réseau test sur lequel nous avons testé les différentes méthodes de reconstruction puis proposé des améliorations est le réseau métabolique d'*Ectocarpus siliculosus*, l'organisme modèle pour l'étude de la biologie des algues brunes [HCP+10]. Sur ce modèle, biologique, différentes sources d'informations sont disponibles.

Génome Les algues brunes sont des straménopiles jouant un rôle important dans la zone intertidale. Le génome d'*E. siliculosus* ayant été séquencé et annoté en 2010. Ce génome est constitué de 1.561 supercontigs (ou scaffolds) de plus de 2.000 paires de bases. 97,4% des séquences d'ARN messagers précédemment séquencés sont retrouvés dans ce génome, indiquant une bonne couverture du séquençage réalisé. Durant tout ce chapitre, le terme "annotation" consistera en l'annotation fonctionnelle du produit d'un gène et non de la recherche de modèles de gènes dans le génome. Les annotations du génome d'*Ectocarpus siliculosus* sont une combinaison d'annotations automatiques, réalisée avec le logiciel EuGene, et manuelles [CSR+10]. On a à disposition des numéros EC, des termes GO, des noms de fonctions, de domaine et les séquences protéiques. Tout cela est disponible via une interface web sur le site ORCAE (`http://bioinformatics.psb.ugent.be/orcae/overview/`

`Ectsi`) qui est régulièrement mis à jour avec de nouvelles annotations fonctionnelles de gènes. Pour la reconstruction du réseau, l'ensemble des données a été téléchargé à partir ce site web le 21 juin 2013. Ces informations vont constituer une première sources pour reconstituer le réseau métabolique d'*Ectocarpus siliculosus*.

Réseau métabolique de référence En complément aux annotations du génome d'*Ectocarpus siliculosus*, pour exploiter les méthodes de reconstruction de réseau métabolique s'appuyant sur un réseau de référence, nous avons recherché un réseau métabolique de bonne qualité (incluant une curation manuelle étendue) et pas trop éloigné phylogénétiquement parlant de l'espèce étudiée. Nous avons choisi d'utiliser le réseau réseau métabolique d'*Arabidopsis thaliana*, AraGEM [dODQP$^+$09], un réseau global qui satisfait les critères listés précédemment. Ce choix a été réalisé car, au moment de la réalisation de l'étude, ce réseau était le plus complet des réseaux d'*Arabidopsis thaliana*. Il est composé de 1.567 réactions et 1.748 métabolites. Ce réseau a été reconstruit en se basant sur des études bibliographiques et un réseau précédent et a été réalisé en utilisant les identifiants présents dans KEGG. Il eut été plus aisé pour la suite de l'étude d'utiliser le réseau AraCyc [MZR03] étant donné que les identifiants auraient été les mêmes entre notre réseau métabolique, basé sur la base de données MetaCyc, et ce réseau basé sur la même base de données. Cependant le réseau AraGEM a été choisi pour sa plus grande qualité, il intègre en effet la plupart des informations contenues dans AraCyc, une curation manuelle étendue ayant été réalisée en plus.

Définition du milieu de culture *Ectocarpus siliculosus* est un organisme vivant dans de l'eau de mer. L'identification de l'ensemble des métabolites présents dans ce milieu de croissance étant compliquée, nous avons choisi de prendre comme milieu de croissance l'eau de mer artificielle utilisée en laboratoire pour cultiver cette algue. Il s'agit d'un milieu de culture nommé Provasoli et dont la composition est détaillée en annexe A dans la colonne "Graines" en association avec quelques cofacteurs. Ce Provasoli contient essentiellement des ions, du nitrate et des vitamines.

Profilage métabolique Des données de profilage métabolique ont été obtenues au cours de plusieurs études [TEP$^+$11] et ont permis d'identifier 51 métabolites que l'on sait être productibles par l'algue (des acides aminés, des acides gras, des sucres, et des polyalcools). Ces métabolites sont listés en annexe A.

La productibilité de ces métabolites par le réseau métabolique sera le critère principal de validation de la fonctionnalité du réseau. Cependant, en étudiant la base de données MetaCyc v17.0 nous nous sommes aperçus que trois d'entre eux (l'acide eicosadienoique, l'acide docosanoique et l'acide erucique) n'étaient produits par aucune réaction de cette base de données. En revanche, deux réactions de KEGG (R08190 et R08184) permettent la production de l'acide eicosadienoique et de l'acide docosanoique mais n'était pas présentes dans la base MetaCyc. Ces deux réactions ont donc été ajoutées à la base de données de référence. Au final, 50 métabolites, parmi les 51 caractérisés chez *Ectocarpus siliculosus*, doivent être productibles par le réseau.

Critères de fonctionnalité Dans ce but, nous avons utilisé *Meneco* pour vérifier que ces cibles étaient productibles à partir des métabolites graines (milieu de culture), selon la sémantique décrite au chapitre 2. Etant donné les limites discutées précédemment de ce cri-

tère qualitatif de productibilité, après la reconstruction globale du réseau, nous avons vérifié que ces métabolites étaient aussi productibles d'un point de vu quantitatif (FBA). Cette vérification quantitative n'a été effectuée qu'à la fin de la complétion du réseau pour plusieurs raisons. Tout d'abord il nous a semblé plus naturel d'effectuer une analyse topologique du réseau avant de s'intéresser aux détails liés à la stœchiométrie des réactions. De plus nous avons intuité que les cycles métaboliques les plus importants du réseau ont été reconstruits lors de la toute première étape de reconstruction du réseau métabolique, les gènes "importants" dans le métabolisme ayant souvent été bien annotés manuellement. D'autre part ASP ne permet pas, actuellement, de réaliser des calculs sur l'ensemble des chiffres réels mais uniquement sur les entiers, rendant difficile voir impossible le fait de tester la productibilité de molécules selon des critères quantitatifs tout en profitant de la très grande efficacité des solveurs ASP.

Données transcriptomiques Enfin, nous possédons des données de transcriptomique et protéomique [TEP+11] qui permettront de valider certaines prédictions de réactions métaboliques intégrées dans le réseau.

3.2 Construction d'une ébauche du réseau : gestion des données hétérogènes

3.2.1 Reconstruction à partir des annotations

L'exploitation des annotations a nécessité un prétraitement important et fastidieux. En effet, les annotations de gènes étaient uniquement disponibles en ligne sous forme de pages html, avec une page par gène. L'ensemble de ces pages est accessible à l'adresse `http://bioinformatics.psb.ugent.be/orcae/overview/Ectsi`. Ces annotations proviennent en grande majorité de recherches de fonctions réalisées manuellement. Les pages pour chaque gène ont donc été créées manuellement et ne sont pas du tout homogènes. Certaines contiennent des informations très précises avec le numéro E.C., des données ontologiques en lien avec GO, des liens vers des réactions présentes dans KEGG, un nom précis et un nom générique, une description précise du gène ou encore une descriptions des différents domaines protéiques retrouvés à partir de la séquence de la protéine prédite. Pour d'autres pages seul un nom de gène plus ou moins précis et non normalisé sera disponible. Il a donc fallu recueillir les différentes informations disponibles au niveau de l'annotation pour pouvoir faire le tri entre les plus précises et celles pouvant être laissées de coté. L'accès direct à la base de données regroupant toutes ces informations nous ayant été refusé, le choix a été fait de télécharger l'ensemble des pages HTML regroupant les annotations et de parser ces fichiers à l'aide de scripts spécialement créés pour l'occasion. Cela a également permis de créer directement des fichiers dans le bon format pour permettre une intégration directe dans la suite de logiciels Pathway tools qui nécessite d'avoir soit des fichiers "maison" contenant l'ensemble des informations disponibles, soit directement des fichiers GeneBank.

La version 17.0 du logiciel *Pathway Tools* [KPR02] a été utilisée pour la reconstruction du draft du réseau métabolique d'*E. siliculosus* à partir des annotations obtenues sur le site internet ORCAE. Ce logiciel permet d'interpréter l'ensemble des informations disponibles dans les annotations pour retrouver quelle(s) réaction(s) de la base de données MetaCyc

(version 17.0) correspond à quelle(s) annotation(s). Le choix s'est basé sur ce logiciel de par sa forte complémentarité avec la base de données MetaCyc et son efficacité dans la gestion d'annotations peu précises, comme lorsque l'on a uniquement le nom d'un gène par exemple.

Cela a permis de reconstruire un premier draft métabolique que l'on nommera EctoGEM-*annot* contenant 1677 réactions et 1889 métabolites.

Comme expliqué précédemment, *Meneco* permet de vérifier la productibilité des 50 métabolites cibles à partir du milieu de culture. Le réseau EctoGEM-*annot* est capable de produire 25 de ces 50 cibles.

Cela suggère que les annotations du génome ne suffisent pas à reconstruire un réseau métabolique fonctionnel. On peut par exemple citer les enzymes appelées "élongases" qui permettent de produire des acides gras poly-insaturés. Celles-ci ne sont annotées que partiellement du fait de la difficulté d'établir leur spécificité à partir seulement de la séquence protéique. Par conséquent, aucun numéro EC ne leur est attribué, et les réactions associées n'ont pas pu être ajoutées au réseau.

3.2.2 Reconstruction à partir d'un réseau métabolique de réference

Une seconde ébauche de réseau métabolique a été reconstruite en n'utilisant plus l'expertise manuelle des spécialistes sur les annotations mais en considérant l'expertise manuelle des spécialistes sur la reconstruction de réseaux métaboliques. L'ensemble des protéines codant pour les réactions contenues dans le réseau métabolique d'*Arabidopsis thaliana*, AraGEM [dQP$^+$10b], ont ainsi été identifiées. Une recherche d'orthologie a ensuite été effectuée entre ces protéines et les protéines prédites chez l'algue brune afin d'extrapoler ces résultats d'orthologie à l'ajout de réactions dans l'ébauche. Cette recherche a été effectuée par les outils InParanoid (version 4.1) [RSS01] et OrthoMCL (version 2.0.9) [LSR03] puis ces résultats ont été combinés en utilisant le logiciel Pantograph (version 0.1.1) [NL13].

L'ébauche métabolique obtenue, appelée EctoGEM-*ortho*, est composée de 786 réactions et 1767 métabolites. *Meneco* indique que le réseau EctoGEM-*ortho* n'est capable de produire aucun des métabolites d'intérêt. Ceci s'explique probablement par la distance phylogénétique entre *Ectocarpus siliculosus* et *Arabidopsis thaliana*, qui a rendu la recherche d'orthologues difficile.

3.2.3 Complémentarité des deux réseaux : outils pour leur fusion

Une fois EctoGEM-*annot* et EctoGEM-*ortho* construits, nous avons choisi de fusionner l'ensemble des deux réseaux pour profiter à la fois de la qualité des annotations manuelles réalisées sur le génome et de la qualité de la reconstruction manuelle réalisée avec le réseau métabolique d'*Arabidopsis thaliana*. Les deux réseaux ont été fusionnés en un réseau appelé EctoGEM-*combined*.

Fusionner deux réseaux métaboliques créés à partir de bases de données de réactions différentes est particulièrement difficile pour une raison simple, les identifiants des réactions et des métabolites ne seront pas les mêmes dans les deux bases de données. Pour résoudre ce problème et réussir à faire correspondre les identifiants de deux métabolites ou de deux réactions identiques, deux outils ont été créés par Guillaume Collet. MeMap transforme chaque identifiant de réaction provenant de MetaCyc (ou possédant un lien depuis

MetaCyc vers une autre base de données) en un identifiant interne. MeMerge fusionne ensuite simplement les différents réseaux qui possèdent désormais des identifiants communs.

Pour réussir à faire correspondre les différents identifiants, ces outils utilisent largement les informations contenues dans la base de données MetaCyc. En effet, cette base de données a l'intérêt énorme de posséder beaucoup de lien entre chaque entité existante et les mêmes entités dans d'autres bases de données. Quand un lien direct existe depuis une réaction présente dans MetaCyc vers une réaction présente dans Kegg, cette association est assez aisée à trouver. Quand ce n'est pas le cas, nous allons pouvoir travailler au niveau des métabolites, en considérant que deux réactions possédant les mêmes réactants et les mêmes produits, il est raisonnable de considérer qu'il s'agit d'une réaction unique. Là encore le lien entre les métabolites pourra se faire directement depuis MetaCyc vers Kegg à l'aide des références croisées entre ces deux bases de données, ou en passant par une base de données intermédiaire spécialisée dans les molécules biochimiques, ChEBI [DdME+08]. Cette base de données contiendra la composition chimique des molécules, des liens externes vers d'autres bases de données mais surtout les synonymes qu'auront les molécules. Ces synonymes permettront de faire le lien entre les molécules présentes dans les différentes bases de données et donc le lien entre les réactions.

Le réseau EctoGEM-$annot$ contient désormais 1785 réactions et 1981 métabolites.

Nous pouvons donc considérer que la recherche d'orthologues a permis d'ajouter 108 réactions par rapport aux annotations, ce qui prouve l'intérêt de cette seconde étape. En revanche, l'ajout de ces réactions n'a rien changé pour la fonctionnalité du réseau ; vis à vis des 50 métabolites d'intérêt, nous avons toujours 25 d'entre eux qui ne peuvent pas être produits.

3.2.4 Conclusion

Ces deux reconstructions d'ébauches métaboliques et leur association ont permis d'utiliser la majorité des informations disponibles dans le génome de l'espèce étudiée. Il en ressort que ces informations ne sont pas suffisantes pour reconstruire un réseau métabolique fonctionnel, lorsque nous sommes en présence d'espèces non classiques et éloignées phylogénétiquement des espèces modèles classiques. Cela est en partie dû au fait que nous nous retrouvons en permanence face à des problèmes de formalisation des données biologiques, des problèmes d'unification d'identifiants entre différentes bases de données, etc. Le développement d'approches de réconciliation efficaces va rapidement devenir nécessaire pour éviter la perte d'information et la multiplication de réseaux métaboliques possédant chacun leurs propres conventions de nommage des identifiants de réactions ou de métabolites.

3.3 Complétion

3.3.1 Identification de réactions à ajouter

Seuls 25 des 50 métabolites d'intérêt pouvant être produits par le réseau résultant de la fusion des deux ébauches métaboliques, il semble donc évident que certaines réactions manquent dans le réseau. Le logiciel $Meneco$ va être utilisé une fois de plus. Ici nous ne nous contenterons pas de vérifier la productibilité de molécules, mais nous allons nous servir de la seconde fonctionnalité de $Meneco$ qui consiste en la résolution du problème combina-

toire traité au chapitre 2. Pour cela, nous avons besoin de quatre sources de données différentes :

- **Un draft métabolique** : le réseau provenant de la fusion précédente
- **Une base de données de réactions métaboliques** : MetaCyc 17.0
- **Une liste de métabolites graines** : les molécules entrant dans la composition de l'eau de mer artificielle
- **Une liste de métabolites cibles** : une liste de 51 métabolites identifiés par différentes techniques biologiques

En suivant un principe de parcimonie, *Meneco* va calculer le nombre minimal de réactions provenant de la base de données à ajouter au réseau pour permettre la production des métabolites cibles à partir des métabolites graines. Étant donné que l'on peut considérer que les cofacteurs sont présents (au moins en quantité infime) dans toutes les cellules d'un organisme et qu'ils ne peuvent être créés à partir de rien, nous avons choisi de rajouter ces molécules nécessaires à la plupart des réactions dans les métabolites graines. Cette approche est en total accord avec les méthodes classiques de modélisations de réseaux métaboliques basées sur une étude topologique des réseaux [HE07]. *Meneco* va ainsi nous indiquer les métabolites cibles ne pouvant pas être produits à partir du réseau et des métabolites graines. Si ce nombre est non nul, on calculera le nombre minimal de réactions à ajouter au réseau, et on listera l'ensemble des groupes de réactions de taille minimale qui pourront compléter le réseau. Nous pouvons également obtenir l'union et l'intersection de ces sets minimaux de réactions.

Le logiciel *Meneco* a donc été utilisé pour calculer le nombre minimal de réactions à ajouter dans le modèle pour permettre la production de l'ensemble des 50 métabolites cibles (c'est à dire les réactions à ajouter pour produire les 25 autres métabolites). La base de données considérée pour identifier les réactions à rajouter est MetaCyc, à laquelle nous avons ajouté les deux réactions de KEGG précédemment citées. Il est ressorti de cette étude que l'ajout de 44 réactions au moins suffit à rétablir la connectivité du réseau vis à vis des 50 métabolites étudiés. Il existe 4320 ensembles de 44 réactions qui permettent de faire cela. L'union de ces 4320 solutions toutes différentes est formée de 60 réactions, signe d'un très grand chevauchement des solutions.

Les améliorations de *Meneco* présentées dans le chapitre 2, qu'elles concernent l'intégration de la réversibilité des réactions dans la modélisation ou le changement de solveur, ont montré leur intérêt pour permettre de réaliser la complétion en un temps raisonnable. Cependant le nombre de solutions proposées est assez important et nécessite un filtrage qui dépasse les questions purement combinatoires posées précédemment.

3.3.2 Filtrage par analyse sémantique

Pour étudier ce chevauchement nous avons réalisé une étude sémantique des différentes solutions pour observer l'importance de ces réactions vis à vis de la fonctionnalité.

En effet certaines réactions vont être mutuellement exclusives les unes des autres car jamais présentes ensemble dans groupe de réactions. En analysant sémantiquement ces réactions et en étudiant les termes GO associés, nous pouvons identifier certaines réactions qui seraient représentatives d'un ensemble de réactions mutuellement exclusives. Pour cela les ancêtres ontologiques des termes GO ont été étudiés, en vu de réduire la taille de l'union des ensembles de réactions.

Nous pouvons les classer en deux groupes : celles qui interviennent dans toutes les solutions (35), et celles n'étant pas ubiquitaires (25). Ces 25 réactions ont été divisées en sept cliques mutuellement exclusives (figure 3.1), et un groupe de quatre réactions avec trois incompatibilités deux à deux. Dans une clique de 4 réactions mutuellement exclusives, par exemple, dès que l'on a une réaction, nous ne pourrons pas avoir les trois autres dans la même solution. Chacune des 4320 solutions sera donc formée des 35 réactions ubiquitaires, d'une des réactions de cliques et de deux réactions du dernier groupe de quatre. Nous avons donc eu l'intuition que certaines réactions étaient équivalentes, notamment au cœur des cliques.

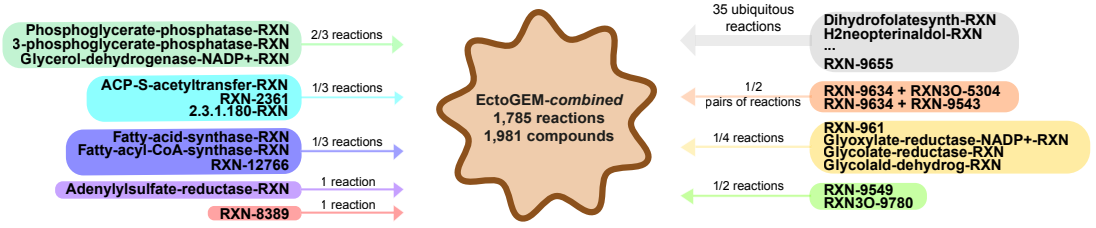


FIGURE 3.1: **Composition des 432 ensembles de 44 réactions candidates à la complétion pour permettre la productibilité des 50 cibles. Chaque ensemble de réactions peut être décomposé en un ensemble de 35 réactions présentes partout auquel nous ajoutons une réaction de chaque groupe représenté, deux réactions du groupe représenté en haut à gauche et une paire de deux réactions.**

Si des cliques contiennent des réactions équivalentes, nous aurons la possibilité de choisir une ou l'autre des réactions de la clique pour limiter l'ajout de réactions non supportées par des preuves biologiques et aussi le nombre de solutions différentes. Cette procédure a permis d'enlever 5 réactions différentes du réseau, qui possédaient déjà leur équivalent, amenant le nombre de solutions différentes de 4320 à 432. Nous avons ajouté ces 55 réactions au réseau précédent pour obtenir EctoGEM-*functional*, qui contient 1840 réactions et 2004 métabolites, et permettent de produire les 50 métabolites identifiés.

3.3.3 Compartimentation

Les annotations du génome d'*Ectocarpus siliculosus* contiennent une information qui n'a pas été prise en compte jusqu'à présent : une prédiction de localisation des protéines effectuée par l'outil HECTAR [GGC08], développé spécifiquement pour les protéines présentes chez les hétérokontes. Cette information a été prise en compte en fin de reconstruction pour définir une localisation supposée de certaines enzymes, bien qu'aucune validation expérimentale n'ai été réalisée chez aucune algue brune à ce jour. D'autre part les gènes présents dans le génome mitochondrial et chloroplastique d'*Ectocarpus siliculosus* ont également servi à prédire la localisation de certaines protéines.

3.4 Curation manuelle du réseau

Afin de valider la reconstruction automatique du réseau métabolique d'*Ectocarpus siliculosus* et d'étudier sa complétude, nous avons effectué une curation manuelle du réseau. Cette curation est basée sur une étude bibliographique par des experts et une étude des profils HMMs pour évaluer l'existence de gènes candidats dans le génome de l'algue. Une attention toute particulière a été apportée aux réactions supportées ni par leurs annotations, ni par des données d'orthologie. Celles-ci sont en effet la cible privilégiée pour de nouvelles annotations géniques amenant à de futures validations fonctionnelles.

3.4.1 Score pour chaque réaction

Pour chaque réaction ajoutée dans le réseau, l'ensemble des enzymes connues dans l'arbre du vivant qui codent pour cette réaction particulière à partir de la base de données ExPASy-ENZYME (`http://enzyme.expasy.org/`) ont été recherchées. Ces séquences enzymatiques ont été alignées par ClustalO (`http://www.clustal.org/omega`) puis des profils HMMs ont été créés pour chaque réaction avec la suite logicielle HMMER (`http://hmmer.janelia.org` version 3.0). La même suite logicielle a été utilisée pour rechercher des séquences correspondantes au profil dans l'ensemble des séquences protéiques prédites par le génome d'*Ectocarpus siliculosus*. Lorsque nous avons obtenu une e-value $\leq 10^{-15}$, nous avons considéré le match comme étant suffisamment sûr pour être présenté aux biologistes pour une éventuelle étude manuelle.

3.4.2 Identification de faux positifs

Parmi ces réactions ne possédant aucune base génétique, nous pouvons citer les réactions ajoutées par Pathway Tools dans le réseau. En effet, lors de la reconstruction initiale, Pathway Tools identifie des voies métaboliques qu'il considère comme étant présentes dans le réseau, que celles-ci soient complètes ou non. Ainsi, pour une voie métaboliques donnée, s'il manque une seule réaction de cette voie, Pathway Tools va probablement ajouter cette réaction au réseau sans associer aucune enzyme. L'ajout de réactions dépendra à la fois de la proportion de réactions manquantes dans une voie donnée et du nombre de voies différentes dans lesquelles sont impliquées les réactions identifiées comme présentes. Ainsi une voie métabolique composée de trois réactions, dont deux ont été retrouvées pourra être considérée comme présente, si ces deux réactions n'interviennent que dans cette voie métabolique, ou absente si ces deux réactions sont associées à de nombreuses voies dans la base de données.

Étant donné la grande distance phylogénétique existant entre *Ectocarpus siliculosus* et les espèces présentes dans MetaCyc, Pathway Tools risque d'inclure un grand nombre de faux positifs. Nous pouvons prendre comme exemple la première voie de dégradation du glycogène (glycocat-PWY dans MetaCyc) formée de sept réactions et considérée comme complète dans EctoGEM-*annot*. Seules trois des réactions sont supportées par des annotations de gènes (numéros E.C. : 2.1.7.1.2, 5.4.2.2 et 3.2.1.20). Deux de ces réactions sont impliquées dans plusieurs autres voies métaboliques comme la dégradation du glucose, du glucose-1-phosphate ou encore de l'amidon. La troisième (3.2.1.20) correspond à une alpha-glucosidase dont la spécificité basée uniquement sur de l'homologie de séquence est difficile à déterminer. Il reste quatre réactions ajoutées par Pathway Tools qui sont très pro-

blablement des faux positifs. Des cas similaires ont été rencontrés dans d'autres voies métaboliques associées au métabolisme des acides aminés. Ces faux positifs ont été identifiés et retirés lors de l'étape de curation manuelle.

3.4.3 Ajout de réactions spécifiques

La curation manuelle a également été utilisée pour rajouter des réactions pour lesquelles les annotations étaient incomplètes et les recherches d'orthologie insuffisantes. On peut par exemple citer l'exemple de la réaction correspondant à la 3-phosphoglycérate-phosphatase, catalysée par une "Purple Acid Phosphatase" (PAP) qui produit du phosphate et du glycérate. Les PAPs sont présentes chez la plupart des eucaryotes, même si les orthologues entre plantes et animaux sont assez éloignés. Aucun gène correspondant aux PAPs n'avait été annoté dans le génome, mais une recherche manuelle a montré que le gène Esi0000_0474 est un bon candidat. Un autre exemple concerne le gène Esi0002_0097. Pathway Tools a considéré qu'une réaction correspondant à une phosphoglycérate-phosphatase devait être présente comme alternative à la production de glycérate, catalysée par une phosphatase alkaline. Le gène Esi0002_0097 est un excellent candidat pour cette enzyme comme le montre les analyses HMMs notamment.

La reconstruction basée sur les analyses d'orthologie apporte également son lot d'informations. On peut par exemple citer deux enzymes, la diamine oxydase (gène Esi0076_0063) et l'alanine aminotransférase (gène Esi0008_0209) qui ont été correctement identifiées par l'analyse des orthologues malgré une annotation insuffisante de ces gènes dans le génome d'*Ectocarpus*. Cette approche n'est pourtant pas parfaite, si l'on regarde par exemple la glycérate kinase impliquée dans la photorespiration. Pour cette réaction, le gène candidat trouvé manuellement est le gène Esi0157_0054. Celui-ci n'a pas été annoté dans le génome d'algue, et aucune protéine n'est associée à cette réaction chez *Arabidopsis thaliana*, empêchant par là même la recherche d'orthologue. Cette réaction et le gène associé ont été ajouté manuellement dans le réseau final.

Enfin, quand cela était possible des données d'expression ont été utilisées pour caractériser certains gènes bien précis. Ainsi, si l'ARN messager d'un gène existe, ceci est un signe que la prédiction du gène est réelle et que nous ne sommes pas en présence d'un pseudogène.

3.4.4 Conclusion

Cette curation manuelle du réseau métabolique a permis de distinguer plusieurs sources de faux positifs dans la reconstruction du réseau, de par l'utilisation de Pathway tools ou encore la présence d'enzymes multi-domaines. Des faux négatifs existant également après la reconstruction des ébauches métaboliques, la complétion du réseau a permis de guider cette curation manuelle lors de la recherche des réactions et des associations gènes-réactions manquantes. Les apports biologiques de cette curation manuelle seront étudiés plus précisément dans le chapitre 4.

3.5 Validation fonctionnelle

3.5.1 Réseau final

Au final le réseau EctoGEM contient 1866 réactions et 2020 métabolites impliqués dans 224 voies métaboliques complètes. Ces chiffres sont du même ordre de grandeur que ceux obtenus pour la reconstruction d'autres réseaux métaboliques d'eucaryotes photosynthétiques :

- *Phaeodactylum tricornutum* : 1719 réactions [FMR+12]
- *Chlamydomonas reinhardtii* : 1500 à 2000 réactions [CMK+09, CGM+11, dQP+10b]
- *Ostreococcus* : ~900 réactions [KYW+12]
- *Arabidopsis thaliana* : 1567 réactions [dQP+10b]

EctoGEM v1.0 est ainsi basé sur l'utilisation de différentes méthodes informatiques complémentaires, complétées par une curation manuelle. Toutes les informations relatives à EctoGEM sont disponibles via un site internet (`http://ectogem.irisa.fr/`) avec les liens vers les différentes annotations génomiques, les données de séquences ou encore les liens vers MetaCyc. Nous considérons que EctoGEM peut être une ressource communautaire, dynamique, flexible et évolutive. En effet n'importe quel chercheur peut désormais avoir accès au réseau et nous donner des informations sur différents aspects du réseau. Un formulaire a été mis en place sur le site internet afin de favoriser et simplifier ce type d'échanges. De plus, l'ensemble des sources d'information ayant amené l'ajout de telle ou telle réaction dans le modèle est disponible, permettant par là même de juger de la crédibilité à apporter à chacune des prédictions.

3.5.2 Analyse par balance de flux

Comme validation globale du réseau EctoGEM-*functional* obtenu nous avons construit une fonction de biomasse spécifique à *Ectocarpus siliculosus*, basée sur une étude bibliographique des résultats de profilage métabolique obtenu chez cette algue. Cette fonction objectif est formée de 30 métabolites issus d'expériences de profilage métabolique décrites dans des publications contenant des valeurs numériques de concentrations de molécules [GDR+10, DGR+11]. Ces 30 métabolites étaient tous présents dans la liste des 50 cibles utilisés pour le gap-filling. La FBA permet de vérifier que le réseau peut produire ces 30 métabolites en quantité équivalente à la réalité biologique. La composition précise de la fonction de biomasse est présentée en table 3.1.

Une fois cette fonction établie, une analyse par balance de flux (décrite précisément en 1.2.2) a été effectuée pour vérifier si le réseau reconstruit est quantitativement capable de produire de la biomasse. Dans le cadre de la reconstruction de ce réseau, nous avons décidé d'utiliser la boite à outil COBRApy [ELPH13] qui utilise le solveur GLPK (GNU Linear Programming Kit).

À l'inverse de *Meneco* qui réalise une analyse topologique qualitative du réseau, les approches numériques telles que le FBA sont basées sur des contraintes quantitatives regardant l'existence de flux quantitatifs capables de produire les métabolites présents dans la fonction objective tout en satisfaisant les contraintes stœchiométriques. Comme discuté dans le chapitre 2, il n'était certain que le réseau complété à l'aide de *Meneco* permettrait de produire de la biomasse avec une approche basée sur la stœchiométrie des réactions. Le

TABLE 3.1: Nom et quantité des molécules prises en compte dans la fonction de biomasse du réseau EctoGEM

molécule	**quantité**
4-amino-butyrate	0.01
arginine	0.1
asparagine	0.75
citrate	19.3055
cysteine	0.3
glucose	2.7615
glutamine	10.7
glutamate	18.85
glycine	3.8
glycerate	0.123
glycerol	2.375
glycollate	0.029
histidine	0.1
Iso-leucine	0.2
alanine	26.6
aspartate	12.5
ornithine	0.1
leucine	0.3
lysine	0.35
mannitol	331.4
methionine	0.3
phenylalanine	0.35
proline	1.35
serine	3.35
succinate	0.878
threonine	0.65
iso-citrate	9.37
tryptophane	0.115
tyrosine	0.125
valine	0.85

modèle métabolique obtenu ici permet de produire de la biomasse globalement à partir de la composition du milieu de culture de l'algue.

Notre analyse montre que la complétion topologique avec un nombre minimal de réactions, combiné à la curation manuelle, résulte en un réseau capable de produire de la biomasse dans des proportions correctes. Étant donné que *Meneco* peut échouer lors de la reconstruction de certains cycles, cela semble indiquer que peu de cycles incomplets existaient lors de la création de l'ébauche métabolique. Ce processus de reconstruction semble donc particulièrement adapté à la création de réseaux métaboliques chez des espèces non classiques pour l'étude de processus globaux.

3.6 Le workflow de reconstruction du réseau

Le but principal de notre approche était la reconstruction automatique d'un réseau métabolique d'un nouvel organisme modèle distant phylogénétiquement des principaux modèles biologiques eucaryotes étudiés jusqu'à présent, tout en minimisant le nombre de faux positifs ajoutés au niveau des réactions et des gaps, ou du moins en facilitant leur détection et ainsi leur retrait. Dans ce but, nous avons implémenté un pipeline intégratif pour la reconstruction d'un réseau métabolique de ce type. Ce processus global peut être généralisé à de nombreux autres organismes du même type.

3.6.1 Description du pipeline

Nous nous basons tout d'abord sur l'utilisation de Pathway Tools qui va prendre en compte les annotations du génome de l'espèce, en particulier les numéros EC et les termes GO, afin de trouver les réactions correspondantes provenant de MetaCyc.

Nous utilisons ensuite Pathologic [LDNS12] afin de retrouver des informations provenant d'orthologie plutôt que d'utiliser les annotations. Pour cela un réseau métabolique reconstruit pour une espèce proche phylogénétiquement ou métaboliquement parlant, et de très bonne qualité, doit être utilisé. Le réseau AraGEM [dQP+10b] d'*Arabidopsis thaliana* a donc été choisi comme référence pour *Ectocarpus siliculosus* car il s'agit d'un réseau métabolique de bonne qualité pour un eucaryote multicellulaire photosynthétique.

Malgré ces deux sources importantes de données biologiques, les deux approches considérées indépendamment ne permettent pas d'obtenir un réseau métabolique fonctionnel. Afin de surmonter cette difficulté, nous avons développé MeMap et MeMerge pour fusionner des réseaux avec des identifiants issus de Metacyc ou Kegg en un réseau unifié.

Afin de compléter ce réseau, nous avons amélioré un outil de gap-filling existant, *Meneco*, en améliorant son efficacité notamment, comme précisé dans le chapitre précédent. Cette amélioration a été nécessaire afin de pouvoir envisager de travailler à l'échelle du génome, et afin de mieux prendre en compte certaines spécificités des réseaux construits pour de nouveaux modèles biologiques susceptibles de contenir de nouveaux gènes et donc de nouvelles réactions métaboliques (grand nombre de gaps et prise en compte des réactions réversibles). Cet outil utilise des résultats de profilage métabolique publiés et suggère des réactions à ajouter au modèle pour satisfaire à des critères topologiques de productibilité des métabolites identifiés. Dans le cas d'*Ectocarpus siliculosus* , le mannitol, par exemple, n'était pas productible par la combinaison des deux réseaux précédents. Il est devenu productible après l'addition d'une seule réaction manquante.

Malgré cela, la combinaison de toutes ces méthodes ne remplace pas la nécessité d'une curation manuelle, bien qu'elle facilite celle-ci. En effet, la dernière étape consiste en une curation manuelle du réseau basée sur des informations provenant d'étude de séquences avec des profils HMMs et/ou de la bibliographie. Cette étape permet également de rajouter des réactions dans le réseau qui n'auraient pas pu être trouvées par les données génomiques ou de profilage métabolique, comme par exemple deux réactions impliquées dans le recyclage du mannitol ou la voie de synthèse des alginates.

La dernière étape consiste à vérifier la fonctionnalité du système à produire quantitativement de la biomasse. Dans le cas d'*Ectocarpus siliculosus* , le système a été capable de produire cette biomasse. Il reste cependant à améliorer le réseau en introduisant de nouveaux métabolites cibles et caractérisant les échanges de co-facteurs dans le cycle cellulaire.

La figure 3.2 résume l'ensemble du workflow et les résultats obtenus lors de la création d'EctoGEM 1.0, le réseau métabolique d'*Ectocarpus siliculosus*.

3.6.2 Outils intégrés dans le pipeline

Comparaison avec les autres pipelines de reconstruction de réseaux algues Comparé aux autres pipelines de reconstruction de réseaux métaboliques utilisés jusqu'à présent chez des algues, qui sont présentés en partie 1.5, notre approche contient un certain nombre de particularités. Tout d'abord nous ne nous sommes pas contentés d'une source unique de données, comme cela a été le cas la plupart du temps, mais nous avons utilisé à la fois les annotations disponibles via l'interface web Orcae et des recherches d'orthologues en s'appuyant sur la qualité du réseau métabolique AraGEM. Ces deux sources de données nous ayant permis de créer une ébauche métabolique de grande taille, nous avons dû utiliser une méthode automatique globale de complétion à la différence d'AlgaGEM et de DiatomCyc. Pour cela, nous avons développé une méthode permettant d'obtenir l'ensemble des complétions minimales, et non uniquement un échantillonnage comme cela a été le cas pour une partie de la complétion réalisée chez *Ostreococcus*. La validation du réseau obtenue a été réalisée en utilisant la technique de Flux Balance Analysis, comme cela a été le cas dans presque tous les réseaux métaboliques d'algues existants, en créant entièrement une fonction de biomasse à l'aide d'une étude bibliographique. Enfin les réactions prédites comme étant présentes dans le réseau ont été étudiées afin de rechercher des gènes pouvant coder pour les enzymes impliquées dans ces réactions. Cela a été réalisé en créant des profils HMMs pour chaque réaction en se basant sur les séquences des enzymes existantes chez d'autres organismes. D'autres méthodes pourraient être utilisées pour effectuer cette recherche telle que la méthode PRIAM [**?**].

Les étapes principales de ce processus global de reconstruction sont présentées en table 3.2 avec les différentes étapes utilisées pour réaliser les autres réseaux métaboliques chez des algues.

Importance des briques du pipeline Ce pipeline global, nous le voyons, est composé de différentes briques qui nous semblent les plus efficaces et les plus précises aujourd'hui. Il est cependant intéressant de remarquer que l'ensemble de ces briques ne sont pas figées et pourraient aisément être remplacées par d'autres méthodes si celles-ci venaient à dépasser celles utilisées. Par exemple, Pantograph pourrait être remplacé par un autre outil d'inférence de réseaux métaboliques à partir de réseaux existants en se basant sur d'autres scores

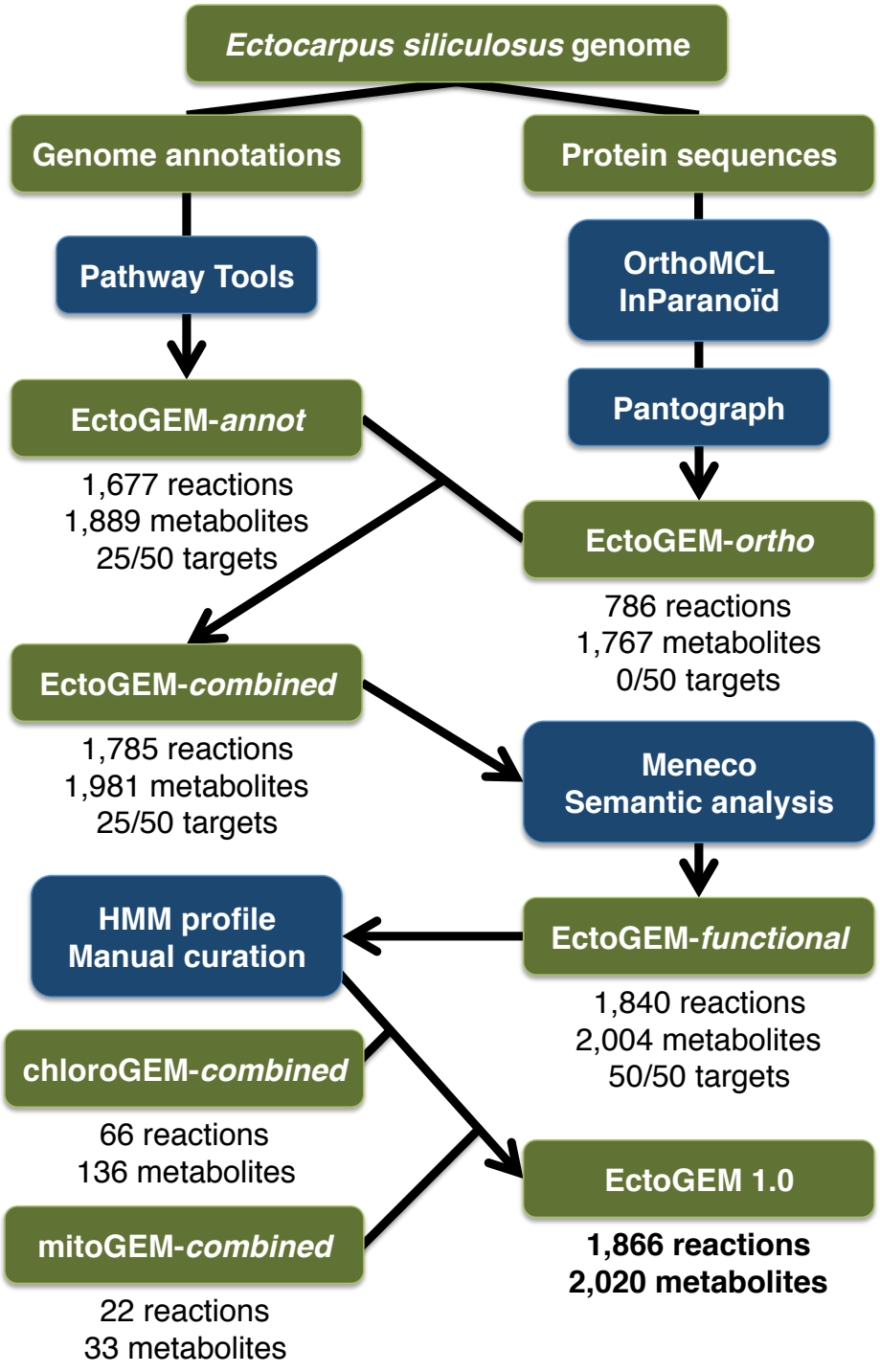


FIGURE 3.2: **Résumé de l'approche intégrative de la reconstruction du réseau métabolique d'*Ectocarpus siliculosus*.** Les boites bleues correspondent aux outils utilisés. Les boites vertes correspondent aux données utilisées et aux différentes versions du réseau existantes.

TABLE 3.2: Comparaison des différentes méthodes de reconstruction de réseaux métaboliques chez des algues avec la méthode proposée pour la reconstruction du réseau d'*Ectocarpus siliculosus*.

Espèce	Nom du réseau	Source de données	Reconstruction initiale	Complétion	FBA ?
Chlamydomonas reinhardtii	AlgaGEM	KEGG	mapping sur KEGG	manuel, composé par composé	Oui
Chlamydomonas reinhardtii	*i*RC1080	réseau pré-existant + bibliographie	Manuel	Effectué mais pas d'informations	Oui
Ostreococcus		KEGG	mapping sur KEGG	Top-down et Bottom-up, pas de solution unique	Oui
Phaeodactylum tricornutum	DiatomCyc	KEGG + Metacyc	Pathway tools	Manuel pour les voies métaboliques connues	Non
Ectocarpus siliculosus	EctoGEM	Orcae + MetaCyc + AraCyc	Pathway tools + Orthologie	Bottom-up automatique, union de toutes les solutions minimales + curation sémantique et manuelle	Oui

qu'uniquement la similarité de séquence. La prédiction des enzymes par des profils HMMs pourrait être remplacée par des outils utilisant des fonctions de score plus précises, en prenant par exemple en compte des domaines protéiques interagissant en trois dimension. *Meneco* pourrait également être remplacé par d'autres outils de complétion si les données nécessaires sont disponibles et que l'efficacité de ces outils est présente.

Ainsi, si la méthode proposée permet bien de reconstruire des réseaux métaboliques, le pipeline en lui même apparaît comme un outil d'aide à l'annotation de réseaux métaboliques plus qu'un outil de reconstruction automatique.

Importance de l'aide à la décision et du caractère évolutif du réseau Ces résultats suggèrent que le pipeline intégratif décrit dans cette thèse est pertinent pour la reconstruction de réseaux métaboliques d'organismes biologiques pour lesquels les données expérimentales sont moins importantes que pour certains modèles biologiques bien établis, et qui sont aussi distants phylogénétiquement par rapport à des organismes tels que E. coli, Arabidopsis, C. elegans et bien d'autres pour lesquels des réseaux métaboliques de bonne qualité sont accessibles. En effet, en combinant des outils de reconstruction automatique et des outils permettant une assistance à la curation manuelle, tout en conservant en permanence des traces de toutes les modifications effectuées, qu'elles soient manuelles ou automatiques, nous obtenons un pipeline global permettant la reproductibilité des manipulations. En se basant sur la même méthodologie, des mises à jour d'EctoGEM pourront être effectuées régulièrement au fur et à mesure que de nouvelles données seront disponibles, telles que de nouvelles annotations de gène, et de nouveaux résultats de profilage transcriptomique, protéomique, et/ou métabolique.

3.7 Conclusion

Ce chapitre nous a permis de décrire le processus global de reconstruction d'un réseau métabolique chez une espèce eucaryote non classique. Si l'étape de complétion d'ébauche métabolique, telle que décrite dans le chapitre précédent, contient un fort problème combinatoire, l'enchaînement des étapes pose, elle, d'autres problèmes. En effet, nous nous sommes retrouvés confrontés a de nombreux problèmes d'unifications de formats entre différents logiciels, d'unifications d'identifiants entre différentes bases de données ou encore de représentations des connaissances. Si tout le processus n'est pas encore entièrement automatisé, cette automatisation est déjà avancée.

Chapitre 4

Contribution du réseau métabolique d'*Ectocarpus siliculosus* pour améliorer les connaissances sur la physiologie des algues brunes

Dans ce chapitre nous présenterons tout d'abord les spécificités du métabolisme d'*Ectocarpus siliculosus* (partie 4.1) avant d'étudier la reconstruction de certaines voies métaboliques importantes chez *Ectocarpus siliculosus* (partie 4.2). Nous continuerons par l'analyse de voies métaboliques mieux connues grâce à l'analyse du réseau métabolique reconstruit dans la partie 4.3, avant de terminer par l'étude de la réannotation de gène permise par la reconstruction du réseau métabolique (partie 4.4). La majorité des apports biologiques présentés dans ce chapitre ont été publiés en 2014 dans Plant Journal [PCD+14]. Ces études ont été réalisées en collaboration avec Simon Dittami et Ludovic Delage de la station biologique de Roscoff.

4.1 *Ectocarpus siliculosus* : ses spécificités

Ectocarpus siliculosus est une algue brune possédant plusieurs spécificités la rendant intéressante à étudier, particulièrement d'un point de vue de biologie systémique. Tout d'abord son génome (et donc son métabolisme) est le résultat d'une endosymbiose secondaire et de nombreux transferts latéraux de gènes [CSR+10, MTS+10b, MTS+10a]. Cela conduit à un métabolisme hybride et à de nouvelles voies métaboliques, certaines des connexions entre les différentes voies n'existant pas chez d'autres organismes.

D'autre part, les algues brunes sont sujettes à de fortes influences de leur environnement sur leurs capacités métaboliques. Elles ont ainsi développé des adaptations métaboliques en rapport avec leur habitat, c'est à dire la zone intertidale. Il est également intéressant de remarquer qu'une souche d'*Ectocarpus* a été découverte vivant dans de l'eau douce en Australie [WK96], entraînant une série d'études passionnantes sur les capacités d'adaptation et d'acclimatation de cette algue aux stress salins.

Comme chez de nombreux organismes, certains exemples d'évolution convergente au niveau d'activités enzymatiques existent, illustrés par le fait qu'une même réaction biochi-

mique est catalysée par des protéines qui n'ont pas de proximité phylogénétique. Il est possible de citer l'exemple de la mannitol-1-phosphatase (M1Pase) [GSM+14]. Ceci n'est pas spécifiques aux algues brunes, mais complexifie l'annotation de génomes. La création de réseaux métaboliques peut aider à la réannotation de génomes lorsque l'on se trouve face à de telles réactions pour lesquelles les enzymes mises en jeu ne peuvent pas être identifiées par similarité de séquence.

De plus il n'est pas possible de réaliser de modification génétique directe ou inverse chez *Ectocarpus siliculosus* . Il est donc difficile d'analyser la fonction physiologique ainsi que l'influence de certains gènes *in vivo* in vivo. La reconstitution de voies métaboliques *in silico* permet d'identifier certains gènes d'intérêt pour guider les futures expérimentations *in vitro* ou *in vivo*.

Enfin *Ectocarpus siliculosus* ne se développe pas seule dans son environnement, mais est associée à de nombreux autres organismes, notamment des bactéries. Pour certains sujets d'étude, il sera donc insuffisant de se concentrer uniquement sur l'algue. Il faudra étudier l'holobionte dans son ensemble et cela sera grandement simplifié par la création ultérieure de méta-réseaux métaboliques impliquant différents organismes.

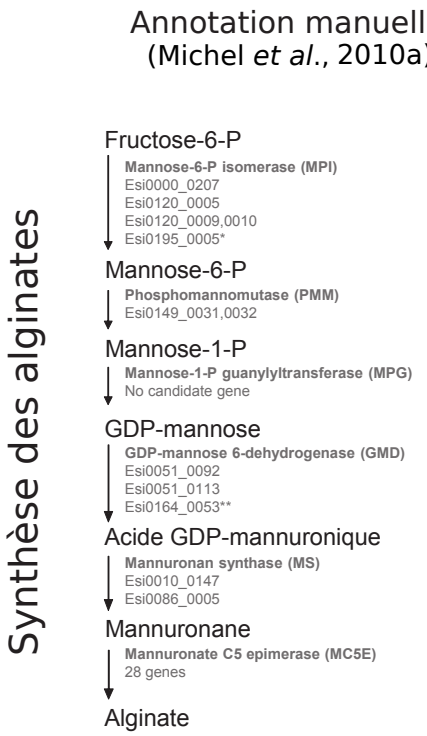
4.2 Analyse de la reconstruction automatique du réseau métabolique

Afin d'avoir une idée générale de la qualité de la reconstruction automatique du réseau métabolique d'*Ectocarpus siliculosus* , nous avons analysé en détail deux voies métaboliques typiques des algues brunes, le cycle du mannitol et la synthèse des alginates. Ces deux voies sont indispensables au métabolisme de l'algue, le mannitol étant utilisé comme forme de stockage du carbone et les alginates étant des composants majeurs de la matrice extracellulaire.

4.2.1 La synthèse des alginates

Une voie potentielle de synthèse des alginates a été proposée par [MTS+10b] et les gènes correspondant ont été annotés dans la base de données Orcae. Dans la voie proposée par [MTS+10b], à partir du fructose-6-phosphate produit par la photosynthèse, la première étape consiste en l'action d'une mannose-6-phosphate isomérase pour produire du mannose-6-phosphate. Celui-ci est ensuite transformé en mannose-1-phosphate par une phosphomannomutase et servira à produire du GDP-mannose par une mannose-1-phosphate guanylyltransférase. Ce GDP-mannose sera ensuite transformé en acide GDP-mannuronique par une GDP-mannose 6-déshydrogénase avant de produire du mannuronane par une mannuronane synthase. Enfin ce mannuronane produira de l'alginate par une mannuronate C5 épimérase. L'ensemble de ces informations est représenté dans la partie gauche de la figure 4.1. Différents gènes candidats existent pour presque toutes ces réactions, seule la transformation du mannose-1-phosphate en GDP-mannose n'est pour l'instant supportée par aucun gène.

Les quatre premières réactions de cette voie de synthèse des alginates ont été bien retrouvées lors de la reconstruction du réseau métabolique. Les gènes proposés sont quasiment équivalents, nous proposons un candidat supplémentaire pour la réaction de transformation du mannose-6-phosphate en mannose-1-phosphate. Une analyse plus poussée



EctoGEM-*functional*
(Cette étude)

Enzyme	Réaction prédite (MetaCyc id)	Gènes assignés
MPI	MANNPISOM-RXN	Esi0000_0207 Esi0120_0005 Esi0120_0009,0010 Esi0195_0002*
PMM	PHOSMANMUT-RXN	Esi0149_0031,0032 Esi0315_0030
MPG	2.7.7.13-RXN	Pas de gène
GMD	GDP-MANNOSE-6-DEHYDROGENASE-RXN	Esi0051_0092 Esi0052_0113 Esi0164_0053**
MS(GT2)	Pas de réaction	Pas de gène
MC5E	Pas de réaction	Pas de gène

FIGURE 4.1: **Comparaison entre la reconstruction automatique et l'annotation manuelle de la voie de synthèse des alginates.** * indique un gène pour lequel l'identifiant a changé dans la base de données en fonction de la mise à jour de l'annotation structurale du génome (Esi0195_0005 est devenu Esi0195_0002). ** indique une protéine dont la fonction a été caractérisée biochimiquement [TVC+11].

de ce gène candidat montre une identité forte (65%) avec des phosphomannomutases présentes chez des plantes [QYQ+07] suggérant qu'*Ectocarpus siliculosus* ne possède pas une, mais deux protéines pour cette fonction. Le gène Esi0051_0113 n'a pas été retrouvé à cause d'un manque d'annotations dans la base de données génomique. Le gène candidat proposé lors de la reconstruction du réseau correspond en effet à une GDP-mannose 6-déshydrogénase, mais celle-ci est connue comme faisant partie d'un virus ayant intégré le génome d'*Ectocarpus siliculosus* [CSR+10].

Les deux dernières réactions de cette voie métabolique n'ont pas été prédites lors de la reconstruction du réseau pour différentes raisons. Tout d'abord les gènes correspondant ne possèdent pas de numéro E.C. dans leur annotation, rendant difficile leur identification et leur assignation lors de l'étape d'étude des annotations pour la reconstruction du réseau. De plus, l'alginate ne faisant pas partie des molécules cibles utilisées lors de la complétion de l'ébauche métabolique par *Meneco*, les réactions enzymatiques associées à cette voie métabolique n'ont pas été considérées pour la complétion du réseau.

4.2.2 Le cycle du mannitol

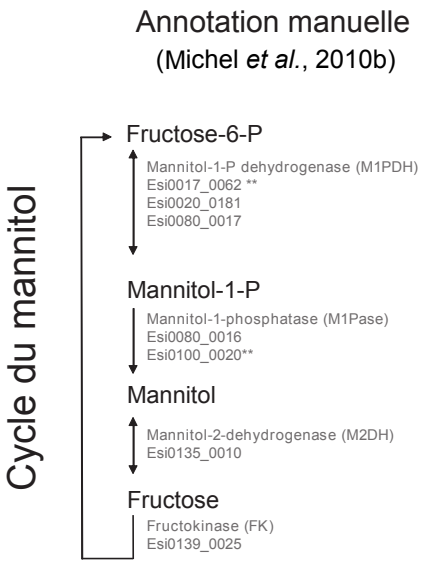
Le cycle du mannitol est composé de quatre réactions différentes, deux pour la synthèse à partir du fructose-6-phosphate, et deux correspondant au recyclage de ce polyalcool pour produire du fructose-6-phosphate. Des gènes candidats ont été identifiés dans le génome d'*Ectocarpus siliculosus* pour ces réactions dans [MTS+10a]. Ces réactions sont présentées dans la partie gauche de la figure 4.2. La première réaction consiste à transformer le fructose-6-phosphate en mannitol-1-phosphate à l'aide d'une mannitol-1-phosphate déshydrogénase. Le mannitol-1-phosphate sert ensuite à la production de mannitol grâce à une mannitol-1-phosphatase. Lors du recyclage du mannitol, nous allons avoir production de fructose par une mannitol-2-déshydrogénase, cet ose étant ensuite reconverti en fructose-6-phosphate par une fructokinase.

La première réaction a été correctement retrouvée et associée aux bons gènes. Le mannitol faisant partie des métabolites cibles, *Meneco* a correctement retrouvé la seconde réaction mais du fait du manque dans les bases de données de séquences protéiques identifiées comme M1Pases, il n'a pas été possible de proposer de gènes candidats pour cette réaction. En effet, une séquence de M1Pase a été identifiée en 1998 chez Eimera [LAF+98] mais sa séquence a trop divergé entre les deux organismes. En revanche aucune des réactions impliquées dans le recyclage du mannitol n'a été retrouvée lors de la création automatique du réseau. Des gènes candidats existants, ils ont été rajoutés manuellement au réseau EctoGEM final.

4.3 Étude de voies métaboliques chez *Ectocarpus siliculosus* à l'aide du réseau métabolique

4.3.1 Synthèse des acides aminés aromatiques

Une étude manuelle d'EctoGEM avant curation montre que, parmi les voies métaboliques de synthèse des acides aminés aromatiques (tryptophane, tyrosine, et phenylalanine) telles que décrites dans [TG10], seule celle de la tyrosine est incomplète. Les trois voies métaboliques de synthèse de la tyrosine présentes dans MetaCyc ("tyrosine biosynthesis I" =



EctoGEM-*functional*
(Cette étude)

Enzyme	Réactions prédite (MetaCyc id)	Gènes assignés
M1PDH	MANNPDEHYDROG-RXN	Esi0017_0062** Esi0020_0181 Esi0080_0017
M1Pase	MANNITOL-1-PHOSPHATASE-RXN	Pas de gène
M2DH	Pas de réaction	Pas de gène
HK	Pas de réaction	Pas de gène

FIGURE 4.2: **Comparaison entre la reconstruction automatique et l'annotation manuelle de la voie du cycle du mannitol.** ** indique une protéine dont la fonction a été caractérisée biochimiquement. [RGD$^+$11, GSM$^+$14]

TYRSYN, "tyrosine biosynthesis II" = PWY-3461 et "tyrosine biosynthesis III" = PWY-6120) montrent la présence d'une seule des trois réactions présentes dans ces voies. La tyrosine étant présente dans les 50 métabolites d'intérêt ayant servi à la reconstruction du réseau, *Meneco* a proposé l'ajout d'une phénylalanine hydroxylase (PAH) afin de permettre la production de tyrosine via une hydroxylation de la phénylalanine(figure 4.4a). Cette suggestion est compatible avec ce qui est observé dans le réseau métabolique DiatomCyc, qui contient la réaction catalysée par une phénylalanine hydroxylase, pour laquelle la protéine associée présente 50% d'identité avec une phenylalanine-4-hydroxylase humaine caractérisée. Cependant, l'analyse du génome d'*Ectocarpus siliculosus* n'a pas permis de retrouver de gène codant pour cette protéine. La voie de production des acides aminés aromatiques a donc été étudiée manuellement afin de mieux la cerner chez cet organisme. Dans MetaCyc, la voie de synthèse I est présente chez *Saccharomyces cerevisiae* et *Escherichia coli*, alors que les voies II et III sont présentes chez les plantes. La principale différence entre ces trois voies réside dans la nature des intermédiaires générés entre le préphénate et la tyrosine (Figure 4.3) :

- La voie microbienne implique le préphénate comme précurseur pour produire du 4-hydroxyphénylpyruvate (HPP) par l'activité d'une déshydrogénase, ce HPP étant ensuite converti en tyrosine par une tyrosine aminotransférase
- La voie présente chez les plantes utilise le préphénate pour produire de l'arogénate par une préphénate aminotransférase (PAT), puis cet intermédiaire est décarboxyler en tyrosine par une arogénate déshydrogénase

La figure 4.3 tirée de [TG10] résume ces différentes informations.

La voie métabolique associée aux plantes est localisée dans le chloroplaste [RPG$^+$09] et les PATs des plantes ont été caractérisées chez *Arabidopsis thaliana* et *Petunia hybrida*

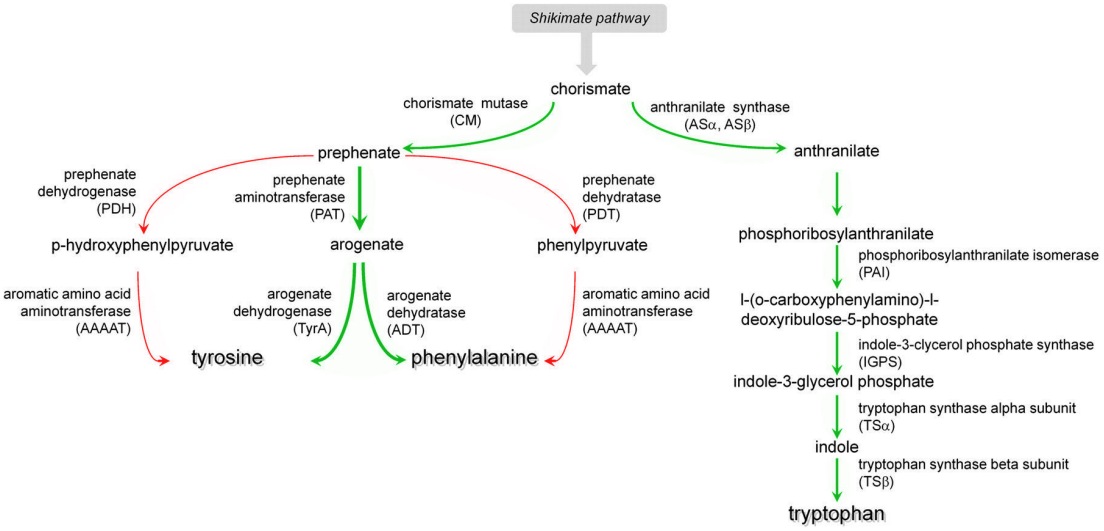


FIGURE 4.3: **Les voies de synthèse des acides aminés aromatiques chez les plantes. [TG10]** Les noms des enzymes sont donnés à côté des flèches, et sont accompagnés des noms abrégés entre parenthèses. Les voies majoritaires de synthèse de la phénylalanine et de la tyrosine chez les plantes sont indiquées par des flèches vertes (voie II pour la tyrosine). La voie de synthèse I de la tyrosine et une voie de synthèse secondaire de la phénylalanine sont représentées par des flèches rouges.

[GGJ+10, MYD11]. Toutefois, des résultats récents semblent indiquer qu'une voie métabolique proche de la voie microbienne pourrait être aussi présente chez les plantes, et que la synthèse de phénylalanine ne serait pas uniquement chloroplastique [YWQ+13]. Trois candidats ont été identifiés dans le génome d'*E. siliculosus* lors de la reconstruction du réseau métabolique pour ces deux voies métaboliques. On note l'absence de candidat pour les PATs, probablement parce que cette protéine n'est pas présente dans le réseau métabolique AraGEM. Cependant une analyse avec la PAT identifiée chez *Arabidopsis thaliana* a permis d'identifier un candidat chez *Ectocarpus siliculosus* (Esi0128_0018). L'ensemble des gènes candidats pour les deux voies est supporté par des EST, et plusieurs d'entre eux ont leur expression qui est réduite lors de stress abiotiques [DSP+09, RDG+14]. La fonction biochimique de ces gènes reste à vérifier expérimentalement. La figure 4.4 (b) représente la voie de synthèse des acides aminés aromatiques proposée après curation manuelle du réseau.

Des analyses phylogénétiques ont été effectuées sur les quatre candidats présents dans le génome d'*Ectocarpus siliculosus* pour étudier l'évolution de cette voie chez les algues brunes. Le logiciel Phylogeny.fr [DGB+08] a été utilisé pour cela. Les alignements protéiques ont été réalisés à l'aide du logiciel MUSCLE avant d'être raffinés avec GBlocks. Une fois les alignements réalisés, les arbres phylogénétiques ont été obtenus par la méthode du maximum de vraisemblance implémentée dans PhyML. Des analyses de bootstrap sur 100 réplicats ont ensuite été réalisées pour estimer la confiance pouvant être accordée à la topologie des arbres obtenus.

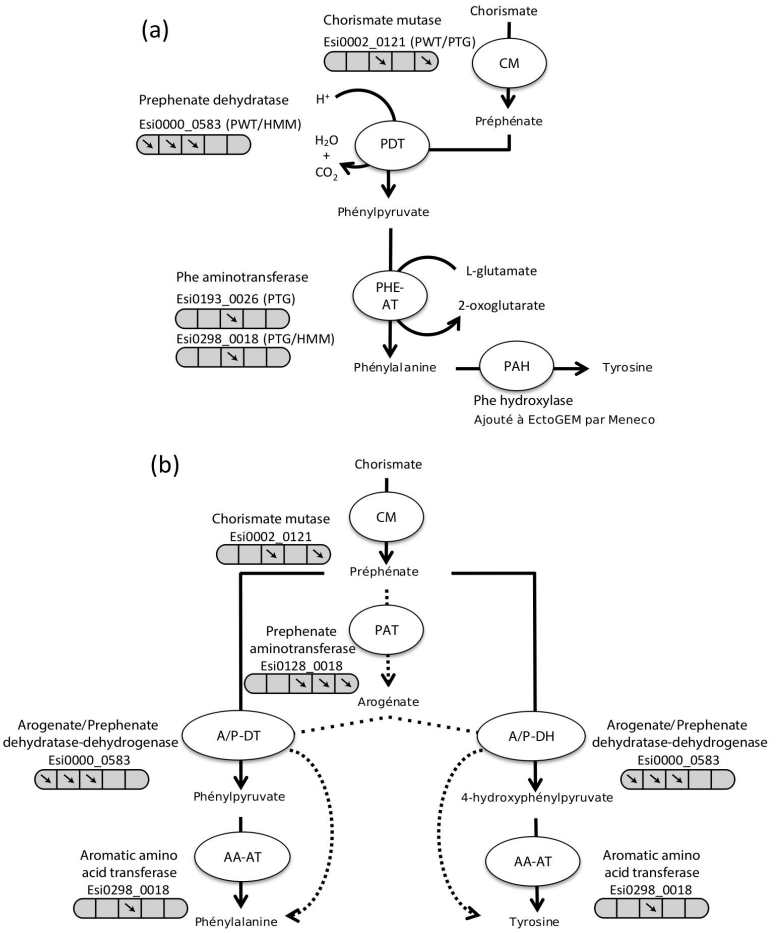


FIGURE 4.4: **Voies de biosynthèse des acides aminés aromatiques avant et après curation manuelle.**(a) Voies métaboliques prédites dans *EctoGEM-functionnal*. Les commentaires associés aux enzymes indiquent comment les gènes ont été identifiés et les réactions prédites : PWT (Pathway Tools), PTG (Pantograph) et HMM. (b) Voies métaboliques obtenues après curation manuelle par analyse comparative avec d'autres espèces. La voie métabolique de l'arogénate est représentée en pointillés. Les changements d'expression significatifs (FDR < 5%) dûs à des stress cuivriques (après 4h et 8h de traitement), hypersalins, hyposalins et oxydatifs (après 6h de traitement pour les trois derniers types de stress) sont indiqués dans cet ordre par les flèches sous les noms des gènes.

La chorismate mutase d'*Ectocarpus* (Esi0002_0121) groupe avec les séquences d'oomycètes et de champignons, alors que les gènes d'algues codant pour les autres enzymes sont plus proches de leurs homologues chez les plantes. Cette observation semble indiquer que la synthèse de phénylalanine et de la tyrosine chez *Ectocarpus siliculosus* se ferait de façon similaire à ce qui se passe chez les plantes terrestres (via l'arogénate) plutôt que directement depuis le préphénate comme chez la levure. Une analyse de la séquence d'Esi0000_0583 montre que cette protéine contient deux domaines, l'un correspondant à une arogénate/préphénate déshydratase (A/P-DT) à l'extrémité N-terminale, et l'autre à une arogénate/préphénate déshydrogénase (A/P-DH) à l'extrémité C-terminale. Cette structure est différente de celles des gènes des plantes terrestres où ces deux fonctions sont catalysées par des enzymes différentes [RPG+09]. Des enzymes potentiellement bi-fonctionnelles A/P-DT et A/P-DH ont été retrouvées chez certains straménopiles comme les oomycètes (*Phytophtora, Albugo*) et chez d'autres organismes plus éloignés comme des haptophytes ou des choanoflagellées (figure 4.5). À l'inverse, chez d'autres straménopiles plus proches phylogénétiquement des algues brunes, comme les diatomées ou les pélagophytes, les deux activités sont codées par des gènes différents, comme c'est le cas aussi chez les algues vertes et rouges. De plus, certains organismes tels que les oomycètes, *M. brevicollis*, et *E. huxleyi* possèdent une enzyme potentiellement tri-fonctionnelle PAT, A/P-DH et A/P-DT. Tous ces résultats semblent indiquer une enzyme ancestrale tri-fonctionnelle étant donné que celle-ci est partagée par deux lignées éloignées, les straménopiles et les opisthokontes. Le gène aurait ainsi été tout d'abord clivé en un gène PAT et un gène bi-fonctionnel A/P-DT et A/P-DH après la divergence entre les oomycètes et les ochrophytes. Les figures 4.6 et 4.7 présentent les analyses phylogénétiques des A/P-DH et A/P-DT respectivement. Ce gène bi-fonctionnel aurait lui même été clivé plus tard chez les diatomées. De plus, le fait que la préphénate aminotransférase soit une partie d'une enzyme trifonctionnelle impliquée dans la synthèse de la phénylalanine et de la tyrosine chez les oomycètes est une indication que le scénario le plus plausible de production de ces acides aminés chez les straménopiles passe par la synthèse d'arogénate. Cependant, la voie de synthèse de la tyrosine connue chez la levure ne peut pas être exclue définitivement.

Enfin, nous avons effectué une recherche d'homologues d'aminotransférases d'acides aminés aromatiques et de phénylalanine hydroxylase (PAH) chez différentes lignées. Le gène codant pour l'enzyme PAH est absent du génome d'*E. siliculosus*, mais a été identifié chez de nombreux autres straménopiles où les deux voies métaboliques pour la production de la tyrosine semblent co-exister, avec ou sans la phénylalanine comme intermédiaire. Cette situation peut également être observée chez les archaeplastides, les opisthokontes, les alvéolés et les cryptophytes. Tout cela semble suggérer une évolution de la production de phénylalanine et de tyrosine indépendante de l'évolution phylogénétique des organismes.

4.3.2 Synthèse du molybdenum

La reconstruction du réseau métabolique d'*E. siliculosus* a également permis d'obtenir de nouveaux indices quant à la synthèse du cofacteur molybdenum (Molybdenum cofactor, Moco). La voie de synthèse du Moco a été correctement reconstruite dans EctoGEM (figure 4.8a) avec trois réactions prédites par Pathway Tools en se basant sur les annotations d'Orcae, et les quatre autres réactions ajoutées par l'algorithme de gap-filling de Pathway Tools. Les quatre réactions ajoutées ont été associées aux gènes correspondant durant la phase

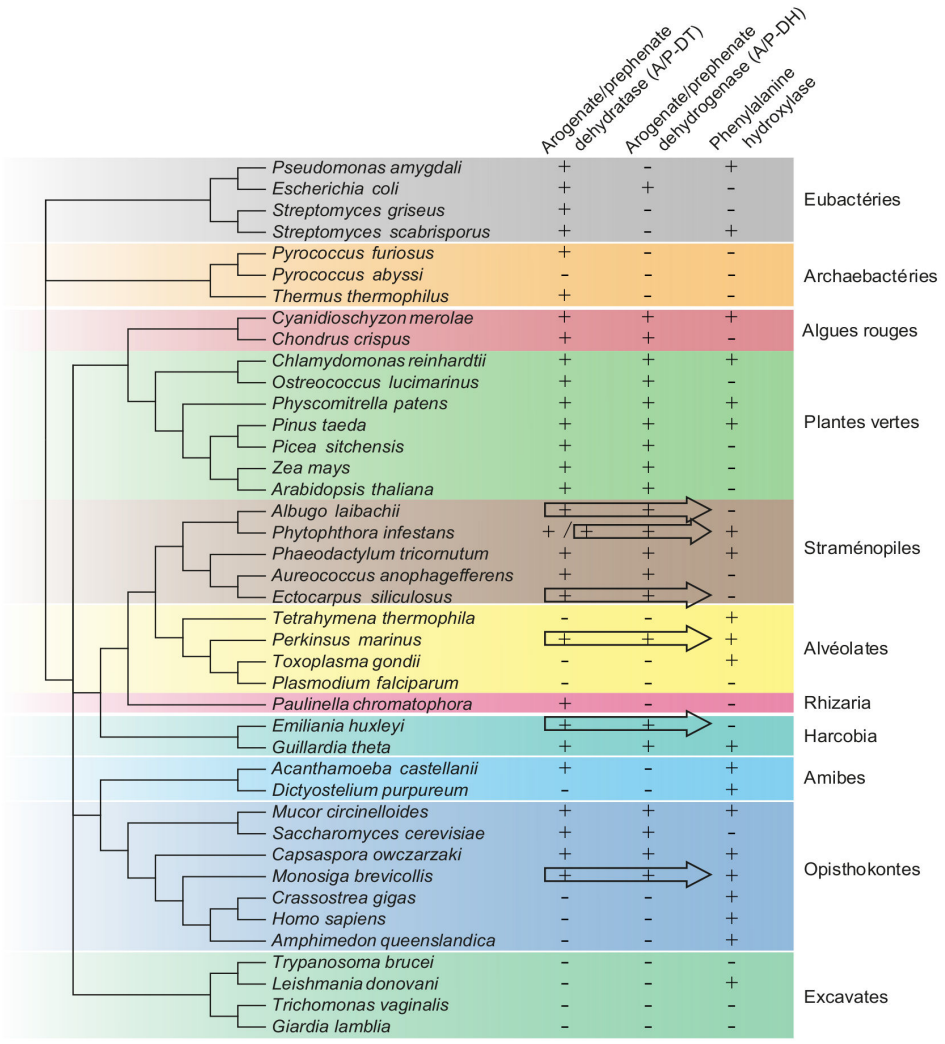


FIGURE 4.5: **Évolution des enzymes impliquées dans la synthèse des acides aminés aromatiques.** L'arbre représente les relations phylogénétiques entre les organismes telles que définies dans Tree Of Life (http ://tolweb.org). La longueur des branches ne représente aucune information. La présence des gènes chez les différents organismes est indiquée par un "+". Les flèches indiquent la fusion des déshydratases (A/P-DT) et déshydrogénases (A/P-DH). *Phytophtora infestans* contient deux déshydratases, l'une étant fusionnée à la déshydrogénase.

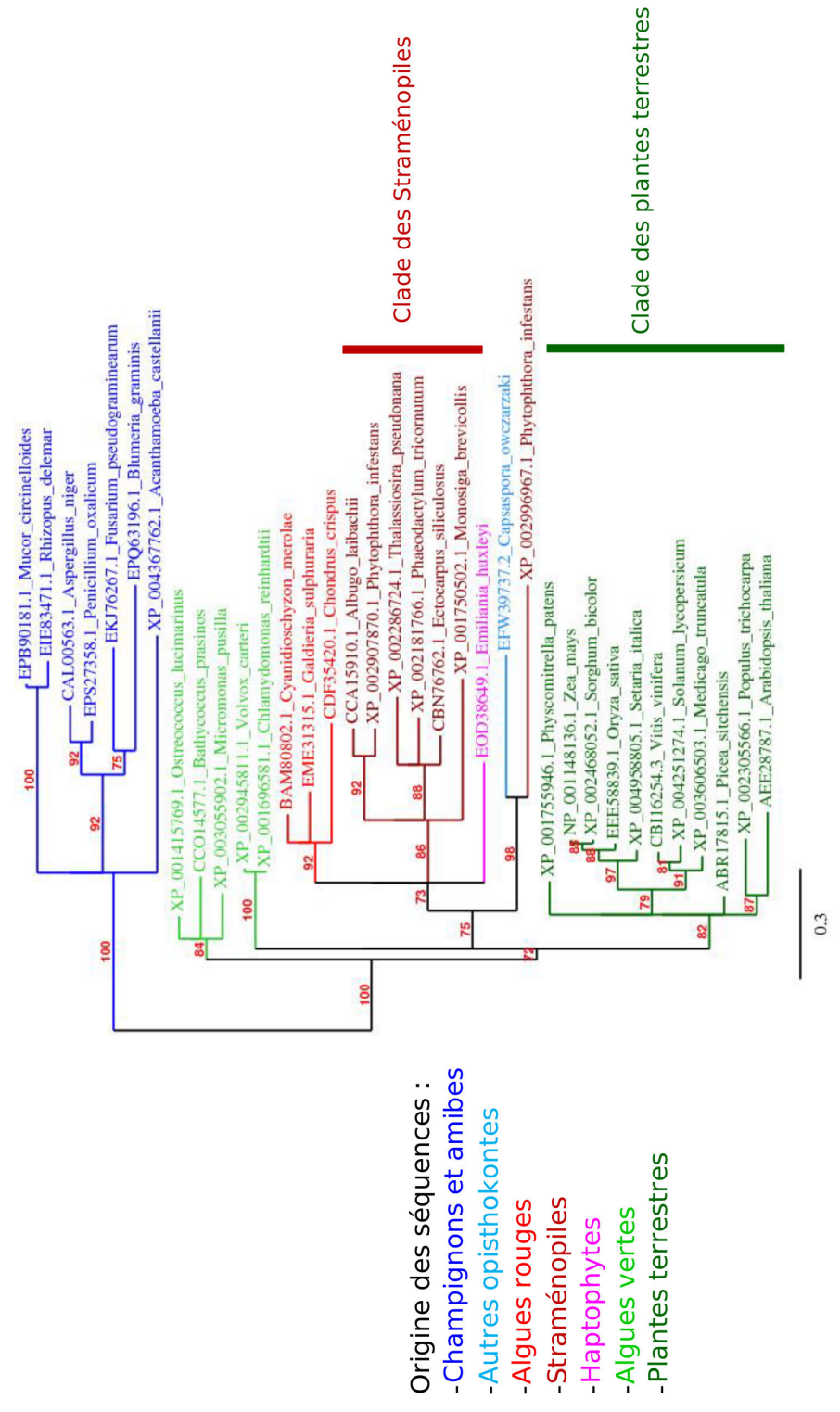


FIGURE 4.6: **Analyse phylogénétique de la A/P-DT d'*Ectocarpus siliculosus* .** Pour chaque séquence, la première partie du nom correspond au numéro d'accession dans la base de données NCBI et la seconde partie correspond au nom de l'espèce où la séquence a été retrouvée. Les couleurs correspondent aux origines phylogénétiques des espèces.

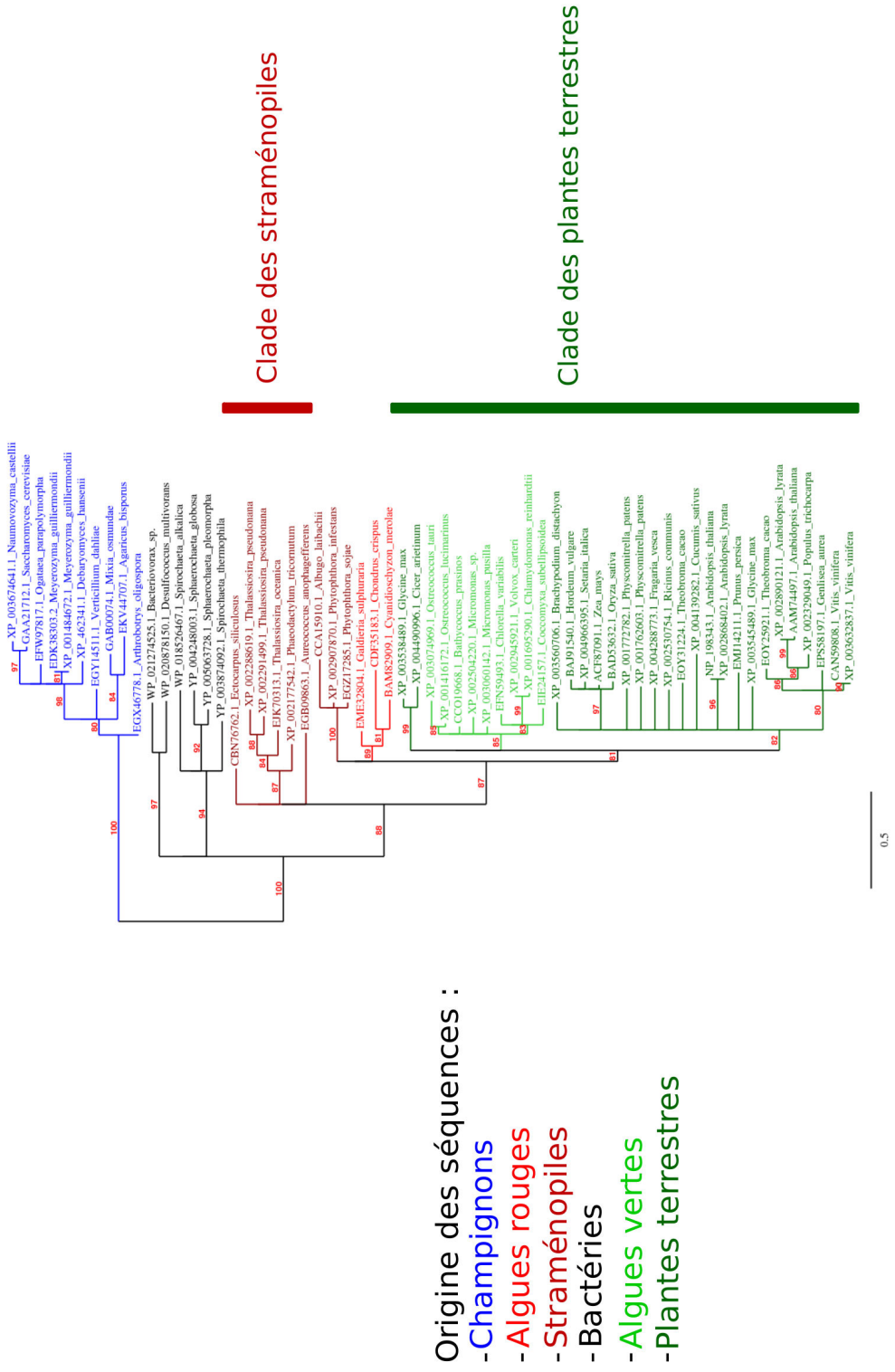


FIGURE 4.7: **Analyse phylogénétique de la A/P-DH d'*Ectocarpus siliculosus* .** Pour chaque séquence, la première partie du nom correspond au numéro d'accession dans la base de données NCBI et la seconde partie correspond au nom de l'espèce où la séquence a été retrouvée. Les couleurs correspondent aux origines phylogénétiques des espèces.

de curation manuelle. Une recherche de gènes associés à cette voie métabolique a mis en avant la présence d'un transporteur de molybdate de type MOT2 [TJGF11] et de six loci génomiques codant pour des protéines qui dépendent probablement du Moco pour leurs activités. Chez les plantes terrestres, le Moco est essentiel au fonctionnement notamment des nitrates réductases et de la xanthine déshydrogénase, impliquées respectivement dans l'assimilation du nitrate, et la dégradation de la purine [SM06]. Chez *E. siliculosus* nous avons retrouvé toutes ces réactions, avec en plus une enzyme de la famille des phosphoadénosine phosphosulfate réductase qui contient aussi un domaine de fixation au Moco.

Tous les gènes candidats impliqués dans la synthèse du molybdenum sont conservés chez les animaux, les plantes et les straménopiles. On a remarqué cependant que les séquences d'*E. siliculosus* sont plus proches de celles des autres straménopiles avec lesquelles elles forment une clade indépendante. La seule exception est pour une des deux molybdopterine synthases potentielles (Esi0290_0003), qui groupe avec les séquences de cyanobactéries lors des analyses phylogénétiques (fig 4.8b). Cela indique que tout en considérant le fait que la biosynthèse du Moco est ancestrale chez les eukaryotes, un deuxième gène codant pour une molybdopterine synthase a été acquis par l'ancêtre commun des straménopiles via un transfert horizontal à partir d'une cyanobactérie. Comme les deux gènes correspondant à cette acitvité enzymatique sont exprimés chez *E. siliculosus* et présents chez les autres straménopiles, il est intéressant de s'interroger quand à leur(s) rôle(s) physiologiques chez ces organismes.

4.4 Réannotation des gènes

Au début de cette thèse, dans la base de données génomique Orcae, plus d'un tiers des gènes ne sont pas annotés précisément ou pas annotés du tout. Un des apports de la reconstruction du réseau métabolique d'*Ectocarpus siliculosus* a été la réannotation d'un certain nombre de ces gènes. En effet, lors des étapes de complétion d'ébauches de réseaux métaboliques, nous ajoutons des réactions dans le réseau qui ne sont pas, la plupart du temps, supportées par une annotation génomique précise (si c'était le cas, normalement, la phase de création de l'ébauche basée sur les annotations aurait dû la trouver). Ainsi, l'étape de complétion pourra attirer l'attention des annotateurs sur une réaction précise pour laquelle les gènes correctement annotés sont manquants. Notre workflow va permettre de proposer des gènes candidats pour ces réactions via l'utilisation des données d'orthologie et les profils HMMs.

Ce manque initial d'annotation peut être dû à différentes raisons. Tout d'abord comme précisé précédemment, le génome d'*Ectocarpus siliculosus* est la résultante d'une endosymbiose secondaire et de nombreux transferts latéraux de gènes. Cela conduit à la présence de voies métaboliques "non classiques" chez cette algue. D'autre part, les algues brunes font partie d'un groupe phylogénétique ayant divergé très rapidement parmi les eucaryotes. Les annotations de génome automatiques basées sur des techniques d'homologie de séquence risquent donc d'échouer du fait de la grande distance phylogénétique entre *Ectocarpus siliculosus* et la plupart des organismes séquencés et annotés jusqu'à présent. De plus, le métabolisme chimérique très particulier à *Ectocarpus siliculosus* engendre le fait que cette espèce possède des voies métaboliques (et donc des gènes) spécifiques à cette espèce. Enfin les expériences de transcriptomique chez *Ectocarpus siliculosus* sont peu nombreuses. Par conséquent, les possibilités de réaliser des méta-analyses afin d'identifier

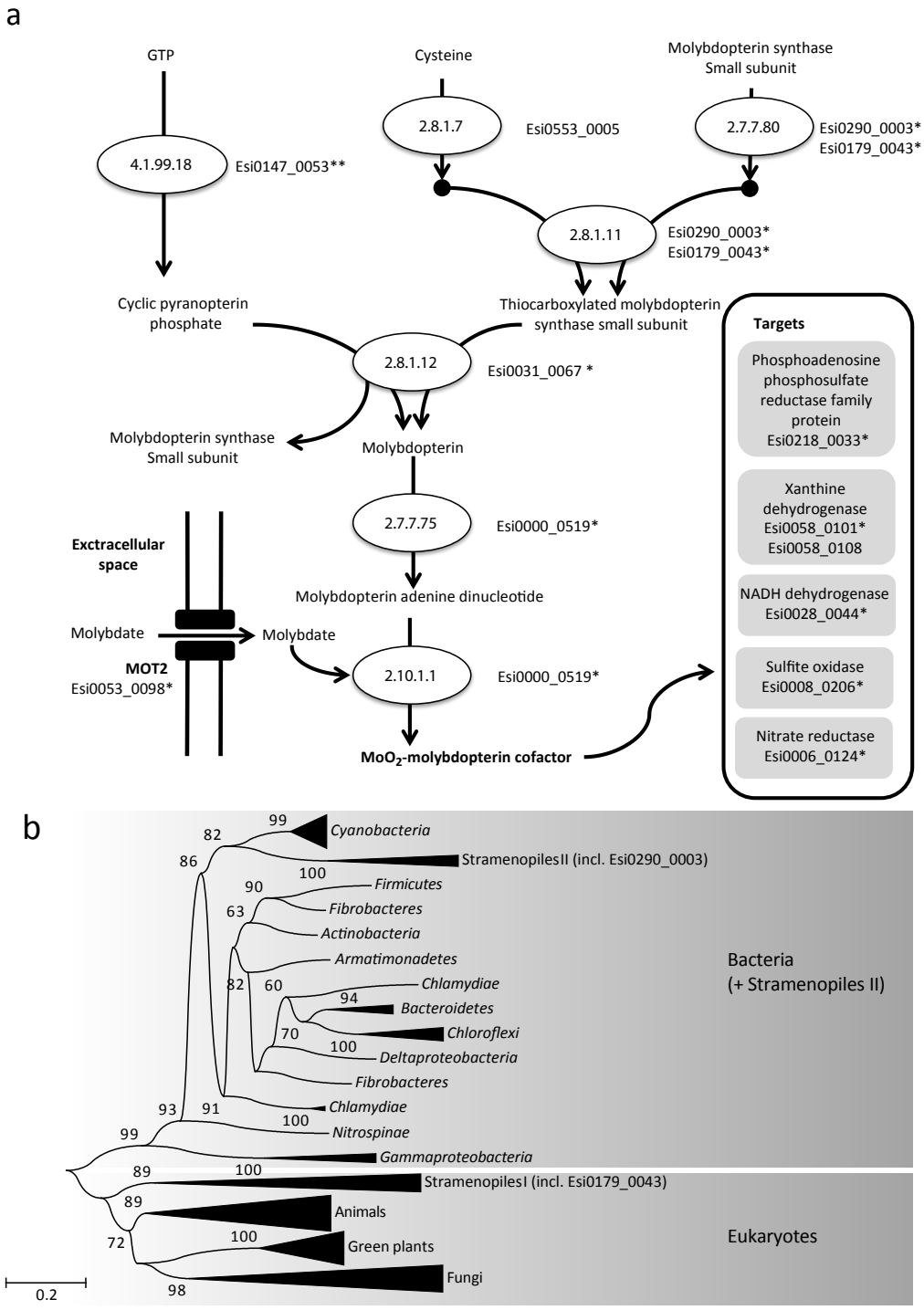


FIGURE 4.8: **Voies de biosynthèse du cofacteur molybdenum.** a) Voies métaboliques curées et protéines associées aux réactions enzymatiques. * marque les gènes exprimés chez *E. siliculosus*. ** marque les gènes significativement (FDR < 5%) réprimés en réponse à un stress hyposalin. Les nombres dans les ovales correspondent aux numéros E.C. des réactions. (b) Position phylogénétique de deux candidats pour l'activité molybdopterin synthase. La figure montre un arbre de maximum de vraisemblance avec 75 séquences sélectionnées dans différentes lignées. Seules les valeurs possédant un bootstrap supérieur à 50 sont indiquées.

des gènes co-régulés susceptibles d'intervenir dans des voies métaboliques communes sont limitées. De la même manière, l'absence de structure en opéron ou en cluster dans le génome (que l'on retrouve chez les bactéries ou dans le métabolisme secondaire des plantes [WGH+12]) empêche de réaliser des prédictions d'annotations en se basant sur la localisation des gènes.

La création du réseau métabolique d'*Ectocarpus siliculosus* a permis de réannoter manuellement, pour le moment, 56 gènes. Aucune autre voie métabolique en particulier n'a été retrouvée associée à ces gènes, ce qui amène à penser que ces réannnotations ont été réalisées pour des gènes correspondant à des protéines impliquées dans différentes voies métaboliques, et des loci dispersés dans le génome. La liste des gènes réannotés est présentée en annexe C.

Toutefois, nous avons remarqué qu'une majorité des gènes réannotés se trouvent dans des voies métaboliques connues comme étant des voies de synthèse de molécules. Cela peut être expliqué par l'étape de complétion du réseau métabolique. En effet, *Meneco* rajoute dans ce réseau des réactions afin de produire certaines molécules. Il est tout à fait possible que ces réactions n'aient pas été ajoutées au cours de la construction de l'ébauche du réseau car elles n'ont pas été retrouvées dans le génome, probablement parce que le(s) gène(s) associé(s) à ces réactions étaient pas, peu, ou mal annotées. Nous retrouvons également quelques réactions catalysant la production de métabolites intervenant dans plusieurs voies métaboliques. De même manière, l'ajout de celles-ci par *Meneco* indique que les gènes étaient mal annotés dans le génome, et leur intégration dans le réseau est sans doute nécessaire pour rendre disponibles un grand nombre de réactions.

4.5 Conclusion

Nous l'avons vu, EctoGEM a d'ores et déjà été utilisé pour compléter certaines annotations au sein du génome d'*Ectocarpus siliculosus* dans la base de données ORCAE, notamment en lien avec certaines voies métaboliques qui ont déjà été étudiées ou sont en cours de considération par des équipes travaillant sur la physiologie d'*Ectocarpus*. D'autre part, le processus de reconstruction du réseau métabolique a permis aux experts de se concentrer sur certaines voies métaboliques particulières apportant ainsi de nouvelles connaissances sur le métabolisme de cette algue, et sur l'évolution des voies métaboliques chez les eucaryotes.

Par la suite, EctoGEM pourra être utilisé pour analyser des changements physiologiques de l'algue brune en réponse à certaines modifications des conditions environnementales. Ce réseau sera également considéré pour intégrer et cartographier des résultats d'expériences de transcriptomique et de profilage métabolique obtenus dans différentes conditions de stress. Ceci devrait permettre d'améliorer la compréhension de l'influence de ces stress sur la physiologie globale de l'organisme par une vision globale, c'est à dire à l'échelle du génome complet.

Conclusion et perspectives

Conclusion

Cette thèse se situe dans le domaine de la bioinformatique. La question abordée est la *reconstruction de réseaux métaboliques* pour des espèces eucaryotes non classiques, en se concentrant particulièrement sur une étape primordiale de cette reconstruction, la complétion d'ébauche métaboliques L'ensemble des résultats obtenus au cours de ce travail ont été appliquées à la reconstruction d'une carte métabolique chez *Ectocarpus siliculosus*, le modèle génomique des algues brunes.

Le problème de complétion d'un réseau métabolique est le suivant. Nous disposons d'information partielles sur les réactions métaboliques qui transforment des molécules chez une espèce données. De plus, nous avons à notre disposition des données qui nous indiquent la présence de certaines molécules, produites par l'organisme d'intérêt (profilage métabolique). Étant donné que ces molécules sont présentes au sein des cellules de l'organisme, il semble logique de supposer que le métabolisme de cet organisme est capable de les produire à partir des nutriments qui composent son milieu de culture. Pour vérifier cela, nous parcourons le graphe créé par les réactions métaboliques reliées entre elles par les molécules impliquées. Ce parcours va se faire à la recherche d'un chemin entre au moins une molécule présente dans le milieu de culture et l'ensemble des molécules identifiées par profilage métabolique. Si ce chemin existe, les informations contenues dans le génome ont suffit à expliquer la présence de ces molécules. Si ce n'est pas le cas, nous allons ajouter des réactions provenant de bases de données de réactions métaboliques dans le réseau afin de rétablir la connectivité. Cet ajout de réaction se fera de manière à modifier le moins possible le réseau existant, en ajoutant le nombre minimal de réactions. Le problème qui doit être résolu consiste donc à proposer un ensemble de réactions métaboliques piochées dans une banque "universelle" de réactions, pour restaurer "in silico" la production de composés cibles.

Au début de cette thèse une méthode de résolution de la version combinatoire de ce problème avait été proposée sous la forme d'une approche de programmation par contraintes déclaratives appelée programmation par ensemble réponses (ou Answer Set Programming, ASP) [ST09]. mais celle-ci n'était pas applicable en routine car elle ne passait pas à l'échelle sur de grandes bases de données. D'autre part cette méthode ne représentait pas fidèlement la réalité biologique de certaines réactions, les réactions réversibles. La première partie de cette thèse a donc consisté en l'amélioration de cette méthode de complétion de réseaux métaboliques sur deux plans : la modélisation des réactions réversibles et les méthodes de recherches de solutions. Ces modifications ont été introduite dans le développement du package python *Meneco*, rapide et aisément utilisable à partir de fichiers dans des formats

sbml classiques. L'intérêt de cette méthode, tant d'un point de vue de l'efficacité des calculs que de la fonctionnalité biologique des réseaux reconstruits, a été validée par des expérimentations sur des réseaux classiques (*E. Coli*) et non-classique (*Ectocarpus siliculosus*).

Une fois la méthode de complétion validée, celle-ci a été utilisée pour reconstruire le réseau métabolique du modèle biologique des algues brunes, *Ectocarpus siliculosus*. Une reconstruction d'un réseau métabolique fonctionnel ne se limitant pas à la seule étape de complétion, nous avons développé un processus global de reconstruction de réseau métabolique. Celui-ci se base sur les données classiquement utilisées lors de la reconstruction de réseaux métaboliques : le génome de l'espèce étudiée (avec les annotations associées) et toute autre source de données biologiques disponibles (métabolomique, transcriptomique, etc).

Pour cela deux ébauches métaboliques différentes seront créées. La première utilisera les informations présentes dans les annotations du génome, que celles-ci soient manuelles ou automatiques. Ces informations seront recoupées avec celles présentes dans les bases de données de réactions métaboliques telles que MetaCyc afin de créer un lien entre un gène et une réaction. Cette étape est réalisée à l'aide du logiciel de référence Pathway Tools. D'autre part, nous utilisons les informations incluses dans les séquences génomiques pour réaliser une seconde ébauche métabolique. Pour cela nous utilisons un réseau métabolique existant chez une espèce proche phylogénétiquement de l'espèce étudiée. Une recherche d'orthologues est alors réalisée entre le génome de l'espèce pour laquelle nous reconstruisons le réseau et celle possédant déjà un réseau métabolique. Le but ici est de transposer les associations entre gènes et réactions ayant déjà été réalisées vers l'espèce d'intérêt. Si dans le premier cas nous cherchons à obtenir le maximum d'informations possible à partir des annotations manuelles du génome, dans le second nous cherchons à profiter de la curation manuelle préexistante dans certains réseaux. Une difficulté ici est de fusionner les deux ébauches métaboliques, en particulier lorsque les sources de données sont différentes, ce qui nécessite de réconcilier les annotations avec des méthodes dédiées.

Nous obtenons alors une ébauche métabolique contenant le maximum d'informations contenues dans le génome. Pour réaliser la complétion de ce réseau métabolique nous utilisons alors une autre source de données (un profilage métabolique), et utilisons l'approche *Meneco* de complétion combinatoire discutée plus haut. Les solutions au problème combinatoire étant nombreuses (mais énumérables), elles sont analysées a posteriori avec des approches sémantiques (et fonctionnelles (validation via des profils HMM pour les réactions introduites). L'ensemble du réseau reconstruit a finalement été validé par une étude fonctionnelle quantitative (FBA) qui a montré sa capacité à produire de la biomasse.

Surtout, les méthodes introduites dans le pipeline complet (alignement de séquences, harmonisation des annotations, complétion combinatoire de réseau, analyse sémantique), permettent de réaliser efficacement une reconstruction de reconstruction de réseau métabolique en tirant partie de la plupart des informations disponibles sur le réseau : annotation du génome, connaissance sur des espèces cousines (réseau métabolique de référence), profils métabolique, et modèles HMM associé à toutes les protéines introduites dans la banque de référence metacyc. Ce pipeline est en cours d'application à d'autres modèles marins ou extrémophiles.

Une fois la complétion réalisée, le réseau a été analysé d'un point de vue topologique pour obtenir de nouvelles connaissances sur le métabolisme d'*Ectocarpus siliculosus*. Cette

étude a permis de faire de nouvelles hypothèses sur la biosynthèse des acides aminés aromatiques ou du molybdenum. Ainsi, une nouvelle voie de synthèse de la tyrosine aurait été identifiée de même qu'une origine probablement bactérienne d'un des gènes présent dans la voie de synthèse du molybdenum. D'autre par la création du réseau métabolique a permis de guider la réannotation de gènes chez *Ectocarpus siliculosus*. En effet, lorsque nous ajoutons des réactions au réseau pendant l'étape de création de l'ébauche à l'aide d'orthologues ou lors de la complétion, cela indique que ces réactions n'ont pas été ajoutées grâce aux annotations. Il y a donc ici une amélioration possible des annotations du génome. Si le transfert d'information est relativement direct lorsqu'il s'agit de l'utilisation des données d'orthologie, cela est beaucoup moins évident à partir de la complétion. Nous proposons alors une liste de réactions à ajouter au réseau mais ces réactions sont totalement détachées de toute preuve de présence de l'enzyme au niveau génétique. Pour palier à cela, lorsque nous proposons la présence d'une réaction dans le réseau, nous associons cette proposition de réaction à un ou plusieurs gènes associés en utilisant la création de profils HMMs à partir des gènes existants chez d'autres espèces.

Perspectives

Différentes perspectives ressortent de ce travail à plus ou moins long terme.

Automatisation du pipeline Tout d'abord, si les différentes briques du processus de reconstruction de réseaux métaboliques sont grandement automatiques, il reste à automatiser l'ensemble du processus pour fournir aux utilisateurs finaux un processus global et automatique de reconstruction nécessitant le moins possible d'étapes manuelles. Cela permettra d'obtenir un outil global utilisable par tous pour la reconstruction de réseaux métaboliques eucaryotes chez des espèces non classiques. Ce travail est d'ors et déjà en cours.

Fonctionnalité La définition de la productibilité par *Meneco*, nous l'avons vu, amène à une sous-approximation de la productibilité des molécules comparé aux méthodes stœchiométriques au niveau des cycles. Une voie d'amélioration de ce point serait d'améliorer la sémantique de productibilité en prenant en compte une partie des cycles utilisés par la sémantique proposée par Acuna Sagot. Une autre grosse piste de recherche consisterait à utiliser les développements en cours de solveurs hybrides ASP+ILP. Cela permettrait de combiner l'efficacité des solveurs ASP avec une définition quantitative de la productibilité. Il faudrait alors combiner des critères quantitatifs et qualitatifs. Mais pour le moment les méthodes de recherche combinatoire ont du mal à "apprendre" des contraintes à partir des solveurs ILP existants.

Amélioration et étude fonctionnelle du réseau métabolique D'autre part, maintenant que le réseau est reconstruit, il devrait pouvoir être utilisé pour réaliser des prédictions sur le métabolisme d'*Ectocarpus siliculosus* en réalisant une étude fonctionnelle précise du réseau. Nous espérons que cette étude aide à comprendre la forte capacité d'adaptation et d'acclimatation de cette algue brune aux stress abiotiques tel que le stress salin. Pour

cela nous comptons utiliser les quelques données de transcriptomique existantes. Ces expériences ont été réalisées en conditions de stress salin ou cuivrique, il sera donc intéressant de voir quelles gènes présents dans le réseau métabolique sont sur-exprimées ou sous-exprimés lors de ce stress. De nouvelles voies de réponses aux stress pourraient alors être identifiées. D'autre part l'espèce d'*Ectocarpus* vivant en eau douce va être très prochainement séquencée. Reconstruire le réseau métabolique de cette espèce par la même méthode et comparer les deux réseaux obtenus pourrait permettre d'obtenir des indications sur la raison de cette adaptation à l'eau douce (apparition de certaines voies métaboliques ou disparition d'autres). Là encore l'étude de la fonctionnalité du réseau a commencé avec sa décomposition en modes élémentaires et en modules réalisée par Annika Röhl et Arne Reimers, respectivement, lors d'une visite au laboratoire "Mathematics in Life Sciences" de Berlin.

De plus de nouvelles données de métabolomique à haut débit devraient être disponibles sous peu. L'identification d'un grand nombre de métabolites chez *Ectocarpus siliculosus* devrait ainsi permettre d'améliorer encore l'étape de complétion du réseau, la technique de complétion utilisée étant d'autant plus précise que le nombre de métabolites dont nous devons expliquer la présence est grand.

Enfin des données de profilage C13 devraient arriver sous peu chez des algues. Il serait alors intéressant de travailler au niveau moléculaire, par exemple en rajoutant des contraintes de traçage des carbones sur les voies métaboliques.

Étude de communautés d'espèces À plus long terme, nous pouvons nous interroger sur la symbiose existant entre *Ectocarpus siliculosus* et de nombreuses souches bactériennes vivant à sa surface. En effet, il a d'ores et déjà été montré que ces bactéries jouent un rôle non négligeable pour le métabolisme de l'algue brune. Reconstruire le réseau "méta-métabolique" de l'algue et de l'ensemble des bactéries vivant en symbiose avec elle. Il est en effet possible que ces bactéries fournissent à l'algue des nutriments et vice-versa. Cet échange pourrait être particulièrement intéressant par exemple lorsque la marée se retire et que l'algue subit de très forts stress au niveau de la zone intertidale.

Pour ce faire, les zones d'échange entre les différents organismes seront particulièrement importantes à modéliser. En effet, ces zones d'échanges et les transporteurs impliqués impliquent des contraintes à la fois qualitatives et quantitatives qui caractérisent les échanges au sein d'une communauté. Identifier automatiquement ces contraintes est à la fois un défi informatique et une donnée très importante en biologie. En effet, il est souvent impossible ou très difficile d'isoler une bactérie précise ou un protiste vivant en communauté pour le faire croître en culture. Une fois que les échanges de molécules entre organismes élucidés, il est possible d'espérer pouvoir créer des milieux de cultures spécifiques en mimant les conditions de vie au sein de la communauté.

Importance des réseaux métaboliques chez de nouveaux modèles Cette thèse a permis de reconstruire un réseau métabolique chez un nouveau modèle, le modèle des algues brunes, *Ectocarpus siliculosus*. Dans un récent article, Fabris et ses collaborateurs [FMC+14] ont étudié le réseau DiatomCyc qu'ils ont reconstruit récemment. Leur étude montre l'intérêt des reconstruire des réseaux métaboliques chez des espèces modèles, notamment chez les straménopiles. En effet, ces réseaux métaboliques pourraient permettre d'élucider certaines

parties de leur métabolisme hybride entre ce que l'on trouve chez les animaux, les plantes et les champignons, notamment.

Bibliographie

[AKMS12] Benjamin Andres, Benjamin Kaufmann, Oliver Matheis, and Torsten Schaub. Unsatisfiability-based optimization in clasp. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 17. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[AKRI13] Shilpi Aggarwal, I. A. Karimi, and Gregorius Reinaldi Ivan. In silico modeling and evaluation of gordonia alkanivorans for biodesulfurization. *Mol. BioSyst.*, 9 :2530–2540, 2013.

[Alo07] Uri Alon. Network motifs : theory and experimental approaches. *Nat. Rev. Genet.*, 8(6) :450–461, 2007.

[ALS+13] Rasmus Agren, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, and Jens Nielsen. The raven toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput Biol*, 9(3) :e1002980, 03 2013.

[AMC+12] Vicente Acuña, Paulo Vieira Milreu, Ludovic Cottret, Alberto Marchetti-Spaccamela, Leen Stougie, and Marie-France Sagot. Algorithms and complexity of enumerating minimal precursor sets in genome-wide metabolic networks. *Bioinformatics*, 28(19) :2474–2483, 2012.

[Bar03] Chitta Baral. *Knowledge Representation, Reasoning, and Declarative Problem Solving*. Cambridge University Press, New York, NY, USA, 2003.

[BBM+14] Thomas Bernard, Alan Bridge, Anne Morgat, Sébastien Moretti, Ioannis Xenarios, and Marco Pagni. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Briefings in Bioinformatics*, 15(1) :123–135, 2014.

[BES+01] V. Badarinarayana, P. W. Estep, J. Shendure, J. Edwards, S. Tavazoie, F. Lam, and G. M. Church. Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol.*, 19(11) :1060–1065, November 2001.

[BFM+07] Scott A Becker, Adam M Feist, Monica L Mo, Gregory Hannum, Bernhard O Palsson, and Markus J Herrgard. Quantitative prediction of cellular metabolism with constraint-based models : the COBRA toolbox. *Nat. Protocols*, 2(3) :727–738, March 2007.

[BNLC13] B. Billoud, Z. Nehr, A. Le Bail, and B. Charrier. Computational prediction and experimental validation of microRNAs in the brown alga Ectocarpus siliculosus. *Nucleic Acids Research*, pages gkt856–, September 2013.

[BPM03] Anthony P. Burgard, Priti Pharkya, and Costas D. Maranas. Optknock : A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6) :647–657, 2003.

[BS05] S. A. Benner and A. M. Sismour. Synthetic biology. *Nat. Rev. Genet.*, 6(7) :533–543, Jul 2005.

[BVM01] Anthony P. Burgard, Shankar Vaidyaraman, and Costas D. Maranas. Minimal reaction sets for escherichia coli metabolism under different growth requirements and uptake environments. *Biotechnology Progress*, 17(5) :791–797, 2001.

[CAB+14] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, Ingrid M. Keseler, Anamika Kothari, Aya Kubo, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, Deepika Weerasinghe, Peifen Zhang, and Peter D. Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/-genome databases. *Nucleic Acids Research*, 42(D1) :D459–D471, 2014.

[CB79] A. Cornish-Bowden. *Fundamentals of enzyme kinetics*. Butterworths, 1979.

[CEG+13] Guillaume Collet, Damien Eveillard, Martin Gebser, Sylvain Prigent, Torsten Schaub, Anne Siegel, and Sven Thiele. Extending the metabolic network of *Ectocarpus Siliculosus* using answer set programming. In Pedro Cabalar and TranCao Son, editors, *Logic Programming and Nonmonotonic Reasoning*, volume 8148 of *Lecture Notes in Computer Science*, pages 245–256. Springer Berlin Heidelberg, 2013.

[CGA+11] S. M. Coelho, O. Godfroy, A. Arun, G. Le Corguille, A. F. Peters, and J. M. Cock. OUROBOROS is a master regulator of the gametophyte to sporophyte life cycle transition in the brown alga ectocarpus. *Proceedings of the National Academy of Sciences*, 108(28) :11518–11523, June 2011.

[CGM+11] Roger L Chang, Lila Ghamsari, Ani Manichaikul, Erik F Y Hom, Santhanam Balaji, Weiqi Fu, Yun Shen, Tong Hao, Bernhard ØPalsson, Kourosh Salehi-Ashtiani, and Jason A Papin. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Molecular Systems Biology*, 7 :518, January 2011.

[CKS01] Nadia Creignou, Sanjeev Khanna, and Madhu Sudan. *Complexity classifications of boolean constraint satisfaction problems*. SIAM, 2001.

[CMA+08] Ludovic Cottret, Paulo Vieira Milreu, Vicente Acuña, Alberto Marchetti-Spaccamela, Fábio Viduani Martinez, Marie-France Sagot, and Leen Stougie. Enumerating precursor sets of target metabolites in a metabolic network. In *Algorithms in Bioinformatics*, pages 233–244. Springer, 2008.

[CMK+09] Nils Christian, Patrick May, Stefan Kempa, Thomas Handorf, and Oliver Ebenhöh. An integrative approach towards completing genome-scale metabolic networks. *Molecular bioSystems*, 5(12) :1889–903, December 2009.

[CPC11] J. Mark Cock, Akira F. Peters, and Susana M. Coelho. Brown algae. *Current Biology*, 21(15) :R573 – R575, 2011.

[CSH+12] Leonid Chindelevitch, Sarah Stanley, Deborah Hung, Aviv Regev, and Bonnie Berger. Metamerge : scaling up genome-scale metabolic reconstructions with application to mycobacterium tuberculosis. *Genome Biology*, 13(1) :r6, 2012.

[CSR+10] J. Mark Cock, Lieven Sterck, Pierre Rouzé, Delphine Scornet, Andrew E. Allen, Grigoris Amoutzias, Veronique Anthouard, François Artiguenave, Jean-Marc Aury, Jonathan H. Badger, Bank Beszteri, Kenny Billiau, Eric Bonnet, John H. Bothwell, Chris Bowler, Catherine Boyen, Colin Brownlee, Carl J. Carrano, Bénédicte Charrier, Ga Youn Cho, Susana M. Coelho, Jonas Collén, Erwan Corre, Corinne Da Silva, Ludovic Delage, Nicolas Delaroque, Simon M. Dittami, Sylvie Doulbeau, Marek Elias, Garry Farnham, Claire M. M. Gachon, Bernhard Gschloessl, Svenja Heesch, Kamel Jabbari, Claire Jubin, Hiroshi Kawai, Kei Kimura, Bernard Kloareg, Frithjof C. Küpper, Daniel Lang, Aude Le Bail, Catherine Leblanc, Patrice Lerouge, Martin Lohr, Pascal J. Lopez, Cindy Martens, Florian Maumus, Gurvan Michel, Diego Miranda-Saavedra, Julia Morales, Hervé Moreau, Taizo Motomura, Chikako Nagasato, Carolyn A. Napoli, David R. Nelson, Pi Nyvall-Collén, Akira F. Peters, Cyril Pommier, Philippe Potin, Julie Poulain, Hadi Quesneville, Betsy Read, Stefan A. Rensing, Andrés Ritter, Sylvie Rousvoal, Manoj Samanta, Gaelle Samson, Declan C. Schroeder, Béatrice Ségurens, Martina Strittmatter, Thierry Tonon, James W. Tregear, Klaus Valentin, Peter von Dassow, Takahiro Yamagishi, Yves Van de Peer, and Patrick Wincker. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature*, 465(7298) :617–21, June 2010.

[DCS+13] Guangyou Duan, Nils Christian, Jens Schwachtje, Dirk Walther, and Oliver Ebenhöh. The metabolic interplay between plants and phytopathogens. *Metabolites*, 3(1) :1–23, January 2013.

[DdME+08] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi : a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl 1) :D344–D350, 2008.

[DET14] Simon M. Dittami, Damien Eveillard, and Thierry Tonon. A metabolic approach to study algal–bacterial interactions in changing environments. *Molecular Ecology*, 23(7) :1656–1660, 2014.

[dFSKF09] Luis F. de Figueiredo, Stefan Schuster, Christoph Kaleta, and David A. Fell. Can sugars be produced from fatty acids ? a test case for pathway analysis tools. *Bioinformatics*, 25(1) :152–158, 2009.

[DGB+08] A Dereeper, V Guignon, G Blanc, S Audic, S Buffet, F Chevenet, J-F Dufayard, S Guindon, V Lefort, M Lescot, J-M Claverie, and O Gascuel. Phylogeny.fr : robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, 36(Web Server issue) :W465–9, 2008.

[DGG+12] Simon M. Dittami, Antoine Gravot, Sophie Goulitquer, Sylvie Rousvoal, Akira F. Peters, Alain Bouchereau, Catherine Boyen, and Thierry Tonon. Towards deciphering dynamic changes and evolutionary mechanisms involved in the adaptation to low salinities in ectocarpus (brown algae). *The Plant Journal*, 71(3) :366–377, June 2012.

[DGR+11] Simon M. Dittami, Antoine Gravot, David Renault, Sophie Goulitquer, Anja Eggert, Alain Bouchereau, Catherine Boyen, and Thierry Tonon. Integrative analysis of metabolite and transcript abundance during the short-term response to saline and oxidative stress in the brown alga Ectocarpus siliculosus. *Plant, cell & environment*, 34(4) :629–642, April 2011.

[DL05] Eric Davidson and Michael Levin. Gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14) :4935, 2005.

[dODN13] Cristiana Gomes de Oliveira Dal'Molin and Lars Keld Nielsen. Plant genome-scale metabolic reconstruction and modelling. *Current Opinion in Biotechnology*, 24(2) :271 – 277, 2013. Food biotechnology • Plant biotechnology.

[dODQP+09] C. G. de Oliveira Dal'Molin, L.-E. Quek, R. W. Palfreyman, S. M. Brumbley, and L. K. Nielsen. AraGEM, a genome-scale reconstruction of the primary metabolic network in arabidopsis. *PLANT PHYSIOLOGY*, 152(2) :579–589, December 2009.

[DPR+11] Simon M. Dittami, Caroline Proux, Sylvie Rousvoal, Akira F Peters, J Mark Cock, Jean-Yves Coppee, Catherine Boyen, and Thierry Tonon. Microarray estimation of genomic inter-strain variability in the genus Ectocarpus (Phaeophyceae). *BMC Molecular Biology*, 12(1) :2, January 2011.

[dQP+10a] Cristiana Gomes de Oliveira Dal'Molin, Lake-Ee Quek, Robin William Palfreyman, Stevens Michael Brumbley, and Lars Keld Nielsen. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant physiology*, 152(2) :579–89, February 2010.

[dQP+10b] Cristiana Gomes de Oliveira Dal'Molin, Lake-Ee Quek, Robin William Palfreyman, Stevens Michael Brumbley, and Lars Keld Nielsen. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant physiology*, 152(2) :579–89, February 2010.

[DSP+09] Simon M Dittami, Delphine Scornet, Jean-Louis Petit, Béatrice Ségurens, Corinne Da Silva, Erwan Corre, Michael Dondrup, Karl-Heinz Glatting, Rainer König, Lieven Sterck, Pierre Rouzé, Yves Van de Peer, J Mark Cock, Catherine Boyen, and Thierry Tonon. Global expression analysis of the brown alga ectocarpus siliculosus (phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress. *Genome Biology*, 10(6) :R66, 2009.

[EIP01] J. S. Edwards, R. U. Ibarra, and B. O. Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, 19(2) :125–130, February 2001.

[ELPH13] Ali Ebrahim, Joshua Lerman, Bernhard Palsson, and Daniel Hyduke. COBRApy : COnstraints-Based reconstruction and analysis for python. *BMC Systems Biology*, 7(1) :74, 2013.

[ENFB+14] Maria Enquist-Newman, Ann Marie E. Faust, Daniel D. Bravo, Christine Nicole S. Santos, Ryan M. Raisner, Arthur Hanel, Preethi Sarvabhowman, Chi Le, Drew D. Regitsky, Susan R. Cooper, Lars Peereboom, Alana Clark, Yessica Martinez, Joshua Goldsmith, Min Y. Cho, Paul D. Donohoue, Lily Luo, Brigit Lamberson, Pramila Tamrakar, Edward J. Kim, Jeffrey L. Villari, Avinash Gill, Shital A. Tripathi, Padma Karamchedu, Carlos J. Paredes, Vineet Rajgarhia, Hans Kristian Kotlar, Richard B. Bailey, Dennis J. Miller, Nicholas L. Ohler, Candace Swimmer, and Yasuo Yoshikuni. Efficient ethanol production from brown macroalgae sugars by a synthetic yeast platform. *Nature*, 505(7482) :239–243, January 2014.

[EP00] J. S. Edwards and B. O. Palsson. The escherichia coli mg1655 in silico metabolic genotype : Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10) :5528–5533, 2000.

[FBC+14] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L. L. Sonnhammer, John Tate, and Marco Punta. Pfam : the protein families database. *Nucleic Acids Research*, 42(D1) :D222–D230, 2014.

[Fin98] Gerald R Fink. Anatomy of a revolution. *Genetics*, 149(2) :473–477, 1998.

[FMC+14] Michele Fabris, Michiel Matthijs, Sophie Carbonelle, Tessa Moses, Jacob Pollier, Renaat Dasseville, Gino J. E. Baart, Wim Vyverman, and Alain Goossens. Tracking the sterol biosynthesis pathway of the diatom phaeodactylum tricornutum. *New Phytologist*, pages n/a–n/a, 2014.

[FMP03] Stephen S. Fong, Jennifer Y. Marciniak, and Bernhard Ø. Palsson. Description and interpretation of adaptive evolution of escherichia coli k-12 mg1655 by using a genome-scale in silico metabolic model. *Journal of Bacteriology*, 185(21) :6400–6408, 2003.

[FMR+12] Michele Fabris, Michiel Matthijs, Stephane Rombauts, Wim Vyverman, Alain Goossens, and Gino J.E. Baart. The metabolic blueprint of phaeodactylum tricornutum reveals a eukaryotic entner-doudoroff glycolytic pathway. *The Plant Journal*, 70(6) :1004–1014, June 2012.

[FS86] David A Fell and J Rankin Small. Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *Biochem. J*, 238 :781–786, 1986.

[GdODQPN11] Cristiana Gomes de Oliveira Dal'Molin, Lake-Ee Quek, Robin Palfreyman, and Lars Nielsen. Algagem - a genome-scale metabolic reconstruction of algae based on the chlamydomonas reinhardtii genome. *BMC Genomics*, 12(Suppl 4) :S5, 2011.

[GDR+10] Antoine Gravot, Simon M. Dittami, Sylvie Rousvoal, Raphael Lugan, Anja Eggert, Jonas Collén, Catherine Boyen, Alain Bouchereau, and Thierry Tonon. Diurnal oscillations of metabolite abundances and gene analysis provide new insights into central metabolic processes of the brown alga Ectocarpus siliculosus. *New Phytologist*, 188(1) :98–110, October 2010.

[GGC08] Bernhard Gschloessl, Yann Guermeur, and J Mark Cock. Hectar : A method to predict subcellular targeting in heterokonts. *BMC Bioinformatics*, 9(1) :393, 2008.

[GGJ+10] Matthieu Graindorge, Cécile Giustini, Anne Claire Jacomin, Alexandra Kraut, Gilles Curien, and Michel Matringe. Identification of a plant gene encoding glutamate/aspartate-prephenate aminotransferase : The last homeless enzyme of aromatic amino acids biosynthesis. *{FEBS} Letters*, 584(20) :4357 – 4360, 2010.

[GKD+07] M H Graham, B P Kinlan, L D Druehl, L E Garske, and S Banks. Deep-water kelp refugia as potential hotspots of tropical marine diversity and productivity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(42) :16576–16580, 2007.

[GKK+11] M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, and M. Schneider. Potassco : The Potsdam answer set solving collection. 24(2) :107–124, 2011.

[GKKS12] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Answer set solving in practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(3) :1–238, 2012.

[GKS12] M. Gebser, B. Kaufmann, and T. Schaub. Conflict-driven answer set solving : From theory to practice. *Artificial Intelligence*, 187-188 :52–89, 2012.

[GSM+14] Agnès Groisillier, Zhanru Shao, Gurvan Michel, Sophie Goulitquer, Patricia Bonin, Stefan Krahulec, Bernd Nidetzky, Delin Duan, Catherine Boyen, and Thierry Tonon. Mannitol metabolism in brown algae involves a new phosphatase family. *Journal of Experimental Botany*, 65(2) :559–570, 2014.

[GT10] Steinn Gudmundsson and Ines Thiele. Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11(1) :489, 2010.

[HCP+10] Svenja Heesch, Ga Youn Cho, Akira F Peters, Gildas Le Corguillé, Cyril Falentin, Gilles Boutet, Solène Coëdel, Claire Jubin, Gaelle Samson, Erwan Corre, Susana M Coelho, and J Mark Cock. A sequence-tagged genetic map for the brown alga Ectocarpus siliculosus provides large-scale assembly of the genome sequence. *The New phytologist*, 188(1) :42–51, April 2010.

[HDB+10] Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9) :977–982, August 2010.

[HE07] Thomas Handorf and Oliver Ebenhöh. Metapath online : a web server implementation of the network expansion algorithm. *Nucleic Acids Research*, 35(suppl 2) :W613–W618, 2007.

[HJFP06] Qiang Hua, Andrew R. Joyce, Stephen S. Fong, and Bernhard Ø. Palsson. Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains. *Biotechnology and Bioengineering*, 95(5) :992–1002, 2006.

[HR14] Joshua J. Hamilton and Jennifer L. Reed. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental Microbiology*, 16(1) :49–59, 2014.

[IEP02] Rafael U. Ibarra, Jeremy S. Edwards, and Bernhard O. Palsson. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912) :186–189, November 2002.

[JBBG08] Fabien Jourdan, Rainer Breitling, Michael P. Barrett, and David Gilbert. Metanetter : inference and visualization of high-resolution metabolomic networks. *Bioinformatics*, 24(1) :143–145, 2008.

[JLI00] J. D. Jordan, E. M. Landau, and R. Iyengar. Signaling networks : the origins of cellular multitasking. *Cell*, 103(2) :193–200, Oct 2000.

[JM61] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3 :318–356, Jun 1961.

[JTA+00] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804) :651–654, Oct 2000.

[KCVGC+05] Ingrid M. Keseler, Julio Collado-Vides, Socorro Gama-Castro, John Ingraham, Suzanne Paley, Ian T. Paulsen, Martín Peralta-Gil, and Peter D. Karp. Ecocyc : a comprehensive database resource for escherichia coli. *Nucleic Acids Research*, 33(suppl 1) :D334–D337, 2005.

[Kee04] Patrick J. Keeling. Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany*, 91(10) :1481–1493, 2004.

[KGS+14] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle : back to metabolism in kegg. *Nucleic Acids Research*, 42(D1) :D199–D205, 2014.

[Kit01] Hiroaki Kitano, editor. *Foundations of systems biology*. MIT Press, Cambridge (Mass.), London, 2001.

[KPK+10] Peter D Karp, Suzanne M Paley, Markus Krummenacker, Mario Latendresse, Joseph M Dale, Thomas J Lee, Pallavi Kaipa, Fred Gilham, Aaron Spaulding, Liviu Popescu, Tomer Altman, Ian Paulsen, Ingrid M Keseler, and Ron Caspi. Pathway Tools version 13.0 : integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 11(1) :40–79, January 2010.

[KPR02] Peter D Karp, Suzanne Paley, and Pedro Romero. The Pathway Tools software. *Bioinformatics*, 18 Suppl 1 :S225–32, January 2002.

[KYW+11] Elias W. Krumholz, Hong Yang, Pamela Weisenhorn, Christopher S. Henry, and Igor G. L. Libourel. Genome-wide metabolic network reconstruction of the picoalga ostreococcus. *Journal of Experimental Botany*, 2011.

[KYW+12] Elias W Krumholz, Hong Yang, Pamela Weisenhorn, Christopher S Henry, and Igor G L Libourel. Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *Journal of Experimental Botany*, 63(6) :2353–62, March 2012.

[LAF+98] Paul Liberator, Jennifer Anderson, Marc Feiglin, Mohinder Sardana, Patrick Griffin, Dennis Schmatz, and Robert W. Myers. Molecular cloning and functional expression of mannitol-1-phosphatase from the apicomplexan parasite eimeria tenella. *Journal of Biological Chemistry*, 273(7) :4237–4244, 1998.

[LBBLP+11] A. Le Bail, B. Billoud, S. Le Panse, S. Chenivesse, and B. Charrier. ETOILE regulates developmental patterning in the filamentous brown alga ectocarpus siliculosus. *The Plant Cell*, 23(4) :1666–1678, April 2011.

[LBG+12] Abdelhalim Larhlimi, Georg Basler, Sergio Grimbs, Joachim Selbig, and Zoran Nikoloski. Stoichiometric capacitance reveals the theoretical capabilities of metabolic networks. *Bioinformatics*, 28(18) :i502–i508, 2012.

[LDNS12] Nicolas Loira, Thierry Dulermo, Jean-Marc Nicaud, and David Sherman. A genome-scale metabolic model of the lipid-accumulating yeast yarrowia lipolytica. *BMC Systems Biology*, 6(1) :35, 2012.

[LKTK12] M. Latendresse, M. Krummenacker, M. Trupp, and P. D. Karp. Construction and completion of flux balance models from pathway databases. *Bioinformatics*, 28(3) :388–396, Feb 2012.

[LLB+01] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A.

Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921, Feb 2001.

[LLK05] Sang Yup Lee, Dong-Yup Lee, and Tae Yong Kim. Systems biotechnology for strain improvement. *Trends in Biotechnology*, 23(7) :349 – 358, 2005.

[LSR03] Li Li, Christian J. Stoeckert, and David S. Roos. Orthomcl : Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9) :2178–2189, 2003.

[MB13] Arne C. Müller and Alexander Bockmayr. Fast thermodynamically constrained flux variability analysis. *Bioinformatics*, 29(7) :903–909, 2013.

[MCDL+13] Laurence Meslet-Cladière, Ludovic Delage, Cédric J-J Leroux, Sophie Goulitquer, Catherine Leblanc, Emeline Creis, Erwan Ar Gall, Valérie Stiger-Pouvreau, Mirjam Czjzek, and Philippe Potin. Structure/Function analysis of a type iii polyketide synthase in the brown alga Ectocarpus siliculosus reveals a biochemical pathway in phlorotannin monomer biosynthesis. *The Plant cell*, 25(8) :3089–103, August 2013.

[MCR+11] Daniel Machado, Rafael Costa, Miguel Rocha, Eugenio Ferreira, Bruce Tidor, and Isabel Rocha. Modeling formalisms in systems biology. *AMB Express*, 1(1) :45, 2011.

[MER+11] Caroline Milne, James Eddy, Ravali Raju, Soroush Ardekani, Pan-Jun Kim, Ryan Senger, Yong-Su Jin, Hans Blaschek, and Nathan Price. Metabolic network reconstruction and genome-scale model of butanol-producing strain clostridium beijerinckii ncimb 8052. *BMC Systems Biology*, 5(1) :130, 2011.

[MGH+09] Ani Manichaikul, Lila Ghamsari, Erik F Y Hom, Chenwei Lin, Ryan R Murray, Roger L Chang, S Balaji, Tong Hao, Yun Shen, Arvind K Chavali, Ines Thiele, Xinping Yang, Changyu Fan, Elizabeth Mello, David E Hill, Marc Vidal, Kourosh Salehi-Ashtiani, and Jason A Papin. Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat Meth*, 6(8) :589–592, August 2009.

[MNP14] Jonathan Monk, Juan Nogales, and Bernhard O Palsson. Optimizing genome-scale network reconstructions. *Nature biotechnology*, 32(5) :447–452, 2014.

[MOMM+12] Shira Mintz-Oron, Sagit Meir, Sergey Malitsky, Eytan Ruppin, Asaph Aharoni, and Tomer Shlomi. Reconstruction of arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proceedings of the National Academy of Sciences*, 109(1) :339–344, 2012.

[MTS+10a] Gurvan Michel, Thierry Tonon, Delphine Scornet, J Mark Cock, and Bernard Kloareg. Central and storage carbon metabolism of the brown alga Ectocarpus siliculosus : insights into the origin and evolution of storage carbohydrates in Eukaryotes. *New Phytologist*, 188(1) :67–81, October 2010.

[MTS+10b] Gurvan Michel, Thierry Tonon, Delphine Scornet, J Mark Cock, and Bernard Kloareg. The cell wall polysaccharide metabolism of the brown alga Ectocarpus siliculosus. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. *The New phytologist*, 188(1) :82–97, October 2010.

[MWK+08] Patrick May, Stefanie Wienkoop, Stefan Kempa, Björn Usadel, Nils Christian, Jens Rupprecht, Julia Weiss, Luis Recuenco-Munoz, Oliver Ebenhöh, Wolfram Weckwerth, and Dirk Walther. Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of chlamydomonas reinhardtii. *Genetics*, 179(1) :157–166, 2008.

[MYD11] Hiroshi Maeda, Heejin Yoo, and Natalia Dudareva. Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogenate. *Nature chemical biology*, 7(1) :19–21, 2011.

[MZR03] Lukas A. Mueller, Peifen Zhang, and Seung Y. Rhee. Aracyc : A biochemical pathway database for arabidopsis. *Plant Physiology*, 132(2) :453–460, 2003.

[NGMS08] Zoran Nikoloski, Sergio Grimbs, Patrick May, and Joachim Selbig. Metabolic networks are np-hard to reconstruct. *Journal of Theoretical Biology*, 254(4) :807 – 816, 2008.

[NL13] David J Sherman Nicolas Loira, Anna Zhukova. Pantograph : A scaffold-based method for genome-scale metabolic model reconstruction. *To appear*, unknown, 2013.

[Pal02] B. Palsson. In silico biology through "omics". *Nat. Biotechnol.*, 20(7) :649–650, Jul 2002.

[PBM03] Priti Pharkya, Anthony P. Burgard, and Costas D. Maranas. Exploring the overproduction of amino acids using the bilevel optimization framework optknock. *Biotechnology and Bioengineering*, 84(7) :887–899, 2003.

[PBM04] Priti Pharkya, Anthony P. Burgard, and Costas D. Maranas. Optstrain : A computational framework for redesign of microbial production systems. *Genome Research*, 14(11) :2367–2376, 2004.

[PCD+14] Sylvain Prigent, Guillaume Collet, Simon M. Dittami, Ludovic Delage, Floriane Ethis de Corny, Olivier Dameron, Damien Eveillard, Sven Thiele, Jeanne Cambefort, Catherine Boyen, Anne Siegel, and Thierry Tonon. The genome-scale metabolic network of ectocarpus siliculosus (ectogem) : a resource to study brown algal physiology and beyond. *The Plant Journal*, pages n/a–n/a, 2014.

[PMSF09] Mark G Poolman, Laurent Miguet, Lee J Sweetlove, and David A Fell. A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant physiology*, 151(3) :1570–81, November 2009.

[PRP+03] Nathan D. Price, Jennifer L. Reed, Jason A. Papin, Sharon J. Wiback, and Bernhard O. Palsson. Network-based analysis of metabolic regulation in the human red blood cell. *Journal of Theoretical Biology*, 225(2) :185 – 194, 2003.

[PRP04] Nathan D Price, Jennifer L Reed, and Bernhard Ø Palsson. Genome-scale models of microbial cells : evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11) :886–897, 2004.

[PVC+10] Akira F. Peters, Serinde J. Van Wijk, Ga Youn Cho, Delphine Scornet, Takeaki Hanyuda, Hiroshi Kawai, Declan C. Schroeder, J. Mark Cock, and Sung Min Boo. Reinstatement of Ectocarpus crouaniorum Thuret in Le Jolis as a third common species of Ectocarpus (Ectocarpales, Phaeophyceae) in Western Europe, and its phenology at Roscoff, Brittany. *Phycological Research*, 58(3) :157–170, May 2010.

[PVPF04] M.G. Poolman, K.V. Venkatesh, M.K. Pidcock, and D.A. Fell. A method for the determination of flux in elementary modes, and its application to lactobacillus rhamnosus. *Biotechnology and Bioengineering*, 88(5) :601–612, 2004.

[PW09] P. E. Purnick and R. Weiss. The second wave of synthetic biology : from modules to systems. *Nat. Rev. Mol. Cell Biol.*, 10(6) :410–422, Jun 2009.

[QYQ+07] Weiqiang Qian, Chunmei Yu, Huanju Qin, Xin Liu, Aimin Zhang, Ida Elisabeth Johansen, and Daowen Wang. Molecular and functional analysis of phosphomannomutase (pmm) from higher plants and genetic evidence for the involvement of pmm in ascorbic acid biosynthesis in arabidopsis and nicotiana benthamiana. *The Plant Journal*, 49(3) :399–413, 2007.

[RC09] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems : applications and challenges. *Briefings in Bioinformatics*, 10(4) :435–449, 2009.

[RDG+14] Andres Ritter, Simon Dittami, Sophie Goulitquer, Juan Correa, Catherine Boyen, Philippe Potin, and Thierry Tonon. Transcriptomic and metabolomic analysis of copper stress acclimation in ectocarpus siliculosus highlights signaling and tolerance mechanisms in brown algae. *BMC Plant Biology*, 14(1) :116, 2014.

[RGD+11] S. Rousvoal, A. Groisillier, S. M. Dittami, G. Michel, C. Boyen, and T. Tonon. Mannitol-1-phosphate dehydrogenase activity in Ectocarpus siliculosus, a key role for mannitol synthesis in brown algae. *Planta*, 233(2) :261–273, Feb 2011.

[RK01] P. R. Romero and P. Karp. Nutrient-related analysis of pathway/genome databases. *Pac Symp Biocomput*, pages 471–482, 2001.

[RPC+06] Jennifer L. Reed, Trina R. Patel, Keri H. Chen, Andrew R. Joyce, Margaret K. Applebee, Christopher D. Herring, Olivia T. Bui, Eric M. Knight, Stephen S.

Fong, and Bernhard O. Palsson. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A*, 103(46) :17480–17484, November 2006. PMID : 17088549 PMCID : PMC1859954.

[RPG+09] P. Rippert, J. Puyaubert, D. Grisollet, L. Derrier, and M. Matringe. Tyrosine and phenylalanine are synthesized within the plastids in Arabidopsis. *Plant Physiol.*, 149(3) :1251–1260, Mar 2009.

[RSS01] Maido Remm, Christian E.V. Storm, and Erik L.L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5) :1041 – 1052, 2001.

[RUR+10] Andrés Ritter, Martin Ubertini, Sarah Romac, Fanny Gaillard, Ludovic Delage, Aaron Mann, J Mark Cock, Thierry Tonon, Juan A Correa, and Philippe Potin. Copper stress proteomics highlights local adaptation of two strains of the model brown alga Ectocarpus siliculosus. *Proteomics*, 10(11) :2074–88, June 2010.

[RVSP03] Jennifer L Reed, Thuy D Vo, Christophe H Schilling, and Bernhard O Palsson. An expanded genome-scale model of escherichia coli k-12 (iJR904 GSM/GPR). *Genome Biol*, 4(9) :R54, 2003. PMID : 12952533 PMCID : PMC193654.

[SDF99] S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks : a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, 17(2) :53–60, Feb 1999.

[SHH12] Samuel M. D. Seaver, Christopher S. Henry, and Andrew D. Hanson. Frontiers in metabolic reconstruction and modeling of plant genomes. *Journal of Experimental Botany*, 63(6) :2247–2258, 2012.

[SKDM07] Vinay Satish Kumar, Madhukar S Dasika, and Costas D Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8 :212, 2007.

[SM06] Günter Schwarz and Ralf R. Mendel. Molybdenum cofactor biosynthesis and molybdenum enzymes. *Annual Review of Plant Biology*, 57(1) :623–647, 2006. PMID : 16669776.

[SSM11a] Rajib Saha, Patrick F. Suthers, and Costas D. Maranas. <italic>zea mays i</italic>rs1563 : A comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE*, 6(7) :e21784, 07 2011.

[SSM+11b] N. Swainston, K. Smallbone, P. Mendes, D. Kell, and N. Paton. The SuBliMinaL Toolbox : automating steps in the reconstruction of metabolic networks. *J Integr Bioinform*, 8(2) :186, 2011.

[ST09] Torsten Schaub and Sven Thiele. Metabolic network expansion with answer set programming. In *Logic Programming 25th International Conference*, pages 312–326. Springer Berlin Heidelberg, 2009.

[TEP+11] Thierry Tonon, Damien Eveillard, Sylvain Prigent, Jérémie Bourdon, Philippe Potin, Catherine Boyen, and Anne Siegel. Toward systems biology in brown algae to explore acclimation and adaptation to the shore environment. *Omics : a journal of integrative biology*, 15(12) :883–892, December 2011. PMID : 22136637.

[TG10] Vered Tzin and Gad Galili. New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Molecular Plant*, 3(6) :956–972, 2010.

[THS+11] Ines Thiele, Daniel Hyduke, Benjamin Steeb, Guy Fankam, Douglas Allen, Susanna Bazzani, Pep Charusanti, Feng-Chi Chen, Ronan Fleming, Chao Hsiung, Sigrid De Keersmaecker, Yu-Chieh Liao, Kathleen Marchal, Monica Mo, Emre Ozdemir, Anu Raghunathan, Jennifer Reed, Sook-Il Shin, Sara Sigurbjornsdottir, Jonas Steinmann, Suresh Sudarsan, Neil Swainston, Inge Thijs, Karsten Zengler, Bernhard Palsson, Joshua Adkins, and Dirk Bumann. A community effort towards a knowledge-base and mathematical model of the human pathogen salmonella typhimurium lt2. *BMC Systems Biology*, 5(1) :8, 2011.

[TJGF11] Manuel Tejada-Jiménez, Aurora Galván, and Emilio Fernández. Algae and humans share a molybdate transporter. *Proceedings of the National Academy of Sciences*, 108(16) :6420–6425, 2011.

[TP10] Ines Thiele and Bernhard ØPalsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1) :93–121, January 2010.

[TVC+11] Raimund Tenhaken, Elena Voglas, J. Mark Cock, Volker Neu, and Christian G. Huber. Characterization of gdp-mannose dehydrogenase from the brown alga ectocarpus siliculosus providing the precursor for the alginate polymer. *Journal of Biological Chemistry*, 286(19) :16707–16715, 2011.

[VBC+13] David Vallenet, Eugeni Belda, Alexandra Calteau, Stéphane Cruveiller, Stefan Engelen, Aurélie Lajus, François Le Fèvre, Cyrille Longin, Damien Mornico, David Roche, Zoé Rouy, Gregory Salvignol, Claude Scarpelli, Adam Alexander Thil Smith, Marion Weiman, and Claudine Médigue. Microscope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Research*, 41(D1) :D636–D647, 2013.

[vIPK+10] Martijn van Iersel, Alexander Pico, Thomas Kelder, Jianjiong Gao, Isaac Ho, Kristina Hanspers, Bruce Conklin, and Chris Evelo. The bridgedb framework : standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics*, 11(1) :5, 2010.

[WGH+12] Thilo Winzer, Valeria Gazda, Zhesi He, Filip Kaminski, Marcelo Kern, Tony R Larson, Yi Li, Fergus Meade, Roxana Teodor, Fabián E Vaistij, et al. A papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, 336(6089) :1704–1708, 2012.

[WK96] J West and G Kraft. Ectocarpus siliculosus (Dillwyn) Lyngb. from Hopkins River Falls, Victoria - the first record of a freshwater brown alga in Australia. *Muelleria*, 9 :29–33, 1996.

[WLW+12] Adam J Wargacki, Effendi Leonard, Maung Nyan Win, Drew D Regitsky, Christine Nicole S Santos, Peter B Kim, Susan R Cooper, Ryan M Raisner, Asael Herman, Alicia B Sivitz, Arun Lakshmanaswamy, Yuki Kashiyama, David Baker, and Yasuo Yoshikuni. An engineered microbial platform for direct

biofuel production from brown macroalgae. *Science*, 335(6066) :308–13, January 2012.

[WQJ13] Na Wei, Josh Quarterman, and Yong-Su Jin. Marine macroalgae : an untapped resource for producing fuels and chemicals. *Trends in Biotechnology*, 31(2) :70 – 77, 2013.

[YHC+04] Hwan Su Yoon, Jeremiah D Hackett, Claudia Ciniglia, Gabriele Pinto, and Debashish Bhattacharya. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular biology and evolution*, 21(5) :809–18, May 2004.

[YWQ+13] Heejin Yoo, Joshua R Widhalm, Yichun Qian, Hiroshi Maeda, Bruce R Cooper, Amber S Jannasch, Itay Gonda, Efraim Lewinsohn, David Rhodes, and Natalia Dudareva. An alternative pathway contributes to phenylalanine biosynthesis in plants via a cytosolic tyrosine : phenylpyruvate aminotransferase. *Nature communications*, 4, 2013.

Table des figures

1.1 **Représentation des principaux processus se produisant dans une cellule.** La cellule reçoit des signaux extérieurs qui engendrent des cascades de signalisation à l'intérieur de la cellule. Ces signaux et les cascades induites sont étudiées dans les *réseaux de signalisation*. Ce signal arrive jusqu'au noyau et engendre une régulation de l'expression des gènes. Cette régulation sera étudiée par les *réseaux de régulation géniques*. Les gènes sont ensuite transcrits en ARN messagers qui pourront être traduits en protéines. Certaines de ces protéines (appelées enzymes) pourront catalyser des réactions métaboliques qui transformeront les molécules présentes à l'intérieur de la cellule ou importées de l'extérieur. Ces réactions sont étudiées dans les *réseaux métaboliques*. 9

1.2 **Représentation graphique d'un réseau métabolique composé de deux réactions.** La première réaction ($R1$) est irréversible et produit B à partir de A. La seconde ($R2$) transforme $B + C$ en D de manière réversible. 11

1.3 **Les méthodes classiques à base de contraintes pour l'analyse de réseaux métaboliques [PRP04].** La construction de l'espace des solutions est représentée au centre, les méthodes d'analyse de cet espace des solutions sont représentées autour. 16

1.4 **Couverture phylogénétique des reconstructions de réseaux métaboliques en février 2013[MNP14].** 17

1.5 **Vue d'ensemble de la procédure de reconstruction de réseaux métaboliques [TP10].** Les étapes 2 à 4 seront itérées jusqu'à ce que les prédictions réalisées par le réseaux métaboliques soient en accord avec les observations biologiques. 18

1.6 **Limitations actuelles au développement rapide de la reconstruction de réseaux métaboliques.** Les éventuels points à améliorer sont indiqués en gris, les méthodes à base de contraintes étant au centre des améliorations [MNP14]. 20

1.7 **Support apporté par les différents outils de reconstruction de réseaux métaboliques, de la création de l'ébauche jusqu'à l'évaluation de la qualité du réseau [HR14].** Les quatre étapes principales et certains des 96 points relevés par [TP10] sont étudiés. 25

1.8 **Photo de sporophyte d'*Ectocarpus siliculosus* en culture.** 38

1.9 **Représentation simplifiée de l'arbre phylogénétique des eucaryotes [CPC11]. Les cinq groupes majeures ayant développé une multicellularité complexe sont représenté en couleur. La longueur des barres de couleur indique le temps relatif approché de développement de la multicellularité dans chaque lignée.** 38

2.1 **Représentation schématique d'un réseau métabolique[AMC+12].** Les nœuds représentent les métabolites et les hyper-arcs représentent les réactions. Les nœuds gris correspondent aux sources et le nœud noir la cible. 43

2.2 **Représentation de la répartition statistique du nombre de métabolites non productibles en fonction de la taille de la base de données utilisée.** Les boites correspondent à la taille du premier et du troisième quartile, la ligne centrale correspond à la moyenne et les barres correspondent aux valeurs extrêmes. . . 45

2.3 **Temps de calcul de Clasp et Unclasp pour la recherche de la taille minimale de complétion.** Les cercles transparents correspondent à la médiane des expérimentations avec Clasp. Les carrés grisés correspondent à la médiane des expérimentations avec Unclasp. Pour chaque expérience, les valeurs minimale et maximale sont reportées avec des lignes verticales. 50

2.4 **Temps de calcul de Clasp pour énumérer l'ensemble des solutions** Les cercles transparents correspondent aux médianes des durées des expérimentations. Les points noirs correspondent aux médianes du nombre de solutions. Les lignes verticales correspondent aux valeurs maximales et minimales. 52

2.5 **Temps de calcul de Clasp pour déterminer les intersections des solutions optimales.** Les cercles transparents correspondent aux médianes des résultats de durée des expérimentations. Les points noirs correspondent à la taille médiane des intersections. Les lignes verticales correspondent aux valeurs maximales et minimales. 53

2.6 Exemple de deux réactions impliquant les mêmes molécules et provenant de la même base de données (MetaCyc), l'une étant réversible et l'autre irréversible. 54

2.7 **Temps de calcul de Clasp et Unclasp pour la recherche de la taille minimale de complétion.** Les cercles correspondent à la médiane des expérimentations avec Clasp. Les carrés correspondent à la médiane des expérimentations avec Unclasp. Les ronds et carrés grisés correspondent à la modélisation avec prise en compte de la réversibilité, les ronds et carrés vides correspondent à la même modélisation sans cette prise en compte. Pour chaque expérience, les valeurs minimale et maximale sont reportées avec des lignes verticales. 56

2.8 **Deux exemples de réseaux métaboliques contenant des cycles.** Les métabolites sont représentés par des ovales verts, les réactions sont représentées par des rectangles rouges. La direction des réactions est représentée par la direction des arcs. La stœchiométrie des réactions n'est pas représentée. . . . 58

2.9 **Exemples de réseaux métaboliques.** Les cercles représentent les métabolites, les rectangles représentent les réactions. Les chiffres sur les arcs représentent la stœchiométrie des réactions. Les métabolites S correspondent aux sources et les métabolites T correspondent aux cibles. 60

2.10 **Tailles des solutions et de l'union des solutions selon la sémantique de productibilité utilisée.** Les croix correspondent aux expériences utilisant la productibilité d'Acuna-Sagot, les carrés celle de *Meneco*. Les courbes noires correspondent à la taille des optimaux, les courbes grises correspondent à la taille des unions. Il n'existe pas de valeur pour l'union avec Acuna pour une base de données de 10.000 réactions, les calculs étant trop long. 62

2.11 **Représentation de la répartition statistique des différentes complétions en fonction de la taille de la base de données utilisée.** La partie a) correspond à la complétion utilisant la sémantique proposée par Acuna et Sagot, la partie droite correspond à la complétion utilisée par *Meneco*. Les boites correspondent à la taille du premier et du troisième quartile, la ligne centrale correspond à la moyenne et les barres correspondent aux valeurs extrêmes. 63

2.12 **Distribution des pourcentages de réactions rajoutées aux réseaux dégradés selon les différentes classes de réactions et le pourcentage de dégradation.** Les boites représentent l'espace interquartile, les moustaches représentent 1.5 espace interquartile. Les ronds représentent les valeurs extrêmes sûres (vides) ou potentielles (pleins). 65

2.13 **Distribution des pourcentages de réactions rajoutées au réseau vide selon les différentes classes de réactions.** Les boites représentent l'espace interquartile, les moustaches représentent 1.5 espace interquartile. Les ronds représentent les valeurs extrêmes sûres (vides) ou potentielles (pleins). Les réactions essentielles à la production de biomasse (au sens de la FVA) sont représentées en bleu, les réactions bloquées qui ne possédent jamais aucun flux non nul en FVA sont représentées en violet, les réactions alternatives pouvant posséder un flux nul ou non nul lors d'optimisation de la fonction objective sont représentées en jaune. 66

3.1 **Composition des 432 ensembles de 44 réactions candidates à la complétion pour permettre la productibilité des 50 cibles. Chaque ensemble de réactions peut être décomposé en un ensemble de 35 réactions présentes partout auquel nous ajoutons une réaction de chaque groupe représenté, deux réactions du groupe représenté en haut à gauche et une paire de deux réactions.** 75

3.2 **Résumé de l'approche intégrative de la reconstruction du réseau métabolique d'*Ectocarpus siliculosus*.** Les boites bleues correspondent aux outils utilisés. Les boites vertes correspondent aux données utilisées et aux différentes versions du réseau existantes. 82

4.1 **Comparaison entre la reconstruction automatique et l'annotation manuelle de la voie de synthèse des alginates.** * indique un gène pour lequel l'identifiant a changé dans la base de données en fonction de la mise à jour de l'annotation structurale du génome (Esi0195_0005 est devenu Esi0195_0002). ** indique une protéine dont la fonction a été caractérisée biochimiquement [TVC+11]. 87

4.2 **Comparaison entre la reconstruction automatique et l'annotation manuelle de la voie du cycle du mannitol.** ** indique une protéine dont la fonction a été caractérisée biochimiquement. [RGD+11, GSM+14] 89

4.3 **Les voies de synthèse des acides aminés aromatiques chez les plantes. [TG10]** Les noms des enzymes sont donnés à côté des flèches, et sont accompagnés des noms abrégés entre parenthèses. Les voies majoritaires de synthèse de la phénylalanine et de la tyrosine chez les plantes sont indiquées par des flèches vertes (voie II pour la tyrosine). La voie de synthèse I de la tyrosine et une voie de synthèse secondaire de la phénylalanine sont représentées par des flèches rouges. 90
4.4 **Voies de biosynthèse des acides aminés aromatiques avant et après curation manuelle.** (a) Voies métaboliques prédites dans *EctoGEM-functionnal*. Les commentaires associés aux enzymes indiquent comment les gènes ont été identifiés et les réactions prédites : PWT (Pathway Tools), PTG (Pantograph) et HMM. (b) Voies métaboliques obtenues après curation manuelle par analyse comparative avec d'autres espèces. La voie métabolique de l'arogénate est représentée en pointillés. Les changements d'expression significatifs (FDR < 5%) dûs à des stress cuivriques (après 4h et 8h de traitement), hypersalins, hyposalins et oxydatifs (après 6h de traitement pour les trois derniers types de stress) sont indiqués dans cet ordre par les flèches sous les noms des gènes. 91
4.5 **Évolution des enzymes impliquées dans la synthèse des acides aminés aromatiques.** L'arbre représente les relations phylogénétiques entre les organismes telles que définies dans Tree Of Life (http ://tolweb.org). La longueur des branches ne représente aucune information. La présence des gènes chez les différents organismes est indiquée par un "+". Les flèches indiquent la fusion des déshydratases (A/P-DT) et déshydrogénases (A/P-DH). *Phytophtora infestans* contient deux déshydratases, l'une étant fusionnée à la déshydrogénase. 93
4.6 **Analyse phylogénétique de la A/P-DT d'*Ectocarpus siliculosus* .** Pour chaque séquence, la première partie du nom correspond au numéro d'accession dans la base de données NCBI et la seconde partie correspond au nom de l'espèce où la séquence a été retrouvée. Les couleurs correspondent aux origines phylogénétiques des espèces. 94
4.7 **Analyse phylogénétique de la A/P-DH d'*Ectocarpus siliculosus* .** Pour chaque séquence, la première partie du nom correspond au numéro d'accession dans la base de données NCBI et la seconde partie correspond au nom de l'espèce où la séquence a été retrouvée. Les couleurs correspondent aux origines phylogénétiques des espèces. 95
4.8 **Voies de biosynthèse du cofacteur molybdenum.** a) Voies métaboliques curées et protéines associées aux réactions enzymatiques. * marque les gènes exprimés chez *E. siliculosus*. ** marque les gènes significativement (FDR < 5%) réprimés en réponse à un stress hyposalin. Les nombres dans les ovales correspondent aux numéros E.C. des réactions. (b) Position phylogénétique de deux candidats pour l'activité molybdopterin synthase. La figure montre un arbre de maximum de vraisemblance avec 75 séquences sélectionnées dans différentes lignées. Seules les valeurs possédant un bootstrap supérieur à 50 sont indiquées. 97

Annexes

Annexe A

Liste des molécules utilisées lors de l'étude de l'efficacité de *Meneco*

TABLE A.1: **Liste des molécules utilisées pour la complétion du réseau métabolique**Les identifiants en majuscule correspondent aux identifiants MetaCyc, les minuscules aux noms des molécules (quand les identifiants n'étaient pas suffisamment parlant)

Graines	**Cibles**
Electron Acceptor	4-amino-butyrate
ACP	EICOSAPENTAENOATE
ADP	AMMONIUM
AMMONIA	ARACHIDIC ACID
AMP	ARACHIDONIC ACID
ATP	ARG
BIOTIN	ASN
CA+2	CIT
CARBON-DIOXIDE	eicosadienoate
CPD-12921	erucate
CL-	myristate
CO-A	γ-linolenate
cobalt chloride	stearidonate
manganese chloride	di-homo-γ-linolenate
molybdate	eicosatetraenoate
cyanocob(III)alamin	palmitoleate
CU+	Cis-vaccenate
CU+2	CYS
Donor-H2	DOCOSANOATE
GDP	GLC
EDTA	GLN
GTP	GLT
HYDROGEN-PEROXIDE	GLUTATHIONE
K+	GLY
MG+2	GLYCERATE

Graines	**Cibles**
NA+	GLYCEROL
NAD	GLYCOLLATE
NADH	HIS
NADP	ILE
NADPH	ALANINE
NITRATE	ASPARTATE
OXYGEN-MOLECULE	ORNITHINE
Pi	LEU
PPI	LINOLEIC ACID
PROTON	LINOLENIC ACID
SULFATE	LYS
THIAMINE	MANNITOL
UDP	MET
Vitamins-B12	OLEATE
COB-I-ALAMIN	PALMITATE
WATER	PHE
ZN+2	PRO
	SER
	STEARIC ACID
	SUC
	THR
	Iso-citrate
	TRP
	TYR
	VAL

Annexe B

Code ASP original de Network-Expansion

En 2009, Torsten Schaub et Sven Thiele proposent une méthode [ST09] permettant de compléter des réseaux métaboliques d'un point de vue qualitatif en s'affranchissant de la nécessité d'obtention de données cinétiques. Par rapport aux approches précédentes, l'idée est de vérifier la fonctionnalité du réseau d'un point de vue qualitatif (accessibilité des métabolites) sans prendre en compte les contraintes induites par la stoechiométrie du système, qui pourront être vérifiée dans un second temps. En reformulant ce problème de manière combinatoire, Schaub et Thiele proposent d'utiliser des méthodes d'optimisation récentes pour énumérer intégralement l'espace des solutions au problème.

Les auteurs sont partis du constat que la grande majorité des réseaux métaboliques reconstruits automatiquement sont incomplets. De plus, les bases de données des réactions métaboliques sont de plus en plus exhaustives et leur utilisation n'en devient que plus pertinente. Enfin, il y a à disposition des données biologiques pour la plupart des organismes biologiques pour lesquels un réseau métabolique est reconstruit ou en cours de reconstruction, et qui sont utilisables pour l'étape de complétion.

B.1 ASP

La modélisation et la résolution du problème sont basée sur des technologies apparues relativement récemment, dite de programmation par ensemble réponse (ou Answer Set Programming, ASP), qui sont connues pour être particulièrement efficace pour résoudre des problèmes NP-dur. La programmation par ensembles réponses (ou Answer Set Programming, ASP) [Bar03, GKKS12] propose un cadre déclaratif pour modéliser des problèmes combinatoires. Les caractères déclaratifs et la haute performance des solveurs ASP actuels permettent de se concentrer sur les problèmes de modélisation plutôt que de rechercher des façons intelligentes de traiter ces problèmes. L'idée de base d'ASP est d'exprimer un problème sous forme logique de manière à ce que les modèles sortant de cette représentation donnent les solutions au problème initial. Les problèmes sont exprimés en tant que programmes logiques et les modèles résultants sont appelés "ensembles réponses" (ou Answer Sets). Ces ensembles-réponses sont identifiées à l'aide de technologies inspirées par les approches SAT et les bases de données. Il est bien évidemment possible de déterminer, à l'aide

d'ASP, si un programme possède un ensemble réponse, mais d'autres modes de raisonnement sont nécessaires pour couvrir l'ensemble des problèmes que l'on peut rencontrer en pratique. Ainsi il sera possible d'identifier aisément l'intersection ou l'union de l'ensemble des ensembles réponses. De la même manière, lister l'intégralité des ensembles réponses pourra être possible.

L'idée générale consiste à représenter les modèles étudiés sous forme de faits logiques afin de pouvoir raisonner sur ces faits par des ensembles de règles et de contraintes. les règles seront de la forme :

Listing 5:

```
1 t :- a_1, ..., a_n.
```

avec t : la tête de la règle et $a_{1...n}$ les atomes du corps de la règles. Dans le cas présent t sera vrai si l'ensemble des atomes du corps de la règle sont vrais. Un grounder va instancier l'ensemble des règles présentes dans un programme ASP à partir des faits logiques afin de transformer l'ensemble en un langage compréhensible par un solveur.

Durant l'ensemble de la thèse, il a été choisi de travailler avec la suite logicielle Potassco, vainqueur de nombreux concours par la grande efficacité de son solveur. Ce solveur (Clasp) a grandement évolué durant les trois années de la thèse et continu encore aujourd'hui d'évoluer : nouvelles heuristiques, nouveaux solveurs, nouvelle syntaxe avec l'arrivée du grounder gringo4.

B.2 Network-Expansion

La reconstruction de réseaux métaboliques est, par nature, incomplète. En effet, celle-ci se base classiquement sur des données d'annotations génomiques qui sont intrinsèquement imprécises, même si l'annotation a été effectuée manuellement. De plus, il peut arriver que certaines parties d'un génome n'aient pas été annotées du tout. Ces deux raisons impliquent qu'un réseau métabolique contient forcément des trous qu'il convient de combler pour coller au mieux à la réalité biologique. L'idée générale des auteurs est qu'une réaction ne peut avoir lieu uniquement que si les réactants de celle-ci sont présents, soit dans un milieu de culture donné, soit comme produit d'une autre réaction. Ainsi, en partant des métabolites présents dans un milieu de culture (les métabolites graines) il devient possible de déterminer le "scope" des graines, c'est à dire l'ensemble des métabolites productibles à partir des graines et des réactions présentes dans un modèle.

En modélisant ce concept en ASP, les auteurs ont réussi à développer un outil permettant de compléter des réseaux métaboliques en utilisant des données expérimentales tel que la présence de certains métabolites cibles dans une cellule. Ainsi, en regardant si ces cibles font partie du scope des graines, le réseau métabolique peut être considéré comme suffisamment complet pour expliquer l'existence de ces cibles. Si ce n'est pas le cas, il conviendra d'ajouter des réactions dans le réseau pour permettre aux cibles de faire partie du scope des graines. Ajouter l'ensemble des réactions provenant d'une base de données de réactions comme MetaCyc permet (généralement) de produire l'ensemble des cibles. Cependant, toute capacité d'étude biologique ultérieure du réseau sera perdue du fait de la trop grande quantité de faux-positifs. Les auteurs proposent donc d'utiliser un principe de parcimonie pour minimiser le nombre total de réactions à ajouter au réseau, et ainsi minimi-

ser les modifications faites à un réseau que l'on considère initialement de bonne qualité. Ce problème est par nature extrêmement combinatoire et donne de très nombreuses solutions. L'utilisation des techniques d'optimisation particulièrement efficaces de la programmation par ensembles réponses permet de résoudre ce problème.

B.3 Modélisation

Un réseau métabolique est représenté comme un graphe dirigé bipartite $G = (R \cup M, E)$ où R et M sont des nœuds représentant respectivement les réactions et les métabolites. Les réactants d'une réaction correspondent à l'ensemble des nœuds reliés par des arêtes entrant dans un nœud R ($reac(r) = \{m \in M | (m,r) \in E\}$) et les produits aux arêtes sortantes ($prod(r) = \{m \in M | (r,m) \in E\}$).

Étant donné un réseau métabolique $(R \cup M, E)$ et un sous ensemble $M' \subseteq M$ de métabolites graines, une réaction $r \in R$ sera "atteignable" depuis M' si l'ensemble de ses réactants est dans le scope de M', c'est à dire si $reac(r) \subseteq M'$. De plus un métabolite $m \in M$ est atteignable depuis M' si $m \in M'$ ou si $m \in prod(r)$ pour au moins une réaction $r \in R$ atteignable depuis M'. Ainsi, le scope de M', noté $\Sigma(M')$, pourra être calculé en temps polynomial.

Une fois le scope d'un ensemble de métabolites défini, il devient possible de regarder de plus près la complétion de réseaux métaboliques. Pour cela, il faut également définir un sous-ensemble de métabolites $T \in M$ correspondant aux métabolites cibles (ou $targets$). On a également un second réseau métabolique de référence $(R' \cup M', E')$ qui correspond à la base de données de réactions métaboliques. On va donc chercher un ensemble de réactions $R'' \subseteq R' \setminus R$ telle que $T \subseteq \Sigma_G(S)$ avec :

$$G = ((R \cup R'') \cup (M \cup M''), E \cup E'')$$
$$M'' = \{m \in M' | r \in R'', m \in reac(r) \cup prod(r)\}$$
$$E'' = \{(m,r) \in E' | r \in R'', m \in reac(r)\} \cup \{(r,m) \in E' | r \in R'', m \in prod(r)\}$$

R'' sera appelé "complétion" de $(R \cup M, E)$ à partir de $(R' \cup M', E')$ par rapport à S et T.

L'ensemble de cette modélisation a été représenté en programmation logique. Le scope d'un réseau métabolite N peut alors être défini de la manière suivante en ASP :

Listing 6:

```
scope(M) :- seed(M).
scope(M) :- product(M,R), reaction(R,N), draft(N), scope(M') : reactant(
    M',R).
```

Ainsi, tous les métabolites graine font partie du scope par nature. Ensuite on défini récursivement qu'un produit M d'une réaction R appartenant à un réseau N fait également partie du scope si tous les réactants M' de cette réaction R font déjà partie du scope. On pourra ainsi décrire un scope pour le draft métabolique ($dscope$) et un scope potentiel qui contient l'union des réactions du draft et de la base de données ($pscope$).

Pour réaliser la complétion il va falloir choisir des réactions provenant de la base de données mais pas inclues dans le draft métabolique N :

Listing 7:

```
{xreaction(R) : not reaction(R,N) : draft(N)}.
```

Le scope des graines du draft métabolique N étendu par la base de données sera donc :

Listing 8:

```
xscope(M) :- seed(M).
xscope(M) :- product(M,R), reaction(R,N), draft(N), xscope(M') :
    reactant(M',R).
xscope(M) :- product(M,R), reaction(R,N), xscope(M') : reactant(M',R).
```

Enfin, il faut que l'ensemble des métabolites cibles puisse être produit, et donc qu'ils fassent partie du scope étendu des graines. Cela se représente par la contrainte d'intégrité suivante :

Listing 9:

```
:- target(M), not xscope(M).
```

L'encodage précédent permet d'obtenir l'ensemble des complétions possibles. Cependant, le nombre de celles-ci étant gigantesque, il y a un gros risque que l'on ne passe pas à l'échelle. Afin de réduire le nombre de solutions possibles, différents ajustements ont été réalisés.

Tout d'abord nous ne nous intéressons qu'à certaines réactions de la base de données, réactions que nous qualifierons d'"intéressantes" :

Listing 10:

```
:- xreaction(R), not ireaction(R).
```

Ces réactions d'intérêt sont toutes les réactions du scope des métabolites graines qui se trouvent en amont des métabolites cibles.

Listing 11:

```
ireaction(R) :- interesting(M), product(M,R), reaction(R,N).
interesting(M) :- target(M), not dscope(M).
interesting(M) :- reactant(M,R), ireaction(R), not dscope(M).
```

Ainsi, une réaction d'intérêt est une réaction produisant un métabolite d'intérêt, c'est à dire :

- Un métabolite cible qui ne peut pas être produit par le draft métabolique
- Un métabolite nécessaire à une réaction d'intérêt et non productible par le draft métabolique.

Ensuite nous nous concentrons sur les réactions qui peuvent être utilisées, c'est à dire celles pour lesquelles tous les réactants sont présents dans l'extension étudiée :

Listing 12:

```
:- xreaction(R), not oreaction(R).
oreaction(R) :- xscope(M) : reactant(M', R), reaction(R,N), not draft(N)
    .
```

Une fois l'ensemble des complétions possibles, on va chercher à réduire au maximum les modifications apportées au réseau en minimisant le nombre de réactions ajoutées au réseau. En ASP cela se définira de la manière suivante :

Listing 13:

```
1 minimize{xreaction(R) : ireaction(R) : not~reaction(R,N)}.
```

Annexe C

Réannotation de gènes

TABLE C.1: Liste des gènes réannotés dans la base de données génomique Orcae après reconstruction du réseau métabolique

Locus	**Numéro EC prédit**	**Prédiction fonctionnelle dans Ectocyc**
Esi0000_0236	EC-2.1.1.163	S-adenosylmethionine :2-demethylmenaquinol methyltransferase
Esi0000_0399	EC-1.1.1	3-oxo-acyl-CoA reductase
Esi0000_0579	EC-5.3.3	2-hydroxyhepta-2,4-diene-1,7-dioate isomerase
Esi0002_0180	EC-4.2.1.46	dTDP-glucose 4,6-dehydratase
Esi0006_0009	EC-5.4.2.3	Phosphoacetylglucosamine mutase (PAGM)
Esi0008_0209	EC-2.6.1.2	Alanine aminotransferase
Esi0009_0077	EC-2.7.1.173	nicotinate riboside kinase
Esi0009_0083	EC-2.7.1.173	nicotinate riboside kinase
Esi0012_0057	EC-3.6.1	dihydroneopterin-PPP pyrophosphohydrolase
Esi0012_0116	EC-3.6.1	dihydroneopterin-PPP pyrophosphohydrolase
Esi0016_0068	EC-2.7.1.173	nicotinate riboside kinase
Esi0016_0192	EC-2.1.1.163	S-adenosylmethionine :2-demethylmenaquinol methyltransferase
Esi0017_0076	EC-2.4.1.67	alpha-D-galactosyl-(1-3)-1D-myo-inositol :raffinose galactosyltransferase
	EC-2.7.1.173	nicotinate riboside kinase
Esi0018_0057	EC-2.7.8.1	ethanolaminephosphotransferase
Esi0021_0019	EC-3.6.1	dihydroneopterin-PPP pyrophosphohydrolase

TABLE C.2: Liste des gènes réannotés dans la base de données génomique Orcae après reconstruction du réseau métabolique (suite)

Locus	**Numéro EC prédit**	**Prédiction fonctionnelle dans Ectocyc**
Esi0031_0020	EC-2.1.2.3	phosphoribosylaminoimidazolecarboxamide formyltransferase
	EC-3.5.4.10	IMP cyclohydrolase
Esi0039_0037	EC-1.6.5.4	Monodehydroascorbate reductase
Esi0039_0059	EC-2.7.1.173	nicotinate riboside kinase
Esi0040_0041	EC-6.2.1.5	succinate—CoA ligase (ADP-forming)
Esi0041_0029	EC-3.6.1	dihydroneopterin-PPP pyrophosphohydrolase
Esi0044_0110	EC-1.3.1.21	7-dehydrocholesterol reductase (=Sterol delta-7 reductase)
Esi0047_0145	EC-4.1.1.22	Histidine decarboxylase
Esi0053_0006	EC-6.1.1.1	Tyrosyl-tRNA synthetase
Esi0059_0045	EC-4.1.3.4	Hydroxymethylglutaryl-CoA lyase
Esi0062_0044	EC-5.1.3	Isomerases acting on Carbohydrates and Derivatives
	EC-5.1.3.18	GDP-mannose 3,5-epimerase
Esi0072_0101	EC-2.7.1.173	nicotinate riboside kinase
Esi0073_0082	EC-1.3.1.70	C-14 sterol reductase (=Delta14-sterol reductase)
Esi0081_0055	EC-3.1.3.67	phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase
Esi0082_0007	EC-5.5.1	Intramolecular lyases
	EC-5.5.1.18	Lycopene epsilon-cyclase
	EC-5.5.1.19	Lycopene β-cyclase
Esi0089_0041	EC-2.1.1.163	Demethylmenaquinone methyltransferase
Esi0091_0014	EC-6.4.1.4	methylcrotonoyl-CoA carboxylase beta chain
Esi0098_0007	EC-2.4.2.7	Adenine phosphoribosyltransferase
Esi0109_0084	EC-4.1.3.36	Naphthoate synthase
Esi0122_0080	EC-1.2.1	pyruvate dehydrogenase
Esi0122_0102	EC-2.1.1.163	Demethylmenaquinone methyltransferase
Esi0143_0028	EC-3.3.2.6	Leukotriene-A(4) hydrolase
Esi0212_0027	EC-1.6.5.4	Monodehydroascorbate reductase
Esi0223_0020	EC-1.3.1.83	geranylgeranyl reductase
	EC-1.3.1.84	acrylyl-CoA reductase (NADPH)

TABLE C.3: Liste des gènes réannotés dans la base de données génomique Orcae après reconstruction du réseau métabolique (fin)

Locus	**Numéro EC prédit**	**Prédiction fonctionnelle dans Ectocyc**
Esi0223_0029	EC-2.4.2.9	Uracil phosphoribosyltransferase
Esi0243_0011	EC-6.1.1.11	Seryl-tRNA synthetase
Esi0346_0014	EC-1.8.7.1	Sulfite reductase
Esi0359_0011	EC-2.7.1.25	Adenylyl-sulfate kinase
Esi0361_0011	EC-2.5.1.9	Riboflavin synthase
Esi0392_0016	EC-2.6.1.62	Adenosylmethionine–8-amino-7-oxononanoate transaminase
Esi0438_0007	EC-2.7.1.35	Pyridoxamine kinase
Esi0673_0003	EC-2.7.1.17	Xylulokinase
Esi0179_0043	EC-2.7.7.80/2.8.1.11	molybdopterin-synthase adenylyltransferase / sulfurtransferase
Esi0147_0053	EC-4.1.99.18	cyclic pyranopterin monophosphate synthase
Esi0031_0067	EC-2.8.1.12	molybdopterin synthase
Esi0000_0519	EC-2.7.7.75/2.10.1.1	molybdopterin adenylyltransferase / hydrolase

Résumé

Durant cette thèse nous nous sommes attachés au développement d'une méthode globale de création de réseaux métaboliques chez des espèces biologiques non classiques pour lesquelles nous possédons peu d'informations. Classiquement cette reconstruction s'articule en trois points : la création d'une ébauche métabolique à partir d'un génome, la complétion du réseau et la vérification du résultat obtenu. Nous nous sommes particulièrement intéressé au problème d'optimisation combinatoire difficile que représente l'étape de complétion du réseau, en utilisant un paradigme de programmation par contraintes pour le résoudre : la programmation par ensemble réponse (ou ASP). Les modifications apportées à une méthode préexistante nous ont permis d'améliorer à la fois le temps de calcul pour résoudre ce problème combinatoire et la qualité de la modélisation. L'ensemble de ce processus de reconstruction de réseau métabolique a été appliqué au modèle des algues brunes, *Ectocarpus siliculosus*, nous permettant ainsi de reconstruire le premier réseau métabolique chez une macro-algue brune. La reconstruction de ce réseau nous a permis d'améliorer notre compréhension du métabolisme de cette espèce et d'améliorer l'annotation de son génome.

Abstract

In this thesis we focused on the development of a comprehensive approach to reconstruct metabolic networks applied to unconventional biological species for which we have little information. Traditionally, this reconstruction is based on three points : the creation of a metabolic draft from a genome, the completion of this draft and the verification of the results. We have been particularly interested in the hard combinatorial optimization problem represented by the gap-filling step. We used Answer Set Programming (or ASP) to solve this combinatorial problem. Changes to an existing method allowed us to improve both the computational time and the quality of modeling. This entire process of metabolic network reconstruction was applied to the model of brown algae, *Ectocarpus siliculosus*, allowing us to reconstruct the first metabolic network of a brown macro-algae. The reconstruction of this network allowed us to improve our understanding of the metabolism of this species and to improve annotation of its genome.