Problèmes d'optimisation combinatoire pour l'étude du métabolisme

Clémence Frioux, Anne Siegel

▶ **To cite this version:**

Clémence Frioux, Anne Siegel. Problèmes d'optimisation combinatoire pour l'étude du métabolisme. Approches symboliques de la modélisation et de l'analyse des systèmes biologiques, ISTE éditions, 2022, Encyclopédie Sciences, 9781789480290. 10.51926/ISTE.9029.ch2 . hal-03885249

HAL Id: hal-03885249
https://inria.hal.science/hal-03885249

Submitted on 5 Dec 2022

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

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

Problèmes d'optimisation combinatoire pour l'étude du métabolisme

Clémence Frioux[1] and Anne Siegel[1]

[1]Univ Rennes, Inria, CNRS, IRISA

La motivation fondamentale de la biologie des systèmes est d'étudier un organisme dans son ensemble plutôt que ses composants isolés, pour comprendre comment des phénotypes complexes émergent des interactions entre ses composants (Kitano, 2002a, 2002b). En pratique, comprendre et modéliser un organisme passe par l'intégration de connaissances et des données multiples et hétérogènes, souvent générées à haut-débit grâce aux technologies dites « omiques » : génomique, transcriptomique, métabolomique, etc. Dans ce chapitre, nous nous focaliserons sur l'information qui peut être extraite des annotations de génomes et de transcrits. Nous expliciterons comment cette information peut-être modélisée à l'échelle du métabolisme pour mieux comprendre la capacité d'un organisme ou d'une communauté d'organismes à réaliser une fonction donnée, en particulier la production de biomasse.

Le métabolisme d'un organisme est représenté sous la forme de réseaux de réactions, supports de nombreuses simulations pour appréhender sa physiologie. Dans la section 1, nous aborderons les notions de dynamique et de fonctionnalité dans les réseaux métaboliques. La section 2 définira les problèmes de complétion associés à ces réseaux. La section 3 démontrera l'applicabilité de ces problèmes à l'étude du métabolisme de communautés d'organismes. Enfin, la section 4 présentera les différentes sémantiques (Booléenne, quasi-stationnaire) qui peuvent être utilisées pour approximer la dynamique d'un réseau métabolique et être exploitées dans les problèmes de complétion de réseaux métaboliques.

1 Dynamique et fonctionnalité d'un réseau métabolique

1.1 Réseaux métaboliques

Le génome encode une grande quantité d'information, dont les séquences correspondants à des protéines appelées enzymes, catalysant des *réactions métaboliques*, c'est-à-dire des transformations de petites molécules (*métabolites* ou *composés*) dans une cellule ou un organisme, catalysées par des protéines appelées *enzymes*. Un exemple classique de réaction est la première étape de la

glycolyse : Glucose + ATP → Glucose-6-phosphate + ADP + H^+. Cette réaction transforme le glucose en glucose-6-phosphate sous l'action de l'enzyme hexokinase. Une cascade de réactions métaboliques, comme la glycolyse qui transforme le glucose en pyruvate, est appelée *voie métabolique*.

Un *réseau métabolique* décrit l'ensemble des réactions associées au métabolisme dans une cellule, un tissu ou un organisme. Ainsi, un réseau métabolique à l'échelle génomique (*GSM*) regroupe toutes les réactions associées à un organisme. Les enzymes qui contrôlent les réactions d'un GSM sont chacune associée à un ou plusieurs gènes du génome de l'organisme considéré. L'essor de l'étude des réseaux métaboliques a donc été porté par l'essor des technologiques de séquençage. Ainsi, le premier GSM publié fut celui de la bactérie *Haemophilus influenzae* (Edwards & Palsson, 1999), obtenu à partir de l'annotation de son génome en 1999. Depuis, des centaines de GSMs ont été construits, pour une grande variété d'organismes (T. Y. Kim et al., 2012).

Formellement, un réseau métabolique peut être décrit par un graphe dirigé biparti $G = (R \cup M, E, stc)$, où R et M sont des ensembles de nœuds correspondants respectivement aux réactions (éventuellement annotées par leurs enzymes) et aux métabolites. Quand $(m, r) \in E$ (respectivement $(r, m) \in E$), avec $m \in M$ et $r \in R$, le métabolite concerné est qualifié de *substrat* (respectivement *produit*) de la réaction r. Les nœuds réactions et métabolites peuvent avoir des arcs entrants et sortants multiples.

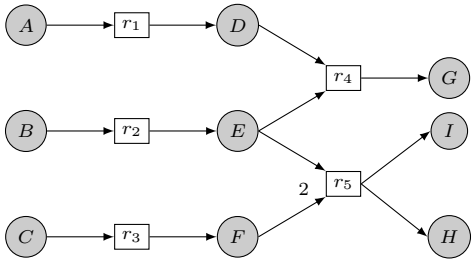
L'étiquetage d'un arc $stc : E \to \mathbb{R}$ indique le *coefficient stoœchiométrique* du composé de la réaction considérée, c'est-à-dire la quantité relative de ce composé qui est impliquée dans la réaction. Ces coefficients sont la base d'une représentation mathématique des réseaux métaboliques sous forme de *matrice stœchiométrique* $\mathbf{S}$ (Maranas & Zomorrodi, 2016), dont les colonnes sont associées aux réactions et les lignes sont associées aux métabolites. Une valeur de la matrice indique la stœchiométrie du composé considéré dans la réaction considérée (0 si le composé n'y participe pas).

De manière générale, la notation G (ou G' etc.) décrira un réseau métabolique, et ses réactions et métabolites seront notés R et M respectivement (ou R', M' etc.). Une autre représentation des réseaux métabolique se fait sous la forme d'hypergraphes (Cottret & Jourdan, 2010) mais elle ne sera pas considérée ici pour faciliter la lecture de ce chapitre.

La figure 1 présente un exemple de réseau métabolique, composé de cinq réactions $R = \{r_1, r_2, r_3, r_4, r_5\}$. La réaction r_4 possède deux substrats : $substrat(r_4) = \{D, E\}$ et un produit $produit(r_4) = \{G\}$. Par défaut, les stœchiométries non indiquées sont de 1. La stœchiométrie associée au métabolite F dans la réaction r_5, $stc(F, r_5) = 2$, signifie que deux molécules de F en plus d'une molécule de E sont nécessaires pour que la réaction r_5 soit catalysée.

1.2 Reconstruction de réseaux métaboliques

Un réseau métabolique est obtenu après un processus de reconstruction visant à capturer l'ensemble des réactions susceptibles d'être catalysées par un organisme, à partir de son information génomique. Dans Thiele et Palsson,



$$
\begin{aligned}
r_1 &: \quad 1A \rightarrow 1D \\
r_2 &: \quad 1B \rightarrow 1E \\
r_3 &: \quad 1C \rightarrow 1F \\
r_4 &: \quad 1D + 1E \rightarrow 1G \\
r_5 &: \quad 1E + 2F \rightarrow 1H + 1I
\end{aligned}
\qquad
\mathbf{S} = \begin{array}{c} \begin{matrix} r_1 & r_2 & r_3 & r_4 & r_5 \end{matrix} \\ \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ +1 & 0 & 0 & -1 & 0 \\ 0 & +1 & 0 & -1 & -1 \\ 0 & 0 & +1 & 0 & -2 \\ 0 & 0 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 & +1 \\ 0 & 0 & 0 & 0 & +1 \end{pmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{matrix}
$$

FIGURE 1 – Réseau métabolique représenté sous forme de graphe biparti. Les cercles représentent des métabolites, les carrés représentent des réactions. Les arcs décrivent les relations de consommation (métabolite vers réaction) et production (réaction vers métabolite). L'étiquetage des arcs (par défaut, 1) décrit la stœchiométrie associée au métabolite associé. Les réactions décrites par le graphe sont les réactions r_1, r_2, r_3, r_4 et r_5. Les coefficients stoechiométriques du réseau sont représentés dans la matrice $\mathbf{S}$, chaque ligne étant associée à la production ou consommation d'un composé fixé, et chaque colonne à la description d'une réaction donnée.

2010, un protocole en 96 étapes est décrit pour obtenir un réseau de qualité, reposant sur une curation manuelle importante du réseau, nécessitant un temps de travail évalué entre plusieurs mois et une année. Ce protocole a été utilisé pour obtenir certains réseaux métaboliques parmi les plus complets, comme ceux de la bactérie *Escherichia coli* (FEIST et al., 2007 ; MONK et al., 2017 ; ORTH et al., 2011 ; REED et al., 2003).

De manière générale, il est possible de séparer le processus de reconstruction en deux grandes étapes, une première non-guidée et automatique, et une seconde guidée et semi-automatique ou manuelle. Ces dernières années, plusieurs plateformes de reconstruction de réseaux métaboliques ont vu le jour pour automatiser cette procédure. Toutes ces plateformes combinent plusieurs méthodes pouvant être classées dans deux catégories : (a) permettre la construction d'une ébauche de réseau, (b) permettre d'améliorer le modèle jusqu'à ce qu'il soit fonctionnel. Ainsi, Pathway Tools (KARP et al., 2016) met à disposition un ensemble d'outils et de méthodes pour construire un réseau métabolique (GSM)

depuis un génome annoté, jusqu'à la simulation de son modèle mathématique. ModelSeed (Devoid et al., 2013 ; Henry et al., 2010) permet la reconstruction automatique de GSMs, et peut être combiné avec le pipeline d'annotation génomique RAST (Aziz et al., 2008), pour les microorganismes et les plantes. Kbase (Arkin et al., 2016) est une plateforme en ligne qui facilite le partage et la collaboration autour de la reconstruction de réseaux métaboliques. Enfin, la RAVEN Toolbox fonctionne sous MATLAB et propose un pipeline complet pour la reconstruction, l'analyse et la visualisation de GSMs (Wang et al., 2018). Au delà de ces plateformes, il existe un grand nombre de méthodes et de boîtes à outils bioinformatiques (Vijayakumar et al., 2017) pour ces applications, telles que la Cobra Toolbox (Heirendt et al., 2019). Elles peuvent être utilisées dans le processus de reconstruction et/ou pour les étapes de curation et/ou pour l'analyse du modèle en résultant (Ebrahim et al., 2013 ; Schellenberger, Que et al., 2011 ; Steffensen et al., 2016).

1.3 De la dynamique d'un réseau métabolique à sa fonctionnalité

La notion de dynamique est indissociable d'un réseau métabolique. Pourtant les réseaux métaboliques en tant que tels ne sont pas des objets dynamiques. Ils présentent simplement l'ensemble des réactions de transformations susceptibles d'être activées par une enzyme dont le gène associé est présent dans le génome d'un organisme.

Une première approche pour décrire la dynamique d'un réseau métabolique consiste à utiliser des modèles mécanistiques de réactions enzymatiques. Désignons par $\mathbf{m}$ le vecteur des concentrations des métabolites M du réseau, et $\mathbf{v}$ le vecteur des vitesses des réactions R du réseau. Une relation de Michaelis-Mentens permet de modéliser le flux $\mathbf{v}_r$ d'une réaction r ayant un composé A pour substrat avec coefficient stoechiométrique 1 par l'équation $\mathbf{v}_r = \frac{v_{max}[A]}{K_m+[A]}$, où v_{max} et K_m sont des constantes spécifiques à la réaction. La dynamique du système est alors décrite par un ensemble d'équations différentielles qui tient compte de la matrice stoechiométrique du système $\mathbf{S}$:

$$\frac{d\mathbf{m}(t)}{dt} = \mathbf{S}\,\mathbf{v}(\mathbf{m}(t)).$$

Simuler une telle dynamique nécessite de connaître l'ensemble des paramètres associés à chaque réaction. L'étude de la cinétique des processus enzymatiques basée sur les équations de Michaelis-Menten a été réalisée pour la bactérie *Escherichi coli* (Mannan et al., 2015). En pratique, utiliser des approches de simulation numérique pour comprendre le fonctionnement d'un réseau métabolique global (GSM) n'est pas envisageable pour la plupart des organismes connus : on ne dispose pas des capacités expérimentales pour identifier les paramètres cinétiques de l'ensemble des enzymes d'un organisme (Mannan et al., 2015).

Pour parvenir à étudier le comportement d'un réseau métabolique en évitant de simuler sa dynamique, la communauté scientifique a introduit différentes ap-

proches, basées sur des hypothèses de stationnarité ou de monotonie, que nous allons décrire dans les différentes parties de ce chapitre. Le principe général commun à ces approches est le suivant : on fait l'hypothèse que le milieu cellulaire (ec.à.d. son environnement) est constitué d'un ensemble de nutriments, appelés graines et notés S. À partir de l'approche utilisée pour modéliser le fonctionnement du système, on prédit qu'un sous-ensemble (réactions et composés) du réseau G (noté de manière générique $active_G(S)$) est activable, à partir du milieu S considéré. Ainsi, on notera $active^s_G(S)$ l'ensemble des réactions activables (resp. composés productibles) à partir des nutriments S sous hypothèse de stationnarité, et $active^t_G(S)$ l'ensemble des réactions (resp. composés) prédits comme étant activables (resp. productibles) à partir des graines, sous hypothèse de monotonie (Frioux et al., 2017). Le lien entre réaction activable et composé productible dans le sous-réseau $active_G(S)$ est intuitif : les composés appartenant aux produits d'une réaction activable sont productibles. Selon les applications, la définition de l'activité du réseau pourra être indifféremment focalisée sur les composés ou les réactions.

Plus précisément, une première abstraction de la fonctionnalité du métabolisme consiste à considérer que, dans un état stationnaire, les composés internes du système ne s'accumulent pas. Cette hypothèse est considérée comme valable sur des périodes de temps relativement long (plusieurs minutes), si on tient compte du fait que les échelles de temps associées aux différents mécanismes cellulaires (signalisation, métabolisme, régulation transcriptionnelle) sont disjointes (Doran, 2013). Sur la base de cette hypothèse, les productions et consommations de chaque métabolite interne sont équilibrées. Le comportement du système est alors contraint par des relations linéaires sur les variables associées aux flux des réactions. L'ensemble des composés productibles et réactions activables $active^s_G(S)$ est calculé en faisant une hypothèse de maximisation de biomasse. Nous reviendrons sur ce modèle en détails dans la section 4.

Dans des phases de croissance, l'hypothèse de stationnarité n'est plus valable, et on peut alors seulement émettre des hypothèses qualitatives sur le comportement du système. On fait alors l'hypothèse que les relations métaboliques sont monotones, c'est-à-dire qu'elles ne peuvent pas être inversées pendant la réponse du système à une perturbation initiale. Ceci permet de modéliser le comportement du système par des règles booléennes, qui activent chaque réaction dès que ses substrats sont disponibles. L'ensemble des métabolites et réactions activés est alors noté $active^t_G(S)$ d'après l'algorithme d'expansion de réseau décrit dans Ebenhöh et al., 2004. Nous reviendrons sur ce modèle en détails dans la section 4.

Ainsi, d'un point de vue dynamique, le passage à l'échelle du métabolisme complet d'une cellule devient possible en faisant des abstractions dans différents cadres formels, qui permettent de modéliser la capacité d'un réseau à produire des composés sous la forme de contraintes et de problèmes d'optimisation. Dans ce contexte, on considère qu'un réseau est *fonctionnel* lorsque les simulations numériques (associées à l'abstraction considérée) permettent de prédire la croissance de l'organisme. Pour cela, une réaction de biomasse (Feist & Palsson, 2010) est généralement ajoutée au modèle ; son contenu en terme de substrats,

produits et leurs stœchiométries respectives varie pour chaque organisme. La précision de la définition de cette fonction est cruciale car son activation est un élément clef pour attester de la qualité du réseau (S. H. Chan et al., 2017). Une difficulté est d'équilibrer les coefficients stœchiométriques des métabolites, dont certains nécessitent des ajustements et constituent des éléments complexes du réseau (Xavier et al., 2017 ; N. Xu et al., 2017). Il existe des réactions de biomasse universelles pouvant être adaptées à chaque espèce (Henry et al., 2010) ; on constate cependant en pratique que les réactions de biomasse fournies dans les réseaux publiés peuvent varier considérablement (Xavier et al., 2017).

2 Comprendre le métabolisme des organismes non-modèles : algorithmes de complétion de métabolisme

2.1 Métabolisme des organismes non modèles

Les organismes dits « modèles » ont été l'objet de nombreuses années d'étude, menant à une maîtrise de ces organismes par et pour les expérimentations, et une compréhension importante de leur physiologie, traduite par la littérature qui lui est associée. C'est soit leur faible complexité biologique, soit la facilité à les utiliser pour des expérimentations qui ont fait de ces organismes des espèces modèles. Parmi elles se trouvent la bactérie *Escherichia coli*, le nématode *Caenorhabditis elegans*, la souris *Mus musculus*, la drosophile *Drosophila melanogaster* ou encore la plante *Arabidopsis thaliana*. Avec l'humain, les organismes modèles ont été les premiers bénéficiaires des techniques dites à haut-débit, en particulier de séquençage. Depuis, l'amélioration des techniques de séquençage et la diminution des coûts qui y sont associés ont permis l'accès aux génomes de milliers d'organismes. Néanmoins, l'étendue des connaissances relatives à ces derniers est moindre par rapport aux organismes modèles. Ainsi, un organisme non-modèle est par définition moins étudié, voire très peu étudié. Sont également inclus dans cette catégorie des organismes difficilement ou non-cultivables, dont les principales (ou uniques) données à disposition les concernant sont les séquences génomiques.

La plupart des organismes séquencés de nos jours sont non-modèles. On note que des groupes taxonomiques sont plus séquencés que d'autres chez les eucaryotes, comme les mammifères (1% des espèces ont leur génome séquencé en 2014) par rapport aux plantes (0.01%) ou aux insectes (0.001%) (Ellegren, 2014). Pour les procaryotes, des discussions actuelles portent sur l'importance et les enjeux liés au développement de nouvelles bactéries modèles (Liu & Deutschbauer, 2018) .

Dans le domaine du métabolisme, la proximité entre un organisme modèle et un organisme non-modèle est exploitée pour prédire des fonctions chez le second à partir des fonctions caractérisées chez le premier, via des recherches de séquences orthologues dans les génomes. L'orthologie repose sur une hypothèse

d'évolution de gènes homologues de deux organismes à partir d'un unique gène chez un ancêtre commun. Des méthodes de reconstruction de réseaux métaboliques se basent ainsi sur l'orthologie en utilisant le métabolisme d'organismes proches de l'espèce d'intérêt (Emms & Kelly, 2015 ; Li et al., 2003 ; Loira et al., 2015). En revanche, la prédiction de fonctions spécifiques à un organisme non-modèle, dite *de novo*, est plus difficile, en raison des difficultés à le cultiver, le modifier génétiquement, ou même l'isoler. En conséquence, les expériences sur les organismes non-modèles sont généralement pratiquées en perturbant globalement leur environnement, là où celles concernant les organismes modèles peuvent être davantage ciblées sur des perturbations génétiques. Cela motive une étude globale du lien entre un phénotype et son environnement et la biologie des systèmes est alors adéquate pour mieux comprendre ces organismes en intégrant de multiples niveaux de connaissance et d'information.

Obtenir un réseau métabolique de qualité pour un organisme est un enjeu important en biologie des systèmes, et particulièrement pour les organismes non-modèles. Comme indiqué dans la Sec. 1.2, des méthodes automatiques existent, mais elles ne peuvent totalement remplacer l'expertise biologique et les modifications manuelles nécessaires aux réseaux métaboliques (Karp et al., 2018). Un réseau métabolique se doit d'être fonctionnel vis-à-vis d'un objectif métabolique défini ; dans de nombreux cas il s'agit de la production de la biomasse (Thiele & Palsson, 2010). La production de biomasse est prédite en utilisant une abstraction sémantique de la fonctionnalité du système (voir Sec. 1.3), qui est souvent basée sur une hypothèse stationnaire modélisée par des contraintes de flux, mais peut être aussi basée sur une hypothèse de monotonie modélisée par des règles booléennes aux niveaux intermédiaires de la reconstruction.

Le réseau métabolique n'est généralement pas fonctionnel après l'intégration des informations issues de l'analyse du génome, et des relations d'orthologie avec les génomes d'autres espèces. On parle souvent d'ébauche de réseau métabolique à cette étape (*draft*). Des réactions peuvent être manquantes, en raison de l'incomplétude des annotations génomiques ou des limites des outils d'annotation. La situation opposée est également possible : les processus de reconstruction ou d'annotation peuvent ajouter des réactions inutiles (faux-positifs). Ces réactions peuvent être isolées dans le réseau, laissant penser à tort qu'il manque d'autres transformations pour les relier.

Dans le cas où le réseau métabolique d'un système ne permet pas de prédire la production de biomasse ou tout autre critère de fonctionnalité, on procède à une étape de *complétion* (ou *gap-filling*) pour identifier des réactions permettant de rétablir la fonctionnalité du système. La complétion est ainsi une part essentielle de la seconde grande étape de reconstruction de réseaux métabolique, semi-automatique et guidée par un objectif métabolique que le modèle doit atteindre. Les incomplétudes du réseau peuvent être identifiées et réparées dans leur ensemble (Satish Kumar et al., 2007 ; Thiele et al., 2014) ou de manière ciblée au sein de voies métaboliques (Prigent et al., 2017). De manière générale, le principe de la complétion est de sélectionner des réactions dans une base de données afin de restaurer la fonctionnalité du modèle vis-à-vis de l'objectif défini. Plusieurs techniques de complétion co-existent, associées à divers

outils bioinformatiques. Un critère généralement partagé par les méthodes est de minimiser le nombre de modifications à faire au modèle, entraînant une minimisation du nombre de réactions à rajouter. Des réactions sont proposées par les algorithmes à partir de la base de données de référence. Une ultime étape est d'assigner aux réactions nouvellement ajoutées un ou plusieurs gènes (Chitale et al., 2016 ; Orth & Palsson, 2010 ; Pan & Reed, 2018 ; Plata et al., 2012).

2.2 Reconstruction du métabolisme d'espèces non-modèles et problèmes de complétion

Plus formellement, résoudre le problème de complétion d'un réseau métabolique ou *gap-filling* consiste à s'assurer qu'un ensemble de réactions objectifs R_T est prédit comme étant activable à partir d'une description du milieu de culture modélisée par un ensemble de composés S appelés *graines*. Comme indiqué précédemment, le concept d'activation dépend de l'abstraction choisie, elle peut-être basée sur des flux, des réseaux booléens ou des contraintes hybrides. La complétion du réseau est réalisée en proposant un algorithme pour sélectionner des réactions additionnelles dans réseau de référence G', une base de données de réactions par exemple (Caspi et al., 2018). En lieu et place de l'utilisation d'un ensemble de réactions objectifs, le focus peut être fait sur la production d'un certain nombre de métabolites cibles. Ces deux possibilités sont équivalentes, puisque s'assurer de la production des substrats d'une réaction objectif revient à s'assurer de l'activation de cette dernière lorsqu'on on fait l'hypothèse que ces différents substrats peuvent être dégradés par l'organisme s'ils sont en surplus. La définition du problème de complétion métabolique peut par ailleurs s'étendre à l'activation de la totalité des réactions du réseau métabolique, c'est-à-dire la production de tous ses métabolites.

Étant donnés une abstraction pour l'activation des réactions du réseau ($active()$), un réseau métabolique $G = (R \cup M, E, stc)$, un ensemble $S \subseteq M$ de métabolites graines décrivant le milieu de culture, un ensemble $R_T \subseteq R$ de réactions cibles, et un réseau de référence $G' = (R' \cup M', E', stc')$, on définit l'*extension du réseau* G par rapport à un ensemble de réactions $R'' \subseteq R'$ comme suit :

$$\begin{aligned}
&\text{extension}(G, G', R'') = ((R \cup R'') \cup (M \cup M''), E \cup E'', stc'') \ , \text{ tel que} \\
&\quad M'' = \{m \in M' \mid \exists\, r \in R'', m \in \text{substrate}(r) \cup \text{product}(r)\} \ , \\
&\quad E'' = E' \cap ((M'' \times R'') \cup (R'' \times M'')), \text{ et} \\
&\quad stc'' = stc \cup stc'.
\end{aligned}$$

Le *problème de complétion métabolique* vise à sélectionner un ensemble de réactions $R'' \subseteq R' \setminus R$ de taille minimale tel que R_T est activable dans l'extension du réseau G à partir de R''. Ainsi, R'' est une *complétion parcimonieuse* de $(R \cup M, E, stc)$ à partir de $(R' \cup M', E', stc')$ vis-à-vis de S et R_T si

$$\begin{aligned}
&\text{completion}(S, R_T, G, G') = \\
&\quad \arg\min_{R'' \subseteq R'} \left(\text{size}\{R'' \subseteq R' |\ R_T \subseteq active_{\text{extension}(G,G',R'')}(S)\}\right).
\end{aligned}$$

L'existence de plusieurs abstractions pour l'activation d'une réaction (quasi-stationnarité, booléenne) se répercute sur les problèmes de complétion de ré-

seaux métaboliques. L'étape de complétion est ainsi fortement dépendante du formalisme utilisé pour modéliser la productibilité dans le réseau. Cependant, dans tous les cas, le problème de complétion de réseaux métaboliques est un problème d'optimisation. Il peut s'agir d'un problème d'optimisation linéaire dans le cas où les contraintes d'activation de réactions considérées sont basées sur les flux. Dans le cas où ces contraintes sont basées sur une relation de monotonie (booléenne), il s'agira de problèmes d'optimisation combinatoire. Différents algorithmes seront décrits dans des sections ultérieures (voir Sec. 4), avec un focus sur les approches combinatoires.

2.3 Valeur ajoutée et limites des problèmes de complétion métabolique : exemple des interactions biotiques

Les algorithmes de complétion permettent d'obtenir des réseaux métaboliques fonctionnels vis-à-vis de l'objectif considéré. Une fois reconstruits, ces réseaux vont pouvoir mettre en lumière le rôle de certaines fonctions, à travers diverses simulations telles que l'effet de KO de gènes sur le métabolisme (Henry et al., 2010). La complétion permet également de découvrir de nouveaux gènes ou de nouvelles fonctions de l'organisme, en analysant le rôle des réactions qui n'avaient pas été identifiées par la reconstruction de l'ébauche de réseau mais qui sont nécessaires à la reproduction des phénotypes attendus (Prigent et al., 2017). Enfin, un réseau complété va permettre d'élaborer des hypothèses plus larges sur son fonctionnement, sa réponse aux perturbations, et sa place dans l'écosystème au milieu duquel l'organisme vit.

Si les approches de complétion ont permis de mieux comprendre le métabolisme de nombreux organismes, elles montrent parfois leurs limites. Il est ainsi important de s'arrêter sur les difficultés liées aux problèmes de complétion métabolique dans leur ensemble.

- Un des risques est d'ajouter des réactions non présentes chez l'organisme (faux-positifs). En effet, il arrive que les réactions suggérées par les opérations de complétion ne puissent pas être associées à une séquence de gène sur le génome des espèces considérées (van der Ark et al., 2017). Cela peut être dû à des limites dans le séquençage puis l'annotation du génome, mais aussi à des prédictions de réactions biaisées par les hypothèses de parcimonie (complétions de taille minimale) réalisées pour explorer efficacement les espaces de recherche de réactions.
- Une seconde difficulté est de choisir de la sémantique d'activation modélisant la fonctionnalité du métabolisme. Ce sont des abstractions, pouvant être plus ou moins adaptées aux spécificités des données traitées et des conditions modélisées.
- Une dernière difficulté est d'identifier correctement l'objectif métabolique à optimiser. Pour les bactéries, il s'agit généralement d'optimiser une production de biomasse dont la composition peut-être évaluée expérimentalement ou d'après la littérature. La fonction de biomasse est plus délicate à définir pour les eucaryotes qui sont notamment susceptibles de ne pas pouvoir vivre individuellement. Et de manière générale, c'est éga-

lement le cas pour la définition de l'objectif métabolique d'un organisme vivant en communauté (microbiote), ou de la communauté elle-même. Il est par exemple possible de combiner plusieurs fonctions de biomasse (Zomorrodi et al., 2014 ; Zomorrodi & Maranas, 2012).

Pour la reconstruction de réseaux métaboliques et pour les problèmes de complétion en particulier, l'enjeu ainsi d'équilibrer l'utilisation d'outils automatiques et l'expertise biologique apportée par la curation manuelle et les vérifications. Les algorithmes de complétion fournissent des hypothèses biologiques sur le fonctionnement de l'organisme mais ces hypothèses doivent être systématiquement rapportées au contexte algorithmique dans lequel elles ont été produites pour en comprendre l'intérêt.

Une illustration des questions biologiques qui peuvent être posées actuellement si on revisite les hypothèses de modélisation, est liée à la notion de symbiote qui fait l'objet d'un intérêt croissant dans la communauté de génomique environnementale. En effet, lorsque le gène associé à une réaction suggérée par une méthode de complétion ne peut pas être identifié dans les données génomiques à disposition, on peut se demander si cette réaction peut être effectivement catalysée par l'organisme. Une explication pour l'absence de gènes est la possible dépendance de ces espèces à d'autres organismes. En effet, la majorité des organismes, et en particulier les eucaryotes, a besoin d'interactions biotiques pour vivre. Un indice dans ce sens est donné par les difficultés souvent rencontrées pour cultiver une espèce non-modèle en laboratoire dans des conditions axéniques (sans micro-organismes) ou de monoculture (Provasoli & Pintner, 1980 ; Tapia et al., 2016). Cela démontre d'une part leurs dépendances vis-à-vis des interactions biotiques, mais complique également les expérimentations pour comprendre ces dépendances. C'est d'ailleurs la forte dépendance métabolique des espèces à d'autres organismes qui peut expliquer les difficultés de culture (Pande & Kost, 2017). Un exemple d'eucaryote difficile à cultiver en conditions axéniques est l'algue brune *Ectocarpus siliculosus*. Tapia et al., 2016 ont montré les dépendances de cet organisme à son microbiote : une culture en conditions antibiotiques entraîne une altération de sa morphologie et de sa physiologie, pouvant être restaurée après inoculation d'isolats bactériens.

D'un point de vue méthodologique, ces exemples complexes invitent à la prudence quant à l'interprétation des résultats de complétion métabolique, et à avoir un esprit très critique vis à vis de la fonction objectif utilisée pour réaliser la complétion, pour ne pas optimiser le modèle vers des fonctions que l'organisme considéré ne peut réaliser lui-même. Plus généralement, les méthodes de reconstruction de réseaux métaboliques apparaissent non seulement intimement liées aux abstractions choisies pour modéliser la fonctionnalité d'un système, mais aussi à la caractérisation de l'espace des réactions possibles pour rendre le métabolisme fonctionnel (base de connaissances, métabolisme d'organismes associés).

3 Métabolisme des microbiotes : vers de nouveaux problèmes d'optimisation

Les organismes non-modèles sont au centre des études des *microbiotes*, en tant qu'hôte ou en tant que symbiote. En effet, en plus d'interactions avec leur environnement abiotique, les organismes interagissent entre eux, formant des communautés qui impactent leur physiologie et les phénotypes. L'intérêt est croissant dans la communauté scientifique pour l'élucidation des mécanismes expliquant ces interactions, en témoigne l'évolution de la littérature traitant de microbiomes et microbiotes ces dix dernières années (214 publications en 2008, plus de 4900 en 2018 [1]). On peut parler d'*écologie des systèmes* pour caractériser le passage à la dimension microbiote de la biologie des systèmes (Muller et al., 2018).

3.1 Génomique des microbiotes

On se réfère à Marchesi et Ravel, 2015 pour introduire les principaux termes liés à ce domaine de recherche. Un *microbiote* est constitué de l'ensemble des micro-organismes dans un environnement. La collection des génomes associés à un microbiote est appelé *métagénome*. Enfin, la combinaison du microbiote, de son métagénome et des conditions environnementales associées est définie comme le *microbiome*. Les organismes appartenant au microbiote d'une espèce (appelée hôte) sont appelés des symbiotes, quelle que soit la nature de leur relation, qui est elle-même appelée symbiose, concept introduit par DeBary en 1879 (Moran, 2006). D'autres termes peuvent être employés comme *holobionte* pour considérer le système composé d'un hôte et ses symbiotes, indépendamment des questions d'évolution souvent associées à ce concept (Webster, 2014). Enfin, ces associations d'organismes sont parfois qualifiées de méta-organismes (Bosch & McFall-Ngai, 2011).

Les différents types d'interactions biotiques pouvant exister entre deux membres d'un microbiote, ou entre un hôte et un symbiote sont résumées par Faust et al., 2012 et Zuñiga et al., 2017. Une symbiose est *mutualiste* quand les deux organismes bénéficient de l'interaction. Si un seul en bénéficie, l'interaction est qualifiée de *commensalisme* ou de *parasitisme*, selon le cas où le second membre n'en tire ni bénéfice ni perte, ou si l'interaction lui est défavorable. Une interaction aux effets négatifs pour les deux membres est une *compétition*. Enfin, le *neutralisme* qualifie une interaction neutre pour les deux partenaires.

L'accès à de grandes quantités de données de microbiomes a induit un changement de paradigme dans l'étude des organismes. Ils ne sont plus considérés comme des individus indépendants mais plutôt comme les membres d'un écosystème complexe, une communauté dans laquelle des interactions ont lieu. Les systèmes hôtes-symbiotes les plus étudiés sont les microbiotes humains, parti-

1. Nombre de publications indexées par Pubmed intégrant "microbiome" ou « microbiota » dans leur titre, obtenu avec la requête suivante ((microbiome[Title] OR microbiota[Title]) AND ("2018/01/01"[Date - Publication] : "2018/12/31"[Date - Publication]))

culièrement intestinal (Gilbert et al., 2018), et celui des plantes, en particulier de la rhizosphère (Hassani et al., 2018). D'autres systèmes sont beaucoup étudiés comme les arthropodes (Gauthier et al., 2015 ; Werren et al., 2008), ou les organismes marins (Blackall et al., 2015 ; Thacker & Freeman, 2012). Étudier les organismes individuellement sans prendre en compte de telles interactions tend à être considéré réductionniste car ils ne vivent pas naturellement isolés de toute interaction biotique (Cavaliere et al., 2017).

La source de données principale pour explorer et comprendre le fonctionnement d'un microbiote est l'information génomique disponible pour ses membres. Les bactéries constituent le principal groupe taxonomique étudié dans un microbiote, mais virus et champignons sont également présents (Baker et al., 2017). Une première approche d'étude d'un microbiote en génomique est basée sur l'identification des espèces en présence. Pour les bactéries, cette approche consiste à séquencer l'ARN ribosomal 16S (métataxonomique) dont le grand degré de conservation entre les espèces sert de support à l'identification d'unités taxonomiques opérationnelles (OTUs). Peu onéreuse, cette technique a néanmoins des limites : la méthode est dirigée (amorces), ne permet généralement pas d'identifier les organismes à l'échelle de l'espèce ou de la souche et ne permet pas d'avoir accès à la fonctionnalité des génomes (Segata, 2018 ; Y. Xu & Zhao, 2018). Or cette dernière est essentielle pour comprendre les interactions entre les organismes.

Le séquençage des génomes en métagénomique permet d'avoir l'information sur « qui fait quoi ? » en plus de « qui est là ? » (Hornung et al., 2018). Ces analyses sont complétées par des données de métatranscriptomique, permettant d'avoir l'information des gènes exprimés dans des conditions définies. La combinaison des deux méthodes est intéressante dans la mesure où la présence d'un gène dans un génome n'implique pas son expression. Cette dernière dépend de nombreux paramètres et de réponses à l'environnement. La difficulté associée à ces méthodes est liée à l'assemblage des séquences et à l'assignement des gènes aux espèces, pouvant conduire à des génomes incomplets d'un part, et à de l'information génomique non assignée d'autre part (Y. Xu & Zhao, 2018). Néanmoins, ces techniques sont indubitablement pertinentes pour étudier les microbiotes dans lesquelles de nombreuses espèces ne sont pas cultivables, constituant ainsi la seule source d'information pour ces organismes.

3.2 Métabolisme des microbiotes : des modèles fusionnés aux modèles compartimentalisés

Le métabolisme est un niveau intéressant pour étudier les microbiotes, en se basant sur les génomes assemblés et annotés des organismes en présence. Reconstruire les réseaux métaboliques des membres d'une communauté est un support pour identifier leurs interactions et expliquer des comportements observés. Parmi les applications de la modélisation du métabolisme d'un microbiote, on peut citer la sélection de communautés (Eng & Borenstein, 2016 ; Frioux, 2018 ; Frioux et al., 2018), le design de communautés synthétiques (Julien-Laferrière et al., 2016), la prédiction et le design de milieux de

culture (Klitgord & Segrè, 2010 ; Zarecki et al., 2014), ou encore l'étude de l'évolution des populations et métabolites au cours du temps (S. H. J. Chan et al., 2017 ; García-Jiménez et al., 2018 ; Zomorrodi et al., 2014). La prédiction de communautés artificielles ou contrôlées remplaçant un microbiote natif est aussi un objectif important pour permettre l'étude d'interactions complexes. Ainsi, des observations faites sur des communautés synthétiques montrent que les interactions sont attendues pour l'échange de ressources métaboliquement coûteuses (Johns et al., 2016).

Des niveaux de modélisation différents sont nécessaires pour ces différents défis (Hanemaaijer et al., 2015 ; Henry et al., 2016 ; Succurro & Ebenhöh, 2018). Un microbiote peut être représenté comme un méta-organisme dans lequel toutes les capacités métaboliques des organismes sont fusionnées, ou par des modèles compartimentalisés, voire des modèles dynamiques. Pour ces derniers, des données additionnelles aux génomes sont nécessaires, notamment l'identification de transporteurs, ou l'incorporation de données expérimentales.

Comme nous allons le voir dans les deux sous-sections suivantes, le niveau de modélisation (compartimentalisé ou non) est lié à la quantité de données disponibles, et à la qualité de ces données (Faria et al., 2016 ; Henry et al., 2016), mais aussi au nombre d'organismes à considérer, allant d'une petite communauté d'une dizaine d'espèces (García-Jiménez et al., 2018) à un ensemble composé de centaines ou quelques milliers d'organismes (Eng & Borenstein, 2016).

3.3 Problème de complétion pour la sélection de communautés dans des microbiotes fusionnés

Un *modèle non compartimentalisé* (ou fusionné) ne prend pas en compte les interactions entre espèces et leur coût énergétique associé : tous les métabolites sont considérés comme un bien partagé de la communauté. Cette modélisation est accessible en se basant uniquement sur les génomes, mais le risque est de surestimer les capacités de la communauté et les mécanismes de coopération. Un exemple d'utilisation de modélisation non compartimentalisée est l'évaluation de l'effet de l'obésité et des maladies inflammatoires chroniques de l'intestin sur les capacités métaboliques du microbiote intestinal par Greenblum et al., 2012. De même, Abubucker et al., 2012 ont proposé un workflow pour explorer le métabolisme d'un microbiote à travers les voies métaboliques en présence ainsi que leur abondance relative.

D'après la définition de Henry et al., 2016, le modèle non-compartimentalisé d'une communauté d'espèce, dit *mixed-bag* ou *gene-soup*, considère un méta-organisme (on parle aussi de *pan-métabolisme*). Les capacités métaboliques des organismes ne sont alors plus examinées individuellement mais dans leur ensemble, dans un unique compartiment abstrait. Formellement, si $G_1, ..., G_N$ sont les réseaux métaboliques d'une communauté d'espèces, le *réseau métabolique non compartimenté de la communauté d'espèces* est le réseau

$$\text{mxdBag}(G_1..G_N) = (\cup M_i \bigcup \cup R_i, \cup E_i, \cup stc_i).$$

En particulier, lorsque deux réactions métaboliques impliquant les mêmes substrats et produits mais avec des stoechiométries différentes apparaissent dans deux réseaux considérés, ces deux réactions différentes sont conservées dans le réseau non compartimenté.

Le problème de complétion de réseaux métaboliques était jusqu'alors considéré individuellement par organisme en biologie des systèmes. Il est possible de revisiter le problème en écologie des système par une approche de sélection de communauté. Ce nouveau problème de complétion consiste à sélectionner des organismes (en proposant ou non des interactions entre eux) susceptibles de coopérer à l'échelle métabolique pour effectuer collectivement un ensemble de fonctions, ou permettre à un organisme hôte de le faire. Cette sélection de communautés peut se faire aux différents niveaux de modélisation préalablement décrits : en considérant une compartimentalisation de la communauté ou un méta-organisme.

Sélectionner une communauté dans une modélisation non-compartimentalisée implique ainsi de sélectionner un nombre minimal d'organismes tels qu'ils peuvent répondre collectivement à un objectif métabolique, sans considérer les réactions de transports ou échanges nécessaires biologiquement. L'espace de recherche est l'ensemble des 2^N sous-familles de la communauté initiale considérée $\{G_{i_1}, ..., G_{i_L}\} \subset \{G_1, ..., G_N\}$. Les organismes sont sélectionnés pour être ajoutés à un système vide (ou composé uniquement d'un organisme hôte) tel qu'un maximum de réactions (respectivement métabolites) cibles soit activés (respectivement accessibles). Contrairement à la complétion pour un réseau métabolique individuel, ce n'est pas le nombre de réactions qui est minimisé mais le nombre d'organismes auxquels appartiennent les réactions intéressantes pour la réalisation de l'objectif métabolique ciblé.

Plus formellement, une communauté optimale en modélisation non-compartimentalisée, pour la production d'un ensemble de réactions $R_T \subset \cup R_i$ à partir de graines $S \subset \cup M_i$, est obtenue par résolution du problème d'optimisation suivant :

$$\begin{aligned}&\text{mxdbagCnity}(S, R_T, G_1..G_N) = \\ &\quad \underset{\substack{\{G_{i_1}..G_{i_L}\} \\ \subset \{G_1..G_N\}}}{\arg\min} \quad \left(\text{size}\left\{\{G_1, .., G_N\} | \, R_T \subseteq \mathit{active}_{\text{mxdBag}(G_{i_1}..G_{i_L})}(S)\right\}\right).\end{aligned}$$

Dans ce problème, l'objectif est d'assurer l'activation d'un ensemble de réactions $R_T \in R$ du réseau métabolique à partir des graines S. Une variante du problème est de s'assurer de l'activité des réactions cibles au sein d'un réseau particulier, par exemple celui de l'hôte associé à un microbiote. Dans ce cas, la condition d'activité devient $R_T \subseteq \mathit{active}_{G_k}(S)$, k étant l'indice associé à cet organisme d'intérêt. Une autre alternative, assez fréquente quand le système considéré est un ensemble de réseaux métaboliques en interaction, est de s'intéresser à la production de composés cibles T plutôt qu'à l'activation de réactions. Dans ce cas là, on change la condition devient $T \subseteq \mathit{active}_{G_k}(S)$ en faisant l'hypothèse qu'un composé est productible s'il appartient aux produits d'au minimum une réaction activable.

Le formalisme non-compartimentalisé étant sujet à peu de contraintes en raison de l'absence de transports ou échanges entre les réseaux, il peut constituer une première étape pour réduire la complexité d'un microbiote et en proposer des membres d'intérêt. Eng et Borenstein, 2016 ont implémenté l'algorithme en programmation linéaire en nombres entiers CoMiDA pour résoudre ce problème avec une sémantique de flux. Cet algorithme a été testé sur 10,000 paires aléatoires de substrats-produits pour identifier des communautés minimales pouvant transformer le substrat en son produit. Néanmoins, pour considérer la sélection de communautés non-compartimentalisées comme une première étape de l'analyse d'un microbiote, il convient d'explorer tout l'espace des solutions du problème d'optimisation et ainsi de pouvoir proposer toutes les solutions optimales. C'est la filtration de ces solutions optimales avec des contraintes supplémentaires qui permettra de combiner plusieurs étapes d'analyses sans perte d'information. Pour appuyer ce point de vue, dans (Frioux et al., 2018), les autrices ont montré que dans l'ensemble de données de Eng et Borenstein, 2016, 86.5% des paires source-produit reliées par un chemin activable dans le réseau non-compartimenté sont en fait activables par plus de 100 communautés optimales différentes, 49.8% plus de 1,000.

3.4 Problème de complétion pour la sélection de communautés compartimentalisées avec échanges minimaux

La modélisation non-compartimentalisée peut être considérée comme une méthode pour étudier globalement la complémentarité métabolique entre les membres d'une communauté. Une limitation intrinsèque est la non prise en compte du coût énergétique représenté par les interactions métaboliques biologiquement. C'est la raison pour laquelle rétablir un compartiment pour chaque espèce est pertinent comme seconde approche pour modéliser le métabolisme associé à un microbiote en écologie des systèmes.

Pour définir formellement un réseau dans ce contexte, on doit introduire l'ensemble des échanges de molécules possibles entre les réseaux métaboliques associés à chaque organisme. Formellement, on note

$$\text{exchg}(G_1..G_N) = \{(r_m, i, j) | m \in M_i \cap M_j,\, i \neq j\}$$

l'ensemble des réactions permettant l'échange de chaque composé m entre la paire de réseaux métaboliques G_i et G_j dès lors que ces deux réseaux contiennent m.

A l'aide de ces échanges, on peut définir un réseau compartimentalisé. On fixe d'abord un ensemble de réactions d'échanges $\mathcal{E} \subset \text{exchg}(G_1..G_N)$ autorisées. Le réseau métabolique compartimentalisé associé à la communauté $G_1..G_N$ et à la famille d'échanges $\mathcal{E}$ est alors obtenu en associant à chaque métabolite ou réaction l'indice correspond à l'organisme auquel il appartient : les métabolites sont dénotés par (m, i), les réactions par (r, j) et les arcs par (e, i, j) ; i et j étant les indices identifiants les organismes. Les arcs des réseaux métaboliques de la communauté incluent les arcs représentant la consommation et la production

de métabolites dans les modèles, ainsi que les réactions d'échanges pour les métabolites sélectionnés. Formellement, le réseau métabolique compartimenté est donc un graphe métabolique un peu spécifique.

$$\text{cptModel}(G_1..G_N, \mathcal{E}) = (\overline{M}, \overline{R}, \overline{E}, \overline{stc})$$

$$\text{avec} \left| \begin{array}{rcl} \overline{M} & = & \cup_{i=1..N}(M_i \times \{i\}) \\ \overline{R} & = & \mathcal{E} \cup_{i=1..N} (R_i \times \{i\}) \\ \overline{E} & = & \{[(m,i),(e_m,i,j)] \,|\, (e_m,i,j) \in \mathcal{E}\} \\ & & \bigcup \cup_{m,i,j}\{[(e_m,i,j),(m,j)] \,|\, (e_m,i,j) \in \mathcal{E}\} \\ & & \bigcup \cup_{i=1..N}\{[(m,i),(r,i)] \,|\, (m,r) \in E_i\} \\ & & \bigcup \cup_{i=1..N}\{[(r,i),(m,i)] | (r,i) \in E_i\}. \end{array} \right.$$

La fonction de stoechiométrie $\overline{stc}$ est définie de manière naturelle en reprenant les valeurs des fonctions stoechiométriques sur les réseaux originaux et en mettant une valeur égale à 1 pour les imports de composés.

Cette définition demande aussi de reformuler la définition d'un milieu de culture (graines). Étant donné un ensemble de métabolites graines S pour toute la communauté, on considère que chaque graine de S a la possibilité d'être importée dans l'organisme considéré en définissant un nouvel ensemble de graines compartimentalisé : $\text{cptSeed}(G_1..G_N, S) = \cup_{1..N}((S \cap M_i) \times \{i\})$. Cela modélise le fait que tous les organismes ne possèdent pas chaque graine dans leur réseau métabolique (en tant que substrat ou produit d'au moins une de leurs réactions).

A l'aide de ces concepts, il devient possible de formuler un problème de sélection de communauté prenant en compte les coûts énergétiques représentés par l'échange de métabolites (import par le receveur, export pour le fournisseur). Il s'agit alors de minimiser le nombre de ces interactions dans la sélection de communautés optimales. Cela crée une nouvelle hypothèse de parcimonie pour le problème de complétion. Le problème général de complétion par sélection de communautés dans un modèle compartimentalisé peut être défini de la manière suivante.

$$\text{cptCnity}(S, R_T, G_1..G_N) =$$
$$\underset{\mathcal{E} \subset \text{exchg}(G_1..G_N)}{\arg\min} \text{weight}\{\mathcal{E} \subset \text{exchg}(G_1..G_N) | \forall r_T \in R_T,\, \exists k \leq N,$$
$$(r_T, i) \in \mathit{active}_{\text{cptModel}(G_1..G_N, \mathcal{E})}(\text{cptSeed}(G_1..G_N, S))\}.$$

Plus précisément, on recherche l'ensemble d'échanges de poids minimal qui assure que toute cible peut être produite dans au moins un compartiment à partir du milieu de culture. Le poids peut être une somme pondérée de scores associés à chaque échange, ou simplement le nombre d'échanges utilisés lorsqu'ils sont tous équiprobables.

La figure 2 illustre l'effet du choix de modélisation de la compartimentalisation pour une sélection de réseaux métaboliques dans une communauté. Cette figure représente trois métabolismes de symbiotes (contours en pointillés) et un métabolisme d'un hôte (pointillés jaunes) pour lequel la réaction cible (à activer)

est r_{2_h} (en vert). Si on avait opté pour une formulation basée sur la production de métabolites, l'objectif aurait été de produire l'ensemble des substrats de la réaction r_{2_h}, c'est-à-dire le métabolite H. Une sélection non-compartimentalisée va proposer trois communautés optimales chacune composée de l'hôte et d'un symbiote : $\{\mathcal{H}, \mathcal{S}_1\}$, $\{\mathcal{H}, \mathcal{S}_2\}$, et $\{\mathcal{H}, \mathcal{S}_3\}$. L'hôte $\mathcal{H}$ est présent dans toutes les communautés car il s'agit de l'hôte qui possède la fonction objectif. Les trois symbiotes peuvent être indépendamment sélectionnés dans cette modélisation car ils possèdent tous une réaction r_3 pouvant produire H et assurer la production de H à partir des substrats $\{A, B\}$ dans un réseau fusionné avec celui de l'hôte. La modélisation compartimentalisée va viser à proposer des échanges pour chacune des trois solutions et va les discriminer. En effet, si les trois symbiotes possèdent une réaction r_3, seul un, $\mathcal{S}_1$, peut activer seul cette réactions, à partir des graines. Les réseaux $\mathcal{S}_2$ et $\mathcal{S}_3$ ont respectivement besoin d'importer A ou $\{A, B\}$ pour activer cette réaction. Cela implique la communauté $\{\mathcal{H}, \mathcal{S}_1)\}$ peut être fonctionnelle avec un seul échange (du composé H) entre le symbiote et l'hôte. Pour les autres communautés il faudrait un échange de A ou deux échanges de A et B supplémentaires. Ainsi, la communauté de coût minimal est $\{\mathcal{H}, \mathcal{S}_1\}$.

Cette modélisation est ainsi plus précise mais également plus coûteuse en terme de calcul. Avec peu d'information sur la nature précise des composés transportés dans et depuis les espèces comme c'est le cas pour la majorité des organismes non modèles, la taille de l'espace de recherche des échanges à considérer est $4^{\sum_{i<j} \text{size}(M_i \cap M_j)}$. Pour résoudre ce problème d'optimisation, Julien-Laferrière et al., 2016 ont proposé un algorithme permettant de générer des communautés synthétiques basée sur la sélection d'espèces bactériennes, de réactions exogènes, et de réactions de transport. L'approche consiste à modéliser le problème par des calculs de plus courts chemins entre des nœuds fixés dans un hypergraphe. Notons que le fait de considérer des hypergraphes dirigés est bien plus complexe que d'étudier les plus courts chemins dans un graphe. En effet, il faut s'assurer que chaque nœud entrant d'un arc tentaculaire (c.à.d. les arcs ayant plusieurs entrées), est activable depuis au moins une source présente dans le milieu de culture. Déterminer un plus court chemin dans un hypergraphe est réalisable en temps exponentiel du nombre de réactions avec plusieurs substrats dans le graphe : il faut réaliser une récursivité sur le graphe pour chaque nœud entrant de chaque arc tentaculaire afin de savoir s'il est activable, ce qui rend le problème NP-difficile. L'approche implémentée résout le problème en particulier lorsque le nombre de bactéries dans la communauté globale est petit, en identifiant les échanges minimaux nécessaires à la production des composés cibles.

Cependant, la combinatoire des échanges et des sous-familles de réseaux rend la recherche directe d'échanges minimaux peu envisageable pour des microbiomes comportant de nombreuses espèces, particulièrement si leurs réseaux métaboliques contiennent un grand nombre de composés partagés. C'est le cas pour des réseaux reconstruits automatiquement à partir de données de métagénomiques, qui ne peuvent être tous individuellement curés et améliorés. Pour

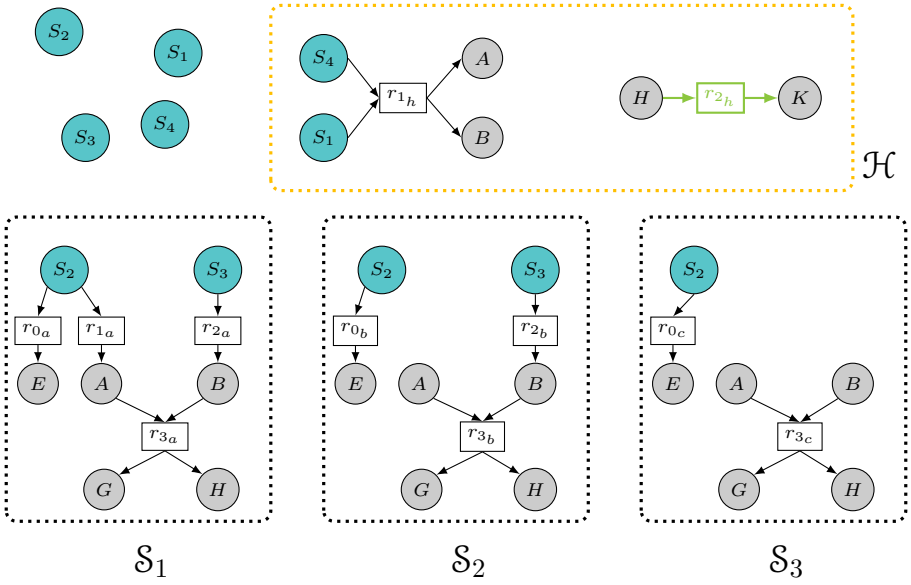


FIGURE 2 – Exemple de sélection de communautés. Figure inspirée de FRIOUX et al., 2018. Cet exemple présente un réseau métabolique d'hôte ($\mathcal{H}$) et trois réseaux métaboliques de symbiotes potentiels ($\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$). L'objectif de la sélection de communautés est d'activer la réaction verte chez l'hôte, soit en produire tous ses subtrats, c'est-à-dire uniquement le métabolite H ici. Une sélection non compartimentalisée *mixed-bag* proposera trois solutions optimale de communautés (l'hôte associé à chacun des symbiotes indépendamment). Une sélection tenant compte des échanges entre compartiments proposera une unique solution optimale $\{\mathcal{H}, \mathcal{S}_1\}$.

résoudre ces problématiques, FRIOUX et al., 2018 ont introduit une heuristique de résolution du problème de sélection de communautés compartimentalisées minimisant les coûts d'interaction. Cela implique de filtrer les communautés de tailles minimales préalablement déterminées, avec des critères supplémentaires basés sur le nombre de composés échangés nécessaires. L'ensemble est modélisé par deux optimisations : minimiser la taille de la communauté non compartimentalisée (comme vu précédemment) puis minimiser les échanges nécessaires pour filtrer ces communautés dans un modèle compartimentalisé. Le problème qui est résolu est alors le suivant :

$$\text{cptCnity}(S, R_T, G_1..G_N) = \left| \begin{array}{l} \underset{\substack{\{G_{i_1}..G_{i_L}\} \subset \{G_1..G_N\} \\ \mathcal{E} \subset \text{exchg}(G_{i_1}..G_{i_L})}}{\arg\min} \\ \end{array} \right| \begin{array}{l} \text{size}\{\, G_{i_1}..G_{i_L} \,\}, \\ \text{size}\{\, \mathcal{E} \subset \text{exchg}(G_{i_1}..G_{i_L}) \,|\forall r_T \in R_T,\, \exists k \leq L, \\ \qquad (r_T, i_k) \in \mathit{active}_{\text{cptModel}(G_{i_1}..G_{i_L}, \mathcal{E})}(\\ \qquad\qquad \text{cptSeed}(G_{i_1}..G_{i_L}, S)). \end{array}$$

Ainsi, on recherche les communautés minimales susceptibles de produire l'ensemble les cibles. Au sein de ces communautés, on identifie les échanges qui sont en nombre minimal pour réaliser cette production. De manière simple, le nombre d'échange peut-être remplacé par un poids associé à chaque échange. Il s'agit ainsi d'une surapproximation du problème précédent mais qui permet de réduire la recherche à deux sous-problèmes solvables en pratique.

4 Sémantique topologique de modélisation : approximation booléenne du fonctionnement du métabolisme

Les sections précédentes ont présenté l'importance des problèmes de complétion, et leurs spécificités en biologie des systèmes et en écologie des systèmes. Leurs descriptions ont été faites dans un cadre général, en utilisant le concept générique d'activation $active_G(S)$ d'un réseau métabolique G pour décrire sa fonctionnalité. Cette section vise à décrire les sémantiques d'activation du métabolisme pouvant être utilisées dans les différents problèmes de complétion présentés dans les sections précédentes pour approcher le comportement du système avec une abstraction adéquate. Deux sémantiques principales d'activation des réactions et des métabolites existent. La première sémantique d'activation, introduite par Ebenhöh et al., 2004 et Handorf et al., 2005, est basée sur un parcours récursif de la structure topologique du graphe métabolique. Les paramètres numériques tels que la stœchiométrie, la cinétique enzymatique ou les flux associés aux réactions sont ignorés au profit de la topologie du graphe et de la connectivité entre ses nœuds (Manor et al., 2014). Cela revient à étudier les points fixes de réseaux booléens monotones.

La seconde sémantique est quantitative et basée sur l'association de flux aux réactions. La distribution ou la non distribution de flux dans les différentes réactions du réseau est la base de la définition de l'accessibilité des réactions associée à cette sémantique. Elle sera détaillée et étudiée dans une section ultérieure (5). Il a été montré dans plusieurs revues que ces deux sémantiques ne sont pas équivalentes, bien que proposant des résultats proches (Kruse & Ebenhöh, 2008). Une illustration concrète de l'impact de ces sémantiques sur le problème de complétion de réseau métabolique, abordé dans la Sec. 2.2, est donnée dans Prigent et al., 2017 où une comparaison sur 10,800 réseaux simulés est réalisée.

4.1 Accessibilité topologique des composés et réactions dans un réseau métabolique

Considérons un réseau métabolique $G = \{R \cup M, E, stc\}$. L'information stœchiomtérique n'est en fait pas nécessaire pour cette sémantique, mais nous la conservons ici par soucis d'homogénéité. Comme expliqué dans les sections précédentes, nous cherchons à définir l'ensemble des composés et réactions activables à partir des métabolites disponibles dans le milieu de culture, notés

$S \subseteq M$, que nous avons appelé graines. Ces graines peuvent être définies à partir de l'expertise d'un biologiste, ou identifiées à l'aide d'algorithmes qui se basent sur les composantes fortement connexes du graphe (Borenstein et al., 2008 ; Carr & Borenstein, 2012).

Le sous-réseau activable dans un réseau métabolique G à partir d'un ensemble de graines S pour la sémantique topologique, est noté $\text{scope}(G, S)$. Il contient l'ensemble des métabolites productibles, appelé *périmètre métabolique du réseau* (ou *scope*). On peut le définir itérativement comme suit :

$$\text{scope}(G, S) = \bigcup_{i \geq 0} \Sigma_G^i(S)$$
$$\text{avec} \left| \begin{array}{rcl} \Sigma_G^0(S) & = & S \\ \Sigma_G^{i+1}(S) & = & \Sigma_G^i(S) \cup \{\text{product}(r \in R \,|\, \text{substrate}(r) \subset \Sigma_G^i(S))\} \end{array} \right.$$

En partant des graines, le périmètre du réseau est donc calculé itérativement jusqu'à ce qu'il atteigne un ensemble stable de composés. Ainsi, un composé appartient au périmètre métabolique des graines d'un réseau s'il appartient à l'ensemble de graines, ou s'il est un produit d'une réaction dont les substrats sont eux-mêmes dans le périmètre (c'est-à-dire activée). Il est important de noter la conjonction implicite dans la récursivité de cette définition : tous les substrats d'une réaction doivent être productibles pour que ses produits le soient. La figure 3 illustre cette propriété : G est activable et H ne l'est pas.

Cette notion de périmètre métabolique peut être revisitée d'un point de vue dynamique. A chaque composé métabolique $m \in M$, associons une variable booléenne $\overline{m} \in \{0, 1\}$. On construit une fonction booléenne monotone qui décrit la dynamique du réseau métabolique comme suit :

$$\phi(\overline{m}) = \bigvee_{r \in R | m \in \text{product}(r)} \bigwedge_{s \in \text{substrate}(r)} \overline{s}.$$

Dans ce contexte, le périmètre métabolique du réseau est le plus petit point fixe de la fonction ϕ pour lequel toutes les variables associées aux graines S sont égales à 1.

Plus généralement, on définit le concept de réactions activables topologiquement en vérifiant que tous ses substrats sont dans le périmètre des graines. Considérant une réaction objectif $r_{obj} \in R$, un réseau métabolique $G = (R \cup M, E, stc)$ et un ensemble de graines S, la réaction r_{obj} est activable si $\text{substrate}(r_{obj}) \subseteq \text{scope}(G, S)$. L'ensemble des réactions activables dans un réseau métabolique G à partir d'un ensemble de graines S pour la sémantique topologique, est noté $active_G^t(S)$.

Les périmètres métaboliques de communautés de réseaux peuvent aussi être facilement définis à partir de ce concept. Dans une communauté non compartimentalisée, comme étudiée à la Sec. 3.3, le périmètre d'un ensemble de graines est défini par :

$$\text{mxdbagScope}(G_1..G_N, S) = \text{scope}(\text{mxdBag}(G_1..G_N), S).$$

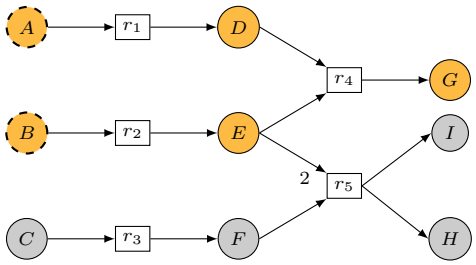


FIGURE 3 – Scope de graines dans un réseau métabolique. Les composés en pointillés sont les graines $S = \{A, B\}$. Les composés jaunes sont dans le scope de ces graines. En particulier, H n'est pas dans le scope car un des substrats de r_5 n'est pas dans le scope.

Par extension, un métabolite m appartient au périmètre d'une communauté compartimentalisée (voir Sec. 3.4) s'il existe un réseau métabolique G_k de la communauté contenant m dans son scope :

$$\begin{aligned}\text{cptScope}(G_1..G_N, \mathcal{E}, S) = \{m \in \cup_{1..N} M_i | \exists k,\\ (m,k) \in \text{scope}(\text{cptSeed}(G_1..G_N, S),\\ \text{cptModel}(G_1..G_N, \mathcal{E}))\}.\end{aligned}$$

Les notions de réactions activables dans des communautés d'organismes compartimentées ou non-compartimentées se dérivent naturellement.

4.2 Activation et cycles

L'activation topologique d'un réseau métabolique peut être interprétée comme une modélisation de la croissance d'un organisme à partir d'un état initial défini par l'activation des composés graines, par opposition à l'état stationnaire. En effet, la définition du périmètre métabolique induit que l'activation de tout composé doit être prouvée par un ensemble de chemins provenant des nutriments que représentent les graines. Il y a donc un lien direct entre le périmètre du réseau métabolique et la capacité du milieu à nourrir le système.

Il faut noter cependant que cette définition empêche la cellule d'être autoalimentée par les cycles présents dans le réseau. Un tel cas est illustré dans la figure 4. Dans cet exemple, l'ensemble des graines est composé des métabolites $\{S_1, S_2, S_3\}$ et permet d'activer les réactions $\{r_0, r_1, r_6\}$, ce qui conduit au périmètre suivant : $\{S_1, S_2, S_3, A, B, G, H\}$. Ce réseau comporte un cycle, composé des réactions $\{r_2, r_3, r_4\}$. En appliquant la définition de la sémantique topologique, le cycle ne peut pas être initié car la réaction r_2 nécessite C en plus de S_3 pour être activée, et le premier dépend lui-même de r_2 pour être produit. L'application de la définition du périmètre induit donc qu'aucun des composants du cycle ne sont activés.

D'un point de vue biologique, on peut cependant noter que, une fois initié, et si S_3 est disponible de manière illimitée, le cycle perdurera dans son activité

car la production de plusieurs molécules de C permet d'alimenter le cycle en plus de la réaction r_5. Le point crucial ici est donc l'initiation du cycle. La modélisation sous-jacente au concept de périmètre métabolique considère que, au vue de l'état initial du système décrit par les graines S, le cycle ne sera jamais initié, à moins qu'une réaction soit manquante dans le système, par exemple une réaction produisant D ou E. Cela est possible car l'annotation des génomes n'est pas infaillible et les génomes eux-mêmes sont de qualité variable, notamment pour les organismes non-modèles. Ainsi, il serait pertinent de chercher à rajouter une telle réaction au réseau, ce qui solutionnerait l'initiation du cycle. C'est l'objet de la complétion de réseaux métaboliques, qui peut être basée sur la sémantique topologique (voir la sous-section suivante).

Une alternative est de considérer que le cycle peut être initié indépendamment de l'état décrit par l'ensemble de graines S. Dans ce cas, une solution est de considérer un état initial étendu, dans lequel certains métabolites sont considérés comme étant déjà productibles, car cette production est connue par la physiologie de l'organisme. C'est souvent le cas pour une certaine classe de métabolites appelés cofacteurs. Ce sont généralement des petites molécules participants à de nombreuses réactions et dont les équilibres reposent sur des cycles. Par exemple, l'ATP, le NAD ou la coenzyme A appartiennent à cette catégorie. Ces molécules peuvent être alors rajoutées dans les graines, ces dernières pouvant alors être distinguées selon qu'elles sont des graines d'initiation (artefactuelles), ou des nutriments. Dans la figure 4, le déblocage du cycle pourrait passer par l'ajout de D ou E aux graines après étude de leur rôle biologique (Christian et al., 2009 ; Handorf & Ebenhoh, 2007). Alternativement, ces composés sont parfois ignorés. Eng et Borenstein, 2016 ont exclu des réactions les métabolites identifiés par Greenblum et al., 2012 et Cottret et al., 2010 ; Julien-Laferrière et al., 2016 ont également ignoré de tels composés.

Une dernière solution est de considérer une sémantique topologique non stringente autorisant le recyclage de composés dans les cycles. Le critère de récursivité du calcul du périmètre est conservé ; la condition d'initiation ne se fait plus seulement avec les graines mais considère les cycles. Il s'agit alors d'autoriser l'activation de tous les composés métaboliques qui prennent la valeur 1 dans au moins un point fixe du réseau booléen associé au graphe métabolique lors de l'activation S, et non plus seulement le point fixe avec le plus petit nombre de valeur 1. Cette nouvelle définition du périmètre a été largement étudiée dans (Acuna et al., 2012 ; Acuña et al., 2009 ; Andrade et al., 2016 ; Cottret et al., 2008), elle s'avère intéressante pour identifier des graines métaboliques pertinentes associées à un réseau.

4.3 Applications

Appliquée à la reconstruction de réseaux métaboliques, Collet et al., 2013 ont complété le réseau métabolique d'*Ectocarpus siliculosus* avec une technique implémentée dans l'outil Meneco (Prigent et al., 2017) qui s'appuie sur une sémantique topologique. Dans le contexte d'études des communautés, l'étude des graphes est non seulement appliquée aux études de cooccurence ou à la modélisa-

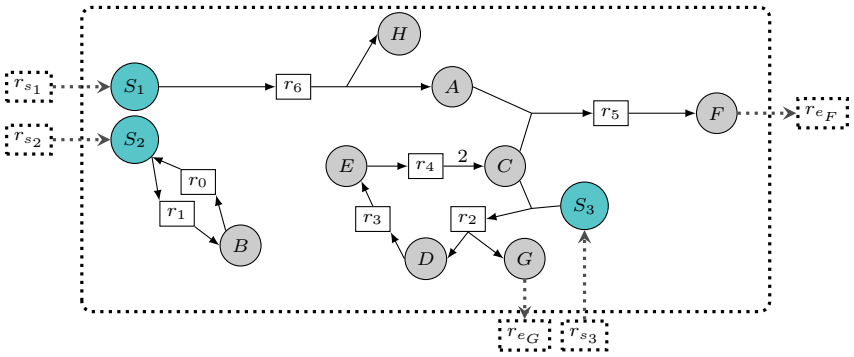


FIGURE 4 – Effet des cycles sur l'activation des composés métaboliques sous la sémantique topologique associée à une abstraction booléenne de la dynamique du réseau métabolique. Les nœuds bleus sont les graines du réseau. Le cycle composé des métabolites $\{S_3, C, D, E, G\}$ et des réactions $\{r_2, r_3, r_4\}$ n'est pas activable car S_3 a besoin d'être combiné avec C, lui-même un produit du cycle, pour activer la réaction r_2.

tion de méta-organismes (HEINKEN et al., 2016), mais elle s'applique également aux études d'interactions dans des modèles compartimentalisés (FRIOUX et al., 2018 ; JULIEN-LAFERRIÈRE et al., 2016) via la sémantique topologique.

Plus précisément, un exemple d'application de la sémantique topologique s'appuyant sur une dynamique booléenne est en lien avec le problème de complétion de réseaux métaboliques individuels présenté à la section 2.2. Différents travaux méthodologiques (COLLET et al., 2013 ; SCHAUB & THIELE, 2009) ont permis de proposer une approche de résolution du problème de complétion suivant une sémantique d'activation topologique à l'aide de programmation logique, plus précisément des technologies de programmation par ensembles-réponses, que nous détaillerons dans la section 7. Pour parvenir à étudier des réseaux dégradés, typiquement ceux associés à des espèces non modèles, le problème de complétion est relaxé pour identifier des complétions de taille minimale telles qu'un maximum de composés cibles sont accessibles. Le problème résolu est donc le suivant :

$$\text{completion}^t(S, T, G, G') = \underset{\{r_1..r_m\} \subseteq G'}{\arg\min} \left| \begin{array}{l} \text{size}(T \setminus \text{scope}(G \cup \{r_1..r_m\}, S) \\ \text{size}(r_1..r_m). \end{array} \right.$$

Notons que Meneco fait le focus sur la productibilité des métabolites plutôt que l'activation des réactions. Néanmoins, il peut s'utiliser indifféremment dans les deux cas, en choisissant pour cibles les substrats des réactions à activer. Le programme en programmation par ensembles-réponses (voir la sec. 7) qui permet de résoudre ce problème a été encapsulé dans un package python appelé Meneco, qui a été testé sur un ensemble de 10,800 réseaux simulés. Cette étude a montré que, en moyenne, une trentaine de réactions suffisent pour restaurer la fonctionnalité d'un réseau métabolique pour la sémantique topologique. Pour

des réseaux faiblement dégradés, une dizaine de réactions sont suffisantes en moyenne, et ces complétions satisfont également la sémantique de flux dans 73% des cas (Prigent et al., 2017).

En suivant une approche similaire de relaxation des contraintes liées aux cibles à produire, Frioux et al., 2018 ont implémenté la sélection de communautés d'organismes présentée dans la section 3.3, à l'aide de problèmes d'optimisation combinatoire, sous une sémantique d'activation topologique. Cette approche permet de résoudre le problème d'optimisation suivant pour la modélisation non-compartimentalisée :

$$\text{mxdbagCnity}^t(S, T, G_1..G_N) = \underset{\substack{\{G_{i_1}..G_{i_L}\} \\ \subset \{G_1..G_N\}}}{\arg\min} \left| \begin{array}{l} \text{size}\left(T \setminus \text{mxdbagScope}(G_{i_1}..G_{i_L}, S)\right) \\ \text{size}\{\, G_{i_1}..G_{i_L} \,\}\,. \end{array} \right.$$

De manière plus générale, le problème de sélection de communauté dans un cadre compartimentalisé (Sec. 3.4) est résolu avec une sémantique d'activation topologique en autorisant une relaxation de la contrainte sur le fait de produire l'ensemble des cibles du système. Il s'agit donc de résoudre le problème suivant.

$$\text{cptCnity}^t(S, T, G_1..G_N) = \underset{\substack{\{G_{i_1}..G_{i_L}\} \\ \subset \{G_1..G_N\}}}{\arg\min} \left| \begin{array}{l} \text{size}\left(T \setminus \text{mxdbagScope}(G_{i_1}..G_{i_L}, S)\right), \\ \text{size}\{\, G_{i_1}..G_{i_L} \,\}, \\ \text{size}\{\, \mathcal{E} \subset \text{exchg}(G_{i_1}..G_{i_L}) \mid \\ \qquad T \cap \text{cptScope}(G_{i_1}..G_{i_L}, \mathcal{E}, S) = \\ \qquad T \cap \text{mxdbagScope}(G_{i_1}..G_{i_L}, S) \;\}\,. \end{array} \right.$$

Ces travaux sur les sélections de communautés ont été appliqués à diverses données de microbiote intestinal dans Frioux et al., 2018. Un premier benchmark sur plus de 2,000 réseaux métaboliques issus du Human Microbiome Project (Engel et al., 2016 ; Human Microbiome Project, 2012) a montré que pour 86,5% des fonctions métaboliques nécessitant une communauté, il existe plus de 100 communautés minimales équivalentes. Pour près de la moitié des fonctions, plus de 1,000 communautés minimales existent. Ces travaux montrent la redondance des fonctions dans un microbiote et la nécessité de screener les fonctions et les bactéries associées. Utiliser les deux méthodologies de sélection de communautés séquentiellement permet de passer à l'échelle et de réduire le nombre de bactéries à considérer (bactéries apparaissant dans au minimum une communauté optimale). Cette réduction de complexité du microbiote permet de se focaliser sur des bactéries importantes pour la fonction considérée. Dès lors, un travail plus approfondi sur ces espèces est envisageable, de même que l'utilisation de modèles mathématiques quantitatifs ne passant pas à l'échelle du microbiote initial.

Plus généralement, ces travaux ont montré que cette sémantique topologique basée sur une approximation booléenne de la dynamique du réseau métabolique est pertinente pour l'étude des organismes non-modèles et des communautés

d'organisme : elle permet de suggérer des réactions pour compléter des réseaux ou réduire des microbiomes dans des situations où les réseaux considérés sont imparfaits en termes d'équilibres de flux ou d'annotation. La flexibilité laissée sur le fait de cibler les modifications du réseaux à étudier (en n'étant pas forcé de produire toutes les cibles attendues, c'est-à-dire tous les composants de la biomasse en même temps), est particulièrement utile pour accompagner la curation d'un réseau métabolique incomplet (Frioux et al., 2018 ; Prigent et al., 2017).

5 Sémantique de flux

Comme expliqué dans les sections précédentes, la seconde grande famille de modélisation de la fonctionnalité d'un réseau métabolique se base sur les méthodes utilisant les flux. On considère que, dans un état stationnaire, les composés internes du système ne s'accumulent pas. Sur la base de cette hypothèse, les productions et consommation de chaque métabolite interne sont équilibrées. Le système vérifie alors un ensemble d'équations du type

$$0 = \frac{d\mathbf{m}(t)}{dt} = \mathbf{S}\,\mathbf{v}(\mathbf{m}(t)) = \mathbf{S}\,\mathbf{v},$$

où $\mathbf{v}$ désigne le vecteur associé aux flux des réactions lorsque les concentrations des métabolites sont supposées constantes. Le comportement du système est alors contraint par des relations linéaires sur les variables associées aux flux des réactions. L'ensemble $active^s_G(S)$ des réactions activables est calculé en faisant une hypothèse de maximisation de biomasse.

5.1 Modéliser la réponse d'un réseau métabolique avec des flux

Étant donné un réseau métabolique $G = (R \cup M, E, stc)$, on associe à chaque réaction $r \in R$ une variable, le flux v_r qui représente son activité, généralement en millimole par gramme de matière sèche par heure ($\text{mmol}\times\text{gDW}^{-1}\times\text{hr}^{-1}$). L'ensemble des flux du système forme le vecteur $\mathbf{v}$. Le flux de l'ensemble des réactions est borné par une limite inférieure (vecteur $\mathbf{b}_l$, *lower bound*) et une limite supérieure (vecteur $\mathbf{b}_u$, *Upper Bound*). La stœchiométrie est prise en compte dans cette modélisation et sa précision est essentielle car le modèle mathématique sous-jacent est la matrice stœchiométrique $\mathbf{S}$.

Le domaine de l'*analyse des flux*, dont la méthode principale est la *Flux Balance Analysis* ou FBA (Orth et al., 2010), consiste à produire des méthodes pour analyser la distribution des flux dans le réseau métabolique. Dans ce domaine, le but est de s'appuyer sur des problèmes d'optimisation linéaires en nombre entier pour trouver une distribution des flux dans le modèle tel que le flux z associé à une réaction objectif r_{obj} soit maximal. Ce problème peut être résolu en programmation linéaire (LP) si le système est considéré à l'état stationnaire, c'est-à-dire si la quantité produite d'un métabolite est égale à la

quantité consommée de ce même métabolite (Orth et al., 2010). La définition suivante (Maranas & Zomorrodi, 2016) présente le problème d'optimisation associé à la FBA comme suit :

$$\begin{aligned} \textit{maximiser} \quad & z = \mathbf{v}_{r_{obj}} \\ \textit{tel que} \quad & \mathbf{S}\mathbf{v} = 0 \qquad (1) \\ & \mathbf{b}_l \leq \mathbf{v} \leq \mathbf{b}_u \qquad (2) \\ & \mathbf{v} \in \mathbb{R}^{\text{size}(R)}. \end{aligned}$$

Si on détaille l'écriture matricielle, en notant $\mathbf{S}_{m,r}$ le coefficient stoechiométrique de chaque composé m apparaissant dans chaque réaction r, ce problème se reformule comme suit :

$$\begin{aligned} \textit{maximiser} \quad & z = \mathbf{v}_{r_{obj}} \\ \textit{tel que} \quad & \sum_{r \in R} \mathbf{S}_{m,r}\mathbf{v}_r = 0, \forall m \in M \\ & (\mathbf{b}_l)_r \leq \mathbf{v}_r \leq (\mathbf{b}_u)_r, \forall r \in R \\ & \mathbf{v}_r \in \mathbb{R}, \forall r \in R. \end{aligned}$$

L'analyse de flux peut être utilisée pour évaluer l'activation d'une réaction selon une sémantique de flux, de même que le concept de périmètre (scope) était utilisé pour évaluer l'activation topologique. Étant donnés une réaction objectif $r_{obj} \in R_{obj}$, un réseau métabolique $G = (R \cup M, E, stc)$ et un ensemble de graines S, l'*ensemble des réactions activées à l'échelle stœchiométrique* est défini comme suit :

$$r_{obj} \in \textit{active}^s_G(S) \quad \text{s'il existe } \mathbf{v} \in \mathbb{R}^{\text{size}(R)} \text{ tel que} \quad \begin{aligned} & \mathbf{v}_{r_{obj}} > 0 \\ & \mathbf{S}\mathbf{v} = 0 \\ & \mathbf{b}_l \leq \mathbf{v} \leq \mathbf{b}_u. \end{aligned}$$

Dans la Fig. 5, on applique une analyse de flux FBA à un exemple de réseau métabolique. Il est à noter que les entrées du réseau métabolisme, correspondant aux graines pour la sémantique basée sur les graphes, sont modélisées par des réactions d'échange. Ces réactions d'échange, telles qu'elles sont illustrées dans la Fig. 5 permettent de représenter les entrées et les sorties entre le réseau métabolique et le milieu de culture. Les nutriments sources sont apportés dans le réseau par des réactions sans substrat (r_1, r_2). De même, la réaction d'export de D, r_6, n'a pas de produit et permet d'équilibrer les flux en assurant que D ne s'accumule pas dans le réseau. Les réactions du réseau qui ne modélisent pas un échange avec le milieu de culture, sont généralement qualifiées d'internes (Schellenberger, Lewis et al., 2011).

5.2 Cycles à l'état stationnaire

Comme précisé ci-dessus, la FBA repose sur une hypothèse d'état stationnaire du système, ce qui implique que l'ensemble des réactions activées à l'échelle

stœchiométrique dépend également de cette hypothèse. C'est une différence majeure avec l'activation topologique, qui, elle, représente l'effet de l'état initial du système. La différence s'observe dans la Fig. 4 que nous avons introduite dans une section précédente. Dans cet exemple, r_{s_1}, r_{s_2} et r_{s_3} sont les réactions qui connectent le réseau avec l'extérieur du système, et nous avions expliqué pourquoi le cycle $\{C, D, E\}$ n'est pas activable en fonction de la sémantique topologique. Dans la Fig. 6, nous reprenons cet exemple en le simplifiant en remplaçant la consommation de C par r_5 par une réaction d'export. Le système d'équation associées au modèle stationnaire $\mathbf{Sv} = 0$ se résout facilement pour donner la solution $\mathbf{v}_{r_{S_2}} = \mathbf{v}_{r_2} = \mathbf{v}_{r_3} = \mathbf{v}_{r_4} = \mathbf{v}_{r_e} = \mathbf{v}_{r_{e_G}}$. La valeur maximale des flux sera donnée par les bornes $\mathbf{b}_u$ et $\mathbf{b}_l$. Dans tous les cas, on en déduit que le cycle $\{C, D, E\}$ est activé selon la sémantique de flux, et ne l'est pas selon la sémantique topologique. Plus précisément, cet exemple nous indique qu'il existe une solution possible à la répartition des flux dans le système à l'état stationnaire. Cela suppose implicitement qu'au moins un des composants du cycle a été produit en quantité suffisante avant de rejoindre l'état stationnaire pour alimenter ce cycle. La sémantique topologique cherche à vérifier s'il existe toujours une explication à l'activation de ces composés depuis l'état initial du système depuis le milieu de culture, et répond par la négative à cette question. On voit donc ici que les deux sémantiques sont complémentaires : la sémantique stoechiométrique décrit le potentiel d'activation à l'état stationnaire tandis que la sémantique topologique s'intéresse aux phases éventuellement transitoires du système.

5.3 Application à la complétion de réseaux métaboliques

La sémantique stœchiométrique est la principale sémantique utilisée pour valider un réseau métabolique après sa reconstruction à l'échelle génomique. Il existe donc des méthodes de complétion qui reposent sur des formalismes stœchiometriques, visant à résoudre le problème suivant :

$$\text{completion}^s(S, R_T, G, G') = \underset{R'' \subseteq R'}{\arg\min} \left| \begin{array}{l} \text{size}(R_T \setminus \mathit{active}^s_{\mathit{extension}(G,G',R'')}(S)) \\ \text{size}(R'') \end{array} \right. .$$

Néanmoins il n'existe pas, à notre connaissance, de méthode de complétion parcimonieuse permettant de résoudre exactement ce problème, c'est-à-dire de cibler une ou plusieurs réactions et s'assurer de leur activation stœchiométrique, comme il est possible de le faire avec l'activation topologique. Il existe néanmoins des heuristiques de résolution non-minimales, permettant de cibler l'activation de tout ou partie du réseau métabolique initial, parfois en ajoutant des réactions d'import ou d'export de composés en plus de réactions de la base de données.

Différentes méthodes de complétion basées sur des sémantiques d'activation stoechiométriques ou booléennes sont comparées dans Prigent et al., 2017, Karp et al., 2018, Pan et Reed, 2018 et Latendresse et Karp, 2018. On notera en particulier que FastGapFill (Thiele et al., 2014), une méthode de

complétion non parcimonieuse mais passant à l'échelle, permet de compléter un réseau métabolique dans sa globalité, avec un nombre raisonnable de réactions. D'autres méthodologies telles que GapFill (Satish Kumar et al., 2007) résolvent une version modifiée de des conditions stœchiométriques introduites ci-dessus, dans laquelle l'accumulation est autorisée : $\mathbf{Sv} \geq 0$.

6 Comparaison des sémantiques : vers une approche hybride

6.1 Complémentarité des abstractions booléennes et stoechiométriques

La sémantique topologique d'activation a été utilisée en complémentarité de la sémantique basée sur les flux dans Laniau et al., 2017 pour étudier les différentes classes de métabolites d'un réseau métabolique.

La Fig. 7 décrit des exemples caractéristiques des différences entre les modélisations booléenne et stœchiométrique de l'activation du métabolisme. Les réactions r_{i_1} et r_{i_2} connectent le réseau avec l'extérieur du système et son milieu de culture. Elles produisent les graines qui sont nécessaires à la production des composés I, E et K par les réactions internes r_3, r_7 et r_{11}. Au niveau stoechiométrique, sous hypothèse de stationnarité, le cycle r_8, r_9, r_{10} peut aussi s'équilibrer et activer la production de K, alors que ca n'est pas le cas si on utilise la sémantique topologique associée à une abstraction booléenne qui requiert d'expliquer la production de L par une réaction venant du milieu de culture. Ces exemples confirment par ailleurs que la sémantique stœchiométrique est fortement sensible aux imprécisions de la stœchiométrie associée aux métabolites, contrairement à la sémantique topologique. Les activations différentes des réactions $\{r_8, r_9, r_{10}\}$ mettent à nouveau en évidence les différences vis-à-vis des cycles. Le sous-réseau jaune montre quant à lui l'effet de l'accumulation de métabolites sur la capacité d'activation de réactions subséquentes et de la production finale d'autres composés. D'une manière générale, l'abstraction booléenne (sémantique topologique) est moins précise, car qualitative, que l'abstraction stœchiométrique, mais elle est également plus résiliente aux imprécisions du réseau, semblant ainsi particulièrement indiquée pour modéliser le métabolisme d'organismes non-modèles.

La sémantique topologique a été comparée à la sémantique stœchiométrique par Kruse et Ebenhöh, 2008. À l'exception de certains composés (principalement des cofacteurs (Ebenhöh et al., 2006)), il est réaliste d'attendre que les métabolites internes sont synthétisés *de novo* à partir de composés issus du milieu de culture. Les auteurs concluent que le calcul du périmètre (scope) du milieu de culture, en particulier en supposant la production de quelques cofacteurs, coincide avec les prédictions issues de méthodes basées sur l'hypothèse de stationnarité (FBA). En conséquence, la sémantique topologique est une bonne alternative pour déterminer la capacité de production de composés dans un réseau métabolique, puisque son calcul est d'une complexité faible et

que cette approche a des résultats similaires à l'approche basée sur les flux. Cette abstraction booléenne est compatible avec l'hypothèse que la cellule est constamment en croissance ou en phase de reproduction, c'est-à-dire à un état non-stationnaire, contrairement aux hypothèses faites par les méthodes basées sur des flux.

6.2 Complétion hybride de réseaux métaboliques

Les deux sémantiques d'activation sont compatibles et peuvent être combinées pour formaliser une sémantique hybride satisfaisant d'une part les contraintes d'alimentation par les graines de la sémantique topologique, et d'autre part les contraintes linéaires de non-accumulation de flux de la sémantique stœchiométrique. Ainsi, il est possible de définir l'ensemble des réactions activable de manière hybride à partir d'un ensemble de graines S dans un réseau métabolique G (Frioux et al., 2017, 2019) en combinant les deux concepts :

$$active^h_G(S) = active^s_G(S) \bigcap active^t_G(S).$$

Cette unification des sémantiques peut être appliquée au problème de la complétion de réseaux métaboliques individuels en visant à rendre un ensemble de réactions R_{obj} activées pour la condition hybride. Les réactions objectifs sont alors activées, si leurs substrats appartiennent au périmètre des graines décrivant le milieu de culture et s'il existe une distribution de flux dans le réseau complété tel qu'un flux positif est possible pour les réactions objectifs.

$$\text{completion}^h(S, R_T, G, G') = \underset{R'' \subseteq R'}{\arg\min} \left| \begin{array}{l} \text{size}(R_T \setminus active^h_{extension(G,G',R'')}(S)) \\ \text{size}(R'') \end{array} \right. .$$

La Fig. 8 illustre la complétion hybride. Une approche pour résoudre ce problème de complétion est détaillée dans Frioux et al., 2017 et Frioux et al., 2019.

7 Résoudre les problèmes de complétion avec la programmation par ensembles-réponses

La programmation par ensembles réponses (*Answer Set Programming* ASP) est un paradigme de programmation logique non monotone particulièrement adapté aux problèmes d'optimisation combinatoire tels que ceux présentés dans ce chapitre sur la complétion, utilisant la sémantique topologique. Le problème à résoudre est décrit sous forme de règles et d'atomes. Contrairement à des algorithmes de programmation itérative, le problème est décrit en tant que tel mais la stratégie de résolution est laissée au solveur. Définir ce qui doit être résolu est à la charge du modélisateur, « comment » cette résolution se fera est géré par le solveur. Cette section propose une introduction à la programmation ASP. Des informations supplémentaires peuvent être obtenues dans le livre de

Gebser et al., 2012. Une distribution ASP est Clingo (Gebser et al., 2016) du projet Potassco [2] avec le solveur Clasp.

Les principaux intérêts d'ASP sont doubles. Le premier est l'expressivité du langage déclaratif proposé : une logique propositionnelle du premier ordre. Une originalité d'ASP est la négation : un atome est faux tant qu'une règle ne l'a pas rendu vrai (raisonnement en monde clos). Le second intérêt est lié aux capacités de résolution par des solveurs efficaces dérivés des techniques de résolution SAT, et à la possibilité de paramétrer la résolution pour la guider, ou explorer l'espace des solutions (union, énumération, intersection des solutions) à l'aide de puissantes heuristiques de résolution. ASP permet de résoudre des problèmes NP (Gebser et al., 2012).

Le problème à résoudre étant décrit, l'approche ASP est pertinente pour cibler des familles de problèmes aux caractéristiques similaires telles que ceux de la complétion métabolique. À partir de cette description, une phase d'instanciation va transformer la représentation du modèle faite par l'utilisateur en un format propositionnel fini. Les ensembles d'atomes satisfaisant les contraintes sont appelés ensembles réponses (*answer sets*). Il est possible de filtrer ces ensembles réponses avec des optimisations.

Un programme logique en ASP est de la forme suivante :

$$\underbrace{a_0}_{head} :- \underbrace{a_1, ..., a_m,\ \text{not}\ a_{m+1}, ...,\ \text{not}\ a_n.}_{body}$$

$\{a_0, ..., a_n\}$ sont des atomes, $:-$ peut être interprété comme un « si ». En conséquence, une règle peut être interprétée comme suit : « la tête (*head*) d'une règle est vraie si le corps (*body*) de la règle est valide ». En particulier, la règle $a_0.$ est toujours vraie, c'est un fait ; et $:- a_0$ est toujours fausse. La négation est importante en ASP. Un atome n'appartenant pas à la tête d'une règle (après instanciation) ne peut être vrai. Ainsi :

```
p.
:- p, not q.
```

n'a pas d'ensembles réponses, alors que

```
p.
q :- p.
```

a l'ensemble réponse $\{p, q\}$.

7.1 Modéliser l'activation booléenne d'une réaction en ASP

ASP permet de représenter un réseau métabolique en terme de réactions et métabolites, mais également des concepts tels que l'accessibilité des composés (scope), comme montré dans l'exemple suivant.

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

2. https ://potassco.org/

La première ligne exprime que les graines appartiennent au scope. La seconde exprime la récursivité de la définition du scope : le produit d'une réaction appartient au scope uniquement si tous les substrats de cette réaction sont eux-mêmes dans le scope. Ainsi, en considérant le réseau métabolique de la figure 3, la définition du scope appliquée aux données suivantes :

```
seed("A"). seed("B").
reaction("r1"). reactant("A", r1). product("D", r1).
reaction("r2"). reactant("B", r2). product("E", r2).
reaction("r3"). reactant("C","r3"). product("F","r3").
reaction("r4"). reactant("D","r4"). reactant("E","r4").
product("G","r4"). reaction("r5"). reactant("E","r5").
reactant("F","r5"). product("I","r5"). product("H","r5").
```

produit l'ensemble réponse suivant :

```
scope("A") scope("B") scope("D") scope("E") scope("G").
```

La programmation ASP a été utilisée pour les problèmes de complétion utilisant la sémantique topologique (Prigent et al., 2017) et hybride (Frioux et al., 2017, 2019). Par concision, les problèmes associés en ASP ne seront pas retranscrits ici. Néanmoins, notons que la programmation ASP a pu associer des contraintes linéaires pour la résolution du second problème grâce à la combinaison du solveur logique avec un solveur linéaire par un propagateur de contraintes (Janhunen et al., 2017).

7.2 Sélection non compartimentalisée de communautés

La sélection non-compartimentalisée telle que présentée dans Frioux et al., 2018 et introduite dans la section 3.3 est réalisée à l'aide la programmation ASP. Le code suivant décrit les étapes essentielles de la description du problème.

Listing 1 – Sélection de communautés de taille minimale dans un modèle non compartimentalisé

```
1  {sel_orga(B) : orga(B)}.
2  mixedbagScope(M) :- seed(M).
3  mixedbagScope(M) :- sel_orga(B); product(M,R); reaction(R,
       B); mixedbagScope(M2) : reactant(M2,R).
4  #minimize { 1@2,T : target(T), not mixedbagScope(T)}.
5  #minimize { 1@1,B : sel_orga(B)}.
```

La ligne 1 implique que la sélection des organismes de la communauté se fait au sein d'un ensemble d'organismes disponibles, un microbiote par exemple. Les deux lignes suivantes définissent le scope non-compartimentalisé des organismes sélectionnés. Les optimisations des lignes 4 et 5 visent à d'abord minimiser le nombre de cibles non productibles, c'est-à-dire qui ne sont pas dans le scope de la communauté sélectionnée, puis à minimiser le nombre d'organismes sélectionnés. Au-delà de l'expressivité dans la description du problème, la programmation ASP est particulièrement pertinente ici par ses capacités d'exploration de l'espace des solutions : proposition d'une solution optimale, énumération de toutes

les solutions optimales, union et intersection de ces solutions (sans passer par l'énumération). Cela permet notamment de combiner ce problème avec le suivant, compartimentalisé, pour filtrer les solutions de communautés par le coût de l'interaction qu'elles impliquent.

7.3 Sélection compartimentalisée de communautés

La sélection dans un modèle compartimentalisé telle que présentée dans ce chapitre (section 3.4) vise à minimiser le coût des interactions en sélectionnant des communautés aux échanges métaboliques minimaux vis-à-vis de l'objectif. Le prédicat *escope* est introduit dans la modélisation pour calculer récursivement le scope du modèle compartimentalisé associé à une famille d'organismes sélectionnés et un ensemble de réactions d'échanges :

Listing 2 – Sélection de communautés de taille minimale et d'échanges minimaux dans un modèle compartimentalisé.

```
1  {sel_orga(B) : orga(B)}.
2  {exchanged(M,O1,O2) : metabolite(M,O1); metabolite(M,O2);
       escope(M,O1); sel_orga(O1); sel_orga(O2); O1!=O2}.
3  escope(M,O) :- seed(M); sel_orga(O).
4  escope(M,O) :- exchanged(M,_,O), sel_orga(O).
5  escope(M,O) :- product(M,R); reaction(R,O); sel_orga(O);
       escope(M2,O) : reactant(M2,R).
6  #minimize{1@3,M : target(M), not escope(M,_)}.
7  #minimize { 1@2,B : sel_orga(B)}.
8  #minimize{1@1,M,O1,O : exchanged(M,O1,O)}.
```

La ligne 2 décrit l'ensemble des composés métaboliques échangeables dans lequel vont être sélectionnés les échanges minimaux. Un métabolite est échangeable s'il est dans le scope individuel de l'organisme fournisseur (possiblement après qu'il ait lui aussi reçu d'autres métabolites), et si ce dernier comme le receveur sont des organismes sélectionnés dans la communauté. En l'absence d'un hôte, les cibles métaboliques doivent être productibles par au moins un des organismes sélectionnés. Dans le cas où un hôte est considéré, les cibles seront attendues comme étant productibles chez l'hôte (ou procurées par un symbiote).

Les exemples ci-dessus démontrent que la programmation par ensembles réponses est un formalisme performant pour décrire et résoudre des problèmes ayant une sémantique booléenne. Les solveurs ASP sont particulièrement adaptés à la grande combinatoire des problèmes présentés ici et permettent notamment, par des modes de résolution spécifiques, d'explorer l'espace de recherche des solutions. Il est ainsi possible d'obtenir des modèles correspondant à l'union ou l'intersection de toutes les solutions optimales sans nécessité l'énumération de ces dernières. Par ailleurs, il est aussi possible de les énumérer si nécessaire. Appliqués aux questions biologiques sous-jacentes des problèmes modélisés, les modes de résolution en ASP engendrent la possibilité de proposer davantage de modèles qu'un unique modèle optimale. Par exemple, il est possible d'obtenir par une résolution efficace, l'ensemble des organismes apparaissant dans une solution optimale lors de la sélection de communautés.

8 Conclusion

Ce chapitre a mis en évidence que les définitions des problèmes de complétion du métabolisme ont une base commune malgré l'existence de diverses sémantiques pour les résoudre. La complétion métabolique permet d'améliorer les réseaux métaboliques individuels, mais également de comprendre les interactions entre ces organismes à l'échelle d'une communautés, ainsi que de sélectionner de telles communautés. La base commune aux problèmes de complétion liés au métabolisme se retrouve également dans leurs formulations en ASP, et se décline pour prendre en compte les spécificités de chacun des problèmes. Ainsi les problèmes de complétion pour des réseaux individuels ou des réseaux d'organismes en interaction peuvent être résolus, et de nouvelles déclinaisons de ces problèmes sont facilement envisageables. Comme présenté dans ce chapitre, il est également possible d'aller plus loin, en rajoutant une syntaxe linéaire à ASP, permettant de résoudre efficacement des problèmes hybrides.

La modélisation du métabolisme présentée dans ce chapitre repose sur les données génomiques associées aux organismes considérés. Il convient de prendre en compte les limites de ces modélisations, tous les gènes d'un organisme n'étant pas exprimés, ou pas dans toutes les conditions. L'aspect régulation de l'expression génique est important à considérer. Il est possible de moduler les réseaux métaboliques avec d'autres données « omiques » pour mieux représenter des conditions particulières. Un exemple est l'utilisation de données de transcriptomique pour prédire le sous-réseau métabolique exprimé. D'une manière générale, associer les différents types de réseaux biologiques, qu'ils traitent de régulation, d'interaction protéique ou de métabolisme, par l'intégration de données « omiques », constitue un socle méthodologique déterminant pour comprendre la biologie des organismes et des communautés.

Références

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S. T., Meth ? ?, B., Schloss, P. D., Gevers, D., Mitreva, M., & Huttenhower, C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome (J. A. Eisen, Éd.). *PLoS Computational Biology*, *8*(6), e1002358. https://doi.org/10.1371/journal.pcbi.1002358

Acuna, V., Milreu, P. V., Cottret, L., Marchetti-Spaccamela, A., Stougie, L., & Sagot, M.-F. (2012). Algorithms and complexity of enumerating minimal precursor sets in genome-wide metabolic networks. *Bioinformatics*, *28*(19), 2474-2483. https://doi.org/10.1093/bioinformatics/bts423

Acuña, V., Chierichetti, F., Lacroix, V., Marchetti-Spaccamela, A., Sagot, M.-F., & Stougie, L. (2009). Modes and cuts in metabolic

networks : Complexity and algorithms. *Biosystems*, *95*(1), 51-60. https://doi.org/10.1016/j.biosystems.2008.06.015

Andrade, R., Wannagat, M., Klein, C. C., Acuña, V., Marchetti-Spaccamela, A., Milreu, P. V., Stougie, L., & Sagot, M.-F. (2016). Enumeration of minimal stoichiometric precursor sets in metabolic networks. *Algorithms for molecular biology : AMB*, *11*, 25. https://doi.org/10.1186/s13015-016-0087-3

Arkin, A. P., Stevens, R. L., Cottingham, R. W., Maslov, S., Henry, C. S., Dehal, P., Ware, D., Perez, F., Harris, N. L., Canon, S., Sneddon, M. W., Henderson, M. L., Riehl, W. J., Gunter, D., Murphy-Olson, D., Chan, S., Kamimura, R. T., Brettin, T. S., Meyer, F., ... Yoo, S. (2016). The DOE Systems Biology Knowledgebase (KBase). *bioRxiv*.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., ... Palsson, B. (2008). The RAST Server : Rapid Annotations using Subsystems Technology. *BMC Genomics*, *9*(1), 75. https://doi.org/10.1186/1471-2164-9-75

Baker, J. L., Bor, B., Agnello, M., Shi, W., & He, X. (2017). Ecology of the Oral Microbiome : Beyond Bacteria. *Trends in microbiology*, *25*(5), 362-374. https://doi.org/10.1016/j.tim.2016.12.012

Blackall, L. L., Wilson, B., & van Oppen, M. J. H. (2015). Coral-the world's most diverse symbiotic ecosystem. *Molecular Ecology*, *24*(21), 5330-5347. https://doi.org/10.1111/mec.13400

Borenstein, E., Kupiec, M., Feldman, M. W., & Ruppin, E. (2008). Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14482-7. https://doi.org/10.1073/pnas.0806162105

Bosch, T. C., & McFall-Ngai, M. J. (2011). Metaorganisms as the new frontier. *Zoology*, *114*(4), 185-190. https://doi.org/10.1016/j.zool.2011.04.001

Carr, R., & Borenstein, E. (2012). NetSeed : a network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment. *Bioinformatics*, *28*(5), 734-735. https://doi.org/10.1093/bioinformatics/btr721

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S., Subhraveti, P., & Karp, P. D. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research*, *46*(D1), D633-D639. https://doi.org/10.1093/nar/gkx935

Cavaliere, M., Feng, S., Soyer, O. S., & Jiménez, J. I. (2017). Cooperation in microbial communities and their biotechnological applications. *En-*

vironmental Microbiology, *19*(8), 2949-2963. https://doi.org/10.1111/1462-2920.13767

Chan, S. H., Cai, J., Wang, L., Simons-Senftle, M. N., & Maranas, C. D. (2017). Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. *Bioinformatics*, *33*(22), 3603-3609. https://doi.org/10.1093/bioinformatics/btx453

Chan, S. H. J., Simons, M. N., & Maranas, C. D. (2017). SteadyCom : Predicting microbial abundances while ensuring community stability (N. D. Price, Éd.). *PLOS Computational Biology*, *13*(5), e1005539. https://doi.org/10.1371/journal.pcbi.1005539

Chitale, M., Khan, I. K., & Kihara, D. (2016). Missing gene identification using functional coherence scores. *Scientific Reports*, *6*(1), 31725. https://doi.org/10.1038/srep31725

Christian, N., May, P., Kempa, S., Handorf, T., & Ebenhöh, O. (2009). An integrative approach towards completing genome-scale metabolic networks. *Molecular BioSystems*, *5*(12), 1889-1903. https://doi.org/10.1039/B915913b

Collet, G., Eveillard, D., Gebser, M., Prigent, S., Schaub, T., Siegel, A., & Thiele, S. (2013). Extending the Metabolic Network of Ectocarpus Siliculosus Using Answer Set Programming. *LPNMR 2013 : Logic Programming and Nonmonotonic Reasoning*, 245-256. https://doi.org/10.1007/978-3-642-40564-8_25

Cottret, L., & Jourdan, F. (2010). Graph methods for the investigation of metabolic networks in parasitology. https://doi.org/10.1017/S0031182010000363

Cottret, L., Milreu, P. V., Acuña, V., Marchetti-Spaccamela, A., Stougie, L., Charles, H., & Sagot, M.-F. (2010). Graph-Based Analysis of the Metabolic Exchanges between Two Co-Resident Intracellular Symbionts, Baumannia cicadellinicola and Sulcia muelleri, with Their Insect Host, Homalodisca coagulata (J. A. Papin, Éd.). *PLoS Computational Biology*, *6*(9), e1000904. https://doi.org/10.1371/journal.pcbi.1000904

Cottret, L., Vieira Milreu, P., Acuña, V., Marchetti-Spaccamela, A., Viduani Martinez, F., Sagot, M.-F., & Stougie, L. (2008). Enumerating Precursor Sets of Target Metabolites in a Metabolic Network. In K. Crandall & J. Lagergren (Éd.), *Algorithms in Bioinformatics* (p. 233-244). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-87361-7_20

Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A. A., & Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods in molecular biology (Clifton, N.J.)*, *985*, 17-45. https://doi.org/10.1007/978-1-62703-299-5_2

Doran, P. M. (2013). *Bioprocess engineering principles*. Elsevier/Academic Press. https://books.google.fr/books?id=wZSylDhgEXMC%7B%5C&%7Dpg=PA664%7B%5C&%7Dlpg=PA664%7B%5C&%7Ddq=flux+analysis+steady+state+assumption%7B%5C&%7Dsource=bl%7B%

5C&%7Dots=fSemZaon-x%7B%5C&%7Dsig=ACfU3U2f2O8iQOeWDsasYAgOOvL2KmS65g%7B%5C&%7Dhl=fr%7B%5C&%7Dsa=X%7B%5C&%7Dved=2ahUKEwjBsa6vjariAhVJ4eAKHaEQBcYQ6AEwCHoECAkQAQ%7B%5C#%7Dv=onepage%7B%5C&%7Dq=flux%20analy

Ebenhöh, O., Handorf, T., & Kahn, D. (2006). Evolutionary changes of metabolic networks and their biosynthetic capacities. *Systems biology*, *153*(5), 354-8. http://www.ncbi.nlm.nih.gov/pubmed/16986316

Ebenhöh, O., Handorf, T., & Heinrich, R. (2004). Structural analysis of expanding metabolic networks. *Genome informatics. International Conference on Genome Informatics*, *15*(1), 35-45. http://www.ncbi.nlm.nih.gov/pubmed/15712108

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy : COnstraints-Based Reconstruction and Analysis for Python. *BMC systems biology*, *7*, 74. https://doi.org/10.1186/1752-0509-7-74

Edwards, J. S., & Palsson, B. O. (1999). Systems properties of the Haemophilus influenzae Rd metabolic genotype. *Journal of Biological Chemistry*, *274*(25), 17410-17416. https://doi.org/10.1074/jbc.274.25.17410

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology and Evolution*, *29*(1), 51-63. https://doi.org/10.1016/j.tree.2013.09.008

Emms, D. M., & Kelly, S. (2015). OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Eng, A., & Borenstein, E. (2016). An algorithm for designing minimal microbial communities with desired metabolic capacities. *Bioinformatics*, *32*(13), 2008-2016. https://doi.org/10.1093/BIOINFORMATICS/BTW107

Engel, P., Kwong, W. K., McFrederick, Q., Anderson, K. E., Barribeau, S. M., Chandler, J. A., Cornman, R. S., Dainat, J., de Miranda, J. R., Doublet, V., Emery, O., Evans, J. D., Farinelli, L., Flenniken, M. L., Granberg, F., Grasis, J. A., Gauthier, L., Hayer, J., Koch, H., ... Dainat, B. (2016). The Bee Microbiome : Impact on Bee Health and Model for Evolution and Ecology of Host-Microbe Interactions. *mBio*, *7*(2), e02164-15. https://doi.org/10.1128/mBio.02164-15

Faria, J. P., Khazaei, T., Edirisinghe, J. N., Weisenhorn, P., Seaver, S. M. D., Conrad, N., Harris, N., DeJongh, M., & Henry, C. S. (2016). Constructing and Analyzing Metabolic Flux Models of Microbial Communities. In *Hydrocarbon and Lipid Microbiology Protocols* (p. 247-273). Springer. https://doi.org/10.1007/8623_2016_215

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial co-occurrence relationships in the Human Microbiome. *PLoS Computational Biology*, *8*(7), e1002606. https://doi.org/10.1371/journal.pcbi.1002606

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., & Palsson,

B. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, *3*, 121. https://doi.org/10.1038/msb4100155

Feist, A. M., & Palsson, B. O. (2010). The biomass objective function. *Current Opinion in Microbiology*, *13*(3), 344-349. https://doi.org/10.1016/j.mib.2010.03.003

Frioux, C. (2018). *Investigating host-microbiota cooperation with gap-filling optimization problems* (thèse de doct.). Université de Rennes 1. Université de Rennes 1. https://hal.inria.fr/tel-01945853

Frioux, C., Fremy, E., Trottier, C., & Siegel, A. (2018). Scalable and exhaustive screening of metabolic functions carried out by microbial consortia. *Bioinformatics*, *34*(17), i934-i943. https://doi.org/10.1093/bioinformatics/bty588

Frioux, C., Schaub, T., Schellhorn, S., Siegel, A., & Wanko, P. (2017). Hybrid Metabolic Network Completion. In *Logic Programming and Nonmonotonic Reasoning : 14th International Conference, LPNMR 2017, Espoo, Finland, July 3-6, 2017, Proceedings* (p. 308-321). Springer International Publishing. https://doi.org/10.1007/978-3-319-61660-5_28

Frioux, C., Schaub, T., Schellhorn, S., Siegel, A., & Wanko, P. (2019). Hybrid metabolic network completion. *Theory and Practice of Logic Programming*, *19*(01), 83-108. https://doi.org/10.1017/S1471068418000455

García-Jiménez, B., García, J. L., & Nogales, J. (2018). FLYCOP : Metabolic modeling-based analysis and engineering microbial communities. *Bioinformatics*, *34*(17), i954-i963. https://doi.org/10.1093/bioinformatics/bty561

Gauthier, J.-P., Outreman, Y., Mieuzet, L., & Simon, J.-C. (2015). Bacterial communities associated with host-adapted populations of pea aphids revealed by deep sequencing of 16S ribosomal DNA. *PloS one*, *10*(3), e0120664. https://doi.org/10.1371/journal.pone.0120664

Gebser, M., Kaminski, R., Kaufmann, B., Ostrowski, M., Schaub, T., & Wanko, P. (2016). Theory solving made easy with clingo. *OpenAccess Series in Informatics*, *52*. https://doi.org/10.4230/OASIcs.ICLP.2016.2

Gebser, M., Kaminski, R., Kaufmann, B., & Schaub, T. (2012). Answer Set Solving in Practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(3), 1-238. https://doi.org/10.2200/S00457ED1V01Y201211AIM019

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4), 392-400. https://doi.org/10.1038/nm.4517

Greenblum, S., Turnbaugh, P. J., & Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(2), 594-9. https://doi.org/10.1073/pnas.1116053109

Handorf, T., & Ebenhoh, O. (2007). MetaPath Online : a web server implementation of the network expansion algorithm. *Nucleic Acids Research*, *35*(Web Server), W613-W618. https://doi.org/10.1093/nar/gkm287

Handorf, T., Ebenhöh, O., & Heinrich, R. (2005). Expanding metabolic networks : Scopes of compounds, robustness, and evolution. *Journal of Molecular Evolution*, *61*(4), 498-512. https://doi.org/10.1007/s00239-005-0027-1

Hanemaaijer, M., Roling, W. F. M., Olivier, B. G., Khandelwal, R. A., Teusink, B., & Bruggeman, F. J. (2015). Systems modeling approaches for microbial community studies : From metagenomics to inference of the community structure. *Frontiers in Microbiology*, *6*(MAR), 213. https://doi.org/10.3389/fmicb.2015.00213

Hassani, M. A., Durán, P., & Hacquard, S. (2018). Microbial interactions within the plant holobiont. *Microbiome*, *6*(1), 58. https://doi.org/10.1186/s40168-018-0445-0

Heinken, A., Ravcheev, D. A., & Thiele, I. (2016). Systems biology of bacteria-host interactions. In *The Human Microbiota and Chronic Disease : Dysbiosis as a Cause of Human Pathology* (p. 113-137). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118982907.ch7

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdóttir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., Magnusdóttir, S., Ng, C. Y., Preciat, G., Žagare, A., Chan, S. H. J., Aurich, M. K., Clancy, C. M., Modamio, J., Sauls, J. T., ... Fleming, R. M. T. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nature Protocols*, *14*(3), 639-702. https://doi.org/10.1038/s41596-018-0098-2

Henry, C. S., Bernstein, H. C., Weisenhorn, P., Taylor, R. C., Lee, J. Y., Zucker, J., & Song, H. S. (2016). Microbial Community Metabolic Modeling : A Community Data-Driven Network Reconstruction. *Journal of Cellular Physiology*, *231*(11), 2339-2345. https://doi.org/10.1002/jcp.25428

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, *28*(9), 977-982. https://doi.org/10.1038/nbt.1672

Hornung, B., Martins dos Santos, V. A., Smidt, H., & Schaap, P. J. (2018). Studying microbial functionality within the gut ecosystem by systems biology. https://doi.org/10.1186/s12263-018-0594-6

Human Microbiome Project, C. (2012). A framework for human microbiome research. *Nature*, *486*(7402), 215-221. https://doi.org/10.1038/nature11209

Janhunen, T., Kaminski, R., Ostrowski, M., Schellhorn, S., Wanko, P., & Schaub, T. (2017). Clingo goes linear constraints over reals and integers. *Theory and Practice of Logic Programming*, 872-888. https://doi.org/10.1017/S1471068417000242

Johns, N. I., Blazejewski, T., Gomes, A. L., & Wang, H. H. (2016). Principles for designing synthetic microbial communities. *Current Opinion in Microbiology*, *31*, 146-153. https://doi.org/10.1016/j.mib.2016.03.010

Julien-Laferrière, A., Bulteau, L., Parrot, D., Marchetti-Spaccamela, A., Stougie, L., Vinga, S., Mary, A., & Sagot, M.-F. (2016). A Combinatorial Algorithm for Microbial Consortia Synthetic Design. *Scientific Reports*, *6*, 29182. https://doi.org/10.1038/srep29182

Karp, P. D., Latendresse, M., Paley, S. M., Krummenacker, M., Ong, Q. D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., Spaulding, A., Fulcher, C., Keseler, I. M., & Caspi, R. (2016). Pathway tools version 19.0 update : Software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, *17*(5), 877-890. https://doi.org/10.1093/bib/bbv079

Karp, P. D., Weaver, D., & Latendresse, M. (2018). How accurate is automated gap filling of metabolic models ? *BMC Systems Biology*, *12*(1), 73. https://doi.org/10.1186/s12918-018-0593-7

Kim, M. K., & Lun, D. S. (2014). Methods for integration of transcriptomic data in genome-scale metabolic models. *Computational and Structural Biotechnology Journal*, *11*(18), 59-65. https://doi.org/10.1016/j.csbj.2014.08.009
NULL

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J., & Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current opinion in biotechnology*, *23*(4), 617-23. https://doi.org/10.1016/j.copbio.2011.10.007

Kitano, H. (2002a). Systems biology : A brief overview. *Science*, *295*(5560), 1662-1664. https://doi.org/10.1126/science.1069492

Kitano, H. (2002b). Computational systems biology. *Nature*, *420*(6912), 206-210. https://doi.org/10.1038/nature01254

Klitgord, N., & Segrè, D. (2010). Environments that induce synthetic microbial ecosystems. *PLoS Computational Biology*, *6*(11), e1001002. https://doi.org/10.1371/journal.pcbi.1001002

Kruse, K., & Ebenhöh, O. (2008). Comparing flux balance analysis to network expansion : producibility, sustainability and the scope of compounds. *Genome Informatics*, *20*, 91-101. https://doi.org/9781848163003_0008[pii]

Laniau, J., Frioux, C., Nicolas, J., Baroukh, C., Cortes, M.-P. M.-P., Got, J., Trottier, C., Eveillard, D., & Siegel, A. (2017). Combining graph and flux-based structures to decipher phenotypic essential metabolites within metabolic networks. *PeerJ*, *5*(10), e3860. https://doi.org/10.7717/peerj.3860

Latendresse, M., & Karp, P. D. (2018). Evaluation of reaction gap-filling accuracy by randomization. *BMC Bioinformatics*, *19*(1), 53. https://doi.org/10.1186/s12859-018-2050-4

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL : Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178-2189. https://doi.org/10.1101/gr.1224503

Liu, H., & Deutschbauer, A. M. (2018). Rapidly moving new bacteria to model-organism status. *Current Opinion in Biotechnology*, *51*, 116-122. https://doi.org/10.1016/J.COPBIO.2017.12.006

Loira, N., Zhukova, A., & Sherman, D. J. (2015). Pantograph : A template-based method for genome-scale metabolic model reconstruction. *Journal of Bioinformatics and Computational Biology*, *13*(02), 1550006. https://doi.org/10.1142/S0219720015500067

Mannan, A. A., Toya, Y., Shimizu, K., McFadden, J., Kierzek, A. M., & Rocco, A. (2015). Integrating Kinetic Model of E. coli with Genome Scale Metabolic Fluxes Overcomes Its Open System Problem and Reveals Bistability in Central Metabolism. *PloS one*, *10*(10), e0139507. https://doi.org/10.1371/journal.pone.0139507

Manor, O., Levy, R., & Borenstein, E. (2014). Mapping the inner workings of the microbiome : genomic- and metagenomic-based study of metabolism and metabolic interactions in the human microbiome. *Cell metabolism*, *20*(5), 742-52. https://doi.org/10.1016/j.cmet.2014.07.021

Maranas, C. D., & Zomorrodi, A. R. (2016). *Optimization methods in metabolic networks*. Wiley.

Marchesi, J. R., & Ravel, J. (2015). The vocabulary of microbiome research : a proposal. *Microbiome*, *3*(1), 31. https://doi.org/10.1186/s40168-015-0094-5

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., Takeuchi, R., Nomura, W., Zhang, Z., Mori, H., Feist, A. M., & Palsson, B. O. (2017). iML1515, a knowledgebase that computes Escherichia coli traits. https://doi.org/10.1038/nbt.3956

Moran, N. A. (2006). Symbiosis. *Current Biology*, *16*(20), 866-871. https://doi.org/10.1016/J.CUB.2006.09.019

Muller, E. E., Faust, K., Widder, S., Herold, M., Martínez Arbas, S., & Wilmes, P. (2018). Using metabolic networks to resolve ecological properties of microbiomes. *Current Opinion in Systems Biology*, *8*, 73-80. https://doi.org/10.1016/j.coisb.2017.12.004

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. a., Nam, H., Feist, A. M., & Palsson, B. Ø. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Molecular systems biology*, *7*(535), 535. https://doi.org/10.1038/msb.2011.65

Orth, J. D., & Palsson, B. (2010). Systematizing the generation of missing metabolic knowledge. *Biotechnology and Bioengineering*, *107*(3), 403-412. https://doi.org/10.1002/bit.22844

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is Flux Balance Analysis ? *Nature biotechnology*, *28*(3), 245-248. https://doi.org/10.1038/nbt.1614

Pan, S., & Reed, J. L. (2018). Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discove-

ries. *Current Opinion in Biotechnology*, *51*, 103-108. https://doi.org/10.1016/j.copbio.2017.12.012

Pande, S., & Kost, C. (2017). Bacterial Unculturability and the Formation of Intercellular Metabolic Networks. *Trends in Microbiology*, *25*(5), 349-361. https://doi.org/10.1016/J.TIM.2017.02.015

Plata, G., Fuhrer, T., Hsiao, T.-L., Sauer, U., & Vitkup, D. (2012). Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nature Chemical Biology*, *8*(10), 848-854. https://doi.org/10.1038/nchembio.1063

Prigent, S., Frioux, C., Dittami, S. M., Thiele, S., Larhlimi, A., Collet, G., Gutknecht, F., Got, J., Eveillard, D., Bourdon, J., Plewniak, F., Tonon, T., & Siegel, A. (2017). Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks (C. Kaleta, Éd.). *PLOS Computational Biology*, *13*(1), e1005276. https://doi.org/10.1371/journal.pcbi.1005276

Provasoli, L., & Pintner, I. J. (1980). Bacteria induced polymorphism in an axenic laboratory strain of Ulva lactuca (Chlorophyceae). *Journal of Phycology*, *16*(2), 196-201. https://doi.org/10.1111/j.1529-8817.1980.tb03019.x

Reed, J. L., Vo, T. D., Schilling, C. H., & Palsson, B. O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome biology*, *4*(9), R54. https://doi.org/10.1186/gb-2003-4-9-r54

Satish Kumar, V., Dasika, M. S., & Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC bioinformatics*, *8*, 212. https://doi.org/10.1186/1471-2105-8-212

Schaub, T., & Thiele, S. (2009). Metabolic network expansion with answer set programming. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5649 LNCS*, 312-326. https://doi.org/10.1007/978-3-642-02846-5_27

Schellenberger, J., Lewis, N. E., & Palsson, B. Ø. (2011). Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical journal*, *100*(3), 544-53. https://doi.org/10.1016/j.bpj.2010.12.3707

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., & Palsson, B. O. (2011). Quantitative prediction of cellular metabolism with constraint-based models : the COBRA Toolbox v2.0. *Nature protocols*, *6*(9), 1290-307. https://doi.org/10.1038/nprot.2011.308

Segata, N. (2018). On the Road to Strain-Resolved Comparative Metagenomics. *mSystems*, *3*(2), e00190-17. https://doi.org/10.1128/mSystems.00190-17

Steffensen, J. L., Dufault-Thompson, K., & Zhang, Y. (2016). PSAMM : A Portable System for the Analysis of Metabolic Models (T. Dandekar,

Éd.). *PLoS Computational Biology*, *12*(2), e1004732. https://doi.org/10.1371/journal.pcbi.1004732

Succurro, A., & Ebenhöh, O. (2018). Review and perspective on mathematical modeling of microbial ecosystems. *Biochemical Society Transactions*, *46*(2), 403-412. https://doi.org/10.1042/BST20170265

Tapia, J. E., González, B., Goulitquer, S., Potin, P., & Correa, J. A. (2016). Microbiota Influences Morphology and Reproduction of the Brown Alga Ectocarpus sp. *Frontiers in Microbiology*, *7*(FEB), 197. https://doi.org/10.3389/fmicb.2016.00197

Thacker, R. W., & Freeman, C. J. (2012). Sponge-Microbe Symbioses. Recent Advances and New Directions. *Advances in Marine Biology*, *62*, 57-111. https://doi.org/10.1016/B978-0-12-394283-8.00002-3

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, *5*(1), 93-121. https://doi.org/10.1038/nprot.2009.203

Thiele, I., Vlassis, N., & Fleming, R. M. T. (2014). fastGapFill : efficient gap filling in metabolic networks. *Bioinformatics (Oxford, England)*, *30*(17), 2529-2531. https://doi.org/10.1093/bioinformatics/btu321

van der Ark, K. C. H., van Heck, R. G. A., Martins Dos Santos, V. A. P., Belzer, C., & de Vos, W. M. (2017). More than just a gut feeling : constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes. *Microbiome*, *5*(1), 78. https://doi.org/10.1186/s40168-017-0299-x

Vijayakumar, S., Conway, M., Lió, P., & Angione, C. (2017). Seeing the wood for the trees : a forest of methods for optimization and omic-network integration in metabolic modelling. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbx053

Wang, H., Marcišauskas, S., Sánchez, B. J., Domenzain, I., Hermansson, D., Agren, R., Nielsen, J., & Kerkhoven, E. J. (2018). RAVEN 2.0 : A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor (C. A. Ouzounis, Éd.). *PLOS Computational Biology*, *14*(10), e1006541. https://doi.org/10.1371/journal.pcbi.1006541

Webster, N. S. (2014). Cooperation, communication, and co-evolution : Grand challenges in microbial symbiosis research. *Frontiers in Microbiology*, *5*(APR), 164. https://doi.org/10.3389/fmicb.2014.00164

Werren, J. H., Baldo, L., & Clark, M. E. (2008). Wolbachia : Master manipulators of invertebrate biology. https://doi.org/10.1038/nrmicro1969

Xavier, J. C., Patil, K. R., & Rocha, I. (2017). Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. *Metabolic engineering*, *39*, 200-208. https://doi.org/10.1016/j.ymben.2016.12.002

Xu, N., Ye, C., Chen, X., Liu, J., & Liu, L. (2017). Genome-scale metabolic modelling common cofactors metabolism in microorganisms. *Journal of Biotechnology*, *251*, 1-13. https://doi.org/10.1016/j.jbiotec.2017.04.004

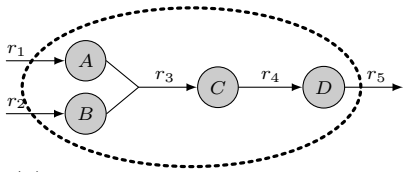
Xu, Y., & Zhao, F. (2018). Single-cell metagenomics : challenges and applications. *Protein & Cell*, *9*(5), 501-510. https://doi.org/10.1007/s13238-018-0544-5

Zarecki, R., Oberhardt, M. A., Reshef, L., Gophna, U., & Ruppin, E. (2014). A novel nutritional predictor links microbial fastidiousness with lowered ubiquity, growth rate, and cooperativeness. *PLoS computational biology*, *10*(7), e1003726. https://doi.org/10.1371/journal.pcbi.1003726

Zomorrodi, A. R., Islam, M. M., & Maranas, C. D. (2014). d-OptCom : Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synthetic Biology*, *3*(4), 247-257. https://doi.org/10.1021/sb4001307

Zomorrodi, A. R., & Maranas, C. D. (2012). OptCom : A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Computational Biology*, *8*(2), e1002363. https://doi.org/10.1371/journal.pcbi.1002363

Zuñiga, C., Zaramela, L., & Zengler, K. (2017). Elucidation of complexity and prediction of interactions in microbial communities. *Microbial Biotechnology*, *10*(6), 1500-1522. https://doi.org/10.1111/1751-7915.12855



(a) *Graphe du réseau métabolique*

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{ccccc} r_1 & r_2 & r_3 & r_4 & r_5 \end{array} \\ \begin{pmatrix} +1 & 0 & -1 & 0 & 0 \\ 0 & +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 \\ 0 & 0 & 0 & +1 & -1 \end{pmatrix} \end{array} = \mathbf{S}$$

(b) *Matrice du réseau métabolique*

$$\begin{aligned} \frac{d[A]}{dt} &= 0 \quad \Leftrightarrow 1 \times \mathbf{v}_{r_1} - 1 \times \mathbf{v}_{r_3} = 0 \\ \frac{d[B]}{dt} &= 0 \quad \Leftrightarrow 1 \times \mathbf{v}_{r_2} - 1 \times \mathbf{v}_{r_3} = 0 \\ \frac{d[C]}{dt} &= 0 \quad \Leftrightarrow 1 \times \mathbf{v}_{r_3} - 1 \times \mathbf{v}_{r_4} = 0 \\ \frac{d[D]}{dt} &= 0 \quad \Leftrightarrow 1 \times \mathbf{v}_{r_4} - 1 \times \mathbf{v}_{r_5} = 0 \end{aligned}$$

(c) *Équations vérifiées par les flux à l'état stationnaire*

$$\mathbf{S} \begin{pmatrix} \mathbf{v}_{r_1} \\ \mathbf{v}_{r_2} \\ \mathbf{v}_{r_3} \\ \mathbf{v}_{r_4} \\ \mathbf{v}_{r_5} \end{pmatrix} = 0$$

(d) *Définition matricielle d'un état stationnaire*

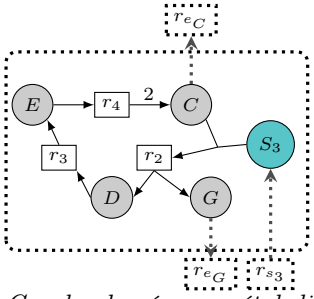
$$\begin{aligned} \text{maximiser} \quad & z = \mathbf{v}_{r_4} \\ \text{tel que} \quad & 0 \leq \mathbf{v}_{r_i} \leq 2 \\ & \mathbf{Sv} = \mathbf{0} \end{aligned}$$

(e) *Problème d'optimisation à résoudre*

$$\begin{pmatrix} \mathbf{v}_{r_1} \\ \mathbf{v}_{r_2} \\ \mathbf{v}_{r_3} \\ \mathbf{v}_{r_4} \\ \mathbf{v}_{r_5} \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{pmatrix}$$

(f) *Problème d'optimisation à résoudre*

FIGURE 5 – **Application de l'analyse de flux à un exemple de réseaux métaboliques** (adapté de M. K. KIM et LUN, 2014). Le système consiste en cinq réactions. Les réactions r_1 et r_2 représentent la production de A et B, les deux graines du système importées depuis le milieu de culture. La réaction r_5 permet d'empêcher l'accumulation du métabolite D. La matrice stœchiométrique associée peut être construite à partir du graphe du réseau en supposant une stœchiométrie égale à 1. L'objectif est de maximiser le flux dans la réaction r_4 produisant D. La résolution de l'équation $\mathbf{Sv} = 0$ représentant l'état stationnaire, en prenant en compte les bornes des flux des réactions, donne une unique solution au système : tous les flux à 2 mmol$\times$gDW$^{-1}\times$hr^{-1}. Le flux dans la réaction r_4 étant positif, cette réaction est activée selon la sémantique de flux.



(a) *Graphe du réseau métabolique*

$$\begin{array}{c} \\ S_2 \\ C \\ D \\ E \\ G \end{array} \begin{array}{c} \begin{array}{cccccc} r_{s_3} & r_2 & r_3 & r_4 & r_{e_C} & r_{e_G} \end{array} \\ \begin{pmatrix} +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & +2 & -1 & 0 \\ 0 & +1 & -1 & 0 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 & 0 \\ 0 & +1 & 0 & 0 & 0 & -1 \end{pmatrix} \end{array} = \mathbf{S}$$

(b) *Matrice du réseau métabolique*

$$\begin{aligned} \frac{d[S_2]}{dt} &= 0 \quad \Leftrightarrow \mathbf{v}_{r_{S_2}} - \mathbf{v}_{r_2} = 0 \\ \frac{d[C]}{dt} &= 0 \quad \Leftrightarrow 2\mathbf{v}_{r_4} - \mathbf{v}_{r_2} - \mathbf{v}_{r_e} = 0 \\ \frac{d[D]}{dt} &= 0 \quad \Leftrightarrow \mathbf{v}_{r_2} - \mathbf{v}_{r_3} = 0 \\ \frac{d[E]}{dt} &= 0 \quad \Leftrightarrow \mathbf{v}_{r_3} - \mathbf{v}_{r_4} = 0. \\ \frac{d[G]}{dt} &= 0 \quad \Leftrightarrow \mathbf{v}_{r_2} - \mathbf{v}_{r_{e_G}} = 0. \end{aligned}$$

(c) *Equations vérifiées par les flux à l'état stationnaire*

$$\mathbf{S} \begin{pmatrix} \mathbf{v}_{rs_2} \\ \mathbf{v}_{r_2} \\ \mathbf{v}_{r_3} \\ \mathbf{v}_{r_4} \\ \mathbf{v}_{r_e} \\ \mathbf{v}_{r_{e_G}} \end{pmatrix} = \mathbf{v}$$

(d) *Vecteur des flux de réactions*

FIGURE 6 – **Impact des sémantiques sur la production des composés dans un cycle**. Le cycle est extrait du réseau métabolique de la figure 4. À l'état stationnaire, la présence de la graine S_2 et la production de C au sein du cycle sont suffisantes pour considérer ce dernier comme activable. Néanmoins, il n'est pas activé topologiquement.

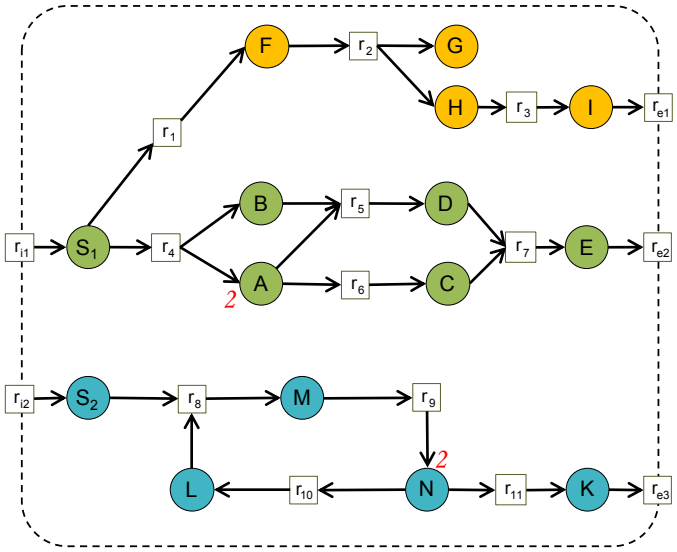


FIGURE 7 – Différences entre les conditions d'activation booléenne (AB) et stœchiométrique (AS). Le sous-réseau jaune est activé à l'échelle booléenne et inactif à l'échelle stœchiométrique en raison de l'accumulation du composé G. Le sous-réseau vert est actif selon les deux conditions. Néanmoins, une variation du coefficient stœchiométrique associé à A (importance du ratio A/B) peut rendre le sous-réseau inactif selon les contraintes stoechiométriques, brisant l'équilibre stœchiométrique attendu à l'état stationnaire. Enfin, le sous-réseau bleu est uniquement activé sous la condition stoechiométrique, la sémantique booléenne nécessitant une activation à partir du milieu de culture de tous les substrats d'une réaction pour qu'elle s'active.

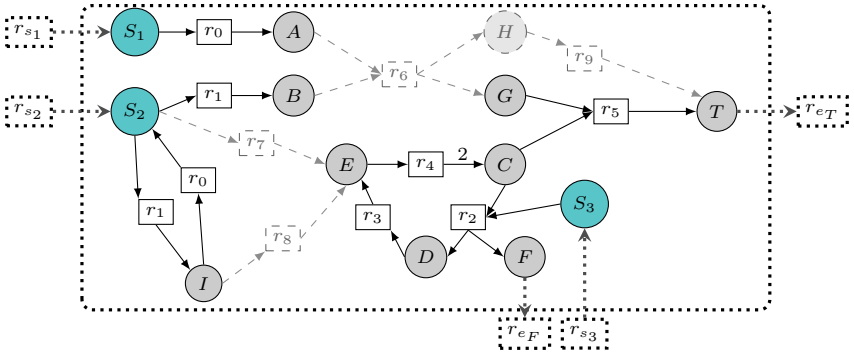


FIGURE 8 – Exemple de complétion hybride du métabolisme. Adapté de FRIOUX et al., 2017 et FRIOUX et al., 2019. Deux solutions de complétions existent $\{r_6, r_7, r_9\}$ et $\{r_6, r_8, r_9\}$. Le problème est également solvable avec les sémantiques individuelles : topologique ou stœchiométrique. Dans le premier cas, r_7 ou r_8 (initiant le cycle) est nécessaire avec r_6. Dans le second cas, r_6 et r_9 (empêchant l'accumulation de H) forment la solution. La solution hybride satisfait les contraintes des deux sémantiques.