L'allocation de ressources dans les systèmes biologiques Un cadre de modélisation pour comprendre et prédire le fonctionnement des organismes vivants

Anne Goelzer

Année 2022

UNIVERSITÉ PARIS-SACLAY

Mémoire présenté pour l'obtention d'une

HABILITATION À DIRIGER DES RECHERCHES

Spécialité : Automatique/Biologie des systèmes

par

Anne GOELZER

L'allocation de ressources dans les systèmes biologiques

Un cadre de modélisation pour comprendre et prédire le fonctionnement des organismes vivants

---

Soutenue le 27 octobre 2022 à l'Université Paris-Saclay

JURY

Rapporteur	M. Paul-Henry Cournède	Professeur des Universités, CentraleSupélec
Rapporteur	M. Jean-Luc Gouzé	Directeur de recherche, INRIA
Rapporteur	M. Gilles Curien	Chargé de recherche, CNRS
Examinatrice	Mme. Stéphanie Heux	Directrice de recherche, INRAE
Examinateur	M. Bertrand Dubreucq	Directeur de recherche, INRAE
Examinatrice	Mme. Anne Siegel	Directrice de recherche, CNRS

Table des matières

Chapitre 1

Introduction

Tout être vivant est constitué d'un très grand nombre de composants en interaction garantissant la croissance et la survie de l'organisme, et ce quel que soit l'organisme (bactérie, plante, mammifère, etc). Les organismes supérieurs comme les plantes ou les mammifères sont composés de milliards de cellules spécialisées, organisés en organes et accomplissant des fonctions spécifiques au sein de l'organisme (cf figure 1.1).

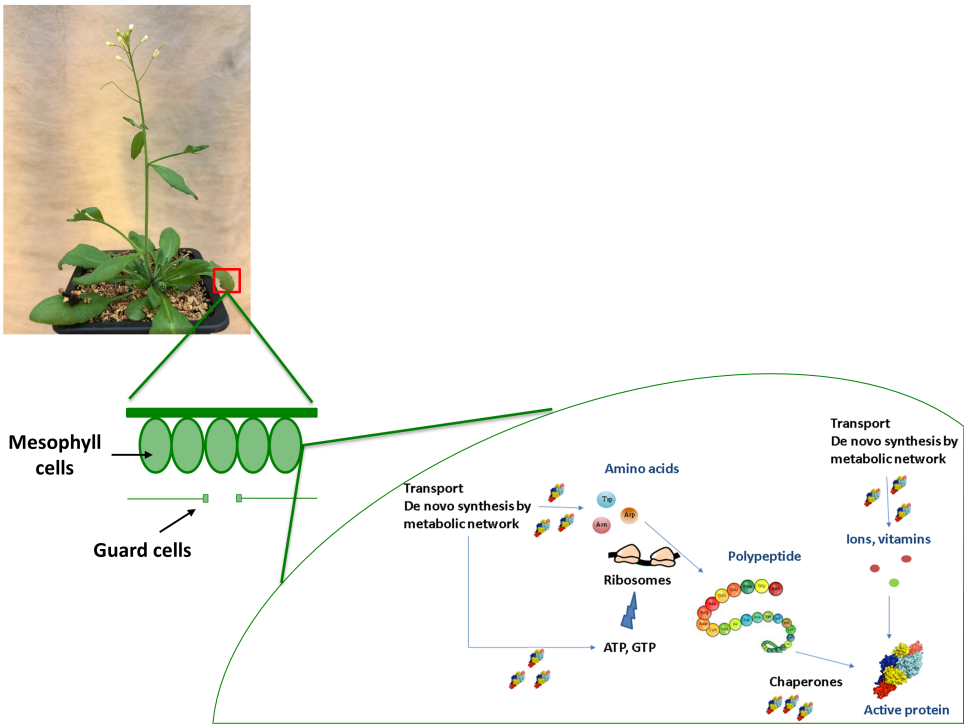

FIGURE 1.1 – **La plante modèle** *Arabidopsis thaliana*

Les microorganismes peuvent s'organiser en communautés de cellules (e.g biofilm bactérien), ce qui leur confère une meilleure résistance vis à vis de compétiteurs mais aussi de substances chimiques [34].
On sait par ailleurs que les décisions d'adaptation d'un organisme à un changement environnemental s'opèrent à l'échelle cellulaire. La cellule, procaryote ou eucaryote reste donc clé dans la

compréhension du fonctionnement global de l'organisme (ou de communautés d'organismes) et de son adaptation à des perturbations. Mieux comprendre son fonctionnement, arriver à intégrer et réconcilier l'ensemble des données acquises (ou à acquérir dans le futur) avec la connaissance des mécanismes moléculaires reste une étape incontournable et un enjeu majeur en biologie.

Depuis les années 1990 et les premiers séquençages d'ADN complets, on assiste à une génération massive de données aux échelles cellulaires. Ces données (par ex. transcriptome, protéome, métabolome, etc.) permettent d'observer l'état interne (ARNm, protéine, métabolite) d'une population de cellules, voire même à l'échelle d'une seule cellule (le *single-cell*). L'accès à ces données a ainsi profondément changé les pratiques des biologistes. Les approches dites réductionnistes, i.e. centrées sur l'étude d'un mécanisme moléculaire particulier en déployant diverses techniques expérimentales, ont peu à peu laissé place à des approches globales où le comportement de la (population de) cellule(s) est analysé [1]. Ce changement d'échelle et les difficultés associées ont contribué à l'émergence d'un nouveau domaine de recherche dans les années 2000, la **biologie des systèmes** [18]. La biologie des systèmes appelle les sciences de l'ingénieur, et en particulier l'automatique, appliquées à la biologie. Il s'agit d'importer le savoir-faire et les techniques issus des sciences de l'ingénieur afin de rationaliser, d'analyser, et prédire le comportement des systèmes biologiques.

Remarque. Dans la littérature, la biologie des systèmes ne désigne pas toujours les mêmes concepts, méthodes et outils selon les communautés. Brièvement, les approches de biologie des systèmes sont souvent regroupées en approches dites *top-down* ou *bottom-up* (voir [5, 13] et les références associées). Les approches *top-down* correspondent (généralement) à des méthodes statistiques de fouilles de données, d'inférence de réseaux et ont donc, comme point de départ, les données omiques. Il s'agit de rechercher des corrélation entre concentrations de molécules afin de formuler de nouvelles hypothèses sur la co-régulation de groupes de molécules, ou sur l'identification d'un biomarqueur représentatif d'une condition spécifique, pour citer quelques exemples d'application. Les approches *bottom-up* se basent sur des modèles (déterministes, stochastiques, etc.) de sous-parties de la cellule, construits sur la base de connaissances biologiques. L'objectif est d'utiliser le modèle afin de mieux comprendre le fonctionnement du (sous)-système biologique, et à terme de combiner les modèles pour aller jusqu'à la prédiction du comportement de la cellule entière. Nous verrons, au cours de cette habilitation à diriger des recherches, que cet objectif est maintenant atteignable dans une certaine mesure ([16], A.11*). L'idée ici n'est pas d'introduire l'ensemble des méthodes *top-down* ou *bottom-up*, mais de souligner que cette distinction existe, et qu'elle est d'autant plus clivante, que ces méthodes sont portées par des communautées différentes. Les méthodes *top-down* sont portées généralement par les communautés des statisticiens et bioinformaticiens, alors que les méthodes *bottom-up* sont portées par les communautés des sciences de l'ingénieur (automatique, analyse des systèmes dynamiques, etc.). Il en résulte souvent des pratiques scientifiques différentes et des difficultés à dialoguer entre communautés, bien que l'objet d'intérêt sous-jacent soit le même (la cellule). Etant ingénieure en Automatique de formation, je me suis naturellement tournée vers les approches *bottom-up* pour modéliser, analyser et prédire le fonctionnement des systèmes biologiques.

1. Dans cette habilitation à diriger des recherches, sauf indication contraire, les données omiques manipulées seront toujours issues d'une population de cellules. Par abus de language et de notation, nous parlerons de *cellule*, mais il s'agira bien toujours d'une population de cellules.

1.1 Modéliser le comportement des systèmes biologiques

A la différence des systèmes industriels, les systèmes vivants n'ont pas été conçus par les hommes. De ce fait, les composants individuels d'une cellule, leurs interactions et leurs réponses à des perturbations environnementales ne sont pas connus, mais identifiés progressivement. On citera dans ce contexte les deux ouvrages de référence sur les bactéries modèles Gram+ *Bacillus subtilis* [37, 38] et Gram- *Escherichia coli* [28] compilant l'état de la connaissance sur ces bactéries sur plus de 30 années de recherche. Regrouper les connaissances (ou une partie de ces connaissances) au sein d'un modèle mathématique permet ainsi de vérifier la cohérence des informations, d'analyser des jeux de données expérimentaux au regard de la connaissance établie, et d'anticiper l'impact de variations environnementales ou internes (e.g. perturbations génétiques) sur le comportement du système biologique. Il s'agit bien d'aller vers des modèles permettant d'analyser le comportement fin des organismes vivants et possédant de grandes capacités de prédiction.

1.1.1 Un compromis entre prédictibilité, complexité et tractabilité

Une cellule aussi simple qu'une bactérie est composée de millions d'entités en constante interaction. Une première idée naïve serait de simuler l'ensemble des évènements apparaissant dans une cellule entière. Or le nombre d'évènements aléatoires au sein d'une cellule entière au cours d'un seul cycle de réplication est de l'ordre de 10^{14-16}, ce qui est impossible à simuler en tant que tel. Dès lors, des simplifications sont nécessaires. A titre d'exemple, on citera le premier simulateur d'une cellule entière pour la bactérie *Mycoplasma genitalium* [16], qui intègre un module déterministe pour simuler la fonction métabolique, ce qui décroit fortement le nombre d'évènements cellulaires à simuler, et des modules stochastiques pour simuler les autres fonctions cellulaires comme la traduction. Ce simulateur constitue une incontestable *milestone* dans la modélisation et la simulation d'une cellule, essentiellement au niveau de la faisabilité opérationnelle, i.e. la capacité à intégrer la connaissance disponible sous forme de modules de simulation et à simuler un très grand nombre d'évènements se passant à différentes échelles cellulaires. Du point de la simulation en elle-même, de nombreux facteurs limitent l'utilisation de ce simulateur : (1) une partie de la configuration cellulaire est fixée a priori, ce qui limite le caractère autonome du simulateur ; (2) le temps de calcul reste important ; (3) la généricité du simulateur est limitée, certaines fonctions cellulaires étant codées en dur dans le simulateur ; (4) enfin, le génome de *M. genitalium* comporte seulement 517 gènes, ce qui reste très loin des 4200 gènes des bactéries modèles *B. subtilis* ou *E. coli*. L'adaptation de ce simulateur pour une bactérie modèle contournant les limitations constatées reste toujours une question d'actualité 10 ans après.
L'exemple de ce simulateur est ainsi illustratif des enjeux et fronts de recherche importants en biologie des systèmes, et plus généralement en biologie prédictive :

- le développement d'un cadre de modélisation flexible, permettant de gérer le compromis entre prédictibilité, complexité et tractabilité [2] du modèle
- l'autonomie des modèles mathématiques développés, en particulier pour les méthodes de modélisation dite *sous contraintes* [32] (voir section 1.1.3, dont l'objectif à terme est de modéliser une grande partie des processus cellulaires [A.17]

2. On définira le terme tractabilité, comme la capacité à simuler un modèle mathématique en un temps raisonnable.

– la représentation et l'encodage d'informations biologiques hétérogènes pour différentes communautés de scientifiques (biologistes, bioformaticiens/biostatisticiens et modélisateurs pour n'en citer que quelques unes) et pour élaborer des modèles prédictifs
– le transfert des modèles mathématiques élaborés sur des espèces biologiques modèles (comme *B. subtilis* ou *A. thaliana*) vers des organismes non modèles ou d'intérêt agronomique (ex : la bactérie *Ralsonia solanacearum* ou la plante *Zea mays*).

Bien entendu, ceci ne peut être réalisé sans connaissance et analyse fine du comportement cellulaire, afin d'identifier des principes de fonctionnement génériques, et donc communs à plusieurs (voire tous les) organismes vivants. Ainsi, pour des raisons liées à la tractabilité du modèle final, ainsi que dans l'objectif à terme de représenter un très grand nombre d'entités biologiques, je me suis tournée vers des modèles déterministes, où on considèrera par exemple des concentrations moyennes de molécules dans une population, plutôt que des modèles stochastiques qui peuvent tenir compte par exemple de la variabilité au sein d'une population, mais sont plus coûteux en terme de temps de calcul et de simulation. Dans ce contexte et dans le cadre de cette HDR, nous nous intéresserons plus particulièrement aux modèles dynamiques basés sur les équations différentielles ordinaires (EDO) et aux modèles sous contraintes appliqués aux échelles cellulaires.

1.1.2 Modèles EDO

Les modèles EDO aux échelles cellulaires ont été plus particulièrement utilisés pour modéliser et simuler le comportement de fonctions biologiques spécifiques au cours du temps. Ces modèles s'écrivent de façon standard comme :

$$\begin{cases} \dfrac{dx(t)}{dt} &= f(p, x(t), t) \\ x(t = t_0) &= x_0 \end{cases}$$

où t désigne le temps, $x \stackrel{\Delta}{=} (x_1, \ldots, x_n)^T$ désigne le vecteur d'états du système, correspondant généralement aux concentrations des molécules comme les métabolites, les ARNm, les complexes enzymatiques, etc. t_0 et x_0 désignent respectivement le temps initial et le vecteur d'états initial. Le vecteur p contient les valeurs des paramètres du système, généralement supposé indépendent du temps. Dans le cas de la modélisation de voie métaboliques, le vecteur p contiendra par exemple les paramètres associés aux cinétiques enzymatiques. Enfin la fonction f définie de $\mathbb{R}^n$ dans $\mathbb{R}^n$ décrit la dynamique du système. La fonction f est généralement non linéaire en fonction de l'état et des paramètres du système.
Les modèles à base d'EDO permettent ainsi d'obtenir la cinétique des concentrations des molécules, sous réserve de connaître la condition initiale x_0, et le vecteur de paramètre p. Néanmoins, la dimension du système d'équations différentielles, même pour une seule voie métabolique, peut rapidement être importante. On citera par exemple les modèles EDO de la glycolyse [6] (18 états, 126 paramètres) ou du métabolisme carboné central d'*E. coli* [20] (47 états, 193 paramètres). En pratique, obtenir les valeurs quantitatives des paramètres nécessite un effort expérimental important. Ceci explique pourquoi les modèles EDOs ont été développés et validés principalement pour des systèmes de petite taille. Du point de vue théorique, du fait de la forte non-linéarité du système, la résolution analytique du système EDO est impossible. En simulation, on constate que, pour des valeurs de paramètres et de conditions initiales réalistes, le système semble avoir de bonnes propriétés : la simulation converge vers un état d'équilibre stable. Néanmoins, l'analyse théorique du comportement du système reste difficile et nécessite souvent des hypothèses

simplificatrices comme le passage à l'équilibre, ou la linéarisation du système autour d'un point de fonctionnement.

1.1.3 Modèles sous contraintes

Développé initialement dans les années 90 [41], la modélisation dite *sous contraintes* formalise les échanges de masse dans la cellule par des contraintes d'égalités et d'inégalités. La loi de conservation de la masse s'exprime, pour le réseau métabolique, à travers la matrice de stœchiométrie Ω de la façon suivante :

$$\Omega\nu = 0,$$

où $\nu \stackrel{\Delta}{=} (\nu_1, \ldots, \nu_m)^T$ désigne le vecteur des flux métaboliques associés aux réactions chimiques composant l'ensemble du réseau métabolique en régime établi. La matrice Ω se déduit en remarquant que l'évolution de la concentration du i-ème métabolite x_i s'écrit en régime établi comme la somme des flux de production moins la somme des flux de consommation de ce métabolite :

$$\frac{dx_i}{dt} = \nu_i - \nu_{i+1} = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \nu_i \\ \nu_{i+1} \end{bmatrix} = 0,$$

en considérant un seul flux de production ν_i et un seul flux de consommation ν_{i+1}. En décrivant l'évolution de l'ensemble des métabolites, on obtient $\Omega\nu = 0$. Cette contrainte d'égalité ainsi que la limitation des flux d'import de nutriments sont au coeur des méthodes de modélisation sous contraintes comme la méthode Flux Balance Analysis [41]. Ces méthodes formalisent le calcul des flux métaboliques et du taux de croissance bactérien comme un problème d'optimisation sous contraintes linéaires. Typiquement la méthode Flux Balance Analysis s'écrit comme le problème de Programmation Linéaire (LP) suivant, et noté P_{fba} par la suite :

$$\begin{aligned} \text{Maximiser} \quad & c^T\nu \\ \nu \in \mathbb{R}^m & \\ \text{tels que :} & \\ & \Omega\nu = 0 \\ & \alpha \leq \nu \leq \beta \end{aligned}$$

où α et β sont des vecteurs de taille adéquate représentant les bornes inférieures et supérieures des flux métaboliques (ou d'un sous-ensemble de flux métaboliques), et c est le vecteur de la fonction objectif. Du point de vue pratique, la formation de la biomasse cellulaire est décrite par une unique réaction chimique agrégée où les substrats et les produits de la réaction représentent l'ensemble des composés moléculaires (e.g. les acides aminés, les (déoxy)-nucléotides, etc.) consommés ou produits lors de la formation d'un gramme de biomasse. L'avantage du problème P_{fba} réside en sa formulation sous forme d'un problème LP, ce qui permet de le résoudre efficacement du point de vue numérique, même pour des grandes dimensions [29, 4]. Les méthodes sous-contraintes peuvent intégrer la globalité des voies métaboliques, et permettent donc de travailler à l'échelle du génome entier.
On notera deux grandes limitations au problème P_{fba} : (1) le choix de la fonction objectif, (2) le fait de devoir fixer des bornes α, β sur les flux pour obtenir une solution finie. De plus, certains flux doivent être imposés ou limités par des valeurs expérimentales pour obtenir une solution quantitative correcte. Au delà de l'effort expérimental nécessaire à la mesure des flux (à chaque fois), imposer des bornes notamment sur les flux d'entrée revient à imposer une partie de la configuration cellulaire, ce qui réduit le caractère autonome du modèle.

Malgré ces limitations, la faible complexité algorithmique du problème P_{fba}, ainsi que le faible nombre de paramètres comparé aux modèles EDOs expliquent pourquoi les méthodes sous contraintes ont été largement étudiées et appliquées pour étudier les configurations cellulaires d'organismes procaryotes et eucaryotes [3]. Il existe également de nombreuses extensions théoriques au cadre de modélisation initial, dont nous ne discuterons pas dans le périmètre de cette HDR, à l'exception de celles discutées dans le chapitre 3. Pour plus de détails sur les autres extensions ou sur les cas d'utilisation de la méthode FBA en ingénierie métabolique notamment, le lecteur pourra se réferer aux articles [33, 35] et aux références citées dans ces articles.

1.2 Problématique de recherche, contributions et organisation du manuscrit

Mes objectifs scientifiques sont d'une part de développer des méthodes et des outils de simulation afin de mieux comprendre et de prédire le fonctionnement des organismes vivants et leur adaptation à des variations environnementales, et d'autre part de développer un nouveau cadre de description des organismes vivants basé sur les approches systèmes, susceptible de mieux structurer la connaissance et les données biologiques, et à terme de faciliter les échanges entre communautés scientifiques. Se faisant, mes activités de recherche ont contribué directement à chacun des quatre enjeux en biologie des systèmes présentés en section 1.1.1. J'ai rassemblé mes principaux travaux de recherche en quatre chapitres :

1. la modélisation, analyse et simulation de (sous)-systèmes cellulaires (Chapitre 2)
2. l'allocation des ressources comme principe de design des cellules et comme outil de prédiction du phénotype cellulaire (Chapitre 3)
3. l'allocation de ressources en régime dynamique (Chapitre 4)
4. l'intégration systémique des données et des connaissances biologiques (Chapitre 5)

Je présenterai enfin dans le chapitre **??** mes perspectives de recherche, notamment mon projet de recherche principal autour de la prédiction du comportement des plantes en conditions environnementales complexes.

Le chapitre 2 correspond à mes travaux les plus anciens, et porte sur le développement et l'analyse de modèles dynamiques détaillés pour étudier le fonctionnement de sous-systèmes biologiques. Au cours de ces travaux, j'ai analysé l'organisation et le fonctionnement de grandes fonctions cellulaires comme le metabolisme [A.1*, B.2*], la réponse au stress oxydatif et à la carence en ions [R.1] ou la traduction des protéines [A.13*]. Ma contribution principale a été d'identifier trois types de modules de fonctionnement dans les voies métaboliques, de les analyser mathématiquement afin d'identifier certaines propriétés remarquables (i.e. entre autres unicité du point d'équilibre, découplage vis à vis de cofacteurs), et finalement de montrer que l'ensemble du réseau métabolique d'une bactérie comme *Bacillus subtilis* peut se décomposer en modules fonctionnels.

Les chapitres 3 et 4 présentent un cadre de modélisation flexible, générique et adéquat pour décrire le fonctionnement des cellules vivantes (procaryotes et eucaryotes), **la méthode de**

3. Le site https ://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms centralise une grande partie des réseaux métaboliques actuellement reconstruits, et permettant ainsi de construire un modèle FBA de l'organisme.

modélisation sous contraintes appelée *Resource Balance Analysis* (RBA). La méthode RBA contourne les limitations majeures de la méthode FBA (cf section précédente), et fournit pour la première fois **un modèle autonome du comportement cellulaire à l'échelle du génome en conditions environnementales complexes**. Le développement du cadre théorique sur les bactéries en régime permanent [C.1, A.3*] et sa validation biologique sur la bactérie *B. subtilis* [A.11*], de l'outil logiciel permettant de générer automatiquement des modèles RBA pour les bactéries [A.18*], ainsi que les extensions du cadre en régime dynamique [A.22*] ou pour les cellules eucaryotes [A.19] sont sans conteste mes contributions scientifiques les plus importantes. De plus, mes travaux préliminaires présentés en perspective de cette HDR (chapitre **??**) montrent que le cadre RBA est aussi adéquat et prometteur pour réconcilier différentes échelles (du gène à l'individu) dans les modèles de plante, et ainsi aller vers des modèles plus autonomes et capables de **prédire le comportement de la plante entière en conditions environnementales complexes** (e.g. en conditions de combinaison de stress).

Le chapitre 5 présente des travaux génériques et transversaux en représentation des connaissances, pour la description des organismes vivants. Les principales contributions de ce chapitre sont le **développement d'ontologies basées sur la décomposition systémique des cellules** [A.16*, A.20] proposée dans ma thèse (et à la base du développement de la méthode RBA). L'originalité de ces ontologies par rapport à l'existant est de lier entités moléculaires, processus cellulaires et modèles mathématiques décrivant le fonctionnement des processus cellulaires, et d'inférer de nouvelles connaissances par raisonnement automatique sur les instances. Ce cadre est très prometteur pour l'intégration de données et de connaissance hétérogènes en biologie. L'ambition est que chaque entité moléculaire mesurable ou chaque donnée déduite d'une mesure (e.g. une durée de vie d'une protéine) puisse être décrite et liée formellement à un objet biologique. Ces ontologies permettront à terme de mieux structurer les entrepôts de données biologiques hétérogènes, et donc d'alimenter les modèles mathématiques développés en biologie des systèmes, notamment les modèles RBA pour ne citer que ceux-là.

Les articles A.1*, B.2*, A.8*, A.13*, A.3*, A.11*, C.9*, A.18*, A.23*, A.9*, A.10* et A.16* fournis en annexe B de ce manuscrit illustrent mes contributions principales dans ces quatre axes de recherche.

Notations utilisées. A^T correspond à la matrice transposée de A. $\mathbb{R}^n_{>0} \triangleq \left\{x \in \mathbb{R}^n \,|x_i > 0 \text{ pour tout } i \in \{1, \cdots, n\}\right\}$, $\mathbb{R}_{>0} \triangleq \mathbb{R}^1_{>0}$, $\mathbb{R}^n_{\geq 0} \triangleq \{x \in \mathbb{R}^n \,|\, x_i \geq 0 \text{ pour tout } i \in \{1, \cdots, n\}\}$ et $\mathbb{R}_{\geq 0} \triangleq \mathbb{R}^1_{\geq 0}$.

Chapitre 2

Modélisation, analyse et simulation de sous-systèmes cellulaires

Les travaux présentés dans ce chapitre sont les plus anciens, et pour certains d'entre eux, ont été initiés avant le début de ma thèse en 2008. Ils ont constitué un socle de connaissances solide sur le fonctionnement des processus cellulaires bactériens, sur lequel je me suis appuyée par la suite pour analyser les résultats de prédiction de mes modèles (voir chapitre 3). Il s'agissait notamment de la reconstruction des voies métaboliques de la bactérie *Bacillus subtilis* et du réseau de régulation associé présentés dans la prochaine section. N'étant pas biologiste de formation, la période 2005-2008 [1] m'a permis de me former à la biologie à travers la lecture d'articles (plus de 400 articles en microbiologie, biologie moléculaire, biochimie, génétique/génomique pour reconstruire les réseaux) et de développer mes premiers modèles mathématiques, principalement basés sur des EDOs, afin d'étudier certaines fonctions cellulaires comme des systèmes contrôlés. A l'exception de l'étude des voies métaboliques faite au cours de ma thèse, le choix des autres fonctions cellulaires étudiées a été motivé par les projets européens interdisciplinaires (BaSysBio, BaSynthec) auxquels je participais. Celles-ci sont néanmoins représentatives des processus cellulaires impliqués dans la croissance (métabolisme, traduction des protéines) et la survie (réponse au stress oxydatif) des cellules vivantes.

Les travaux présentés dans ce chapitre ont été menés en collaboration avec V. Fromion (MaIAGE), et m'ont permis de co-encadrer trois thèses interdisciplinaires mélant microbiologie et modélisation : M. Celton [T.2] en collaboration avec l'UMR SPO (S. Dequin, C. Camarasa), O. Borkowski [T.3] en collaboration avec l'institut MICALIS (M. Jules, S. Aymerich), C. Cousin [T.4] en collaboration avec l'institut MICALIS (I. Mijakovic). Les trois étudiants étaient microbiologistes de formation. Je les ai encadrés principalement sur les aspects liés au développement, à l'analyse, à la calibration des modèles mathématiques développés au cours de leur thèse, ainsi que sur le traitement et l'analyse de leurs données expérimentales. Ces premières expériences d'encadrement de thèses m'ont permis d'acquérir des connaissances sur les dispositifs expérimentaux, à dialoguer avec une communauté très différente de la mienne, et à former des étudiants aux approches de biologie des systèmes. En parallèle, j'ai aussi pu constater la difficulté à concevoir l'ensemble d'expériences adéquates pour démontrer un phénomène biologique, ainsi que le temps nécessaire aux étudiants à apprendre leur métier de biologiste.

1. 2008 est l'année du début de thèse et de la publication [A.1*]

2.1 Identification de modules fonctionnels dans les voies métaboliques

A partir de 2005, l'objectif de mes travaux était de mieux comprendre l'organisation des cellules vivantes, et leurs mécanismes d'adaption, i.e. leurs régulations. Compte tenu de la complexité d'une cellule, l'idée initiale consistait à commencer par un sous-système cellulaire pour lequel le niveau de connaissance était important. Notre choix s'est porté sur le métabolisme de la bactérie modèle des Gram+ *Bacillus subtilis* plutôt que la bactérie modèle des Gram- *Escherichia coli*. Au delà du niveau important de connaissance accumulée sur *B. subtilis*, la raison de ce choix a été la proximité de nos collègues microbiologistes (MICALIS, INRAE) spécialistes de cette bactérie sur Jouy-en-Josas, avec lesquels nous avons collaboré dans le cadre de plusieurs projets européens de recherche (BaSysBio, Basynthec, ITN ProteinFactory) et de plusieurs co-encadrements de thèse (O. Borkowski [T.3], C. Cousin [T.4], A. Kalantari [T.5], M. Zaarour [T.7]). J'ai reconstruit manuellement les voies métaboliques du métabolisme primaire de *B. subtilis* ainsi que l'ensemble des régulations associées : transcriptionnelles, traductionnelles, post-traductionnelles et enzymatiques [A.1*]. Ce modèle de connaissance est continuellement enrichi, et contient actuellement environ 275 mécanismes de régulation à l'échelle de la cellule entière [A.12]. Le réseau de régulation génétique des voies métaboliques a une structure hiérarchique. Peu de régulateurs ont un rôle pléiotrope, c'est-à-dire qu'ils contrôlent un grand nombre de gènes ou qu'ils induisent une cascade de régulation. A l'inverse, la plupart des régulateurs contrôlent uniquement un nombre limité de gènes souvent impliqués dans la même voie métabolique. Cette structure hiérarchique de contrôle n'est pas spécifique à *B. subtilis*, puisque des résultats comparables ont été obtenus sur la bactérie modèle des Gram- *Escherichia coli* [24]. Mais l'organisation des voies métaboliques n'est révélée que lorsque le réseau de régulation génétique est connecté aux voies métaboliques. En effet, plus de la moitié des gènes impliqués dans le métabolisme sont sous le contrôle direct d'un métabolite via un mécanisme de régulation comme un facteur de transcription ou un riboswitch (voir [A.1*]). L'activité de la majorité des mécanismes de régulation est modulée par un métabolite. Il existe donc une boucle de rétroaction du métabolisme sur le réseau génétique ([A.1*], [19]).

Nous avons alors identifié deux motifs de régulation récurrents (cf. figure 2.1) possédant des propriétés mathématiques (et biologiques) remarquables : unicité du régime d'équilibre, découplage du régime d'équilibre vis à vis de cofacteurs, rôle clé des enzymes irréversibles dans la définition de ces motifs. De plus, les variations des composants des modules en réponse à des perturbations des entrées/sorties peuvent être prédites qualitativement. De part leurs propriétés mathématiques, ces motifs correspondent à des modules fonctionnels, dont l'objectif est d'assurer l'adaptation de la voie métabolique à des variations de flux situées en amont ou en aval du module. L'adaptation (ou la régulation) est dite locale, dans la mesure où le métabolite clé impliqué dans la régulation génétique appartient à la voie métabolique (cf. figure 2.1). La régulation globale d'une voie métabolique est définie comme l'ensemble des régulations non locales.

Cette notion de module a été approfondie dans mes travaux de thèse [T.1], publiés sous forme de chapitre d'ouvrage [B.2*] en analysant de façon systématique les régimes d'équilibre de différentes interconnections de modules métaboliques. Nous avons ainsi déterminé si le régime d'équilibre d'une interconnection existait structurellement ou non, et dans ce dernier cas, quelles étaient les conditions supplémentaires nécessaires pour assurer l'existence du régime d'équilibre. Lorsqu'on analyse l'ensemble des voies métaboliques de *B. subtilis* décrites dans [A.1*], le ré-

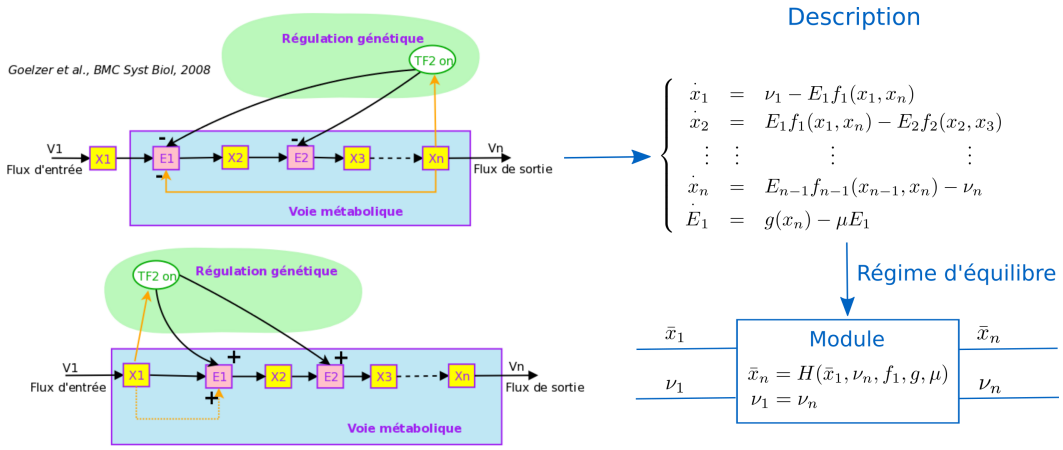

FIGURE 2.1 – **Les deux grands types de modules présents dans les voies métaboliques, et un exemple de formalisation pour une régulation locale de voie métabolique par le métabolite final x_n. Les concentrations de métabolites et d'enzymes sont notées respectivement x_* et E_*, les flux ν_*. Les fonctions f_* désignent les caractéristiques cinétiques (i.e. l'activité) des enzymes, et g la fonction de production de la première enzyme de la voie métabolique (voir [T.1, B.2*] pour plus de détails sur ces fonctions).**

gime d'équilibre existe structurellement dans la majorité des cas, ce qui est, en soit, absolument remarquable compte tenu des non-linéarités fortes du modèle mathématique associé. Une exception notable concerne la voie du métabolisme carboné central, que nous avons étudié dans la thèse de C. Cousin [T.4]. Dans ce cas, l'existence du régime d'équilibre est conditionnée au fait que les caractéristiques cinétiques croissantes de deux enzymes (la phosphofructokinase et la pyruvate kinase, encodées respectivement par les gènes *pfkA* et *pykA*) possèdent une unique intersection. Or les gènes *pfkA* et *pykA* forment un opéron chez *B. subtilis*. Les quantités d'enzymes évoluent ainsi de façon coordonnée, ce qui facilite l'existence de ce point d'équilibre. Cet exemple illustre parfaitement l'intérêt de ce type d'analyse : identifier mathématiquement les acteurs clés du fonctionnement du système et les mettre en perspective du point de vue biologique.

Discussion. A ma connaissance, il n'existe pas d'études comparables à [A.1*, B.2*] d'une telle ampleur, à savoir la décomposition systématique du réseau métabolique en modules fonctionnels et l'analyse de leur régime d'équilibre du point de vue théorique. Une perspective théorique à ce travail serait d'étudier la stabilité des modules et de leurs interconnexions. Ce problème reste difficile. Peu de résultats existent dans la littérature et ont été obtenus dans des cas rarement réalistes du point de vue biologique (e.g. une voie métabolique où toutes les enzymes sont irréversibles) [3]. L'obtention de résultats généraux semble aussi limité, car conditionné à la structure de la voie métabolique choisie. Du point de vue expérimental, la thèse [T.4] a montré la difficulté réelle à venir perturber (de façon biologique) la régulation du métabolisme carboné central chez *B. subtilis*. Un des objectifs de cette thèse était de perturber le couplage en opéron des deux gènes *pfkA* et *pykA* identifiés comme clés dans le fonctionnement de cette voie métabolique. La difficulté d'obtenir une souche viable a constitué en soit une première validation de notre approche. Toutefois, la validation biologique de notre analyse théorique reste toujours d'actualité. La thèse [T.4] a aussi soulevé une autre limitation expérimentale :

le fait de disposer d'outils permettant de réaliser biologiquement [2] et d'observer des variations faibles ou modérées des entités biologiques autour d'un régime d'équilibre.

2.2 Adaptation de sous-systèmes cellulaires à des perturbations

Tout au long de ma thèse et dans les années qui ont suivi, j'ai été amenée à développer moi-même (pour les deux premiers modèles (*i*)-(*ii*)) ou à encadrer (modèles *iii*) le développement des modèles de sous-systèmes cellulaires afin de comprendre leur fonctionnement en présence de perturbations :

(*i*) l'adaptation de la voie du métabolisme carboné central de *B. subtilis* lors d'un shift nutritionnel [D.2, A.6, T.4],

(*ii*) la réponse au stress oxydatif et à la carence en fer chez *B. subtilis* [R.1],

(*iii*) l'adaptation de la voie centrale des carbones de la levure *Saccharomyces cerevisiae* lors de perturbations de NADPH [T.2, A.7, A.8*].

Le but ici n'est pas de détailler l'ensemble de ces modèles mais de souligner leur apport dans ma réflexion scientifique.

Le modèle de la voie du métabolisme carboné central de *B. subtilis* est composé de 67 équations différentielles ordinaires non-linéaires. Ce modèle a été étudié analytiquement à travers l'analyse en modules fonctionnels décrite dans la section précédente, et numériquement en simulation. L'analyse théorique a permis d'identifier (a) les points clés du contrôle de la voie centrale des carbones (b) les régulations manquantes, qui ont été confirmées expérimentalement dans [T.4], (c) et des conditions suffisantes pour obtenir un point d'équilibre unique (discutées dans la section précédente). En simulation, en utilisant des paramètres extraits de la littérature, nous avons également obtenu un régime d'équilibre, et ce malgré le nombre important d'équations différentielles non-linéaires à simuler. Ces simulations démontrent, une fois encore, que les systèmes biologiques possèdent des propriétés structurelles remarquables. En effet, il est rare d'obtenir un régime d'équilibre pour un système non-linéaire de grande dimension.

Le modèle de la réponse au stress oxydatif et de la carence en ion de *B. subtilis* est un modèle de connaissance (cf. figure 2.2). Il m'a permis, en premier lieu, d'apprendre les effets pléiotropes des différents types de stress au sein de la cellule, de mieux comprendre le rôle de chaque régulateur impliqué, et comment les effets combinés de ces différents régulateurs permettent *in fine* d'assurer l'adaptation de la cellule au stress. Ce modèle de connaissance sera utilisé par la suite afin d'étudier si les prédictions issues de nos modèles de cellules entières sont réalistes du point de vue biologique (voir chapitre 3).

Enfin, le modèle de la voie centrale des carbones de *S. cerevisiae*, développé dans le cadre de la thèse de M. Celton [T.2], m'a permis de me former à la modélisation sous contraintes (voir section 1.1.3). En conditions fermentaires, la conservation de la masse pour les couples redox NAD+/NADH et NADP+/NADPH induit des contraintes fortes sur la production de sous-produits comme le glycérol, l'éthanol. La modélisation sous contraintes était donc le cadre

2. Lors de la thèse [T.4], la technique CRISPR-Cas9 [10], maintenant standard en ingénierie des génomes pour modifier *à façon* le code génétique, était encore en émergence et n'avait pas encore été largement adoptée par la communauté des biologistes.

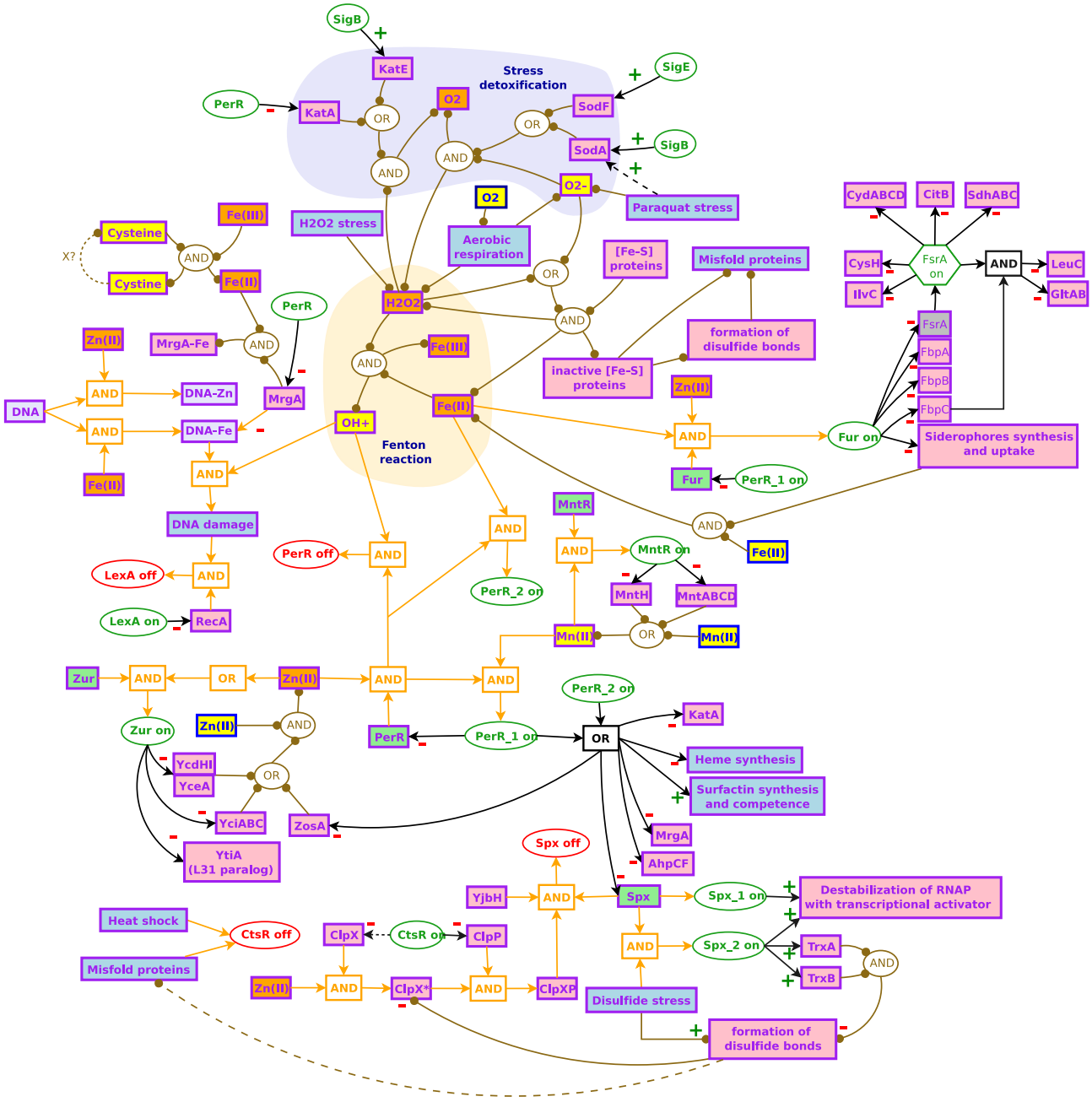

FIGURE 2.2 – **Modèle de connaissance de la réponse au stress oxydatif, stress disulphide, et en carences en fer et manganèse de *B. subtilis*. Figure extraite de [R.1].**

de modélisation adéquat. En collaboration avec nos collègues biologistes de l'UMR SPO de l'INRA (S. Dequin, C. Camarasa), nous avons donc développé un modèle sous contraintes de *S. cerevisiae* spécialisé aux conditions fermentaires, i.e. où seules les isoformes des enzymes actives en conditions fermentaires. En combinant le modèle et des expériences biologiques dédiées, nous avons identifié les mécanismes impliqués dans la réponse de cette levure à des perturbations graduelles de NADPH en conditions oenologiques [T.2, A.7, A.8*]. Au delà des résultats obtenus, ce travail est représentatif de la biologie des systèmes, où les approches de modélisation et de biologie expérimentales sont réellement combinées et permettent de progresser dans la compréhension du fonctionnement du système biologique.

2.3 Le taux de croissance comme un régulateur global de l'expression des gènes

Les travaux de [A.1*] ont montré l'existence d'une structure de contrôle modulaire dans les voies métaboliques, où les quantités d'enzymes dans les voies métaboliques sont modulées en fonction des concentrations de métabolites clés situés en aval ou en amont de la voie métabolique (voir Figure 2.1). Or, au regard de la connaissance actuelle, seulement environ la moitié des gènes impliqués dans le métabolisme sont sous le contrôle direct d'un mécanisme de régulation génétique. Les gènes restant sont dits *constitutifs*, i.e. exprimés continuellement et de façon non contrôlée. Cela peut sembler surprenant car de nombreux gènes constitutifs sont présents dans les voies de synthèse d'acides aminés. Or la demande en acides aminés augmentant avec le taux de croissance, comment ces voies peuvent-elles s'adapter sans mécanisme de contrôle explicite ? Cette question a été explorée dans la thèse d'O. Borkowski [T.3]. Nous avons développé un modèle dynamique décrivant les étapes de la traduction des protéines à partir des ARN messagers (ARNm) :

$$\dot{P}_i(t) = m_i(t)\frac{K_{1i}R_{free}(t)}{K_{2i} + R_{free}(t)} - \mu(t)P_i(t)$$

où P_i, m_i correspondent respectivement aux concentrations de la protéine et de l'ARNm, R_{free} représente la concentration du complexe d'initiation de la traduction, μ est le taux croissance. K_{1i} et K_{2i} sont des paramètres constants spécifiques à chaque ARNm et dépendant uniquement des constantes d'association, de dissociations ou de translocation du ribosome le long de l'ARNm. La concentration R_{free} agit ainsi sur la synthèse de l'ensemble des protéines. En phase exponentielle de croissance (à l'équilibre), on obtient une relation entre le taux de croissance, et les concentrations des protéines, des ARNm et R_{free} : $P_i = \frac{m_i}{\mu}\frac{K_{1i}R_{free}}{K_{2i} + R_{free}}$. Différentes séquences TIRs (Translation Initiation Region) situées en amont des codons start sur l'ARNm permettent d'obtenir différentes valeurs pour K_{1i} et K_{2i}, et donc d'obtenir différents profils d'expression en fonction du taux de croissance. En utilisant des données de protéomique quantitative et de transcriptomique pour cinq milieux de culture différents, nous avons pu estimer le profil de R_{free} en fonction du taux de croissance, ainsi que les paramètres K_{1i} et K_{2i} pour $\approx$1000 transcrits [A.13*] (cf. figure 2.3). Cela nous permet d'obtenir une banque de TIRs pour lesquels nous pouvons (théoriquement) prédire la quantité de protéine produite pour un taux de croissance et une quantité d'ARNm donnés.

La transcription est régulée de manière analogue par le taux de croissance, à travers la quantité d'ARN polymérase libre dans la cellule, différents types de promoteurs (i.e. les motifs de fixation de l'ARN polymérase sur l'ADN) possédant des affinités différentes, le(s) premiers(s)

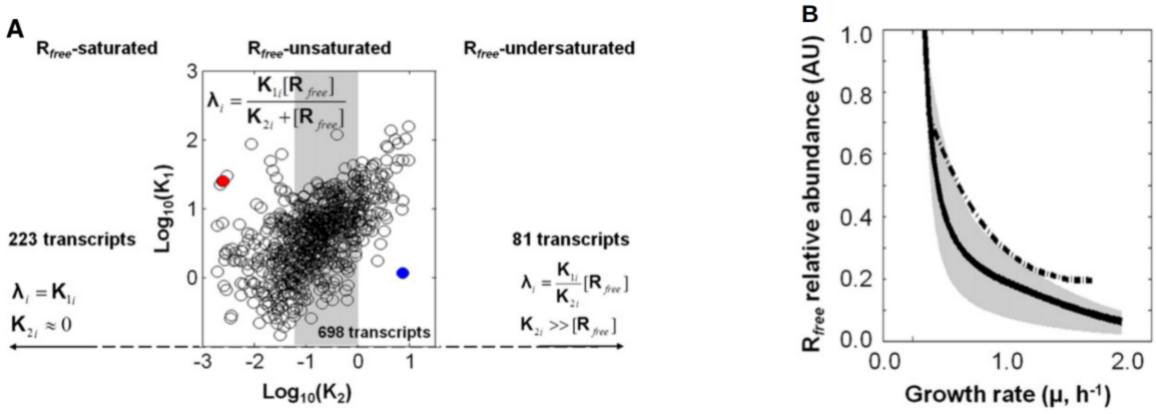

FIGURE 2.3 – **(A). Valeurs de K_{1i} vs. K_{2i} en échelle logarithmique pour chacun des 698 transcripts pour lequel l'efficacité de traduction $\lambda_i = \frac{K_{1i}R_{free}}{K_{2i}+R_{free}}$ suit une loi de type Michaelis-Menten. (B). Estimation de la concentration R_{free}. Figure extraite de [A.13*].**

nucléotide(s) des transcrits, le nombre de copies du gène présents dans la cellule (voir [12] pour un modèle préliminaire reprenant certains de ces éléments). Chaque gène possède donc un profil d'expression spécifique en fonction du taux de croissance, codé *en dur* dans les séquences des promoteurs et des TIRs. En caractérisant de façon systématique les profils d'expression des gènes constitutifs, on peut disposer de banques natives de bio-briques (les promoteurs et les TIRs) d'intérêt pour la Biologie de Synthèse permettant, en théorie, de moduler « à façon » le profil d'expression d'un gène d'intérêt.

2.4 Conclusion

Les travaux présentés dans ce chapitre ont été fondateurs à plus d'un titre. Tout d'abord, la reconstruction (manuelle) du réseau métabolique de *B. subtilis* et de l'ensemble de ses régulations, de la réponse au stress oxydatif et à la carence en ions m'ont permis, en premier lieu, d'investir le domaine biologique, n'étant pas issue moi-même de ce domaine. Ensuite, l'existence de modules fonctionnels, i.e. de sous-systèmes, dans les voies métaboliques ainsi que dans de nombreuses autres fonctions cellulaires comme la réponse au stress replace l'Automatique et l'Ingénierie des systèmes comme des outils pertinents pour étudier les systèmes biologiques. Les travaux [A.1*, B.2*, R.1, A.13*] ont ainsi initié mes deux principales directions de recherche actuelles :

(A) Comment une organisation modulaire a-t-elle pu émerger dans les cellules ? (voir chapitre 3 et 4)

(B) Comment représenter efficacement les organismes vivants par des approches systèmes ? (voir chapitre 5.2)

Principales publications de ce chapitre

A.13* O. Borkowski, A. Goelzer, M. Schaffer, U. Mäder, S. Aymerich, M. Jules, V. Fromion. Translation elicits a growth-rate dependent and genome-wide, differential production of proteins in *Bacillus subtilis*. *Molecular Systems Biology*, 12(5) :870, 2016.

A.8* M. Celton, A. Goelzer, C. Camarasa, V. Fromion, S. Dequin. A constraint-based model analysis of the metabolic consequences of increased nadph oxidation in *Saccharomyces cerevisiae*. *Metabolic engineering*, 14(4) :366-379, 2012.

B.2* A. Goelzer and V. Fromion. Towards the modular decomposition of the metabolic network. In V. Kulkarni, editor, *A Systems Theoretic Approach to Systems and Synthetic Biology I : Models and System Characterizations*, pages 121-152. Springer, 2014.

A.1* A. Goelzer, F. Bekkal Brikci, I. Martin-Verstraete, P. Noirot, P. Bessières, S. Aymerich, V. Fromion. Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Systems Biology*, 2 :20, 2008.

Chapitre 3

L'allocation des ressources comme principe de design des cellules et comme outil de prédiction

Suite aux travaux sur la reconstruction des voies métaboliques [A.1*], la structure du contrôle des voies métaboliques de *B. subtilis* apparaît donc modulaire, hautement coordonnée et hiérarchisée en fonction de quelques signaux métaboliques pléiotropes bien choisis. En Automatique, une telle structure de contrôle est appelée commande décentralisée. En pratique, pour obtenir un correcteur décentralisé par des algorithmes classiques, il est nécessaire d'imposer la structure du correcteur par des contraintes, en plus des contraintes classiques de performance et de robustesse liées au cahier des charges. Ce parallèle entre Automatique et Biologie nous a conduit à rechercher des contraintes intrinsèques au fonctionnement des cellules vivantes, et donc de fait structurelles.
Intuitivement on peut identifier la loi de conservation de la masse au sein de la cellule, formalisée dans la méthode FBA, par l'ensemble de contraintes égalités $\Omega\nu = 0$ (cf. section 2.2). Du point de vue de l'Automatique, le problème d'optimisation P_{fba} associé à la méthode FBA correspond à une représentation en boucle ouverte de la cellule. Plus les flux à travers les voies métaboliques sont importants, et plus la cellule pourra produire de biomasse. Or le volume de la cellule est limité, ce qui contraint, de fait, la quantité de composés cellulaires maximale admissible dans la cellule. Enfin, et de façon plus profonde, la cellule est un système qui s'autoreproduit. Elle doit produire non seulement les composés de bases de la biomasse (acides aminés, energie, (déoxy)-nucléotides, etc.), les machines moléculaires qui produisent et assemblent ces composés pour produire la biomasse (e.g. enzymes, transporters, ARN polymérase, etc.) mais aussi les machines qui produisent les machines moléculaires (e.g. l'appareil de traduction). Ainsi la capacité de production de chaque machine moléculaire doit être suffisante pour assurer sa fonction. A titre illustratif, prenons l'exemple d'une enzyme $\mathbb{E}_i$ irréversible de concentration E_i. La capacité de production de l'enzyme doit être suffisante pour produire son flux métabolique ν_i, soit :

$$\nu_i \leq k_{E_i} E_i$$

où le coefficient $k_{E_i} > 0$ correspond à l'efficacité de l'enzyme. De plus, l'enzyme $\mathbb{E}_i$ doit être produite par l'appareil de traduction, elle va occuper un certain volume dans le cytosol. Il s'agit donc de ré-introduire les concentrations des machines moléculaires dans un cadre de modélisation sous contraintes, d'identifier et de formaliser leurs contraintes de fonctionnement sous

forme d'un problème d'optimisation sous contraintes. Ces travaux de formalisation avaient été initiés dans ma thèse [T.1] et consolidés dans les articles [C.1, A.3*, A.4], pour aboutir à une nouvelle méthode de modélisation sous contraintes appelée Resource Balance Analysis (RBA). Dans les sections suivantes, je distinguerai la méthode RBA d'un modèle RBA. La méthode RBA correspondra au cadre méthodologique que nous avons développé. Un modèle RBA correspondra au problème d'optimisation de type RBA, dédié à un organisme spécifique. Ce modèle contient ainsi une description de la cellule de l'organisme considéré.

Ces travaux ont été menés en collaboration avec les membres permanents de l'équipe BioSys (V. Fromion et L. Tournier en particulier), mon directeur de thèse G. Scorletti de l'École centrale de Lyon, les thésards (M. Zaarour [T.7], A. Bulović [T.8], O. Bodeit [T.9]) et post-doctorants (S. Fischer) que j'ai directement supervisés, et mes collaborateurs biologistes pour l'acquisition des données nécessaires à la validation expérimentale du modèle RBA de *B. subtilis* (MICALIS : E. Prestel, M. Jules ; Université de Greifswald : J. Muntel, D. Becher ; ETHZ : V. Chubukov, U. Sauer). A. Bulović, O. Bodeit et S. Fischer ont des profils pluridisciplinaires en (bio)-informatique, mathématiques appliquées et biologie des systèmes de part leur formation et expériences passées, alors que M. Zaarour est biologiste. En plus des aspects techniques relevant de la modélisation, j'ai co-encadré les thèses au niveau scientifique, en définissant les objectifs, les différents points d'étape des travaux de recherche, et, le cas échéant, en proposant des solutions pour contourner des difficultés.

3.1 La méthode Resource Balance Analysis (RBA)

La méthode RBA formalise mathématiquement les interactions et l'allocation des ressources entre les entités cellulaires sous forme de contraintes d'égalités et d'inégalités. Nous avons ainsi identifié trois grandes classes de contraintes, représentées sur la figure 3.1 : En phase exponentielle de croissance,

(I) le réseau métabolique doit produire tous les précurseurs métaboliques nécessaires à la production de la biomasse (égalités C_1 en vert) ;

(II) la capacité de production de toutes les machines moléculaires doit être suffisante pour assurer leur fonction, à savoir catalyser la réaction chimique à un flux suffisant (inégalités C_2 en bleu pour les enzymes et les transporteurs, en jaune pour les machines moléculaires associés aux autres processus macromoléculaires) ;

(III) la densité intracellulaire des compartiments et l'occupation des membranes sont limitées (inégalités C_3 en orange) ;

(IV) la conservation de la masse est assurée pour tous les types de molécules (égalités C_1 en vert).

Afin de formaliser ces trois grandes classes de contraintes, on considère que la cellule se compose de l'ensemble d'entités suivantes :

(i) N_y machines moléculaires, réparties en N_m enzymes et transporteurs impliqués dans les voies métaboliques $\mathbb{E} \triangleq (\mathbb{E}_1, \ldots, \mathbb{E}_{N_m})$ aux concentrations $E \triangleq (E_1, \ldots, E_{N_m})^T$ et associées aux flux métaboliques $\nu \triangleq (\nu_1, \ldots, \nu_{N_m})^T$; et N_p machines macromoléculaires $\mathbb{M} \triangleq (\mathbb{M}_1, \ldots, \mathbb{M}_{N_p})$ impliquées dans les processus macromoléculaires comme l'appareil de traduction, aux concentrations $M \triangleq (M_1, \ldots, M_{N_p})^T$;

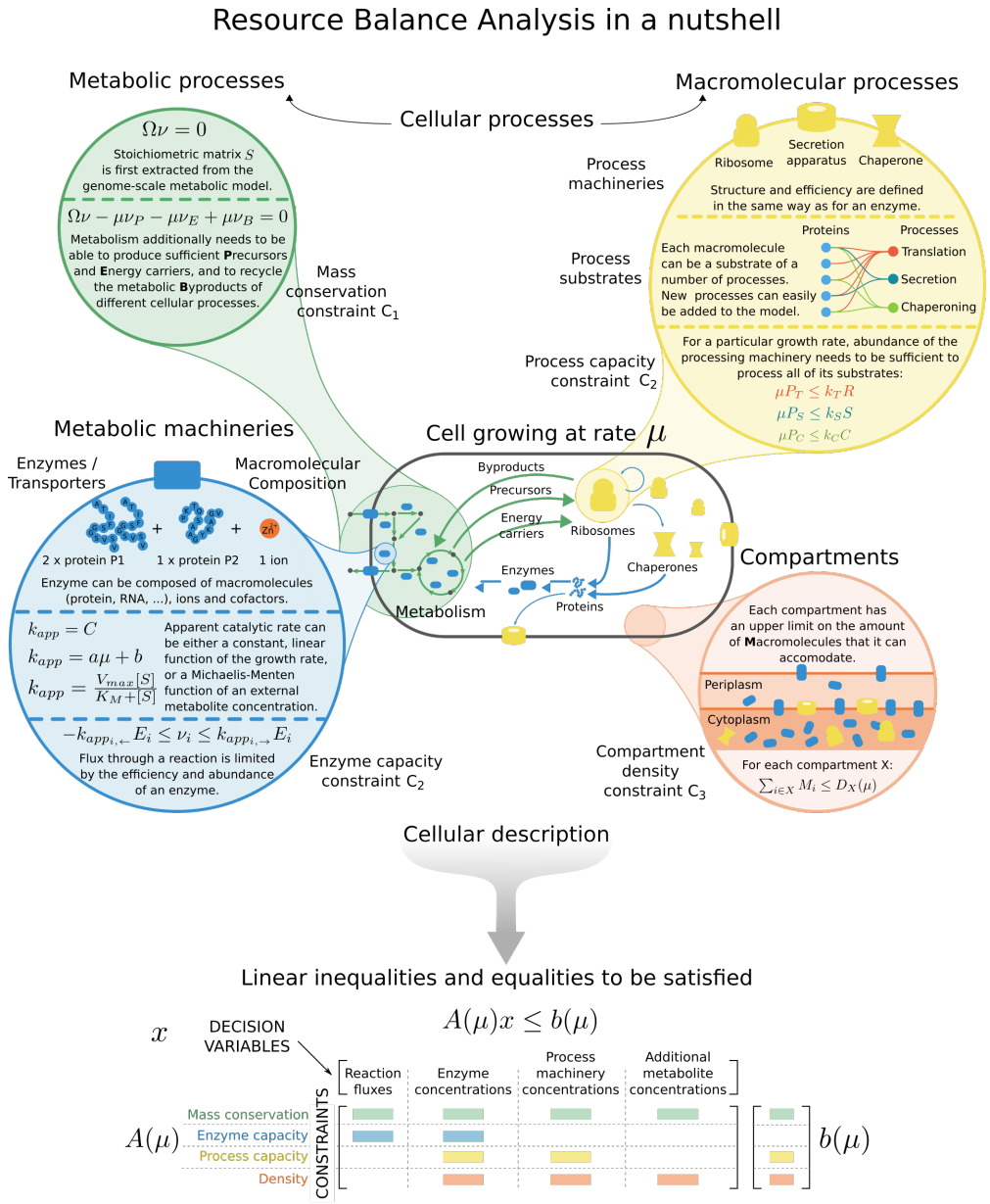

FIGURE 3.1 – **La méthode RBA. Figure extraite de [A.18*].**

(*ii*) N_g protéines $\mathbb{P}_G \triangleq \{\mathbb{P}_{G_1}, \ldots, \mathbb{P}_{G_{N_g}}\}$ pour lesquelles le processus cellulaire dont elles dépendent n'est pas spécifié. $P_G \triangleq (P_{G_1}, \ldots, P_{G_{N_g}})^T$ correspond à l'ensemble des concentrations de $\mathbb{P}_G$;

(*iii*) N_s métabolites $\mathbb{S} \triangleq (\mathbb{S}_1, \ldots, \mathbb{S}_{N_s})$ aux concentrations $S \triangleq (S_1, \ldots, S_{N_s})^T$. Parmi l'ensemble $\mathbb{S}$, on distingue un sous-ensemble de métabolites $\mathbb{B} \triangleq (\mathbb{B}_1, \ldots, \mathbb{B}_{N_b})$ ayant une concentration fixe $\bar{B} \triangleq (\bar{B}_1, \ldots, \bar{B}_{N_b})^T$.

Par définition, chaque composante de Y est positive. Pour un taux de croissance $\mu \geq 0$ donné, et pour une concentration $P_G \in \mathbb{R}^g_{>0}$ de protéines donnée, on cherche une distribution de ressources (Y, ν) faisable, i.e. qui satisfait les contraintes suivantes :

$$
\begin{array}{ll}
\text{Trouver} & Y \in \mathbb{R}^y_{\geq 0}, \nu \in \mathbb{R}^m, \\
\text{tels que} & \\
(C_1) & -\Omega\nu + \mu(C^S_Y Y + C^S_B \bar{B} + C^S_G P_G) = 0 \\
(C_2) & \mu(C^M_Y Y + C^M_G P_G) - K_T Y \leq 0 \\
 & -K^{'}_E Y \leq \nu \leq K_E Y \\
(C_3) & C^D_Y Y + C^D_G P_G - \bar{D} \leq 0
\end{array}
$$

où :

- Ω est la matrice de stœchiométrie du réseau métabolique de taille $N_s \times N_m$, où Ω_{ij} désigne le coefficient stœchiométrique du métabolite $\mathbb{S}_i$ impliqué dans la j-ème réaction ;
- C^S_Y (resp. C^S_G) est une matrice de taille $N_s \times N_y$ (resp. $N_s \times N_g$) où chaque coefficient $C^S_{Y_{ij}}$ (resp. $C^S_{G_{ij}}$) correspond au nombre de métabolites $\mathbb{S}_i$ consommés (ou produits) pour la synthèse d'une machine $\mathbb{Y}_j$ (resp. d'une protéine $\mathbb{P}_{G_j}$). $C^S_{Y_{ij}}$ et $C^S_{G_{ij}}$ sont donc positifs, négatifs ou nuls si $\mathbb{S}_i$ est produit, consommé ou non impliqué dans la synthèse de la machine moléculaire $\mathbb{Y}_j$ ou de la protéine $\mathbb{P}_{G_j}$;
- C^S_B est une matrice de taille $N_s \times N_b$ où chaque coefficient $C^S_{B_{ij}}$ correspond au nombre de métabolites $\mathbb{S}_i$ consommés (ou produits) pour la synthèse d'un $\mathbb{B}_j$;
- la matrice K_T (respectivement K_E, $K^{'}_E$) de taille $N_p \times N_p$ (respectivement $N_m \times N_m$) est une matrice diagonale où chaque coefficient K_{T_i} (respectivement K_{E_i}, $K^{'}_{E_i}$) est positif et correspond à l'efficacité de la machine moléculaire $\mathbb{M}_i$ (respectivement les efficacités de l'enzyme $\mathbb{E}_i$ dans le sens forward et reverse), i.e. la vitesse de production du processus par unité de machine moléculaire.
- C^M_Y (resp. C^M_G) est une matrice de taille $N_p \times N_y$ (resp. $N_p \times N_g$) où chaque coefficient $C^S_{Y_{ij}}$ correspond au nombre d'acides aminés de la machine $\mathbb{Y}_j$ (resp. de la protéine $\mathbb{P}_{G_j}$). Dans certains cas (par exemple pour les contraintes de repliement des protéines), le nombre d'acides aminés peut être multiplié par un coefficient, comme la fraction du protéome nécessitant l'aide d'une chaperone pour être replié ;
- $\bar{D}$ est un vecteur de taille N_c où le coefficient $\bar{D}_i$ désigne la densité totale d'entités cellulaires admissible dans le i-ème volume ou la i-ème surface. Les densités s'expriment en nombre d'acides aminés par unité de volume ou de surface [25].
- C^D_Y (resp. C^D_G) est une matrice de taille $N_c \times N_y$ (resp. $N_c \times N_g$) où chaque coefficient $C^D_{Y_{ij}}$ (respectivement $C^D_{G_{ij}}$) correspond à la densité de la machine $\mathbb{Y}_j$ (resp. $\mathbb{P}_{G_j}$) dans le i-ème volume ou la i-ème surface.

Ce problème d'optimisation, que nous appellerons $P_{rba}(\mu)$, est convexe pour μ fixé [C.1, A.3*], et correspond à un problème de Programmation Linéaire. À μ fixé, l'espace des solutions faisables défini par $P_{rba}(\mu)$, correspond à l'ensemble des phénotypes cellulaires possibles. Parmi ces phénotypes, on peut s'intéresser en particulier à celui qui obéit à un principe d'utilisation parcimonieuse des ressources, i.e. le phénotype pour lequel le taux de croissance est maximal. On peut d'ailleurs montrer que ce taux de croissance μ^* existe et est tel que pour tout $\mu \leq \mu^*$, $P_{rba}(\mu)$ est faisable et pour tout $\mu > \mu^*$, $P_{rba}(\mu)$ n'a pas de solution [C.1, A.3*]. Ainsi, le taux de croissance maximal μ^* peut être calculé en résolvant une série de problème LP pour différentes valeurs de μ. En pratique, on utilise un algorithme de type dichotomie pour sélectionner les valeurs de taux de croissance à tester. Pour un milieu donné, on peut donc calculer le taux de croissance maximal, et l'allocation en ressources associée, i.e. les flux métaboliques ν^*, et les concentrations de machines moléculaires Y^*.

Remarque. Pour simplifier les notations, j'ai choisi de présenter dans cette HDR la formulation $P_{rba}(\mu)$ intégrant uniquement les aspects μ dépendant. Une formulation plus générale du problème $P_{rba}(\mu)$ intégrant explicitement les vitesses de dégradation des macromolécules, i.e. le *turnover* des ARNm et des protéines, est donnée dans [A.19]. L'impact du turnover des macromolécules en terme de coût en ressources devient important pour la cellule en phase stationnaire (i.e. à $\mu = 0$). Le méthode RBA permet d'intégrer ces coûts facilement, de même que les machines moléculaires impliquées, à savoir les protéases et les ARNases.

Des modèles sous contraintes plus autonomes. A la différence des modèles FBA (cf. section 1.1.3), la configuration de la cellule pour une condition environnementale donnée n'est plus imposée à travers des limitations sur certains flux métaboliques a priori (bornes α et β de P_{fba}), mais résulte d'un compromis d'allocation parcimonieuse en ressources. Dans la méthode RBA, la condition environnementale correspond aux concentrations en nutriments dans le milieu de culture. Elle est modélisée à travers les efficacités des transporteurs localisés sur la membrane plasmique. On considère que l'efficacité du transporteur est une fonction croissante de type Michaëlis-Menten de son (ou ses) nutriment(s) extracellulaire(s) :

$$K_{E_i} = \frac{k_{max}^{E_i} S_j}{K_m^{E_i} + S_j}$$

où S_j est la concentration extracellulaire du nutriment, $k_{max}^{E_i} > 0$ est l'efficacité maximale du transporteur, et $K_m^{E_i} > 0$ la constante d'affinité. En fonction des concentrations extracellulaires en nutriments, le coût d'import du nutriment change. La configuration finale prédite résulte donc d'un compromis entre l'investissement lié au transport et à l'assimilation des nutriments et le bénéfice obtenu en terme de croissance cellulaire. A titre d'exemple, le modèle RBA de *B. subtilis* prédit correctement l'utilisation des transporteurs haute et basse affinité en fonction de la valeur de la concentration extracellulaire de nutriment (voir [A.3*]).

Liens avec les modèles phénoménologiques de type Monod [26]. Une première validation des contraintes intégrées consiste à comparer le comportement du modèle RBA avec des modèles phénoménologiques bien établis. Par exemple, le modèle phénoménologique pour la croissance bactérienne développé par J. Monod [26] décrit la croissance bactérienne en fonction de la concentration en nutriment limitante S par la relation suivante :

$$\mu = \mu_{max} \frac{S}{K_m + S}$$

où les paramètres μ_{max} (taux de croissance maximal) et K_m sont déterminés expérimentalement. En simulant un modèle RBA pour différentes valeurs de concentration d'un nutriment, le modèle RBA recouvre bien un comportement macroscopique de type Monod (cf figure 3.2). De même, le modèle RBA reproduit correctement l'évolution des grands pools de protéines

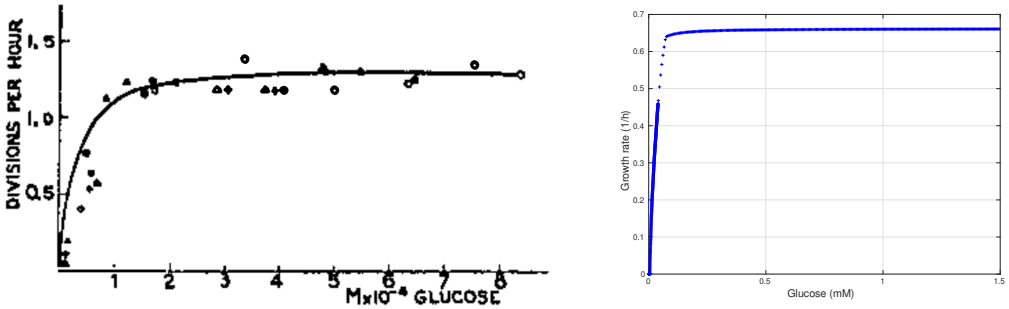

FIGURE 3.2 – **Comparaison des taux de croissance bactérienne en fonction de la concentration extracellulaire de glucose par le modèle de Monod (gauche, figure extraite de [26]) et par le modèle RBA de *B. subtilis* (droite).**

intracellulaires (ribosomales vs non-ribosomales) en fonction du taux de croissance [A.4]. Ces résultats démontrent que les contraintes intégrées dans la méthode RBA sont suffisantes pour recouvrir les comportements phénoménologiques connus chez les bactéries. Cela signifie également qu'on dispose d'une méthode permettant de générer un modèle phénoménologique pour n'importe quel nutriment.

Positionnement de la méthode RBA vis-à-vis d'autres méthodes existantes. Depuis la première publication de la méthode RBA en 2009 [C.1], d'autres méthodes de modélisation sous contraintes intégrant un principe d'allocation de ressource à l'échelle du génome ont été proposées : ME-FBA [30], et cFBA [27]. Elles diffèrent principalement de la méthode RBA par le type de processus cellulaires qui sont considérés, ou le fait que certaines contraintes soient saturées (contrainte d'égalité) ou potentiellement non saturées (contrainte d'inégalité). Ces méthodes peuvent se reformuler sous forme d'un problème RBA, et aucune n'a été validée expérimentalement à l'échelle du génome.

3.2 Validation expérimentale du modèle RBA de *B. subtilis*

La méthode RBA a ensuite été validée expérimentalement en 2015 sur la bactérie *Bacillus subtilis* [A.11*]. Pour cela, nous avons planifié les expériences biologiques à réaliser, et nos collègues de l'université de Greifswald et de l'ETH Zürich ont généré un jeu de données dédié à la calibration et à la validation du modèle RBA de *B. subtilis*. Ce jeu de données était composé de protéomique quantitative pour les protéines cytosoliques, des flux de nutriments consommés et de sous-produits excrétés, et ce pour cinq milieux de culture conduisant à une large gamme de taux de croissance de 0.3h^{-1} à $1.6\ \text{h}^{-1}$. Quatre milieux de culture ont été utilisés pour calibrer les paramètres du modèle (principalement les matrices des efficacités K_T, K_E et la concentration des protéines de maintenance P_G) par un algorithme de type moindres carrés et le cinquième pour le valider. En particulier, nous avons estimé les vitesses apparentes d'environ 600 enzymes (matrices K_E) pour ces 5 milieux (cf figure 3.3). Lors de la calibration de la méthode, une première étape a été d'analyser les données, notamment les données de

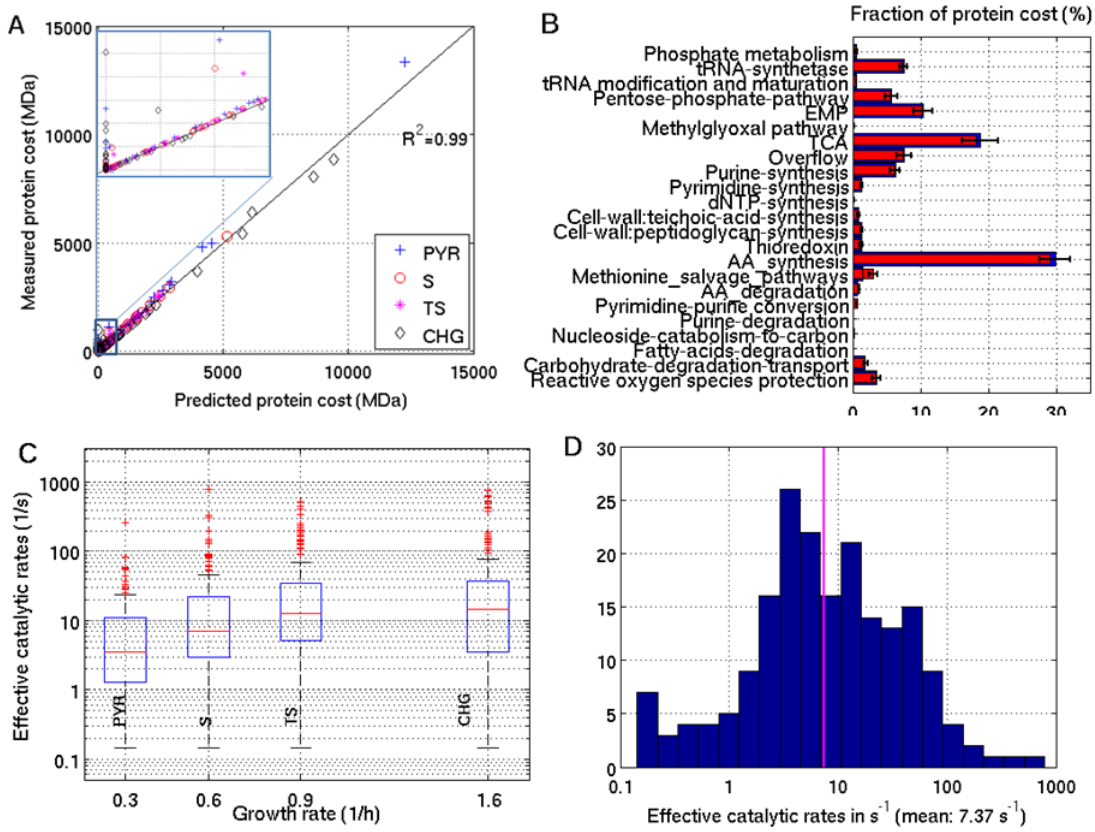

FIGURE 3.3 – **Convergence de la procédure d'estimation des paramètres. (A) Coûts en protéines (en MDa) prédits et mesurés pour les processus cellulaires pour les conditions dédiées à la calibration des paramètres (PYR, S, TS et CHG) ; (B) Comparaison entre les coûts en protéines prédits (bleu) et mesurés (rouge) pour les processus cellulaires en condition S. (C) Boxplot des vitesses apparentes des enzymes estimées sur les conditions PYR, S, TS et CHG. (D) Distribution empirique des vitesses apparentes des enzymes (en s^{-1}) estimées en condition S. La ligne en magenta représente la médiane des vitesses estimées. Figure extraite de [A.11*].**

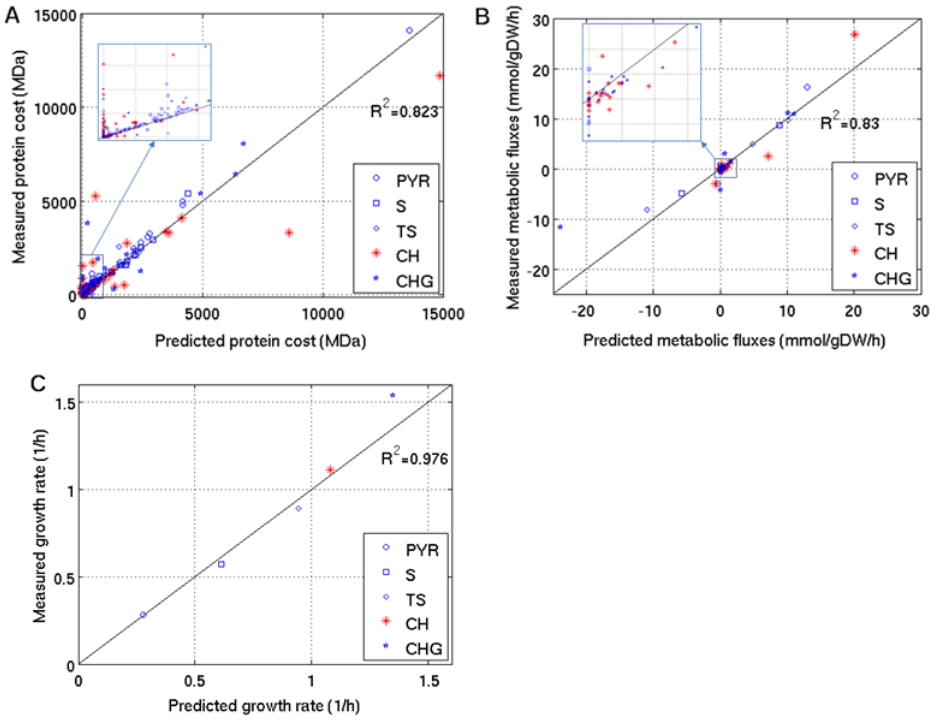

FIGURE 3.4 – **Évaluation de la capacité de prédiction du modèle RBA pour les 5 conditions PYR, S, TS, CH, CHG. (A) Coût en protéine mesuré vs prédit (en MDa) pour les processus cellulaires ($R^2 = 0.89$, $p < 10^{-101}$). (B) Flux mesurés vs prédits de nutriments importés et de sous-produits excrétés (en mmol/gDW/h, gDW = gramme de poids sec) ($R^2 = 0.83$, $p < 10^{-29}$). Les valeurs négatives de flux correspondent aux flux de sous-produits excrétés. (C) Taux de croissance mesuré vs prédit (en h^{-1}) ($R^2 = 0.98$, $p < 10^{-2}$). Figure extraite de [A.11*].**

protéomique quantitative, afin de déterminer entre autre la présence éventuelle de biais et le modèle du bruit de mesure associé à ce type de données. En effet, le jeu de données généré était le premier jeu de données de protéomique quantitative chez *B. subtilis* et le développement du protocole d'acquisition associé a été un réel challenge pour nos collègues de Greifswald [A.10*]. L'évaluation de la qualité du jeu de données de protéomique nous a ainsi permis de concevoir un algorithme d'estimation des paramètres plus adapté au bruit de mesure intrinsèque à ces données. La modèle RBA calibré a montré une grande capacité de prédiction, à la fois pour les milieux d'apprentissage PYR, S, TS et CHG, et pour le milieu de validation CH (voir figure 3.4). A ce jour, la méthode RBA est la seule méthode de modélisation sous contraintes intégrant un principe d'allocation de ressource qui a été validée expérimentalement et capable de prédire quantitativement l'allocation des ressources d'une bactérie.

Pour poursuivre nos travaux sur la validation expérimentale de la méthode RBA, nous avons caractérisé dans la thèse de M. Zaarour [T.7], en collaboration avec nos collègues de MICALIS (V. Sauveplane, M. Jules), l'impact sur le taux de croissance de la sur-expression d'une protéine « gratuite », i.e. qui ne contribue pas directement à la croissance de la bactérie. Par définition, cette protéine gratuite appartient à l'ensemble $\mathbb{P}_G$ dans la méthode RBA. L'objectif de [T.7] était d'affiner la capacité de prédiction de la méthode RBA vis-à-vis de variations du groupe de protéines $\mathbb{P}_G$, et aussi de caractériser dans une certaine mesure le coût en ressources associé au volume cellulaire occupé par une protéine.

Par la suite, nous avons rapidement cherché à rendre disponible la construction de modèles RBA

à la communauté des biologistes, bioinformaticiens ou modélisateurs (voir section suivante).

3.3 Génération automatique de modèles RBA pour les procaryotes

Le modèle RBA de *B. subtilis* a été développé manuellement, en parallèle du développement du cadre théorique. La construction du modèle nécessite d'extraire et de croiser différentes informations comme la composition en composés moléculaires et la localisation de chaque machine moléculaire $\mathbb{Y}_j$, la description des processus cellulaires considérés, etc.
Pour faciliter la création des modèles RBA pour d'autres bactéries, j'ai conçu et supervisé le développement d'un pipeline (appelé RBApy), implémenté en un package Python open-source dans le cadre de la thèse [T.8] d'A. Bulović en collaboration avec nos collègues de l'Université de Humboldt de Berlin (E. Klipp) et du post-doc de S. Fischer financé sur le projet Lidex-IMSV de l'université de Paris-Saclay. RBApy permet de construire, calibrer et simuler un modèle RBA pour les procaryotes [A.18*] (cf. figure 3.5). La construction d'un modèle RBA nécessite un modèle métabolique annoté et disponible dans le format standard Systems Biology Markup Language (SBML) [14]. Ce modèle métabolique doit contenir une description des réactions chimiques des voies métaboliques et des gènes codant les enzymes catalysant les réactions. RBApy télécharge automatiquement la dernière version de la séquence protéique annotée de l'organisme d'intérêt, croise le modèle SBML et la séquence protéique afin d'extraire automatiquement les informations nécessaires à la construction d'un modèle RBA : composition des enzymes en acides aminés, vitamines et ions, le nombre de sous-unités par complexe enzymatique, et leur localisation. En sortie, nous obtenons une première version du modèle, ainsi que des fichiers d'aide contenant les informations manquantes ou ambigues que RBApy a détectées. L'utilisateur peut ensuite renseigner les fichiers d'aide pour lever les ambiguités. RBApy met également à disposition de l'utilisateur des algorithmes pour la calibration du modèle à partir de données biologiques et des méthodes de résolution efficaces du problème d'optimisation sous-jacent. Enfin, le package inclut des fonctions pour interfacer les résultats de simulation avec des outils de visualisation comme Escher maps [17] pour la visualisation des flux métaboliques, ou Proteomaps [22] pour la répartition des protéines prédites par catégories fonctionnelles.

Validation. Afin de valider RBApy, nous avons d'abord regénéré un modèle RBA pour *B. subtilis* et comparé les résultats de simulation entre le modèle original et ce nouveau modèle. Nous obtenons une réconciliation parfaite, ce qui valide en soit le logiciel. Nous avons ensuite généré un modèle RBA calibrés pour la souche sauvage *E. coli K12* en utilisant uniquement les informations et des données de protéomique et de fluxomique disponibles dans la littérature ([31, 40, 36]). La comparaison des simulation et des mesures expérimentales montre une bonne corrélation tant pour les valeurs de flux métaboliques ($R^2 = 0.89$), que pour le nombre de copies de protéines par cellule ($R^2 = 0.65$) [A.18*]. Ceci démontre que RBApy est capable de générer un modèle RBA capable de prédictions quantitatives correctes pour une nouvelle bactérie. A notre connaissance, RBApy est le seul logiciel permettant ces fonctionnalités. A titre de comparaison, la seule alternative, le logiciel COBRAme [23], a permis de regénérer uniquement un modèle ME-FBA pour la bactérie *E. coli K12*, pour laquelle le cadre ME-FBA avait été développé initialement [30]. De plus, le logiciel est peu flexible car certains processus cellulaires sont codés en dur, ce qui limite par exemple l'ajout de nouveaux processus cellulaires dans le modèle.

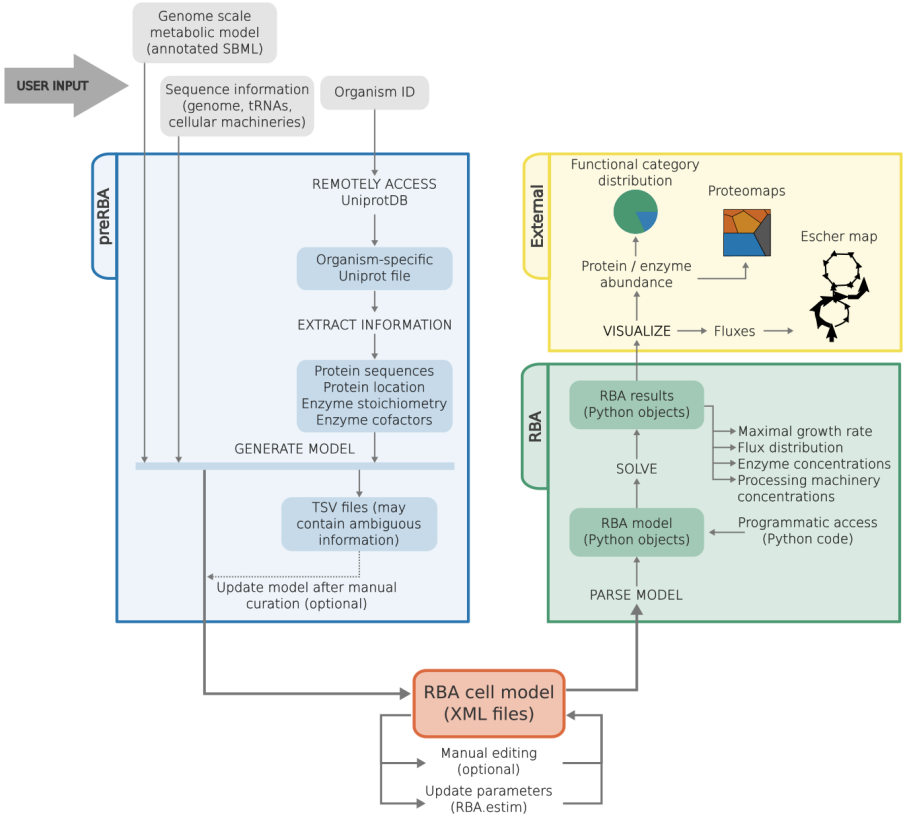

FIGURE 3.5 – **Architecture du package RBApy. Figure extraite de [A.18*].**

Modèles RBA pour d'autres bactéries. A ce jour, nous avons utilisé RBApy pour générer des modèles RBA d'une souche d'*E. coli* modifiée pour fixer le CO2 comme source de carbone [A.18*], des bactéries *Ralstonia solanacearum* en collaboration avec C. Baroukh (LIPME, INRAE), *Streptomyces coelicolor* [ST.2], et la cyanobactérie *Synechocystis sp. PCC6803*. RBApy est également utilisé par d'autres équipes de recherche, pour générer un modèle RBA de la bactérie *Cupriavidus necator* [15].

Perspectives. Par la suite, l'environnement RBApy sera étendu pour inclure les différentes extensions théoriques de la méthode RBA, en particulier la méthode dRBA (cf. chapitre suivant), mais aussi des interfaces conviviales pour explorer ou modifier le modèle [CI.11], ou encore la génération de modèles RBA pour les cellules eucaryotes comme la levure ou les cellules de plante ou humaines (voir la section 3.5 et le chapitre **??**).

3.4 Prédiction des configurations cellulaires en conditions environnementales complexes

Enfin, dans [C.1, A.3*], nous avions montré que la méthode RBA prédisait l'activation ou la répression de certaines réactions métaboliques en fonction de la composition du milieu simulé, et que les gènes codant pour ces réactions appartenaient à des régulons connus et reconstruits dans [A.1*]. En utilisant le modèle RBA calibré, nous avons alors exploré de façon systématique les configurations cellulaires prédites en fonction du milieu extracellulaire dans [A.15]. Nous avons ainsi pu déterminer les règles logiques Booléennes régissant l'activation des flux métaboliques (et donc des enzymes catalysant ces flux) en fonction des différents milieux extracellulaires. Les premiers résultats obtenus sur le métabolisme central de *B. subtilis* sont très encourageants, car les règles inférées correspondent aux régulation locales connues (voir section 2.1 et [A.1*]). De même, la hiérarchie prédite d'utilisation des carbohydrates majoritairement utilisé par *B. subtilis* est cohérente avec celle observée dans [26]. Ainsi, un grand nombre de régulations génétiques présentes dans les voies métaboliques sont cohérentes avec un principe d'allocation parcimonieuse des ressources. Explorer systématiquement les régles d'activation des flux reste également un moyen complémentaire de calibrer/valider la méthode RBA, en comparant les hiérarchies prédites et observées d'utilisation des carbohydrates, acides aminés, etc.

3.5 Extension du cadre RBA pour les cellules eucaryotes

A la différence des cellules procaryotes, la cellule eucaryote est organisée en compartiments cellulaires tels que la mitochondrie ou le chloroplaste. En fonction des conditions environnementales, le nombre de compartiments cellulaires varie en fonction des besoins cellulaires. Pour tenir compte de ces phénomènes, j'ai introduit dans [A.19] de façon explicite la taille des compartiments cellulaires grâce à de nouvelles variables de décision notées f, où chaque composante du vecteur f correspond au volume du compartiment cellulaire normalisé par la volume cellulaire total. De fait, les composantes du vecteur f varient entre 0 et 1. Le problème RBA pour les cellules eucaryotes se formalise de la façon suivante :
$P^e_{rba}(\mu)$: Pour un taux de croissance $\mu \geq 0$ donné, et pour une concentration $P_G \in \mathbb{R}^g_{>0}$ de protéines donnée, on cherche une distribution de ressources (Y, ν, f) faisable, i.e. qui satisfait

les contraintes suivantes :

$$
\begin{array}{ll}
\text{Trouver} & Y \in \mathbb{R}^{N_y}_{\geq 0}, \nu \in \mathbb{R}^{N_m}, f \in \mathbb{R}^{N_c}_{\geq 0}, \\
(C_1) & -\Omega\nu + \mu(C^S_Y Y + C^S_G P_G + C^S_B \bar{B} + C^S_F f \hat{B}) = 0 \\
(C_2) & \mu(C^M_Y Y + C^M_G P_G) - K_T Y \leq 0 \\
 & -K^{'}_E Y \leq \nu \leq K_E Y \\
(C_{3a}) & C^D_Y Y + C^D_G P_G - C^D_F f \leq 0 \\
(C_{3b}) & C^F_F f - \bar{C} = 0 \\
(C_{3c}) & \underline{f}_V \leq I_V f \leq \bar{f}_V
\end{array}
$$

où :

- (C_1) et (C_2) correspondent aux mêmes contraintes que $P_{rba}(\mu)$.
- (C_{3a}) correspond aux contraintes liées à l'occupation des compartiments cellulaires (membranes et/ou volume interne) par les machines moléculaires. Selon les organelles, les contraintes pourront être saturées (contraintes égalités) ou non (contraintes inégalités), au choix du modélisateur. Cette contrainte est comparable à la contrainte (C_3) de $P_{rba}(\mu)$.
- (C_{3b}) contient des contraintes additionnelles permettant de décrire (i) la structure du compartiment cellulaire, comme par exemple lier la surface et le volume du compartiment à travers le ratio volume/surface ; (ii) de relier les volumes de l'ensemble des N_c compartiments, le volume du cytosol f^0, et le volume cellulaire total : $f^0 + \sum_{i=1}^{N^c} f^i = 1$.
- (C_{3c}) contient des contraintes sur le volume minimal et maximal des compartiments cellulaires vis -à-vis du volume cellulaire total. Les vecteurs $\underline{f}_V$ et $\bar{f}_V$ contiennent les valeurs normalisées minimales et maximales des compartiments cellulaires et la matrice I_V est la matrice identité telle que $f_V = I_V f$.

Le problème $P^e_{rba}(\mu)$ est un problème de programmation linéaire (LP), et reste très proche de la formulation pour les cellules procaryotes. À $\mu \geq 0$ fixé, $P^e_{rba}(\mu)$ donne l'ensemble des phénotypes de la cellule compatibles avec la condition environnementale testée. Parmi l'ensemble des phénotypes possibles, on pourra chercher le phénotype compatible avec une allocation parcimonieuse des ressources, i.e. le phénotype correspondant au taux de croissance maximal μ^*, calculé en résolvant une série de problème LP pour différentes valeurs de μ.

Remarque. Les contraintes additionnelles liées à la prise en compte du turnover des protéines et des ARNm ne posent a priori aucune difficulté technique, et se déduisent directement en suivant le même raisonnement que pour les cellules procaryotes [A.19].

Des modèles RBA pour les eucaryotes. Le fait que les formulations de la méthode RBA pour les cellules procaryotes et eucaryotes soient très proches nous permet de réutiliser le logiciel RBApy (en intégrant des modifications liées à la gestion des compartiments cellulaires) pour générer des modèles RBA pour les eucaryotes : la levure *S. cerevisiae* dans le cadre de la thèse d'O. Bodeit [T.9], la cellule mésophylle de la plante *Arabidopsis thaliana* que je développe. Un des objectifs de la thèse [T.9] est d'évaluer la capacité de prédiction de $P^e_{rba}(\mu)$ sur un organisme eucaryote unicellulaire, et de revisiter certaines questions biologiques comme l'effet Crabtree [1]

1. L'effet Crabtree se caractérise par le fait que la levure en condition aérobie et à haute concentration de glucose fermente, i.e. produise de l'éthanol, au lieu de respirer.

du point de vue de l'allocation des ressources. Je présenterai dans le chapitre **??** les premiers résultats obtenus pour *A. thaliana* et discuterai des conséquences en terme de modélisation multi-échelle de la plante.

3.6 Conclusion

Le développement et la validation biologique de la méthode RBA sont incontestablement deux contributions majeures, initiées pendant mes travaux de thèse, et poursuivies par la suite. Depuis les développements théoriques de 2009-2011 et la validation en 2015, plusieurs directions de recherche ont été explorées dans l'équipe BioSys pour étendre le cadre de la méthode RBA du point de vue théorique :

(A) aux fluctuations stochastiques de l'expression des gènes (voir la thèse [11] co-encadrée par M. Dinh, G. Scorletti et V. Fromion) ;

(B) en intégrant des contraintes de thermodynamique et de cinétique enzymatique plus fines pour prédire les concentrations des métabolites. Ces travaux sont menés principalement par mes collègues M. Dinh, S. Peres et V. Fromion. Il s'avère que le problème d'optimisation calculant simultanément les flux métaboliques, les concentrations des enzymes et des métabolites est non-convexe [9] ;

(C) aux conditions dynamiques notamment à travers la thèse [T.6] que j'ai encadrée (voir chapitre suivant).

Au delà de ces extensions théoriques, un modèle RBA a aussi été couplé à un modèle du chemostat par mes collègues M. Dinh et V. Fromion dans [9]. Sous certaines hypothèses raisonnables des caractéristiques cinétiques des transporteurs, le problème d'optimisation associé couplant RBA et chemostat reste convexe et peut se résoudre numériquement. Mes collègues ont alors montré que les configurations cellulaires dans le chemostat pour différents taux de dilution coincident avec les configurations mesurées. Ils utilisent actuellement ce modèle pour étudier des questions d'écologie microbienne, pour déduire notamment sous quelles conditions différentes populations de microorganismes peuvent co-exister au sein d'un chemostat.

La méthode RBA pour la biologie prédictive. Les résultats présentés dans ce chapitre démontrent tout le potentiel de la méthode RBA à prédire la configuration cellulaire des bactéries en conditions environnementales complexes. D'abord, et de façon remarquable, la plupart des grands processus cellulaires ou voies métaboliques semblent régulés en accord avec le principe de parcimonie en ressources. Clairement il existe des exceptions, notamment au niveau de la voie centrale des carbones. Par exemple, le cycle de Krebs de *B. subtilis* est sur-dimensionné, révélant ainsi qu'un autre compromis se joue à ce niveau. Néanmoins, lorsqu'on compare la configuration de la cellule prédite lors de limitations nutritionnelles, comme lors d'une carence en fer, et celle déduite du réseau de régulation (cf. section 2.2), les deux configurations coïncident.
La méthode RBA prédit donc la configuration cellulaire optimale au sens de la parcimonie des ressources, i.e. les concentrations de machines moléculaires, la répartition de flux métabolique et le taux de croissance maximal dans un milieu donné. La composition en machines moléculaires prédite (composées principalement de protéines pour la plupart d'entre elles et de protéines et d'ARNr pour les ribosomes) varie ainsi en fonction du milieu extracellulaire simulé. Nous obtenons ainsi une **prédiction autonome de la composition de la biomasse en fonction**

du taux de croissance, en terme de composition en acides aminés, mais aussi en terme de toute molécule participant à l'activité de la machine moléculaire (ions, cofacteurs, vitamines, etc.). Ainsi, la biosynthèse ou le transport des vitamines et cofacteurs, le transport des ions, s'adapte en fonction de la composition du milieu extracellulaire. A ce titre, la méthode RBA est donc prometteuse pour adresser les challenges de biologie prédictive liés par exemple à la **prédiction des phénotypes d'organismes vivants en conditions de stress combinés**.

Vers un continuum entre prédiction versus simulation. Certains processus cellulaires n'obéissent pas au principe d'utilisation parcimonieuse des ressources. En simulation, il s'agirait de construire un cadre où certaines parties de la cellule seraient simulées de façon fine, en ajoutant des contraintes de fonctionnement (par exemple les régulations génétiques, des quantités minimales de protéines à produire, etc.), et d'autres parties prédites seulement sur la base de contraintes RBA. On obtiendrait ainsi un continuum entre prédiction pure du comportement des cellules sur la base de contraintes RBA, et simulation pure où le comportement de la population de cellules serait totalement imposé.

Principales publications de ce chapitre

A.19 A. Goelzer, V. Fromion. RBA for eukaryotic cells : foundations and theoretical developments, BioRxiv, 2019.

A.18* A. Bulović, S. Fischer, M. Dinh, F. Golib, W. Liebermeister, C. Poirier, L. Tournier, E. Klipp, V. Fromion, A. Goelzer. Automated generation of bacterial resource allocation models. *Metabolic Engineering*, 55 :12-22, 2019.

A.11* A. Goelzer, J. Muntel, V. Chubukov, M. Jules, E. Prestel, R. Nolker, M. Mariadassou, S. Aymerich, M. Hecker, P. Noirot, D. Becher, V. Fromion. Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic Engineering*, 32 :232-243, 2015.

A.3* A. Goelzer, V. Fromion, G. Scorletti. Cell design in bacteria as a convex optimization problem. *Automatica*, 47(6) :1210-1218, 2011.

C.1 A. Goelzer, V. Fromion, and G. Scorletti. Cell design in bacteria as a convex optimization problem. *Proceedings of the 48th IEEE Conference on Decision and Control*, pages 4517-4522, December 2009.

Chapitre 4

Allocation des ressources en régime dynamique et application à la biologie de synthèse

La méthode RBA présentée au chapitre précédent a été développée en régime établi, i.e. en phase exponentielle de croissance à un taux de croissance constant. Dans ce chapitre, je vais présenter l'extension de la méthode RBA en conditions dynamiques (appelée dRBA), et ses conséquences pour les biotechnologies et la biologie de synthèse.

Contrairement au chapitre précédent où nous avions été jusqu'à la validation expérimentale de la méthode, les résultats présentés dans ce chapitre sont essentiellement théoriques, et réalisés principalement avec mon collègue V. Fromion, et dans le cadre de la thèse en automatique de G. Jeanne que je supervisais en collaboration avec mes collègues du laboratoire L2S de CentraleSupélec (S. Tebbani, D. Dumur).

4.1 RBA en conditions dynamiques

Mes premiers travaux sur ce thème ont débuté dès 2015, à la suite de la validation biologique de la méthode RBA. J'avais développé une première extension théorique de la méthode RBA aux conditions dynamiques (dRBA) dont les résultats en simulation semblaient très prometteurs. Par exemple, les cinétiques de disparition des substrats des milieux de culture simulées correspondaient bien à celles mesurées dans [A.11*]. De même, le phénomène de diauxie, i.e. deux phases de croissance successives correspondant à l'utilisation successive de deux sources carbones, était qualitativement correctement prédit (cf. Figure 4.1 à titre illustratif).

Plus récemment, nous avons consolidé le cadre théorique de la méthode dRBA dans [A.23*]. La méthode dRBA intègre des équations différentielles discrétes décrivant l'évolution au cours du temps des machines moléculaires, des composés du milieu extracellulaire, et du volume total de la population bactérienne, ainsi que des contraintes RBA à chaque pas de discrétisation. Lors du développement du cadre dRBA, nous avons été particulièrement vigilants sur (*i*) le schéma de discrétisation, (*ii*) le fait que la conservation de la masse doit être satisfaite à chaque pas de temps, et (*iii*) la tractabilité du problème d'optimisation. Pour réduire les difficultés numériques possibles (instabilités liées à des étapes excessives de discrétisation), nous avons

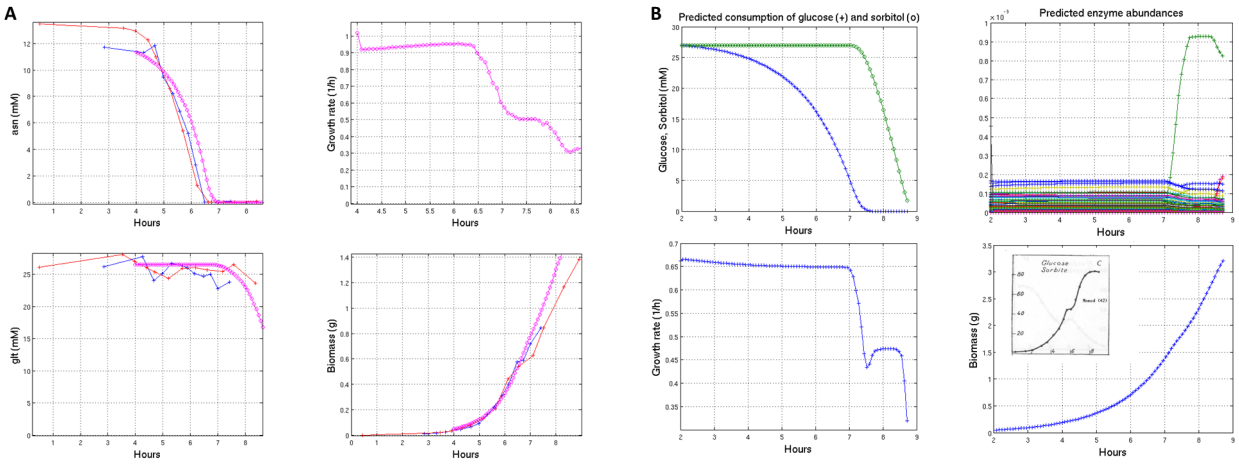

FIGURE 4.1 – **A. Cinétique prédite de disparition de quelques substrats (asparagine : asn, glutamate : glt, en mM), taux de croissance(1/h) et quantité de biomasse (g) en milieu riche (voir [A.11*]). Les courbes bleues et rouges correspondent aux cinétiques mesurées, et en magenta celles prédites. B. Cinétique de croissance prédite dans un milieu contenant deux sources de carbone (glucose et sorbitol en mM), taux de croissance (1/h), et concentration d'enzymes en mM par gramme de poids sec. Les deux simulations ont été obtenues en résolvant un problème dRBA avec un algorithme de commande prédictive.**

d'abord discrétisé le problème continu en utilisant un schéma de discrétisation implicite d'Euler. Ensuite, le point (iii) était particulièrement crucial pour espérer résoudre un problème dRBA sur plusieurs pas de temps, avec un modèle à l'échelle du génome comme celui de *B. subtilis* [A.11*]. La difficulté était d'obtenir une formulation convexe du problème, afin de pouvoir le résoudre de façon efficace. Pour cette raison, nous avons considéré des nombres de molécules n_* au lieu de concentrations, qui sont par définition non linéaires et égales à $\frac{n_*}{X}$ avec X le volume d'une bactérie ou d'un ensemble de bactéries. Les contraintes qui doivent être satisfaites à chaque temps k sont définies par :
P_{drba_k} : Pour $\tilde{Z}^k \in \mathbb{R}^{y+g+1}_{\geq 0}, \tilde{A}^k \in \mathbb{R}^s_{\geq 0}, \geq 0$ fixés,

$$
\begin{array}{ll}
\text{Trouver} & \tilde{u}^k, \tilde{Z}^{k+1} \in \mathbb{R}^{y+g+1}_{\geq 0}, \tilde{\nu}^k \in \mathbb{R}^m, \tilde{A}^{k+1} \in \mathbb{R}^a_{\geq 0}, \\
\text{soumis à} & \\
(C^k_1) & -\Omega\tilde{\nu}^k + C^S_U\tilde{u}^k + C^S_Z\tilde{Z}^{k+1} = 0 \\
(C^k_{2a}) & C^M_U\tilde{u}^k - K_T\tilde{Z}^{k+1} \leq 0 \\
(C^k_{2b}) & -K_E\tilde{Z}^{k+1} \leq \tilde{\nu}^k \leq K_E\tilde{Z}^{k+1} \\
(C^k_{3a}) & C^{D,c}_U\tilde{u}^{k,c} = 0 \\
(C^k_{3b}) & C^{D,s}_U\tilde{u}^{k,s} \leq 0 \\
(C^k_{4a}) & \tilde{Z}^{k+1} = \tilde{Z}^k + \tilde{u}^k\delta t \\
(C^k_{4b}) & \tilde{u}^k_G = \bar{P}_G u^k_X \\
(C^k_{4c}) & \tilde{A}^{k+1} = \tilde{A}^k + \tilde{\nu}^k_a\delta t
\end{array}
$$

Les composantes du vecteur $\tilde{Z}^{k+1} \triangleq (\tilde{Y}^{k+1}, \tilde{P}^{k+1}_G, X^{k+1})^T$ correspondent respectivement au nombre de machines moléculaires, au nombre de protéines $\mathbb{P}_G$, au volume de la population de

bactéries au temps $k+1$. Le vecteur $\tilde{A}^{k+1}$ de taille N_a correspond au nombre de molécules de substrats et de produits excrétés dans le milieu extracellulaire au temps $k+1$. Les variables de décision $\tilde{u}^k \triangleq (\tilde{u}_Y^k, \tilde{u}_G^k, u_X^k)^T$ correspondent respectivement au nombre de machines $\mathbb{Y}$, $\mathbb{P}_G$ produites, et à l'augmentation de volume entre le temps k et $k+1$. Le vecteur $\tilde{\nu}_a^k$ correspond aux flux importés ou excrétés de métabolites extracellulaires au temps k. De plus, on suppose que la trajectoire de $\tilde{P}_G$ est constante au cours du temps et égale à $\bar{P}_G$, ce qui conduit à la contrainte (C_{4b}^k). Ainsi, pour X_0, A_0 et $\bar{P}_G$ donnés, le problème d'optimisation (appelé dRBA) suivant :

Trouver
pour tout $k \in 0 \dots N-1$,
$(\tilde{u}^k), (\tilde{Z}^{k+1}) \in \mathbb{R}_{\geq 0}^{y+g+1}$,
$(\tilde{\nu}^k) \in \mathbb{R}^m, (\tilde{A}^{k+1}) \in \mathbb{R}_{\geq 0}^a$,
$\tilde{Y}^0 \in \mathbb{R}_{\geq 0}^y$
soumis à

$$\begin{aligned} &(\text{drba})_{(k)_{0\dots N-1}} \\ &X^0 = X_0,\ \tilde{A}^0 = A_0,\ \tilde{P}_G^0 = \bar{P}_G X_0 \\ &C_U^{D,c}\tilde{Y}^{0,c} \leq 0,\ C_U^{D,s}\tilde{Y}^{0,s} \leq 0 \end{aligned}$$

nous donne accès à l'évolution de l'ensemble des trajectoires possibles. On notera que nous avons ajouté les conditions à $t(0) = 0$ relativement peu contraintes. On recherche des quantités de machines moléculaires qui satisfont les contraintes de densité de volume et de surface ($C_U^{D,c}\tilde{Y}^{0,c} \leq 0$, $C_U^{D,s}\tilde{Y}^{0,s} \leq 0$). Les trajectoires des entités moléculaires sont alors calculées sur l'ensemble des pas de temps en maximisant ou en minimisant un certain critère. Dans [A.22*], nous avons exploré quel était le critère à maximiser, ou si d'éventuelles contraintes étaient à rajouter afin d'obtenir le comportement attendu d'une population bactérienne en mode batch, i.e. une phase exponentielle de croissance. Maximiser la biomasse finale (X^N) sur un intervalle de temps fixé et un critère d'arrêt donné ($\tilde{A}^N = A_N$, $A_N \geq 0$ étant fixé) ne permet pas d'obtenir une phase de croissance exponentielle. En revanche, dès que quelques compétiteurs sont présents dans le milieu et consomment des ressources extracellulaires, la maximisation de la biomasse finale permet de retrouver la phase exponentielle de croissance.

Commande optimale et commande prédictive. Le problème enoncé ci-dessus correspond à chercher la trajectoire optimale maximisant la biomasse sur un horizon de temps donné. On résout numériquement un problème de commande optimale discrétisé sous contraintes mixtes entre l'état et la commande (contraintes C_{2a}^k et C_{2b}^k). L'avantage est d'obtenir les trajectoires optimales sur l'intervalle de temps d'intérêt. L'inconvénient réside en la taille du problème d'optimisation associé. En effet, comme toutes les contraintes d'égalités et d'inégalités, et le critère à maximiser sont linéaires, le problème dRBA est un problème de programmation linéaire. Le nombre de variables de décision et de contraintes augmente linéairement avec le nombre de pas de temps. Cependant, même si la taille du problème dRBA augmente linéairement, elle peut rapidement devenir très grande. Par exemple, pour le modèle calibré de [A.11*], le problème dRBA construit sur 10 (resp. 70) pas de temps contient 18580 (resp. 130060) variables de décision, 12320 (resp. 86240) contraintes d'inégalités et 10960 (resp. 76720) contraintes d'égalités. Une alternative consiste à simuler la trajectoire globale par un algorithme de commande prédictive à horizon glissant [2]. En Automatique, la commande prédictive est une optimisation en temps réel qui vise à trouver la commande à appliquer à un système discret pour maximiser un critère étant donné son état actuel et la trajectoire de référence calculée hors-ligne.

A chaque pas de temps k, on résout un problème de type dRBA [1] pour les n_H pas de temps suivants. Ceci nous donne un vecteur de commande à appliquer aux n_H pas d'échantillonnage $(\tilde{u}^k, \tilde{u}^{k+1}, \ldots, \tilde{u}^{k+n_H})$, dont on n'applique que la première composante, i.e. la commande $\tilde{u}^k$. Puis on passe au pas de temps suivant $k \leftarrow k+1$ et on réitère l'opération. On résout ainsi environ N problèmes d'optimisation de taille réduite, bien plus rapides à résoudre. Numériquement les trajectoires obtenues par les algorithmes de commande optimale et de commande prédictive sont très proches. Une perspective intéressante serait de le démontrer formellement du point de vue théorique.

Perspectives. A très court terme, il s'agit donc d'une part de valider expérimentalement la méthode dRBA, en collaboration avec nos partenaires biologistes de l'UMR MICALIS de l'INRAE, et d'autre part d'étendre le cadre théorique aux cellules eucaryotes.

4.2 Optimisation conjointe d'un bioprocédé et des souches bactériennes

Les résultats préliminaires obtenus en 2015 nous ont permis d'initier la thèse de G. Jeanne [T.6]. Dans cette thèse, il s'agissait d'établir une preuve de concept où le bioprocédé et la souche bactérienne sont optimisés simultanément pour maximiser la production d'un composé d'intérêt. Cette thèse s'appuie sur un nouveau modèle de bioprocédé couplant (i) une description macroscopique du fonctionnement des bioprocédés utilisés traditionnellement pour la commande de bioréacteurs et (ii) une description du comportement des microorganismes intégrant les échelles infra-cellulaires. Traditionnellement, le comportement des microorganismes au sein d'un bioprocédé est décrit par un modèle phénoménologique de type Monod [26]. Nous avons substitué ce modèle phénoménologique par un modèle dynamique de cellule simplifiée reprenant les éléments clés de la méthode RBA : (i) les voies métaboliques catalysées par des machines moléculaires, (ii) la synthèse des machines moléculaires (y compris les ribosomes) ; (iii) la densité cytosolique est constante [2]. Le problème d'optimisation associé correspond à un problème de Mayer (le critère à optimiser ne dépend que du temps final) avec des contraintes mixtes sur l'état et la commande du système. La résolution analytique de ce type de problème d'optimisation est difficile [42]. Nous nous sommes donc tournés vers des méthodes numériques pour résoudre le problème d'optimisation. Fort de nos travaux sur l'expression des gènes (voir section 2.1 et 2.3), nous avons également considéré (i) différentes stratégies de contrôle de l'expression des gènes (libre, gènes constitutifs pilotés par le taux de croissance) à travers des contraintes de fonctionnement supplémentaires et (ii) différents types de bioprocédés opérant en mode batch ou fed-batch (voir [T.6, C.7, C.9*]).

1. Le problème dRBA de commande prédictive est identique à celui associé à la commande optimale où la biomasse finale est maximisée (sans critère d'arrêt), à l'exception des conditions au bord à $t(0) = k$ qui sont maintenant imposées, et telles que le vecteur d'état à l'instant k corresponde au vecteur d'état calculé au pas précédent.

2. Chez les bactéries en forme de bâtonnet, la densité cytosolique est effectivement constante [21]. Pour des bactéries en forme de coque, ou pour les levures, on peut observer des variations de densité.

4.3 Vers la conception assistée de microorganismes

La preuve de principe obtenue dans la thèse [T.6] doit être consolidée et étendue à l'échelle du génome, et à plus long terme validée expérimentalement. L'idée serait de développer un *workflow* d'outils méthodologiques et expérimentaux permettant d'optimiser conjointement le bioprocédé et les microorganismes : (*i*) prédiction des profils d'expression optimaux des microorganismes et de la stratégie de contrôle du bioprocédé pour maximiser la production d'un composé d'intérêt ; (*ii*) analyse des trajectoires optimales d'expression au regard de la décomposition en modules fonctionnels de [A.1*] pour déterminer un sous-ensemble de gènes à modifier ; (*iii*) déterminer les bio-briques dans la banque de promoteurs et de TIRs [A.13*] permettant de réaliser les profils d'expression optimaux. (*iv*) construire les souches modifiées de microorganismes et re-adapter la souche par évolution dirigée. (*v*) évaluer les stratégies de contrôle optimales du bioprocédé prédites en utilisant les microorganismes optimisés.

Principales publications de ce chapitre

A.23* A. Goelzer, V. Fromion. Optimization and games theory to investigate the bacterial behaviors in batch mode at the genome scale. *pre-print*, 2019.

C.9* G. Jeanne, A. Goelzer, S. Tebbani, V. Fromion, D. Dumur. Towards a realistic and integrated strain design in batch bioreactor. In *Proceedings of the 57th IEEE Conference on Decision and Control*, pages 2698-2703, Miami, USA, 2018.

C.7 G. Jeanne, A. Goelzer, S. Tebbani, V. Fromion, D. Dumur. Dynamical resource allocation models for bioreactor optimization. In *Foundations of Systems Biology in Engineering (FOSBE)*. Chicago, USA, 2018. 51(19) : 20-23.

Chapitre 5

Intégration de connaissances et de données biologiques hétérogènes

Ce chapitre regroupe un ensemble de travaux transversaux à mes activités de recherche en modélisation, avec tout d'abord un ensemble d'outils dédiés au traitement et à la visualisation de données biologiques. Ces outils ont été développés conjointement aux modèles présentés dans les chapitres précédents. Puis je présenterai mes activités de recherche en représentation des connaissances pour la biologie. Ces travaux ont été initiés dans le cadre du projet Lidex-IMSV de l'université Paris-Saclay (coordonné par V. Fromion), et menés avec mes collègues informaticiennes du Laboratoire de Recherche en Informatique (LRI) (C. Froidevaux, F. Saïs) et de MIA-Paris (J. Dibie). Je coordonnais le développement des travaux, et supervisais directement l'ingénieur et le post-doctorant (A. Ferré, V. Henry) dans ces travaux de recherche. Ces activités de recherche se poursuivent actuellement dans le cadre de la thèse en informatique d'O. Inizan [T.10], que je co-encadre en collaboration avec F. Saïs (LRI), D. Symeonidou (MISTEA,INRAE) et V. Fromion.

5.1 Outils pour le traitement, l'analyse et la visualisation de données biologiques

Dans le cadre des projets de biologie de système et de synthèse auxquels j'ai participé, j'ai développé de nombreux outils de traitement de données biologiques (transcriptomique, protéomique, métabolique, fluxomique, live cell array). Ce volet d'activités est historiquement le plus ancien et constitue le socle sur lequel s'est appuyé le développement des modèles [A.8*, A.11*, A.13*]. Si les algorithmes de traitements de données sont standards en eux-mêmes (par exemple un filtre de Kalman étendu pour le traitement de données LCA dans [A.9*]), l'originalité réside en l'analyse fonctionnelle des données. Les données analysées comme la transcriptomique dans [A.5] ou la protéomique dans [A.10*] sont systématiquement mises en perspective vis-à-vis de l'organisation de la cellule (réseaux de régulation, organisation chromosomique sous forme d'opérons, et modules fonctionnels définis dans [A.1*]). Grâce aux résultats théoriques obtenus dans [T.1, B.2*], les variations qualitatives des composants des modules fonctionnels (principalement les métabolites, les flux métaboliques et les concentrations d'enzymes) peuvent être prédites. Nous avons pu ainsi analyser des jeux de données hétérogènes composés de transcriptomique, fluxomique et métabolique acquis pour différents milieux de culture [T.1, A.6], et déterminer si les variations observées étaient cohérentes avec le comportement qualitatif attendu des modules

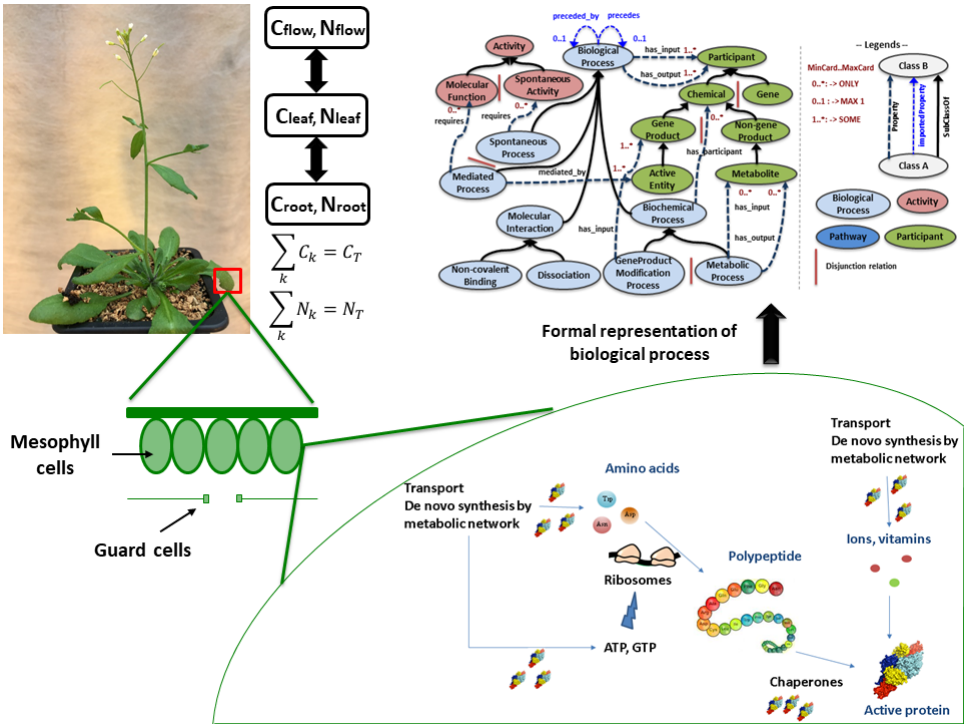

FIGURE 5.1 – **Représentation systémique multi-échelle d'une plante. (Gauche) La plante est composée d'organes, et chaque organe est composé de plusieurs populations de cellules spécialisées (par exemple les cellules mésophylles dans la feuille). (Bas) Les cellules sont constituées de composés cellulaires (par exemple métabolites, protéines, ribosomes, etc.) en interaction à travers les processus biologiques (par exemple la traduction). (Droite). Représentation formelle des processus biologiques par des ontologies.**

fonctionnels des voies métaboliques.

5.2 Vers une représentation formelle des organismes vivants

La question de la description des systèmes vivants n'est pas nouvelle et recoupe des efforts importants de diverses communautés, en particulier la communauté bioinformaticienne. Malgré les progrès réalisés, une partie de la question semble échapper aux approches actuelles. De façon plus précise, le problème ne réside pas dans la description des entités composant les systèmes vivants (par exemple génomes, ARN messagers, protéines, etc.) mais plutôt sur la façon de décrire les liens plus ou moins complexes qui les relient. Dans ce contexte, la biologie des systèmes adjoint au « catalogue des entités cellulaires », la description des processus biologiques dans lesquels interviennent ces entités. Cette façon de faire n'est pas originale en soi puisqu'elle est commune en Science de l'Ingénieur, où il s'agit, pour un système donné, de rattacher ses composantes aux sous-systèmes et de conditionner les propriétés de ces composantes aux sous-systèmes.

Par ailleurs, l'approche système est pertinente dans le champ biologique, au regard de la nature modulaire du vivant. Elle est déjà largement utilisée pour décrire les organismes supérieurs, à l'échelle de l'individu. Prenons l'exemple de la plante (voir Figure 5.1) : une plante est composée d'organes interconnectés par le système vasculaire (xylème/phloème) et assurant chacun une

fonction spécifique pour la plante (par exemple, la feuille fixe le CO2 atmosphérique). Chaque organe se décompose ensuite en différents tissus composés de plusieurs types de cellules, chaque type de cellules a des spécificités propres et constitue un sous-système à part entière. Il en résulte une organisation modulaire. Ce type de décomposition a donné lieu à des modèles mathématiques à compartiments comme les modèles écophysiologiques (à l'échelle de l'individu) ayant une grande capacité de prédiction en mode nominal (non perturbé) [39]. Les progrès récents en biologie des systèmes ont montré que l'approche système reste également adéquate et pertinente pour décrire le fonctionnement des échelles infra-cellulaires. Nous avons montré par exemple dans [A.1*] que le réseau métabolique peut être découpé en sous-systèmes (au sens des Sciences de l'Ingénieur). L'approche système est à la base de la méthode RBA et plus largement à la base des premiers modèles mathématiques de cellules entières ayant une capacité de prédiction importante (voir [A.11*] et [16]). Cela signifie que les principes sous-jacents à ces modèles (ici, la systémique) captent bien le fonctionnement du système global.

Ces dernières années, une part significative de mon activité de recherche a consisté à mobiliser l'approche système du point de vue formel pour décrire un organisme vivant. Ainsi dans [A.16*], en collaboration avec nos collègues du LRI, nous avons construit la première ontologie *systémique*, BiPON, où les processus biologiques sont représentés formellement sous forme de systèmes au sens des Sciences de l'Ingénieur, i.e. un système possédant des entrées, des sorties et un modèle mathématique associé décrivant le comportement du système. Les entités cellulaires comme les métabolites ou les protéines sont définies et typées comme des entrées et des sorties des processus biologiques, et les modèles mathématiques peuvent être associés aux processus biologiques par raisonnement automatique sur BiPON. Nous nous sommes attachés à définir une représentation la plus générique possible, ce qui se traduit par un ensemble minimal de classes et de propriétés dans BiPON. Comme preuve de principe, nous avons décrit un ensemble de processus biologiques suffisamment représentatif de la complexité des processus bactériens : les processus impliqués dans l'expression des gènes. Dans les articles [C.5, A.20], nous avons encore progressé dans le niveau d'abstraction de l'ontologie, afin d'utiliser au mieux les capacités des raisonneurs automatiques pour enrichir l'information contenue dans l'ontologie. Cette nouvelle ontologie systémique, appelée BiPOm, a été utilisée pour décrire le fonctionnement du cycle de Calvin d'*Arabidopsis thaliana*, à savoir les réactions métaboliques ainsi que les processus post-traductionnels impliqués dans la formation des complexes enzymatiques, notamment la RuBisCO. De plus, dans [A.20], nous avons montré qu'un réseau métabolique entier, celui d'*E. coli*, pouvait être instancié dans l'ontologie, et que le raisonnement automatique restait possible malgré le grand nombre d'instances (plus de 10,000). Il s'agit maintenant d'étendre BiPOm, en intégrant typiquement certaines classes et propriétés de BiPON, pour pouvoir représenter l'ensemble des processus biologiques d'une cellule vivante. Ce travail fait d'ailleurs l'objet de la thèse d'O. Inizan [T.10], thèse qui vient de débuter dans l'équipe.

A terme, il s'agit ici de construire un cadre générique de représentation des cellules (et de façon plus large des organismes vivants) formalisant la connaissance biologique (par exemple le réseau métabolique, les réseaux de régulations, etc.) sous forme d'ontologies et d'utiliser cette ontologie pour organiser les données expérimentales associées : les données de différente nature obtenues à l'échelle du génome entier (les « omiques » : transcriptome, protéome, fluxome, métabolome, phénome, etc) ; les données issues de technologies de séquençage haut-débit (RNA-seq, Chip-seq, Ribosome profiling, etc.), de traitement d'images, de spectres RMN, de chemotypage. Ces données sont généralement acquises dans le temps et/ou associées à des variations

de conditions expérimentales. Il s'agit donc de décrire un ensemble (très) significatif des parties de la cellule et des paramètres qui vont avec, d'être capable de le faire « pour n'importe quel organisme » et d'avoir la possibilité d'ancrer sur l'objet biologique n'importe quel type de données expérimentalement acquises ou prédites (par exemple un motif de fixation d'un facteur de transcription prédit). Au sein des collectifs d'ingénieurs que j'ai coordonnés (les CATI MIAGO et SYSMICS, cf. section **??** de la première partie de cette HDR), ce dernier point a déjà fait l'objet de reflexions approfondies. Pour l'ensemble des données mentionnées ci-dessus, nous avons étudié comment décrire (de façon générique) les procédures d'acquisition des données expérimentales, et si les données produites et traitées pouvaient être associées à une classe déjà présente dans les ontologies BiPON/BiPOm [S.4]. Du point de vue pratique, une partie significative des composantes, des propriétés et des paramètres associés est soit déjà accessible dans des bases de données comme Uniprot [1], CheBI [8], etc., soit prédictible sur la base de méthodes bioinformatiques, de statistiques, ou bien encore obtenue sur la base de la combinaison de ces méthodes et des données expérimentales. La modélisation systémique correspond ainsi à un changement de paradigme où les processus biologiques (i.e. les sous-systèmes) sont remis au centre de la description de la cellule. La rupture est conceptuelle, mais du point de vue technique, il s'agit avant tout de récupérer l'information existante sur les composants cellulaires, et de renseigner les parties manquantes, et plus particulièrement les processus biologiques. La modélisation systémique constitue ainsi un cadre adéquat pour l'intégration de données hétérogènes et vient compléter les méthodes d'intégration statistiques de données hétérogènes en apportant de nouvelles informations (les processus biologiques).

Dans le cadre du projet IMSV, nous avons développé un premier prototype d'environnement informatique permettant une intégration systémique des données -omiques hétérogènes et des connaissances biologiques pour la bactérie *B. subtilis*, et structuré autour de trois grandes entités :

1. l'ontologie BiPON décrivant formellement les processus cellulaires d'un organisme de façon systémique,
2. un entrepôt de données hétérogènes structuré par l'ontologie et regroupant de la connaissance biologique, des données expérimentales (par ex., séquençage, transcriptomique, etc.), des prédictions bioinformatiques (par ex. détections de motifs, etc.) et des modèles mathématiques associés aux processus cellulaires,
3. un simulateur stochastique de l'expression des gènes [A.21] couplé à l'ontologie/entrepôt pour extraire l'information nécessaire.

Ce prototype ne contient actuellement qu'un nombre limité de processus cellulaires, mais il a vocation à être étendu à la cellule entière. Dans tous les cas, le cadre proposé semble pertinent et adéquat pour représenter et intégrer des connaissances et des données biologiques hétérogènes.

5.3 Conclusion

Les travaux sur la représentation des systèmes vivants présentés dans ce chapitre seront poursuivis. Traditionnellement, les ontologies liées à la biologie cellulaire, comme la Gene Ontologie [7], sont utilisées principalement comme vocabulaire contrôlé et exploitent peu le raisonnement automatique. Dans ce contexte, les ontologies BiPON et BiPOm sont innovantes et un levier formidable de compilation des connaissances par rapport à l'existant, car d'une part elles relient molécules, processus biologique et modèles mathématiques du point de vue formel, et d'autre

part elles s'appuient sur le raisonnement automatique pour inférer de nouvelles connaissances. A terme, l'objectif est de démontrer qu'un organisme vivant, comme la bactérie *B. subtilis*, ou une cellule de la plante *A. thaliana* peut être décrit dans sa totalité en suivant notre approche de la section 5.2, et que des entrepôts de données structurés par nos ontologies peuvent être également construits. Ce dernier point nécessitera des moyens en ingénierie, et devra exploiter au mieux les infrastructures ou bases de données existantes au sein des unités de recherche (MaIAGE, IJPB, IPS2).

Principales publications de ce chapitre

A.20 V. Henry, F. Saïs, O. Inizan, E. Marchadier, J. Dibie, A. Goelzer*, V. Fromion. BiPOm : a rule-based ontology to represent and infer molecule knowledge from a biological process-centered viewpoint *BMC Bioinformatics*, 21 :327, 2020. (* corresponding author)

A.16* V. Henry, A. Goelzer*, A. Ferré, S. Fischer, M. Dinh, V. Loux, C. Froidevaux, V. Fromion. The Bacterial interlocked Process ONtology (BiPON) : a systemic multi-scale unified representation of biological processes in prokaryotes. *Journal of Biomedical Semantics*, 8(1) :53, 2017. (* corresponding author)

A.10* J. Muntel, V. Fromion, A. Goelzer, S. Maass, U. Mäder, K. Büttner, M. Hecker, D. Becher. Comprehensive absolute quantification of the cytosolic proteome of *Bacillus subtilis* by data independent, parallel fragmentation in liquid chromatography/mass spectrometry (LC/MSE). *Molecular and Cellular Proteomics*, 13(4) :1008-1019, 2014.

A.9* L. Aïchaoui, M. Jules, L. Le Chat, S. Aymerich, V. Fromion, A. Goelzer. Basylica : a tool for automatic processing of a bacterial live cell array. *Bioinformatics*, 28(20) :2705-2706, 2012.

Bibliographie

[1] UniProt : the universal protein knowledgebase. *Nucleic acids research*, 45(D1) :D158–D169, 2016.

[2] F. Allgöwer and A. Zheng. *Nonlinear model predictive control*, volume 26. Birkhäuser, 2012.

[3] M. Arcak and E. D. Sontag. Diagonal stability of a class of cyclic systems and its connection with the secant criterion. *Automatica*, 42(9) :1531–1537, 2006.

[4] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization : analysis, algorithms, and engineering applications*. MPS/SIAM Series on Optimization, 2001.

[5] F. J. Bruggeman and H. V. Westerhoff. The nature of systems biology. *TRENDS in Microbiology*, 15(1) :45–50, 2007.

[6] C. Chassagnole, N. Noisommit-Rizzi, J.W. Schmid, K. Mauch, and M. Reuss. Dynamic modeling of the central carbon metabolism of escherichia coli. *Biotechnology and bioengineering*, 79(1) :53–73, 2002.

[7] Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1) :D258–D261, 2004.

[8] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI : a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1) :D344–D350, 2007.

[9] M. Dinh and V. Fromion. RBA like problem with thermo-kinetics is non convex. *arXiv preprint arXiv :1706.01312*, 2017.

[10] J. Doudna and E. Charpentier. The new frontier of genome engineering with crispr-cas9. *Science*, 346(6213), 2014.

[11] Marouane Ait El Faqir. *Prédiction de la structure de contrôle de bactéries par optimisation sous incertitude*. PhD thesis, Ecole Centrale Lyon ; Université de Lyon, 2016.

[12] L. Gerosa, K. Kochanowski, M. Heinemann, and U. Sauer. Dissecting specific and global transcriptional regulation of bacterial gene expression. *Molecular systems biology*, 9(1), 2013.

[13] A. Heinken and I. Thiele. Systems biology of host–microbe metabolomics. *Wiley Interdisciplinary Reviews : Systems Biology and Medicine*, 7(4) :195–219, 2015.

[14] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, et al. The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4) :524–531, 2003.

[15] M. Jahn, N. Crang, M. Janasch, A. Hober, B. Forsström, K. Kimler, A. Mattausch, Q. Chen, J. Asplund-Samuelsson, and E. Hudson. Protein allocation and utilization in the versatile chemolithoautotroph cupriavidus necator. *bioRxiv*, 2021.

[16] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival Jr, N. Assad-Garcia, J. I. Glass, and M. W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2) :389–401, 2012.

[17] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson. Escher : a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS computational biology*, 11(8) :e1004321, 2015.

[18] H. Kitano. *Foundations of systems biology*. The MIT Press Cambridge, Massachusetts London, England, 2001.

[19] O. Kotte, J. B. Zaugg, and M. Heinemann. Bacterial adaptation through distributed sensing of metabolic fluxes. *Molecular systems biology*, 6(1) :355, 2010.

[20] O. Kotte, J.B. Zaugg, and M. Heinemann. Bacterial adaptation through distributed sensing of metabolic fluxes. *Molecular systems biology*, 6(1) :355, 2010.

[21] H. E. Kubitschek, W. W. Baldwin, S. J. Schroeter, and R. Graetzer. Independence of buoyant cell density and growth rate in *Escherichia coli*. *Journal of bacteriology*, 158(1) :296–299, 1984.

[22] W. Liebermeister, E. Noor, A. Flamholz, D. Davidi, J. Bernhardt, and R. Milo. Visual account of protein investment in cellular functions. *Proceedings of the National Academy of Sciences*, 111(23) :8488–8493, 2014.

[23] C. Lloyd, A. Ebrahim, L. Yang, Z. King, E. Catoiu, E. O'Brien, J. Liu, and B. Palsson. Cobrame : A computational framework for genome-scale models of metabolism and gene expression. *PLoS computational biology*, 14(7) :e1006302, 2018.

[24] H.-Wu. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, and A.-P. Zeng. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic acids research*, 32(22) :6643–6649, 2004.

[25] A. G. Marr. Growth rate of escherichia coli. *Microbiology and Molecular Biology Reviews*, 55(2) :316–333, 1991.

[26] J. Monod. Recherches sur la croissance des cultures bacteriennes. 1942.

[27] M. Mori, T. Hwa, O. C. Martin, A. De Martino, and E. Marinari. Constrained allocation flux balance analysis. *PLoS computational biology*, 12(6) :e1004913, 2016.

[28] F. Neidhart. *Escherichia coli and salmonella : cellular and molecular biology*. American Society of Microbiology Press, Washington D.C., USA, 2nd edition, 1996.

[29] Y. Nesterov. *Introductory lectures on convex optimization : a basic course*. Kluwer Academic Publishers, 2004.

[30] E. J. O'brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. Ø. Palsson. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9(1) :693, 2013.

[31] J. Orth, T. Conrad, J. Na, J. Lerman, H. Nam, A. Feist, and B. Palsson. A comprehensive genome-scale reconstruction of escherichia coli metabolism-2011. *Molecular systems biology*, 7(1) :535, 2011.

[32] J.D. Orth, I. Thiele, and B.Ø. Palsson. What is flux balance analysis ? *Nature biotechnology*, 28(3) :245–248, 2010.

[33] S. Park, D. Yang, S. Ha, and S. Lee. Metabolic engineering of microorganisms for the production of natural compounds. *Advanced Biosystems*, 2(1) :1700190, 2018.

[34] A. Penesyan, M. Gillings, and I. Paulsen. Antibiotic discovery : combatting bacterial resistance in cells and in biofilm communities. *Molecules*, 20(4) :5286–5298, 2015.

[35] A. Sahu, M. Blätke, J. Szymański, and N. Töpfer. Advances in flux balance analysis by integrating machine learning and mechanism-based models. *Computational and Structural Biotechnology Journal*, 2021.

[36] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, and M. Heinemann. The quantitative and condition-dependent *Escherichia coli* proteome. *Nature biotechnology*, 34(1) :104, 2016.

[37] A. Sonenshein, J. Hoch, and R. Losick. *Bacillus subtilis and other gram-positive bacteria. Biochemistry, physiology and molecular genetics.* American Society of Microbiology, 1993.

[38] A. Sonenshein, J. Hoch, and R. Losick. *Bacillus subtilis and its closest relatives, from genes to cells.* American Society of Microbiology, 2001.

[39] A. Tuzet, A. Perrier, and R. Leuning. A coupled model of stomatal conductance, photosynthesis and transpiration. *Plant, Cell & Environment*, 26(7) :1097–1116, 2003.

[40] B. R. B. H. Van Rijsewijk, A. Nanchen, S. Nallet, R. J. Kleijn, and U. Sauer. Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. *Molecular systems biology*, 7(1), 2011.

[41] A. Varma and B. O. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli W3110*. *Applied and environmental microbiology*, 60(10) :3724–3731, 1994.

[42] R. Vinter. *Optimal control*. Springer Science & Business Media, 2010.