Introduction

0.1 Modèle *in silico* pour étudier la microbiologie

Les simulations informatiques sont utiles pour décrire le fonctionnement des populations d'un écosystème bactérien. Au commencement de la microbiologie, les scientifiques se concentraient sur les maladies infectieuses, affectant à la fois les hommes et les animaux. Plus tard, les chercheurs et chercheuses ont découvert que certaines bactéries étaient responsables de certaines pathologies, mais aussi, des processus de fermentations comme dans le vin (**?**). Afin de les analyser biologiquement, une première donnée **omique**, permettant de caractériser le vivant à l'échelle moléculaire, a été utilisé : la génomique. Les ADN ont été séquencé, générant des lectures et les premiers algorithmes informatiques d'assemblage de ces lectures en contigs. Puis, de nouvelles données omiques portant sur l'analyse des ARN (transcriptomique), des protéines (protéomique) et de la biochimie (métabolomique) émergèrent en masse pour comprendre le fonctionnement des populations bactériennes, accélérant le développement en parallèle de **modèles informatiques**. Un modèle informatique représente de façon simplifiée un processus ou une réalité, il analyse les données d'entrée, dans ce cas présent les données omiques, afin de proposer des sorties, pouvant être numériques.

A l'échelle de bactéries qui sont cultivables, en mono- ou en co-culture, ces modèles informatiques servent principalement à analyser les données générées à haut débit. Dès lors où les données proviennent d'un microbiote, composé d'un ensemble de micro-organismes pouvant atteindre quelques milliards, cultiver ces ensembles devient difficile. Les modèles informatiques vont pouvoir analyser ces ensembles et surtout générer des hypothèses testables sur le fonctionnement de ces écosystèmes. Au sein de ces écosystèmes, il existe des communautés de bactéries dans lesquelles des mécanismes complexes d'interactions sont présents, rendant difficile leurs analyses. Parmi les phénomènes à expliquer on retrouve les **interaction bactériennes**, que sont des relations entre les organismes au sein d'une même communauté. Ces interactions ont des effets positifs, négatifs ou neutres sur les organismes qui composent la communauté (**?**). Ces interactions sont importantes et ont différents rôles selon l'écosystème : protection contre des pathogènes de l'intestin (**?**), recyclage des matières organiques dans l'écosystème du sol (**??**) ou encore, libération de composés d'arômes (**?**) responsable de la qualité du vin par exemple (**?**). Révéler ces interactions en laboratoire est possible en élaborant des co-cultures par paire d'espèces (**?**) mais peut devenir complexe lors de communautés de grande taille. Ainsi, il existe un réel besoin de créer des modèles informatiques permettant de mettre en avant ces interactions.

En somme, le développement de ces modèles informatiques est donc stimulé par les données omiques, permettant de les analyser, mais également permet de générer de nouvelles hypothèses testables en laboratoire, et donc potentiellement, créer de nouveaux modèles biologiques. Lier ces modèles informatiques au données biologiques n'est pas un processus trivial. Cependant, un domaine appelé la **biologie des systèmes** propose un moyen d'étudier ces écosystèmes, à l'aide des données biologiques en entrée.

0.2 La biologie des systèmes comme support de l'étude du métabolisme

La biologie des systèmes associe un organisme à un système et consiste à étudier le système dans son ensemble et non comme un assemblage de gènes et de protéines (**?**).

Comme énoncé précédemment et selon la définition de (**?**), la biologie des systèmes utilise des données omiques en entrée. Elle se décompose en 4 parties : la section *omiques*, qui constitue la connaissance biologique ; la section *computationelle*, qui construit un modèle informatique pour effectuer des simulations selon les hypothèses des biologistes et des données disponibles ; la section *analyse*, qui formule des hypothèses et des prédictions sur le systèmes ; et enfin la section *technologique*, qui consiste à établir un protocole expérimental pour vérifier les résultats du modèle. Tout comme la microbiologie a évolué vers l'étude d'écosystème, la biologie des systèmes s'est adaptée également et permet **l'écologie des systèmes**. Le système d'étude comprend désormais l'ensemble des individus de l'écosystème (**?**). La prochaine étape consiste à identifier le moyen pour caractériser les interactions bactériennes à l'aide de la biologie des systèmes. Nous savons que les interactions métaboliques entre bactéries ou avec son hôte consistent en l'échange de molécules impactant ou non le phénotype du récepteur (la bactérie ou l'hôte) (**?**). Or, la biologie des systèmes concentre son étude à plusieurs niveaux que sont les gènes, les protéines et les métabolites. Dans la littérature, un réseaux métabolique, qui est une approximation de l'ensemble des réactions que compose un individu (métabolisme), est décrit comme un ensemble de réactions biochimiques sous forme d'une association de gène-protéine-réaction (GPR) (**?**). Ainsi, créer un modèle du **métabolisme** à l'aide de réseaux métaboliques à l'échelle du génomes (GEM) semble être une solution pour caractériser les communautés bactériennes. A l'échelle d'un individu, il existe de nombreuses méthodes informatiques et mathématiques permettant l'analyse métabolique, comme par exemple l'analyse basée sur des flux (**?**) ou encore sur les graphes (**?**). La principale question partiellement résolue est l'étude du métabolisme à l'échelle de la communauté. Les méthodes usuellement utilisées sont peu adaptées à cause de coûts calculatoires importants dûs au nombre d'interaction bactérienne et à la faible disponibilité des données. De ce fait, l'étude haut débit des communautés bactériennes *in silico* nécessite le développement de nouvelles méthodes informatiques et mathématiques, discrètes et numériques, en mettant en avant les interactions bactériennes *via* l'étude du métabolisme.

0.3 Objectifs de la thèse

Du point de vue de la modélisation, le défi principal est la création d'un modèle avec le meilleur compromis entre précision des résultats, versatilité du modèle, temps de calcul en fonction des données en entrée et explicabilité du modèle. À cette problématique s'ajoute le nombre de mécanismes métaboliques que l'on souhaite modéliser qui augmente avec le nombre d'organismes que le système compte. Ainsi, analyser des communautés naturelles rapidement, avec une parfaite compréhension de l'ensemble du métabolisme de chaque individu au sein de la communauté permettant la prédiction de nouveaux processus d'interaction est utopique. Dans cette thèse, nous avons identifié deux grands formalismes permettant d'analyser des communautés microbiennes et de mettre en avant des interactions. Il existe des méthodes de modélisation numérique du métabolisme apportant la précision requise mais pouvant être limitées par la disponibilité des données et de la connaissance biologique en entrée ainsi que par le temps de calcul. Ces limites sont contournées par les approches discrètes du métabolisme, dégradant la précision numérique mais apportant la capacité d'analyser des communautés naturelles rapidement. Ainsi, l'objectif principal de la thèse est de trouver un compromis entre ces méthodes numériques et discrètes en développant une approche hybride de modélisation explicable du métabolisme des communautés microbiennes. Les chapitres **??**, **??**, et **??** présentent

les apports de la thèse pour répondre à cet objectif.

Le chapitre **??** est une revue de la littérature concernant les différents moyens de représenter le métabolisme ainsi que les méthodes d'analyses existantes.

Le chapitre **??** décrit comment créer un modèle numérique de communauté explicable à partir de données hétérogènes. Pour cela, nous avons à notre disposition 3 souches modèles permettant la production de fromage : *L. lactis*, *L. plantarum* et *P. freudenreichii*, ainsi que des données de croissances bactériennes, de concentrations de métabolites dans des conditions de cultures pures et de co-culture. Nous verrons dans un premier temps comment ces données hétérogènes couplées à notre stratégie itérative basée sur l'analyse de l'équilibre des flux (**?**) que nous avons développée permettent d'obtenir des résultats conformes aux données expérimentales. Et dans un second temps, comment un modèle numérique permet de générer des hypothèses testables sur l'expression des voies métaboliques ainsi que sur les interactions bactériennes mises en jeu durant la fabrication du fromage.

Le chapitre **??** décrit comment pseudo-quantifier des interactions bactériennes et comparer des communautés de grande taille entre elles. À l'aide de l'approche par raisonnement, nous avons établi des règles logiques nous servant à calculer le potentiel de coopération et de compétition de communautés naturelles. Afin de tester le passage à l'échelle, nous utiliserons des données appartenant aux écosystèmes de (i) la feuille et de la racine de la plante modèle *A. thaliana*, (ii) au sol et (iii) de l'intestin pour créer des communautés synthétiques de tailles variables. Puis dans un second temps une comparaison avec les outils existants sera faite.

Le chapitre **??** décrit comment accroître les résultats du modèle discret présenté dans le chapitre **??**. Nous présenterons deux prototypes permettant d'affiner les sorties du modèle discret en (1) sélectionnant des communautés bactériennes candidates et en (2), intégrant de la temporalité.

Enfin, le chapitre **??** montre les conclusions et les discussions de la thèse.

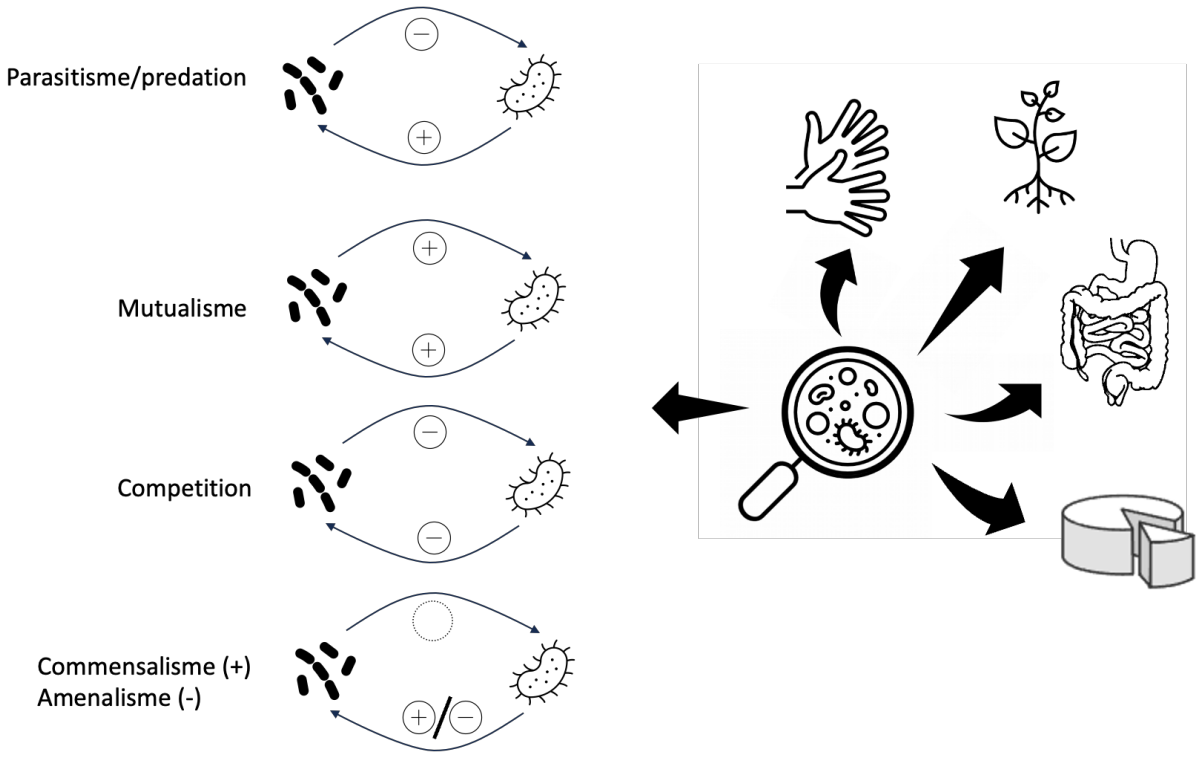

FIGURE 1 – **Présence de bactéries dans divers écosystèmes et illustrations des interactions bactériennes.** Un + (respectivement −) signifie une interaction positive (respectivement négative). Un cercle en pointillé représente une interaction sans effet.