Chapitre 1

Conclusion et perspectives [*ML* – NE PAS LIRE]

Discussion

Discussion partie informatique

Dans le monde de la biologie, l'identification des mécanismes sous-jacents expliquant une observation est un défi majeur. D'après (**?**), les modèles dit "boîtes-noires" permettent difficilement une interprétation du résultat du modèle, et ainsi ne correspondent pas à la demande des biologistes. A l'opposé, nous retrouvons des modèles "boites blanches" qui, en plus d'identifier des mécanismes responsables de l'observable et d'obtenir une bonne prédiction, permettent d'émettre des hypothèses testables en laboratoire. De ce fait, nous avons développé deux modèles informatiques répondant aux doubles enjeux des biologistes : bonnes capacités prédictives et pouvoir d'explicabilité. En reprenant la terminologie de (**?**), nos modèles assurent donc la *causalité*, *i.e.* l'explicabilité. Parmi les autres critères de (**?**), nos modèles répondent également à la *transparence*, c'est-à-dire, qu'ils peuvent être auditables par les biologistes. Ce critère traite la question de comment le modèle fonctionne, au niveau de l'algorithme, des paramètres et du modèle entier.

Le modèle numérique présenté dans le chapitre **??** modélise un écosystème microbien étudié expérimentalement pour l'élaboration d'un fromage industriel. Pour rappel, notre stratégie itérative consistait à calibrer chaque souche métabolique pour inférer des potentiels d'interaction en communauté. De part cette stratégie, nous avons montré que la prédiction de notre modèle numérique était de bonne qualité. Nos résultats de simulation, à l'échelle de chaque souche métabolique et en communauté, sont numériquement proches des données expérimentales. En raison de cette précision, de la généralité du modèle et de l'utilisation explicite de mécanismes biologiques, nous appelons ce modèle numérique un modèle numérique mécanistique. De plus, la *causalité* du modèle est assurée par l'identification de voies métaboliques activées, de flux de consommation et ou de production ainsi que par la contribution de chaque modèle métabolique. Concernant la souche de *L. plantarum*, nous avons observé pour la première fois en lait l'activation de la voie hétérolactique, représenté par un flux non-nul de la voie de la transketolase. De plus,nous avons observé une utilisation du citrate par *L. plantarum* et une préférence d'utilisation du lactose au regard du lactate par *P. freudenreichii*. Notre modèle numérique explique ces résultats par une valeur de xflux de consommation de ces composés. Nous avons également implémenté un partage équilibré des ressources, lorsqu'elles deviennent limitantes,

entre les organismes, affinant la prédiction des concentrations des métabolites. Ce mécanisme a permis de mettre en avant les interactions bactériennes révélées dans le chapitre **??**. Toutes ces explications fournies par le modèle permettent ainsi de générer des hypothèses testables expérimentalement, qualifiant ainsi notre modèle numérique de modèle explicable. Enfin, nous assurons la *transparence* en explicitant les noms des métabolites, des réactions et des voies métaboliques.

Le modèle discret du chapitre **??** modélise l'interaction métabolique entre organismes au sein d'un écosystème microbien et prédit le potentiel de compétition et de coopération entre ces organismes. Avec l'approche par raisonnement, nous avons formalisé des définitions biologiques sous forme de règles et contraintes logiques : à partir d'un ensemble nutritif nous calculons le potentiel métabolique pour chaque souche et en communauté ; *scope* ; les échanges métaboliques ; *exchanged* ; les substrats en compétition ; *polyopsonistic*. Cette formalisation explicite nous permet de calculer des potentiels d'interaction, compétition et coopération, et garantit une bonne confiance dans les prédictions de notre modèle. D'une part, nos prédiction permettent d'obtenir des résultats similaires ceux fournis les outils numériques existants qui font référence, SMETANA et MICOM, pour les mêmes données. D'autre part, notre logiciel réussit une batterie de tests induits par un benchmark très complet, que nous avons conçu pour valider les prédictions et scores spécifiques d'un écosystème, évaluer le passage à l'échelle et vérifier le criblage haut débit des communautés.

Tout comme le modèle numérique mécanistique, ce modèle discret assure également la *causalité* : le potentiel de coopération est expliqué par le nombre de métabolites échangés ainsi que par les espèces intervenant dans l'échange ; tandis que le potentiel de compétition est expliqué par le nombre de consommateurs associés à un substrat limitant. Ainsi, le proxy de la coopération, représenté par la notion de métabolites échangeables, est expliqué en termes biologiques ainsi : parmi deux organismes A et B, un métabolite est échangeable si, il est productible par l'organisme A et consommable par l'organisme B uniquement lors d'un échange. De même, la compétition est expliquée en termes biologiques : un substrat nutritionnel est en compétition si ce dernier est limitant et co-consommé. Toutes ces explications issues de notre modèle permettent de générer des hypothèses testables en laboratoire. La production d'un métabolite échangé entre organismes coopérants A et B peut être détecté en métabolomique par LCMS ou par phénotypage biochimique ciblé, et associé avec la présence ou l'absence de l'espèce A. La consommation d'un métabolite par deux organismes en compétition peut être testé de façon similaire.

Enfin, notre modèle assure également la *transparence* car chaque règle logique du modèle est lisible, vérifiable et critiquable par les biologistes. En reprenant la notion biologique de coopération plus haut, un métabolite est dit échangé s'il est dans le *scope* d'un producteur, qu'il soit reactant d'un consommateur et qu'il ne soit pas dans le *scope* du consommateur.

Les modèles explicables développés dans cette thèse ont été élaborés pour répondre avec précision à des enjeux biologiques précis. Dans les deux cas, identifier la bonne méthode correspondante à ses enjeux n'a pas été trivial mais la conclusion d'une analyse menée en concertation avec les collègues biologistes. Le critère que les modèles doivent être prédictifs et explicables ne permet pas en soi de choisir entre un modèle numérique et discret.

L'approche discrète développée répond à de multiples enjeux que sont : la résolution de problèmes combinatoires permettant de passer à l'échelle de communautés naturelles, d'émettre des hypothèses sur de potentielles interactions bactériennes, d'apporter une

explication structurelle des phénomènes observés, d'obtenir des *scope* de métabolites productibles et consommées par un plusieurs organismes et d'obtenir une précision des résultats dû à l'inférence des règles biologiques. Chacun des modèles mécanistiques que nous avons développés permet de répondre à ces enjeux plus ou moins efficacement. Le modèle numérique a démontré sa polyvalence pour répondre à ses enjeux et est particulièrement adapté pour donner une dimension numérique de l'explicabilité, de scope métaboliques et des hypothèses testables. Lee modèle par raisonnement est quant à lui pertinent pour sa capacité à déterminer des potentiels d'interaction pour des grandes communautés en résolvant des problèmes combinatoires et de passer ainsi à l'échelle. De plus, ce modèle procure une explication détaillée et rapide du fonctionne de la commauté : par exemple, l'ensemble des métabolites échangés ainsi que les ensembles de producteurs et de consommateurs intervenant pour l'échange de ce métabolite. En somme, chaque enjeu peut être résolu par un modèle, et donc, un unique modèle hybride n'est pas forcément nécessaire.

Notre étude permet cependant d'évaluer l'opportunité que représente la construction des modèles hybrides pour des écosystèmes microbiens complexes. Quelles auraient été les propriétés d'un modèle hybride, combinant les avantages de la précision numérique d'un jumeau numérique avec celles du passage à l'échelle de communautés naturelles d'un modèle discret. Dans un soucis de satisfaire les contraintes explicitées dans le paragraphe 1, ce nouveau modèle hybride doit toujours être explicable, en donnant de bonnes prédictions numériquement proches des valeurs expérimentales et qu'il puisse identifier des potentiels d'interaction à grande échelle dans le but de générer des hypothèses testables sur les voies métaboliques activées, les espèces impliquées dans un échange métabolique ou encore celles en réelle compétition.

Deux pistes peuvent être explorées : améliorer le passage à l'échelle du modèle numérique ou enrichir le modèle discret. Nous avons opté pour le second choix en proposant un premier enrichissement temporel. En utilisant la logique temporel (**?**), il est possible, à l'aide de règles logiques temporelles, d'identifier quand un métabolite est productible. Cet ajout temporel permettrait d'affiner la prédiction du potentiel de compétition : deux espèces sont en compétition si pour un même substrat limitant, au moins deux espèces le consomment *au même moment*. Nous avons par la suite permis la sélection de communauté pour la construction d'un design expérimental. Cet ajout permet d'améliorer l'explicabilité du modèle puisque chacun des modèles proposés par le modèle discret est expliqué par les critères de sélection en amont.

Le modèle numérique représente le but final à atteindre de part sa précision numérique. Pour améliorer son passage à l'échelle, les paramètres inférés pour chaque souche métabolique peuvent être déduits à partir d'un méta-modèle pré-entraîné **(ref)**. Dans un second temps, le flux de consommation et de production peut être calculés à partir de la concentration des métabolites d'intérêt à chaque point de temps. Enfin, nous pouvons discrétiser le temps et calculer un FBA dans un intervalle de temps plus important, réduisant ainsi le temps de calcul mais augmentant l'approximation entre deux pas de temps.

[*ML* – expliciter pourquoi c'est dur LCMS, car c'est la résultant production-consommation]
[*ML* – 'prise en compte de la temporelle' dernier paragraphe]

Discussion partie biologique

La fermentation bactérienne est au cœur de la fabrication fromagère et elle se caractérise par une acidification du milieu, une production de composés d'arômes et enfin par une croissance bactérienne. Les échanges de métabolites entre espèces favorisent le fonctionnement de l'écosystème.

L'approche discrète du chapitre **??** identifie des hypothèses biologiques en comparant

des ensembles, de métabolites produits et consommés, par entité biologique au sein d'un écosystème. Parmi ces ensembles, nous avons calculé les ensembles de métabolites : (i) échangeables entre différentes souches, (ii) déjà présents dans le milieu extra-cellulaire, (iii) nouvellement disponibles libérés au cours d'une fermentation bactérienne par exemple, (iv) pouvant être limitant. L'analyse comparée de ces ensembles a permis de prédire les espèces en compétition, les échanges et les souches impliquées dans les échanges. Ces échanges peuvent impacter la croissance des bactéries, la production de composés d'intérêts et le pH. En plus de l'identification d'interactions en comparant des ensembles, nous avons calculé des potentiels de coopération et de compétition, ce qui ajoute une dimension pseudo-quantitative à l'interaction et permet de comparer des communautés de divers écosystèmes entre eux.

La dimension quantitative d'une interaction bactérienne est apportée par le modèle numérique mécanistique du chapitre **??**. Cette dimension est caractérisée par la prédiction de la concentration des métabolites, et la quantité de biomasse produite, correspondant au taux de croissance de l'espèce. Les interactions sont identifiées en analysant les flux de métabolites (i) entrant dans la cellule bactérienne et (ii) sortant vers le milieu extracellulaire. Par exemple, à l'échelle de chaque cellule, tous les modèles métaboliques ont un flux de consommation du lactose non nul, suggérant une compétition pour ce dernier. Ou encore, le lactate est produit par les bactéries lactiques et consommé par *P. freudenreichii* symbolisé respectivement par un flux de lactate sortant et entrant. Ces mécanismes révèlent une coopération entre les bactéries lactiques, *L. plantarum* et *L. lactis*, et la bactérie propionique *P. freudenreichii*. Cette analyse est faite grâce aux méthodes d'analyse par contrainte des flux (FBA, FVA) (**??**)sous l'hypothèse de maximisation de la biomasse. Il est possible de faire varier cette fonction objective, par exemple maximisation de la synthèse d'ATP ou de lactate, la distribution des flux changerait et donc, de nouvelles hypothèses d'interactions verraient le jour.

Ces interactions ont des impacts cinétiques sur la concentration des composés d'arômes et sur la densité bactérienne. Au cours du temps, la concentration et la densité bactérienne varient, et une croissance nulle ou ralentie d'une souche peut-être expliquée par une compétition pour un substrat ou par un précurseur limitant, au sens biologique. La production d'un composé d'arôme peut également être ralentie ou annulée, suggérant également une compétition. Ces hypothèses ont été mises en avant par une analyse dynamique de l'équilibre des flux (DFBA) (**?**) et a permis de proposer des valeurs de contribution relative de chaque espèce à la consommation et la production de composés d'intérêts. Nous avons ainsi suggéré une indication des producteurs et des consommateurs pour chaque composé pouvant être testés en laboratoire. La validation biologique pourrait être réalisée par un knockout des gènes correspondants. En plus de ces impacts, nous avons pu déduire quelles sont les espèces participant le plus à l'acidification du milieu, au regard de notre modèle de pH.

Les deux approches, discrètes et numériques que nous avons développées dans cette thèse, ont permis de générer des hypothèses testables, affinant la compréhension des interactions qui favoriserait la croissance bactérienne et/ou la production de métabolites. A partir de ces hypothèses, nous discuterons des expérimentations biologiques permettant de valider ou d'invalider les hypothèses afin d'améliorer les modèles explicatifs. Nous traiterons uniquement le cas où des génomes annotés et les souches correspondantes sont à notre disposition.

Les approches numériques et discrètes que nous avons développées ont permis de

détecter des interactions bactériennes. Dans le cas d'un échange de métabolites entre deux souches, nous avons pour les deux approches la bactérie productrice et consommatrice ainsi que le sens de l'échange. Par exemple, le lactate, mis en avant par notre implémentation du dFBA, est produit par les bactéries lactiques et consommés par *P. freudenreichii*, la phénylalanine, révélée par SMETANA (**?**) et MiCOM (**?**), produite par *L. plantarum* et consommée par *P. freudenreichii*. Ces deux approches approches permettent également de nous indiquer des souches co-consommatrice de substrats qui peuvent conduire à une compétitionentre souches si c'est substrats devenaient limitants. Par exemple, le lactose, co-consommé par les consortium, ou encore la glycine, produite par *L. plantarum* et co-consommée par *L. lactis* et *P. freudenreichii*. Toutes ces interactions révélées *in silico* avec des outils *a priori* ou *sans a priori* peuvent être validées expérimentalement au moyen d'une comparaison entre des co-cultures et des cultures pures. Afin de vérifier l'impact de ces interactions, la comparaison se porterait sur la croissance des bactéries ainsi que sur le dosage de métabolites.

Dans le cas d'une hypothèse de co-consommation d'un substrat, une décroissance bactérienne et/ou une diminution de la concentration de ce métabolite devraient être observées en co-culture. Afin de vérifier les souches impliquées, un marquage isotopique à froid de ce métabolite suivi d'une spectroscopie avec la méthode Raman (**??**) permettrait d'identifier les consommateurs de ce métabolite. Les activations et les inhibitions de productions métaboliques peuvent être mises en avant avec les données métatranscriptomique ou métaprotéomiques. Pour révéler expérimentalement ces échanges d'intermédiaires métaboliques (1), il faut valider indépendamment la production de ce métabolite par l'espèce productrice en culture pure et la consommation par l'espèce consommatrice avec un marquage isotopique.

Retour d'expérience de la thèse

Au cours de la collaboration établie avec les biologistes dans cette thèse, nous avons pu mettre en évidence deux façons d'exploiter les modèles développés. Dans un premier temps, seul le *résultat* du modèle est important afin de vérifier expérimentalement les prédictions. Par exemple, au sein du chapitre **??**, nous avons mis en avant de nouvelles interactions métaboliques potentielles, à la fois avec notre modèle avec *a priori* et les outils *sans a priori*, pouvant être validées ou réfutées grâce aux moyens que nous avons développés plus haut. Dans un second temps, pour une réexploitation des résultats de la thèse, la propriété de *généricité* des outils et du savoir-faire est importante. Cette généricité permet l'utilisation pour une autre question de recherche ou avec des données d'entrées différentes. Par exemple, les outils développés des chapitres **??** et **??** ont permis l'utilisation de réseaux métaboliques appartenant à n'importe quels écosystèmes, et permettent de s'intéresser à la compétition, coopération, sélection de communautés selon différents critères biologiques etc. Également, le degré de *diffusion* des logiciels développés est un facteur à prendre en compte pour un travail interdisciplinaire. En effet, le déploiement de nos outils sur une plateforme, telle que Galaxy (**?**) par exemple, permettrait de garantir un accès simplifié pour les biologistes et d'approfondir leur recherche sur la compréhension des interactions bactériennes.

Bilan des contributions

Au sein des travaux de cette thèse nous avons développé plusieurs méthodes d'analyse de communautés microbiennes applicables à des communautés appartenant à un écosystème simplifié, ou bien, à des communautés simulées provenant d'écosystèmes naturels

complexes, en proposant dans un premier temps de caractériser la coopération et la compétition au sein d'une communauté et dans un second temps, de proposer un modèle de sélection de communauté pour permettre l'analyse numérique de communautés intéressantes.
Notre méthodologie numérique se base sur la connaissance *a priori* du système biologique d'étude ainsi que sur la grande disponibilité des données. Ainsi, notre approche itérative devrait être applicable pour toute communauté où ces contraintes sont respectées. En effet, le raffinement de chaque modèle est un processus universel consistant à ajouter, éteindre ou modifier des réactions. En se basant sur l'analyse de la variance des flux et de la connaissance avec *a priori*, les chemins métaboliques utilisés ont pu être déterminés et la cohérence de ces résultats vis à vis de la littérature a été vérifiées. De plus, l'ajustement des mécanismes biologiques, tels que la consommation du lactose ou du lactate, à travers l'optimisation des paramètres, a été déduit en observant les courbes de simulations dynamiques. Enfin, nous avons montré qu'il était possible, à partir d'un mécanisme équitable de partage des ressources et d'un "tunning" à l'échelle de chaque bactérie, de mettre en avant des interactions bactériennes en communauté et d'obtenir des prédictions numériques satisfaisantes. Ces résultats ont fait l'objet d'un papier soumis dans metabolic engineering.

Nous avons par la suite cherché à définir les limites de l'analyse des interactions microbiennes dans le cas le plus pauvre, c'est-à-dire, avec uniquement des données génomiques. En basant notre méthodologie développée sur le principe de raisonnement et construit comme une extension du logiciel MISCOTO (**?**), les propriétés d'explicabilité des modèles et le passage à l'échelle de communautés de grande taille ont été conservés. Grâce à cette approche d'abstraction booléenne, nous avons caractérisés des potentiels de coopération et de compétition permettant une comparaison de communautés entières et non deux à-deux. Même si la méthode n'a pas la précision des approches numériques, elle compense par sa capacité de criblage au débit, permettant d'effectuer un premier filtre de caractérisation de l'écosystème microbien. Lors de la comparaison avec des méthodes quantitatives, nous avons démontré que la tendance de nos scores, issu d'un raisonnement booléen, était corrélée avec des méthodes basées sur des méthodes par contrainte. De part le faible temps d'exécution de notre approche, ces résultats confortent l'utilisation d'une méthode moins précise mais peu coûteuse pour permettre d'identifier manuellement des ensembles considérés pertinents du point de vue de la compétition, de la coopération, des substrats limitants ou bien des métabolites échangeables. Enfin, nous démontrons l'indépendance de notre approche vis-à -vis de la méthode de reconstruction des réseaux métaboliques ainsi que de l'écosystème d'origine des génomes.

Nous avons par la suite cherché à enrichir ces modèles logiques du point de vue de l'identification de communautés pertinentes en ajoutant des contraintes fournies par l'utilisateur ou l'utilisatrice. En effet, dans un monde idéal, un modèle hybride où le formalisme logique et numérique communiquent, permettrait à la fois une sélection de communautés pertinentes et une analyse numérique de ces dernières. La méthode de (**?**) est une première étape dans la construction d'un tel modèle qui sélectionne uniquement des communautés minimales, en minimisant le nombre d'échanges métaboliques au sein de la communauté. En s'inspirant de cette méthode, nous avons contruit un prototype qui, à partir d'un ensemble de réseaux métaboliques, l'utilisateur ou l'utilisatrice peut sélectionner une communauté en filtrant sa taille, sa composition taxonomique, le nombre de métabolites cibles devant être produits. Par ailleurs, l'ordre et le choix d'optimisation est également libre à

l'utilisateur ou l'utilisatrice. Grâce à ce travail, la taille de l'ensemble de communautés trouvés, respectant les règles et les contraintes logiques définies en amont, est réduite et l'utilisation d'un modèle numérique précis et performant sur des communautés de petites tailles peut être ainsi utilisé. Ce travail, vers un modèle hybride, est l'objet d'un papier scientifique en cours de rédaction.

Perspectives

Avec ces trois chapitres, nous avons contribué à l'avancement des analyses des communautés microbiennes sur le plan numérique et discret. Cependant plusieurs points d'amélioration subsistent. A court terme, valoriser d'avantage l'intégration des données métatransciptomiques serait interessant. En effet, dans le chapitre **??**, nous avons utilisés ces données hétérogènes pour valider des ineractions mises en évidence par SMETANA en regardant l'expression des gènes associés à la production et à la consommation des métabolites présumés échangés. Il pourrait être intéressant de contraindre les réseaux métaboliques avec les données méta-transcriptomiques avec différentes méthodes (**???**) et de les comparer avec nos données simulées au niveau des voies métaboliques exprimées ainsi que par rapport aux données de métabolomiques. De plus, nous pourrions nous demander si seul les régulations des expressions géniques suffisent pour obtenir un modèle numérique précis, ce qui aurait pour conséquence de réduire le temps passé au raffinement et à la calibration de modèles métaboliques. Ces données pourrait également servir de contrainte *a priori* pour les modèles discrets du chaptitre **??** et **??**. En effet, inférer des règles logiques sur l'expression de voies métaboliques connues peut changer les potentiels d'interactions de communautés microbiennes.

À moyen terme, la construction d'un modèle hybride des écosystèmes microbiens porrait être fait. Même si nous avons conclu qu'un modèle hybride n'est pas nécessaire, ce type d'approche pourrait apporter une plus-value. En effet, les métabolites suivis en dynamique dans le modèle numérique du chapitre **??** proviennent d'une question biologique. Coupler ce modèle numérique à une approche de criblage, comme le modèle logique du chapitre **??**, permettrait de fournir des métabolites d'interêts en entrée. Une difficulté mis en évidence, est le grand nombre de composés échangés détecté par une telle approche. Dans le but de filtrer ces solutions, un premier apport du modèle numérique serait de vérifier la faisabilité de production des composés identifiés par le modèle logique. Cette faisabilité pourrait-être transmise aux modèles logiquescomme une donnée d'entrée sous forme de règles et ou contraintes. Nous aurions a terme, un modèle hybride couplé où le formalisme numérique et le formalisme discret communiquent permettant l'analyse de communautés microbiennes.

Enfin à plus long terme, la construction d'un modèle de jumeau numérique du processus de la fermentation bactérienne. Selon la définition de L'INRAE, un jumeau numérique est "défini comme une représentation numérique de l'object ou du système d'intérêt (physique ou biologique ; e.g. cellule, tissu, plante, animal), incluant des variables décrivant son état et son évolution de façon dynamique" [1]. Nous avons des données d'entrée dynamique sur le processus de fermentation (métatranscriptomique, métabolomique et cinétique), un algoritme de modélisation et de simulation. Une caractérisation manquante est la boucle de rétro-action du modèle sur l'état du processus modélisé : *i.e.* une prise de décision.

1. `https://digitbio.hub.inrae.fr/thematiques/vers-le-jumeau-numerique#Jumeaux`