Modélisation à base d'agents des capacités fonctionnelles des communautés microbiennes

Laboratoire : UMMISCO, UMI 209, SU/IRD.

Directeur de thèse : Jean-daniel Zucker (DR IRD, HDR SU)

Co-directeur de thèse : Alexis Drogoul (DR IRD, HDR SU)

Encadrant : Eugeni Belda (IR IRD)

École doctorale : EDITE ED 130 de SU

Mots-clés : Modélisation multi-agents, Métagénomique, Annotation fonctionnelle, Assimilation de données, Bioinformatique, Plateforme GAMA.

Resumé:

Le microbiote intestinal (l'ensemble des micro-organismes qui peuplent le tube digestif) joue un rôle essentiel dans notre santé. Il participe à la digestion de nutriments complexes, façonne notre réponse immunitaire et synthétise des molécules qui interagissent avec l'hôte humain. La diversité de cet écosystème est énorme, et son répertoire génétique est de plusieurs ordres de grandeur supérieur à celui du génome humain. Les technologies omiques, en particulier le séquençage massif, ont permis d'étudier son rôle dans des pathologies telles que l'obésité, le diabète de type 2 (T2D) ou les maladies coronariennes. Ces études ont montré que la composition du microbiote intestinal peut être plus prédictive de l'état de santé que la variation génétique du génome humain. Cela ouvre la porte à la médecine personnalisée ou prédictive en combinant le profil clinique de l'individu avec la composition de son microbiote intestinal. **Le problème majeur qui ressort de ces études est le manque d'explication biologique** ou mécanistique derrière les associations qui émergent de l'analyse statistique. C'est le principal problème des algorithmes de machine learning, connus comme des "boîtes noires" car bien qu'ils soient très performants ils sont extrêmement complexes, ce qui rend les modèles difficiles à interpréter biologiquement. L'interprétabilité est également nécessaire pour l'acceptabilité par la communauté des cliniciens afin que ces modèles puissent être utilisés dans le cadre de la médecine personnalisée. C'est **pourquoi il est nécessaire de développer des méthodes d'analyse qui explorent les mécanismes biologiques à l'origine des variations de composition du microbiome intestinal en lien avec les pathologies complexes humaines**. C'est dans ce contexte qu'une approche systémique basée sur la modélisation de différents réseaux moléculaires (métaboliques, régulateurs, interaction protéine-protéine) qui intègre des données omiques pourrait contribuer à ces besoins. La modélisation à base de contraintes (méthodes COBRA) utilisant les réseaux métaboliques à l'échelle du génome (GSMN) et

IRD, Institut de Recherche pour le Développement
32 avenue Henri Varagnat, 93143, Bondy cedex
www.ird.fr

techniques d'optimisation linéaire (FBA, FVA, pFBA) permet de simuler in-silico le comportement des écosystèmes microbiens en termes d'activités cellulaires individuelles, ainsi que ses interaction. Néanmoins, sont limitées dans le sens où elles fournissent un instantané de l'état métabolique défini par la topologie du réseau en termes de réactions biochimiques. Dans ce contexte, la modélisation à base les agents (ABM) est une approche de modélisation qui, appliquée aux communautés microbiennes, permet de représenter chaque microbe comme une entité qui peut évoluer dans un environnement simulé, fournissant une plateforme pour modéliser les interactions entre différents microbes et entre les microbes et l'environnement. GAMMA (GIS Agent-based Modelling Architecture) est une plateforme open-source pour la GPA développée depuis 2007 à UMMISCO-IRD qui fournit un environnement intégré pour le développement de modèles hétérogènes à grande échelle avec un fort support de la dimension spatiale et son propre langage de modélisation qui a été utilisé dans domaines comme la modélisation du trafic et des problèmes de transport, l'atténuation des risques, et la gestion des ressources dans les systèmes socio-environnementaux. L'objectif du présent projet de thèse est dans un premier temps d'étendre la plateforme GAMA à la modélisation 3D des communautés microbiennes de l'intestin humain. Il s'agira d'abord de concevoir des skills pour représenter le comportement d'un agent microbien. Dans un second temps il s'agira d'intégrer des données métagénomiques quantitatives avec GSMM des microbes intestinaux provenant de différents espaces génomiques, des méthodes COBRA comme cadre de simulation des activités métaboliques, et des données environnementales (alimentation, pH, température, densités cellulaires, etc.) et OMIC (protéomique, métabolomique, transcriptomique) comme variables pour paramétrer le comportement de la communauté. Les défis en termes de modélisation sont liés à la représentation d'une bactérie comme un agent informatique spatialisé. Cet environnement de modélisation sera utilisé pour modéliser la dynamique microbienne décrite par des ensembles de données métagénomiques longitudinales publiques afin d'évaluer les résultats des simulations et d'apprendre des paramètres supplémentaires sur le comportement du système. Enfin, le cadre de simulation sera intégré à des modèles prédictifs de maladies dérivés d'algorithmes d'apprentissage automatique pour aider à l'interprétabilité fonctionnelle des relations écologiques entre les entités microbiennes de ces modèles et proposer des stratégies d'intervention (consortiums microbiens nutritionnels, thérapeutiques) vers des résultats ciblés.

Summary (English)

The intestinal microbiota (all the microorganisms that inhabit the digestive tract) plays an essential role in our health. It participates in the digestion of complex nutrients, shapes our immune response and synthesizes molecules that interact with the human host. The diversity of this ecosystem is enormous, and its genetic repertoire is several orders of magnitude larger than the human genome. Omics technologies, in particular next-generation sequencing (NGS), have made it possible to study its role in pathologies such as obesity, type 2 diabetes (T2D) or coronary heart disease. These studies have shown that the composition of the gut microbiota may be more predictive of health status than genetic variation in the human genome. This opens the door to personalized or predictive medicine approaches by combining an individual's clinical profile with the composition of

IRD, Institut de Recherche pour le Développement
32 avenue Henri Varagnat, 93143, Bondy cedex
www.ird.fr

their gut microbiota. The major problem that emerges from these studies is the lack of biological or mechanistic explanation behind the associations that emerge from the statistical analysis. This is the main problem with machine learning algorithms, known as "black boxes'' because although they are very powerful they are extremely complex, making models difficult to interpret biologically. Interpretability is also necessary for acceptability by the clinical community so that these models can be used in personalized medicine. Therefore, there is a need to develop analytical methods that explore the biological mechanisms behind variations in gut microbiome composition in relation to complex human pathologies. Is in this context that a systemic approach based on the modeling of different molecular networks (metabolic, regulatory, protein-protein interaction) that integrates OMICS data could contribute to these needs. Constraint-based modeling (COBRA methods) using genome-wide metabolic networks (GSMN) and linear optimization techniques (FBA, FVA, pFBA) allows to simulate in-silico the behavior of microbial ecosystems in terms of individual cellular activities, as well as its interaction. Nevertheless, they are limited in the sense that they provide a snapshot of the metabolic state defined by the network topology in terms of biochemical reactions. In this context, agent-based modeling (ABM) is a modeling approach that, when applied to microbial communities, allows to represent each microbe as an entity that can evolve in a simulated environment, providing a platform to model interactions between different microbes and between microbes and the environment. GAMMA (GIS Agent-based Modelling Architecture) is an open-source platform for ABM developed since 2007 at UMMISCO-IRD that provides an integrated environment for the development of large-scale heterogeneous models with strong support for the spatial dimension and its own modelling language that has been used in areas such as modelling of traffic and transportation problems, risk mitigation, and resource management in socio-environmental systems. The objective of this thesis project is to extend the GAMA platform to the 3D modeling of human gut microbial communities. In a first step, it will be necessary to design skills to represent the behavior of a microbial agent. In a second step, we will integrate quantitative metagenomic data with GSMM of gut microbes from different genomic spaces, COBRA methods as a framework for simulating metabolic activities, and environmental data (food, pH, temperature, cell densities, etc.) and OMIC data (proteomics, metabolomics, transcriptomics) as variables to parameterize the behavior of the community. The modeling challenges are related to the representation of a bacterium as a spatialized computing agent. This modeling environment will be used to model the microbial dynamics described by public longitudinal metagenomic datasets to evaluate the results of the simulations and learn additional parameters about the behavior of the system. Finally, the simulation framework will be integrated with disease predictive models derived from machine learning algorithms to aid in the functional interpretability of ecological relationships between microbial entities in these models and propose intervention strategies (nutritional, therapeutic microbial consortia) towards targeted outcomes.

IRD, Institut de Recherche pour le Développement
32 avenue Henri Varagnat, 93143, Bondy cedex
www.ird.fr

Contexte et motivation

Le microbiote intestinal, défini comme l'ensemble des micro-organismes qui peuplent le tube digestif (bactéries, archées, virus, champignons et microorganismes eucaryotes tels que les protozoaires) joue un rôle essentiel dans notre santé à de multiples niveaux. Il participe à la digestion de nutriments complexes en molécules simples pouvant être absorbées par l'épithélium intestinal, il façonne notre réponse immunitaire et ainsi que synthétise ou module la production d'une pléthore de molécules potentiellement bioactives qui interagissent avec l'hôte humain. La diversité de ces communautés microbiennes est énorme, comme l'ont montré différentes études de séquençage à grande échelle telles que l'International Human Microbiome Project (iHMP) aux États-Unis, le projet FrenchGut en France ou le projet MetaHIT en Europe. Le répertoire génétique qu'elles représentent au niveau individuel est plusieurs ordres de grandeur supérieur à celui du génome humain [1,2].

Le développement de différentes technologies omiques, en particulier le séquençage massif, a rendu possible de multiples études associatives pour caractériser la variation de la composition taxonomique et fonctionnelle du microbiote intestinal dans de multiples pathologies telles que l'obésité, le diabète de type 2 (T2D) ou encore les maladies coronariennes et cardiovasculaires [3–6]. La combinaison d'une analyse statistique classique avec des algorithmes de machine learning a montré que la composition du microbiote intestinal peut être plus prédictive de la transition vers différentes pathologies que la variation génétique du génome humain [7]. Cela ouvre la porte à la médecine personnalisée ou médecine prédictive en combinant notamment le profil clinique de l'individu avec la composition de son microbiote intestinal [8]. De même, le développement de différentes techniques à permis de quantifier le profil des métabolites bioactifs dans différents fluides (sérum, urine, plasma, échantillons fécaux). La mise en relation de la composition du microbiote intestinal défini par séquençage massif, permettent et les données issues de la métabolomique a permi d'identifier des associations fonctionnelles potentielles entre les variables du microbiome (espèces, fonctions) et les métabolites bioactifs[9].

Le problème majeur qui ressort de ces études est le manque d'explication biologique ou mécanistique derrière la multitude d'associations qui émergent de l'analyse statistique, à savoir pourquoi le microbiote varie en fonction de différentes conditions environnementales ou quelle serait la conséquence fonctionnelle de ces variations dans la production ou la consommation de différents métabolites et macromolécules entre les composants du microbienne, ainsi qu'entre le microbiome et l'hôte humain. C'est le principal problème des différents algorithmes de prédiction basés sur le machine learning (réseaux neuronaux, forêt aléatoire, machines à vecteurs de support, etc.), qui sont connus d'être des "boîtes noires" car bien qu'ils soient très performants ils sont extrêmement complexes, ce qui les rend difficiles à interpréter biologiquement. L'interprétabilité est également nécessaire dans le contexte de l'acceptabilité par la communauté des cliniciens afin que ces modèles puissent être utilisés dans le cadre de la médecine personnalisée [8,10]. C'est **pourquoi il est nécessaire de développer des méthodes d'analyse qui explorent les mécanismes biologiques à l'origine des variations de composition du microbiome intestinal en lien avec les pathologies complexes humaines**. Ceci pourrait à terme, permettre de concevoir des

IRD, Institut de Recherche pour le Développement
32 avenue Henri Varagnat, 93143, Bondy cedex
www.ird.fr

stratégies visant à manipuler la composition du microbiome pour améliorer l'état de santé des individus.

Approches méthodologiques

C'est dans ce contexte qu'une approche systémique basée sur la modélisation de différents types de réseaux moléculaires (métaboliques, régulateurs, interaction protéine-protéine) qui intègre des données quantitatives issues de diverses technologies omiques pourrait contribuer à une compréhension mécanistique des capacités fonctionnelles du microbiome intestinal, ainsi que des interactions potentielles avec l'hôte. Dans ce contexte, la modélisation à base de contraintes (méthodes COBRA) utilisant les réseaux métaboliques à l'échelle du génome (GSMN) permet de simuler in-silico le comportement des écosystèmes microbiens en termes de consommation et de production de composés chimiques par les membres de la communauté, ainsi que ses interactions [11–13]. En simulant le comportement cellulaire dans différentes conditions à l'aide de techniques d'optimisation linéaire (FBA, FVA, pFBA), il est possible d'identifier les consortiums microbiens qui maximisent la production d'un métabolite donné d'intérêt, ou qui maximisent la dégradation de contaminants chimiques à des fins de biorémédiation [14,15]. Malgré leur polyvalence, les méthodes COBRA sont en quelque sorte limitées à la simulation de l'état métabolique du système (individu, communauté) dans le sens où elles fournissent un instantané de l'état métabolique défini par la topologie du réseau en termes de réactions biochimiques. Les approches informatiques basées sur des simulations itératives mettant à jour les conditions du système sur la base des flux métaboliques prédits permettent d'approcher des comportements dynamiques sur des communautés microbiennes de petite taille [16–18].

Dans ce contexte, la modélisation à base les agents (ABM) est une approche de modélisation qui, appliquée aux communautés microbiennes, permet de représenter chaque microbe comme une entité qui peut évoluer dans un environnement simulé, fournissant une plateforme pour modéliser les interactions entre différents microbes et entre les microbes et l'environnement [19]. Plusieurs plates-formes ABM spécifiquement conçues pour modéliser les biofilms microbiens sont disponibles, mais elles sont limitées à un petit nombre d'agents [20] et très peu intègrent les méthodes GSMM et COBRA dans l'étape de modélisation [21]. GAMA (GIS Agent-based Modelling Architecture) est une plateforme open-source pour la GPA développée depuis 2007 à UMMISCO-IRD qui fournit un environnement intégré pour le développement de modèles hétérogènes à grande échelle avec un fort support de la dimension spatiale et son propre langage de modélisation qui a été utilisé dans différents domaines de recherche comme la modélisation du trafic et des problèmes de transport, l'atténuation des risques, la gestion de crise et la gestion des ressources dans les systèmes socio-environnementaux [22].

Objectif

L'objectif du présent projet de thèse est dans un premier temps d'étendre la plateforme GAMA à la modélisation 3D des communautés microbiennes de l'intestin humain. Il s'agira notamment de concevoir des skills pour représenter le comportement d'un agent microbien [24]. Dans un second temps il s'agira d'intégrer des données métagénomiques quantitatives

IRD, Institut de Recherche pour le Développement
32 avenue Henri Varagnat, 93143, Bondy cedex
www.ird.fr

avec GSMM des microbes intestinaux provenant de différents espaces génomiques, des méthodes COBRA comme cadre de simulation des activités métaboliques des membres de la communauté, et des données environnementales (alimentation, pH, température, densités cellulaires, etc.) et OMIC (protéomique, métabolomique, transcriptomique) comme variables pour paramétrer le comportement de la communauté. Les défis en termes de modélisation sont liés à la représentation d'une bactérie comme un agent informatique spatialisé. Cet environnement de modélisation sera utilisée pour modéliser la dynamique microbienne décrites par des ensembles de données métagénomiques longitudinales publiques afin d'évaluer les résultats des simulations et d'apprendre des paramètres supplémentaires sur le comportement du système. Enfin, le cadre de simulation sera intégré à des modèles prédictifs de maladies dérivés d'algorithmes d'apprentissage automatique (Predomics [23]) pour aider à l'interprétabilité fonctionnelle des relations écologiques entre les entités microbiennes de ces modèles et proposer des stratégies d'intervention (consortiums microbiens nutritionnels, thérapeutiques) vers des résultats ciblés.

Références

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
2. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
3. Aron-Wisnewsky, J. *et al.* Major microbiota dysbiosis in severe obesity: fate after bariatric surgery. *Gut* **68**, 70–82 (2019).
4. Vieira-Silva, S. *et al.* Statin therapy is associated with lower prevalence of gut microbiota dysbiosis. *Nature* **581**, 310–315 (2020).
5. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
6. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
7. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
8. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-Generation Machine Learning for Biological Networks. *Cell* **173**, 1581–1592 (2018).
9. Koh, A. & Bäckhed, F. From Association to Causality: the Role of the Gut Microbiota and Its Functional Products on Host Metabolism. *Mol Cell* **78**, 584–596 (2020).
10. Yu, M. K. *et al.* Visible Machine Learning for Biomedicine. *Cell* **173**, 1562–1565 (2018).
11. Durot, M., Bourguignon, P.-Y. & Schachter, V. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol. Rev.* **33**, 164–190 (2009).
12. Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129–143 (2009).
13. Thiele, I. & Palsson, B. ?. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* **5**, 93–121 (2010).
14. Thiele, I. & Palsson, B. ?. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* **5**, 93–121 (2010).

IRD, Institut de Recherche pour le Développement
32 avenue Henri Varagnat, 93143, Bondy cedex
www.ird.fr

15. Bauer, E. & Thiele, I. From Network Analysis to Functional Metabolic Modeling of the Human Gut Microbiota. *mSystems* **3**, (2018).
16. Zomorrodi, A. R., Islam, M. M. & Maranas, C. D. d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth Biol* **3**, 247–257 (2014).
17. Harcombe, W. R. *et al.* Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep* **7**, 1104–1115 (2014).
18. Popp, D. & Centler, F. µBialSim: Constraint-Based Dynamic Simulation of Complex Microbiomes. *Frontiers in Bioengineering and Biotechnology* **8**, (2020).
19. Hellweger, F. L., Clegg, R. J., Clark, J. R., Plugge, C. M. & Kreft, J.-U. Advancing microbial sciences by individual-based modelling. *Nat Rev Microbiol* **14**, 461–471 (2016).
20. Koshy-Chenthittayil, S. *et al.* Agent Based Models of Polymicrobial Biofilms and the Microbiome-A Review. *Microorganisms* **9**, 417 (2021).
21. Bauer, E., Zimmermann, J., Baldini, F., Thiele, I. & Kaleta, C. BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLoS Comput Biol* **13**, e1005544 (2017).
22. Taillandier, P. *et al.* Building, composing and experimenting complex spatial models with the GAMA platform. *Geoinformatica* **23**, 299–322 (2019).
23. Prifti, E. *et al.* Interpretable and accurate prediction models for metagenomics data. *Gigascience* **9**, (2020).
24. Leveau, J. H. J., Hellweger, F. L., Kreft, J.-U., Prats, C. & Zhang, W. Editorial: The Individual Microbe: Single-Cell Analysis and Agent-Based Modelling. *Front Microbiol* **9**, 2825 (2018).

IRD, Institut de Recherche pour le Développement
32 avenue Henri Varagnat, 93143, Bondy cedex
www.ird.fr