

Projet de recherche doctorale : Modèles à variables latentes et inférence variationnelle pour analyser la structure du microbiote intestinal

La communauté microbienne complexe qui vit dans l'intestin humain, plus connue sous le nom de microbiote intestinal, présente de nombreuses fonctions importantes pour son hôte et est désormais reconnue comme un facteur crucial du maintien de la santé. Ainsi, lorsque ce microbiote est altéré, il peut jouer un rôle dans l'apparition, la chronicité et la sévérité de nombreuses maladies intestinales et extra-intestinales. Actuellement, la composition et les fonctions du microbiote intestinal sont déterminées par les technologies de séquençage de nouvelle génération. Cela représente une quantité massive de données car un seul microbiote humain héberge 25 fois plus de gènes que le génome humain. Les micro-organismes de l'intestin sont très diversifiés et également très sensibles aux changements environnementaux. Il a donc été suggéré que le microbiote pourrait être utilisé comme outil médical pour le diagnostic, le pronostic, et même la prédiction de la réponse au traitement. En effet, de nombreuses études médicales ont confirmé le fort potentiel du microbiote intestinal en tant que biomarqueur. Cependant, la structure spécifique du microbiote intestinal a été peu prise en compte jusqu'à présent et c'est l'un des maîtres de ce projet.

Objectifs de la thèse

Les défis visés par ce projet sont les suivants : (i) utiliser les données de séquençage du microbiote intestinal pour prédire des résultats médicaux d'intérêt, notamment la réponse au traitement, avec des erreurs raisonnables (tâche difficile en raison du petit nombre d'observations par rapport au nombre de variables) ; (ii) fournir des résultats interprétables qui peuvent être utilisés en pratique à des fins médicales.

Une première ligne de travail consiste à développer des approches dédiées à la spécificité des données microbiennes, qui sont des données d'abondance (données discrètes) et parcimonieuses (beaucoup de valeurs nulles car de nombreuses bactéries sont présentes seulement chez certains patients) avec une structure taxonomique. Plus précisément, chaque bactérie appartient à plusieurs groupes qui sont ordonnés hiérarchiquement du plus précis (espèce) au moins précis (domaine). Bien que cette information taxonomique soit connue, la manière dont elle est liée à l'impact d'une bactérie sur son hôte n'est pas claire. L'un de nos principaux objectifs est d'étudier le rôle du microbiote intestinal en tant que prédicteur d'un résultat médical d'intérêt, notamment la réponse à l'immunothérapie dans le cancer ou le développement de complications après une chirurgie gastro-intestinale [6].

De plus, bien qu'il existe une grande diversité parmi les compositions microbiennes individuelles, certains microbiomes peuvent contenir différentes bactéries mais présenter des similitudes en termes d'impact global sur leur hôte. Par conséquent, nous cherchons à proposer un modèle dans lequel nous pourrions reconstruire des états discrets non observés qui encoderaient une représentation des différentes fonctions internes du microbiome humain. L'hypothèse sous-jacente à cette ligne de recherche est qu'il existe des caractéristiques globales du microbiome qui permettent de stratifier les patients. Cette hypothèse est cohérente avec la compréhension actuelle de l'auto-organisation d'écosystèmes complexes et de leurs interactions avec l'hôte, voir [5]. La récupération de cette représentation non observée de chaque microbiome individuel permettrait d'établir une correspondance entre le microbiome de tout nouveau patient et l'un des groupes déjà identifiés afin de prédire, par exemple, sa réponse à un nouveau traitement.

Approche scientifique

Nous proposons d'analyser la dépendance statistique de la composition microbienne en utilisant l'hypothèse qu'elle provient d'états non observés de faible dimension. Dans ce contexte, les procédures d'inférence nécessitent la distribution a posteriori des états latents compte tenu des observations, ce qui est un problème difficile à résoudre. Une approche classique consiste à approximer ces distributions à l'aide de méthodes de chaînes de Markov ou de Monte Carlo séquentiel, voir par exemple [3] et les références qui y figurent. Cependant, ces méthodes basées sur la simulation sont gourmandes en ressources informatiques et montrent des limites même pour des dimensions relativement petites. Par conséquent, nous proposons d'étudier des alternatives plus attrayantes sur le plan computationnel.

Approches Poisson-lognormal pour des données d'abondance avec une structure taxonomique : développement d'un cadre statistique adapté à la spécificité des données de microbiote. Récemment, [2] ont proposé un cadre générique utilisant des variables latentes gaussiennes multivariées et ont développé un algorithme d'inférence variationnelle efficace pour qui est une alternative aux méthodes de Monte Carlo classiques. Les auteurs ont également proposé des solutions spécifiques pour les données de comptage. Le premier objectif de ce projet est d'étendre de tels résultats et de proposer de nouvelles approches adaptées aux données d'abondance où la composition taxonomique de chaque échantillon est observée à différents niveaux. Cette extension naturelle

de l'inférence variationnelle avec un espace latent gaussien multivarié fournirait une structure de dépendance simple pour décrire les données microbiennes.

De plus, très peu de résultats ont été proposés pour obtenir des garanties théoriques quant à l'estimation des structures latentes obtenues à l'aide de méthodes d'inférence variationnelle. Récemment, dans le contexte des séries temporelles, quelques résultats ont établi que sous de faibles hypothèses sur le vrai modèle, les espérances conditionnelles des états latents sachant les observations peuvent être correctement estimées en utilisant l'inférence variationnelle [1]. L'extension de ces résultats aux données d'abondance avec structure taxonomique est loin d'être évidente mais serait d'un intérêt significatif pour fournir une structure latente microbienne explicable.

Etat latent discret pour expliquer les fonctions taxonomiques : modélisation et inférence des états discrets qui encodent les caractéristiques principales du microbiote humain. En ce qui concerne le lien entre la réponse thérapeutique d'un individu et son microbiome, quelques bactéries candidates ont été identifiées parmi les variables microbiennes comme potentiellement pertinentes pour la prédiction grâce à des approches non paramétriques, voir [6]. Ces résultats motivent la conception de modèles alternatifs avec des procédures d'inférence statistique dédiées qui permettraient d'atteindre un tel niveau de prédiction avec une meilleure interprétabilité.

Nous proposons d'améliorer les performances de prédiction en utilisant le fait qu'il est connu que la composition en espèces d'un échantillon microbien fluctue fortement. En effet, ces fluctuations modifient le profil fonctionnel des données d'abondance des taxons. Par conséquent, une autre perspective de recherche visant l'identification de bactéries ayant des fonctions similaires consiste à utiliser des modèles à données latentes avec des espaces d'état discrets. Dans un tel cas, l'espace discret est constitué d'un nombre fini de vecteurs, également appelés "embeddings", décrivant formellement chaque fonction spécifique d'un échantillon microbien. L'inférence variationnelle pour ces modèles n'a trouvé que très récemment quelques solutions telles que des auto-encodeurs efficaces [7]. Cependant, pour traiter l'étape de quantification, les auteurs de [7] ont proposé une approximation grossière du gradient de la fonction de perte et ces modèles sont basés sur une distribution préalable autorégressive très complexe sur les états discrets. Nous proposons d'utiliser des modèles d'espace d'état discret pour fournir une description explicable de toutes les fonctions microbiennes encodées dans les embeddings. En étendant les résultats de [4], nous introduirons donc un modèle de données latentes permettant un apprentissage variationnel de bout en bout.

Adéquation du sujet avec l'Institut Sciences du Calcul et des Données

Le sujet présenté vise à développer une collaboration interdisciplinaire autour d'une thématique aux enjeux sociétaux majeurs, la compréhension du lien complexe entre microbiote intestinal et diverses pathologies. Notre équipe réunit deux statisticiens et un spécialiste du microbiote pour répondre aux questions statistiques, numériques et médicales soulevées par ce projet. En effet, notre objectif est de développer des méthodes d'analyse de données médicales complexes pour fournir des outils fiables et interprétables à la communauté médicale, mais il soulève également des défis théoriques et méthodologiques importants d'un point de vue statistique et numérique. En particulier, l'inférence des modèles à variables latentes s'accompagne généralement d'un important coût computationnel. Nous porterons une attention particulière au choix d'une modélisation et de procédures adaptées pour obtenir un compromis raisonnable entre la performance des méthodes et leur coût numérique. Nous veillerons également à proposer des implémentations efficaces des algorithmes développés tout au long de la thèse. Plusieurs jeux de données sont déjà disponibles pour ce projet afin d'étudier le potentiel du microbiote intestinal pour prédire (i) le développement de complications septiques après une chirurgie gastro-intestinale pour la maladie de Crohn, et (ii) la réponse à l'immunothérapie dans plusieurs cancers. En parallèle, de nombreux autres jeux de données sont actuellement générés pour tester davantage les approches méthodologiques construites dans ce projet.

Nous pensons que la thématique et les objectifs de notre projet s'inscrivent parfaitement dans l'axe de développement d'algorithmes pour l'analyse de données de l'ISCD, et plus précisément dans cet appel à projets qui requiert un co-encadrement entre deux unités de recherches distinctes.

Equipe d'encadrement. Sylvain Le Corff (directeur de thèse HDR) est statisticien, ses travaux portent principalement sur les modèles génératifs, les modèles à variables latentes et les méthodes de simulation. Anna Bonnet (co-encadrante) est statisticienne spécialisée dans les applications en biologie et en santé. Ils superviseront ensemble le développement mathématique et l'implémentation des algorithmes. Harry Sokol (co-encadrant) est hépato-gastro-entérologue, spécialiste du microbiote et des maladies de l'intestin. Il sera le référent pour identifier les questions médicales pertinentes, guider les choix de modélisation et interpréter les résultats. Les publications [1], [4], et [6] des encadrant.e.s sont en lien avec le sujet de thèse.

References

- [1] Mathis Chagneux, Élisabeth Gassiat, Pierre Gloaguen, and Sylvain Le Corff. Amortized backward variational inference in nonlinear state-space models. *arXiv preprint arXiv:2206.00319*, 2022.
- [2] Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.
- [3] Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer, 2020.
- [4] M. Cohen, G. Quispe, Sylvain Le Corff, C. Ollion, and É Moulines. Diffusion bridges vector quantized variational autoencoders. *Proceedings of the 39th International Conference on Machine Learning*, pages 4141–4156, 2022.
- [5] Anna Heintz-Buschart and Paul Wilmes. Human gut microbiome: function matters. *Trends in microbiology*, 26(7):563–574, 2018.
- [6] Clément Julien, Emre Anakok, Xavier Treton, Maria Nachury, Stéphane Nancey, Anthony Buisson, Mathurin Fumery, Jérôme Filippi, Léon Maggiori, Yves Panis, Philippe Zerbib, Yves François, Anne Dubois, Charles Sabbagh, Amine Rahili, Philippe Seksik, Matthieu Allez, Jérémie H. Lefevre, REMIND Study Group Investigators, Sylvain Le Corff, Anna Bonnet, Laura Beyer-Berjot, and Harry Sokol. Impact of the ileal microbiota on surgical site infections in crohn’s disease: A nationwide prospective cohort. *J Crohns Colitis*, 2022.
- [7] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in neural information processing systems*, 2017.