NUTRISS
CENTRE NUTRITION, SANTÉ ET SOCIÉTÉ

Atelier Parlons Bouffe du 30 mai 2025

L'intelligence artificielle (IA) pour analyser les données omiques en nutrition

Présenté par Elsa Rousseau, Département d'informatique et de génie logiciel, Faculté des sciences et de génie, Centre NUTRISS, INAF, Université Laval.

Quelques définitions et une mise en contexte ont été présentées en guise d'introduction à l'atelier. Le terme « omiques » provient du suffixe « ome », qui fait référence à la « totalité ». L'ajout du terme « omique » à un terme moléculaire réfère à l'évaluation complète d'un ensemble de molécules. La première discipline omique à avoir fait son apparition est la génomique. Cette discipline s'attarde à l'étude de génomes entiers, soit l'ensemble des gènes composant un organisme. D'autres exemples comme l'épigénomique, la métabolomique et la protéomique ont été mentionnés.

Les données issues de l'étude de génomes sont plus précisément des données de séquençage, qui sont constituées de lectures (*reads*), et qui représentent en quelque sorte les morceaux individuels permettant de créer le casse-tête en entier. L'analyse peut se faire sur les morceaux du casse-tête assemblés ou séparés. Différentes techniques analytiques, dont la spectrométrie de masse, permettent d'obtenir différents types de données omiques.

Dans les dernières années, le microbiote a suscité un fort intérêt dans la communauté scientifique pour son rôle en tant que modérateur des interactions entre la nutrition et la santé. En effet, les recherches montrent que des altérations au niveau des communautés microbiennes sont associées à diverses maladies. Dans la littérature, ce sont les bactéries les plus étudiées à ce jour. Cependant, il existe d'autres entités telles que les champignons, les protozoaires et les virus.

Qu'est-ce qu'un phage?

Un phage est un virus de bactérie.

Les phages constituent les entités les plus abondantes sur Terre, mais leur rôle demeure peu connu à ce jour. Des études montrent que des altérations dans la composition des populations de phages intestinales sont associées à différentes maladies. Un thème de recherche exploré par la professeure Rousseau est l'influence des phages sur l'écosystème intestinal.

L'IA pour la recherche de biomarqueurs (apprentissage supervisé)

Extraction de signatures à partir de données omiques

L'IA interprétable pour l'identification de biomarqueurs permet notamment :

- D'utiliser des modèles interprétables, c'est-à-dire qui permettent de comprendre les raisons d'une prédiction
- L'identification des variables les plus importantes utilisées par ces modèles

Des exemples de modèles interprétables : arbres de décision, forêts aléatoires, etc.

Plus concrètement, des exemples de projets de recherche

Étude de l'impact d'interventions nutritionnelles reposant sur la consommation de bleuets sur le microbiote et la santé cardiométabolique

La professeure Rousseau et son équipe ont mené un projet de recherche à partir de données provenant d'une intervention nutritionnelle qui visait à évaluer l'effet de la consommation de petits fruits sur le microbiote et la santé cardiométabolique (Projet des membres chercheurs Marie-Claude Vohl, Charles Couillard et André Marette). Les individus recrutés dans l'étude avaient un syndrome métabolique et devaient consommer quotidiennement des bleuets. Les bleuets sont reconnus pour leur taux élevé en polyphénols et leur consommation est associée à une meilleure santé vasculaire et métabolique. La durée de l'intervention était de 8 semaines.

L'objectif du projet de la professeure Rousseau visait à identifier les phages et les bactéries biomarqueurs de la nutiriton des individus par des approches d'IA. Pour ce faire, des échantillons de fèces ont été obtenus aux semaines 0 (pré-intervention) et 8 (post-intervention) et le séquençage de ces échantillons a été réalisé.

Un traitement bio-informatique a permis l'obtention de tableaux d'abondance taxonomique des bactéries et des phages dans le but de prédire si les échantillons étaient étiquetés pré ou post intervention selon l'abondance des taxons choisis.

Les résultats obtenus montrent des performances acceptables pour les bactéries, mais moins bonnes pour les phages. Les raisons sous-jacentes sont notamment le faible nombre d'échantillons disponibles, les données éparses et la plus grande diversité de phages comparativement à la diversité des bactéries. Principaux constats : Besoin de plus d'échantillons et viser des études d'intervention davantage contrôlées. On peut toutefois questionner le nombre adéquat d'échantillons requis pour obtenir de bons modèles d'IA.

Données sur le microbiote intestinal des Inuit du Nunavik

Un autre exemple de projet a été présenté lors de l'atelier. Cette fois-ci, l'équipe a étudié les phages dans le microbiote intestinal des Inuit consommant une diète plus traditionnelle versus une diète industrialisée.

Le projet a mené au développement d'une méthodologie permettant d'identifier des signatures de taxons plutôt que des taxons isolés pour le style de vie. Pour ce faire, le projet a identifié le taxon le plus important pour la prédiction, puis l'a retiré, et l'entraînement a été refait sans ce taxon. Le processus a été fait de nouveau afin d'extraire tous les taxons importants.

Projets cliniques en nutrition

Ce type de méthodologie a été utilisé dans d'autres projets de recherche visant à prédire les métabolites plasmatiques, notamment en tant que biomarqueurs de la consommation de lait et de fromage dans le cadre d'études randomisées et contrôlées et d'études partiellement contrôlées. Dans le cadre de ces études, les métabolites sélectionnés sont les métabolites les plus prédictifs dans le jeu de données (signature métabolomique).

Analyses multivariées versus univariées

Un modèle multivarié possède plusieurs variables explicatives, contrairement au modèle univarié, qui possède une seule variable explicative.

Une méta-analyse a été effectuée sur un grand ensemble de jeux de données de métabolomique. Dans environ 45% des jeux de données, le modèle multivarié semble meilleur, dans 21% des cas le modèle univarié est meilleur, puis dans environ 33% des cas les modèles multivariés et univariés sont équivalents.

Dans les cas où on a plus d'échantillons que de variables, il est préférable d'utiliser un modèle multivarié. Lorsque l'on a plus de variables que d'échantillons, ça dépend !

L'IA pour explorer les données omiques de façon non supervisée

La professeure Rousseau a terminé son atelier en donnant un exemple de projet qui visait l'étude d'échantillons du microbiote des océans prélevés à travers le globe. La mesure de métadonnées a été effectuée en lien avec l'environnement (profondeur, température, taux de salinité, etc.). L'équipe s'est intéressée au développement d'une métrique de distance entre les échantillons basée sur les distances de Wasserstein. Les résultats pour cette métrique ont été comparés à d'autres métriques existantes, telle que les distances en k-mers, reflétant le contenu partagé en k-mers, soit les séquences de k-nucléotides entre chaque paire d'échantillons. Les séquences nucléotidiques du génome communes entre 2 échantillons ont permis de déterminer la similitude génomique entre ces deux échantillons. Les distances en k-mers sont très liées aux distances physiques entre les échantillons.

La distribution des distances de Wasserstein est plus étalée que celle des k-mers, ce qui se traduit par plus de nuances dans les données.

Ces méthodes semblent prometteuses dans le domaine et pourraient entre autres servir dans l'étude du microbiote intestinal.