Spectrométrie de masse et intelligence artificielle pour cartographier le vivant

Résumé La spectrométrie de masse (SM) est une technique d'analyse très sensible permettant d'analyser des composés organiques avec une limite de détection pouvant atteindre la femtomole. Les récents progrès d'analyse des peptides par SM en tandem (SM/SM) permettent d'établir l'identité de milliers de protéines contenues dans des échantillons aussi complexes que les microbiotes. Toutefois, gérer d'énormes quantités de données est un vrai défi et une chance à saisir. L'intelligence artificielle (IA) joue un rôle clé pour l'interprétation des spectres et leur attribution à des séquences de peptides. L'IA révolutionne également l'analyse des données sur ces millions de composés chimiques qui aboutit à l'identification des protéines, des organismes qui les ont produites et bien d'autres informations. Cet article décrit comment une cartographie détaillée des microbiotes humains peut être obtenue à l'aide de la SM, cette technique permettant d'identifier un nombre important de protéines présentes dans les échantillons contenant ces microbiotes (fèces, salive…). Ces informations peuvent être exploitées en associant l'analyse métaprotéomique à l'IA pour identifier les organismes présents ainsi que leurs profils fonctionnels. Des biomarqueurs peuvent ainsi être identifiés en exploitant les données obtenues, permettant la sélection des cibles les plus pertinentes pour des traitements thérapeutiques.

Mots-clés **Spectrométrie de masse, intelligence artificielle, métaprotéomique, biomarqueurs.**

Abstract **Mass spectrometry and artificial intelligence to map living**

Mass spectrometry (MS) is a highly sensitive technique for analysing organic compounds with a detection limit of up to femtomole. Recent advances in peptide analysis using tandem mass spectrometry (MS/MS) have made it possible to establish the identity of tens of thousands of proteins contained in samples as complex as microbiota. However, managing huge quantities of data is a real challenge and an opportunity to be seized. Artificial intelligence (AI) is playing a key role in interpreting spectra and assigning them to peptide sequences. AI is also revolutionising the analysis of data on these millions of chemical compounds, leading to the identification of proteins, the organisms that produced them, and much other functional information. This article describes how a very detailed mapping of the different human microbiota can be obtained, making it possible to identify a very large number of proteins contained in samples taken from these microbiota (faeces, saliva, etc.), thus making it possible to identify the organisms present and their functional profiles using metaproteomics coupled with AI. Biomarkers can thus be identified by exploiting the data, enabling the selection of the most relevant targets for therapeutic treatments.

Keywords **Mass spectrometry, artificial intelligence, metaproteomics, biomarkers.**

L'étude des différents microbiotes humains a connu un développement fulgurant au cours des dernières années, mettant en avant leurs rôles prépondérants dans le fonctionnement du corps humain [1]. Un microbiote humain correspond à l'ensemble des bactéries, champignons, virus et archées présents dans une partie spécifique du corps humain telle que l'intestin ou la peau par exemple, il a été démontré que les micro-organismes présents y accomplissent des tâches utiles, voire essentielles à la survie de l'individu hôte. Par conséquent, l'étude du microbiote et de ses altérations qualitatives et/ou fonctionnelles (dysbiose) est devenue essentielle pour la recherche médicale, en particulier pour le microbiote intestinal, étroitement associé à l'apparition et au développement de plusieurs maladies [2]. Pour mieux comprendre et prédire ces maladies, de nombreuses analyses visant à détecter de potentiels biomarqueurs ont été conduites. En effet, les biomarqueurs sont des indicateurs biologiques mesurables dont la présence ou l'absence peut être utilisée pour diagnostiquer une maladie, prédire et/ou suivre son évolution, identifier différentes sous-populations de patients, ainsi que prévoir et surveiller leur réponse à des traitements. L'ensemble des travaux réalisés en cancérologie est un important corpus illustrant ce domaine [3]. Dès lors, l'alliance entre biomarqueurs et microbiotes humains apparaît comme un élément-clé pour la médecine du futur, pouvant favoriser et améliorer le diagnostic clinique et les prédictions pronostiques.

Bien que plusieurs biomarqueurs candidats ont été mis en évidence dans différentes études, peu ou aucun basé sur le microbiote n'a été mis en œuvre dans la pratique clinique. En effet, des facteurs de confusion entre les études masquent souvent les véritables caractéristiques des communautés microbiennes et peuvent donc conduire à des conclusions peu fiables en raison du caractère très épars des ensembles de données microbiennes. Par conséquent, des méthodes informatiques pour l'intégration des données d'analyse du microbiote provenant de différentes cohortes sont urgentes. Pour y parvenir, deux défis majeurs se posent : l'acquisition des données biologiques à un niveau de précision très fin, et le traitement de ces énormes quantités de données. Pour y répondre, la spectrométrie de masse à haute résolution se révèle de plus en plus une technique d'analyse pertinente. En effet, grâce aux développements récents, les spectromètres de masse de dernière génération de type LC-MS/MS permettent d'établir l'identité de plusieurs dizaines de milliers de protéines dans des échantillons aussi complexes que les microbiotes [4].

D'autre part, l'émergence de l'intelligence artificielle (IA) représente une véritable opportunité dans ce domaine. Permettant de gérer les énormes quantités de données

produites par spectrométrie de masse, mais aussi les problèmes liés aux données de microbiotes (caractère épars, composite et à haute dimension), l'IA joue déjà un rôle clé pour l'interprétation et l'analyse des données, permettant d'obtenir une cartographie des différents microbiotes humains [5]. Ainsi, nous décrivons ici comment une cartographie très détaillée des différents microbiotes humains peut être obtenue, comment des biomarqueurs peuvent être identifiés, et comment les cibles les plus pertinentes pour des traitements thérapeutiques peuvent être sélectionnées.

La spectrométrie de masse

Fonctionnement

La spectrométrie de masse (SM) est une technique d'analyse physico-chimique permettant la détection, l'identification et la quantification de molécules par mesure de leurs masses [6], avec une limite de détection pouvant atteindre la femtomole (10^{-15} mole). Le principe de l'analyse par SM consiste en la conversion de molécules en ions qui sont ainsi manipulés par des champs électriques et magnétiques et séparés en fonction de leur rapport masse/charge (m/z). Pour ce faire, trois composants du spectromètre de masse sont nécessaires : la source d'ionisation qui ionise les molécules à caractériser, l'analyseur de masse qui trie les ions en fonction de leur ratio masse/charge (m/z), et le détecteur qui mesure l'intensité des ions pour chaque valeur m/z. Ces ions sont produits par méthodes d'ionisation (dites « douces ») de la molécule, consistant à l'ajout d'un ion (H^+ par exemple) ou à la soustraction d'un électron à la molécule d'intérêt (ionisation électrospray par exemple). Il est ensuitepossible de fragmenter ces ions (dits précurseurs) en ions fils, dans le cadre d'une analyse par spectrométrie de masse en tandem (MS/MS, ou MS^2), permettant d'augmenter la capacité d'analyse des molécules. Des informations structurales sur la molécule introduite peuvent ainsi être obtenues, représentées sous la forme d'un spectre de masse, un graphique représentant l'intensité des ions en fonction de leur rapport m/z. La spectrométrie de masse est très souvent couplée aux méthodes séparatives telles que les chromatographies en phase gazeuse ou liquide, qui permettent une séparation optimale des composés d'un échantillon afin de le décomplexifier.

Les récents progrès dans ce domaine, tels que les innovations technologiques introduites avec le spectromètre de masse à temps de vol (« time of flight », TOF) ou l'analyseur Orbitrap, permettant de détecter précisément plusieurs dizaines de molécules contenues dans des échantillons très complexes, rendent possible l'application de la SM à des échantillons humains extrêmement complexes.

Application au vivant : la métaprotéomique

L'analyse des protéines d'un organisme, ou protéomique, a toujours été considérée comme pertinente pour cartographier les acteurs moléculaires du vivant [7]. En effet, les protéines constituent les briques élémentaires des organismes vivants et rendent compte du fonctionnement d'un organisme par les différentes fonctions accomplies. Jusqu'à récemment, la protéomique n'était néanmoins pas aussi largement utilisée que l'analyse des gènes d'un organisme (génomique), principalement en raison d'un coût plus élevé et de son accessibilité limitée pour les non-spécialistes. Ces limites ayant été résolues par les récents progrès technologiques, la protéomique a connu des progrès fulgurants en parallèle au développement de la SM. Désormais, ces progrès permettent l'émergence d'une nouvelle approche qui révolutionne l'analyse des microbiotes humains : la métaprotéomique.

La métaprotéomique étudie le contenu protéique d'échantillons complexes, tels que ceux de microbiotes humains. Les différentes étapes d'une analyse métaprotéomique sont détaillés en *figure 1*. L'étape initiale consiste à extraire efficacement les protéines de l'échantillon complexe, puis à les digérer en peptides. En raison de leurs propriétés chimiques similaires, en comparaison aux protéines, les peptides se prêtent mieux à la séparation par chromatographie liquide et à l'analyse ultérieure par spectrométrie de masse. Chaque analyse métaprotéomique produit des dizaines, voire des centaines de milliers de spectres de masse qui sont ensuite utilisés pour l'identification des peptides et des protéines. Les données obtenues sont donc ceux de peptides qui seront assemblés pour retrouver la cartographie protéique, taxonomique et fonctionnelle de l'échantillon initial. En fournissant une cartographie complète de l'état fonctionnel d'un microbiote, tout en permettant de rendre compte de sa composition taxonomique avec la même précision que la métagénomique, la métaprotéomique permet de pallier les faiblesses de cette dernière, notamment en termes de rapidité d'analyse et de l'observation réelle des fonctions assurées.

Bien que nous ne sommes encore qu'aux prémices de l'application de la métaprotéomique, les dernières études démontrent son rôle plus que prometteur quant à la découverte de biomarqueurs pour le diagnostic, le suivi et l'évaluation de traitements de nombreuses maladies. On peut notamment citer son potentiel dans le diagnostic des maladies inflammatoires chroniques de l'intestin (MICI). Grâce à la métaprotéomique, le paysage taxonomique et fonctionnel des microorganismes du microbiote intestinal de divers phénotypes de patients atteints de MICI a pu être caractérisé et comparé à celui de patients sains, permettant d'identifier les différences taxonomiques et fonctionnelles, ainsi que les différences de profils d'abondance protéiques, signatures de ces maladies [8]. Néanmoins, le potentiel de la métaprotéomique et son application en clinique restent encore restreints [9].

De fait, deux principaux défis se posent : d'une part, analyser le plus de protéines dans un temps le plus court possible, et d'autre part, interpréter les données en attribuant les signaux enregistrés aux bonnes protéines, et les protéines aux bons taxa. Le premier défi est en passe d'être résolu avec la dernière génération de spectromètres de masse qui permettent d'identifier un protéome global en moins de dix minutes. Quant au deuxième défi, l'application à la métaprotéomique de l'intelligence artificielle, qui bénéficie d'un ensemble de théories et techniques qui ne cessent de s'améliorer au fil du temps, est prometteuse.

L'intelligence artificielle

Principes généraux

L'intelligence artificielle (IA) rassemble l'ensemble des sciences, techniques, théories et technologies qui permettent de mimer, d'augmenter ou d'étendre l'intelligence humaine à l'aide de machines [10]. Ce terme est souvent confondu avec celui d'apprentissage automatique (« machine learning »), qui est un champ d'étude de l'IA ayant pour but de permettre à des machines d'apprendre à partir d'un jeu de données spécifiques en utilisant des modèles mathématiques. Cette forme d'apprentissage vise à résoudre diverses tâches, comme

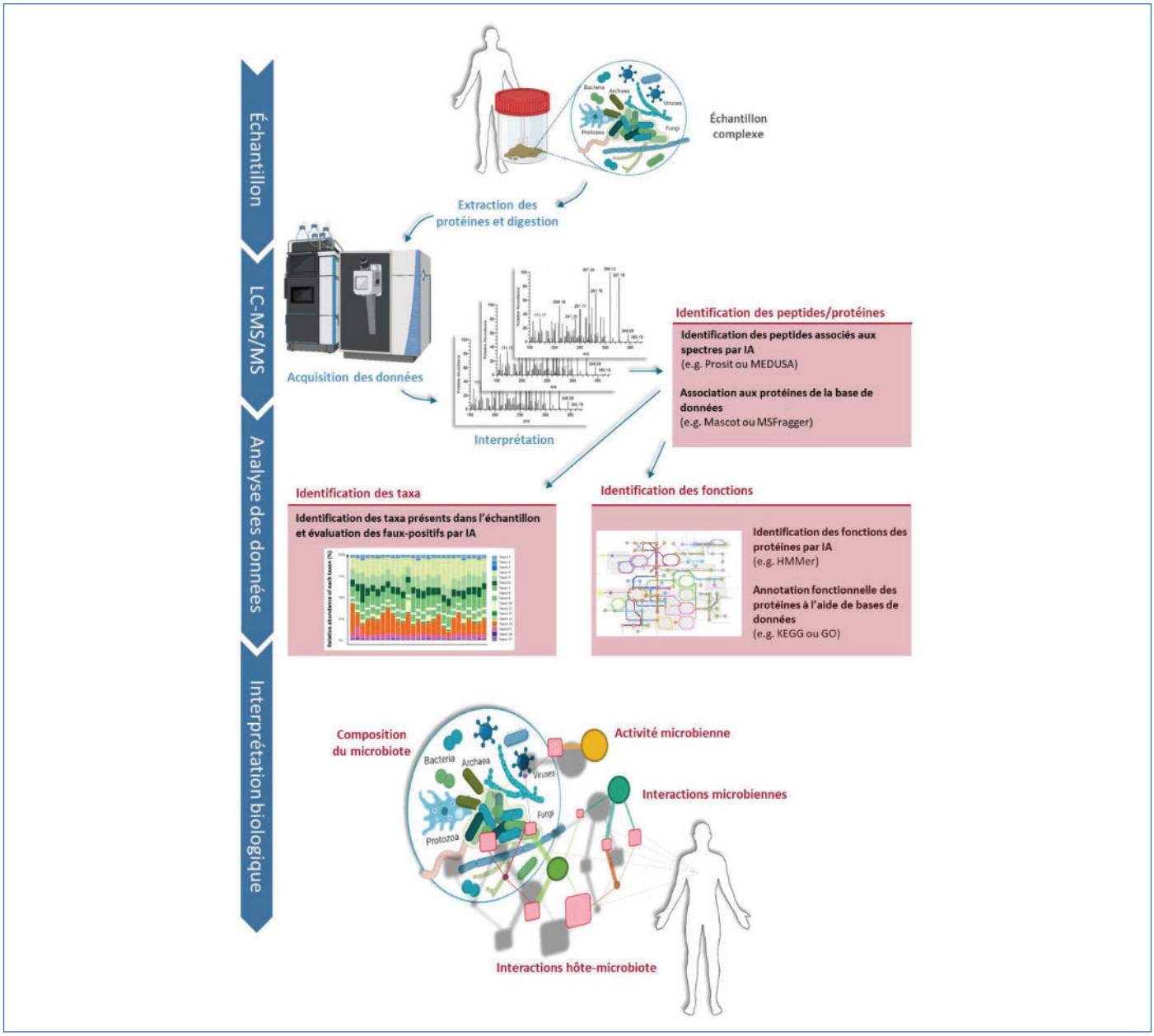

Figure 1 - Différentes étapes d'une analyse métaprotéomique et apports de l'intelligence artificielle.

Données d'entraînement			
Caractéristiques			Sortie
Spore	***Flagelle***	***Forme***	***Type Bactérie***
Non	Non	coques	*Streptococcus*
Oui	Oui	bacilles	*Bacillus*
Non	Oui	bacilles	*Pseudomonas*
Non	Non	coques	*Streptococcus*
Non	Non	spirochètes	*Treponema*
Oui	Oui	bacilles	*Bacillus*
Non	Non	coques	*Staphylococcus*
Non	Non	spirochètes	*Leptospira*
Non	Oui	bacilles	*Salmonella*
Oui	Oui	bacilles	*Clostridium*
Oui	Oui	bacilles	*Bacillus*
Données de test			
Oui	Oui	bacilles	?

Tableau I - Composition des données d'entraînement et de test de l'arbre de décision.

l'identification d'une bactérie en fonction d'un ensemble de caractéristiques définies. Il existe différents modèles mathématiques, allant de simples à complexes, qui démontrent des performances variables selon le jeu de données d'entraînement et de l'objectif recherché.

Parmi les modèles les plus simples et intuitifs pour comprendre le principe de l'apprentissage automatique se trouve l'arbre de décision. Les données d'entraînement et de test du modèle correspondent à la liste des différentes caractéristiques (« features ») connues sur les données ainsi que sur la tâche à réaliser ou sortie (« output »). Pour illustrer le propos, les données peuvent être représentées comme indiqué dans le *tableau I*.

Le modèle d'arbre de décision est entraîné sur les données d'entraînement, ce qui lui permet de créer un ensemble de choix de valeurs de caractéristiques (automatiquement calculés comme ceux maximisant la performance du modèle) afin de classifier les données (prédiction de l'« output »). Une manière plus simple d'appréhender cette tâche est de la représenter sous la forme d'un graphique (*figure 2*). Les chiffres à la fin des branches de l'arbre représentent les

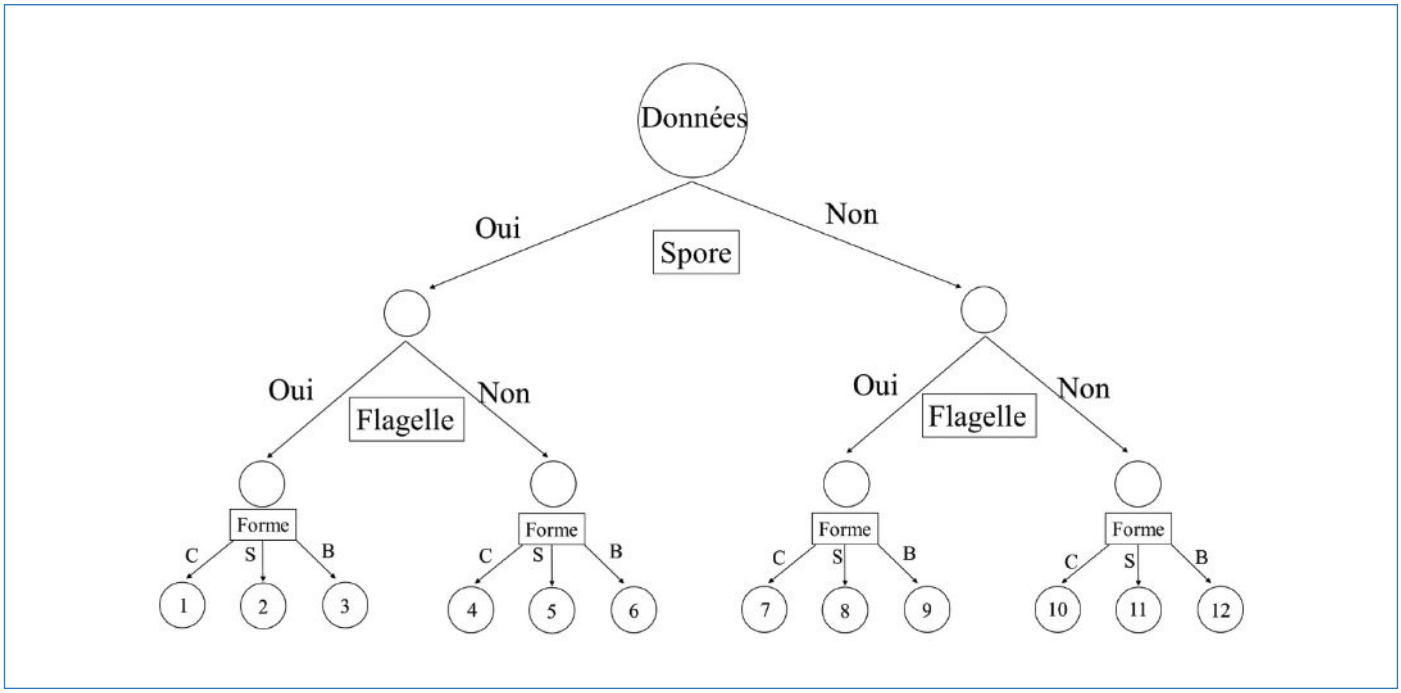

Figure 2 - Arbre de décision obtenu après entraînement sur les données d'entraînement. Chaque nœud représente les données qui seront divisées en d'autres nœuds en fonction des valeurs des « features ». C : coques ; S : spirochètes ; B : bacilles.

Chemin	Prédiction
1	Hors modèle
2	Hors modèle
3	75 % *Bacillus* 25 % *Clostridium*
4	Hors modèle
5	Hors modèle
6	Hors modèle
7	Hors modèle
8	Hors modèle
9	50 % *Pseudomonas* 50 % *Salmonella*
10	66,66 % *Streptococcus* 33,33 % *Staphylococcus*
11	50 % *Treponema* 50 % *Leptospira*
12	Hors modèle

Tableau II - Chemins de l'arbre de décision et prédictions associées.

différentes prédictions possibles, résumées dans le *tableau II*. Chaque prédiction est accompagnée de probabilités d'appartenance à chaque classe (type de bactérie dans notre exemple), calculées à partir des données d'entraînement. Les prédictions n'ayant pas été observées dans les données d'entraînement sont qualifiées d'hors modèle, et par conséquent ne sont pas assorties de probabilités.

Ainsi à l'aide de ce modèle, les données de test auront pour prédiction le chemin 5, correspondant à une probabilité de 25 % que ces caractéristiques appartiennent à la bactérie *Clostridium*, et une probabilité de 75 % qu'elles appartiennent à la bactérie *Bacillus*.

Ce cas d'application de l'apprentissage automatique répond à un problème de classification de données. Néanmoins, il existe bien d'autres problèmes pouvant être résolus par les différents modèles développés depuis plusieurs décennies. On peut notamment citer les modèles visant à réduire le bruit dans les données enregistrées, ou encore ceux ayant pour but de détecter des erreurs dans les données pouvant résulter d'une mauvaise qualité des données ou à un excès de bruit, par exemple. L'apprentissage automatique se révèle ainsi être un outil idéal pour gérer les immenses quantités de données générées dans différents contextes, notamment les énormes jeux de données biologiques produits par l'analyse en spectrométrie de masse.

Application à la métaprotéomique

Le premier défi en protéomique, et par conséquent en métaprotéomique, est l'interprétation des spectres obtenus en sortie du spectromètre de masse, de sorte à pouvoir obtenir la liste des peptides présents dans l'échantillon analysé [11].

Jusqu'à très récemment, les méthodes d'interprétation implémentées dans les moteurs de recherche des bases de données, tels que MSFragger ou Mascot, étaient exclusivement basées sur des approches statistiques [12]. La validation des interprétations ainsi obtenues est essentielle de par le fait que la comparaison des spectres expérimentaux et théoriques (générés par ces moteurs de recherche) peut engendrer des faux positifs, c'est-à-dire une mauvaise attribution d'un spectre SM/SM à un peptide. La stratégie utilisée pour évaluer ces faux positifs est appelée « target-decoy » (TDA, « target decoy approach »). Cette stratégie consiste en la quantification des interprétations obtenues dans une base de données protéique cible (target) d'une part, et dans une base de données protéique leurre (decoy) d'autre part. Les séquences leurres représentent les mauvaises attributions ou faux positifs et peuvent être générées de différentes manières, soit de manière aléatoire ou en inversant les séquences. Ainsi, le nombre de séquences leurres attribuées fournit une estimation de faux positifs, ce qui permet de calculer le taux de

fausses découvertes (FDR, « false discovery rate »). L'objectif est de limiter les faux positifs lors de l'attribution des spectres aux peptides ; dans cette optique, la valeur de FDR souvent retenue est de 1 %. Cependant, bien que cette approche soit souvent très performante, elle présente des limites, notamment lorsque le taux observé de faux positifs dans des échantillons à composition connue est anormalement élevé du fait de la forte densité de l'information enregistrée et de la taille atypique de la base de données.

Pour résoudre ce défi, des approches basées sur l'apprentissage automatique ont été développées récemment. Un exemple est le moteur de recherche MEDUSA, proposé en 2022, qui utilise des modèles de classification tels que le « random forest » et le « gradient boosting », pour enlever les pics isotopiques et simplifier le spectre, i.e. le « deisotoping » [13]. Le déisotopage consiste à supprimer les pics correspondant aux isotopes lourds naturels dans les spectres de masse, afin de réduire la complexité des données et améliorer l'annotation des spectres.

De nombreuses autres méthodes pourraient également être appliquées à ces données pour en réduire le bruit et ainsi améliorer l'annotation. On peut notamment citer les méthodes de réduction de bruit, ainsi que les méthodes dites de classification non supervisée, permettant d'annoter des données en fonction de leurs caractéristiques, sans connaître leurs véritables annotations, notamment la séquence peptidique.

Suite à l'identification des peptides dans l'échantillon, se pose le problème d'attribuer ces peptides aux protéines d'origine. En métaprotéomique, ce problème est particulièrement complexe car un peptide peut être commun à plusieurs protéines d'un même organisme, isoformes ou pas, et de plus, une protéine identifiée peut être partagée par plusieurs taxa, ce qui rend difficile l'attribution des peptides aux différents organismes.

Actuellement, différentes méthodes sont utilisées en métaprotéomique, regroupées en deux groupes principaux :

- les méthodes redondantes, ou non parcimonieuses, consistent à attribuer les peptides à toutes les protéines possibles, et donc ont tendance à sur-interpréter les résultats ;
- les méthodes parcimonieuses attribuent chaque peptide à une protéine unique selon des critères définis et sont donc plus justes en principe. L'approche de parcimonie la plus connue est celle du « rasoir d'Ockam », basée sur « l'hypothèse suffisante la plus simple est la plus vraisemblable ». Ce principe attribue chaque peptide à la protéine la plus abondante parmi toutes celles auxquelles il peut être théoriquement attribué ; l'abondance d'une protéine étant la somme des signaux des peptides identifiés comme lui appartenant.

Le développement des méthodes avancées de quantification est donc crucial pour une interprétation précise. Afin de quantifier la présence de chaque espèce dans l'échantillon et d'identifier leur activité fonctionnelle, il est donc essentiel de développer des méthodes robustes et efficaces de quantification et de parcimonie. Il est évident que l'apprentissage automatique permettra dans les années à venir d'améliorer ces procédés. Notamment, l'utilisation de modèles probabilistes, en particulier les méthodes dites bayésiennes [14], sont à envisager car elles permettent, entre autres choses, d'extraire des informations cruciales à partir de petits ensembles de données et de traiter les données manquantes. Considérant la probabilité comme une mesure de la confiance ou de la vraisemblance de l'occurrence d'un événement, il est alors possible de choisir un scénario probabiliste maximisant cette confiance ; le scénario étant en l'occurrence l'attribution des peptides aux différentes protéines.

Un dernier problème que l'on pourrait citer concerne les données manquantes, un défi bien connu en apprentissage automatique. En effet, certaines protéines ne sont pas présentes dans les bases de données (protéines inconnues ou mal annotées), rendant leur identification impossible. Pour résoudre cette problématique, des méthodes d'apprentissage automatique utilisant les réseaux de neurones, telles que DeepNovo [15], ont commencé à être utilisées pour le séquençage *de novo* des peptides, permettant de s'affranchir de la nécessité d'une base de données parfaite.

Un autre défi majeur à relever concerne l'annotation fonctionnelle. En effet, les bases de données sont loin de contenir une annotation fonctionnelle suffisamment précise pour chaque protéine, ce qui rend difficile l'interprétation des données. L'approche la plus connue pour annoter fonctionnellement une protéine dont la fonction n'a pas été confirmée expérimentalement est basée sur la recherche par similarité de séquence [16]. En général, les protéines homologues, i.e. ayant les mêmes fonctions, présentent un fort degré de similarité de séquence. L'une des méthodes d'apprentissage automatique les plus connues dans ce domaine utilise un modèle de Markov caché, qui est un modèle statistique représentant les probabilités de transition d'un état à l'autre. Dans ce cas, le modèle est entraîné sur un alignement de plusieurs séquences ayant la même fonction, afin d'identifier les motifs signatures de la fonction de ces séquences. En appliquant ce modèle à des séquences à fonctions inconnues, il est possible d'identifier, avec un certain degré de confiance, la fonction que pourrait avoir une protéine.

Néanmoins, il reste encore beaucoup à faire dans ce domaine, et les récents développements en apprentissage automatique permettront sans aucun doute une bien meilleure annotation fonctionnelle des protéines. On peut notamment citer la révolution AlphaFold [17], dont la première version date de 2018, permettant de prédire la structure en trois dimensions d'une protéine uniquement à l'aide de sa séquence. L'utilisation de cet outil pour l'annotation fonctionnelle semble prometteuse, en permettant une analyse par similarité de structure. En effet, sachant que les protéines adoptent une structure tridimensionnelle qui leur permet d'assurer leur fonction biologique, utiliser les prédictions d'AlphaFold pour prédire la fonction d'une protéine à travers sa structure, ou plus particulièrement la structure de son site actif, semble être la clé pour résoudre ce problème.

Il y a donc de nombreux défis à relever pour permettre à la métaprotéomique de jouer un rôle clé dans la médecine de demain, notamment à travers l'identification de biomarqueurs, ces biomarqueurs étant basés sur le microbiote et identifiés par analyse métaprotéomique. Néanmoins, l'ensemble de ces défis apparaît comme des opportunités à saisir. De plus, en observant la vitesse fulgurante à laquelle l'IA se développe, ces défis sont voués à être résolus dans un futur proche. Une fois ces défis relevés, l'analyse métaprotéomique, couplant l'analyse par SM et l'interprétation de ces données par IA, permettra d'obtenir une véritable cartographie du vivant, et par extension permettra une utilisation massive en tant qu'outil de diagnostic. Cela modifiera les approches cliniques classiquement utilisées et ouvrira la porte à une médecine de précision offrant des traitements plus personnalisés.

Vers une cartographie du vivant

L'utilisation combinée de la spectrométrie de masse et de l'intelligence artificielle (SM-IA) permet dès aujourd'hui d'obtenir une cartographie détaillée de différents microbiotes. Cette cartographie détaille les micro-organismes présents ainsi que leurs activités fonctionnelles, en étudiant les activités et fonctions des protéines qui composent ces micro-organismes. Elle permet également de comprendre l'activité fonctionnelle de l'hôte. Les informations obtenues sont cruciales pour comprendre l'activité du microbiote et la réponse de l'hôte en vue d'identifier de potentiels biomarqueurs. Exploiter les signatures de microbiotes ainsi obtenues représente un potentiel conséquent pour l'identification de biomarqueurs de diagnostic de maladies. Ces biomarqueurs pourraient ainsi être utilisés, combinés aux données cliniques et à d'autres biomarqueurs pour mieux comprendre et caractériser les pathologies. Il est cependant important de noter qu'un autre défi qui se posera dans ce cas sera celui d'évaluer l'impact de la variabilité interindividuelle observée dans la composition taxonomique et fonctionnelle des microbiotes sur les biomarqueurs mis en évidence par ces approches métaprotéomiques. Cela constituera également un défi pouvant être relevé à l'aide de l'intelligence artificielle, en vue de faciliter le transfert de ces biomarqueurs des laboratoires de recherche vers les équipes cliniques [18].

En conséquence, l'approche SM-IA pour l'identification de biomarqueurs aura des implications significatives pour le diagnostic et la médecine personnalisée.

[1] A. El-Sayed, L. Aleya, M. Kamel, Microbiota's role in health and diseases, *Environ. Sci. Pollut. Res. Int.*, **2021**, *28*, p. 36967-983.

[2] L. Grenga *et* al., Taxonomical and functional changes in Covid-19 faecal microbiome could be related to SARS-CoV-2 faecal load, *Environ. Microbiol.*, **2022**, *24*, p. 4299-316.

[3] E. Pons-Tostivint, A. Lugat, J.-F. Fontenau, M.G. Denis, J. Bennouna, STK11/LKB1 modulation of the immune response in lung cancer: from biology to therapeutic impact, *Cells*, **2021**,*10*, 3129.

[4] B.A. Rappold, Review of the use of liquid chromatography-tandem mass spectrometry in clinical laboratories: part I, Development, *Ann. Lab. Med.*, **2022**, *42*, p. 121-140.

[5] O. Pible *et al.*, Estimating relative biomasses of organisms in microbiota using "phylopeptidomics", *Microbiome*, **2020**, *8*, 30.

[6] B. Domon, R. Aebersold, Mass spectrometry and protein analysis, *Science*, **2006**, *312*, p. 212–217.

[7] J.L. Harry *et al.*, Proteomics: capacity versus utility, *Electrophoresis*, **2000**, *21*, p. 1071-81.

[8] C. Henry *et al.*, Modern metaproteomics: a unique tool to characterize the active microbiome in health and diseases, and pave the road towards new biomarkers - Example of Crohn's disease and ulcerative colitis flare-ups, *Cells*, **2022**, *11*, 1340.

[9] J. Armengaud, Metaproteomics to understand how microbiota function: the crystal ball predicts a promising future, *Environ. Microbiol.*, **2023**, *25*, p. 115-125.

[10] O. Pallanca, J. Read, Principes généraux et définitions en intelligence artificielle, *Arch. Mal. Coeur Vaiss – Prat.*, **2021**, *2021*, p. 3-10.

[11] S. Kim, N. Gupta, P.A. Pevzner, Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases, *J. Proteome Res.*, **2008**, *7*, p. 3354-63.

[12] A.T. Kong, F.V. Leprevost, D.M. Avtonomov, D. Mellacheruvu, A.I. Nesvizhskii, MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics, *Nat. Methods*, **2017**, *14*, p. 513-520.

[13] D.A. Boiko *et al.*, Fully automated unconstrained analysis of high-resolution mass spectrometry data with machine learning, *J. Am. Chem. Soc.*, **2022**, *144(32)*, p. 14590-606.

[14] B.K. Hackenberger, Bayes or not Bayes, is this the question?, *Croat. Med. J.*, **2019**, *60*, p. 50-52.

[15] N.H. Tran, X. Zhang, L. Xin, B. Shan, M. Li, De novo peptide sequencing by deep learning, *Proc. Natl. Acad. Sci. USA*, **2017**, *114*, p. 8247-52.

[16] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res.*, **2011**, *39*, p. W29-37.

[17] J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold, *Nature*, **2021**, *596*, p. 583-589.

[18] J. Mehrer, C.J. Spoerer, N. Kriegeskorte, T.C. Kietzmann, Individual differences among deep neural network models, *Nat. Commun.*, **2020**, *11*, 5725.

Hamid HACHEMI[*1,2], doctorant, **Lucia GRENGA**[1,2], chercheuse, et **Jean ARMENGAUD**[1,2], directeur de recherche au CEA.

[1]Université Paris-Saclay, CEA, INRAE, Département Médicaments et technologies pour la santé (DMTS), SPI, Bagnols-sur-Cèze.
[2]Laboratoire Innovations technologiques pour la détection et le diagnostic (Li2D), Université de Montpellier, Bagnols-sur-Cèze.

* Hamid.Hachemi@cea.fr