Protéotypage haut-débit de microorganismes par spectrométrie de masse en tandem

Madisson Chabas

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biologie Santé

École doctorale Sciences Chimiques et Biologiques pour la Santé (CBS2 ED n°168)

Unité de recherche : Laboratoire d'Innovations Technologique pour la Détection et le Diagnostic (Li2D)

Protéotypage haut-débit de microorganismes par spectrométrie de masse en tandem

Présentée par Madisson CHABAS

Le 15 décembre 2023

Sous la direction de Béatrice ALPHA-BAZIN et Jean ARMENGAUD

Devant le jury composé de

Mme Odile SCHILTZ, Directeur de recherche CNRS, Toulouse	Rapporteure
Mme Ana VARELA COELHO, Chargé de recherche, Lisbonne	Rapporteure
M. Christophe FLAHAUT, Maitre de Conférence, Université Artois, Lens	Examinateur
M. Franck VANDERMOERE, Directeur de recherche CNRS, Montpellier	Président du jury
M. Jean ARMENGAUD Directeur de recherche CEA, Bagnols-sur-Cèze	Co-directeur de thèse
Mme Béatrice ALPHA- BAZIN, Directeur de recherche CEA, Bagnols-sur-Cèze	Directrice de thèse

UNIVERSITÉ DE MONTPELLIER

Remerciements

Je tiens à remercier l'ensemble des membres du jury qui ont accepté de juger mon travail. Je remercie Dr Odile Schiltz et Dr Ana Varela Coelho pour avoir accepté d'être rapporteur ainsi que Dr Christophe Flahaut et Dr Franck Vandermoere d'avoir accepté d'être examinateur. Je vous remercie tous d'avoir pris le temps de lire et évaluer mon travail.

Je tiens particulièrement à remercier mes directeur et co-directeur de thèse Béatrice Alpha-Bazin et Jean Armengaud pour tout ce qu'ils m'ont apporté durant ces trois années de thèse. Merci à toi Béatrice d'avoir cru en moi en me prenant d'abord en stage puis en me proposant cette thèse. Je te remercie pour tout ce que tu m'as appris que ce soit d'un point de vue professionnelle mais aussi personnelle, d'avoir su me donner les clés pour m'épanouir pleinement et évoluer dans mes travaux et de m'avoir soutenue. Jean, je te remercie également de m'avoir donné l'opportunité de pouvoir réaliser cette thèse, je te remercie pour tes conseils, ta sympathie et ta disponibilité.

Je tiens également à remercier Laurent Bellanger pour sa gentillesse et ses conseils et de m'avoir donné la chance de profiter des ressources du laboratoire.

Je tiens à remercier Dr Gérald Culioli ainsi que Dr Christophe Hirtz d'avoir accepté de faire partie de mon comité de thèse et de leur bienveillance durant nos échanges.

Je remercie toutes les personnes du laboratoire avec qui j'ai pu travailler de près ou de loin et d'avoir rendu cette expérience professionnelle très enrichissante et je vous remercie pour votre bonne humeur. Cela a été un plaisir de travailler à vos côtés durant ces 3 années. Merci à Jean-Charles Gaillard, Guylaine Miotello, Mélodie Kielbasa, Alexia Breysse, Alicia Nouvel, Olivia Ardizzoni, Dana Coic, Gauthier Landerer, Clément Lozano, Thibaut Dumas, Justine Gricourt, Christine Almunia, Fabrice Gallais, Lucien Capuano, Yves Brignon, Anastasia Dewolf, Virgine Nouvel, Stéphanie Debroas, Anne Desplan, Joëlle Illiano, Olivier Pible, Fabienne Gas, Sylvie Ruat, Hamid Hachemi, Lucia Grenga, Pauline Hardouin, Virginie Jouffret, Karim Hayoun, Karen Cullota, Charlotte Foissard, Yannick Delcluze, Pascale Richard, Florence Laffont, Lauriane Plouinec.

Je tiens particulièrement à remercier l'équipe de protéomique. Merci à toi, Jean-Charles Gaillard avec qui nous avons fait avancer la science comme tu aimes bien dire et pour tous tes conseils et explications en spectrométrie de masse et bien sûr pour ta bonne humeur. Je tiens également à te remercier Guylaine Miotello pour ton aide et ta bienveillance. Je remercie également Mélodie Kielbasa pour ton aide en spectrométrie de masse et ta bonne humeur. Merci à Olivier Pible pour toutes tes explications et ton aide. Merci à Clément Lozano pour m'avoir

aidé et pour ta patience à répondre à toutes mes questions. Merci également à Pauline Hardouin pour tes conseils.

Je te remercie Karim Hayoun d'avoir été d'une grande aide durant mon stage et mon début de thèse, je te remercie pour ta gentillesse, ta patience et ton écoute.

Je tiens bien sûr à remercier les jeunes du Li2D qui sont devenus mes amis. Merci à toi Alexia, mon binôme depuis le début, cela été un plaisir d'être dans le même bureau depuis le début entre fou rire, entraide et soutiens je ne pouvais espérer mieux en binôme. Je tiens également à remercier Alicia et Mélodie le quatuor du début, je vous remercie pour votre bonne humeur, vos conseils et votre soutien. Merci à vous Gauthier, Clément, Thibaut, Olivia, Dana, Pauline cela a été un réel plaisir de travailler avec vous mais surtout de passer des moments supers avec vous avec tous et les fou rire qui vont avec !

Et bien sûr je tiens à remercier ma famille sans qui tout cela n'aurait pu arriver. Je tiens particulièrement à remercier mes parents qui m'ont poussé à réaliser cette thèse et qui ont été d'un soutien indéfectible et qui ont toujours cru en moi et encore merci pour votre aide ! Je remercie particulièrement mon frère qui a été d'un réel soutien pour moi, tu as toujours cru en moi et tu as toujours été de très bons conseils et me pousses toujours à me dépasser. Je tiens à remercier ma cousine Eva, qui a été d'un soutien indéfectible, à l'écoute et toujours de bons conseils et je te remercie également pour ta patience quand il a fallu écouter mes présentations. Je remercie grandement tout le reste de ma famille qui a été un vrai soutien, mes cousins Alexis, Kevin, Théo, Loan, Mathis, mes grands-mères, mes tantes Christèle, Hélène et Emilie, mes oncles Fabien, Claude et Didier et ma belle-sœur Romane. Et je tiens particulièrement à remercier mes amis de toujours qui m'ont soutenu, merci à Xavier, Alizé, Carla, Nicolas, Rémi, Laura, Mickael, Dimitri, Maxime, Adrien, Guylain, Marine et Rémi.

Résumé

Protéotypage haut-débit de microorganismes par spectrométrie de masse en tandem

Identifier rapidement des microorganismes est essentiel dans le domaine du diagnostic clinique, des contrôles sanitaires et alimentaires, et du criblage pour applications biotechnologiques. Améliorer les méthodes d'identification afin qu'elles soient plus rapides et sensibles est un enjeu de taille. Actuellement, le protéotypage par spectrométrie de masse MALDI-TOF est la méthode de référence pour les isolats bactériens dans les laboratoires de microbiologie clinique. Toutefois, cette méthodologie n'est pas en mesure de traiter la plupart des isolats environnementaux et des agents pathogènes opportunistes en raison d'une base de données de spectres expérimentaux incomplète. En enregistrant beaucoup plus d'informations sur les séquences au niveau des peptides, le protéotypage par spectrométrie de masse en tandem est capable d'identifier la position taxonomique de n'importe quel micro-organisme dans l'arbre de la vie, et peut s'avérer hautement discriminant au niveau des sous-espèces. Cette thèse a pour objectif d'adapter et rendre haut-débit une approche d'identification de microorganismes sans a priori par protéotypage de microorganismes, développée au laboratoire. La phylopeptidomique repose sur l'enregistrement de données de séquences peptidiques par spectrométrie de masse en tandem et l'association de ces données taxonomiques, permettant d'identifier l'organisme. Afin de réduire les coûts et le temps d'analyse, deux méthodes ont été inventées et testées durant ma thèse pour identifier tout type de microorganismes en quelques minutes d'analyse. La première méthode permet de multiplexer sans marquage plusieurs échantillons en créant un mélange contenant des fractions de chaque isolat qui différent en hydrophobicité. La robustesse, la reproductibilité et les limites de cette méthode ont été évaluées. La seconde méthode utilise les capacités de rapidité d'une analyse par infusion directe, permettant de réduire le temps d'analyse à 36 secondes de spectrométrie.

Mots clés : microorganismes, identification, protéines, protéotypage, spectrométrie de masse en tandem, taxonomie

Abstract

High-throughput proteotyping of microorganisms by tandem mass spectrometry

Rapid identification of microorganisms is essential in clinical diagnostics, health and food quality controls, and screening for biotechnology applications. Improving identification methods to make them faster and more sensitive is a major challenge. Currently, proteotyping by MALDI-TOF mass spectrometry is the reference method for isolates. However, this methodology is unable to handle most environmental isolates and opportunistic pathogens due to an incomplete database of experimental spectra. By recording much more sequence information at the peptide level, proteotyping by tandem mass spectrometry is able to identify the taxonomic position of any microorganism in the tree of life, and can be highly discriminating at the subspecies level. The aim of this thesis is to adapt and render high-throughput an approach without any a priori for microorganism identification by proteotyping developed in the laboratory. Phylopeptidomics is based on the recording of peptide sequence data by tandem mass spectrometry and their association with taxonomic information, enabling the organism to be identified. In order to reduce costs and analysis time, two methods were invented and tested during my thesis's work to identify any type of microorganism in just a few minutes of analysis. The first method enables sample multiplexing without labeling, by creating a mixture containing fractions for each isolate differing in hydrophobicity. The robustness, reproducibility and limitations of this method were evaluated. The second method takes advantage of the speed of direct infusion analysis, reducing analysis time to 36 seconds of spectrometry measurement.

Keywords: microorganisms, identification, high-throughput, proteotyping, tandem mass spectrometry

Table des matières

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

- ACN : Acétonitrile
- ADN : Acide Désoxyribonucléique
- AF : Acide formique
- AFLP : Polymorphisme de longueur des fragments amplifiés
- ANI : Identité nucléotidique moyenne
- API : Indice de Profil Analytique
- ARN : Acide ribonucléique
- ATCC : American Type Culture Collection
- BCYE : Milieu de culture aux extraits de levures (Buffered Charcal Yest Extract)
- BI : Inférence bayésienne
- CEA : Commissariat à l'Energie Atomique et aux Energies Alternatives
- CID : Dissociation induite par collision
- DART : Analyse directe en temps réel
- DDA : Acquisition dépendante des données (Data Dependant Acquisition)
- DDH : méthode taux d'hybridation ADN – ADN
- DIA : Acquisition indépendante des données (Data Independant Acquisition)
- DSMZ : Deutsche Sammlung von Mikroorganismen und Zellkulturen
- dNTP : désoxyribonucléoside triphosphate
- ECD : Dissociation par capture d'électrons
- ESI : Source d'ionisation par electrospray
- ETD : Dissociation par transfert d'électrons
- FAIMS : High Field Asymmetric Waveform Ion Mobility Spectrometry
- FAME : acide gras méthylester
- FASP : Filter Aided Sample Preparation
- FTIR : Fourier Transform InfraRed spectroscopy
- GC/MS : chromatographie gazeuse couplée à la spectrométrie de masse
- HCD : Higher energy Collision
- ICNP : International Code of Nomenclature of Prokaryote
- ICSP : International Committee on Systematics of Procaryotes
- IMS : Ion Mobility Spectrometry
- iST : improved Sample Technology
- iTOL : interactive Tree Of Life
- iTRAQ : isobaric Tags for Relative and Absolute Quantification
- LAMP : Loop mediated isothermal amplification
- LB : Lysogenic Broth
- LC : Chromatographie liquide
- Li2D : Laboratoire d'innovations technologique pour la Détection et le Diagnostic
- *m/z* : Rapport masse sur charge
- MAFFT : Multiple Alignement using Fast Fourrier Transform
- MAGs : Metagenome Assembled genomes
- MALDI : Matrix Assisted Laser Desorption/ Ionisation
- MEGA : Molecular Evolutionary Genetic Analysis

- ML : Maximum Likewood
- MLST : Multilocus Sequencing Typing
- MP : Maximum Parcimony
- MRM : Multiple Reaction Monitoring
- MS : Spectrométrie de masse
- MS/MS : Spectrométrie de masse en tandem
- NCBI nr : Base de données National Center for Biology Information non redundant
- NCBI : National Center for Biology Information
- NGS : New Generation Sequencing
- nLC : nano liquid chromatography
- PCR : Polymerase chain reaction
- PRM : Parallel Reaction Monitoring
- PSM : Peptide Spectrum Matches
- Q : Quadrupole
- QIT : trappe ionique
- qPCR : PCR quantitative (ou PCR en temps réel)
- RAPD : Random amplified polymorphic DNA
- RAPD : Random Amplification of Polymorphic DNA
- RDP : Ribosomal Database Project
- RFLP : Restriction Fragment Length Polymorphism
- SCX : Strong Cation Exchange
- SILAC : Stable isotope labelling by amino acids in cell culture
- SISPROT : Simple and intergrated spintip based protein digestion and three-dimensional peptide fractionation technology
- SP3 : Single-pot, solid-phase-enhanced sample preparation
- SPE : Solide Phase Extraction
- SPEED : Sample Preparation by Easy Extraction and Digestion
- SPi : Species Proteotyping index
- SRM : Single Reaction Monitoring
- TA : Température ambiante
- Taq : *Thermophilus aquaticus*
- TFA : Trifluoroacetic acid
- TIC : Total ion chromatography
- TLA : Total Automatisation Laboratory
- TMT : Tandem Mass Tag
- TOF : Time of Flight
- TSM : Taxon Spectrum Match
- VOC : Volatile organic compounds
- WGS : Whole genome sequencing

I. Introduction générale

I.1. Les microorganismes

1.1. Un peu d'histoire sur la découverte des microorganismes …

Les êtres humains ont plus de microorganismes dans leurs corps que de cellules. Ils nous entourent et sont présents en tout lieu. Mais que signifie le mot microorganismes, quand ont-ils été découverts et quels organismes sont regroupés dans ce terme.

Les microorganismes sont qualifiés d'organisme vivant invisible à l'œil nu individuellement et signifie étymologiquement « petits organismes » (Fuerst, 2014). Dès l'Antiquité, les microorganismes étaient soupçonnés d'être responsables d'épidémie décimant les habitants des différents empires (Opal, 2010). Les scientifiques et penseurs de cette époque n'ont pu découvrir l'origine de ces épidémies en raison du caractère invisible des microorganismes. Ce n'est qu'en 1683, avec le développement du microscope apporté par Antoni van Leeuwenhoek, que les microorganismes ont pu être observés. Il a été le premier à documenter correctement des observations « d'animalcules » grâce à un microscope qu'il avait lui-même confectionné (Gest, 2004). Durant le XIXème siècle, l'un des grands pionniers de la microbiologie, Louis Pasteur, chimiste d'origine, s'est pris de passion pour les microorganismes étant persuadé que de « petits êtres vivants » étaient à l'origine de nombreux mécanismes. Il démontre dans un premier temps l'implication des microorganismes dans le processus de fermentions lactique et alcoolique dans la fabrication des vins et bières, avant d'identifier les microorganismes à l'origine de certaines maladies. Le développement des vaccins pour lutter contre les infections causées par certains pathogènes le fut connaître de tous. En parallèle des recherches de Pasteur, un médecin de campagne allemand, Robert Koch s'intéresse également aux microorganismes. A travers l'étude de la maladie du charbon qui attaque les troupeaux de moutons du pays, il observe des bacilles au microscope et développe une méthode de culture permettant la pousse de ce bacille, *Bacillus anthracis*, (Ibrahim et al., 2021) et démontre que la bactérie associée est l'agent responsable de la maladie. Cette découverte propulse Robert Koch au rang de scientifique reconnu. Il découvrira de nombreux agents pathogènes à l'origine de maladies dont *Mycobacterium tuberculosis* nommé bacille de Koch qui lui vaudra le prix Nobel de médecine en 1905 (Blevins and Bronze, 2010). En 1878, Charles-Emmanuel Sédillot, médecin militaire français, proposa de les nommer « microbes » (Sédillot, 1878). Longtemps les virus n'étaient pas visibles au microscope et n'étaient donc pas caractérisés. La mise en place de l'ultra filtration utilisant un filtre Chamberland a permis de donner un nom à ces entités : les virus (Ibrahim et al., 2021) et ont été caractérisés au XIXème siècle. Le génie et la perspicacité de

grands scientifiques aura permis de découvrir et mettre un nom sur ces « petits organismes ». Les grandes découvertes au fil des époques sont résumées dans la Figure 1.

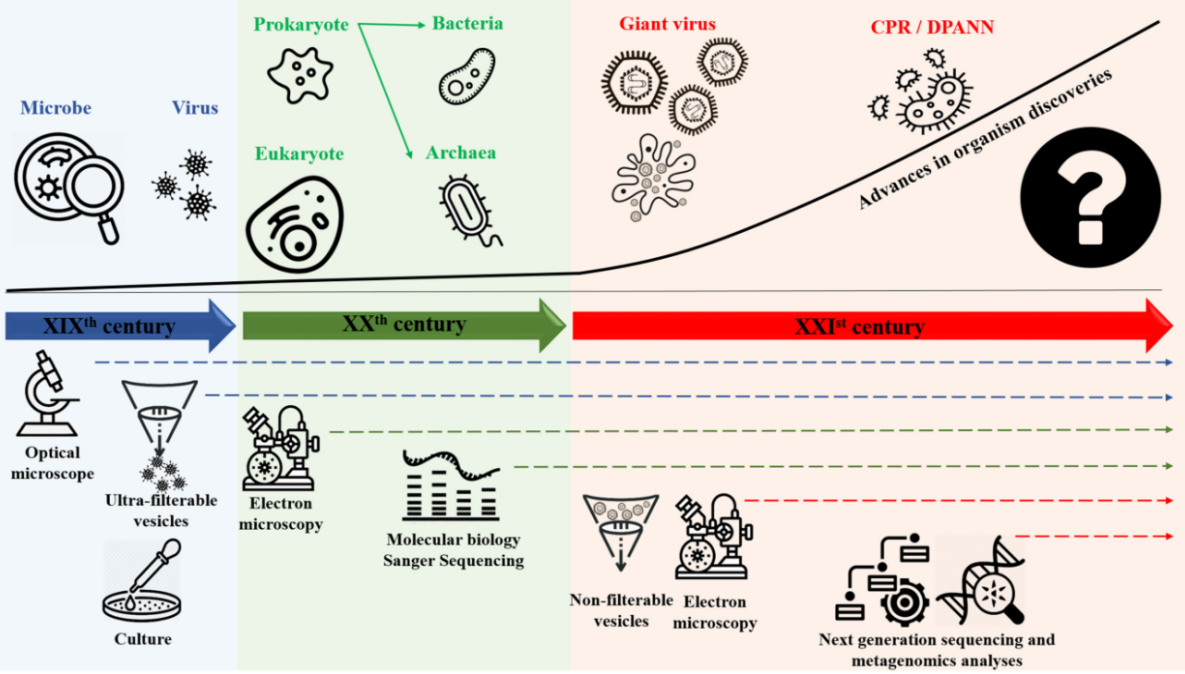


Figure 1 : Les découvertes des microorganismes au fil des siècles à travers les avancées scientifiques (tirée de Ibrahim et al., 2021)

1.2. La diversité des microorganismes

100 milliards c'est le nombre représentant les bactéries qui vivent dans notre intestin, 1 million c'est le nombre d'espèces différentes que l'on retrouve dans 1g de terre (Gans et al., 2005; Vitorino and Bessa, 2018). Les microorganismes constituent un réel pilier de la vie sur Terre. Parmi les microorganismes on retrouve les procaryotes qui regroupent les archées et les bactéries, les eucaryotes qui regroupent les champignons, levures, parasites et enfin les virus. Au sein de chaque groupe de microorganismes, la diversité est incroyable.

Rien que pour les bactéries, 42 phyla sont reconnus par le comité international de systématique des procaryotes (ICSP, « International Committee on Systematics of Prokaryotes ») dont le journal « International Journal of Systematic and Evolutionary Microbiology » (IJSEM) est l'organe de dissémination de l'information (Oren and Garrity, 2021). Les phyla constituent un rang taxonomique regroupant des espèces bactériennes qui ont des caractères similaires permettant de les regrouper ensemble. Parmi eux, on peut retrouver les Actinomycetota, les Bacillota, les Protéobactéries, les Bacteroidota. Les Bacillota anciennement appelées les

Firmicutes (Oren and Garrity, 2021) constituent l'un des phyla les plus abondants avec les Bacteroidetes. Les Bacillota regroupent les bactéries ayant une enveloppe cellulaire avec un taux de peptidoglycane élevé qu'on associe aux bactéries à coloration de Gram positive telles Bacillus, Staphylococcus, et Listeria. Le phylum des Bacteroidetes, récemment renommé Bacteroidota, contient un grand nombre de bactéries répandues dans l'environnement. Ces bactéries dégradent les polysaccharides contribuant à la libération d'énergie à partir de fibres alimentaires et à la sécrétion de CAZymes (Larsbrink and Sara McKee, 2020). Le groupe Actinomycetota anciennement appelé Actinobacteria est principalement composé de bactéries à coloration de Gram positive. Il y a cependant une grande diversité au sein de ce phylum en terme de morphologie, physiologie et capacité métabolique. Les morphologies des membres de ce phylum peuvent varier de la forme coccoïde avec par exemple les Micrococcus, bâtonnet-coccoïde (par exemple, Arthrobacter), hyphes fragmentées (par exemple, Nocardia) ou à des mycéliums ramifiés hautement différenciés (par exemple, Streptomyces) (Gao and Gupta, 2012). La sporulation des organismes est un mécanisme connu chez les bactéries comme chez *Bacillus subtilis* par exemple et un peu moins bien caractérisée chez les champignons et pourtant présent. C'est une stratégie largement utilisée par une grande variété d'organismes pour s'adapter aux changements de leurs niches environnementales individuelles et survivre dans le temps et/ou dans l'espace jusqu'à ce qu'ils rencontrent des conditions acceptables pour la croissance végétative (Huang and Hull, 2017).

Les archées étaient anciennement regroupées avec les bactéries du fait de leurs ressemblances morphologiques mais ont ensuite été distinguées grâce à l'analyse de la séquence de leur ARN ribosomique, puis de leur génome. Certaines sont connues pour vivre dans des environnements extrêmes, comme *Thermococcus gammatolerans*, capable de croitre à une température de 105°C et de supporter une dose instantanée de 5 000 Gy sans perte de viabilité (Zivanovic et al., 2009).

Les virus, qui ont été caractérisés plus récemment, sont des entités biologiques dépourvues de la fonction cruciale de synthèse des protéines, et ont besoin d'un hôte cellulaire pour se multiplier. Ils sont d'une extrême diversité et mutent très fréquemment générant un grand nombre de souches pour une même espèce de virus. La considération des virus en tant qu'être vivant ou non a longtemps été discuté puisqu'ils dépendent toujours d'un hôte pour se répliquer et vivre (Selosse, 2022). Les virus sont des entités biologiques nanométriques constitués d'une molécule d'ADN ou ARN enfermée dans une capside protectrice pour leur dissémination. Les virus peuvent également être nus ou enveloppés c'est-à-dire qu'ils ont soit seulement une nucléocapside, soit une enveloppe composée de phospholipides. En 2022, 11 273 espèces de

virus réparties en 264 familles et 2818 genres étaient recensés sur le site du comité international de taxonomie des virus. Ils sont responsables de nombreuses maladies comme celle de Ebola, celle de la Covid-19, la grippe, la gastro-entérite, la fièvre jaune. La plupart des virus ont une taille de l'ordre de 10 à 50 nanomètres. Dans les années 2000, ont été découverts les premiers virus géants ayant une taille similaire à des bactéries avec par exemple les Mimivirus ayant une taille de 450 nm (Moreira and Brochier-Armanet, 2008; Raoult et al., 2004). Dans la figure 2, la diversité au sein de chaque groupe de microorganismes est représentée (Vitorino and Bessa, 2018).

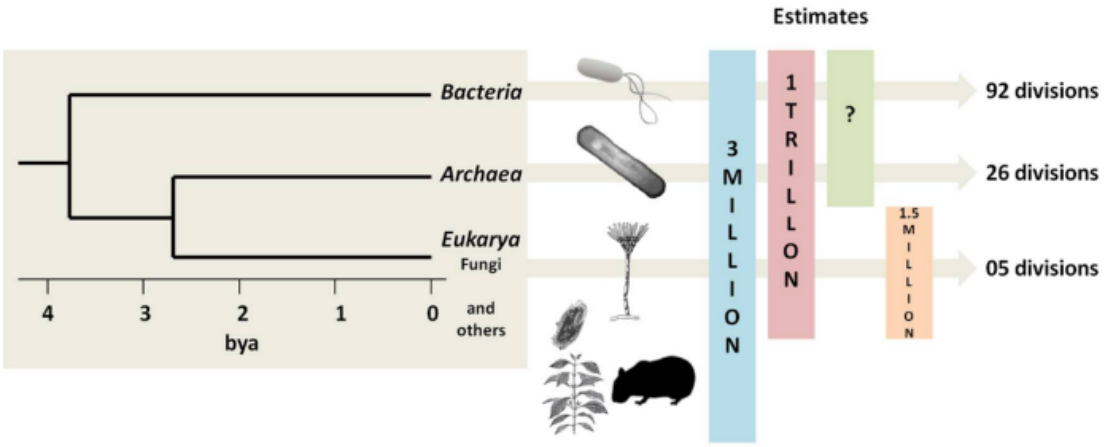


Figure 2 : Les estimations de la diversité des organismes selon les trois règnes (tirée de Vitorino and Bessa, 2018)

Sur la représentation de l'arbre phylogénétique des différents groupes bactéries, archées et eucaryotes en fonction du temps en milliards d'années (billion years ago (bya)), les bactéries sont le groupe le plus ancien et le plus riche en biodiversité suivi des archées et des champignons. 3 millions d'espèces d'organismes vivants sont recensés en revanche l'estimation sera largement supérieure et serait de 1 000 milliards d'organismes vivants.

Beaucoup d'espèces différentes ont déjà été cultivées, identifiées ou caractérisées mais cela n'est que la partie émergée de l'iceberg, il resterait encore 99 % des microorganismes à découvrir. En effet, d'après une étude réalisée en utilisant les données moléculaires de 35 000 sites, l'estimation du nombre de microorganismes est bien plus grande que celle que l'on connait actuellement. Un trillion (10^{12}) d'espèces microbiennes serait abrité sur la Terre (Locey and Lennon, 2016). Les dernières avancées technologiques avec les méthodes omiques, notamment la métagénomique, ont permis la découverte de nombreux organismes non

cultivables augmentant le nombre de microbes identifiés et classifiés à 10^5 (Vitorino and Bessa, 2018).

I.2. Une affaire de taxonomie avant tout

La taxonomie, la classification et la phylogénie sont des domaines étroitement liés, qui étudient la diversité et les relations entre les organismes vivants.

2.1. Définition et histoire de la taxonomie

La taxonomie est une science qui a pour objectif l'étude de la diversité du monde vivant. Le mot taxonomie a été proposé par Augustin Pyrame de Candolle en 1813 (De Candolle, 1815) et provient du grec *taxis* signifiant ordre, arrangement et *nomos* signifiant loi renvoyant directement à la science des lois de classification des êtres vivants. Cette science va consister à hiérarchiser les différents êtres-vivants en fonction de différents caractères afin de comprendre leur évolution et caractéristiques. Ainsi après une description de chaque organisme réalisé par diverses méthodes que l'on détaillera dans la prochaine partie, le microorganisme est attribué à un groupe, que l'on nomme taxon. La classification actuelle comprend 8 rangs taxonomiques majeurs : domaine, règne, embranchement, classe, ordre, famille, genre, espèce. Avant d'arriver à cette classification, de nombreuses hypothèses et règles ont été émises par différents pionniers de la science.

Au XVIIIème siècle, Carl von Linné posa les fondements de la taxonomie moderne par l'étude des plantes devenant le pionnier de la taxonomie (Sentausa and Fournier, 2013) même si au IVème siècle avant J-C, les philosophes grecs Aristote et Théophraste ont été les premiers à classifier les animaux et les plantes, respectivement. Le naturaliste suédois von Linné, aura permis à travers ses différents ouvrages notamment *Systema naturae*, de nommer et classer pas moins de 6000 espèces végétales et 4400 espèces animales (Linné, 1759). Il se reposait sur des critères de ressemblance morphologique pour classer les espèces découvertes. Cependant, le naturaliste suédois avait une position fixiste, c'est-à-dire qu'il considérait que les espèces vivantes n'évoluaient pas et qu'elles étaient la pure création de Dieu. Il aura fallu attendre la théorie de l'évolution proposée par Jean-Baptiste de Lamarck et Charles Darwin pour parler de notion d'évolution. Charles Darwin est grandement connu pour sa découverte de la sélection naturelle et donc de l'acquisition de certains caractères pour chaque espèce afin de survivre à l'environnement qui l'entoure. Il met donc en avant le fait qu'au sein d'une espèce, un caractère

peut évoluer pour s'adapter et se transmettre aux générations suivantes. Ainsi dans son ouvrage *l'Origine des espèces* paru en 1859 (Darwin, 1869), il y fait une classification généalogique classant les espèces selon leur degré d'apparentement évolutif, c'est-à-dire qu'il fait une classification tenant compte à la fois de la généalogie des espèces et de leurs distances phénotypiques. Il apporte la conception d'une descendance avec des modifications, ce qui modèlera la pensée phylogénétique. Ferdinand Cohn a permis de classifier les bactéries en genre et espèces, travail qui lui a valu la médaille Leeuwenhoek en 1885 et la médaille linnéenne en 1895 (Sentausa and Fournier, 2013).

2.2. La taxonomie des microorganismes et classification

L'étude de la taxonomie microbienne qui englobe la classification et la nomenclature et en finalité l'identification occupe une place primordiale dans le domaine de la microbiologie (Ciccarelli et al., 2006). Cette science est au cœur de l'identification microbienne puisqu'elle va permettre dans un premier temps de décrire et caractériser l'espèce, ensuite de la classer taxonomiquement c'est-à-dire d'établir un lien entre les différentes espèces existantes et enfin de définir un nom de l'espèce selon la nomenclature microbienne. Depuis l'évolution de notre compréhension du monde microbien, il est absolument nécessaire d'avoir un système efficace pour organiser et catégoriser la vaste diversité des microorganismes. Ainsi des règles et des représentations ont été mises en place pour représenter la classification de ces microorganismes. Charles Darwin a été le premier à représenter schématiquement la classification des êtres-vivants à travers un arbre dans son ouvrage *l'Origine des espèces* (Darwin, 1869). En 1866, Ernst Haeckel illustra le premier arbre phylogénétique et proposa le terme de phylogénie composé d'espèces réelles et de taxa supérieurs (Kutschera, 2011). Dans cet arbre, le tronc de l'arbre représente l'ancêtre commun d'où émergent les autres formes de vie. En 1968, il ajoute un troisième règne nommé Protiste regroupant les mycètes, les algues, les protozoaires et les bactéries et le place aux côtés des règnes plantes et animaux. En 1969, Robert Harding Whittaker propose un système de classification des organismes constitué de 5 règnes : Animalia, Plantae, Fungi, Protista et Monera (Hagen, 2012). La présence d'un noyau ou non (eucaryote ou procaryote), la composition de la paroi cellulaire, le niveau d'organisation des cellules (unicellulaire ou multicellulaire) et le type de nutrition sont les critères sur lesquels il se repose pour proposer sa classification. Enfin Carl Woese a permis en 1977 de définir un nouveau règne, les Archées menant ainsi à la classification actuelle en trois règnes : les eucaryotes, les bactéries et les archées. Toutefois, certains chercheurs s'accordent à dire que les virus peuvent constituer une branche de l'arbre de vie à part entière (Boyer et al., 2010; Nasir and Caetano-Anollés, 2015).

La taxonomie microbienne a pu réellement se développer grâce aux avancées techniques et méthodologiques de caractérisation des microorganismes.

2.3. Les caractères utilisés pour la classification bactérienne

2.3.1. Les méthodes phénotypiques

i) Les caractères morphologiques

L'aspect général des colonies obtenues par cultures de type clonage (isolats) peut donner des informations sur l'identification de par la forme de la colonie (ronde, bords irréguliers), la couleur, l'aspect (translucide, rugueux, lisse), l'odeur. Un milieu de culture particulier peut être utilisé pour permettre de cibler un type de bactérie. Les renseignements de type morphologique (coques, bacilles, taille) et de mobilité (présence d'un flagelle) peuvent être observés avec un microscope. Enfin la coloration de Gram, développée par Hans Christian Gram en 1884, permet de mettre en évidence la constitution de la paroi bactérienne ou non permettant la classification des bactéries en deux groupes : les Gram positifs (présence d'une grande quantité de peptidoglycane) et les Gram négatifs (peu de peptidoglycane et une membrane externe). D'autres tests comme celui de Ziehl-Neelsen permettent également de donner des informations sur le type de bactéries par la résistance ou non à l'acide-alcool (Alauzet, 2009). La cytométrie en flux est une technologie haut-débit en plein essor consistant à classer et sélectionner les microorganismes sur la base de leurs caractéristiques physiques (taille et forme). L'étude récemment publiée dans Nature a montré l'intérêt et l'efficacité de combiner les données multiparamétriques obtenues par la cytométrie de flux à des données obtenues par une méthode génomique (van de Velde et al., 2022). Ces observations des caractéristiques ont beaucoup aidé et aident les taxonomistes à classifier les microorganismes. Ces techniques ne sont cependant pas assez précises pour pouvoir identifier des microorganismes au niveau espèce.

ii) Les caractères biochimiques

Les activités enzymatiques de différents microorganismes sont un moyen de pouvoir différencier les différentes souches de bactéries. Ainsi des tests biochimiques sont capables par différentes réactions de déterminer l'identité de l'organisme présent. L'utilisation d'une source

de carbone, l'utilisation de divers glucides, la production de certains métabolites, la production d'enzymes spécifiques, la sensibilité aux antibiotiques ou encore la présence de CO2 nécessaires à leur survie sont des critères qui vont permettre de savoir s'il s'agit de bactéries de type aérobie/anaérobie, oxydative ou fermentative, antibio-résistante, etc... Les galeries d'identification API commercialisées par la société Biomérieux ont révolutionnées le monde de la bactériologie dans les années 1970 et peuvent toujours être utilisées dans certains cas de contamination agroalimentaire par exemple. Le test Becton Dickinson repose sur le même genre de réactions spécialisées pour les bactéries de type Enterobacteriacae et divers autres bâtonnets à Gram négatif (Alauzet, 2009).

iii) L'utilisation des acides gras

La caractérisation des acides gras peut être utilisée pour classifier et identifier les microorganismes. Les acides gras ont de nombreuses fonctions notamment des fonctions structurales car ils constituent des éléments essentiels des membranes cellulaires, peuvent être impliqués dans la signalisation cellulaire, et servir de matériau de stockage de l'énergie dans les cellules (Li et al., 2010). Ils comprennent un grand groupe de composés chimiquement hétérogènes. On peut retrouver les acides gras impliqués dans la fluidité de la membrane (chaîne carbonées), les glycérolipides ou phospholipides qui sont les principaux constituants des procaryotes et eucaryotes, les glycérolipides qui sont impliqués dans le stockage chez les procaryotes et beaucoup d'autre encore (de Carvalho and Caramujo, 2018). Les études utilisant les acides gras pour l'identification des archées sont très rares du fait de l'utilisation d'isoprénoides au lieu des acides gras pour la constitution de leur membrane (de Carvalho and Caramujo, 2018). Chaque souche bactérienne possède une empreinte spécifique puisque chacune a un profil d'acides gras unique. En effet, les acides gras peuvent être saturés, insaturés, linéaires, ramifiés, contenant des groupes hydroxyles ou méthyles ou des groupes cyclopropanes et propre à chaque espèce bactérienne agissant comme une empreinte digitale. Il est important de prendre en compte que les conditions de croissance des bactéries (milieu, température de culture, …) peuvent modifier leurs compositions en acide gras.

Les acides gras méthyles esters (FAMEs, de l'anglais *fatty acid methyl esters*) ont été utilisés comme biomarqueurs pour l'identification de bactéries aérobies, anaérobies et anaérobie facultative dans des systèmes de traitements des eaux usées. Ils ont par exemple montré dans leur étude que les saturés et hydroxy FAMEs étaient exclusivement associés aux bactéries aérobies. Les bactéries facultativement aérobies présentaient plusieurs biomarqueurs FAMEs

notamment les insaturés, ramifiés, cyclopropanes et acides gras hydroxylés (Quezada et al., 2007). Dans une étude utilisant les chromatogrammes d'ions totaux (TIC de l'anglais *total ion chromatography*) pour l'étude des acides gras de trois espèces bactériennes, 9 pics ont été obtenus pour *Escherichia coli*, 6 pour *Bacillus subtilis* et 13 pour *Francisella novicida*. Il a été démontré que des acides gras hydroxylés étaient présents chez *E. coli* et *F. novicida* et non chez *B. subtilis*, les acides carboxyliques étaient uniquement présents chez *E. coli* et non chez *F. novida* et *B.subtilis* (Li et al., 2010).

Plusieurs méthodes peuvent être utilisées impliquant l'estérification et la chromatographie en phase gazeuse /spectrométrie de masse (GC/MS) ou l'analyse de désorption/ionisation laser assistée par matrice (MALDI). Une étude utilisant une analyse directe en temps réel (DART) couplée à un spectromètre de masse a permis de définir les empreintes de 10 espèces bactériennes dont cinq Gram + et cinq Gram – . Les profils des spectres MS montrent que certains pics correspondent à la matrice et d'autres sont spécifiques à chaque espèce (Cody et al., 2015). L'avantage de cette méthode DART peut fournir une analyse rapide. En 2022, des chercheurs d'une équipe hongroise ont montré que les conditions de culture (la température) avaient un effet sur le profil des acides gras du genre Pseudomonas, non seulement quantitativement mais également qualitativement, introduisant ainsi un biais au niveau de l'interprétation par la suite dû aux profils présents dans les bases de données (Mező et al., 2022).

2.3.2. Les méthodes génomiques

i) Le contenu en G+C de l'ADN

L'ADN est constitué de paires de bases de nucléotides qui sont la guanine (G), la cytosine (C), la thymine (T) et l'adénine (A). La guanine et la cytosine sont appariées et la thymine et l'adénine sont appariées dans l'ADN double brin et liées par des liaisons hydrogènes. La teneur en GC ou le pourcentage de G+C est utilisée pour mesurer la composition en nucléotides du génome. Les paires GC sont préférées aux paires AT qui sont moins stables lors de l'exposition à des températures élevées (Hu et al., 2022). La température de fusion ou de dénaturation de l'ADN est dépendante du contenu en G+C. Cette caractéristique constitue un paramètre important pour la classification des microorganismes puisque c'est un trait variable selon les organismes. En effet, les bactéries étant en constante adaptation, elles modulent l'utilisation de leurs nucléotides (Mann and Chen, 2010). Une multitude de facteurs liés à l'histoire de l'évolution et à l'environnement sont responsables de ces différences (Mann and Chen, 2010).

La taille du génome, l'abondance d'oxygène et de l'azote et l'absorption d'ADN étranger par conjugaison font partie de ces critères et peuvent faire varier la teneur en GC de 13 % à plus de 75 % chez différents genres de bactéries (Bohlin et al., 2017). Par exemple sur la Figure 3, les Actinobacteria ont globalement une teneur en %GC plus importante donc supérieur à 50% contrairement aux Firmicutes qui sont à moins de 50% (Bohlin et al., 2017). Dans une étude, il a été démontré que les organismes bactériens avec une teneur en GC élevée de l'ordre de 49 % ont de plus grands génomes (en moyenne 4 Mb) et sont souvent retrouvés chez les bactéries qui ne dépendent pas d'un organisme pour vivre. En revanche, les organismes intracellulaires dépendant d'un autre organisme pour vivre nommés endosymbionte ont des génomes plus petits (de l'ordre de 1.4 Mb) et une teneur en % GC moins élevée de l'ordre de 30% (Mann and Chen, 2010). Chez les bactéries, en général les différences de la teneur en GC au sein d'un même genre ne dépassent pas le seuil de 10 % et un seuil inférieur à 5 % pour estimer que les isolats appartiennent à la même espèce (Alauzet, 2009).

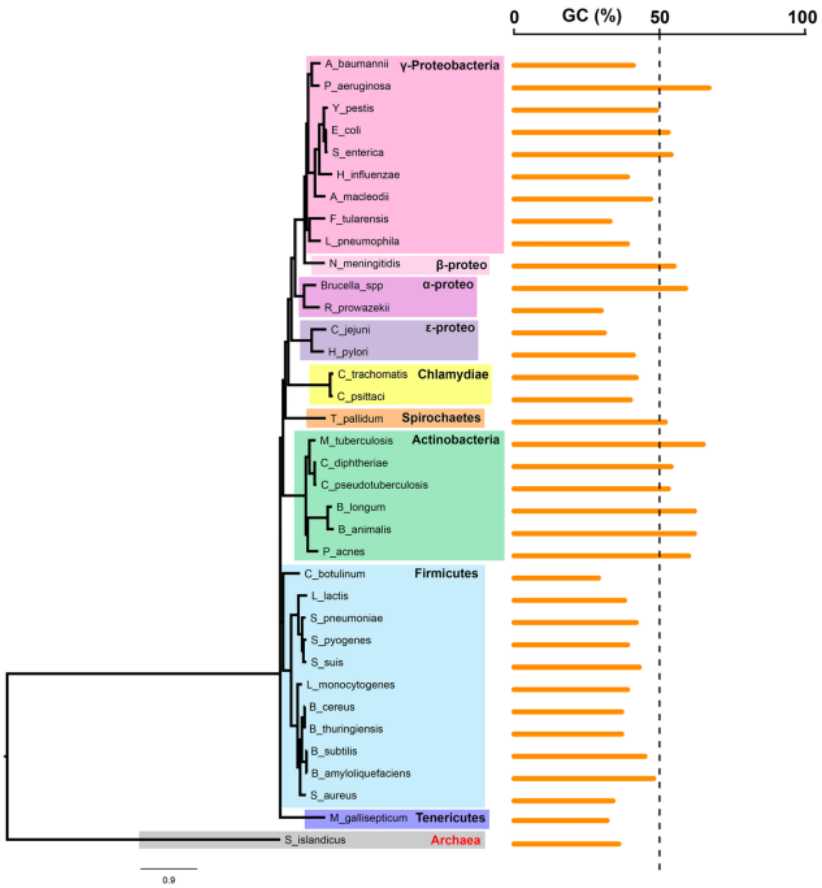


Figure 3 : Les variations de la teneur en G+C en fonction des différents phyla bactériens (tirée de Bohlin et al., 2017)

L'arbre phylogénétique basé sur une analyse ARNr 16S démontre les variations du pourcentage de GC en fonction du phylum, du genre et de l'espèce pour tous les microbes inclus dans leur étude.

ii) Détermination des taux d'hybridation ADN-ADN

La méthode des taux d'hybridation ADN-ADN (DDH de l'anglais *DNA-DNA hybridization)* repose sur la détermination de la similarité de séquence entre les génomes à comparer. Pour mesurer ce taux, la méthode se base sur le principe de renaturation de l'ADN. En effet, lorsque l'on chauffe l'ADN, les brins se dénaturent et se séparent pour donner des ADNs simple brin. Ainsi l'enjeu va être de créer un ADN hybride entre les ADNs des espèces à comparer et mesurer leur degré d'appariement comme décrit sur la figure 4. Plus les deux simples brins d'ADNs vont être complémentaires plus l'espèce sera proche. A l'inverse, plus l'hybridation sera partielle ou quasiment non complémentaire, moins l'espèce sera proche (Figure 4). Une valeur seuil de 70 % a été définie pour délimiter le niveau espèce (Goris et al., 2007). Lors de l'hybridation ADN-ADN, la stabilité thermique des appariements est étudiée, le ΔTm est la différence de température de fusion de l'hybridation entre les hybrides homologues et hétérologues formés sous condition standard. Cette technique fonctionne bien dès lors que les comparaisons d'organismes se font au niveau espèce. En revanche lorsque la comparaison se fait pour des espèces qui ne sont pas du même genre, de la même famille ou du même ordre, la méthode n'est pas suffisamment sensible. Cette méthode a largement été utilisée pour estimer la parenté génomique entre les micro-organismes. Elle a été considérée comme le critère de référence pour la délimitation des espèces de procaryotes. Ce critère était recommandé pour le dépôt de la souche avant d'être remplacé par les méthodes de séquençage où la comparaison est directement faite à partir des séquences des brins d'ADNs.

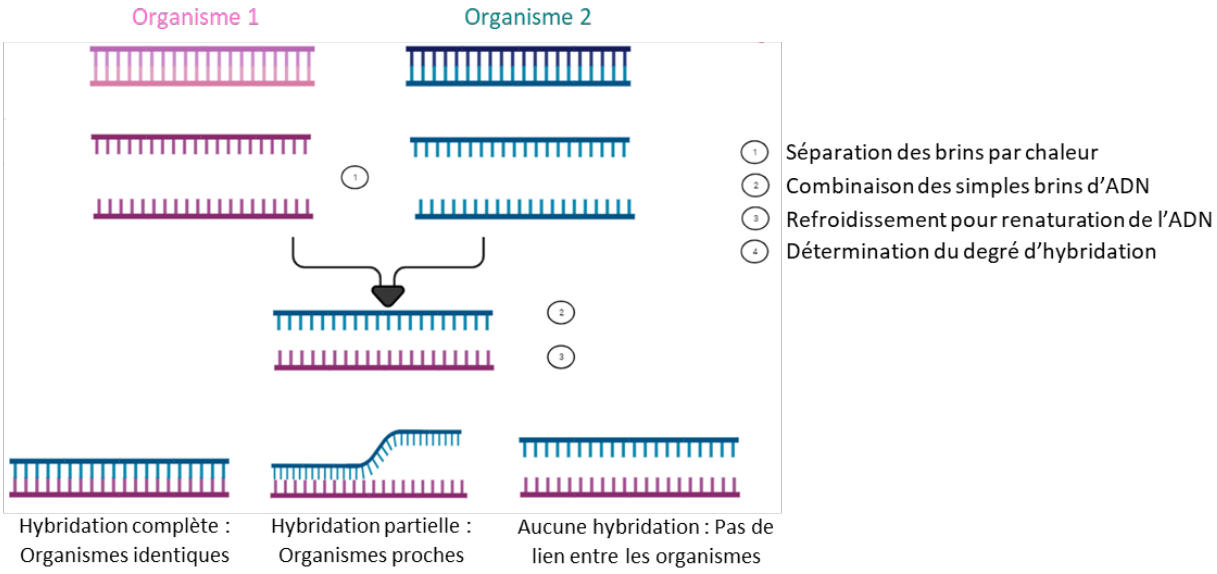


Figure 4 : Stratégie d'hybridation ADN-ADN pour l'identification d'un organisme

Les ADNs de chaque organisme sont chauffés. Les brins se dénaturent et se séparent pour donner des ADNs simple brin. Un ADN hybride est créé entre les ADNs des espèces à comparer pour mesurer leur degré d'appariement. Plus les deux simples brins d'ADNs vont être complémentaires plus l'espèce sera proche. A l'inverse, plus l'hybridation sera partielle ou quasiment non complémentaire, moins l'espèce sera proche.

En effet, la méthode d'hybridation ADN-ADN est de plus en plus considérée comme obsolète à l'heure où les méthodes de génomique sont très utilisées. Ainsi la méthode d'identité nucléotidique moyenne (ANI de l'anglais *average nucleotide identity*) entre deux génomes a été créée et reflète réellement la méthode DDH mais de manière plus moderne en utilisant le séquençage des génomes. L'ANI consiste à mesurer le taux de similarité de séquence au niveau nucléotidique entre les régions codantes de deux génomes et un seuil de similarité a été fixé pour assigner deux organismes à une même espèce. Il doit être supérieur ou égal à 95% (Goris et al., 2007; Konstantinidis and Tiedje, 2005; Richter and Rosselló-Móra, 2009). Une étude a permis de caractériser à haut débit 90 000 procaryotes en utilisant une méthode basée sur l'ANI appelé FastANI qui estime l'ANI à l'aide d'une cartographie approximative des séquences sans alignement et beaucoup plus rapide qu'une approche basée sur l'alignement (Jain et al., 2018).

iii) L'ARNr 16S

En 1977, Carl Woese et son équipe ont fait une découverte qui révolutionna l'identification des microorganismes (Woese and Fox, 1977). Cette révolution est basée sur le séquençage de l'ARNr des petites sous-unités ribosomiques, 16S, 5S, 23S pour les prokaryotes et 18S pour les

eucaryotes (notamment les champignons). Le gène de l'ARNr 16S est le plus utilisé pour le séquençage des bactéries du fait d'un optimum d'informations phylogéniques apportées par ce gène dont l'évolution en terme de séquence retrace l'évolution des microorganismes. Le ribosome est un complexe ribonucléoprotéique composé de protéines et d'ARN et est composé de deux sous-unités 70S (grosse sous-unité) et 30S (petite sous-unité). Dans la sous-unité 30S, on trouve l'ARNr 16S qui est encodé par 1500 nucléotides. L'ARNr 16S est constitué de neuf régions hypervariables et de régions conservées (Figure 5). C'est en utilisant la combinaison de ces deux types de régions que la caractérisation de l'espèce se base (Renvoisé et al., 2013) puisque la région ultra conservée est utilisée pour définir les amorces pour l'amplification par PCR (de l'anglais *Polymerase Chain Reaction*) du fragment à séquencer. De ce fait, l'utilisation d'amorces spécifiques permet de pouvoir se fixer au gène codant l'ARNr 16S de n'importe quel type de bactéries et de pouvoir séquencer ces séquences et ainsi couvrir toutes les bactéries présentes dans l'échantillon pour caractériser les nouvelles souches. Cette méthode est beaucoup utilisée dans le cadre d'étude des microorganismes composant un microbiote (Wensel et al., 2022) (O'Dwyer et al., 2019) ou pour la caractérisation de nouvelles souches (Lagier et al., 2012). De plus, les séquences des gènes ARNr 16S étant de 1500 paires de bases, cela permet d'être plus rapide en temps de séquençage par rapport à un séquençage du génome entier.

Le séquençage du gène de l'ARNr 16S est très utilisé en phylogénie car la vitesse d'évolution des divergences génétiques est assez lente permettant ainsi de reconstruire la phylogénie à partir des ancêtres et la molécule est systématiquement présente chez tous les microorganismes. Pour la caractérisation des espèces, la séquence obtenue est comparée à celles présentes dans les bases de données de type SILVA (Quast et al., 2012), Ribosomal Database Project (RDP) (Cole et al., 2009), GreenGenes (McDonald et al., 2012) qui diffèrent par leur taille et leur résolution en terme de nombre de taxons. SILVA et RDP donnent une identification au rang taxonomique genre et GreenGene est une base de données de petite taille utilisée uniquement pour les bactéries et les archées (Balvočiūtė and Huson, 2017). Cette méthode présente donc de nombreux avantages mais a quelques limites. Le coût des appareils et donc des échantillons ainsi que le manque d'automatisation de ces méthodes sont les premières limitations pour une mise en place de cette méthode en routine en microbiologie clinique. Les biais expérimentaux telle que l'extraction de l'ADN qui peut être insuffisante, la contamination de l'ADN soit au niveau de la préparation des échantillons (environnement, consommables, réactifs) soit au niveau de l'étape d'amplification par PCR (produits PCR d'une précédente analyse : cross contamination) sont aussi sources de limites puisqu'ils peuvent impacter l'identification finale en entraînant des faux positifs par exemple (Boers et al., 2015). Des solutions existent pour

limiter les étapes de contamination et augmenter la sensibilité et la spécificité au niveau de la préparation des échantillons avec des optimisations faites sur les protocoles d'extraction ou en utilisant des kits commerciaux (Sune et al., 2020). Les amorces sont définies de telle sorte à amplifier spécifiquement la région conservée mais il arrive que la région ne soit pas 100 % conservée se traduisant ainsi par une inexactitude et des ambiguïtés au niveau de l'identification (Boers et al., 2015). Certains laboratoires développent des échantillons standards pour limiter les biais de cette méthode appelée « *synthetic microbial community* ». De plus il arrive que certaines espèces aient plusieurs copies du gène ARNr16S dans leur génome entraînant ainsi des ambiguïtés au niveau des analyses, notamment les analyses quantitatives. Enfin, certaines espèces ont peu de variations entre elles au niveau des régions variables empêchant ainsi la discrimination des espèces. C'est par exemple le cas de certaines espèces de *Streptocoques* (*Streptococcus mitis* et *Streptococcus pneumoniae*), de certaines espèces d'*Entérobactéries* tels *Escherichia coli* et *Shigella spp* ou encore d'espèces de *Bacillus* comme *Bacillus cereus* et *Bacillus anthracis* (Renvoisé et al., 2013). Malgré les limites qui viennent d'être évoquées, l'approche du gène codant pour l'ARNr16S reste la plus fréquemment utilisée.

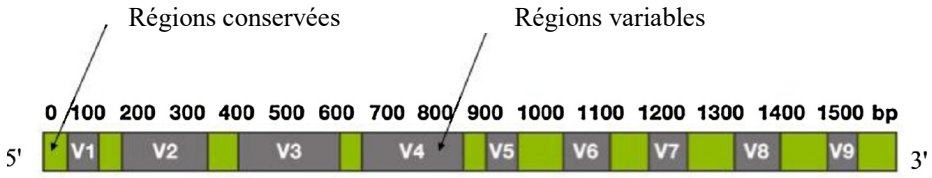


Figure 5 : Schéma de la composition d'un gène codant pour l'ARNr 16S (tirée de Renvoisé et al., 2013)

Le gène codant pour l'ARNr 16S est composé d'environ 1500 paire de bases et est constitué de régions conservées entre espèces et de régions variable entres espèces. Il est utilisé pour l'identification de souches bactériennes et archées.

iv) Séquençage des génomes entiers (whole-genome sequencing)

Le whole genome sequencing (WGS) consiste à étudier l'ensemble de la séquence ADN d'un organisme. Les premières méthodes de séquençage proposées par Sanger (Sanger et al., 1977) et Gilbert (Gilbert and Maxam, 1973) ont marqués l'histoire (prix Nobel 1980) et ont permis d'acquérir des connaissances sur les gènes au sein des génomes. Le séquençage Maxam-Gilbert se base sur le radiomarquage de l'ADN et le traitement chimique permettant de couper la chaîne de nucléotides au niveau des bases modifiées. Celui de Sanger s'inspire de la façon dont une cellule recopie son propre ADN. L'utilisation d'analogues chimiques de

didésoxynucléotides (ddNTP) permet de générer des fragments d'ADN avec pour terminaison chaque analogue marqué correspondant à chaque base (A, T, C et G). Après migration sur un gel polyacrylamide, l'ordre de chaque base est déterminé et permet ainsi de reconstituer le génome. Des améliorations ont été apportées tel que le remplacement d'éléments radiomarqués par des éléments fluorescents. En 2003, après 13 ans de recherche à travers le monde, le projet de séquençage du génome humain a abouti pour ce génome de 3 milliards de nucléotides (pour un coût de 3 milliards d'euros). Le séquençage de génome restait tout de même coûteux en main-d'œuvre et en argent et mal adapté à une utilisation de routine (Heather and Chain, 2016).

Les technologies de séquençage de nouvelle génération, NGS (New Generation Sequencing), ont révolutionné le séquençage en terme de coût et de rapidité. Actuellement 1 génome humain entier peut être séquencé en une journée pour un coût de moins de 1000 euros soit presque une baisse de 10 000 fois le coût d'avant 2003.

Deux méthodes de séquençage se basant sur deux technologies différentes sont très utilisées. Ces deux méthodes ont pour base commune l'extraction de l'ADN des organismes étudiés (bactéries, champignons, …) puis la préparation de librairies c'est-à-dire que l'ADN de chaque organisme va être découpé en fragments. Ces fragments vont ensuite être liés à un adaptateur qui contient un « barcode » unique pour permettre de multiplexer les échantillons. Une fois la librairie de séquences ADN composée, une des deux technologies de séquençage peut être utilisée. La première est la méthode développée par l'entreprise Illumina. Elle utilise des amorces universelles marquées avec un fluorochrome qui vont se lier à chaque extrémité des brins d'ADN. Ces brins d'ADN sont ensuite amplifiés pour avoir une plus grande profondeur d'analyse et sont séquencés. C'est grâce aux séquences marquées que la reconstitution des brins d'ADNs est réalisée. Les séquences chevauchantes peuvent être ordonnées pour reconstituer le génome entier (Fedurco, 2006; Hilt and Ferrieri, 2022; Voelkerding et al., 2009). La seconde méthode développée par Oxford Nanopore Technologies utilise des nanopores intégrés dans une membrane et un courant électrique appliqué à travers le nanopore. Lorsqu'un brin d'ADN va passer à travers le nanopore, les bases nucléotidiques modifient le courant électrique, chaque base de manière caractéristique, permettant ainsi de séquencer directement en temps réel le génome (Branton et al., 2008; Hilt and Ferrieri, 2022). Dans le cas de ces deux méthodes, le génome doit ensuite être reconstitué. Pour cela, différentes méthodes existent pour avoir une identification :

- L'assemblage de référence qui va réaliser un alignement des fragments d'ADN étudiés sur un génome de référence connu

- L'assemblage de novo où tous les fragments d'ADN sont assemblés en contig (ensemble spécifique de fragments d'ADN)
- L'assemblage hybride des assemblages de novo et de référence

Des améliorations majeures par rapport au séquençage de première génération ont donc été apportées et ont permis d'acquérir un plus grand nombre de données à moindre coût et ainsi permet à plus large échelle la possibilité d'étudier beaucoup plus de génomes. Cela a permis de rendre les NGS accessibles et contribue donc à l'identification de nombreux microorganismes (Park and Kim, 2016). Cette méthode permet de classer les microorganismes en mettant à disposition la séquence ADN complète d'un organisme. Ainsi on peut se baser sur ces informations pour voir de possibles adaptations, l'évolution de caractères ancestraux entre différents genres.

En début juillet 2023, la base de données NIH comptabilisait 30 211 eucaryotes, 527 904 procaryotes dont 513 292 bactéries et 14 612 archées et 66 161 virus et un mois après 30 530 eucaryotes, 567 228 procaryotes et 66 429 virus sont comptabilisés. Cela montre l'enrichissement des bases de données quotidiennement grâce aux nouvelles méthodes d'identification et caractérisation des organismes.

(https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/)

2.4. Les méthodes utilisées pour les représentations phylogénétiques

Les évolutions, les similarités et différences entre les organismes sont représentées visuellement à travers des arbres phylogénétiques obtenus par des méthodes de constructions phylogénétiques basées sur des marqueurs nucléotidiques ou protéiques. Ainsi les relations entre les taxa ou séquences et leurs ancêtres communs vont être représentés à travers un arbre phylogénétique (Hall, 2013).

Plusieurs étapes sont nécessaires à l'élaboration d'un arbre phylogénétiques pour pouvoir calculer les distances séparant les différentes espèces. La première consiste en l'identification et l'acquisition de séquences d'ADN ou protéiques similaires. La seconde va réaliser l'étape d'alignement entre les séquences similaires à comparer. La troisième consiste à estimer le degré de similarité entre les séquences alignées. La dernière étape consiste à construire l'arbre phylogénétique représentant visuellement les distances entre toutes les séquences analysées.

Représenter un arbre phylogénétique peut permettre soit de caractériser une nouvelle souche et ainsi identifier sa phylogénie soit de comparer plusieurs organismes pour avoir leur degré de similarité entre eux. Ainsi pour pouvoir réaliser ces comparaisons deux moyens sont utilisés :

- Dans le cas de la caractérisation d'une souche inconnue : utiliser le séquençage du génome entier ou de l'ARNr 16S pour obtenir la séquence ADN de ce nouvel organisme et l'intégrer aux bases de données
- Dans le cas d'une simple étude phylogénétique où les organismes sont déjà connus : utiliser les séquences ADN ou protéiques déjà disponibles dans les bases de données publiques comme GenBank et NCBI (de l'anglais *National Center for Biotechnology Information*)

Il est à noter que pour l'étude des séquences protéiques, des logiciels de traduction à partir de la séquence nucléique tel que Translate existent pour obtenir la séquence protéique.

Pour réaliser l'alignement des séquences, il faut avoir des séquences similaires c'est-à-dire qu'elles vont provenir d'un ancêtre commun et donc forcément avoir des régions en commun. Le moyen le plus fiable pour identifier des séquences similaires est BLAST (pour *Basic Local Alignment Search Tool*) (Altschul, 1997; Hall, 2013; Kapli et al., 2020).

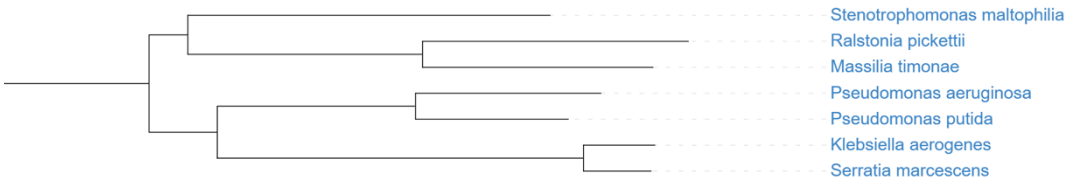
Afin de savoir si les organismes sont proches ou non, il convient de réaliser un alignement des séquences (Hall, 2013). Plusieurs méthodes d'alignements sont possibles. La plus connue est une approche progressive utilisée notamment par les logiciels MUSCLE (Edgar, 2004), CLUSTAL W (Thompson et al., n.d.) ou encore MAFFT pour *Multiple Alignment using Fast Fourier Transform* (Katoh et al., 2019), qui estime en premier lieu la similarité de chaque paire de séquence pour produire un arbre phylogénétique approximatif. Ensuite, l'ajout de séquences plus éloignées se fait à partir de l'arbre approximatif pour donner un arbre final. Les méthodes basées sur la cohérence estiment tous les alignements par paire et conservent un enregistrement des solutions alternatives à haut score pour chaque paire de séquences. Elles tentent ensuite d'identifier l'alignement global maximisant la cohérence entre toutes les paires. T-Coffee (Di Tommaso et al., 2011) et ProbCons (Do et al., 2005) utilisent ce type de méthodes d'alignements. Les méthodes basées sur la cohérence sont plus lentes mais plus précises que les méthodes progressives. Enfin des méthodes se basant sur la statistique peuvent être utilisées. Bali-Phy et StatAlign sont des logiciels utilisant cette méthode. L'alignement est ensuite utilisé par différents modèles mathématiques permettant de représenter la phylogénie des organismes étudiés.

Deux catégories de modèles de représentations phylogénétiques sont utilisées pour la reconstruction d'arbres :

- Les méthodes basées sur la distance : elles impliquent le calcul d'une distance génétique entre chaque paire d'espèces (sur la base de la comparaison de leurs séquences alignées) et l'utilisation itérative de la matrice de résultante pour construire un arbre. La plus populaire est l'algorithme NJ pour *Neighbors Joining* (Saitou, 1987). Elle est basée sur l'évolution minimale signifiant que l'arbre sera plus fiable s'il y a moins d'évolution. Ces méthodes peuvent donner un mauvais résultat sur les espèces éloignées, les grandes distances étant difficiles à estimer.
- Les méthodes basées sur les caractères incluent le maximum de vraisemblance aussi appelé ML pour *Maximum likelihood*, le maximum de parcimonie appelée « Maximum Parsimony » (MP) (Mount, 2008) ainsi que l'inférence Bayésienne (BI pour *Bayesian inference*). Le ML évalue la probabilité que le modèle proposé et l'histoire hypothétique donnent lieu à l'ensemble de données observées. Le MP calcule le nombre minimum de changements de nucléotides ou d'acides aminés nécessaires pour expliquer les données en utilisant chaque topologie d'arbre possible (Kapli et al., 2020). Cette méthode requiert le plus petit nombre de changements évolutifs. La parcimonie est connue pour être plus sujette que les méthodes de vraisemblance aux erreurs systématiques. Enfin, le BI (Yang and Rannala, 1997) se base sur le calcul des probabilités postérieures des arbres phylogénétiques par la combinaison d'une probabilité a priori avec la fonction de vraisemblance.

Enfin la dernière étape est de représenter ces données mathématiques brutes sous forme d'arbres phylogénétiques. Ainsi plusieurs représentations sont possibles notamment l'arbre le plus courant dit *rectangular phylogram* (Figure 6.A) où les nœuds internes sont représentés par des lignes verticales (les nœuds représentant des groupes). Les modèles dits de radiation explosive (Figure 6.B) sont également utilisés. Des logiciels tels que MEGA pour *Molecular Evolutionary Genetic Analysis* ou encore itol pour *Interactive tree of life* peuvent être utilisés pour générer ces différentes représentations.

A)



B)

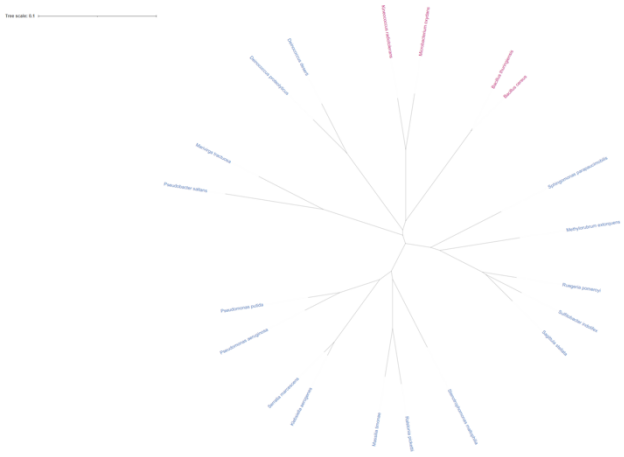


Figure 6: Exemples de représentations phylogénétiques

6A. Représentation d'un arbre phylogénétique de façon « rectangulaire ». 6B. Représentation d'un arbre phylogénétique de façon « explosive ».

Toutes ces méthodes sont donc utilisées pour pouvoir représenter la phylogénie des organismes. La figure 7 représente une synthèse des différentes étapes menant à la construction d'un arbre phylogénétique.

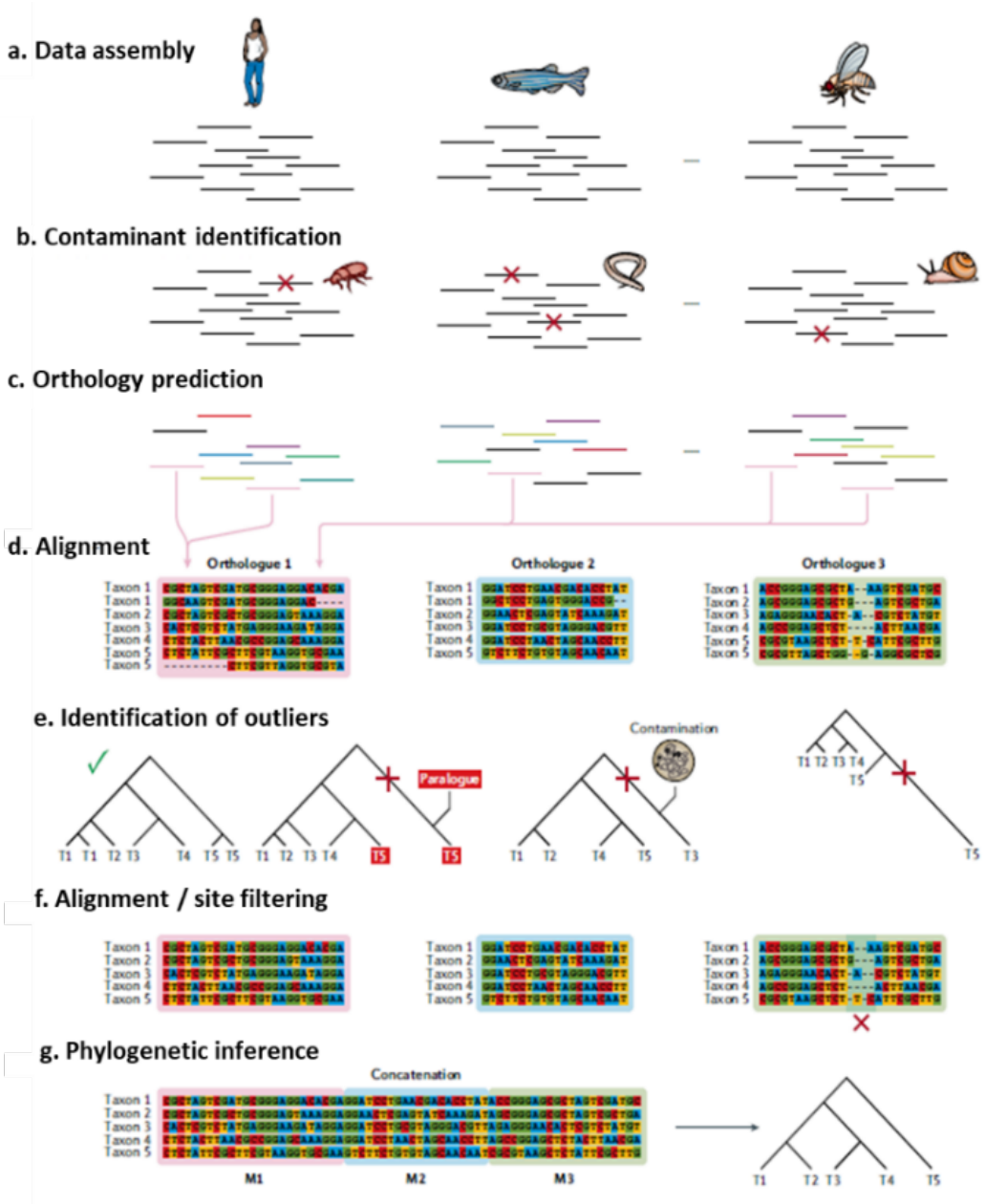


Figure 7 : Résumé des étapes pour l'établissement d'un arbre phylogénétique (inspirée de Kapli et al., 2020)

Plusieurs étapes sont nécessaires pour mener à la construction d'un arbre phylogénétique. Dans un premier temps les séquences génomiques de chaque organisme sont obtenues. Une suppression des contaminants est réalisée afin d'éliminer des problèmes d'identification. Suite à cela des alignements vont être réalisés par l'utilisation de logiciels d'alignements multiples et une étape de filtrage des données est réalisée pour pouvoir ensuite générer l'abre phylogénétique par le calcul des distances entre chaque organisme.

2.5. La nomenclature

La taxonomie consiste donc à identifier et classifier les microorganismes et pour terminer nommer les organismes identifiés. En effet, il est primordial d'adopter une nomenclature

précise pour chaque organisme identifié afin qu'il soit reconnu de tous. Carl von Linné avait introduit les bases de la taxonomie actuelle en proposant des catégories hiérarchiques (Règne, Phylum, Classe, Ordre, Famille, Genre et Espèce) et une nomenclature dite binomiale composée d'un nom pour le genre et d'une épithète pour désigner l'espèce. Jusqu'à 1947, les microorganismes étaient classifiés dans le *Botanical code*. En 1958, ils ont été reclassifiés dans le *Revised Edition of International Code of Nomenclature of Bacteria and Viruses* aujourd'hui appelé *International Code of Nomenclature of Prokaryotes* (ICNP). Les cyanobactéries sont toujours classifiés dans le Code Botanique en raison de leur capacité de photosynthèse les reliant aux plantes (Hugenholtz et al., 2021). L'ICNP permet de fixer des règles précises pour la nomenclature des procaryotes. Ces règles ont été fixées par un comité international appelé *International Committee on Systematics of Procaryotes* (ICSP) et peuvent être améliorées et corrigées lors de discussions et d'un vote de ce comité (Oren et al., 2022). Entre 2008 et 2019, pas moins de 45 corrections ont été apportées à ICNP.

La nomenclature actuelle propose de nommer les organismes par des noms latin ou latinisés. Le nom de l'espèce peut être associé à une personne, un lieu ou encore un caractère phénotypique. *Escherichia coli* est un exemple de cette nomenclature, Escherich est le nom du bactériologiste qui l'a découvert et *coli* représentant le colon, organe d'où la bactérie a été isolée et découverte. Autre exemple, *Staphylococcus aureus* a été nommée comme telle en rapport avec ses caractères phénotypiques (coques et couleur de la colonie jaune doré).

Avant de pouvoir nommer officiellement un organisme et qu'il soit validé par l'ICNP, des prérequis sont absolument nécessaires (Lagier et al., 2018):

- Disposer de la séquence du gène ARNr 16S de l'organisme avec une analyse BLAST basée sur la base NCBI nucléotides (avoir au moins 1300 pb)
- Le seuil de similitude de la séquence ARNr 16S avec l'espèce la plus proche phylogénétiquement doit être en-deçà de 98,7%
- Les données issues de l'hybridation ADN-ADN (longtemps méthode de référence mais maintenant considérée comme dépassée de par son manque de reproductibilité entre différents laboratoires et son manque de rentabilité) et le pourcentage du taux G+C% doivent être renseignés
- La séquence du gène ARNr 16S doit être établie et soumise à la base de donnée GenBank
- Décrire les souches avec des caractères morphologiques, phénotypiques

- Disposer de la souche (doit être cultivable) et son type doit être déposé dans au moins deux instituts de collecte de souches (type DSMZ, Institut Pasteur, ATCC)

L'*International Journal of Systematic and Evolutionary Microbiology* (IJSEM) est le journal de référence permettant de valider les nouvelles espèces et où la plupart des espèces sont décrites. Si la publication d'une nouvelle espèce est faite dans un autre journal, elle doit être validée officiellement par l'IJSEM. Les taxa validés ayant été décrits avant 1980 ont été publiés dans *Approved Lists of Bacterial Names*.

Sur la prédiction de 10 millions d'espèces bactériennes, seulement 15 000 ont été cultivées (Lagier et al., 2018). En effet, le problème de nomenclature des espèces non cultivables et analysées par des méthodes ne nécessitant pas de culture est réellement présent. Avec l'émergence des méthodes omiques comme la métagénomique, les organismes ne sont pas forcément cultivés et pourtant peuvent être identifiés comme de nouvelles espèces. En 1994, un statut taxonomique provisionnel de « *Candidatus* » a été proposé pour les taxa non cultivés. Depuis 1995, plus de 700 noms « *Candidatus* » sont recensés. Cependant cette proposition n'a pas été largement adoptée puisqu'elle ne représente que 4.9% des 45 414 taxa dans la base données *Genome Taxonomy* (Hugenholtz et al., 2021). Il est donc impératif de prévoir une nomenclature pour ces types d'organismes non cultivables qui se multiplient au fil des grands programmes de métagénomique. Après de nombreuses discussions, deux propositions ont été faites en ce sens et sont présentées dans la figure 8 (Murray et al., 2020a) :

- Accepter de valider les souches non-cultivées
- Créer une nomenclature pour ce type de souche et un répertoire dédié à ces souches.

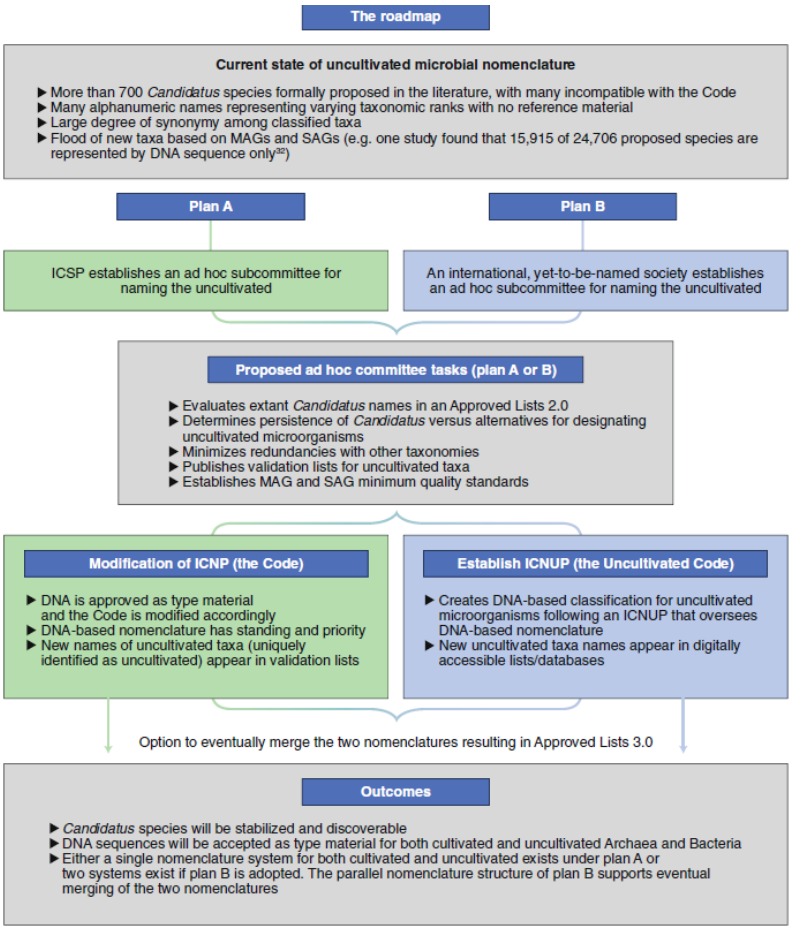


Figure 8 : Proposition d'une nouvelle nomenclature pour les espèces non cultivées (tirée de Murray et al., 2020)

Un arbre décisionnel pour la nomenclature d'espèces non cultivées a été proposé et se compose en deux parties soit la validation des souches non-cultivées ou encore la création d'une nomenclature pour ce type de souche et un répertoire dédié à ces souches.

I.3. L'identification des microorganismes pour la détection et le diagnostic

Les microorganismes caractérisés, nommés et classifiés par les méthodes exposées dans le paragraphe partie I.2 permettent d'enrichir les bases de données. L'identification des microorganismes dans un contexte clinique ou dans le cas d'une contamination agroalimentaire est primordiale pour le traitement. Ces types de contexte nécessitent d'avoir des méthodes rapides, applicables en routine et bon marché. Dans un contexte de criblage de microorganismes, notamment pour l'identification des microorganismes composant un écosystème, les méthodes utilisées peuvent être dans un premier temps rapides pour pouvoir se focaliser par la suite sur des organismes d'intérêts pour lesquels des méthodes plus longues et

plus coûteuses peuvent être mises en oeuvre. Les méthodes traditionnelles de diagnostic tels que la culture ou les tests biochimiques souvent utilisés dans le cas de contamination agroalimentaire restent valables dans certains contextes. Toutefois de nouvelles approches moléculaires permettent d'obtenir des informations sans précédent sur l'identification et le typage des bactéries y compris sur des échantillons complexes. Ces approches comprennent des méthodes simples basées sur l'amplification de l'ADN par PCR et des méthodes plus complexes basées sur l'analyse de fragments de restriction, le séquençage de gènes ciblés et de génome entier. Au cours des années 2000, l'identification par spectrométrie de masse à temps de vol par désorption/ionisation laser assistée par matrice (MALDI-TOF MS) a révolutionné la microbiologie médicale et est devenue la méthode de référence utilisée en routine dans ce domaine. Cette méthodologie est rapide, sensible et bon marché. Les différentes méthodes d'identification pour la détection des microorganismes seront décrites dans ce paragraphe.

3.1. Les techniques traditionnelles

Longtemps, la culture a représenté un moyen d'identification des microorganismes rencontrés en routine en diagnostic clinique et peut encore être utilisée de nos jours. Elle permet d'avoir une réponse relativement rapide puisque la croissance pour la majorité des bactéries pathogènes chez les humains est obtenue en 24 à 72h avec une exception pour certains types de bactéries comme *Francisella spp* (3-5 jours), *Brucella spp* (14 jours) (Doern, 2000) (Peyroux, 2022). Parmi les milieux utilisés, on distingue les milieux non sélectifs (ex : Columbia broth, Lysogenic Broth) qui favorisent la pousse de nombreux microorganismes et les milieux sélectifs (Austin, 2017) permettant de cibler un genre ou une espèce spécifique. Cette méthode est principalement utilisée en clinique pour les souches bien caractérisées et où les milieux spécifiques permettent de donner une réponse d'indentification plus rapide. Dans le cas de croissance difficile en plus des nutriments fondamentaux (Bonnet et al., 2020), l'ajout de certaines substances organiques spécifiques appelés facteurs de croissance (bases purines et pyrimidines, acides aminés, vitamines, sang, liquide ruminal) peuvent être nécessaires.

D'autres paramètres tels que la température d'incubation, le pH du milieu, la source de carbone, la résistance à des antibiotiques, à des antiseptiques sont aussi de possible moyens de sélection. Ainsi des phages ont récemment été introduits pour cibler un type de bactéries (Kim et al., 2021) ou pour améliorer la culture de *Streptococcus agalactiae* (Uchiyama et al., 2018). Le milieu Chapman connu pour sa forte concentration en chlorure de sodium sélectionne les organismes halophiles comme les *Staphylococcus*, *les Micrococcus*, les *Enterococcus* ou

encore les *Bacillus*. Le milieu BCYE (pour *Buffered Charcoal Yeast Extract*) est lui utilisé pour l'isolement des *Légionnelles* et plus particulièrement *Legionella pneumophila*. Des efforts sont même toujours déployés pour découvrir de nouveau milieux sélectifs d'intérêt pour les échantillons cliniques (Al-blooshi et al., 2021) (Ochoa et al., 2019) .

Les méthodes phénotypiques telles les galeries API (paragraphe 2.3.1) permettent l'identification d'environ 2000 phénotypes de microorganismes avec une précision pouvant aller jusqu'à 93% mais souffrent d'un manque de discrimination des nouvelles espèces (Jin et al., 2011).

3.2. Les techniques moléculaires

Dans le cas de l'identification de souches plus complexes et difficilement cultivable comme par exemple les souches environnementales, les approches moléculaires peuvent être utilisées. La plupart des méthodes moléculaires se basent sur l'analyse des variations de séquence de l'ADN entre les différentes espèces avec soit des étapes d'amplification soit du séquençage (Franco-Duarte et al., 2019). L'identification peut se caractériser par deux questions :

- Est-ce-que l'organisme recherché est présent dans l'échantillon ? Dans ce cas des approches ciblées sont mises en oeuvre pour rechercher un organisme spécifique.
- Quels organismes sont présents dans un échantillon donné ? Dans ce cas-là, il est fait appel à des méthodes sans a priori comme le séquençage du gène codant pour l'ARNr 16S ou le séquençage complet.

3.2.1. Approches ciblées

La plupart des méthodes ciblées se basent sur l'amplification d'un gène. Les méthodes décrites ci-dessous peuvent être utilisées dans le cadre d'identification dans un contexte clinique où le délai de réponse doit être rapide, dans un contexte de sécurité alimentaire ou encore dans la lutte contre le bioterrorisme.

i) PCR, qPCR et LAMP

La PCR est une méthode d'amplification d'un gène qui par amplification enzymatique permet l'obtention d'un grand nombre de copies identiques d'un fragment d'ADN menant à l'identification (Franco-Duarte et al., 2019). La PCR utilise des amorces qui sont modélisées à partir des séquences ADN des organismes recherchés à l'aide d'outils d'alignements de

séquences. La PCR est constituée de trois étapes essentielles. Les brins d'ADN sont d'abord séparés par chauffage (dénaturation). Ensuite lors de l'étape d'hybridation, les amorces viennent s'hybrider par complémentarité au brin d'ADN. Enfin, l'étape d'élongation permet de compléter la synthèse de la copie du brin d'ADN à partir de l'amorce grâce à la Taq polymérase (*Thermophilus aquaticus*) (enzyme de synthèse d'ADN) et aux nucléotides présents dans le milieu. Toutes ces étapes nécessitent des variations de température réalisées par des thermocycleurs (Figure 9.A).

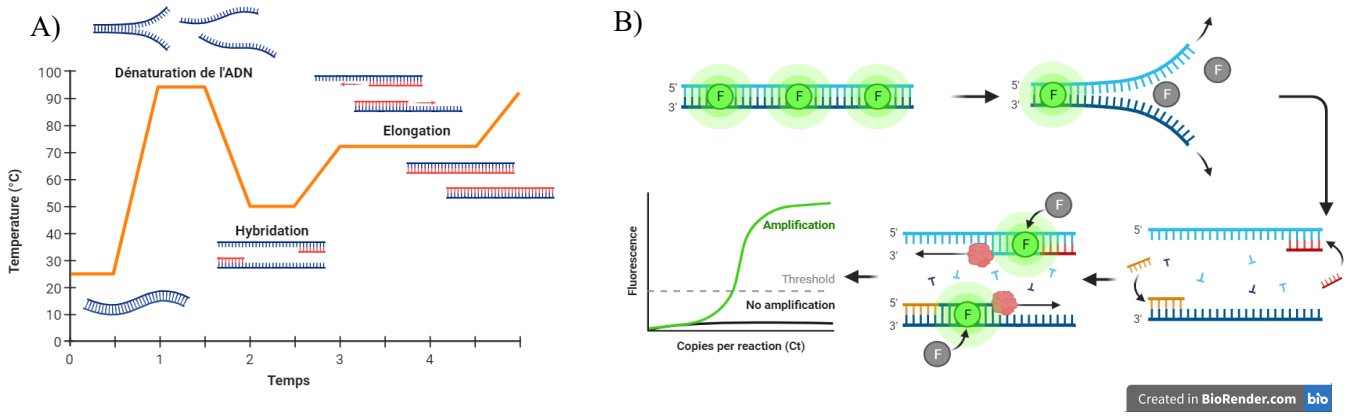


Figure 9 : Les différentes étapes nécessaires à la PCR traditionnelle (A) et la qPCR (B)

(A)La PCR est constituée de trois étapes essentielles. Les brins d'ADN sont d'abord séparés par chauffage (dénaturation) puis les amorces viennent s'hybrider par complémentarité au brin d'ADN. L'étape d'élongation permet de compléter la synthèse de la copie du brin d'ADN à partir de l'amorce grâce à la Taq polymérase (*Thermophilus aquaticus*) et aux nucléotides présents dans le milieu. Toutes ces étapes nécessitent des variations de température réalisées par des thermocycleurs. (B) La PCR en temps réel repose sur le même principe que la PCR et utilise des fluorophores qui vont s'intercaler entre les brins d'ADN permettant ainsi de suivre l'amplification de l'ADN.

La PCR quantitative (qPCR) ou PCR en temps réel est préférée à la PCR traditionnelle car elle présente une meilleure sensibilité et la possibilité de suivre en temps réel l'amplification des brins d'ADN grâce à l'utilisation d'un fluorophore (Figure 9.B). L'utilisation de fluorophores s'intercalant entre les brins d'ADN ou de sondes ADN spécifiques marquées avec un fluorochrome qui émet une fluorescence lors de l'hybridation avec le brin permet de suivre l'amplification de l'ADN. Cette technique peut être quantitative ou semi-quantitative pour quantifier la quantité d'ADN présente dans l'échantillon (Franco-Duarte et al., 2019; Kralik and Ricchi, 2017).

La technique de PCR LAMP (amplification isotherme à médiation par boucle) développée par Notomi en 2000 amplifie l'ADN avec une spécificité, une efficacité et une rapidité élevées dans des conditions isothermes (Notomi, 2000). Elle est de plus en plus utilisée comme alternative à la PCR traditionnelle. Dans cette méthode, la séquence cible est amplifiée à une température constante entre 60 et 65°C et utilise trois paires d'amorces et une polymérase ayant une activité de déplacement de brin élevée en plus d'une activité de réplication. Par des phénomènes d'hybridation des amorces et de formation de boucles, le brin d'ADN est amplifié (Geojith et al., 2011) (Figure 10). Cette méthodologie a par exemple été utilisée pour analyser *Mycobacterium tuberculosis* sur un terrain où les ressources sont limitées notamment dans les pays en développement (Geojith et al., 2011) . Elle a également été utilisée pour l'identification d'agents pathogènes responsables de la parodontite et préférée à une PCR traditionnelle en raison de sa plus grande sensibilité et de l'identification d'un plus grand nombre d'agents pathogènes simultanément. De plus la réponse de détection est rapide puisque cette méthode ne nécessite pas de culture (Lenkowski et al., 2021). La technologie LAMP ne nécessitant que peu de matériel permet une utilisation en ambulatoire en comparaison avec la PCR. Enfin l'utilisation de certaines amorces marquées par un fluorophore améliore la détection et permet le multiplexage si différents fluorophores sont utilisés (Hardinge and Murray, 2019). Cette technique permet d'avoir une réponse rapide et est réalisable sur le terrain mais reste une approche ciblée.

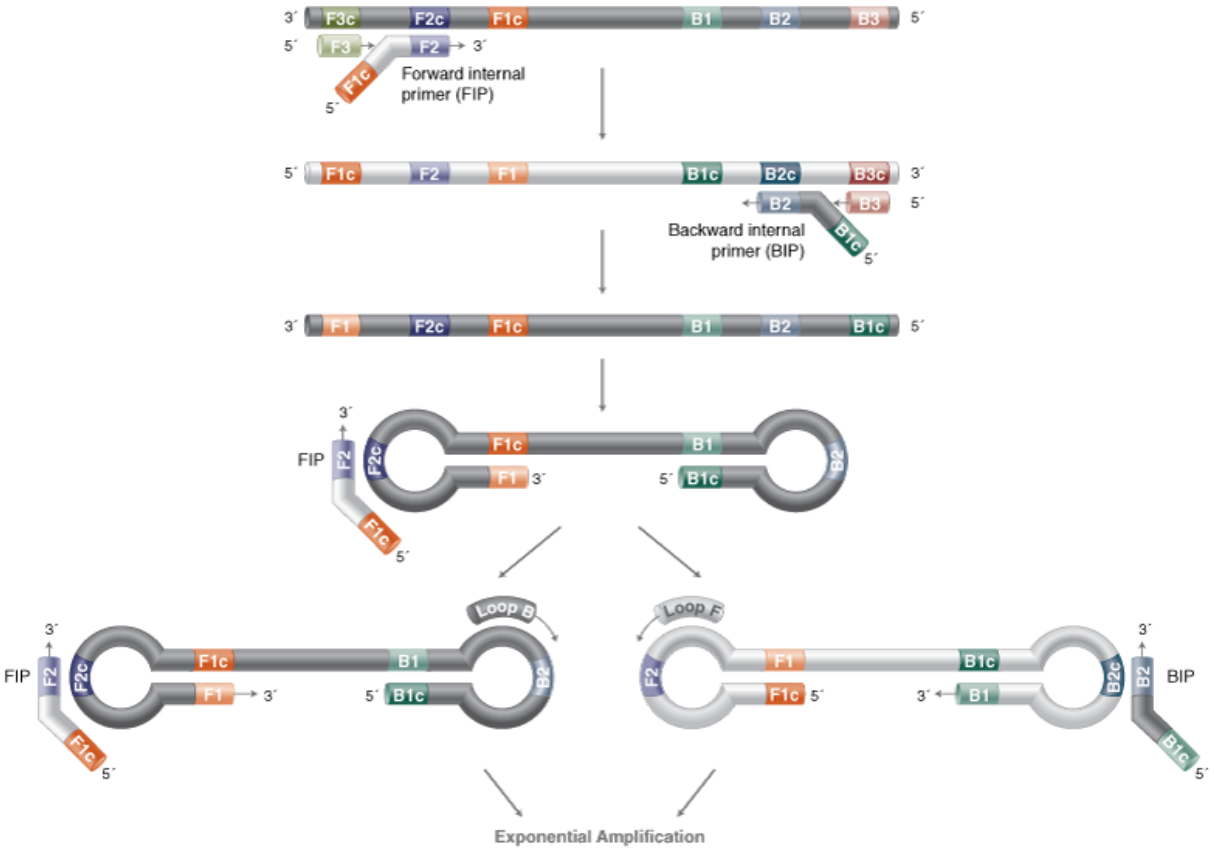


Figure 10 : Schéma de la stratégie de PCR LAMP (tirée du site New Englands Biolabs)

La méthode LAMP va amplifier la séquence cible à une température constante entre 60 et 65°C et utilise trois paires d'amorces et une polymérase ayant une activité de déplacement de brin élevée en plus d'une activité de réplication. Par des phénomènes d'hybridation des amorces et de formation de boucles, le brin d'ADN est amplifié et l'identification est réalisée.

ii) RFLP, AFLP et MLST

Le polymorphisme de longueur des fragments (RFLP pour *Restriction Fragment Lenght Polymorphism)* et l'AFLP pour *Amplified Fragment Lenght Polymorphism* sont des méthodes d'analyse génétiques reposant sur la détection de polymorphisme génétique inter ou intra espèces.

La méthode RFLP est utilisée comme une caractéristique des ADN permettant de les distinguer les uns des autres. L'ADN extrait est hydrolysé par des enzymes de restriction qui le coupe à des endroits spécifiques en différents fragments de longueur différentes. Si le site de restriction est intact alors la coupure sera réalisée, en revanche s'il y a une modification, la coupure ne sera pas réalisée et cela entraînera la coupure de fragments de tailles différentes. La séparation

des fragments par électrophorèse sur gel d'agarose permet de visualiser la différence de taille des fragments entre des ADN. Cette méthode est très souvent utilisée dans un contexte épidémiologique afin de voir si les souches infectantes sont les mêmes pour les patients infectés (Franco-Duarte et al., 2019).

Le principe de la méthode AFLP est similaire à la méthode RFLP mais diffère par l'ajout d'une étape d'amplification PCR sélective de l'ADN digéré. Des adaptateurs se lient aux fragments d'ADN sélectionnés et l'amplification de ces fragments est réalisée. Comme pour la RFLP, les fragments sont ensuite analysés sur électrophorèse d'agarose pour observer les différences. L'AFLP offre une meilleure sensibilité que la RFLP (Franco-Duarte et al., 2019).

La méthode de typage de séquences multilocus (MLST pour *MultiLocus Sequencing Typing*) est le polymorphisme de séquence de fragments. La MLST va permettre de caractériser une espèce ou des sous-espèces grâce au séquençage de 7 gènes dit « de ménages » (house-keeping genes) qui sont des gènes stables dans le temps et en nombre suffisant pour distinguer des souches entre elles. Cette méthode peut être utilisée dans les études épidémiologiques (Urwin and Maiden, 2003). Une récente étude montre la performance de la méthode dans un contexte épidémique en Malaisie où la mélioïdose, maladie tropicale, sévissait dans le pays. La mélioïdose est endémique en Asie du Sud-Est et en Australie du Nord et est de plus en plus préoccupante en Malaisie. L'agent causal est *Burkholderia pseudomallei* (Arushothy et al., 2020). L'analyse MLST des isolats cliniques de *B. pseudomallei* de tous les États de Malaisie a révélé une faible diversité et une association étroite avec les isolats d'Asie du Sud-Est montrant que l'agriculture et les activités de domestication étaient des voies à risque d'infection.

L'utilisation de la PCR est très utilisée pour la détection d'organismes cibles. On verra par la suite que de coupler la PCR à l'analyse du gène codant de l'ARNr 16S permet de travailler sans a priori et donc d'avoir une approche large spectre.

3.2.2. Approches large spectre

Il s'agit de répondre à la question : quels sont les organismes présents dans l'échantillon. L'intérêt des techniques d'analyse sans a priori est que même pour des souches inconnues et donc non caractérisées, l'espèce peut être identifiée comme nouvelle espèce avec son placement dans l'arbre phylogénétique.

i) RAPD

La méthode d'amplification aléatoire d'ADN polymorphe (RAPD pour *Random Amplification of Polymorphic DNA*) est une méthode se basant sur l'utilisation de la PCR. Elle utilise des amorces courtes d'environ 8 à 12 nucléotides non spécifiques appelés amorces *random* rendant cette méthode sans à priori. Les résultats de la réaction d'amplification PCR de l'ADN polymorphe amplifié au hasard fournissent un profil unique pour chaque bactérie permettant l'identification. Cette méthode est applicable directement sur les bactéries, sans besoin d'isolement de l'ADN et peut être appliquée sur les Gram + et les Gram - (Franco-Duarte et al., 2019). Cepedant du fait de l'utilisation d'amorces non spécifiques, la méthode n'est pas reproductible.

ii) Métagénomique

La métagénomique est un domaine d'application utilisant le séquençage comme moyen d'identification. Elle vise à étudier le microbiome composant un écosystème. Cette méthode séquence tous les génomes contenus dans un échantillon sans passer par l'étape de culture et d'isolement. Deux types d'analyses sont possibles, la métagénomique ciblée utilisant des marqueurs de détection comme par exemple la résistance à un antibiotique et la métagénomique classique qui est la plus utilisée pour la caractérisation des microbiotes et séquence tous les génomes présents. Le séquençage de tous les génomes génère des fragments d'ADN appelés lecture de séquences ou « *reads* ». Les fragments sont ensuite assemblés en *contigs* selon plusieurs méthodes : l'assemblage *greedy*, l'assemblage par recouvrement « *Overlap-Layout-Consensus* » et De-Brujn. Ensuite les contigs sont combinés pour reformer le génome complet. Cet assemblage passe par l'utilisation de bases de données d'assemblage de génomes MAGs pour *Metagenome-Assembled Genomes* dans le cas de microorganismes environnementaux en mélange. Cette approche résumée dans la figure 11 est très souvent utilisée pour la caractérisation des microorganismes composant un écosystème car elle est applicable sur échantillon direct et ne nécessite pas d'étape de culture, ce qui est un avantage pour les microorganismes dont la croissance est difficile. Dans certaines études cliniques, le signal humain écrase le signal microbien. Des solutions se développent pour palier à cette limitation en utilisant des méthodes se basant sur les différences de structure cellulaire et des variations génétiques qui diffèrent entre les humains et les microorganismes. L'utilisation de centrifugation différentielle, de la cytométrie en flux, de tampon de lyse sélectif sont autant de

solutions utilisées pour palier au niveau du signal humain (Shi et al., 2022). La métagénomique contribue également à la découverte de virus qui peuvent être compliqués à détecter par les méthodes de culture ou de méthodes avec a priori (Carbo et al., 2021; Mokili et al., 2012). La métagénomique a grandement contribué à la compréhension des fonctions microbiennes dans les écosystèmes grâce à des méthodes centrées sur le génome, l'annotation des fonctions, l'assemblage du métagénome et le regroupement dans des échantillons hétérogènes. Ces deux derniers points restent cependant un défi. Le développement de nouvelles plateformes d'analyse et de séquençage générant des séquences long-read aident à la compréhension de la taxonomie et des fonctions des microbes dans l'environnement (Taş et al., 2021).

La métagénomique peut devenir ciblée notamment si l'on ne veut étudier que les bactéries au sein d'une communauté, comme dans le cas du séquençage de l'ARNr 16S (cf. paragraphe I.2.3.2) pour l'identification des bactéries.

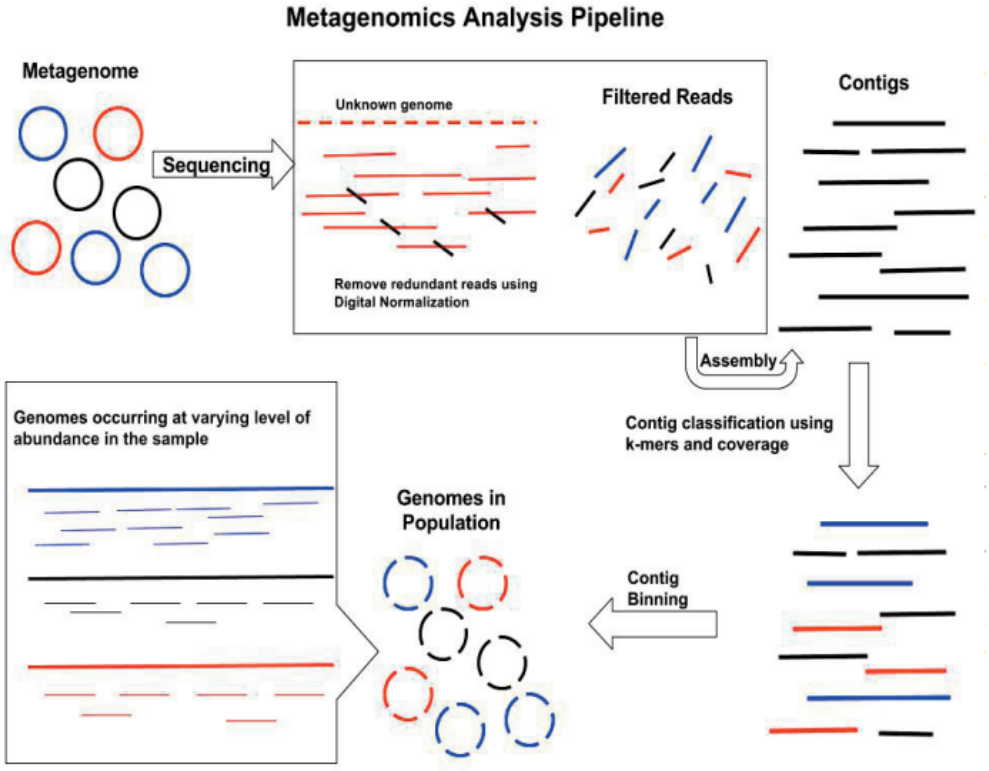


Figure 11 : Schéma de la stratégie de reconstruction des génomes par métagénomique (tirée de Ghurye, 2016)

Les cercles de différentes couleurs représentent plusieurs génomes bactériens indiquant que plusieurs individus forment un même organisme. Après le séquençage, les lectures redondantes sont supprimées grâce à la normalisation numérique. Les lectures filtrées sont ensuite assemblées en contigs et classées à l'aide de k-mers et de statistiques de couverture. Les contigs de chaque groupe sont ensuite regroupés pour former des projets de séquences génomiques pour l'organisation.

3.3.Les techniques d'identifications émergentes

D'autres technologies développées plus récemment peuvent être utilisées pour l'identification des microorganismes mais sont encore peu utilisées.

La production d'acides gras volatils par les bactéries sous l'influence des conditions de culture a été observée en 1921 démontrant la voie de détection des microorganismes grâce aux composés organiques volatils (VOC). Sur la base de cette observation, des techniques plus récentes utilisant les propriétés des VOC ont émergées. L'utilisation d'une chromatographie en phase gazeuse couplée à un spectromètre de masse est la méthode de référence pour la détection des VOC. Des études visant à trouver des VOC signatures de bactéries se développent notamment dans le cas de sepsis mais l'utilisation de différents milieux influe et perturbe l'identification finale. Des outils tels l'IMS (Ion Mobility Spectrometry) sont destinés à miniaturiser ces méthodes de détection pour pouvoir les utiliser en routine (Kunze-Szikszay et al., 2021) et a été démontré à travers une étude de souches antibiorésistantes (Steppert et al., 2021). Des applications dans le domaine environnemental sont également possibles avec l'exemple de la caractérisation d'une vingtaine de VOCs émis par des bactéries bénéfiques du sol comme *Pseudomonas sp.* pour lutter contre les phytopathogènes du sol (Huang et al., 2020; Montes-Osuna et al., 2022).

D'autres méthodes utilisant la spectroscopie infrarouge à transformée de Fourier (FTIR) peuvent détecter les microorganismes. En effet, lorsqu'un échantillon (isolat bactérien) est exposé à un faisceau infrarouge, le comportement vibratoire des molécules va être différent en fonction de l'absorption des molécules et donc de sa composition. Cette méthode est sans marquage, rapide, non destructive et adaptable au haut-débit (Zarnowiec et al., 2015). De même la spectroscopie Raman gagne de l'intérêt pour l'identification des microorganismes du fait du changement de fréquence des molécules après émission d'une lumière sur un milieu. Cependant les recherches pour l'utilisation de ces méthodes ne sont qu'à leurs débuts (Rebrosova et al., 2022; Wang et al., 2021).

Enfin la microfluidique, domaine en pleine expansion ces dernières années, s'intéresse également à la détection des microorganismes. La microfluidique intègre des méthodes physiques d'extraction reposant sur par exemple des différences de taille, la diafiltration, la séparation acoustique ou encore la conception particulière de la structure des microcanaux au sein de la puce microfluidique (Spatola Rossi et al., 2023; Zhang et al., 2018). Les avantages des systèmes microfluidiques sont la détection rapide, la facilité d'utilisation, les faibles coûts et la sensibilité (Nasseri et al., 2018). Une équipe a mis au point un outil intégrant la méthode

de résonance plasmonique surface (SPR) et la fluorescence (Figure 12.A). Les anticorps fixés sur l'or permettent la capture des organismes et la SPR et la fluorescence permettent la détection. Ainsi ils ont pu détecter *E. coli* et *S. aureus* de manière fiable en seulement 20 minutes (Zhao et al., 2019). Une autre équipe a mis au point un outil basé sur l'utilisation d'un smartphone-microscope et la détection par celui-ci de signaux fluorescents prévenants des bactéries. Ils utilisent une sonde d'acide nucléique peptidique (PNA) spécifique à l'espèce permettant ainsi d'être perçu par le microscope fluorescent couplé au smartphone (Figure 12.B). Les PNA sont des molécules de synthèse analogues des acides nucléiques qui permettent une forte spécificité. Cette méthode a permis d'identifier des bactéries du genre *Cronobacter* (Müller et al., 2018). Ces méthodologies sont prometteuses cependant l'utilisation d'anticorps spécifiques ou d'une sonde spécifique fait qu'il y a un a priori sur la détection (Zhao et al., 2019).

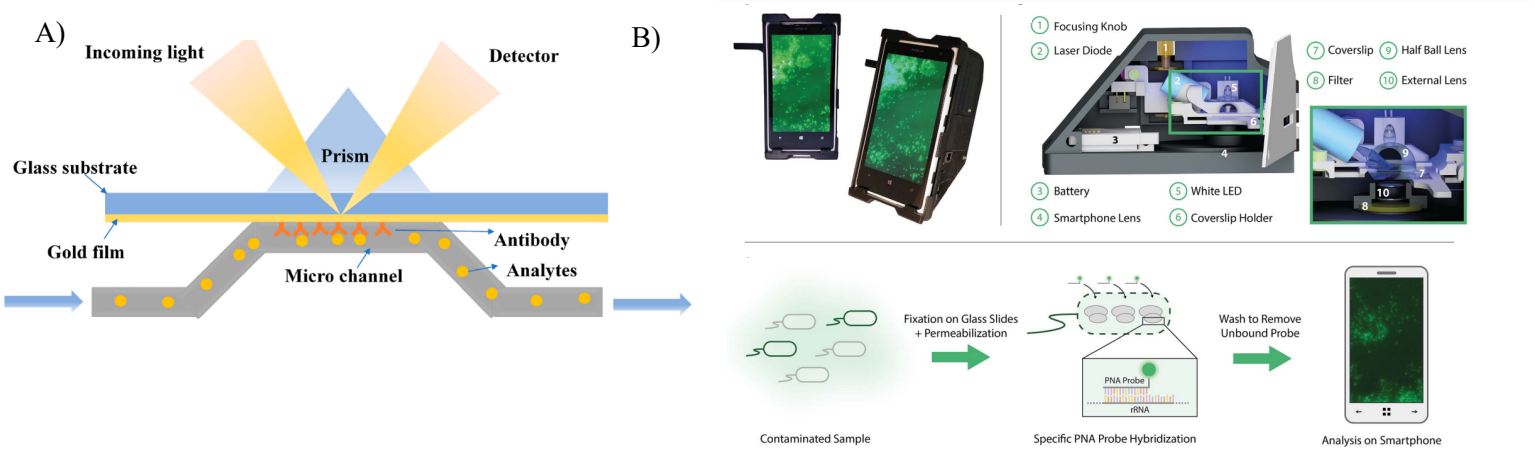


Figure 12 : Exemples de méthodes de microfluidique utilisées pour la détection de pathogènes se basant sur la SPR (A) ou encore sur l'utilisation de la fluorescence et d'un smartphone (B) (tirée de (Zhao et al., 2019 et Müller et al., 2018)

(A) Cet outil microfluidique repose sur l'utilisation de la méthode SPR pour permettre l'identification des microorganismes. Des anticorps spécifiques à un analyte sont fonctionalisés sur un film d'or et grâce à la méthode SPR, le comportement vibratoire généré si un analyte se fixe à l'anticorps modifier la valeur dectectée et ainsi identifier l'organisme. (B) Représentation schémation de la procédure de détection bactérienne. L'échantillon contaminé est fixé sur des lames de verre puis la membrane bactérienne est perméabilisée afin de permettre la pénétration de la sonde PNA dans les bactéries. Les sondes PNA, qui sont marquées avec un fluorphore, vont se lier spécifiquement à l'ARNr cible s'il est présent dans l'échantillon. Ainsi lors d'une étape de lavage, les sondes non liées seront éliminés et les sondes liées aux bactéries cibles vont emettre une fluorescence et vont être visualisées à l'aide du microscope basé sur un smartphone illustré en (A).

3.4. Une technique de spectrométrie de masse : le protéotypage par MALDI-TOF MS

A partir de 2009, la spectrométrie de masse type MALDI-TOF s'est imposée comme méthode de référence dans les laboratoires de microbiologie clinique mais aussi en tant que méthode de criblage pour la caractérisation de souches composant un microbiote.

Le MALDI-TOF MS combine une technique d'ionisation douce MALDI et un analyseur TOF pour *Time of Flight* ou temps de vol comme indiqué Figure 13. La source d'ionisation MALDI permet la désorption et l'ionisation des molécules entières permettant l'ionisation des protéines. Pour cela, l'échantillon cellulaire est directement mélangé à une solution saturée d'une matrice organique tel que l'α-cyano-4-hydroxycinnamique puis le mélange est déposé sur une plaque de dépôt MALDI en métal. Les microorganismes sont lysés par cette matrice et les protéines sont libérées. Les protéines et la matrice vont co-cristalliser sous l'effet de l'évaporation du solvant de la matrice. L'échantillon se retrouve alors sous forme solide. Suite à cela un flash laser, produit généralement par un laser à azote, va énergétiser le mélange matrice / débris cellulaires. Les molécules de la matrice vont absorber l'énergie du laser et la transmettre de façon optimale aux protéines entières pour les ioniser, et donc favoriser la désorption des protéines ionisées. Les ions de protéines entières sont subitement introduits dans un tube de vol maintenu sous vide et attirés jusqu'au détecteur. Ils sont analysés en fonction de leur ratio masse sur charge le long de leur trajet dans le tube de vol. Les ions ayant une masse faible atteignent plus vite le détecteur que les ions de masse élevée. A l'issue de l'analyse, on obtient une empreinte spectrale ou encore *fingerprint*, spectre associé spécifiquement à un organisme correspondant aux protéines entières, basiques et abondantes, telles que les protéines ribosomiques, la protéine HU, et certaines chaperons (Suarez, 2013). Des bases de données sont constituées à chaque analyse d'une souche connue permettant d'enrichir les bases de données des empreintes spectrales de chaque souche analysée. L'identification repose sur la comparaison d'une empreinte spectrale expérimentale aux empreintes spectrales contenues dans les bases de données. On appelle ce type d'analyse le protéotypage par MALDI-TOF MS basé sur la signature protéique unique d'un organisme pour pouvoir l'identifier par spectrométrie de masse (Emele et al., 2019; Ojima-Kato et al., 2023). Plusieurs appareils permettent de réaliser ces analyses dont le Vitek MS commercialisé par bioMérieux et le MALDI biotyper commercialisé par Bruker Daltonics. (Carbonnelle and Nassif, 2011; Suarez, 2013). En fonction des différents systèmes, l'interprétation du logiciel pour donner l'identification est exprimée de façon différente. Le MALDI Biotyper de Bruker donne un score d'identification lors du résultat de matching compris entre 0 et 3, si le score est inférieur à 1.7 alors il n'y a pas d'identification, si le score est compris entre 1.7 et 2 alors l'identification est

considérée au niveau genre et enfin si le score est supérieur à 2, le niveau espèce est validé. Le Vitek MS lui aussi utilise un score de confiance (exprimé en %) pour valider ou non l'identification. La fourchette de probabilité ente 60 et 99% permet la validation de l'organisme et plus la probabilité se rapproche de 99% plus l'identification est validée. En dessous de 60%, l'organisme est considéré comme non identifié. L'identification pour les bactéries Gram positives comme les *Staphylocoques* ou encore les levures peut être limitée du fait de spectres de mauvaise qualité. Cela peut être dû à la structure des bactéries Gram positives qui sont connues pour être difficiles à casser (paroi épaisse) mais aussi à une faible extraction des protéines contenues dans l'échantillon (Tsuchida et al., 2020). Des protocoles de traitement de l'échantillon avant l'analyse ont été mis en place pour pallier à ces limites notamment en utilisant de l'acide formique qui va favoriser l'extraction de protéines (Jang and Kim, 2018). Le Vitek MS possède une base de données de 42 000 spectres de références comptant 15 466 souches, et 2500 espèces alors que la base de données proposée par le Biotyper Microflex possède 150 000 spectres dont 6903 souches et 2461 espèces. La résolution de l'appareil de masse est meilleure pour le Vitek MS puisqu'elle est de 5000 contre 3500 pour le BioTyper Microflex (Jang and Kim, 2018).

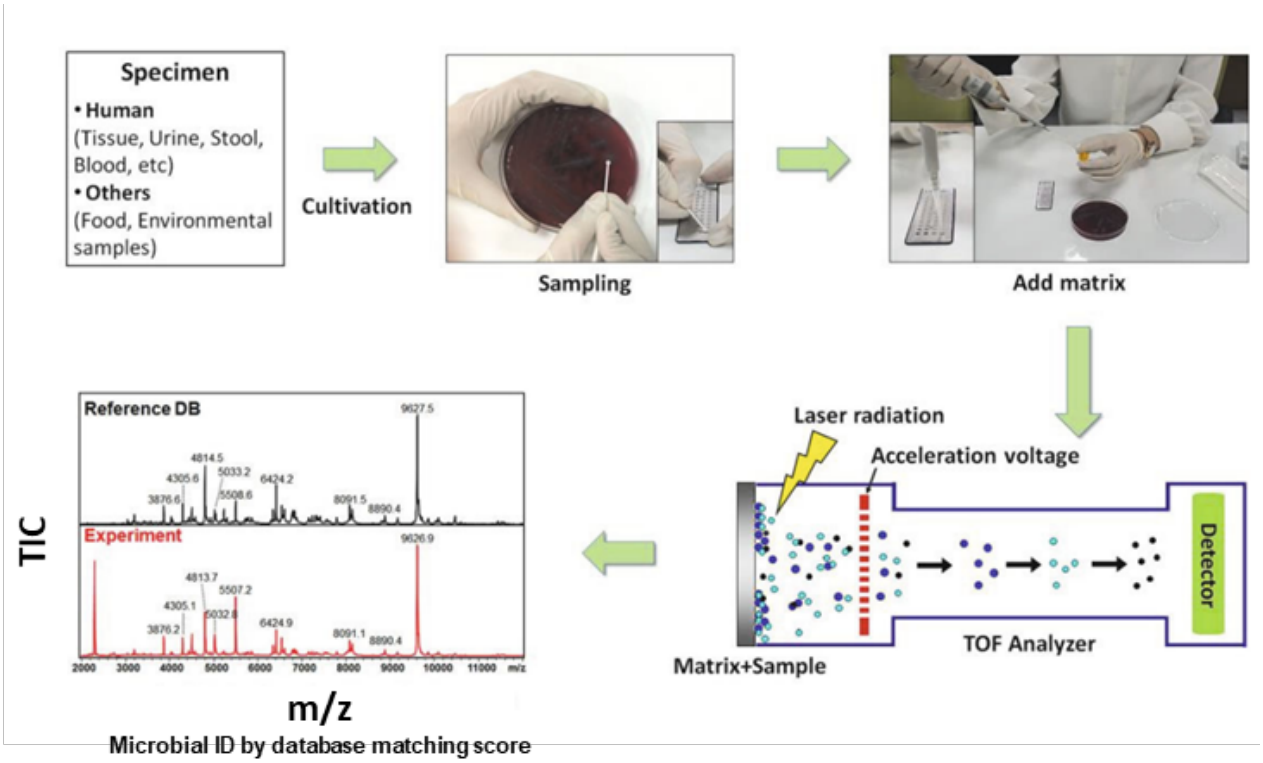


Figure 13 : Stratégie MALDI-TOF MS pour l'identification microbienne (inspirée de Jang and Kim, 2018)

A partir d'échantillons cliniques ou environnementales par exemple, des étalements sont réalisés sur boîte. Les colonies bactériennes sont prélevées puis déposés sur plaque métallique MALDI. Une matrice organique est directement mélangée à l'échantillon et ces derniers vont co-cristaliser. Chaque dépôt est ensuite soumis à l'action d'un laser. Les molécules de la matrice vont absorber l'énergie du laser et la transmettre de façon optimale aux protéines entières pour les ioniser, et donc favoriser la désorption des protéines ionisées. Les ions de protéines entières sont subitement introduits dans un tube de vol maintenu sous vide et attirés jusqu'au détecteur. Ils sont analysés en fonction de leur ratio masse sur charge le long de leur trajet dans le tube de vol.Ensuite les empreintes spectrales obtenues vont être comparées aux empreintes spectrales contenues dans les bases de données pour donner une identification.

Pour encore réduire le temps de réponse de l'identification, il est possible d'analyser directement l'échantillon clinique telles des hémocultures ou encore de l'urine en s'affranchissant de l'étape de culture comme dans le cas de sepsis où le diagnostic doit être le plus rapide possible. Des kits commerciaux sont mis en place pour chaque type d'échantillon clinique qui sont optimisés pour obtenir une meilleure identification en utilisant différentes étapes de centrifugation par exemple pour séparer les cellules humaines des microorganismes. Sepsityper de Bruker, Vitek MS blood culture de Biomérieux sont des exemples de kit commerciaux. Une équipe a mis en place un protocole combinant l'utilisation de plusieurs

milieux de culture et une sonication de 5 minutes pour arriver à un résultat de 100% d'identification en ayant un échantillon obtenu en 15 minutes contrairement à une application MALDI-TOF sur colonie nécessitant au préalable au moins 24h de culture sur milieu agar (Oviaño et al., 2018). En revanche, dans ce type d'analyse directe, si l'infection est polymicrobienne alors l'espèce majoritaire sera probablement identifiée mais les autres microbes ne le seront pas. Seule l'identification d'une espèce sera donnée avec un score permettant la validation. D'autres organismes seront identifiés mais avec un score faible. De plus en plus d'études recherchent également à trouver des signatures de résistance aux antibiotiques afin d'adapter rapidement les traitements antibiotiques (Florio et al., 2020).

Le MALDI-TOF MS est très utilisé en microbiologie clinique mais du fait de la nécessité de spectres de référence, son application à des souches environnementales est limitée.

Le MALDI-TOF MS est donc une révolution dans le domaine de l'identification. L'identification des microorganismes peut être fait jusqu'au niveau espèce (Dhiman et al., 2011). Malgré le coût important de la machine, le coût de l'analyse de l'échantillon est très rentable (moins de 1€) et en plus très rapide (en 6 minutes environ par échantillon) faisant donc de cette méthode une référence pour les analyses de routine (Bizzini and Greub, 2010).

I.4. Le protéotypage par spectrométrie de masse en tandem : l'utilisation de la protéomique bottom-up comme facteur discriminant pour l'identification taxonomique

La spectrométrie de masse est un outil puissant pour l'identification des microorganismes. Le protéotypage par spectrométrie de masse en tandem (MS/MS), basé sur des approches de protéomique, peut être utilisé pour obtenir des informations de séquences sur les protéines et palier aux limites du protéotypage par MALDI-TOF de cellules entières pour l'identification rapide de tout type de microorganismes.

4.1. La protéomique bottom up ou shotgun

La protéomique consiste en la description de l'ensemble des protéines d'une cellule, d'un tissu, ou organisme, à un moment donné et avec des conditions données. L'ensemble des protéines est appelé protéome et la protéomique est la science dédiée à son identification, quantification et caractérisation. Dans cette discipline, la spectrométrie de masse (MS) est une technique analytique de référence. On distingue trois types d'analyse protéomique: l'approche top-down, qui va étudier les protéines entières, l'approche bottom-up (ou shotgun), qui étudie les peptides issus d'une digestion enzymatique des protéines et va permettre de remonter à l'identité des

protéines et l'approche middle-down, qui consiste à analyser des peptides de grande taille, c'est-à-dire un intermédiaire entre les deux autres approches précédemment citées. L'approche top-down permet de répondre à des questions structurales sur les protéines, notamment pour la détermination des modifications des protéines et pour l'analyse des protéoformes. L'approche bottom-up est très utilisée car la chromatographie et la spectrométrie de masse sont plus sensibles et plus performantes au niveau de peptides qu'au niveau des protéines, permettant une analyse quantitative d'un grand nombre de protéines de l'échantillon (Dupree et al., 2020).

4.1.1. Le principe de la spectrométrie de masse

La spectrométrie de masse est une méthode analytique de mesure de masse.

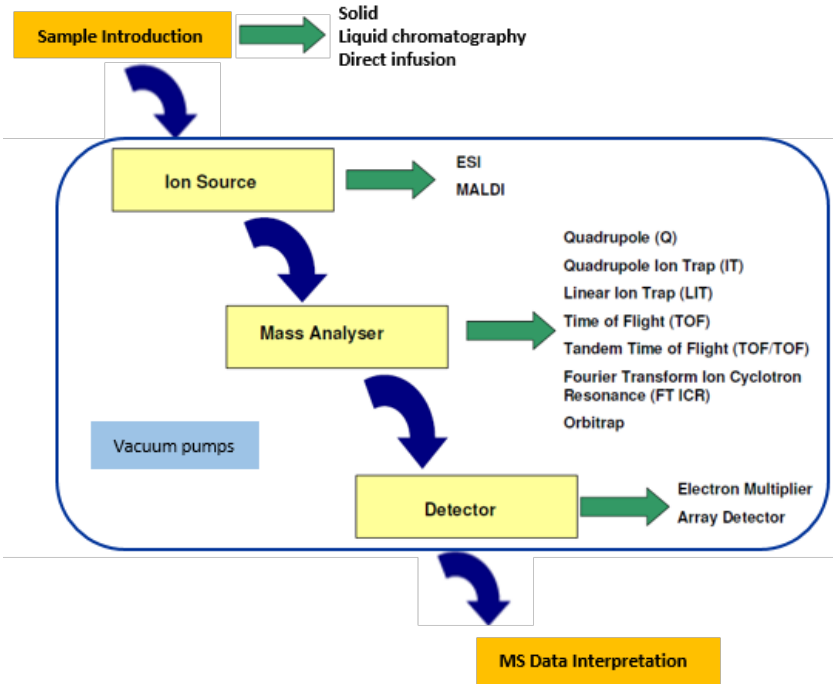


Figure 14: Schéma des composants d'un spectromètre de masse (inspiré de Graham et al., 2007)

Le spectromètre de masse se compose principalement de trois parties : la source d'ions, l'analyseur de masse et le détecteur. L'échantillon peut être injecté dans le spectromètre de masse de différentes manières notamment par une séparation par chromatographie en amont ou par une injection directe. Suite à cela l'échantillon va être ionisé par une source d'ions qui peut être MALDI ou ESI. Un analyseur de masse va permettre de séparer les ions en fonction de leur rapport masse/charge (*m/z*). Plusieurs types d'analyseurs de masse existent et peuvent être utilisés au sein d'un même appareil. Enfin le détecteur enregistre le signal d'intensité des ions pour chaque *m/z* possible. Dans certains cas, les signaux sont très complexes et nécessite une transformée de Fourrier pour permettre d'obtenir les intensités pour chaque possible *m/z*.

Le principe de la spectrométrie de masse réside dans la séparation en phase gazeuse de molécules chargées (ions) en fonction de leur rapport masse/charge (*m/z*). L'appareil se compose pricipalement de trois parties : la source d'ions, l'analyseur de masse et le détecteur (Figure 14), dans un environnement sous vide (Graham et al., 2007).

a) L'introduction de l'échantillon dans la source peut être faite de deux manières :

- directement sous forme gazeuse, solide (dépôt sur plaque MALDI,…) ou liquide (infusion directe). La méthode d'injection par infusion directe est très utilisée pour l'analyse en protéine entière. Ce type d'injection se fait souvent à l'aide d'un pousse-seringue directement dans la source ESI. Les avantages de cette technique sont d'obtenir une vue d'ensemble des espèces ionisables sans séparation, ce qui peut être représentatif de la qualité de l'échantillon. Ce mode est très souvent utilisé pour des études de métabolomique et de lipidomique où la signature de métabolites ou lipides est utilisée pour la détection de biomarqueurs (Marques et al., 2022; Mount, 2008). En revanche, dans le cas d'échantillons complexes, ce mode d'injection est beaucoup moins utilisé du fait de l'absence d'étape de séparation de l'échantillon générant des spectres moins informatifs qu'avec une séparation. Des méthodes sont toutefois mises en place en pour maximiser la qualité des spectres soit en améliorant la qualité d'acquisition des données avec par exemple la source de mobilité ionique FAIMS (Meyer et al., 2020) ou encore en améliorant les données acquises tels que la mise en place d'algorithmes de corrections réalisés sur plusieurs échantillons biologiques pouvant être en *open source* avec la mise en œuvre de méthodes de filtrage basées sur l'affectation d'éléments, l'erreur instrumentale et la soustraction de blancs par exemple.(Kirwan et al., 2014; Kozlova et al., 2022; Zielinski et al., 2018).
- indirectement par couplage avec une méthode séparative telle que la chromatographie en phase liquide majoritairement en polarité de phase inversée, l'électrochromatographie capillaire, la chromatographie en phase gazeuse….

b) La source d'ionisation est le lieu de vaporisation et d'ionisation des molécules. En biologie, les deux techniques d'ionisation courantes sont la désorption-ionisation de type MALDI et la désolvatation-ionisation de type *Electrospray Ionisation* (ESI) car elles n'engendrent que peu ou pas de fragmentation de la molécule durant le processus d'ionisation. La source MALDI a été décrite dans le paragraphe 3.4. La source ESI

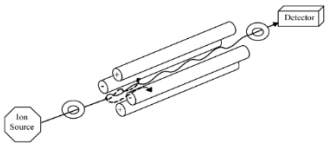
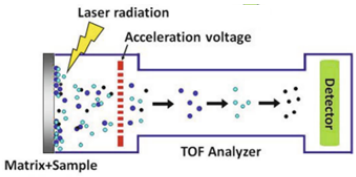
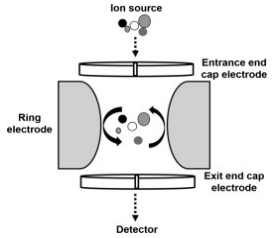
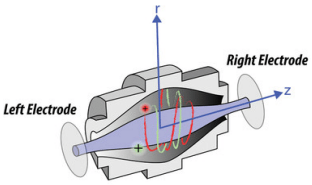
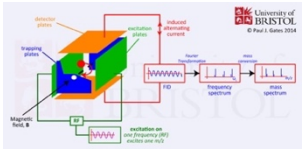
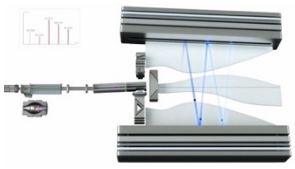
électrospray ou par électronébulisation a été développée en 1984 par John Bennet Fenn (prix Nobel en 2002). Son principe se base sur la dispersion d'un liquide sous forme de microgouttelettes chargées électriquement. Par l'application d'un haut potentiel électrique positif ou négatif, la formation de gouttelettes de plus en plus petites est réalisée, asséchant progressivement le solvant, et libérant les ions de molécules biologiques sans les fragmenter et qui sont envoyés dans le spectromètre de masse (Graham et al., 2007).

c) L'analyseur sépare les ions générés en fonction de leur rapport *m/z* avant leur détection. Plusieurs types d'analyseurs existent et ils se différencient par leur principe de séparation des ions. Les analyseurs les plus souvent utilisés sont le quadripôle (*Quadrupole* ; Q), le piège à ions (*Ion Trap* ; IT)), l'analyseur à temps de vol (*Time of Flight* ; TOF) et l'Orbitrap (Tableau 1). Ils se différencient par leurs caractéristiques et performances analytiques notamment en terme de résolution, précision, gamme de masse (El-Aneed et al., 2009). Plusieurs analyseurs peuvent être assemblés pour former des spectromètres de masse hybrides cumulant les avantages de chaque analyseur. Par exemple, un quadripôle peut être utilisé comme un filtre à ions très précis, en amont d'un piègeage sur un analyseur de type Orbitrap. En particulier lors de l'analyse protéomique, les informations sur les séquences des peptides sont obtenues grâce une analyse de spectrométrie de masse en tandem (MS/MS). Les ions produits dans la source sont séparés et après la détermination initiale de la masse, des ions spécifiques sont sélectionnés et soumis à une fragmentation par collision. Dans un appareil dit triple-quadripôle, le premier analyseur sert à séparer les ions formés dans la source et à sélectionner un ion parent d'intérêt, la fragmentation de cet ion sélectionné a lieu dans le deuxième analyseur, et le troisième analyseur sert à séparer les ions fragments générés et établir le ratio *m/z* des fragments ainsi générés.

d) Le détecteur collecte les ions sortants de l'analyseur, amplifie le signal, et quantifie l'intensité associée à chaque *m/z*. Les spectres de masse qui en résultent précisent l'intensité du courant ionique détecté en fonction du *m/z*.

Tableau 1 : Exemples d'analyseurs utilisés en spectrométrie de masse (tirée de El-Aneed et al., 2009, p.; Jang and Kim, 2018; Savaryn et al., 2016; Thomas, 2019, Heil et al., 2023; Pan et al., 2020)

Plusieurs exemples d'analyseurs sont présentés dans ce tableau notamment le quadripôle, le temps de vol, le piège à ions, l'orbitrap.

Nom de l'analyseur	Schéma de l'analyseur	Description	Résolution possible
Quadrupole (Q)		Cet analyseur est composé de deux paires d'électrodes cylindrinques parallèles sur lesquelles est appliquée un potentiel sinusoïdal, potentiel opposé deux à deux permettant la focalisation des ions. Le champ électrique quadrupole créé entraîne une oscillation des ions dont seuls les ions ayant une trajectoire stable seront analysés par le détecteur.	2 000 (à *m/z* 400)
Time Of Flight (TOF)		L'analyseur TOF mesure le temps de vol d'un ion, préalablement accéléré dans un champ électrostatique au travers d'un tube de vol, région libre de champ. Le temps de vol est corrélé au rapport *m/z*. Un réflectron est introduit pour l'extraction retardée des ions et l'amélioration de la résolution.	20 000 (à *m/z* 400)
Ion Trap (IT)		Le piège ionique (IT) est un piège ionique correspond à l'analyseur quadripôle mais en 3 dimensions. Cet analyseur est constitué d'une électrode annulaire et deux calottes sphériques. A ces trois éléments sont appliquées des tensions, f0 pour les calottes et -f0 pour l'électrode annulaire. Ces valeurs de tensions permettent d'accumuler les ions formés au sein de la trappe puis de les éjecter.	5 000 (à *m/z* 400)
Orbitrap		L'Orbitrap est composé d'une électrode creuse à l'intérieur de laquelle est placée coaxialement une électrode en forme de fuseau. Le. champ électrostatique quadripolaire dirigé vers l'électrode centrale provoque le piégeage des ions qui tournent autour de l'électrode centrale dans un mouvement orbital en fonction de leurs rapports *m/z*.	500 000 (à *m/z* 200)
FT-ICR		L'analyseur est constitué d'une cellule, dite trappe de Penning, délimitée par des champs électriques et confinée dans un champ magnétique. Son principe consiste à piéger puis à exciter les ions dans cette cage électromagnétique.	1 000 000 (à *m/z* 400)
Astral		La caractéristique de l'analyseur Astral est que la transmission d'ions se fait presque sans perte : >80% des ions qui entrent sont détectés. L'injection d'un paquet d'ions aligné et le guidage précis du mouvement des ions en trois dimensions sur une longue piste asymétrique permet une sensibilité et une résolution élevée (Thermo).	80 000 (à *m/z* 524)

4.1.2. La spectrométrie de masse en tandem (MS/MS) et la fragmentation

a) La spectrométrie de masse en tandem

Elle permet d'obtenir des informations précises sur la structure des ions parents qui sont fragmentés, et notamment lors de l'analyse de peptides d'obtenir leur séquence. Elle consiste à effectuer une sélection d'ions précurseurs (parents), la fragmentation de ces ions sélectionnés, et l'analyse des fragments. Cela implique d'avoir une cellule de collision adaptée où les ions fragments (ions fils) sont générés.

b) La fragmentation

Pour obtenir des informations structurales il faut fournir aux ions stables de l'énergie supplémentaire pour provoquer leur fragmentation. Il existe plusieurs modes de fragmentation, qui reposent sur des mécanismes différentes (Révész et al., 2023).

- Le mode *collision-induced dissociation* (CID), qui est le plus courant, utilise la collision entre des atomes d'un gaz inerte comme l'hélium, l'argon, l'azote et les ions précurseurs sélectionnés. (Johnson and Carlson, 2015). L'énergie de fragmentation peut varier : fragmentation de basse énergie (quelques électron-volt (eV)) et fragmentation de haute énergie (quelques keV). Dans les spectromètres de masse possédant un analyseur de type orbitrap le mode higher energy C-trap dissociation (HCD), qui est un mode de type CID dans son principe, permet d'énergiser plus rapidement les ions précurseurs, permettant d'obtenir plus d'ions fragments.
- Les techniques de fragmentation à médiation telles que la dissociation par capture d'électrons (ECD) et la dissociation par transfert d'électrons (ETD) utilisent l'interaction entre des électrons et des ions réactifs donneurs ou accepteurs d'électrons.

L'activation par collision est la méthode de fragmentation la plus utilisée en protéomique, permettant d'obtenir des données de séquences des peptides. Une nomenclature de fragmentation a été mise en place pour désigner et catégoriser les fragments en fonction de l'endroit de la rupture : la rupture de la liaison peptidique C(O)-N(H) conduit à la formation des séries d'ions fragments b, si la charge est retenue du côté N-terminal du peptide, et y, si la charge est retenue du côté C-terminal du peptide (produits majoritaires des modes CID/HCD) sont les plus abondants ; la rupture de la liaison C-C produit les ions fragments a et x, et la rupture de la liaison C-N conduit aux ions fragments c et z, avec rétention de la charge en N- et C- terminal du peptide, respectivement (produits des modes ETD et ECD) (Brodbelt, 2016). Le

Q Exactive est un spectromètre de masse hybride combinant deux types d'analyseurs de masse (quadripôle et orbitrap) (Figure 15) (Brodbelt, 2016). Cette configuration permet d'accroitre la vitesse de balayage de l'appareil, et d'utiliser de la chromatographie ultra rapide.

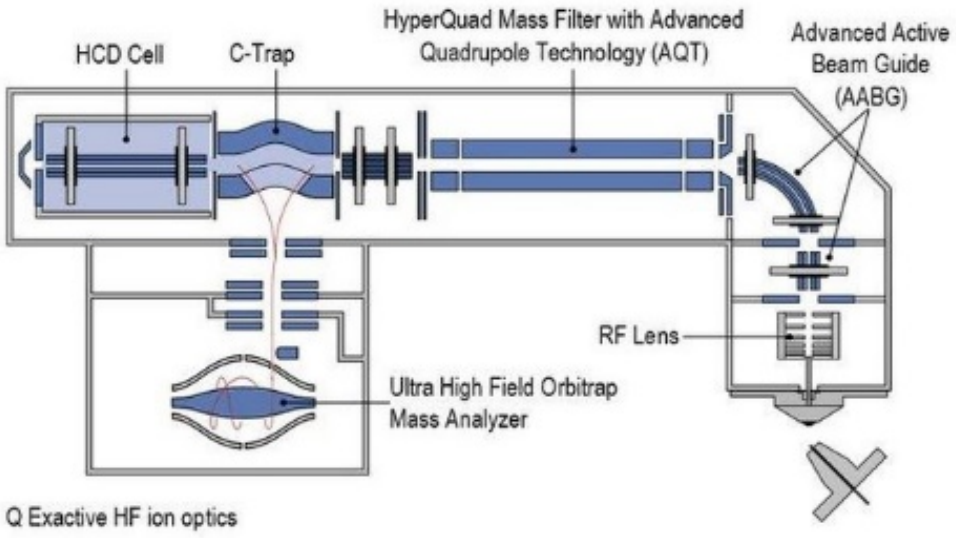


Figure 15 : Schéma du spectromètre de masse Orbitrap Q-exactive HF (tirée de Eliuk et Marakov, 2015)

L'Orbitrap Q-exactive HF est composé de deux analyseurs de masse avec une première analyse sur un quadripôle. Les ions sélectionnés à cette étape vont être fragmentés dans la cellule HCD puis accumulés dans la C-trap avant d'être analysés par le second analyseur : l'orbitrap.

L'utilisation de la spectrométrie de masse en tandem se fait très souvent en couplage avec une chromatographie liquide en amont de l'analyse MS/MS pour permettre la séparation des peptides de manière la plus résolutive possible en phase inverse selon l'hydrophobicité. Cela permet de générer ainsi des spectres MS/MS sur des entités peptidiques uniques, donc beaucoup plus simples à interpréter.

4.1.3. L'acquisition des données de MS/MS

Les données MS/MS peuvent être acquises selon plusieurs modes incluant :

- Le mode *Data Dependant Acquisition* (DDA)
- Le mode *Data Independant Acquisition* (DIA)

L'acquisition par le mode PRM (*Parralel reaction monitoring*) est également possible et va permettre la sélection d'ions pour l'analyse.

Le mode DDA est historiquement plus utilisé en protéomique bottom-up alors que le mode DIA, développé dès 2010 par Geiger *et al.* connaît un grand essor depuis quelques années (Krasny and Huang, 2021).

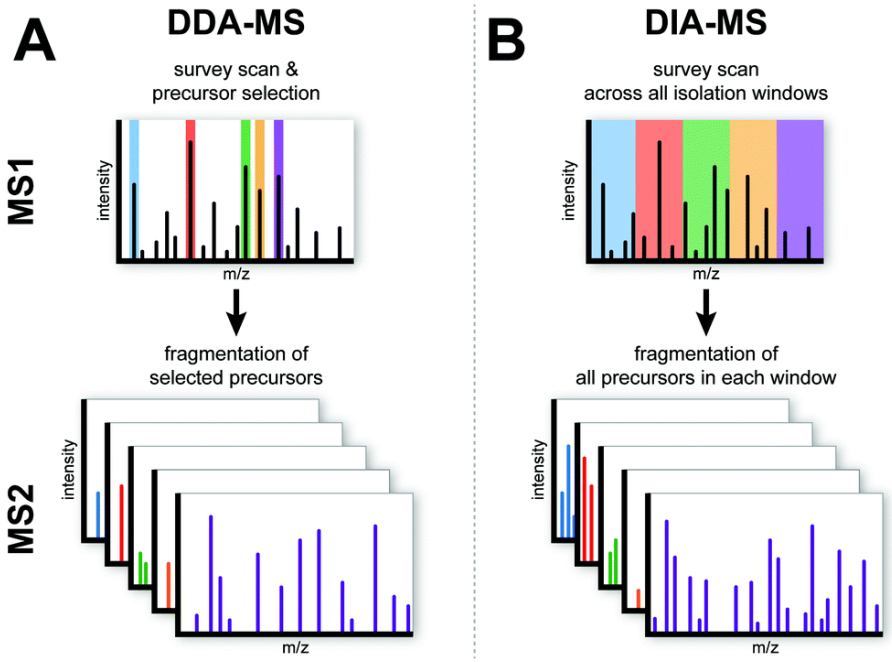


Figure 16 : Représentation schématique des deux stratégies d'acquisition des données de masse en mode DDA (A) et mode DIA (B) (tirée de Krasny and Huang, 2021)

(A) Le mode DDA sélectionne les ions précurseurs, en les priorisant par leur abondance (surlignés en couleur bleu, rouge, vert et violet), puis ces ions vont être les uns à la suite des autres fragmentés pour donner les ions fils. (B) Le mode DIA utilise des cycles d'analyse en fragmentant un ensemble de précurseurs sélectionnés sur une fenêtre de *m/z* étroite (fenêtres de couleur bleu, rouge, vert, orange et violet) et fait l'acquisition des données MS/MS de tous les ions fragments générés.

- En mode DDA le spectromètre de masse est paramétré pour sélectionner puis fragmenter les ions en alternant cycle MS et MS/MS. La sélection des ions est basée sur leur intensité. Le nombre d'ions précurseurs sélectionnés pour fragmentation (Figure 16.A) est prédéfini par l'utilisateur : on parle de stratégie Top N. Ainsi à

chaque cycle d'analyse on sélectionne les uns après les autres les N ions les plus intenses du spectre MS. N dépend de la vitesse de balayage du spectromètre de masse, par exemple, le Top 20 peut être utilisé pour l'instrument Q-exactive HF (Thermo Fisher). Cependant, le fait que la sélection d'un nombre défini d'ions précurseurs favorise les plus abondants aboutit à ce que certains ions précurseurs pertinents biologiquement ne soient pas analysés s'ils sont faiblement abondants. Une stratégie supplémentaire dite d'exclusion dynamique des ions précurseurs peut permettre d'éviter la sélection d'un ion précurseur déjà analysé pendant un temps donné (Kreimer et al., 2016). L'approche DDA est aussi parfois critiquée en terme de reproductibilité analytique (Fernández-Costa et al., 2020; Krasny and Huang, 2021), puisque il y a un échantillonage des ions choisis pour la MS/MS.

- En mode DIA pour chaque cycle d'analyse l'instrument se concentre sur une fenêtre de *m/z* étroite de précurseurs et fait l'acquisition des données MS/MS de tous les ions fils résultant de la fragmentation de cet ensemble d'ions précurseurs et conduisant à des spectres MS/MS chimériques. Les fenêtres sont échelonnées sur toute la gamme de masse avec la collecte systématique des données MS/MS. Le mode DIA permet de ne pas se limiter aux ions précurseurs les plus abondants (Figure 16. B). Pour pouvoir utiliser l'approche DIA, il est important de disposer de spectromètres de masse haute résolution avec une vitesse de balayage élevée. Cette approche permet une meilleure reproductibilité technique et une meilleure sensibilité (Collins et *al*., 2017, Gillet et *al*., 2012.; Krasny and Huang, 2021).

4.1.4. Interprétation des données acquises en DDA et DIA

a) En mode DDA

Lors de l'analyse par LC-MS/MS des peptides, des spectres MS/MS sont obtenus. Le fichier *mascot generic file* (mgf) généré contient pour chaque spectre le rapport *m/z* de l'ion précurseur ainsi que la liste des masses des ions fragments correspondants, la charge de l'ion précurseur et son temps de rétention. Pour l'identification des protéines un algorithme de recherche (appelé moteur de recherche) est utilisé. Il effectue la recherche selon des informations fournies par l'utilisateur incluant la base de données, l'enzyme utilisée et bien d'autres paramètres repris dans la section iii) ci-après.

Les masses expérimentales issues de l'analyse sont comparées à l'aide du moteur de recherche aux masses dites théoriques obtenues par digestion in-silico des protéines contenues dans les bases données sélectionnées avec l'enzyme utilisée dans l'expérience. Lors de cette comparaison les spectres MS/MS attribués à des séquences peptidiques sont appelés PSMs pour *Peptide Spectrum Matches*. Les peptides identifiés sont ensuite associés à des protéines par inférence (Granholm and Käll, 2011; Meyer, 2021).

i) Les bases de données

Les bases de données de séquences de protéines sont obtenues à partir des séquences génomiques des organismes référencés établies par séquençage global de leur génome (WGS) et annotation. Ces bases de données peuvent être généralistes comme NCBI et UniprotKb (*Universal Protein Resource*), qui sont des bases régulièrement mises à jour. NCBI propose la base de données publique complète des séquences de nucléotides appelée GenBank (Sayers et *al*., 2019) comptant en juin 2023, 243 560 863 séquences de protéines. UniProtKB est une base de connaissances complète sur les séquences de protéines qui se compose de deux sections : UniProtKB/Swiss-Prot, qui contient des entrées annotées manuellement, et UniProtKB/TrEMBL, qui contient des entrées annotées par ordinateur. Elle utilise une banque de données appelée Swiss Prot et compte à ce jour 248 842 690 séquences. Elles peuvent aussi être spécifiques c'est-à-dire qu'elles vont par exemple se limiter à un seul organisme. Des bases de données spécifiques aux résistances aux antibiotiques ou aux toxines existent par exemple (Akarsu et *al*., 2019; Alves et *al*., 2022). NCBI met aussi à disposition une base de données avec les séquences protéiques non redondantes appelée NCBInr.

ii) Les moteurs de recherche et paramètres d'interprétations

De nombreux moteurs de recherche existent, dont certains sont commercialisés comme Mascot Daemon (Perkins et *al*., 1999) et Sequest (Diament and Noble, 2011) de chez Matrix Science et Thermo Scientific, respectivement, et d'autres sont libres d'emploi comme X !Tandem (Bjornson et *al*., 2008), ou encore Andromeda (Cox et *al*., 2011).

Le moteur de recherche effectue la recherche selon des informations spécifiées par l'utilisateur portant sur les paramètres expérimentaux utilisés dont les principaux sont les suivants pour le moteur de recherche Mascot Daemon.

- La base de donnée utilisée
- L'enzyme utilisée pour la protéolyse des protéines qui influe sur les sites de coupure

- Le nombre autorisé de coupures manquées (*missed cleavage*)
- Les modifications chimiques fixes et variables attendus : carbamidométhylation des cystéines, oxydation des méthionines…
- Le type de spectromètre de masse utilisé qui peut influer sur le type d'ions fils mesurés
- La gamme de charge des précurseurs à considérer
- Les tolérances en masse sur les précurseurs (MS) et ions fragments (MS/MS) en lien avec la résolution spécifiée pour l'acquisition de chacune de ces entités.

iii) La validation

Lors de la recherche pour l'attribution des spectres aux protéines, un score de validation statistique est donné définissant le degré de confiance. En fonction des moteurs de recherche le calcul du score est différent. Par exemple X ! Tandem utilise une e-value pour « *expectation value* » soit la probabilité de commettre une erreur et Mascot utilise une p-value pour « *probability value* » signifiant la probabilité que l'attribution du spectre soit aléatoire. Dans l'approche protéomique bottom-up, les protéines sont identifiées à partir des peptides identifiés. La sélection des protéines candidates dans la base de données peut se présenter selon plusieurs cas de figure : i) Si plusieurs peptides sont attribués à une même protéine candidate ils permettent sa validation. ii) Si un même peptide est attribué à plusieurs protéines il est désigné comme peptide partagé. Dans le cas où le principe de parcimonie est appliqué : le peptide partagé est attribué à la protéine la plus abondante (Hayoun, 2020).

La validation des identifications obtenues est primordiale. La comparaison des spectres expérimentaux et théoriques peut parfois générer une mauvaise attribution d'un spectre MS/MS à une séquence, c'est-à-dire des faux positifs. Le nombre de faux positifs dépends, toutes choses égales par ailleurs, de la taille du jeu de données. Il est donc important de préciser leur fréquence d'occurrence en pourcentage. Une stratégie appelée target-decoy est le plus souvent utilisée pour évaluer les taux de faux positifs obtenus par le pipeline d'interprétation (Elias and Gygi, 2007). L'utilisation d'une base de données aléatoire (decoy) pour l'interprétation des spectres MS/MS, tel que par exemple une base inversée, permet le calcul du taux de faux positifs appelé FDR pour False Discovery Rate. Dans beaucoup de travaux de protéomique, un FDR de 1% est considéré, montrant en général un taux de confiance très élevé dans les résultats d'identification obtenus par cette méthodologie.

b) En mode DIA

L'extraction des données et l'identification des peptides à partir des données générées peuvent être réalisées en utilisant des librairies spectrales générées par une extension des données

protéomiques obtenues par acquisition préalable en mode DDA. Désormais, des alternatives utilisant des prédictions directement à partir de la base de données peuvent être mises en œuvre (Pino et al., 2020). Les logiciels permettant le traitement des données acquises par DIA sont en constante amélioration comme observé pour DIA-NN (Demichev et al., 2022) ou encore MSFragger-DIA (Yu et al., 2023).

4.2. L'interprétation des données de protéomique shotgun pour l'analyse taxonomique

La protéomique shotgun peut être utilisée pour l'identification taxonomique des organismes et on peut parler dans ce cas là de protéotypage par nLC-MS/MS. Cette identification va dépendre de l'attribution des peptides aux protéines par rapport aux références contenues dans les bases de données. Ainsi à partir des protéines identifiées, l'information peut remonter jusqu'à découvrir l'organisme qui les a produites. L'avantage de la méthode est qu'en cas d'analyse d'un organisme inconnu il est taxonomiquement identifié selon ses voisins les plus proches qui partagent un certain nombre de séquences de protéines (Karlsson, 2015).

Le protéotypage sans a priori utilise une base de données référençant l'ensemble des microorganismes séquencés. Ainsi l'assignation des peptides se fait par comparaison à la base de données où un spectre MS/MS est assigné à un peptide (Hayoun, 2020). Ainsi des associations vont être faites reliant les peptides aux données taxonomiques donnant ainsi une identification. L'utilisation de logiciels efficaces capables de reconnaître les peptides discriminants pour identifier les différents microorganismes est primordial. Selon les logiciels, l'analyse taxonomique se base sur diverses stratégies. En effet, l'utilisation unique de peptides spécifiques, de peptides partagés et spécifiques, de comptage spectral sont des informations utilisées par ces différents pipelines. Par exemple, Unipept est une application web développée pour l'analyse de données métaprotéomiques. Le protéotypage par Unipept repose sur l'association de chaque peptide identifié à un taxon via la base de donnée UniProtKB. La spécificité taxonomique des peptides tryptiques utilise l'approche de l'ancêtre commun le plus proche (Mesuere et al., 2016b, 2012). D'autres logiciels vont se baser sur l'utilisation des peptides spécifiques à un organisme. Proteoclade, est un outil utilisant les peptides spécifiques pour l'assignation des peptides à un organisme donc à l'identification. Cet outil récupère les lignées taxonomiques complètes de la base de données NCBI et s'interface directement avec l'API Uniprot pour le téléchargement et la concaténation des bases de données et crée ainsi une sous-base de données plus spécifique appelée Proteoclade Database (Mooradian et al., 2020). Cependant l'utilisation uniquement des peptides spécifiques peut avoir des limites

d'identification comme dans le cas d'organismes très proches comme *Escherichia* et *Shigella*. Ainsi des méthodes utilisent également l'information des peptides partagés en plus des peptides spécifiques pour l'identification. C'est le cas de l'outil MiCId (Alves et al., 2022, 2018) qui calcule des score de similarité (e-value) pour les peptides partagés en faisant d'abord un score sur les peptides identifiés puis une e-value unifiée pour les microorganismes identifiés. Récemment, cet outil propose une extension de l'outil pour l'identification de protéines impliquées dans l'antibio-résistance. D'autres outils comme PIPASIC vont utiliser une combinaison d'informations sur l'assignation des peptides à une base de données et le comptage spectral pour l'identification (Penzlin et al., 2014). TCUP est un outil permettant de donner la composition taxonomique mais peut également donner des informations sur l'antibio-résistance en utilisant l'attribution de tous les peptides et l'abondance relative. Cet outil va dans un premier temps attribuer les peptides aux protéines par un alignement des peptides à une base de données génomique puis réaliser un alignement des peptides sur une base de données d'antibio-résistance (Boulund et al., 2017). Enfin pour augmenter la sensibilité de l'analyse, des logiciels utilisent plusieurs étapes d'analyses visant à réduire la taille des bases de données à chaque étape. TaxIT et Proteome2pathogene utilisent deux étapes d'analyse pour l'identification de l'échantillon. En effet, TaxIT réalise une première analyse d'attribution des peptides permettant la sélection des espèces les plus pertinentes contenues dans l'échantillon puis réalise une seconde analyse sur une base de données réduite aux espèces sélectionnées à la première étape (Kuhring et al., 2020). Proteo2pathogene utilise une approche sur deux rangs taxonomiques différents. En effet, une première analyse est réalisée avec une identification au niveau genre puis une base de données réduite à l'ensemble des espèces du genre validé au tour précédent est utilisée pour analyser de nouveau les peptides. La phylopeptidomique développé au laboratoire Li2D (Pible et al., 2020) est également une méthode d'identification sans a priori des microorganismes prenant en compte les peptides partagés et spécifiques et dont l'analyse peut se dérouler en plusieurs étapes d'analyses pour augmenter la sensibilité de l'identification pouvant ainsi permettre des discriminations entre des souches très proches. Cette dernière méthode sera détaillée dans un chapitre ultérieur. De nombreux outils qui ne cessent de s'améliorer sont donc nécessaires à une bonne identification des microorganismes.

4.3. L'utilisation de la signature peptidique pour le protéotypage des microorganismes

Le protéotypage des microorganismes qui signifie littéralement typer par les protéines (Grenga et al., 2019) est un concept qui a été décrit par Karlsson en 2015 (Karlsson, 2015). La

protéomique se caractérisant par l'étude de l'expression des gènes donc des protéines permet un lien direct avec le phénotype des microorganismes étudiés. La caractérisation des microorganismes avec la classification et l'identification est un des aspects du protéotypage. Dans ce cadre, l'utilisation des données protéiques est uniquement faite dans le but de discriminer différents groupes d'organismes taxonomiques ou leurs fonctionnalités (Grenga et al., 2019). Le protéotypage par MALDI-TOF MS largement utilisé en routine pour l'identification de microrganismes présente certaines limites, notamment pour l'identification de souches non caractérisées. Le protéotypage par MALDI-TOF MS utilise une empreinte protéique basée sur les protéines abondantes et basiques, mais ne comprend en général que de l'ordre de 60 à 70 signaux (masses de protéines entières). L'utilisation d'une approche bottom-up par nLC-MS/MS permet une plus grande sensibilité et repose sur l'identification de milliers de peptides L'analyse par spectrométrie de masse en tandem par l'approche bottom-up est une alternative explorée par le laboratoire d'accueil de mes travaux de thèse. Ainsi une spécificité plus grande est obtenue avec le protéotypage MS/MS permettant une meilleure discrimination notamment au niveau d'espèces très similaires, voire de typer des sous-espèces. Par exemple, la résolution du protéotypage par nLC-MS/MS a permis de différencier taxonomiquement les espèces très proches du Mitis Group du genre *Streptococcus* (*pneumoniae, pseudopneumoniae et mitis)* avec un nombre de peptides uniques pour chaque espèce de plus de 200 peptides (Karlsson et al., 2018a) .

Le protéotypage par MS/MS permet une vision globale de l'activité des microorganismes à partir d'isolats microbiens ou d'échantillons complexes permettant l'élucidation des caractéristiques taxonomiques pour l'identification mais également des informations caractéristiques importantes sur la résistance aux antibiotiques ou encore sur l'utilisation de voies métaboliques de certaines espèces dont voici quelques exemples. Une étude menée par Hayoun *et al*, montre la réalisation d'un criblage sur les microorganismes contenus dans une piscine de stockage de combustible usagé par protéotypage pour pouvoir isoler de potentielles souches disposant d'un catalyseur enzymatique utilisable pour des applications biotechnologiques (Hayoun et al., 2020b). Cette étude est une application du protéotypage par MS/MS aux études environnementales qui sont plus difficiles avec le protéotypage par MALDI-TOF MS du fait de l'inexistence de base de données de référence par exemple. Le protéotypage par bottom-up a été appliqué pour l'identification d'espèces marines isolées de la côte méditerranéenne et a permis de caractériser pour la première fois à l'échelle protéomique un représentant du phylum Balneolaeota. Cette étude montre une nouvelle fois la capacité du protéotypage par MS/MS à caractériser rapidement les isolats même les plus atypiques (Lozano et al., 2022). Une application récente du protéotypage appelé « *Paleoproteotyping* » s'intéresse

à la caractérisation des microorganismes dans l'objectif de répondre à des questions d'ordre historique comme par exemple les conditions possibles de la mort de Pauline Jaricot, catholique française donc le cœur était conservé, ainsi que les conditions de conservation des reliques (Bourdin et al., 2023).

Le protéotypage peut se faire par l'utilisation de biomarqueurs peptidiques pour la détection rapide de microorganismes comme avec l'étude de la détection du virus Sars-CoV-2 en 3 minutes de mesure (Gouveia et al., 2020) ou encore avec la détection de pathogènes respiratoires (Karlsson et al., 2020). Dans cette dernière étude, la découverte de biomarqueurs chez des pathogènes responsables de maladies respiratoires comme *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Moraxella catarrhalis* ou encore *Staphylococcus aureus* ont permis d'identifier dans un contexte clinique sur des échantillons cliniques les différentes espèces. Cela montre la puissance de l'utilisation de la spectrométrie de masse et l'approche protéomique pour l'identification de pathogènes sans a priori.

Enfin des optimisations de méthodes de protéotypage permettent l'augmentation du débit des différentes étapes expérimentales du protéotypage notamment avec l'optimisation des méthodes de lyse et de digestion (Hayoun et al., 2020a, 2019) mais aussi l'évaluation des méthodes d'identification des microorganismes par protéotypage en développant des échantillons standards complexes contenant un mélange d'espèces microbiennes (Mappa et al., 2023a).

I.5. Les stratégies haut-débit utilisées pour l'identification rapide de microorganismes en protéomique bottom-up

L'identification des microorganismes doit être rapide non seulement pour répondre à des questions cliniques afin d'adapter les traitements mais également dans le cas de criblage de microorganismes pour le développement d'applications biotechnologiques. L'augmentation du débit est donc une priorité et peut se décliner à deux niveaux : l'augmentation du débit au niveau de la préparation des échantillons et au niveau de l'analyse nLC-MS/MS.

5.1. Au niveau de la préparation des échantillons

5.1.1. Amélioration des protocoles de préparation des échantillons et automatisation

La préparation des échantillons regroupe toutes les étapes de la culture à l'obtention de l'échantillon pour analyse MS : le digestat peptidique. A chacune des étapes que ce soit celle de culture, de lyse ou encore digestion, le débit peut être augmenté en passant par l'adaptation à des formats de manipulation pour le haut-débit, voire à de l'automatisation. Au niveau de la culture, des automates complets de laboratoire (TLA de l'anglais *total laboratory automatisation*) comme le WASPLAB commercialisé par Copan, le PreLUD (I2a) ou encore le BD Kiestra (BD Diagnostic) permettent d'automatiser les étapes d'ensemencement, les préparations des lames pour coloration, la traçabilité (système code barre) et permettent également une meilleure reproductibilité. D'autres automates spécialisés pour la culture liquide sont également utilisés et adaptés au haut-débit en utilisant un bras robotique central et/ou des plaque 96 puits (Kurokawa and Ying, 2017; Ross et al., 2004).

Pour les étapes de lyse et de digestion enzymatique une équipe allemande a développé un protocole rapide d'extraction des protéines puis de digestion appelé SPEED pour « Sample Preparation by Easy Extraction and Digestion » basé sur l'utilisation de l'acide trifluoroacétique (TFA) pour la lyse des cellules et l'obtention très rapide d'un lysat (Doellinger et al., 2020). Des techniques de digestion sont aussi améliorées pour réduire au maximum le temps et augmenter le débit. Le *Filter Aided Sample Preparation* (FASP), est une méthode de digestion qui repose sur l'utilisation d'une membrane filtrante permettant la rétention des protéines et c'est sur laquelle les étapes de réduction, alkylation et digestion sont faites. La technique a été améliorée en agissant sur la taille du seuil de coupure du filtre (fa-SPEED) permettant de réduire le temps des étapes de centrifugation. L'adaptation du format FASP aux plaques 96 puits a aussi permis d'augmenter le débit (Loroch et al., 2022; Potriquet et al., 2017). La méthode de digestion *Single-pot solid-phase-enhanced sample preparation* (SP3) est une méthode de digestion liquide qui est très efficace et s'est généralisée récemment et présente des rendements supérieurs aux méthodes de digestion en gel par exemple (Hayoun et al., 2020a). Elle repose sur l'utilisation de billes magnétiques qui sont fonctionnalisées de façon à capturer les protéines à leur surface. Suite à l'ajout de solvants et d'une enzyme protéolytique, les peptides vont être obtenus. Les adaptations de la méthode SP3 au format plaques 96 puits ont été décrites et permettent de réduire le temps de digestion. Des automates de manipulation de liquides du type de la plateforme Bravo (Agilent) permettent de réaliser les étapes de réduction, alkylation, digestion sans aucune intervention (Figure 17) (Hayoun et al., 2020a) (Müller et al., 2018). Une équipe a de plus rendu la méthode SP3 encore plus rapide sous le nom de USP3 en optimisant des paramètres variables comme la quantité de billes, la quantité de l'enzyme protéolytique (Dagley et al., 2019).

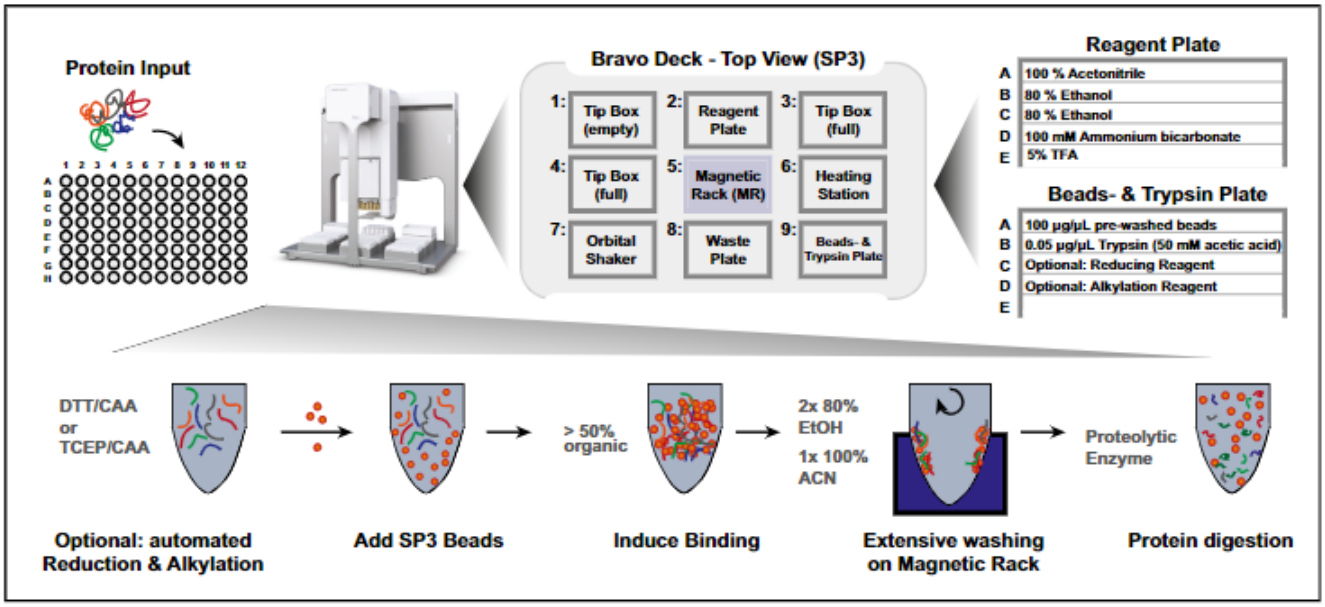


Figure 17 : Fonctionnement de l'automate Bravo de Agilent pour l'application à une digestion SP3 automatisée (tirée de Müller et al., 2018)

La méthode de digestion SP3 peut être automatisé en utilisant notamment l'automate Bravo développé par Agilent. La digestion se déroule avec un format microplaque 96 puits. A l'intérieur de cet automate, les solvants et le matériel utile à la réalisation de la digestion sont contenus dans différentes cases. La méthode SP3 est donc réalisé de manière automatisé avec une prmière étape optionnelle de réduction et d'alkykation par l'ajout de DTT /CAA ou TCEP/CAA puis les billes SP3 sont ajoutés à l'échantillon puis différentes étapes d'ajout de solvants sont réalisées avant la rétention des billes sur l'aimant magnétique pour éliminer le surnageant. Après l'ajout de l'enzyme protéolytique et des étapes d'élimination du solvant, les peptides sont obtenus.

Par ailleurs l'outil *Simple and integrated spintip-based protein digestion and three-dimensional peptide fractionation technology* (SISPROT) regroupe deux technologies de séparation, des billes *Strong Cation Exchange* (SCX) et une membrane C18 dans un cône de pipette. Cet outil intègre les étapes de réduction, alkylation, digestion et fractionnement (Chen et al., 2016). La même équipe a ensuite développé 3D-SISPROT qui combine trois types de séparation, l'utilisation de la séparation *Strong Anion Exchange* (SAX) est faite pour la digestion et pour le fractionnement puis l'utilisation de la séparation par phase inverse à pH basique permet le fractionnement (Chen et al., 2017). Enfin la technologie *improved Sample Technology* (iST) développée par Preomics permet également de réaliser les étapes de la lyse à la purification des échantillons dans un automate avec l'utilisation de leur kit iST (Kulak et al., 2014).

Toutes ces améliorations et automatisation de méthode de préparations des échantillons sont primordiales pour l'augmentation du débit.

5.1.2. Analyse d'échantillons en simultanée par marquage chimique

Le multiplexage en protéomique est une technique qui a été introduite en 1999 (Suarez et al., 2015). L'introduction de variants isotopiques (léger/lourd) au niveau des protéines ou peptides permet le mélange de plusieurs échantillons et par conséquent une analyse MS unique. En effet, la modification chimique ou métabolique sur les protéines ou peptides permet une identification des échantillons en fonction du variant isobarique introduit, cela peut être comparé à un système à code-barres (Arul and Robinson, 2019). La différence de masse en fonction des variants isotopiques va permettre d'identifier l'échantillon. Les stratégies de marquage peuvent reposer sur plusieurs principes : chimique, enzymatique ou métabolique. La méthode chimique est celle qui est le plus souvent utilisée du fait de sa spécificité à se lier à des résidus spécifiques ou terminaux des peptides et l'efficacité des réactions de marquage. De plus, elle ne nuit pas aux propriétés physicochimiques des peptides.

Le marquage métabolique, une des approches les plus connues est la méthode *stable isotope labelling by amino acids in cell culture* (SILAC). Cette méthode repose sur l'incorporation métabolique des acides aminées avec des isotopes lourds et légers dans les protéines durant l'étape de culture. Les cellules sont marquées de manière différentielle et c'est lors de l'analyse par spectrométrie de masse que l'observation de différence de masse sera faite (Chen et al., 2015; Emadali and Gallagher-Gambarelli, 2009).

Le marquage chimique peut reposer sur l'utilisation de marqueurs isobares qui se fixent aux groupements amino ou thiol libres des peptides protéolytiques, sans créer de différence de masse dans le cas de marqueurs isobares. Dans ce dernier cas les étiquettes sont le plus souvent formées de 3 groupes. Le premier groupe dit groupe rapporteur est composé du marqueur isobare qui aura une masse variable lors de l'analyse en spectrométrie de masse, le second groupe, équilibreur de masse, assure l'équilibre des masses et garantit que le poids moléculaire global reste le même avec les différents ions rapporteurs et le troisième groupe est le groupe réactif qui va lier l'étiquette avec les peptides. Le marquage *Tandem Mass Tag* (TMT) (Thompson et al., 2003) est l'approche de marquage la plus utilisée et est commercialisée sous forme de kits par Thermo Fisher. Cette approche de multiplexage peut désormais analyser jusqu'à 16 échantillons en simultanée avec le kit TMT 16 pro reagents (Li et al., 2020) (Figure 18.A). Le marquage par *isobaric Tags for Relative and Absolute Quantification* (iTRAQ) (Ross et al., 2004) permet l'analyse de 4 à 8 échantillons en même temps (Figure 18.B). Les différents réactifs d'un type de marquage utilisent le même principe de fixation mais diffèrent dans la distribution et l'emplacement des isobares dans l'étiquette.

Ces méthodes sont souvent utilisées pour comparer des protéomes obtenus dans des conditions différentes mais peu pour augmenter le débit d'identification de microorganismes du fait du coût des marqueurs.

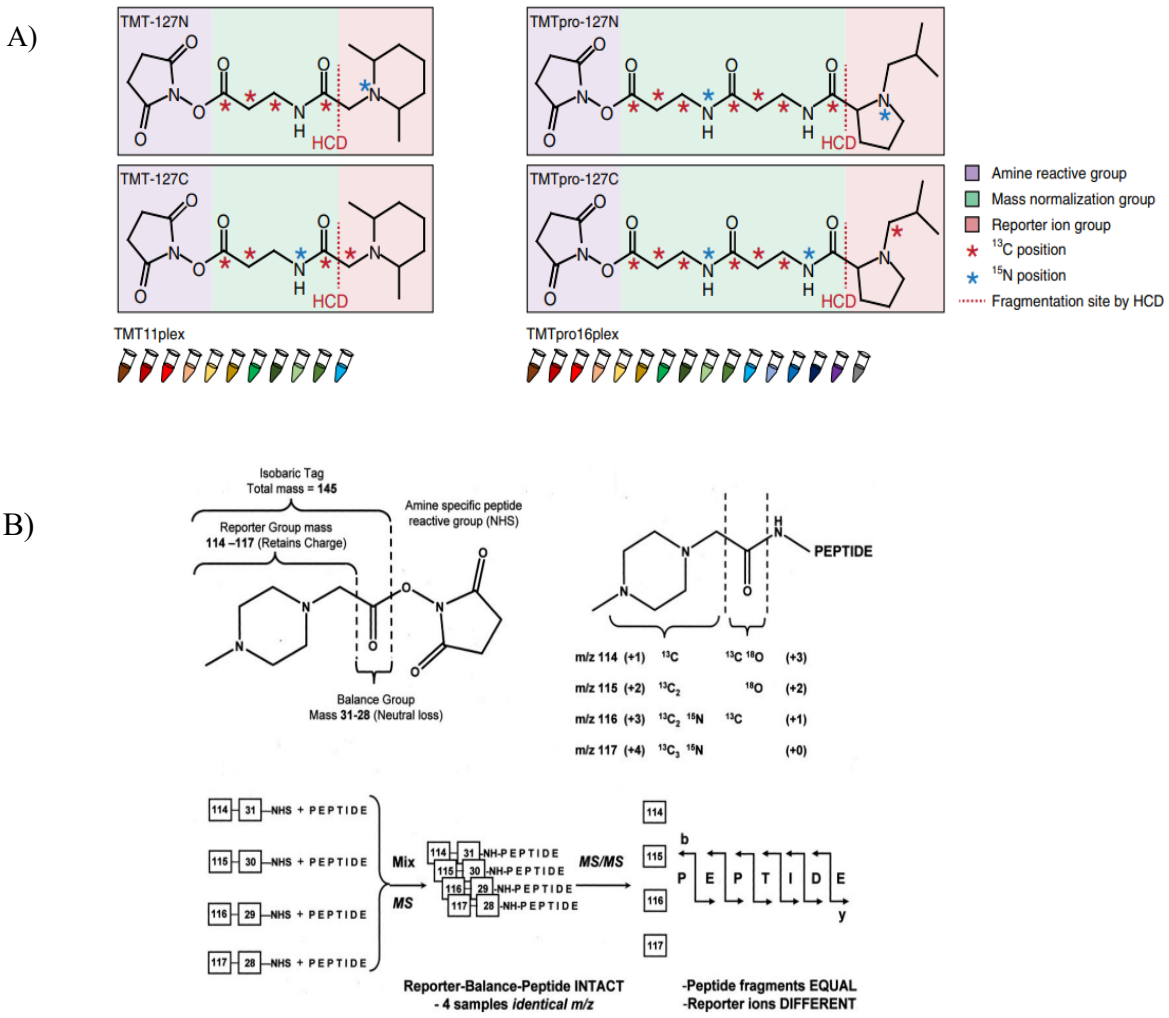


Figure 18 : Réactifs pour les stratégies de marquage par TMT (A) et iTRAQ (B) (tirée de Ross et al., 2004)

Le marquage chimique repose sur l'utilisation de marqueurs isobares qui se fixent aux groupements amino ou thiol libres des peptides protéolytiques. Ces étiquettes sont formées de 3 groupes. Le premier groupe dit groupe rapporteur est composé du marqueur isobare qui aura une masse variable lors de l'analyse en spectrométrie de masse, le second groupe, équilibreur de masse, assure l'équilibre des masses et garantit que le poids moléculaire global reste le même avec les différents ions rapporteurs et le troisième groupe est le groupe réactif qui va lier l'étiquette avec les peptides. (A) Le marquage TMT peut permettre l'analyse de 16 échantillons en parralèle. (B) Le marquage iTRAQ permet l'analyse de 4 à 8 échantillons en parrallèle.

5.2. Au niveau de l'analyse nLC-MS/MS

5.2.1. Les méthodes de multiplexage en protéomique ciblée

Le multiplexage peut également être réalisé au cours de l'analyse par spectrométrie de masse. En effet deux modes d'analyses sont possibles : le mode *Selected Reaction Monitoring* / *Multiple Reaction monitoring* (SRM/MRM) ainsi que le mode *Parallel Reaction Monitoring* (PRM). Pour la méthode SRM le premier analyseur quadripôle (Q1) sélectionne par son *m/z* le peptide précurseur à fragmenter dans le deuxième quadripôle servant de cellule de collision. Le troisième quadripôle (Q3) sélectionne un ion fragment déterminé. On parle alors de transition père→fils. La MRM enregistre plusieurs transitions père→fils pour un même précurseur pour augmenter la fiabilité de l'interprétation. Pour la méthode PRM l'ion précurseur sélectionné dans le quadrupôle est fragmenté dans la cellule de collision HCD après transfert vers la C-trap pour accumulation d'ions. Les fragments fils produits sont ensuite transférés dans la C-Trap pour être refocalisés avant d'être éjectés et détectés dans l'Orbitrap. La SRM peut être réalisée sur un spectromètre de masse à basse résolution de type triple quadripôle alors que la PRM nécessite un instrument de haute précision/haute résolution de type quadrupôle-Orbitrap tel que par exemple l'Orbitrap-Q-exactive HF (Elschenbroich and Kislinger, 2011; Saleh et al., 2019).

En sélectionnant un ou plusieurs peptides définis, le temps d'analyse est donc raccourci et l'identification est donc beaucoup plus rapide. C'est ce qui a d'ailleurs été démontré par une étude sur le Sars-CoV2, où des séquences peptidiques ont été ciblées comme étant biomarqueurs du virus et ont permis une identification du virus dans des échantillons de salive en 3 minutes de MS/MS (Gouveia et al., 2020). Les deux modes d'analyse sont schématisés Figure 19.

L'analyse de type MRM ou PRM peut être précédée de l'immunocapture des molécules cibles par le biais d'anticorps. La spectrométrie de masse est ensuite utilisée pour l'analyse rapide de biomarqueurs (Chenau et al., 2011; Dupré et al., 2021).

Ces méthodes sont forcément ciblées et ne s'appliquent pas à une démarche sans a priori pour l'identification de microorganismes sans connaissances préalables.

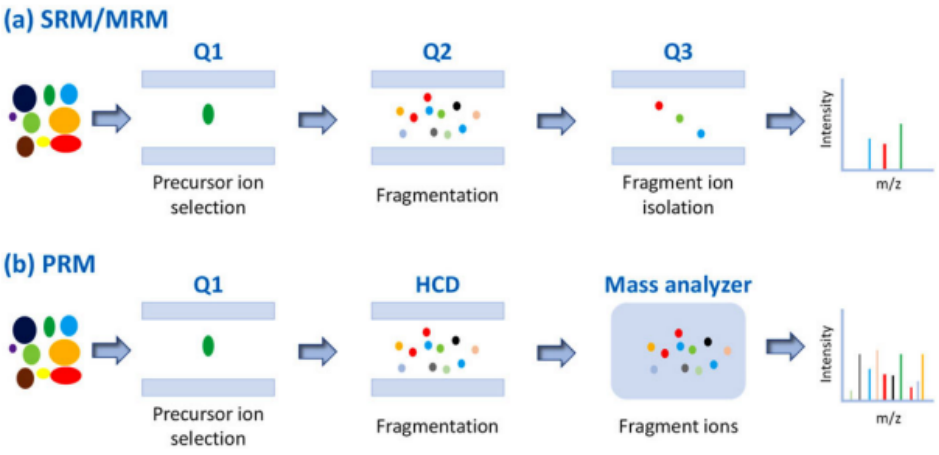


Figure 19 : Représentation schématique d'analyse ciblée par stratégie de multiplexage au niveau de l'analyse MS/MS : stratégie SRM/MRM (A) et PRM (B) (tirée de Saleh et al., 2019)

Pour la méthode SRM le premier analyseur quadrupôle (Q1) sélectionne par son *m/z* le peptide précurseur à fragmenter dans le deuxième quadrupôle servant de cellule de collision. Le troisième quadrupôle (Q3) sélectionne un ion fragment déterminé. On alors parle de transition père→fils. La MRM, appelé multiplexe SRM, est une variante de la SRM pour laquelle il est possible d'alterner la détection des transitions pour augmenter le nombre de peptides détectés. Pour la méthode PRM l'ion précurseur sélectionné dans le quadrupôle est fragmenté dans la cellule de collision HCD. Les fragments fils sont ensuite transférés dans la C-Trap pour être refocalisés avant d'être éjectés et détectés dans l'Orbitrap.

5.2.2. Diminution du temps d'analyse

L'augmentation du débit passe également par l'optimisation de méthodes pour diminuer au maximum le temps des différentes étapes de l'analyse nLC-MS/MS notamment l'étape de chromatographie liquide.

i) L'optimisation du temps de séparation des échantillons par chromatographie liquide en amont de l'analyse par spectrométrie de masse

De multiples méthodes intervenant au niveau de la séparation de l'échantillon traditionnellement réalisée par la nLC sont développées. Evosep™ est une technologie de chromatographie permettant de réduire le temps de cette étape. Cette technique reposant sur l'utilisation d'un EvoTip où l'échantillon est retenu sur la membrane C18 et va être élué rapidement par l'utilisation de 4 pompes basses pression puis une pompe haute-pression couplée à un système de valves rotatives améliore le débit de l'étape de séparation en amont de l'analyse par le spectromètre de masse (Bache et al., 2018a). Des évolutions ont été faites sur l'Evosep permettant l'analyse de protéomes en 8 min et ont également démontré que

l'utilisation du marquage TMT couplé à l'Evosep One augmentait le débit (Krieger et al., 2019). Thermo Scientific a également développé un système de chromatographie liquide appelée Vanquish Neo qui peut fonctionner à un débit du nano jusqu'à 100 µL/min et à une forte pression permettant ainsi de pouvoir analyser des protéomes en 8 min (Zheng, 2021). Zip Chip™ est également une méthode de séparation reposant sur la séparation des analytes par leur charge et leur masse. Cette technologie intègre l'électrophorèse capillaire (CE) et l'ionisation par electrospray (ESI) dans un seul dispositif microfluidique pour préparer et séparer rapidement des échantillons biologiques et les injecter directement dans le spectromètre de masse. Elle permet d'analyser un échantillon en 15 minutes (Rinas et al., 2019). Des méthodes basées sur l'utilisation de double colonnes en parallèle se sont développées. Une colonne charge l'échantillon pendant que l'autre colonne réalise un lavage et se prépare pour analyser le prochain échantillon et vis-versa permettant ainsi de réduire par deux le temps d'analyse des échantillons (Hayoun et al., 2020a; Hosp et al., 2015; van der Laan et al., 2020).

ii) Utilisation de l'infusion directe

Récemment, des travaux portant sur l'étude du protéome par infusion directe ont été menés. En effet, une étude a permis par le couplage de l'infusion directe et de la source FAIMS, source de mobilité ionique, (Thermo Fisher) ainsi qu'un logiciel d'identification CsoDIAq de pouvoir identifier 2000 protéines du protéome Hela à partir de 1 µg d'échantillon en 4.5 minutes (incluant l'analyse de l'échantillon et l'acquisition des données) (Jiang et al., 2022). Une autre étude a permis de démontrer l'analyse rapide d'un protéome avec une quantification de plus de 500 protéines en quelques minutes (environ 3.5 protéines par secondes) par Infusion Directe – *Shotgun Proteome Analysis* (DI-MS) et une acquisition de données DIA. L'utilisation de la source FAIMS a permis de tripler le nombre d'identification de peptides (Meyer et al., 2020).

iii) L'évolution des appareils de masse

L'évolution des appareils de masse contribue également à l'augmentation de la rapidité et de l'exhaustivité des analyses. En effet, en agissant sur certains paramètres de masse comme la vitesse de balayage, on arrive à avoir plus d'informations en moins de temps. C'est le cas de l'Orbitrap Exploris 480 commercialisé par Thermo Fisher qui a une résolution maximale de 480 000, une plage de *m/z* de 40-6000 ainsi qu'une vitesse de balayage pouvant atteindre 40 Hz contre une résolution de 240 000, une plage de *m/z* de 50 - 6000 pour le Q-exactive HF (Thermo Fisher). Grâce à la vitesse de balayage améliorée de l'Exploris 480, le même nombre

d'informations que sur un appareil type Q exactive HF peut être obtenu en deux fois moins de temps. Des éléments ajoutés en amont de l'appareil de masse comme la source FAIMS peuvent améliorer la sensibilité de l'analyse en filtrant certains ions (Bekker-Jensen et al., 2020). En juin 2023, Thermo Fisher a présenté un nouveau spectromètre de masse l'Orbitrap-Astral présentant une résolution en masse de plus de 480 000 et une vitesse d'acquisition des spectres MS/MS supérieur à 200 Hz. De même, Brucker propose un instrument (Tims-TOF ULTRA) avec une vitesse d'acquisition des spectres MS/MS de 300 Hz. Ces performances techniques permettent de réduire considérablement le temps des analyses en ayant des informations similaires aux deux autres appareils cités ci-dessus avec des gradients de 8 minutes au lieu de 30 min à 60 min, et donc un énorme gain en débit.

I-6. Contexte et objectifs de la thèse

6.1. La détection des microorganismes au Li2D

Le Laboratoire d'Innovations technologiques pour la Détection et le Diagnostic (Li2D) a pour mission le développement de méthodologies et technologies pour la détection des agents pathogènes ou toxiques présents dans l'environnement, dans des échantillons ou tissus biologiques, et la découverte et la validation de biomarqueurs pour le diagnostic. Les méthodes proposées sont basées sur de la détection immunologique (format bandelette ou ELISA), génétique (tests qPCR) et par spectrométrie de masse en tandem. La détection des microorganismes intéresse les différents domaines, clinique, environnemental et est aussi en lien avec la lutte contre le bioterrorisme. La plateforme ProGénoMix du Li2D est une palteforme labellisée par le réseau GIS IBISA (Infrastructures et Biologie Santé et Agronomie, https://www.ibisa.net/) et est équipé de 5 spectromètres de masse de haute résolution permettant de mener à bien des prestations d'analyses en protéomique, protéogénomique et métaprotéomique, ainsi que ses propres projets de recherche et développement. Beaucoup de projets portent sur la détection de microorganismes, notamment avec des études sur la composition et la dynamique de microbiote basées sur des données de métaprotéomique, des développements technologiques au niveau bioinformatique (traitements, intégration de données) et également sur des développements de méthodologies et leurs applications pour la détection de microorganismes par spectrométrie en tandem. C'est dans ce cadre que s'inscrit ma thèse qui a pour but le développement de méthodes de protéotypage haut-débit de microorganismes par spectrométrie de masse en tandem.

6.2. La phylopeptidomique

6.2.1. Principe

Le protéotypage par MS/MS passe obligatoirement par l'interprétation des données par des logiciels de bioinformatique permettant l'identification des microorganismes. Dans les paragraphes précédents (paragraphe I-4), plusieurs logiciels ou outils d'identification ont été présentés. Certains se basent sur les peptides spécifiques à un organisme, d'autres prennent en compte les peptides partagés et spécifiques, d'autres se focalisent sur les activités antibiotiques. Le nombre de séquençage des génomes augmente considérablement de par la facilité et la réduction des coûts du séquençage du gène de l'ARNr 16S et du séquençage du génome entier augmentant ainsi le nombre de séquences dans les bases de données. Cependant, la conséquence de l'augmentation des bases de données se ressent sur certaines espèces qui partagent une très grosse partie de leur séquence génomique comme certaines espèces de *Bacillus* et qui présentent donc un nombre de peptides discriminants/ spécifiques qui diminue considérablement au fil de l'apport de nouveaux génomes (Mappa et al., 2023b). La phylopeptidomique est une méthodologie métaprotéomique conçue au Li2D, qui par l'interprétation informatique des données relie les protéines aux organismes qui les produisent. C'est une approche sans a priori, de protéotypage des microorganismes qui en plus d'être sensible et fiable permet la quantification, et est basée sur les peptides spécifiques et les peptides communs, donc est immun à l'augmentation des bases de données de séquences de génomes.

La phylopeptidomique est basée sur une approche protéomique de type bottom-up pour laquelle les protéines extraites sont digérées par l'enzyme trypsine. Le pool de peptides appelé digestat peptidique est ensuite analysé par spectrométrie de masse en tandem. Les données de spectrométrie de masse sont interprétées à l'aide d'un ensemble de scripts et infrastructure de bases de données développés au laboratoire, µorgID (brevet WO2015019245A1). Cette infrastructure utilise le moteur de recherche Mascot Daemon pour l'interprétation des spectres MS/MS et donc la détermination des PSMs. La procédure de phylopeptidomique sert à identifier les taxonomies en associant à chaque spectre MS/MS les données taxonomiques associées aux différentes séquences peptidiques qui y sont liées (TSMs : Taxa-Spectrum Matches) et la biomasse relative des microorganismes présents.

La proposition d'identification donnée par le pipeline µorgID repose sur l'utilisation d'un modèle mathématique, appelé signature phylopeptidomique, prenant en compte l'ensemble des données TSMs et des peptides spécifiques attribués à chaque rang taxonomique pour un

organisme. Ainsi pour chaque organisme, une signature phylopeptidomique décrit la proportion de peptides partagés avec les autres organismes et leur phylogénie respective (Pible et al., 2020). De plus la phylopeptidomique permet d'estimer avec précision l'abondance relative des microorganismes contenus dans un mélange, et peut en principe s'appliquer dans un contexte de métaprotéomique. En effet, dans l'étude menée par Pible *et al*, un mélange artificiel contenant 2 bactéries proches (*Shighella flexneri* et *Salmonella bongori*) a été analysé. La Figure 20 illustre de façon claire la différence de signature dans le cas d'un microorganisme seul ou en présence d'un mélange.

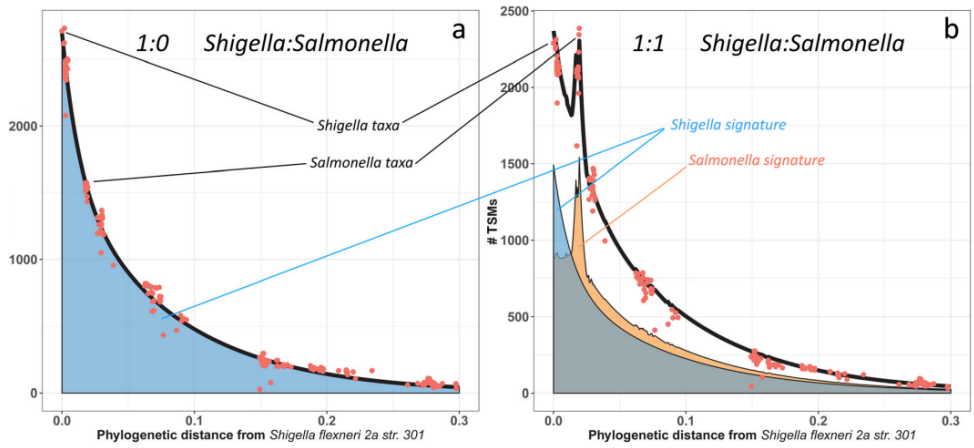


Figure 20 : Signature phylopeptidomique de deux échantillons comprenant dans un cas uniquement l'espèce *Shigella flexneri* (A) et dans le second cas un mélange de *Shigella flexneri* et *Salmonella bongori* (B) (tirée de Pible et al., 2020)

La signature phylopeptidomique permet une meilleure estimation de la biomasse que l'utilisation de peptides spécifiques uniquement, des TSMs seuls ou encore du comptage spectral comme présenté dans la Figure 21. Dans cette figure, la différence d'estimation de la biomasse en fonction des paramètres d'identification est représentée.

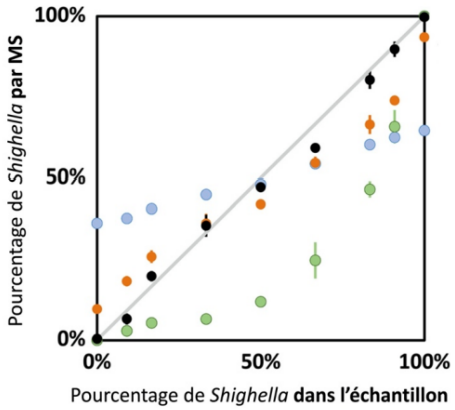


Figure 21 : Estimation de la biomasse de *Shigella flexneri* avec différents moyens de quantification

L'estimation de la biomasse *Shigella flexneri* est réalisée avec différents moyens de quantification : l'utilisation des TSMs uniquement (points bleus), des peptides spécifiques uniques (points verts), du comptage spectral (points oranges) et de la signature phylopeptidomique (points noirs) (tirée de Pible et al., 2020)

6.2.2. Applications de la phylopeptidomique au protéotypage de microorganismes

Les travaux de Mappa *et al.* sur le mélange complexe Mix 24 constitué d'un mélange de bactéries environnementales et pathogènes provenant de 20 genres différents appartenant à 5 phyla bactériens et incluant des espèces proches, ont démontré la capacité d'identification très résolutive de la phylopeptidomique (Mappa et al., 2023b). La méthodologie s'est montrée aussi performante que ce soit sur des isolats cliniques microbiens (Gouveia et al., 2020; Grenga et al., 2019) ou sur des isolats environnementaux (Hayoun et al., 2020b; Lozano et al., 2022). Elle peut aussi s'appliquer au domaine de la métaprotéomique (Armengaud, 2023; Grenga et al., 2022; Petit et al., 2020) qui a pour but d'analyser la dynamique du microbiote ou la composition taxonomique microbienne au sein d'échantillons cliniques type sputum de patients atteints de la mucoviscidose (Hardouin et al., 2022) ou encore sur des échantillons environnementaux types carottes de sol (Jouffret et al., 2021).

La phylopeptidomique est une méthode puissante permettant une identification taxonomique ultra-rapide à partir des informations peptidiques obtenus sur des spectromètres de masse de haute résolution. L'approche de protéomique bottom-up inclut une étape de chromatographie liquide pour décomplexifier le pool très complexe des peptides à analyser. Dans le contexte de criblages d'isolats de microorganismes des améliorations techniques doivent être apportées pour réduire le temps d'analyse nLC-MS/MS.

6.2.3. Les objectifs de la thèse

Le projet de thèse, cofinancé par la région Occitanie et le CEA, a été réalisé au laboratoire Li2D (Direction de la Recherche Fondamentale, Institut Joliot, Département Médicaments et Technologies pour la Santé (UMR0496), SPI) au Centre CEA de Marcoule. Il s'inscrit dans le développement de la phylopeptidomique au laboratoire afin de rendre cette technologie accessible au diagnostic clinique et environnemental. L'objectif est de tester de nouveaux concepts pour l'approche de phylopeptidomique pour l'identification haut-débit de microorganismes par protéotypage par spectrométrie de masse en tandem.

Les travaux antérieurs au laboratoire (travaux de Thèse de Karim Hayoun) ont permis d'optimiser la préparation des échantillons pour l'analyse par spectrométrie de masse afin de réduire le temps de ces étapes. Les conditions optimisées de lyse des microorganismes sont applicables avec une bonne efficacité à tous les types de microorganismes, les bactéries Gram négatives, Gram positives, mais aussi les champignons et les levures (Hayoun et al., 2019). La digestion SP3 utilisant des billes magnétiques et réalisée sur plaque 96 puits est adaptée au haut-débit (Hayoun et al., 2020a). L'utilisation d'un système dual de chromatographie permet d'éliminer le temps de rééquilibration de la colonne chromatographique réduisant ainsi le temps d'analyse nLC-MS/MS. Augmenter significativement le débit d'analyse nLC-MS/MS nécessite de réduire drastiquement le temps d'analyse nanoLC-MS/MS alloué à chaque échantillon, notamment pour diminuer soit le nombre d'analyses de spectrométrie de masse soit le temps de chaque analyse pour avoir moins de temps d'occupation de l'appareil de spectrométrie de masse.

Dans ce manuscrit décrivant et discutant les travaux de thèse, plusieurs méthodes pour l'identification haut-débit d'isolats sont présentées sous la forme de manuscrits scientifiques publiés ou soumis à publication. Dans le premier chapitre de résultats, la preuve de concept d'une méthode de multiplexage sans marquage pour l'identification haut-débit d'isolats de microorganismes par phylopeptidomique est présentée. Le principe qui repose sur la juxtaposition de signaux MS/MS appartenant à chaque isolat est mis en place par le fractionnement des peptides par chromatographie sur phase inverse sur HPLC. Elle est basée sur le fractionnement pour l'analyse de 21 organismes en une analyse unique de spectrométrie de masse en tandem sur un gradient de 60 minutes. Dans le second chapitre de résultats, la méthode de multiplexage est optimisée en terme de chromatographie pour la rendre plus facile d'utilisation. De plus, la robustesse de la méthode est démontrée par l'analyse d'un grand nombre de mélanges de 6 isolats microbiens. Enfin dans le troisième chapitre de résultats, un

concept de protéotypage flash d'isolats est présenté sous la forme d'un manuscrit « *technical brief* ». Ce protéotypage flash est basé sur le spectromètre de masse Orbitrap Exploris 480 utilisé en mode infusion directe permettant de s'affranchir de la chromatographie liquide et permet une réduction drastique du temps d'analyse de spectrométrie de masse en tandem par échantillon et aussi une réduction des coûts. L'ensemble de ces travaux permet de faire progresser le domaine du protéotypage des microorganismes en offrant la possibilité d'un débit d'analyse d'isolats très élevé.

II. Résultats

II-1 : Développement d'une méthode de multiplexage sans marquage pour le protéotypage d'isolats microbiens

L'identification de microorganismes est cruciale que ce soit pour répondre à des questions de diagnostic clinique, pour l'exploration de la diversité microbienne ou encore pour le criblage de microorganismes à des fins d'applications biotechnologiques. Les techniques moléculaires se sont beaucoup développées du fait de leur capacité d'analyse à haut-débit, de leur précision et surtout de l'affranchissement de l'étape de culture. Pourtant cette dernière étape reste primordiale non seulement pour enrichir les banques de souches mais aussi pour disposer des souches en vue de les étudier et pouvoir découvrir des propriétés d'intérêt. La culturomique a été développée pour favoriser la croissance des souches difficiles à cultiver. Cette méthode produit un très grand nombre d'isolats qui doivent être caractérisés et identifier. Le MALDI-TOF MS est la méthode de référence pour ce type de criblage du fait de sa rapidité et de son faible coût. Cependant cette méthode présente des limites notamment l'incapacité d'identifier des souches non présentes dans la base de données des spectres de référence. Le protéotypage par nanoLC-MS/MS peut palier aux limitations du protéotypage par MALDI-TOF MS. La phylopeptidomique développée au laboratoire qui permet le protéotypage de microorganismes sans a priori est sensible mais coûteuse en temps de spectrométrie de masse pour un criblage de microorganismes. Le développement de méthodes d'identifications rapides, sensible et pouvant limiter les coûts est primordial pour aller vers le haut-débit.

Une méthode permettant l'augmentation du débit des analyses de phylopeptidomique est proposée. Elle a pour concept l'analyse de plusieurs isolats en une analyse unique de spectrométrie de masse appelée multiplexage sans marquage. Cette méthode repose sur l'ajout d'une étape de fractionnement par HPLC en amont de l'analyse nanoLC-MS/MS. Ainsi, après la lyse de chaque isolat, la digestion enzymatique de l'extrait protéique est réalisée, les peptides sont fractionnés en phase inverse et de multiples fractions d'hydrophobicité différentes sont obtenues pour chaque isolat. Le concept est d'associer un isolat à une fraction d'hydrophobicité spécifique pour pouvoir les ré-associer par la suite lors de la séparation par chromatographie liquide couplée à la MS/MS. A travers la mise en place d'une stratégie de traitements des données de phylopeptidomique adaptée au multiplexage menant à la définition d'un index baptisé « SPi » pour *Species Proteotyping index*, nous sommes capables d'identifier 21 organismes en une analyse unique de 60 minutes de spectrométrie de masse en tandem. La preuve de concept de la méthode a été démontrée sur un mélange composé de bactéries Gram positives et Gram négatives comprenant aussi des organismes du même genre et aussi un même organisme dans plusieurs fractions. La méthode a montré qu'elle est suffisamment sensible pour

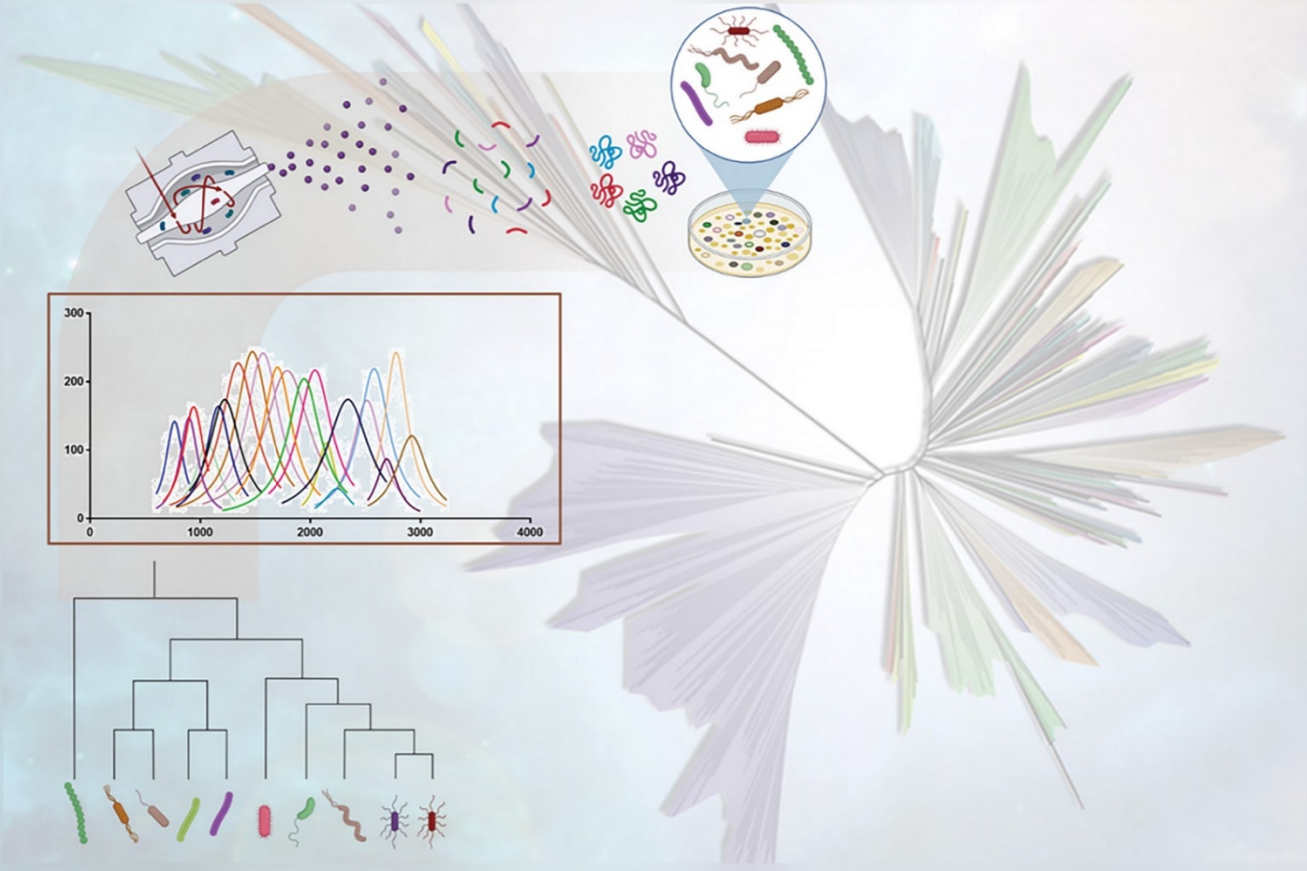
distinguer au sein du mélange deux organismes du même genre qu'il s'agisse de *Pseudomonas*, de *Bacillus* et de *Deinococcus*. *Ralstonia pickettii* qui était un isolat introduit par deux fractions d'hydrophobicité différentes dans le mélange a été parfaitement identifié dans les deux fractions de manière distincte.

Cette méthode permet de multiplexer l'analyse d'échantillons sans recourir à des réactifs chimiques. Cette approche rapide et performante de protéotypage par MS/MS permet de diminuer de façon importante les coûts d'analyse MS/MS pour chaque échantillon par une forte réduction du nombre d'analyses. Une demande de brevet a été déposée par le CEA pour cette méthode innovante en Avril 2023 (Demande n° EP23305557.3). Puis, ces travaux ont été publiés par le journal « Analytical chemistry » (https://doi.org/10.1021/acs.analchem.3c01975). Une page de couverture a également été publiée dans le journal Analytical Chemistry illustrant le principe de la méthode et mettant en valeur l'originalité de notre concept.

pubs.acs.org/ac

analytical chemistry

September 5, 2023 Volume 95 Number 35



ACS Publications
Most Trusted. Most Cited. Most Read.

www.acs.org

Label-free multiplex proteotyping of microbial isolates

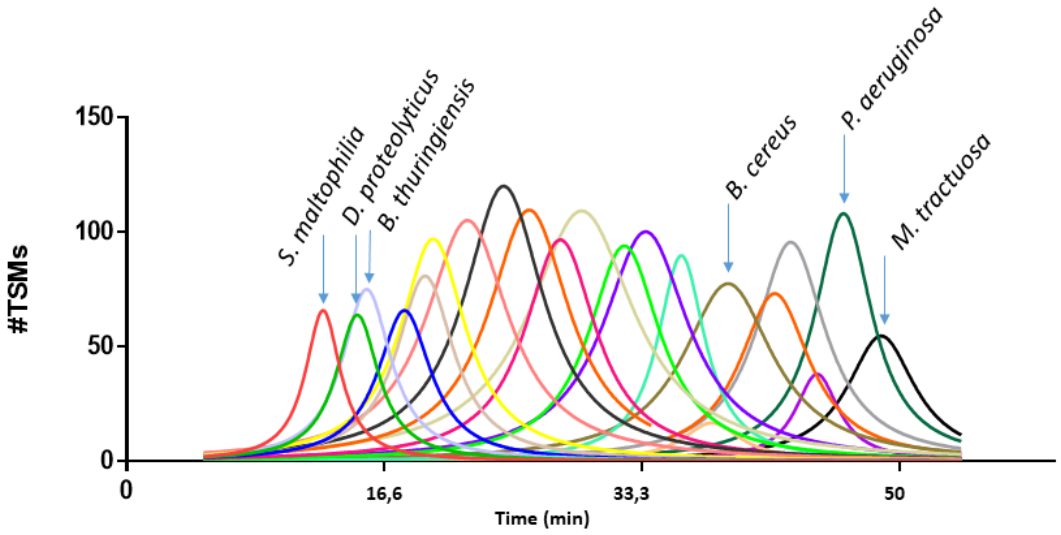
Madisson Chabas[1,2], Olivier Pible[1], Jean Armengaud[1*#], Béatrice Alpha-Bazin[1*#]

[1]Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, 30200 Bagnols-sur-Cèze, France .[2]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Université de Montpellier, F-30207 Bagnols sur Cèze, France.

[#]These co-authors should be considered as co-last authors.

*Authors to whom correspondence should be addressed.

Keywords: proteotyping, tandem mass spectrometry, proteomics, bacterial isolates, concatenation, taxonomy, multiplex



Abstract

To meet clinical diagnostic needs and for general microbiological screening, it is essential to be able to accurately and rapidly identify any microorganisms from complex microbiota. To gain insights into the individual components of microbiota, culturomics has been proposed as a means to systematically test hundreds of possible cultivation conditions and generate numerous microbial isolates with very distinct characteristics. High-throughput identification methods must now be developed to quickly screen these isolates. Currently, most multiplexing methods involve labeling, which comes at a cost. In this article, we present an innovative label-free multiplexing method for the identification of microorganisms using tandem mass spectrometry. The method is based on off-line reverse-phase fractionation of individual peptidomes. Multiplexing is achieved by mixing fractions of staged hydrophobicity, thus each sample is mapped to specific elution times. In this proof-of-concept study, multiplexed samples were analyzed by tandem mass spectrometry in a single run, and microorganisms present in the mixture were resolved by phylopeptidomics proteotyping. Using this methodology, up to 21 microorganisms could be identified in a single 60-min run performed with a Q-Exactive HF high-resolution mass spectrometer, resulting in a rate of one microorganism identified per 3 min of mass spectrometry, without any need for the use of labeling reagents. This approach opens new perspectives for the application of high-throughput proteotyping of bacteria using tandem mass spectrometry in large culturomics projects.

Introduction

For clinical diagnostics, and to explore microbial diversity, understand the environment, and screen new catalysts for biotechnological purposes, it is crucial to be able to accurately and rapidly identify microorganisms in the form of isolates. Whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) is well-established as a means to achieve this type of identification based on polypeptide fingerprints (Suarez, 2013). This rapid proteotyping method is suited for known clinical pathogens, but its performance declines when isolates are not closely related to those listed in the experimentally-established profiles. Tandem mass spectrometry-based proteotyping is a valuable complementary methodology to quickly classify atypical isolates, and has been shown to be applicable even with mixtures of microorganisms (Hayoun et al., 2020b; Pible et al., 2020). Massive identification of isolates, especially environmental microorganisms, is important if we wish to better characterize microbiomes. However, the cost and speed of the methodology used to identify isolates must be optimized for successful applications.

Classical tandem mass spectrometry-based proteotyping involves four steps: i) extraction of proteins from the biological material, ii) their proteolysis with trypsin to generate peptides, iii) identification of the peptides by high-resolution tandem mass spectrometry coupled to reverse-phase chromatography, and iv) interpretation of the peptide list to taxonomically identify the species. As it is based on low molecular-weight peptides rather than whole proteins, this methodology is much more sensitive than whole-cell MALDI-TOF MS (Mappa et al., 2023a). Furthermore, it provides very precise taxonomical identification through the identification of thousands of peptide sequences. Several approaches have been proposed for the interpretation of the list of peptides identified. Unipept-based proteotyping associates each peptide identified with a taxon via the UniProtKB database based on the lowest common ancestor approach (Mesuere et al., 2012). Alternatively, TCUP uses two analysis steps to identify the sample, the first step is attribution of the spectra to peptides / proteins and then to a taxon; then peptides are searched against a database of resistance genes (Boulund et al., 2017). A third approach, based on ProteoClade software, uses specific peptides to assign peptides to an organism (Mooradian et al., 2020). Like TCUP, TaxIT also applies a two-step analysis, with initial selection of the most relevant species contained in the sample, followed by searches of a reduced database based on the species selected to allow precise identification (Kuhring et al., 2020).

Recently, phylopeptidomics-based proteotyping was proposed as a means to take advantage of taxon-specific peptides and peptides shared by closely- and distantly-related organisms at all possible taxonomic

ranks (Pible et al., 2020). Based on predictable specific signatures for each microorganism present in the sample when querying a comprehensive database, this approach allows microorganisms to be identified even in complex mixtures, and their relative biomasses quantified. The approach was applied to screen isolates from several environments (Petit et al., 2020), can be amenable to high-throughput (Hayoun et al., 2020a), and is effective even on poorly-documented branches of the tree of life (Lozano et al., 2022). The limit of detection of this methodology has been recently reported to be 4 x 10^4 colony-forming units from a sample volume of 1 mL(Mappa et al., 2023a)

Analyzing multiple samples in a single nanoLC-MS/MS run would in principle save instrument time and would increase the throughput. In the present proof-of-concept study, we propose an innovative multiplex proteotyping methodology for use with microbial isolates that involves no protein or peptide labeling. The method relies on prefractionation of the peptidomes from each sample, pooling of specific fractions, and analysis of the pool in a single analytical nanoLC-MS/MS run. Specific MS/MS signals from each organism can be discriminated from the other signals by their hydrophobic characteristics. We also developed a specific proteotyping concept to assign taxonomical information to the signals based on the hydrophobicity of the peptides identified. In this article, we document the performance of the approach for the simultaneous analysis of 21 bacterial samples with a single 60-min gradient nanoLC-MS/MS run. Our approach significantly decreases the per-isolate cost of identification.

Experimental section

1.Microbial cultures and production of peptide digests

The 20 microbial strains were grown under aerobic conditions in liquid media until the stationary phase was reached (Supplementary Table S1). Cells were treated and disrupted as described (Hayoun et al., 2019). Tryptic peptides were produced in 96-well plate format as previously reported (Hayoun et al., 2020a). They were quantified with the Pierce Quantitative Colorimetric Peptide Assay as recommended by the supplier (Thermo Scientific, product number 23275). The efficiency of the tryptic digest was checked a posteriori by analysis of the average miss-cleavages of peptides detected by tandem mass spectrometry.

2. Off-line HPLC fractionation of peptide digests and assembly of M11 and M21

Peptide digests (25 µg) were fractionated off on a ZORBAX StableBond C18 column (Agilent) with particle size 5 µm, average pore size 300 Å, length 15 cm and internal diameter 4.6 mm with a 30-min linear gradient developed from 2.5% to 50% of acetonitrile with 0.1% formic acid at a flow rate of 400 µL.min-1. Fractions were collected every 30 s for 30 min, then dried down and resuspended in 50 µL TFA 0.1%. A 40 µL-volume of each of the chosen fractions were combined in a single tube. M11 and M21 samples were a mixture of eleven 1-min fractions and twenty-one 30-s fractions, respectively (Table 1).

Table 1: Composition of the concatenated fractions samples making up M11 and M21

	Strain	Fraction number
MIX M11	*Ralstonia pickettii*	F1
	Sphingomonas yabuuchiae	F3
	Microbacterium oxydans	F5
	Ralstonia pickettii	F7
	Stenotrophomonas maltophilia	F9
	Serratia marcescens	F11
	Kineococcus radiotolerans	F13
	Sagittula stellata	F15
	Pseudopedobacter saltans	F17
	Pseudomonas aeruginosa	F19
	Methylobacterium extorquens	F21
MIX M21	*Stenotrophomonas maltophilia*	F1
	Deinococcus proteolyticus	F2
	Bacillus thuringiensis	F3
	Serratia marcescens	F4
	Massilia timonae	F5
	Klebsiella aerogenes	F6
	Pseudomonas putida	F7
	Sagittula stellata	F8
	Ralstonia pickettii	F9
	Kineococcus radiotolerans	F10
	Ruegeria pomeroyi	F11
	Deinococcus deserti	F12
	Oceanibulbus indolifex	F13
	Sphingomonas yabuuchiae	F14
	Microbacterium oxydans	F15
	Bacillus cereus	F16
	Ralstonia pickettii	F17
	Pseudopedobacter saltans	F18
	Methylobacterium extorquens	F19
	Pseudomonas aeruginosa	F20
	Marivirga tractuosa	F21

3.Tandem mass spectrometry and interpretation for proteotyping at the species taxonomical rank

NanoLC-MS/MS analyses were conducted using a Q-Exactive HF tandem mass spectrometer (Thermo Fisher Scientific) in data-dependent mode with a 60-min gradient acquisition as described (Trapp et al., 2016). Proteotyping was performed as previously described (Hirtz et al., 2022; Lozano et al., 2022). Briefly, a database derived from the National Center

for Biotechnology Information non- redundant (NCBInr) database and comprising only a restricted number of representatives per species as presented earlier (Grenga et al., 2022) was queried using Mascot Daemon version 2.6.1 search engine (Matrix Science). A second database comprising all the identified genera from the first round search and their descendants was then created and searched for. A third query was then performed on an even more reduced database comprising only the annotated genomes of the strains belonging to the species identified in the second query. For MS/MS spectrum-to-peptide assignment, the following parameters were applied in the first search: mass tolerance of 3 ppm and 0.02 Da on parent ion and secondary ions, respectively, 2+ or 3+ as possible peptide charges, a maximum of one missed cleavage, carbamidomethylation of cysteine as fixed modification, oxidation of methionine as variable modification, trypsin as proteolytic enzyme. In the second and third searches, the same parameters were used, except that the mass tolerance was 5 ppm on parent ions and a maximum of two missed cleavages was allowed. The p-values used for peptide validation were 0.3, 0.15, and 0.05 in homology threshold mode for the first, second, and third search rounds, respectively. Peptide sequences were mapped to taxa at the species, genus, family, order, class, phylum, and superkingdom taxonomical ranks [3], resulting in Taxon-to-Spectrum Matches (TSMs). Taxonomies were identified based on these TSMs and from the number of taxon-specific peptide sequences (spePEP). For each 1-min window, the TSMs and spePEPs for the microbial species identified were extracted and normalized relative to the total number of bacterial TSMs and spePEPs per sample, respectively. Only species for which TSMs and spePEPs were assigned in more than three consecutive acquisition time intervals were retained. A filter combining the two parameters (1x TSMs + 2x spePEPs) for each 1-min window for each species was used to calculate the retention time (center value) and the r-squared value as an indication of confidence. This “Species Proteotyping index” (SPi) filter varied depending on the retention time.

4.Data availability

All mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository under dataset identifiers PXD035870 and 10.6019/PXD035870.

Results

Strategy for multiplex proteotyping of microbial isolates based on hydrophobicity fractionation of peptides

The label-free multiplexing protocol described here is a new method that we have developed to increase the throughput of tandem mass spectrometry proteotyping of microbial isolates. As shown in **Figure 1**, the concept involved the production of peptide mixtures from each individual microbial isolate. These peptide mixtures were then fractionated based on their hydrophobicity, and multiple fractions were subsequently concatenated to associate a specific reverse-phase chromatography retention time with each isolate. After merging the different peptide fractions, a single nanoLC-MS/MS analysis was performed to sequentially generate MS/MS spectra for each organism over a low- to high-hydrophobicity sequence. To establish the nature of each microorganism in this type of multiplexed nanoLC-MS/MS run, a highly discriminative proteotyping approach must be employed, ensuring identification without false-positives at each hydrophobicity window. Here, MS/MS spectra were first assigned to peptide sequences, resulting in peptide-to-spectrum-matches (PSMs). The TSMs at the different taxonomic ranks could then be extracted by mapping the peptides to all theoretical proteomes contained in the database. Taxon-specific peptides were also obtained based on the search for the lowest common ancestor. TSMs and taxon-specific peptides were then used to identify the microorganisms.

Bacterial isolates can be proteotyped from a short acquisition window, even with a low number of MS/MS spectra

In order to check whether the tryptic peptides obtained after proteolysis of the soluble proteomes of different species are similar when eluted from the reverse-phase chromatography, we analyzed the peptidomes from three representative species– *Pseudomonas putida, Klebsiella aerogenes, and Ralstonia pickettii* – in a 60-min gradient. The resulting datasets comprised 37 139, 35 759, and 27 339 MS/MS spectra, respectively. Query of the comprehensive NCBInr database resulted in 9 706, 9 137, and 7 401 unique peptide sequences, respectively. The total number of TSMs extracted at the species level was 21 825 for *P. putida*, 20 681 for *K. aerogenes,* and 11 856 for *R. pickettii*. Importantly, the general profile of the peptides eluted from the reverse-phase chromatography was quite similar for the three peptidomes.

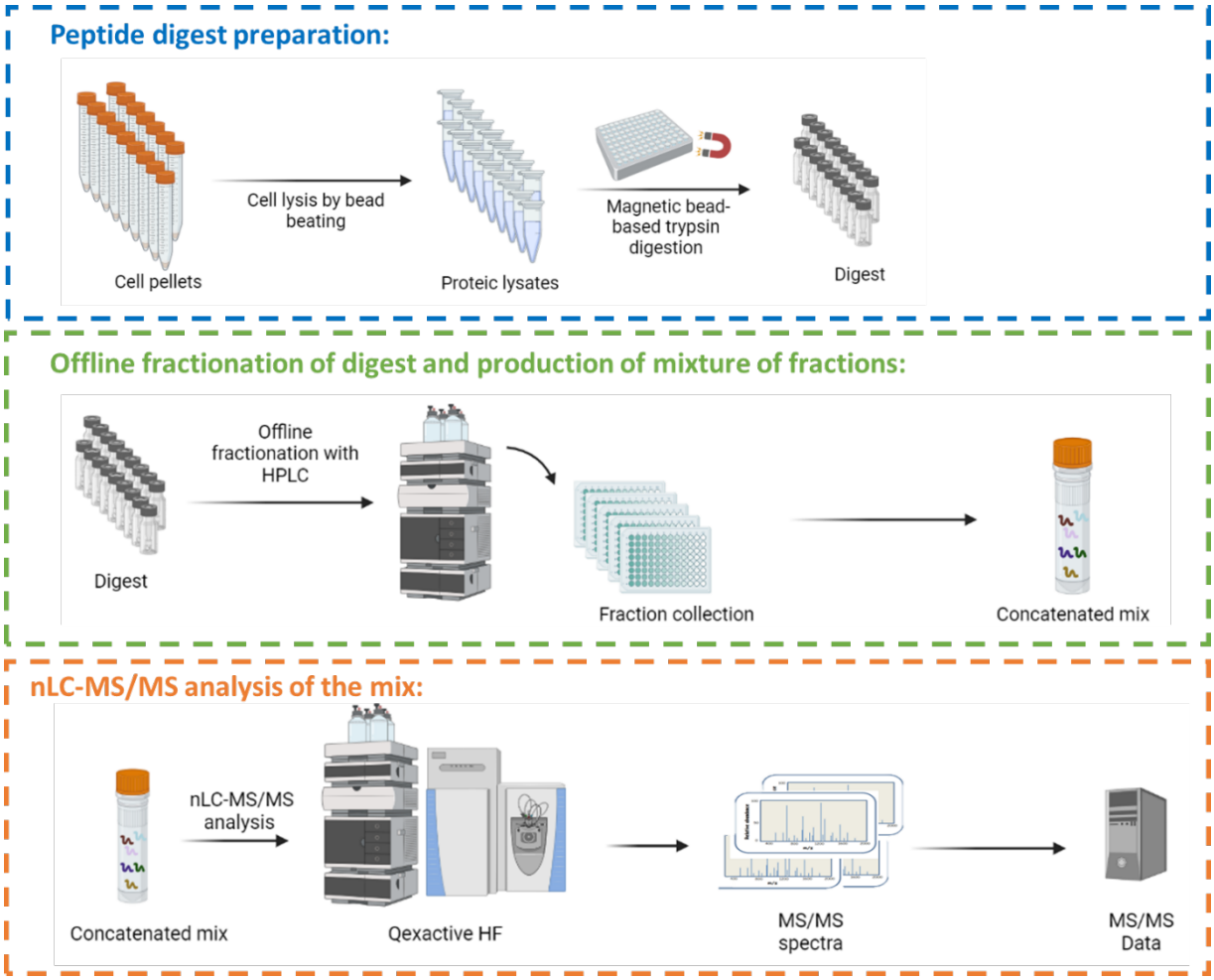


Figure 1. Workflow for multiplex proteotyping of bacterial isolates using Phylopeptidomics.

For our label-free proteotyping concept, we assessed whether short, 1-min, windows of tandem mass spectrometry data were sufficient to confidently identify the expected species. The *P. putida* peptidome dataset was split into 60 short windows corresponding to 1 min of acquisition time. Each of these small datasets was proteotyped without a priori. The *P. putida* species was identified in each of the fractions, based on at least 48 TSMs and 16 spePEPs. As shown in **Figure 2 (Panel A),** an important number is observed for most fractions with an average of 396 ± 73 TSMs. However, the first five fractions and the last five fractions – the most hydrophilic and hydrophobic peptides, respectively – produced lower numbers of TSMs. Therefore, the TSM signal was at a high level (>200) over a very broad separation window. The number of spePEP sequences for each fraction was similarly distributed, with an average of 227 ± 44 spePEPs per fraction, excluding the five most hydrophilic and five most hydrophobic fractions. Similar results were observed with the datasets acquired on *K. aerogenes* (357 ± 95 TSMs and 184 ± 66 specific peptides) and *R. pickettii* (257 ± 73 TSMs and 13 ± 7 specific peptides), as shown in **Figure 2 (Panel B & C)**. These results indicated that a 1-min acquisition with the Q-Exactive HF instrument should be sufficient to identify the expected species by proteotyping.

Subsequently, the peptidome for the pure *R. pickettii* isolate was resolved into 60 x 30-s fractions with a 30-min gradient. As an example, we selected one of the fractions, F1 – eluting at 12 min and which will be used after for the assemblage of the mixture M11 – for analysis by nanoLC-MS/MS with a 60-min acetonitrile gradient. The resulting dataset was subdivided into 60 small sub-datasets corresponding to 1-min acquisition windows, and each of these sub-datasets was proteotyped. For 11 contiguous fractions, the interpretation undoubtedly pointed to

the presence of *R. pickettii*. These results confirm that a 1-min acquisition window provides enough information to proteotype the expected species. **Figure 3** shows the distribution of the number of TSMs allowing the identification of *R. pickettii* plotted against the retention time. These results show that it is possible to identify an isolate using a single HPLC fraction, equivalent to a reduced proportion of the eluted peptides.

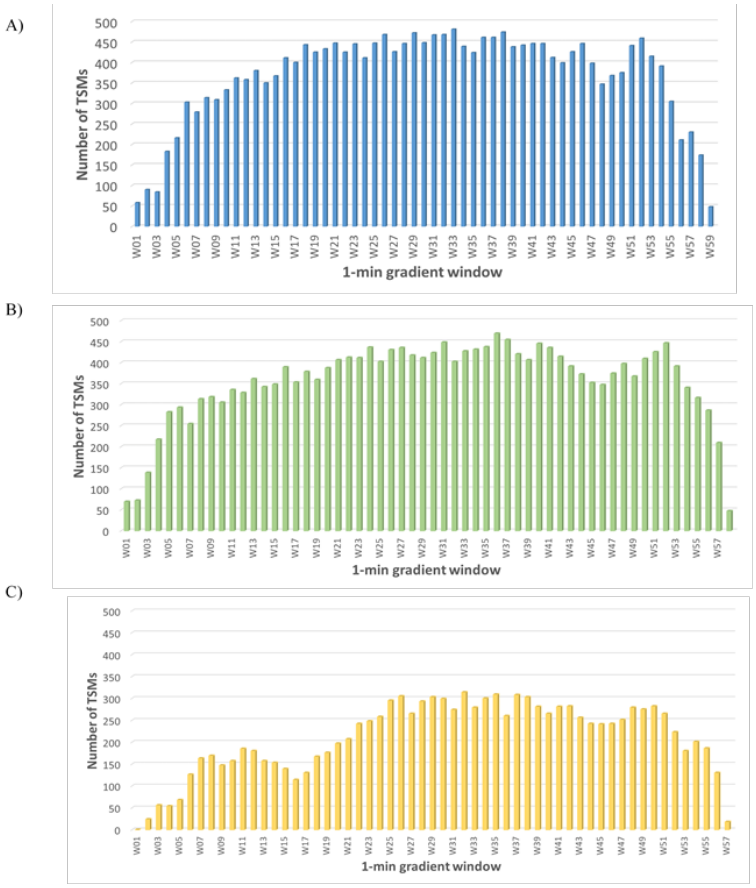


Figure 2. Distribution of TSMs attributed to *P. putida (A), K. aerogenes (B)* and *R. pickettii* (C) in 1-min acquisition windows. A) Distribution of total TSMs for 1-min gradient intervals over the whole 60-min chromatographic run for *P. putida*, B) Distribution of total TSMs for 1-min gradient intervals over the whole 60-min chromatographic run for *K. aerogenes*. C) Distribution of total TSMs in 1-min gradient intervals over the whole 60-min chromatographic run for *R. pickettii*. Data were obtained by shotgun proteomics. W: window.

Multiplex proteotyping applied to 10 isolates analyzed in a single nanoLC-MS/MS run.

In addition to the *R. pickettii* peptidome, nine other microbial peptidomes were resolved individually by reverse-phase chromatography with a 30-min gradient of acetonitrile, systematically collecting 30-s fractions. A sample consisting of specific fractions of these 9 peptidomes was assembled, together with two fractions from the *R. pickettii* peptidome. Thus, a duplicate of one of the isolates at different chromatographic elution times was included. The mixture of peptides (M11) was then injected into a nanoLC reverse-phase column coupled to a tandem mass spectrometer and subjected to a 60-min gradient and MS/MS analysis. The resulting MS/MS dataset comprised 30 881 MS/MS spectra. Once again, the data were split into 1-min acquisition windows for interpretation. We propose to name the combination of TSMs and spePEPs at a given retention time the "Species Proteotyping index" (SPi), as both parameters contribute to the proteotyping result. To increase the specificity of this index, a weighting factor of two was applied to spePEPs. Ten species were identified when all the fractions were merged, and only the species with SPi values > 0 over more than three consecutives acquisition time intervals were retained.

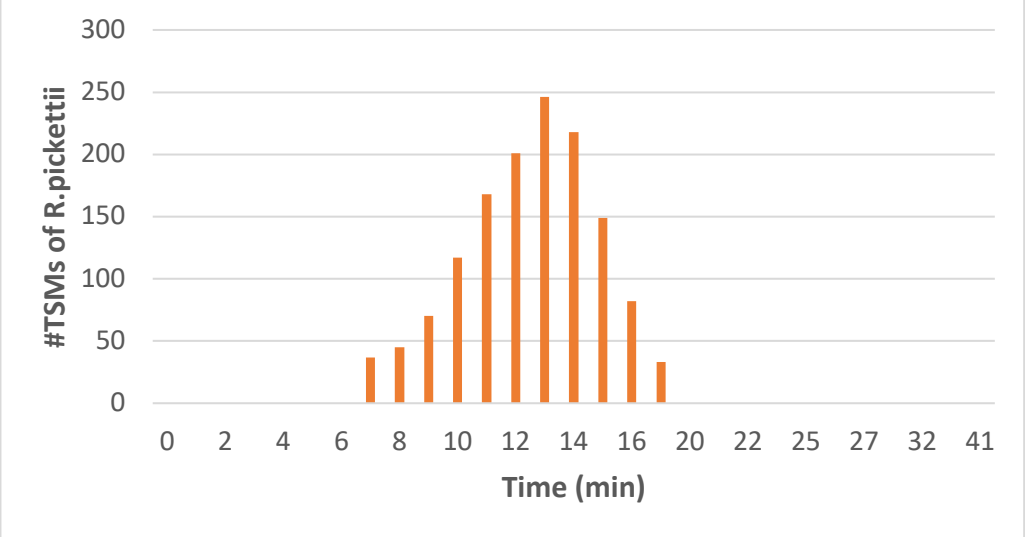


Figure 3. Number of TSMs attributed at different retention times for a single *R. pickettii* HPLC fraction. Number of TSMs pointing to the identification of *R. pickettii* in fraction 1 from *R. pickettii* for each 1-min acquisition period.

The ten species identified perfectly matched those expected for this M11 sample.

To establish the link between each fraction and the corresponding isolate, a proteotyping-retention time corresponding to the SPi maximum was established for each isolate. To do so, SPi values were plotted according to the 1-min windows for each of the ten species contained in the mixture. **Figure 4** shows the plots obtained for two species, namely *Sagitulla stellata* and *R. pickettii*. As observed from this figure, well-defined elution peaks were obtained. The proteotyping interpretation profile for each species has a bell shape representative of the elution of the peptides from each specific fraction included in the mixture. The shape of these retention-time-resolved proteotyping results allows the center of the elution peak to be precisely defined through experimental–data curve fitting with a Lorentzian function. As expected, a single peak was observed for *S. stellata* fraction F15 with a retention time (*S. stellata* SPi maximum) at 37.6 min, whereas two well-separated peaks were observed for *R. pickettii* at 12.4 min and 22.4 min (*R. pickettii* SPi maximum). These results are in full agreement with the two fractions, F1 and F7, from this organism contributing to the M11 assemblage. Lorentzian curve fitting was applied to each peak corresponding to the eight other fractions of M11 (Supplementary Figure S1) to determine the respective proteotyping elution time: *Sphingomonas yabuuchiae* F3 at 16.3 min, *Microbacterium oxydans* F5 at 19.3 min, *Stenotrophomonas maltophilia* F9 at 26.1 min, *Serratia marcescens* F11 at 29.7 min, *Kineococcus radiotolerans* F13 at 33.5 min, *Pseudopedobacter saltans* F17 at 41.5 min, *Pseudomonas aeruginosa* F19 at 45.3 min, and *Methylobacterium extorquens* F21 at 49.3 min (Supplementary Table S2). As expected, these proteotyping-retention times were increasing with the fraction numbers from the initial chromatography runs from which M11 was assembled. **Figure 5** shows the relationship between fraction numbers and SPi maximum. A second polynomial regression fitting gave $y = 0.7018x^2 + 94.781x + 662.8$, with a coefficient of determination (r-squared) of 0.9997. The average difference between two consecutive proteotyping-retention times was thus 221 (±21) s, allowing clear distinction between the fractions. Consequently, within the F1-F21 range, a larger number of fractions corresponding to even shorter hydrophobic windows could be introduced to allow greater multiplexing.

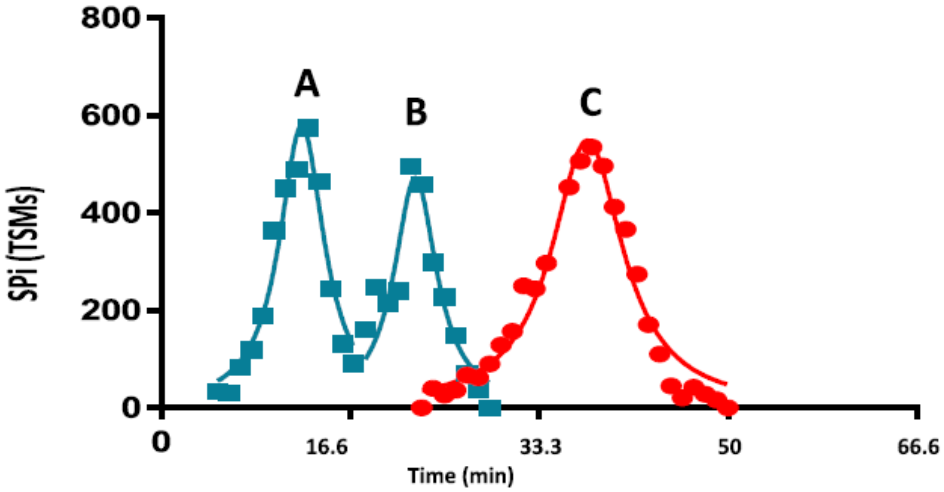


Figure 4. Scatter plot of Species Proteotyping index (SPi) for *R. pickettii* and *Sagitulla stellata*.

Label-free multiplex proteotyping is successful on fractions assembled from 20 isolates

We next test the multiplexing capacity of our method by assembling the M21 chimeric sample by mixing 21 distinct 30-s fractions from 20 individual peptidomes prepared from isolates. Like for M11, two fractions from the *R. pickettii* isolate were included in this assemblage as a control. The peptidome assemblage was once again analyzed by tandem mass spectrometry with a 60-min gradient, resulting in a dataset comprising 28 883 MS/MS spectra. Proteotyping, carried out as described for the M11 mixture, revealed the presence of 20 distinct species in the dataset: *Bacillus cereus, Bacillus thuringiensis, Deinococcus deserti, Deinococcus proteolyticus, K. radiotolerans, K. aerogenes, Massilia timonae, Marivirga tractuosa, M. extorquens, M. oxydans, Oceanibulbus indoliflex*, *P. aeruginosa, P. putida, P. saltans, R. pickettii, Ruegeria pomeroyi, S. stellata, S. marcescens, S. yabuucchiae,* and *S. maltophilia.* Notably, several pairs of isolates belonging to the same genus were identified, and discriminated between at the species level: two *Pseudomonas*, two *Bacillus*, and two *Deinococcus*. **Table 2** shows the attribution of each organism to its corresponding fraction. In the first four columns, the theoretical time is calculated as an approximation of the expected retention time. These times were calculated using the second polynomial equation curve for M21. The low fraction and high fraction limits indicate the theoretical limits of the range for each fraction. The Lorentzian nonlinear regression of the SPi values for each microorganism was used to determine the retention time for each species identified. The 21 SPi maximum for the different species agreed perfectly with the chronological order, with values between the theoretical low and high limits. We used the second polynomial regression ($y = 0.394x^2 + 99.935x + 654.68$; $r^2 = 0.9988$) to determine the respective origins of each of the identified species, *i.e.,* to which initial fraction they belong. As expected, the retention time agreed perfectly with the attribution of the fraction. The difference between consecutive proteotyping-retention times was 108 (±32) s on average. This result is in full agreement with the time difference obtained between two consecutive proteotyping-retention times for M11. As for M11, the presence of two distinct fractions of *R. pickettii* was confirmed by the two corresponding SPi peaks.

Discussion

The aim of this study was to develop a label-free multiplexing strategy for analysis of microbial isolates by tandem mass spectrometry. Indeed, tandem mass spectrometry-based proteotyping has an important role to play in the massive identification of isolates, especially environmental microorganisms (Lozano et al., 2022) through deeper characterization of microbiomes (Van Den Bossche et al., 2021), or medically-relevant but poorly characterized pathogens (Grenga et al., 2019). Improving the throughput of this methodology requires robust sample preparation methods that can be applied to any sample (Hayoun et al., 2019), and miniaturization of the experimental load upstream of the mass spectrometry step (Hayoun et al., 2020a). Additional improvements to throughput, such as multiplexing samples to reduce

the per-sample costs of mass spectrometry are of considerable interest. Multiplex isobaric labeling has been successfully used to increase throughput, however, the cost of the chemical reagents is .

detrimental for massive use. The innovative label-free multiplex strategy described here can be applied to proteotype microbial isolates

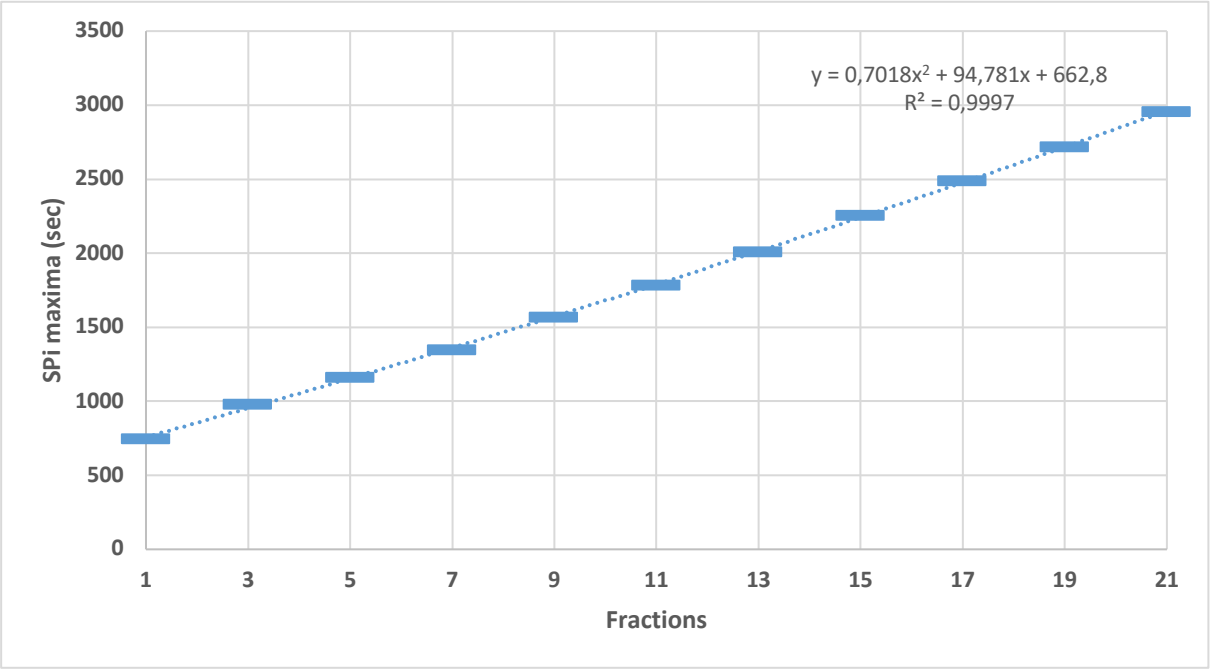


Figure 5. The fraction number correlates almost perfectly with the corresponding proteotyping-retention time value (SPi maxima). A second polynomial regression fitting was established using the theoretical fractions with the SPi maximum in chronological order; the equation y = 0.7018x2 +94.781x + 662.8 was used to determine the SPi identified for each fraction.

We have demonstrated its capacity to simultaneously analyze 21 bacterial isolates within a single 60-min gradient nanoLC-MS/MS run, resulting in significant time and cost savings compared to 21 separate nanoLC-MS/MS analyses. The label-free multiplexing concept we developed relies on prefractionation, where the peptide digest from each isolate to be characterized is first resolved by reverse-phase chromatography. This off-line peptide separation should ideally be performed with the same separation mode and in similar conditions to the on-line separation applied during the final nanoLC-MS/MS analysis. Here, HPLC at normal flow was used with a reverse-phase column with distinct characteristics to the column used with the nanoflow LC-MS/MS system. Although there were notable differences between the two chromatography runs, the elution profile of the fractions was consistent across both systems. To collect peptide fractions, we used a HPLC system equipped with an automatic collector allowing fraction collection in 96-well plates. Numerous samples could be fractionated with the automated procedure, but operator handling of the 96-well plate is required after fractionation. In the future, this sample preparation protocol could be fully automated to produce and store a single specific fraction for each isolate, thus minimizing labware-related costs. Because reverse-phase chromatography with HPLC is a highly robust and reproducible technique, this sample preparation step resulted in a robust procedure. Indeed, all of the microbial digests analyzed in this study presented similar chromatographic profiles during prefractionation, and each peptide fraction was clearly distinguishable along the acetonitrile gradient during the second chromatography run. Despite the differences in chromatographic conditions, overlap between adjacent fractions was quite low during the second chromatography run.

Table 2. Correlation between SPi maximum and fraction origin for the 20 species identified in the M21 assemblage. Values obtained for mix M21 are shown. The first column lists the experimental fraction used to create the mix. The 'Expected SPi max' is a theoretical value calculated using the correlation curve. The 'low' and 'high' fraction limits correspond to the range for each fraction. The fraction identified must fall within the range. The 'Measured SPi maximum' corresponds to the value obtained from the nonlinear regression curve, corresponding to the maximum of the SPi. The 'fraction identified' column corresponds to the theoretical calculation of the fraction number from the SPi maximum measured.

Fraction used	Expected SPi maximum (s)	SPi low fraction limit (s)	SPi high fraction limit (s)	Bacterial species identified	Measured SPi maximum (s)	Fraction identified
1	755.009	704.746	805.469	*S. maltophilia*	767	1
2	856.126	805.469	906.98	*D. proteolyticus*	896	2
3	958.031	906.98	1009.279	*B. thuringiensis*	939	3
4	1060.724	1009.279	1112.366	*S. marcescens*	1065	4
5	1164.205	1112.366	1216.241	*M. timonae*	1165	5
6	1268.474	1216.241	1320.904	*K. aerogenes*	1222	6
7	1373.531	1320.904	1426.355	*P. putida*	1345	7
8	1479.376	1426.355	1532.594	*S. stellata*	1474	8
9	1586.009	1532.594	1639.621	*R. pickettii*	1570	9
10	1693.43	1639.621	1747.436	*K. radiotolerans*	1701	10
11	1801.639	1747.436	1856.039	*R. pomeroyi*	1790	11
12	1910.636	1856.039	1965.43	*D. deserti*	1943	12
13	2020.421	1965.43	2075.609	*O. indolifex*	2042	13
14	2130.994	2075.609	2186.576	*S. yabuuchiae*	2146	14
15	2242.355	2186.576	2298.331	*M. oxydans*	2243	15
16	2354.504	2298.331	2410.874	*B. cereus*	2342	16
17	2467.441	2410.874	2524.205	*R. pickettii*	2514	17
18	2581.166	2524.205	2638.324	*P. saltans*	2579	18
19	2695.679	2638.324	2753.231	*M. extorquens*	2694	19
20	2810.98	2753.231	2868.926	*P. aeruginosa*	2780	20
21	2927.069	2868.926	2985.409	*M. tractuosa*	2921	21

The label-free multiplexing concept was validated in our study using two mixtures of peptide fractions, M11 and M21 – comprising 11 and 21 fractions, respectively. These mixtures were subjected to a 60-min gradient analysis on a Q-Exactive HF tandem mass spectrometer incorporating a high-field Orbitrap analyzer. Microorganisms could be identified at the species level from acquisition data recorded over just 1 min thanks to the high number of informative MS/MS spectra acquired. Even higher performance can be expected with more recent tandem mass spectrometers that can deliver more MS/MS spectra per unit of time. Here, we identified the equivalent of 21 hydrophobicity-resolved microorganisms in a single 60-min tandem mass spectrometry acquisition, resulting in a rate of identification of one microorganism every 3 min of mass spectrometry.

In terms of proteotyping performances, the high potential of tandem mass spectrometry has been extensively demonstrated [11,18,19]. The results presented here confirm the discriminative power of this methodology even in a complex situation where a high number of isolates are mixed together. Microorganisms from the same genus were easily distinguished at the species level, as shown for three genera: *Pseudomonas* with *Pseudomonas aeruginosa* and *Pseudomonas putida;* *Deinococcoccus* with *Deinococcus deserti* and *Deinococcus proteolyticus;* and *Bacillus* with *Bacillus cereus* and *Bacillus thuringiensis*. In the latter case, the two bacterial species are generally

difficult to phylogenetically discriminate as their 16S rRNA share more than 99% sequence identity. To illustrate the application of the 1 min window acquisition method to a wide range of microorganisms, we were able to identify the fungus Cordyceps confragosa with 345 TSMs and 48 species-specific peptides for a 1 min window eluting at 16 min (F5) and 430 TSMs and 83 species-specific peptides for the 1 min window eluting at 25 min (F11). We also identified the yeast Saccharomyces cerevisiae yeast with 484 TSMs and 386 species-specific peptides and 509 TSMs and 400 species-specific peptides for the same two windows. This shows that the methodology developed in the present study can be applied to any kingdom.

A possible drawback of the approach presented could be the presence of exactly the same species in neighboring samples. For example, if F15, F16, and F17 originated from samples containing exactly the same *R. picketti* species, the Lorentzian fitting procedure might be difficult to perform, as an unexpectedly broad peak would be obtained. The possibility of such an occurrence could be included in the data-treatment procedure to ascertain whether this is the case. Moreover, dubious identification could be easily verified by analyzing a new fraction. A second possible drawback is the presence in one of the fractions of a mixture of microorganisms rather than a true isolate. Nevertheless, tandem mass spectrometry proteotyping has been shown to be capable of handling mixtures of microorganisms, and thus should be able to tackle such samples. We chose not to document such a case in this proof-of-concept study due to the numerous potential cases and ratios that would need to be analyzed. Future investigation could be undertaken to analyze whether different ratios of microorganisms and different pairs or trios of microorganisms, differing in terms of phylogenetic distances, could be identified after multiplexing. Another possible drawback of the methodology is the off-line fractionation that may appear as demanding additional work by the operator compared to online fractionation. However, in a clinical settings, off-line fractionation if done by a simple automate could be cost-effective as much simpler and robust than a more complex system.

In conclusion, our innovative label-free multiplex proteotyping approach can efficiently identify up to 21 microorganisms in a single 60-min nanoLC-MS/MS run. There is plenty of room for improvement of the methodology, and compatibility with higher levels of multiplexing can probably be achieved with the latest generation of tandem mass spectrometers. This study involved a variety of bacterial isolates from diverse phyla, including Gram-positive and Gram-negative bacteria. The methodology is perfectly streamlined with an integrated culturomics pipeline where MALDI-TOF mass spectrometry could be used to rapidly dereplicate isolates, and label-free multiplex proteotyping by tandem mass spectrometry could be used to identify the relevant microorganisms at the species level.

Acknowledgments

MC thanks the Région Occitanie and the CEA for supporting part of her PhD fellowship (grant 20007404/ALDOCT-001066). The authors also acknowledge support from the French National Agency for Research (ANR) through the "EndOMiX" (grant number ANR-19-CE34-0009) and "Dyn-Microbiome" (grant number ANR-20-CE34-0012) projects that contribute to the development of proteomics and proteotyping expertise in the research team. We warmly thank Karim Hayoun and Charlotte Mappa for their contributions to the biological material used in this study, and Karen Culotta for help with programming Python scripts. We also thank Guylaine Miotello, Mélodie Kielbasa, and Jean-Charles Gaillard for their help for operating tandem mass spectrometers.

References

(1) Suarez, S. Ribosomal Proteins as Biomarkers for Bacterial Identification by Mass Spectrometry in the Clinical Microbiology Laboratory. J. Microbiol. Methods 2013, 7.
(2) Hayoun, K.; Pible, O.; Petit, P.; Allain, F.; Jouffret, V.; Culotta, K.; Rivasseau, C.; Armengaud, J.; Alpha-Bazin, B. Proteotyping Environmental Microorganisms by Phylopeptidomics: Case Study Screening Water from a Radioactive Material Storage Pool. Microorganisms 2020, 8 (10), 1525. https://doi.org/10.3390/microorganisms8101525.
(3) Pible, O.; Allain, F.; Jouffret, V.; Culotta, K.; Miotello, G.; Armengaud, J. Estimating Relative Biomasses of Organisms in Microbiota Using "Phylopeptidomics." Microbiome 2020, 8 (1), 30. https://doi.org/10.1186/s40168-020-00797-x.
(4) Mappa, C.; Alpha-Bazin, B.; Pible, O.; Armengaud, J. Evaluation of the Limit of Detection of Bacteria by Tandem Mass Spectrometry Proteotyping and Phylopeptidomics. Microorganisms 2023, 11 (5), 1170. https://doi.org/10.3390/microorganisms11051170.
(5) Mesuere, B.; Devreese, B.; Debyser, G.; Aerts, M.; Vandamme, P.; Dawyndt, P. Unipept: Tryptic Peptide-Based Biodiversity Analysis of

Metaproteome Samples. J. Proteome Res. 2012, 11 (12), 5773–5780. https://doi.org/10.1021/pr300576s.
(6) Boulund, F.; Karlsson, R.; Gonzales-Siles, L.; Johnning, A.; Karami, N.; AL-Bayati, O.; Åhrén, C.; Moore, E. R. B.; Kristiansson, E. Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics. Mol. Cell. Proteomics 2017, 16 (6), 1052–1063. https://doi.org/10.1074/mcp.M116.061721.
(7) Mooradian, A. D.; van der Post, S.; Naegle, K. M.; Held, J. M. ProteoClade: A Taxonomic Toolkit for Multi-Species and Metaproteomic Analysis. PLOS Comput. Biol. 2020, 16 (3), e1007741. https://doi.org/10.1371/journal.pcbi.1007741.
(8) Kuhring, M.; Doellinger, J.; Nitsche, A.; Muth, T.; Renard, B. Y. TaxIt: An Iterative Computational Pipeline for Untargeted Strain-Level Identification Using MS/MS Spectra from Pathogenic Single-Organism Samples. J. Proteome Res. 2020, 19 (6), 2501–2510. https://doi.org/10.1021/acs.jproteome.9b00714.
(9) Petit, P. C. M.; Pible, O.; Eesbeeck, V. V.; Alban, C.; Steinmetz, G.; Mysara, M.; Monsieurs, P.; Armengaud, J.; Rivasseau, C. Direct Meta-Analyses Reveal Unexpected Microbial Life in the Highly Radioactive Water of an Operating Nuclear Reactor Core. Microorganisms 2020, 8 (12), 1857. https://doi.org/10.3390/microorganisms8121857.
(10) Hayoun, K.; Gaillard, J.-C.; Pible, O.; Alpha-Bazin, B.; Armengaud, J. High-Throughput Proteotyping of Bacterial Isolates by Double Barrel Chromatography-Tandem Mass Spectrometry Based on Microplate Paramagnetic Beads and Phylopeptidomics. J. Proteomics 2020, 226, 103887. https://doi.org/10.1016/j.jprot.2020.103887.
(11) Lozano, C.; Kielbasa, M.; Gaillard, J.-C.; Miotello, G.; Pible, O.; Armengaud, J. Identification and Characterization of Marine Microorganisms by Tandem Mass Spectrometry Proteotyping. Microorganisms 2022, 10 (4), 719. https://doi.org/10.3390/microorganisms10040719.
(12) Hayoun, K.; Gouveia, D.; Grenga, L.; Pible, O.; Armengaud, J.; Alpha-Bazin, B. Evaluation of Sample Preparation Methods for Fast Proteotyping of Microorganisms by Tandem Mass Spectrometry. Front. Microbiol. 2019, 10, 1985. https://doi.org/10.3389/fmicb.2019.01985.
(13) Trapp, J.; Almunia, C.; Gaillard, J.-C.; Pible, O.; Chaumot, A.; Geffard, O.; Armengaud, J. Proteogenomic Insights into the Core-Proteome of Female Reproductive Tissues from Crustacean Amphipods. J. Proteomics 2016, 135, 51–61. https://doi.org/10.1016/j.jprot.2015.06.017.
(14) Hirtz, C.; Mannaa, A. M.; Moulis, E.; Pible, O.; O'Flynn, R.; Armengaud, J.; Jouffret, V.; Lemaistre, C.; Dominici, G.; Martinez, A. Y.; Dunyach-Remy, C.; Tiers, L.; Lavigne, J.-P.; Tramini, P.; Goldsmith, M.; Lehmann, S.; Deville de Périère, D.; Vialaret, J. Deciphering Black Extrinsic Tooth Stain Composition in Children Using Metaproteomics. ACS Omega 2022, 7 (10), 8258–8267. https://doi.org/10.1021/acsomega.1c04770.
(15) Grenga, L.; Pible, O.; Miotello, G.; Culotta, K.; Ruat, S.; Roncato, M.; Gas, F.; Bellanger, L.; Claret, P.; Dunyach-Remy, C.; Laureillard, D.; Sotto, A.; Lavigne, J.; Armengaud, J. Taxonomical and Functional Changes in COVID -19 Faecal Microbiome Could Be Related to SARS-COV -2 Faecal Load. Environ. Microbiol. 2022, 1462-2920.16028. https://doi.org/10.1111/1462-2920.16028.
(16) Van Den Bossche, T.; Arntzen, M. Ø.; Becher, D.; Benndorf, D.; Eijsink, V. G. H.; Henry, C.; Jagtap, P. D.; Jehmlich, N.; Juste, C.; Kunath, B. J.; Mesuere, B.; Muth, T.; Pope, P. B.; Seifert, J.; Tanca, A.; Uzzau, S.; Wilmes, P.; Hettich, R. L.; Armengaud, J. The Metaproteomics Initiative: A Coordinated Approach for Propelling the Functional Characterization of Microbiomes. Microbiome 2021, 9 (1), 243. https://doi.org/10.1186/s40168-021-01176-w.
(17) Grenga, L.; Pible, O.; Armengaud, J. Pathogen Proteotyping: A Rapidly Developing Application of Mass Spectrometry to Address Clinical Concerns. Clin. Mass Spectrom. 2019, 14, 9–17. https://doi.org/10.1016/j.clinms.2019.04.004.
(18) Heyer, R.; Benndorf, D.; Kohrs, F.; De Vrieze, J.; Boon, N.; Hoffmann, M.; Rapp, E.; Schlüter, A.; Sczyrba, A.; Reichl, U. Proteotyping of Biogas Plant Microbiomes Separates Biogas Plants According to Process Temperature and Reactor Type. Biotechnol. Biofuels 2016, 9 (1), 155. https://doi.org/10.1186/s13068-016-0572-4.
(19) Karlsson, R.; Gonzales-Siles, L.; Gomila, M.; Busquets, A.; Salvà-Serra, F.; Jaén-Luchoro, D.; Jakobsson, H. E.; Karlsson, A.; Boulund, F.; Kristiansson, E.; Moore, E. R. B. Proteotyping Bacteria: Characterization, Differentiation and Identification of Pneumococcus and Other Species within the Mitis Group of the Genus Streptococcus by Tandem Mass Spectrometry Proteomics. PLOS ONE 2018, 13 (12), e0208804. https://doi.org/10.1371/journal.pone.0208804.

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c01975.
Strains used for the composition of the assemblages (Table S1); correlation between SPi maximum and fraction origin for the 11 identified species in M11 mix (Table S2); and Lorentzian curve fitting of each organisms corresponding to each 11 fractions of M11 (Figure S1) (PDF).

II-2 : Développement d'une méthode de multiplexage simplifiée sans marquage pour le protéotypage de mélanges de six isolats en une analyse unique par spectrométrie de masse en tandem

Dans le but de rendre plus facile d'utilisation la méthode de multiplexage sans marquage présentée dans la partie II-1, il serait nécessaire d'éviter l'utilisation d'un appareillage HPLC pour obtenir les fractions. Ici, nous proposons une autre façon d'obtenir un fractionnement pour le multiplexage sans marquage. Ce protocole de simplification se base sur l'utilisation de colonnes SPE (Extraction en phase solide) C18 s'éluant par centrifugation appelé en anglais *spin colonne* avec le même principe de séparation des peptides que celui utilisé pour le fractionnement HPLC. Ce format a été choisi pour obtenir des fractions ayant une hydrophobicité différente pour chacun des isolats sur la base de plusieurs critères : la séparation des peptides, la facilité de mise en oeuvre, un équipement matériel simple et peu coûteux et la possibilité d'automatisation future. De plus, la reproductibilité et la robustesse et les divers scénarios ambigus pouvant survenir en cas d'analyse sur des isolats inconnus sont exposés. La présence d'isolats en quantité différente de peptides ou la présence d'un même organisme dans plusieurs fractions du mélange qu'elles soient séparées ou consécutives seront expliqués.

Dans ce travail, 32 mélanges composés de peptides provenant de six isolats ont été constitués dont 23 mélanges utilisent les six isolats dans chaque fraction avec 23 combinaisons différentes et 9 mélanges utilisés pour évaluer les différents cas de figures ambigües auxquels on peut être confronté avec des isolats inconnus comme la présence d'un même isolat dans plusieurs fractions ou encore la différence de quantité de peptides dans chaque fraction. *Ruegeria pomeroyi*, *Oceanibulbus indolifex*, *Sagittula stellata*, *Klebsiella aerogenes*, *Marivirga tractuosa* et *Saccharomyces cerevisiae* ont été utilisés pour constituer les différentes combinaisons de mélange peptidique.

Ainsi les 23 mélanges ont été réalisés dans le but de démontrer la reproductibilité de la méthode de multiplexage. En effet, une stratégie de traitement des données de multiplexage sans marquage a été mise en place et présenté dans la partie II-1 menant à la définition du SPi. Ici, nous avons voulu montrer que pour chaque fraction était associé un SPi spécifique qui était reproductible. C'est ce qui a été observé à travers l'analyse de ces 23 mélange puisque sur les 23 combinaisons le même SPi était retrouvé pour chaque fraction et l'identification des organismes était dans le bon ordre. Cela montre donc la reproductibilité des résultats et donc la robustesse de la stratégie de traitement des données mise en place.

Dans un second temps, la robustesse de la méthode a également été illustrée par la création de 9 mélanges chacun réalisés avec soit la présence d'un même organisme dans plusieurs fractions du mélange soit avec des quantités de peptides différentes dans chaque fraction. La création de 4 mélanges comprenant un même organisme dans plusieurs fractions avec au moins une fraction d'écart a montré que l'identification était claire et que la distinction du même organisme dans chaque fraction était faite avec la distinction des 2 pics correspondant aux deux fractions permettant d'avoir un Spi spécifique et donc une association de la fraction à l'organisme de départ. 3 autres mélanges ont été créé avec la présence d'un même organisme dans plusieurs fractions consécutives. L'introduction d'un même organisme dans des fractions consécutives a montré que nous étions capables par notre stratégie de traitement des données de détecter ce cas-là du fait de l'élargissement du pic d'élution du peptidome de ce microorganisme. Ainsi en fonction de la valeur de largeur du pic d'élution, il est possible de dire si le même organisme est présent dans 2 ou 3 fractions consécutives. Enfin 2 autres mélanges ont été crée dans le but d'évaluer la différence de quantité de peptides dans chaque fraction. Après analyse cela n'a pas montré d'impact sur l'identification des microorganismes ni sur les SPi obtenus pour chaque fraction.

Les 32 mélanges ont été analysés sur un gradient de 20 minutes rendant ainsi équivalent le temps d'analyse avec la méthode par fractionnement HPLC (60 minutes pour 21 organismes). De plus le choix d'inclure *S. cerevisiae* a été fait pour montrer l'application de notre méthode à d'autres types d'organismes que les bactéries.

Dans ce chapitre l'adaptation et la validation de la méthode de protéotypage par fractionnement des peptidomes à un format plus facile d'utilisation ne nécessitant pas une instrumentation HPLC coûteuse et le savoir faire associé, ainsi que sa reproductibilité et la robustesse ont été exposés. Ces travaux ont montré la puissance de la méthode : l'identification de six microorganismes en une analyse unique de spectrométrie de masse de 20 minutes. Cela permet donc d'identifier un isolat en 3.3 minutes d'analyse MS. Ce manuscrit est soumis à publication au journal « Journal of Proteomic Research ».

A simplified label-free method for proteotyping sets of six isolates in a single tandem mass spectrometry analysis

Madisson Chabas[1,2], Jean Armengaud[1*#], Béatrice Alpha-Bazin[1*#]

[1]Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, 30200 Bagnols-sur-Cèze, France. [2]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Université de Montpellier, F-30207 Bagnols sur Cèze, France.

[#]These co-authors should be considered as co-last authors.

*Authors to whom correspondence should be addressed.

Keywords: proteotyping, tandem mass spectrometry, proteomics, bacterial isolates, concatenation, taxonomy, multiplex

Abstract

High-throughput identification of microorganisms is key for clinical diagnostics and microbiology in general. Multiplexing samples prior their detection is an attractive solution to reduce the costs and the time to results. Recent studies have demonstrated the discriminative power of proteotyping based on tandem mass spectrometry. As this technology can quickly identify the most probable taxonomical position of any microorganism, even if not yet previously characterized in terms of taxonomy, its application on environmental isolates or new emergent threats has been promoted. Here, we present a simplified label-free multiplexing method for the proteotyping of isolates by tandem mass spectrometry allowing the identification of six microorganisms in a single 20 min analytical run. The method allows obtaining similar throughput as whole-cell MALDI-TOF but is more discriminant and is applicable to any microorganism. The strategy relies on spin column fractionation to obtain fractions with different hydrophobicity for each of the six isolates. The assemblage of these 6 fractions is then analyzed by mass spectrometry. The interpretation establishes the link between each identified taxon with the initial sample based on the hydrophobic characteristics of the measured peptides. The robustness of the methodology has been tested on 32 different sets of six organisms as well as several worst-scenario assemblages with differences in sample quantities or the presence of the same organisms in multiple fractions. This study paves the way for deployment of the tandem mass spectrometry-based proteotyping methodology in microbiology laboratories.

Introduction

One of the most fundamental facet of microbiology is to identify and classify microorganisms in order to systematically explore microbial life, understand its evolution, improve fundamental knowledge, and exploit its catalytic resources. While microbiologists propose new DNA-based naming system for microbes to allow naming uncultivable microorganisms (Murray et al., 2020b), the field is debating the importance of isolating new isolates that will allow reproducibility of experiments for fundamental research and more direct biotechnological exploitation. Microbiome research has gained momentum as the tools to analyse complex samples comprising several hundreds of microorganism species have significantly progressed allowing both taxonomical description

and functional analysis(Armengaud, 2022). In order to gain more insights into the microbiomes, identifying and isolating the key players are crucial (Bilen, 2020). Culturomics explores the high-throughput screening of laboratory conditions to culture and isolate numerous microorganisms with the objective of establishing parameters to grow hitherto uncultivable microorganisms (Lagier et al., 2012). Important efforts at automatizing and standardizing the operations have been done (Antonios et al., 2021; Diakite et al., 2020). Indeed, culturomics has been shown highly valuable to extend the known repertoire of isolated archaea and bacteria from the human body as pathogens or commensals (Bilen et al., 2018; Vanstokstraeten et al., 2022). In such approach, dereplicating the isolates and quickly identifying them is key to concentrate efforts at the most interesting ones. For this, whole-cell MALDI-TOF mass spectrometry approach has been shown efficient as this methodology is rapid and low cost (Christie-Oleza et al., 2013). However, the identification is frequently erroneous as soon as the recorded spectrum is not corresponding to previously recorded spectra in the database. In order to improve culturomics throughput, novel methodology for quick identification of environmental isolates should be incorporated into the workflow. The same considerations apply well for clinical diagnostic of atypical pathogens or for epidemiological surveys requiring more information that just typing the species (Grenga et al., 2019) .

As an alternative to whole-cell MALDI-TOF, tandem mass spectrometry-based proteotyping has been shown to yield more peptide sequence information to rapidly identify atypical isolates and classify closely-related microorganisms at the subspecies level with its invaluable discriminative power. The methodology can even afford mixtures of microorganisms that usually represent a deadlock for whole-cell MALDI-TOF (Hayoun et al., 2020b). High-throughput application of this recent technology has been shown feasible in 96-well plate format and robust in terms of bacterial loads (Hayoun et al., 2020a). The applications of tandem mass spectrometry proteotyping have covered clinical diagnostics for the discrimination of *Pneumococcus* species within the Mitis group (Karlsson et al., 2018b), culture-free identification of *Francisella* directly from hare carcasses (Witt et al., 2020), detection of SARS-CoV-2 coronaviruses (Dollman et al., 2020), screening of marine isolates (Lozano et al., 2022), identification of the microbial flora of cystic fibrosis patients (Hardouin et al., 2022) and authentication of historical remains (Bourdin et al., 2023). Moreover, the methodology can be cost effective as recently shown with the possibility of multiplexing samples without the need of any reagent (Chabas et al., 2023). Typically, the identification of twenty-one organisms in a single tandem mass spectrometry analysis of 60 min was reported, thus a yield of an identification per 3 min. This method relies on separation of peptidomes by their hydrophobic characteristics, mixture of different fractions of the various isolates, and phylopeptidomics-based proteotyping. This last step relies on the assignation of taxonomical information for all the identified peptides (Pible et al., 2020). Taxon-specific peptides are sequences that are uniquely found in a given taxon, giving specificity to the identification. Taxon-to-Spectrum Matches (TSMs) entities correspond to the number of spectra assigned to a given taxon and thus represent a proxy of the abundance of the identified taxa. The methodology takes into account all the assigned peptides shared either by closely- or distantly-related organisms present in the generic database used to interpret the tandem mass spectrometry spectra at all possible taxonomic ranks. We reported the use of reverse phase separation by HPLC with a C18 reverse phase chromatographic column for resolving the peptidomes into fractions prior to their mass spectrometry analysis (Chabas et al., 2023). However, different means allowing similar separation can be used to generate fractions (Manadas et al., 2010). Magnetic beads (Deng et al., 2021), Solid-Phase Extraction columns (Herraiz and Casal, 1995), stage tips (Rappsilber et al., 2007) and spin columns (Dimayacyac-Esleta et al., 2015) have been shown to allow separation of peptides into fractions with defined characteristics without the need for a costly instrument, thus these approaches represent interesting alternatives to make cost effective and user friendly the label free, multiplexing proteotyping approach.

Here, we propose a novel approach to multiplex samples for tandem mass spectrometry proteotyping without the need of costly reagents based on reverse phase spin columns. We describe the multiplexing of six organisms that can be analyzed within 20 min of tandem mass spectrometry. In addition, we explore the robustness of the method by exemplifying several worst-case scenario such as the presence of identical organisms in various fractions or fractions differing in biomasses.

Materials & Methods

1.Biological material

The five bacteria used in this study (*Klebsiella aerogenes*, *Oceanibulbus indolifex (*recently renamed *Sulfitobacter indoliflex)*, *Marivirga tractuosa*, *Sagitulla stellata* and *Ruegeria pomeroyi)* were from a commercial source and cultivated as mentioned earlier (Chabas et al., 2023). *Saccharomyces cerevisiae* was from commercial baker's yeast bought in supermarket that was solubilized as follows: 108 mg in 25 mL of PBS at pH 7.4 (GIBCO). Pellets cells were obtained after two centrifugations for the removing of residual liquid medium as described in Chabas et al,. 2023 (Chabas et al., 2023). Resulting pellets were weighed and stored at -20 °C until use.

2.Preparation of peptide digests from isolates

Peptides were prepared as previously detailed(Chabas et al., 2023). Briefly, a specific volume of 1x lithium dodecyl sulfate sample loading buffer (Thermo Fisher Scientific) prepared without dyes and supplemented with 5% beta-mercaptoethanol (v/v) was added to each cell pellet (100 µL per 1.7 mg wet biomass). After their disruption, protein lysates were stored at -20 °C until use. Peptide digests were obtained by SP3-based proteolysis performed in a 96 well-plate format as described by Hayoun et al (Hayoun et al., 2020a, 2019). Lysate (20 µL) was mixed with 40 µg of Sera Mag beads (4µL), formic acid (12 µL) and CH_3CN to a final concentration of 85%. Successive washes with 70 % ethanol and CH3CN were made to purify trapped proteins using a Smart2 MBS (Tecan) neodymium magnetic rack. Finally, beads were resuspended in 10 µL of 50 mM NH_4HCO_3 supplemented with 0.01% Protease Max surfactant (Promega) and 1 $\mu g.\mu L^{-1}$ trypsin gold (Promega), and incubated for 15 min at 50°C. After removal of the paramagnetic beads, the total volume was adjusted to 50 µL with aqueous 0.1% TFA and the resulting peptide solution was acidified by addition of trifluoroacetic acid (0.5% final concentration). The concentration of peptides in the samples was measured using a Pierce colorimetric Peptide Assay, as recommended (Thermo Scientific).

3.Spin-column reverse-phase fractionation of peptide digests

Peptide digests were fractionated by C18 micro spin columns (Harvard Apparatus) with particle size of 10 µm and average pore size of 300 Å as recommended by the supplier. First the column was washed twice with a solution containing 80 % of CH_3CN and twice with a 0.5 % trifluoroacetic acid (TFA) – H2O solution. A quantity of 40 µg of peptides was loaded onto the column and centrifugated during 1 min at 1000 x g. The eluate was load again onto the column and centrifugated. Then, a step gradient from 1% to 32% CH_3CN in 0.1% formic acid was applied by incrementing 1% at each step. A fraction was collected at each step after a centrifugation step for 1 min at 1000 x g, giving a total number of 32 fractions. The quantity of peptides from each fraction was established by Nanodrop (Thermo Fisher) measurement at 214 nm. Fractions with 9% (F1), 12% (F2), 15% (F3), 18% (F4), 21% (F5) and 25% (F6) CH_3CN + 0.1% formic acid were used for the assemblage of mixtures.

4. Assemblage of different mixtures of 6 fractions

Identical quantities of peptides (100 ng) from six fractions of different hydrophobicity were pooled to obtain different mixtures (M601 to M623) representing combinations of the six microorganisms. Additional mixtures (M624 and M625) were assembled with different peptide quantities. Finally, seven mixtures (M626 to M632) were assembled with less than six organisms, by repeating the same organism for multiple hydrophobic fractions. Mixtures and their exact composition are listed in Table S1.

5. Tandem mass spectrometry

Assembled peptide mixtures (M601-M632) were analyzed by nanoLC-MS/MS with an ultimate 3000 nanoLC system (Thermo Fisher Scientific) coupled to a Q-Exactive HF tandem mass spectrometer as described (Trapp et al., 2016). Briefly, peptides were desalted on a reverse-phase PepMap 100 C18 µ-precolumn. Then, they were separated on a nanoscale PepMap 100 C18 nanoLC column with a flow rate of 0.3 $\mu L.min^{-1}$ following a two-slope 20-min gradient with 4-25% B from 0 to 17 min and 25-32% B from 17 to 20 min. Mobile phase A consisted of an aqueous solution of 0.1% (v/v) formic acid in water; phase B consisted of 0.1% formic acid in 100% CH_3CN. The data-dependent mode (DDA) was conducted with an MS acquisition range of 350 to 1500 *m/z.* The 20 most abundant precursor ions were selected for fragmentation, applying a 10-s dynamic exclusion window and a 1.6-*m/z* isolation window and 8.3e5 for the intensity threshold.

6 .MS/MS data interpretation for multiplex proteotyping at the species taxonomical rank

Tandem mass spectrometry proteotyping was performed as previously described (Chabas et al., 2023). Briefly, ion peak lists were extracted using Mascot Daemon software, version 2.6.0 (Matrix Science) generating a MGF file per sample. Then, the mgf file was split into 20- sec acquisition time window by an in-house python script. MS/MS spectra for each of the resulting files were interpreted using Mascot version 2.6.1 (Matrix Science) against the NCBInrS database (Grenga et al., 2022). A cascade search for the taxonomical analysis was applied as described in the previous study(Chabas et al., 2023) for each 20 sec splitted files. Peptide sequences were mapped to taxa at the species, genus, family, order, class, phylum, and superkingdom taxonomical ranks, as previously described (Hayoun et al., 2020b; Pible et al., 2020), resulting in Taxon-to-Spectrum Matches (TSMs). TSMs and taxon-specific peptide sequences (spePEP) were used for the taxonomic identification. The deconvolution of TSMs and spePEPs from an assemblage of fractions was performed as previously described from these 20 sec windows(Chabas et al., 2023). Only species identified with more than three consecutive acquisition time intervals were selected. Then, the Species Proteotyping index (SPi) was calculated by combining the two parameters (1xTSMs + 2 x SpePEP) and used for the calculation of the retention time. A non-linear regression curves of SPi as function of retention time was used for the determination of the retention time using GraphPad Prism version 6.0 (GraphPad software). Average SPi peak width was calculated by half-height section x half-width section of the modeled peaks. For quality control, MS/MS spectra datasets were also interpreted against a database comprising the protein sequences of the six microorganisms.

7. Mass Spectrometry and Proteomics Data

All mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository under the dataset identifiers PXD044755 and 10.6019/PXD044755.

Results

1. Strategy for multiplexing microbial isolates using C18 spin columns

We proposed a strategy to produce different fractions of peptides from microbial isolates using C18 reverse phase spin columns eluted with solvents with different hydrophobic characteristics. We exemplified this strategy with a set of six isolates that will be analyzed all at once by tandem mass spectrometry. **Figure 1** shows the different steps of the methodology. For each isolate, proteins are extracted and proteolyzed. The peptides from each digest are fractionated using their hydrophobicity characteristic with C18 spin column with simple application of solvents and centrifugation. Six fractions were selected (F1-F6) corresponding to different percentage of acetonitrile. For each of the six isolates a specific fraction is selected, taking care that each fraction is different for the six isolates in order that a specific reverse phase chromatography retention time is associated with each isolate. Then, the six fractions of peptides are mixed together resulting in six pools of peptides with different hydrophobicity characteristics that will eluted sequentially from a reverse phase chromatography. The mixture of peptides is then analyzed by a single run of nanoLC-MS/MS. To identify the organisms present, the file containing the MS/MS spectra recorded along their retention time is sliced into 20 sec portions. Taxonomical analysis of each of the subfiles by phylopeptidomics identifies the organisms present at a given retention time. Then, a specific index, named Species Proteotyping index (SPi), is calculated for each of the organisms identified in the whole dataset for each window of 20 sec. The SPi versus time data can be fitted to a Lorentzian model to obtain the maximum of the peak, called SPi max, corresponding to the retention time of each isolate. As indicated in **Figure 2**, six well-defined elution peaks are clearly distinguished for an experimental mixture that has been analyzed with a nanoLC-MS/MS run operated with a 20 min gradient. In this analytical run six organisms have been identified at the species level: *Sagittula stellata*, *Ruegeria pomeroyi*, *Oceanibulbus indolifex*, *Saccharomyces cerevisiae*, *Klebsiella aerogenes* and *Marivirga tractuosa.* Each organism is associated to

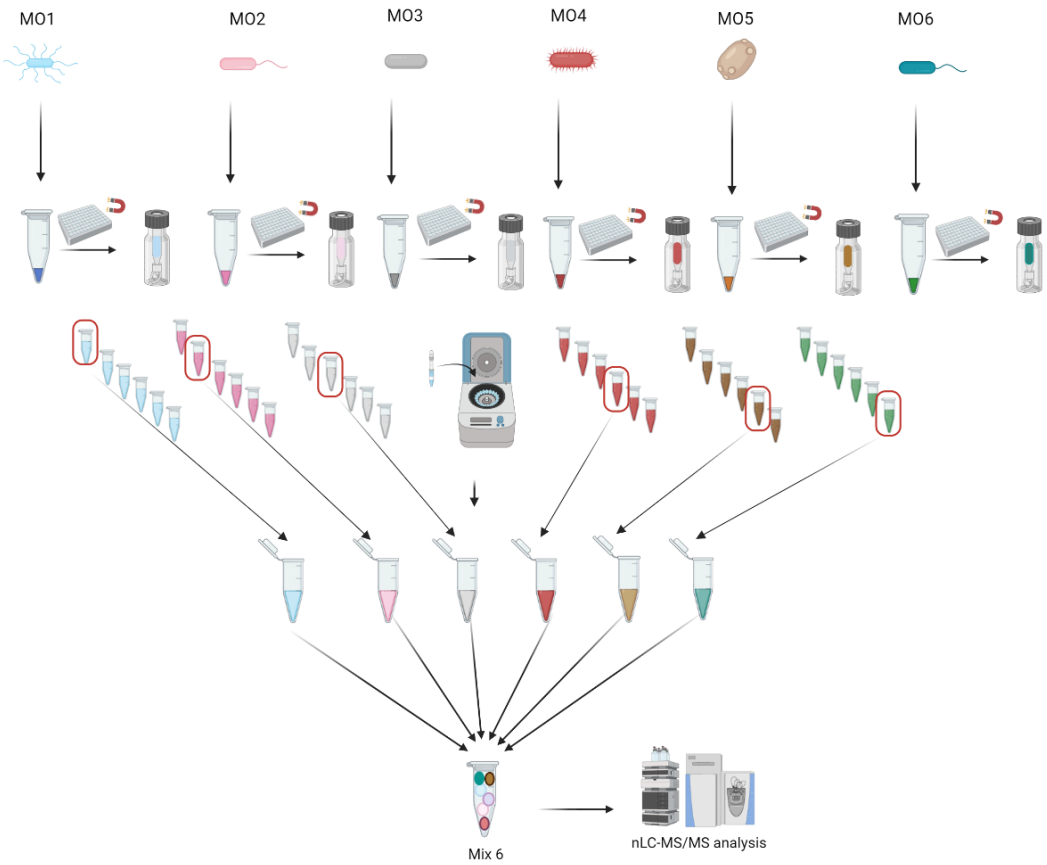


Figure 1. Workflow of data interpretation for multiplex proteotyping of bacterial isolates using spin column as fractionation way. The four main steps are schematized: i) extraction of proteins from each isolate and trypsin proteolysis, ii) fractionation of peptides by their hydrophobicity by simple centrifugation, iii) selection and assemblage of six fractions, and iv) nanoLC-MS/MS of the mixture.

a characteristic peak of elution and shows a SPi-max clearly defined and well separated from the others. Thus, a specific retention time can be assigned to each of the organisms. The order of assignation of the species to each isolate is in the expected order. These results shows that the proof of concept of multiplexing isolates for tandem mass spectrometry using a C18 spin column is validated. Last, a correlation curve showing the SPi-max index as exemplified in **Figure S2** confirmed the assignation order.

2. The label-free multiplexing method is robust and reproducible

To test the robustness and possible limits of the multiplexing method, the six peptide fractions that were collected systematically for the six isolates were used to create a diversity of combinations of assemblages. A total of 23 different mixtures were created with a peptide quantity for each fraction normalized at 100 ng. Each of these 23 multiplexed samples were analyzed with a 20 min nanoLC-MS/MS gradient. For all of the 23 mixtures, the 6 expected species were systematically identified and their order along the chromatography was found matching perfectly with the expected order. When analyzing the data obtained for these 23 assemblages, we found that the SPi-max intensities associated to each fraction was relatively constant for the five first fractions, but decreased for the more hydrophobic fraction. Indeed, the integrated signal of the modeled SPi peak is in average for the 23 assemblages 2120 ± 264 TSMs x sec for the first fraction, 2187 ± 438 TSMs x sec for the second fraction, 2687 ± 559 TSMs x sec for the third fraction, 2595 ± 563 TSMs x sec for the fourth fraction, 2241 ± 716 TSMs x sec for the fifth fraction, but only 1132 ± 538 TSMs x sec for the sixth fraction. Overall, the five first fractions have an average SPi-max of 2366 ± 257 TSMs x sec. This lower level of signal for the sixth fraction is not interfering in the correctness of the identification of the species, nor the correctness of the SPi-max measurement. The low taxonomical value of the most hydrophobic fraction is due to i) lower level of MS signals due to poor ionization of hydrophobic peptides, and ii) low number of assigned MS/MS spectra recorded because of poor fragmentation.

Figure S2 presents the results for the 23 mixtures with representation of the SPi values of each organism plotted as a function of time. **Figure 3** shows a boxplot of the measured retention times for the six peaks and the 23 mixtures. For the six fractions, the median SPi-max is at 554, 666, 780, 894, 1010, and 1159 sec of chromatography, respectively. The range of values are [536-570], [639-682], [764-803], [880-908], [993-1026], and [1143-1170], respectively. Thus, a clear separation of SPi-max retention times for each of the six fractions and a perfect linearity of retention times as a function of the increasing fraction number are observed in **Figure 3.** Interestingly, the lowest and the highest values of retention for each SPi peak are very close and within less than 43 sec at maximum, showing the high reproducibility of retention times for any of the six possible fractions and whatever the microorganism origin of the fractions. In average, the difference between the maximum and the minimum values is 34 sec. Based on these observations, a reference table of retention times associated to each of the six fractions could be established (**Table 1**). With the exact same experimental analytical set-up, these values can be used to directly associate a SPi-max retention time to one of the six possible fractions, thus a specific isolate. Interestingly, we noted that the width of each elution peak is similar regardless the organism found for the five first fractions with a width of 78 ± 8 sec in average. For the most hydrophobic fraction, the width is 45 ± 10 s as the intensity of this peak is half of the others.

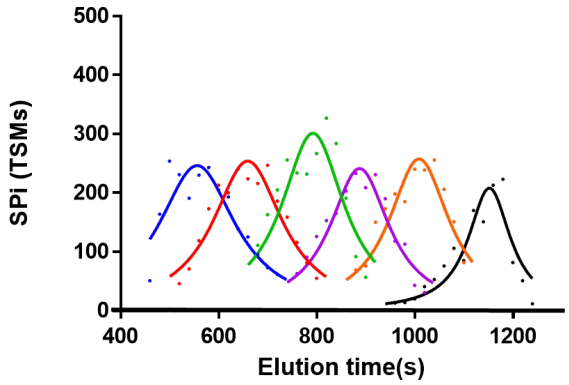


Figure 2. Scatter-plot of SPi values of each of the six organisms in 20-sec acquisition time window. SPi values of each organisms composing the mix are represented as function of 20 sec acquisition time windows. Fitting curves are represented by colored lines as follows: *S. stellata* (blue), *R. pomeroyi* (red), *O. indolifex* (green), *S. cerevisiae* (purple), *K. aerogenes* (orange), and *M. tractuosa* (black).

3. The multiplex proteotyping method can manage different quantities of peptides

An important question for a routine application of the methodology is whether all the six fractions should be equal in terms of peptide quantities or if differing quantities can be afford. To test the possible limits, two mixtures of six fractions were assembled with different peptide quantities.

Mix M624 comprises six peptide fractions: 280 ng of peptides from *K. aerogenes* (first fraction), 60 ng of *M. tractuosa* (second fraction), 100 ng of *O. indolifex* (third fraction), 340 ng of *R. pomeroyi* (fourth fraction), 180 ng of *S. cerevisiae* (fifth fraction), and 80 ng of *S. stellata* (sixth fraction). Its analysis by a 20 min gradient nanoLC-MS/MS and interpretation is indicated **Figure 4** (**Panel A)**. Six organisms could be detected and their corresponding SPi values were plotted against the elution time. Six peaks were clearly distinguished. As expected because of the differing quantities of peptides per fraction, differences in term of SPi signals are observed between organisms. For example, the SPi-max intensity associated to *K. aerogenes* (280 ng of peptides) is the double of that associated to *M. tractuosa* (60 ng). Regarding the integration of the whole SPi peak signal, the signal of the first species is 3.16 times higher than that of the second, a ratio corresponding exactly to the one expected from the injected quantities. The same was observed for *R. pomeroyi* with 3.17 times higher than the SPi value of other organisms excepted for *K. aerogenes*.

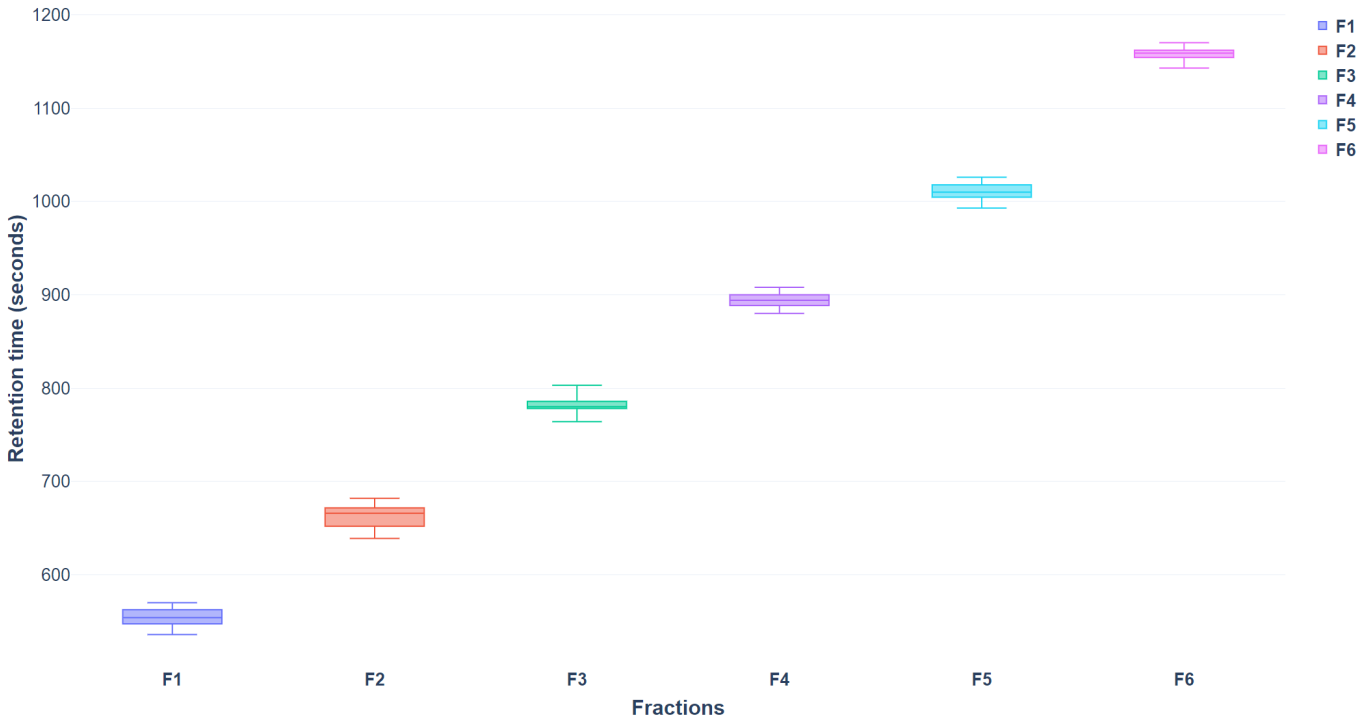


Figure 3. Retention times of each of the six fractions of the 23 mixtures. The box plot was established with all retention time value of the 23 mixtures as function of each fraction (F1, F2, F3, F4, F5, F6). For the six fractions, the median is at 554, 666, 780, 894, 1010, and 1159 sec of chromatography, respectively. The minimum and maximum values are [536-570], [639-682], [764-803], [880-908], [993-1026], and [1143-1170], respectively.

Mix M625 is the assemblage of six peptide fractions: 140 ng of peptides from *S. stellata* (first fraction), 120 ng of *R. pomeroyi* (second fraction), 40 ng of *M. tractuosa* (third fraction), 160 ng of *S. cerevisiae* (fourth fraction), 160 ng of *K. aerogenes* (fifth fraction), and 100 ng of *O. indolifex* (sixth fraction). **Figure 4 Panel B** shows the results of its analysis with the same previous condition. Once again, the six organisms were correctly identified and their elution peaks were clearly distinguished. The order of their respective retention times is in perfect agreement with the composition of the mixture and selected fractions. As indicated in the figure, here the SPi signal is relatively more similar between the peaks but a slight drop of SPi values were noted for *M. tractuosa* (third fraction) and *O. indolifex* (sixth fraction) as expected as their quantities are lower than the others with 40 and 100 ng, respectively.

The retention times associated to each fraction for the two mixtures are 562 ± 7 sec; 673 ± 7 sec; 777 ± 0 sec; 884 ± 4 sec; 1016 ± 5 sec; and 1163 ± 1 sec for the least to the most hydrophobic fractions. These values are perfectly matching the elution windows defined for the mixtures made with fractions with equal amounts of peptides which are indicated in **Table 1**. Thus, we observed in these two examples, that the difference of quantities of peptides per fraction does not affect the modeled retention time of the identified organism associated to each fraction.

4. The method is robust even if a same organism is present in several fractions in the mix

Another important question for a routine application of the methodology is whether the methodology is robust and delivers the expected identification when in the series of six samples to multiplex, several are the same species. Seven mixtures of six fractions were assembled to represent several scenarios. First, we analyzed whether the presence of the same organism in several distant fractions is disturbing the identification procedure. The analyses of the four first mixtures are presented in **Figure 5**. M629 comprises only 3 organisms: *R. pomeroyi*, *S. cerevisiae* and *K. aerogenes*. Each of the three organisms contributed to two fractions. As shown in **Figure 5 (Panel A),** two SPi peaks are clearly distinguished for each of these three organisms. The peaks at 556 and 885 sec corresponds to *R. pomeroyi.* At 788 and 1155 sec the peaks indicate the presence of *S. cerevisiae* and the peaks at 678 and 998 sec are

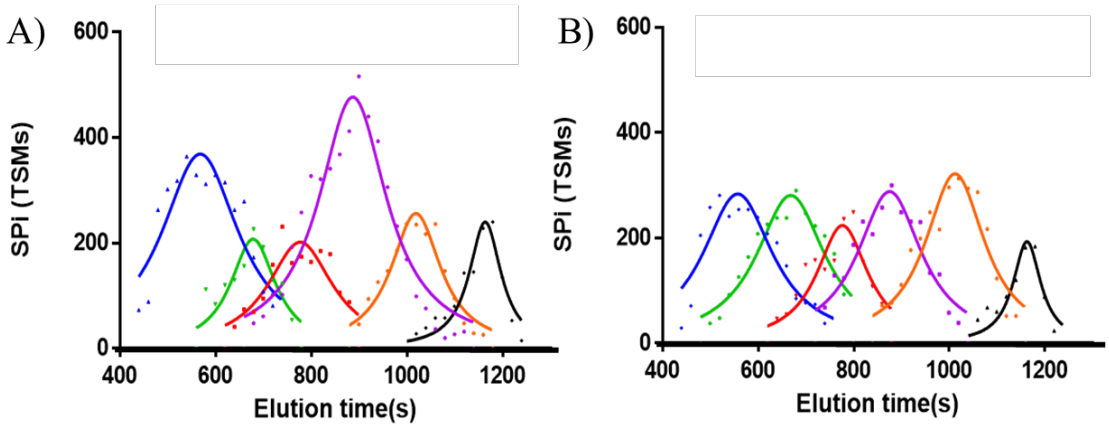


Figure 4. Scatter plots of Species Proteotyping index of mixtures with different quantities. Panel A. The SPi values of each organism are plotted and fit by Lorentzian model. The blue color is associated to *K. aerogenes* with a peptide quantity of 280 ng, the successive peaks are associated to *M. tractuosa* (60ng), *O. indolifex* (100 ng), *R. pomeroyi* (340 ng), *S. cerevisiae* (180ng) and *S. stellata* (80ng). Panel B. The SPi values of each organism are plotted and fit by Lorentzian model. The blue line is associated to *S. stellata* with a peptide quantity of 140 ng, the following peaks are associated to *R. pomeroyi* (120 ng), *M. tractuosa* (40 ng), *S. cerevisiae* (160 ng), *K. aerogenes* (160 ng) and *O. indolifex* (100 ng).

Table 1. Reference values of retention time associated to each of the six fractions. An average of all retention times for the 23 mixtures obtained for each specific fraction was calculated as reference values. The difference between the minimum and the maximum experimental values are used to determine the range of time of retention times for each of the six fractions with the retention time low range and the retention time high range values as indicated.

Fractions	Retention time reference value	Retention time low range value	Retention time high range value
1	554	520	588
2	664	630	698
3	782	748	816
4	894	860	928
5	1012	978	1046
6	1158	1124	1192

from *K. aerogenes*. As previously observed, the last peak (F6) associated to *S. cerevisiae* shows a decreasing of SPi intensity.

Indeed, SPi intensity for the last peak (F6) is at 119 and the average of SPi intensity for last fraction is 202 ± 69. M630 contains peptides corresponding to four organisms: *K. aerogenes*, *O. indolifex*, *R. pomeroyi* and *S. stellata* (**Figure 5, Panel B**). Two organisms have two distinct peaks: *O. indolifex* at 790 and 1017 sec, and *R. pomeroyi* at 659 and 1162 sec. The remaining two organisms are identified by a single peak observed at 549 sec for *S. stellata* and 901 sec for *K. aerogenes*. Fractions from four organisms were assembled for Mix M632: *K. aerogenes*, *S. cerevisiae*, *R. pomeroyi* and *S. stellata*. As shown in **Figure 5 (Panel C),** two organisms are identified with two peaks: *R. pomeroyi* at 656 and 1164 sec, and *K. aerogenes* at 795 and 1002 sec. For the remaining two organisms, a single peak was observed at 543 sec for *S. cerevisiae* and 901 sec for *S. stellata*. Mix M632 corresponds to peptides from four organisms: *R. pomeroyi*, *K. aerogenes*, *S. cerevisiae*, and *S. stellata*.

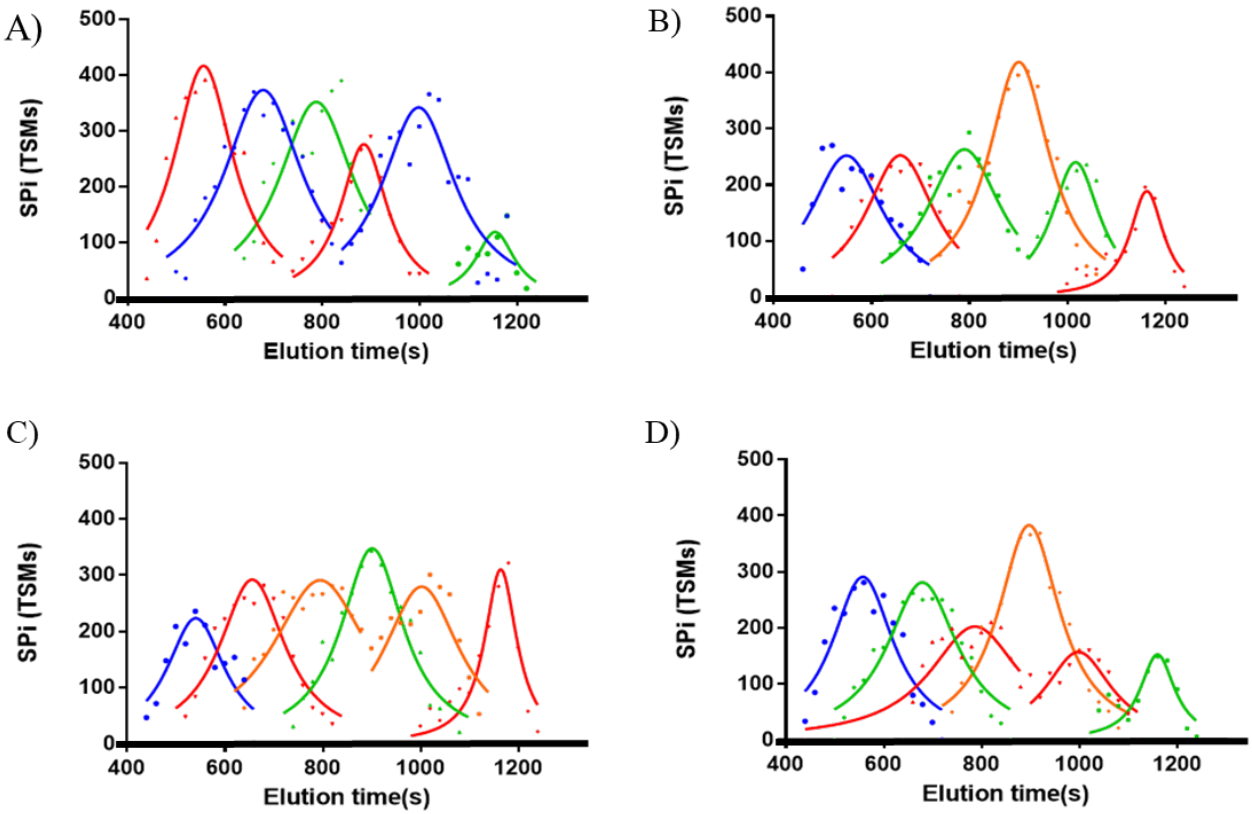


Figure 5. Scatter plots of Species Proteotyping index of mixtures containing a same organism in several fractions. SPi values of a given microorganism are plotted with a distinct color as a function of 20-s window acquisition time. Peptide quantities for each of the six fractions were normalized (100 ng). The Lorentzian model fit of SPi values for each organism is indicated with a colored line. The microorganism present twice is represented in red. *R. pomeroyi*, *K. aerogenes*, *S. cerevisiae*, *R. pomeroyi*, *K. aerogenes* and *S. cerevisiae* (Panel A), *S. stellata*, *R. pomeroyi*, *O. indolifex*, *K. aerogenes*, *O. indolifex* and *S. cerevisiae* (Panel B), *S. cerevisiae*, *R. pomeroyi*, *K. aerogenes*, S.stellata, *K. aerogenes*, and *R. pomeroyi* (Panel C), and *R. pomeroyi*, *K. aerogenes*, *S. cerevisiae*, *S.stellata*, *S. cerevisiae* and *K. aerogenes* (Panel D) are assigned to each peak in left-to-right order.

Figure 5 (Panel D) shows that two peaks can be distinguished for two organisms: *S. cerevisiae* at 786 and 998 sec and *K. aerogenes* at 679 and 1159 sec. For the remaining two organisms, a single peak was observed at 557 sec for *R. pomeroyi* and 897 sec for *S. stellata*. For the four mixtures, the identified organisms and the order of the retention time of the corresponding fractions are in perfect agreement with the composition of this mix. For these four assemblages, the average retention time for each fraction is: 551 ± 6 sec, 668 ± 12 sec, 789 ± 4 sec, 896 ± 7 sec, 1004 ± 9 sec and 1160 ± 4 sec for the least to the most hydrophobic fractions. These retention times are in perfect agreement with the reference values mentioned in **Table 1** with a CV <1%, showing again the robustness of the method.

Another more complex scenario was tested: three mixtures containing a same organism with two or three adjacent fractions were assembled and analyzed. The M626 assemblage comprises four organisms: *O. indolifex*, *S. cerevisiae*, *R. pomeroyi* and *K. aerogenes*. **Figure 6 (panel A)** shows the SPi values of each selected organisms plotted as a function of time, with a point every 20 sec acquisition time window. Four peaks are observed, the first peak being wider than the other peaks. Indeed, its width is 156 sec while the average of the other peaks is 78 ± 8 sec in average. The same observation can be done for the mix M627 represented on **Figure 6 (panel B),** where three organisms were identified. A single peak was associated to *K. aerogenes*, two peaks were attributed to *S. cerevisiae,* and a very wide peak was assigned to *R. pomeroyi* with a width of 128 sec. A diminution of the SPi intensity with an amplitude of 82 TSMs for the last peak (F6) associated to *S. cerevisiae* is observed. The average in term of SPi

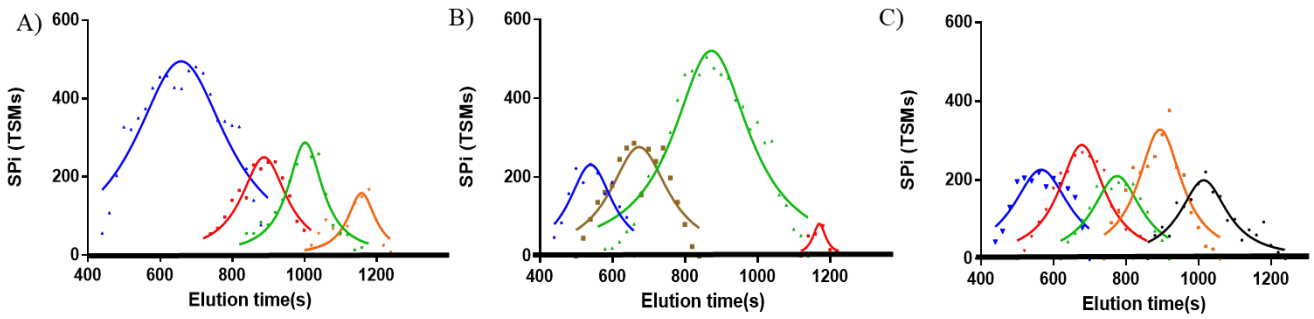


Figure 6. Scatter plot of Species Proteotyping index of mixtures containing the same organism in successive fractions. SPi values of a given microorganism are plotted with a distinct color as a function of 20-s window acquisition time. Peptide quantities for each of the six fractions were normalized (100 ng). The Lorentzian model fit of SPi values for each organism is indicated with a colored line. Panel A: M626 assemblage of *O. indolifex* (blue), *S. cerevisiae* (red), *R. pomeroyi* (green) and *K. aerogenes* (orange). Panel B: M627 mixture of *S. cerevisiae* (blue), *K. aerogenes* (brown), *R. pomeroyi* (green), and *S. cerevisiae* (red). Panel C: M628 assemblage *K. aerogenes* (Blue), *S. stellata* (red), *R. pomeroyi* (green), *O. indolifex* (orange) and *S. cerevisiae* (black). Large wide peak is an indicator of the presence of a same organism in consecutive fractions.

intensity for the last fraction is 202 ± 69 TSMs. It may be due to the previous larger peak associated to *R. pomeroyi*. M628 assemblage is made of peptide fractions from five organisms: *K. aerogenes, S. stellata, R. pomeroyi*, *O. indolifex* and *S. cerevisiae* are detected at 566 sec, 677 sec, 775 sec, 893 sec and 1014 sec, respectively, as indicated in **Figure 6 (panel C)**. For all the peaks, the width is in agreement with a single fraction. Only five peaks are observed but the more hydrophobic peak corresponding to *S. cerevisiae* can be assigned to the fifth and sixth fractions taking into account the width and the start and end values of the peak. In these last worst-case scenarios, we have highlighted that the width of the peak gives directly the information whether the same organism is present in a single fraction only or in several consecutive fractions in the assemblage. In conclusion, the six retention times expected and the peak width if less than six SPi-max are measured allow the final assignation of the organisms to each of the six initial samples.

Discussion

Isolating and identifying quickly microorganisms is pivotal for diagnostic purposes, searching for uncharacterized branches of the Tree of Life, or screening novel catalysts of interest in biotechnology. Tandem mass spectrometry-based proteotyping is an impressively accurate methodology, but improving its throughput is key to popularizing this approach. Its limit of detection recently reported for the bacterium *Salmonella bongori*, 40 thousands colony-forming units, is attractive as low amount of material is sufficient for the species delineation (Mappa et al., 2023). Furthermore, the methodology works whatever the physiological state of the bacteria (Mappa et al., 2023). Here, the aim of the study was to make cost-effective an innovative label-free multiplex proteotyping protocol that we previously developed (Chabas et al., 2023). While the first description of the methodology was relying on fractionation of the peptidome obtained from each isolate by reverse phase chromatography using HPLC, here we have proposed a simplified alternative based on C18 spin columns. We also evaluated the robustness of the methodology by exploring different worst-case scenarios.

A selected peptide fraction from each isolate is differentiated from the fractions arising from the other isolates by its hydrophobic characteristics. Because HPLC requires equipment and specific expertise, we proposed here to fraction the peptidomes with C18 spin columns that would require only a centrifuge. In principle, this approach can be fully automatized and is cost-effective. However, the approach is less resolute than HPLC. Here, six fractions differing in hydrophobicity are used for creating a mixture representative of six

isolates and analysed into a single 20 min nanoLC-MS/MS run. We noted that each of the resulting peaks of elution of species-specific peptides are clearly identified. Without any optimisation and automation, the fractionation step took about 1 h for treating in parallel six isolates. The current yield of identification is around 3 min of tandem mass spectrometry per isolate. This results in a rather competitive methodology in terms of costs. However, the time to results is higher and in our hands is strongly depending on the data treatment step that would require further optimization and use of more powerful computing systems. Some years ago, it has been shown that with a specific set-up for injecting samples and preformed gradient, the proteomic analysis of more than 200 samples per day was possible (Bache et al., 2018b), thus 7.2 min of tandem mass spectrometry per sample. Here, the objective being different, we could reduce the time per isolate with a classical nanoLC-MS/MS set-up. Combining our label-free multiplexing approach with such system able to pre-form gradient could result in even better results. Furthermore, the use of double-barrel liquid chromatography system coupled to the tandem mass spectrometer has been shown to be highly efficient in reducing the mass spectrometry time per sample (Hayoun et al., 2020a; Hosp et al., 2015). Such option that can be easily implemented for our strategy could further reduce the costs of mass spectrometry per sample.

The best practice to establish a novel analytical method is to test its robustness whatever the conditions. Here, the robustness of the methodology was tested by using 23 sets of six isolates with the same organisms but with different combinations. Fractions used for the artificial combinations were normalized with a peptide quantity of 100 ng per added fraction, resulting into mixtures of 600 ng of peptides injected per nanoLC-MS/MS. The peptides of each of the six microorganisms present in a mixture can delineate a specific taxon-retention time on the final reverse-phase chromatography. The dispersion of each of these six retention times observed for this set of 23 mixtures is minimal and shows a very good reproducibility. If similar conditions of nanoLC-MS/MS are used, reference values for the six expected species-retention times can be defined. Whether difference in peptide quantities in the six fractions could influence the results was also tested in this work. Introduction of fractions with different quantities in the range of 60-280 ng did not introduce any noticeable bias. Nevertheless, the normalization with a peptide quantity of 100 ng in each fraction is recommended, especially for the most hydrophobic fraction. Indeed the total number of SPi for the sixth fraction is always lower compared to the results from the five first fractions.

Finally, another potential difficulty was considered: the presence of the same organism in multiple isolates in the same multiplex set of six samples. Indeed, many culturomics project have to deal with dereplication of isolates to avoid waste of time in characterizing the same biological object several times (Kapinusova et al., 2022). The same appears true for clinical diagnostic where several samples commonly contain the same type of pathogens. Two types of mixtures were created: four sets with at least an organism repeated in two different fractions and three with the same organism repeated but in neighboring fractions. Once again, this worst-case scenario did not affect the identification success and the correct assignation to the expected initial samples. In this case, the width of peaks assigned to a given species, i.e. its retention time range, and its average retention time give information on the presence of the same organisms in neighboring fractions. As an example, we noted that in the case where three neighboring fractions contain the same organism, the SPi-max value for this organism corresponds exactly to the fraction in the middle, but the width of the peak is enlarged. Also, if the number of organisms is inferior to the number of introduced fractions, i.e. six here, a simple ratio of the width of peaks and the average of the expected width can confirm the presence of the same organism in neighboring fractions. Thus, these three parameters should be taken into account to determine the number of fractions containing the same organism. For mixtures composed with the same organism in different but not successive fractions two peaks of elution for the organism can be easily distinguished with width in full agreement with the average value. In such case, two SPi-max are obtained.

Interestingly, the method makes no use of pre-recorded mass spectra databases such as with whole-cell MALDI-TOF mass spectrometry (Suarez, 2013), but is rather based on the whole known diversity through the current database of annotated full genome sequences. It is thus applicable to any of the organisms already genome sequenced and their relatives, as well as uncharacterized branches of the tree of life because partial conserved information can be always retrieved (Armengaud, 2022). Regarding the range of application of the methodology, we have tested

whether bacteria or yeasts can be easily identified. Specifically, *S. cerevisiae* was introduced in several scenarios to prove that the method applies equally well on Bacteria and Eukaryota. *S. cerevisiae* was always identified. The extraction of peptides from *S. cerevisiae* or from the bacteria is identical as based on an optimized protocol previously developed for complex samples including both types of material (Hayoun et al., 2019). However, we observed less signal in terms of abundance (TSMs) and taxon-specificity (SpePEPs) compared to the signals for bacteria while the quantity of peptides was identical. This comes from the dynamic range of the proteomes that differ between Eukaryota and Bacteria on the one hand, and the number of genomes for this specific branch of the tree of life that may lower the *Saccharomyces cerevisiae*-specific peptides on the other hand.

In conclusion, the methodology presented here enables robust and cost-effective multiplexing of samples for microbial identification by tandem mass spectrometry proteotyping. The results show how simple the methodology is because the fractionation step before the analysis can be performed with C18 spin columns. Based on this affordable technology, the fractionation of peptidomes can be fully automatized to reduce operations and time to result. The current protocol allows the identification of 6 organisms in a single 20-min nanoLC-MS/MS analytical run, but there is room for improvement as the latest generation of tandem mass spectrometers have several fold more capacities than the mass spectrometer used here. This protocol paves the way for proteotyping of isolates by tandem mass spectrometry in clinical settings for the diagnosis of difficult organisms that are not yet in whole-cell MALDI-TOF databases or the culturomics pipeline dedicated to the discovery of new microorganisms.

Acknowledgements

MC thanks the Région Occitanie and the CEA for supporting part of her PhD fellowship. The authors also acknowledge support from the French National Agency for Research (ANR) through the "EndOMiX" (grant number ANR-19-CE34-0009) and "Dyn-Microbiome" (grant number ANR-20-CE34-0012) projects that contribute to the development of proteomics and proteotyping expertise in the research team.

Associated content – Supporting information

Table S1. List of assemblages and species for each of their six fractions.

Figure S1. Correlation of SPi maxima obtained for the six microorganisms from the M610 assemblage. The six microorganisms, their fractions and their SPi maxima expressed in seconds are: *Sagittula stellata* (F1, 555.7), *Ruegeria pomeroyi* (F2, 658.8), *Oceanibulbus indoliflex* (F3, 791.7), *Saccharomyces cerevisiae* (F4, 887.0), *Klebsiella aerogenes* (F5, 10008.0), and *Marivirga tractuosa* (F6, 1152.0). The correlation is shown with dotted line.

Figure S2. Scatter-plot of SPi values of each of the six organisms in 20-sec acquisition time windows for the assemblages M601 to M623. SPi values of each organisms are represented by colored dots for each assemblage as function of 20 sec acquisition time windows. Fitting curves are represented by colored lines for the different microorganisms.

References

(1) Murray, A. E.; Freudenstein, J.; Gribaldo, S.; Hatzenpichler
, R.; Hugenholtz, P.; Kämpfer, P.; Konstantinidis, K. T.; Lane, C. E.; Papke, R. T.; Parks, D. H.; Rossello-Mora, R.; Stott, M. B.; Sutcliffe, I. C.; Thrash, J. C.; Venter, S. N.; Whitman, W. B.; Acinas, S. G.; Amann, R. I.; Anantharaman, K.; Armengaud, J.; Baker, B. J.; Barco, R. A.; Bode, H. B.; Boyd, E. S.; Brady, C. L.; Carini, P.; Chain, P. S. G.; Colman, D. R.; DeAngelis, K. M.; de los Rios, M. A.; Estrada-de los Santos, P.; Dunlap, C. A.; Eisen, J. A.; Emerson, D.; Ettema, T. J. G.; Eveillard, D.; Girguis, P. R.; Hentschel, U.; Hollibaugh, J. T.; Hug, L. A.; Inskeep, W. P.; Ivanova, E. P.; Klenk, H.-P.; Li, W.-J.; Lloyd, K. G.; Löffler, F. E.; Makhalanyane, T. P.; Moser, D. P.; Nunoura, T.; Palmer, M.; Parro, V.; Pedrós-Alió, C.; Probst, A. J.; Smits, T. H. M.; Steen, A. D.; Steenkamp, E. T.; Spang, A.; Stewart, F. J.; Tiedje, J. M.; Vandamme, P.; Wagner, M.; Wang, F.-P.; Yarza, P.; Hedlund, B. P.; Reysenbach, A.-L. Roadmap for Naming Uncultivated Archaea and Bacteria. Nat. Microbiol. 2020, 5 (8), 987–994. https://doi.org/10.1038/s41564-020-0733-x.

(2) Armengaud, J. Metaproteomics to Understand How Microbiota Function: The Crystal Ball Predicts a Promising Future. Environ. Microbiol. 2022, 1462-2920.16238. https://doi.org/10.1111/1462-2920.16238.

(3) Bilen, M. Strategies and Advancements in Human Microbiome Description and the Importance of Culturomics. Microb. Pathog. 2020, 149, 104460. https://doi.org/10.1016/j.micpath.2020.104460.

(4) Lagier, J.-C.; Armougom, F.; Million, M.; Hugon, P.; Pagnier, I.; Robert, C.; Bittar, F.; Fournous, G.; Gimenez, G.; Maraninchi, M.; Trape, J.-F.; Koonin, E. V.;

La Scola, B.; Raoult, D. Microbial Culturomics: Paradigm Shift in the Human Gut Microbiome Study. Clin. Microbiol. Infect. 2012, 18 (12), 1185–1193. https://doi.org/10.1111/1469-0691.12023.
(5) Antonios, K.; Croxatto, A.; Culbreath, K. Current State of Laboratory Automation in Clinical Microbiology Laboratory. Clin. Chem. 2021, 68 (1), 99–114. https://doi.org/10.1093/clinchem/hvab242.
(6) Diakite, A.; Dubourg, G.; Dione, N.; Afouda, P.; Bellali, S.; Ngom, I. I.; Valles, C.; Tall, M. lamine; Lagier, J.-C.; Raoult, D. Optimization and Standardization of the Culturomics Technique for Human Microbiome Exploration. Sci. Rep. 2020, 10 (1), 9674. https://doi.org/10.1038/s41598-020-66738-8.
(7) Bilen, M.; Dufour, J.-C.; Lagier, J.-C.; Cadoret, F.; Daoud, Z.; Dubourg, G.; Raoult, D. The Contribution of Culturomics to the Repertoire of Isolated Human Bacterial and Archaeal Species. Microbiome 2018, 6 (1), 94. https://doi.org/10.1186/s40168-018-0485-5.
(8) Vanstokstraeten, R.; Mackens, S.; Callewaert, E.; Blotwijk, S.; Emmerechts, K.; Crombé, F.; Soetens, O.; Wybo, I.; Vandoorslaer, K.; Mostert, L.; De Geyter, D.; Muyldermans, A.; Blockeel, C.; Piérard, D.; Demuyser, T. Culturomics to Investigate the Endometrial Microbiome: Proof-of-Concept. Int. J. Mol. Sci. 2022, 23 (20), 12212. https://doi.org/10.3390/ijms232012212.
(9) Christie-Oleza, J. A.; Maria Piña-Villalonga, J.; Guerin, P.; Miotello, G.; Bosch, R.; Nogales, B.; Armengaud, J. Shotgun NanoLC-MS/MS Proteogenomics to Document MALDI-TOF Biomarkers for Screening New Members of the Ruegeria Genus. Environmental Microbiology. 2013.
(10) Grenga, L.; Pible, O.; Armengaud, J. Pathogen Proteotyping: A Rapidly Developing Application of Mass Spectrometry to Address Clinical Concerns. Clin. Mass Spectrom. 2019, 14, 9–17. https://doi.org/10.1016/j.clinms.2019.04.004.
(11) Hayoun, K.; Pible, O.; Petit, P.; Allain, F.; Jouffret, V.; Culotta, K.; Rivasseau, C.; Armengaud, J.; Alpha-Bazin, B. Proteotyping Environmental Microorganisms by Phylopeptidomics: Case Study Screening Water from a Radioactive Material Storage Pool. Microorganisms 2020, 8 (10), 1525. https://doi.org/10.3390/microorganisms8101525.
(12) Hayoun, K.; Gaillard, J.-C.; Pible, O.; Alpha-Bazin, B.; Armengaud, J. High-Throughput Proteotyping of Bacterial Isolates by Double Barrel Chromatography-Tandem Mass Spectrometry Based on Microplate Paramagnetic Beads and Phylopeptidomics. J. Proteomics 2020, 226, 103887. https://doi.org/10.1016/j.jprot.2020.103887.
(13) Karlsson, R.; Gonzales-Siles, L.; Gomila, M.; Busquets, A.; Salvà-Serra, F.; Jaén-Luchoro, D.; Jakobsson, H. E.; Karlsson, A.; Boulund, F.; Kristiansson, E.; Moore, E. R. B. Proteotyping Bacteria: Characterization, Differentiation and Identification of Pneumococcus and Other Species within the Mitis Group of the Genus Streptococcus by Tandem Mass Spectrometry Proteomics. PLOS ONE 2018, 13 (12), e0208804. https://doi.org/10.1371/journal.pone.0208804.
(14) Witt, N.; Andreotti, S.; Busch, A.; Neubert, K.; Reinert, K.; Tomaso, H.; Meierhofer, D. Rapid and Culture Free Identification of Francisella in Hare Carcasses by High-Resolution Tandem Mass Spectrometry Proteotyping. Front. Microbiol. 2020, 11, 636. https://doi.org/10.3389/fmicb.2020.00636.
(15) Dollman, N. L.; Griffin, J. H.; Downard, K. M. Detection, Mapping, and Proteotyping of SARS-CoV-2 Coronavirus with High Resolution Mass Spectrometry. ACS Infect. Dis. 2020, 6 (12), 3269–3276. https://doi.org/10.1021/acsinfecdis.0c00664.
(16) Lozano, C.; Kielbasa, M.; Gaillard, J.-C.; Miotello, G.; Pible, O.; Armengaud, J. Identification and Characterization of Marine Microorganisms by Tandem Mass Spectrometry Proteotyping. Microorganisms 2022, 10 (4), 719. https://doi.org/10.3390/microorganisms10040719.
(17) Hardouin, P.; Pible, O.; Marchandin, H.; Culotta, K.; Armengaud, J.; Chiron, R.; Grenga, L. Quick and Wide-Range Taxonomical Repertoire Establishment of the Cystic Fibrosis Lung Microbiota by Tandem Mass Spectrometry on Sputum Samples. Front. Microbiol. 2022, 13, 975883. https://doi.org/10.3389/fmicb.2022.975883.
(18) Bourdin, V.; Charlier, P.; Crevat, S.; Slimani, L.; Chaussain, C.; Kielbasa, M.; Pible, O.; Armengaud, J. Deep Paleoproteotyping and Microtomography Revealed No Heart Defect nor Traces of Embalming in the Cardiac Relics of Blessed Pauline Jaricot. Int. J. Mol. Sci. 2023, 24 (3), 3011. https://doi.org/10.3390/ijms24033011.
(19) Chabas, M.; Pible, O.; Armengaud, J.; Alpha-Bazin, B. Label-Free Multiplex Proteotyping of Microbial Isolates. Anal. Chem. 2023, acs.analchem.3c01975. https://doi.org/10.1021/acs.analchem.3c01975.
(20) Pible, O.; Allain, F.; Jouffret, V.; Culotta, K.; Miotello, G.; Armengaud, J. Estimating Relative Biomasses of Organisms in Microbiota Using "Phylopeptidomics." Microbiome 2020, 8 (1), 30. https://doi.org/10.1186/s40168-020-00797-x.
(21) Manadas, B.; Mendes, V. M.; English, J.; Dunn, M. J. Peptide Fractionation in Proteomics Approaches. Expert Rev. Proteomics 2010, 7 (5), 655–663. https://doi.org/10.1586/epr.10.46.
(22) Deng, W.; Sha, J.; Plath, K.; Wohlschlegel, J. A. Carboxylate-Modified Magnetic Bead (CMMB)-Based Isopropanol Gradient Peptide Fractionation (CIF) Enables Rapid and Robust Off-Line Peptide Mixture Fractionation in Bottom-Up Proteomics. Mol. Cell. Proteomics 2021, 20, 100039. https://doi.org/10.1074/mcp.RA120.002411.
(23) Herraiz, T.; Casal, V. Evaluation of Solid-Phase Extraction Procedures in Peptide Analysis. J. Chromatogr. A 1995, 708 (2), 209–221. https://doi.org/10.1016/0021-9673 (95)00388-4.
(24) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for Micro-Purification, Enrichment, Pre-Fractionation and Storage of Peptides for Proteomics Using StageTips. Nat. Protoc. 2007, 2 (8), 1896–1906. https://doi.org/10.1038/nprot.2007.261.
(25) Dimayacyac-Esleta, B. R. T.; Tsai, C.-F.; Kitata, R. B.; Lin, P.-Y.; Choong, W.-K.; Lin, T.-D.; Wang, Y.-T.; Weng, S.-H.; Yang, P.-C.; Arco, S. D.; Sung, T.-Y.; Chen, Y.-J. Rapid High-PH Reverse Phase StageTip for Sensitive Small-Scale Membrane Proteomic Profiling. Anal. Chem. 2015, 87 (24), 12016–12023. https://doi.org/10.1021/acs.analchem.5b03639.
(26) Hayoun, K.; Gouveia, D.; Grenga, L.; Pible, O.; Armengaud, J.; Alpha-Bazin, B. Evaluation of Sample Preparation Methods for Fast Proteotyping of Microorganisms by Tandem Mass Spectrometry. Front.

Microbiol. 2019, 10, 1985. https://doi.org/10.3389/fmicb.2019.01985.
(27) Trapp, J.; Almunia, C.; Gaillard, J.-C.; Pible, O.; Chaumot, A.; Geffard, O.; Armengaud, J. Proteogenomic Insights into the Core-Proteome of Female Reproductive Tissues from Crustacean Amphipods. J. Proteomics 2016, 135, 51–61. https://doi.org/10.1016/j.jprot.2015.06.017.
(28) Grenga, L.; Pible, O.; Miotello, G.; Culotta, K.; Ruat, S.; Roncato, M.; Gas, F.; Bellanger, L.; Claret, P.; Dunyach-Remy, C.; Laureillard, D.; Sotto, A.; Lavigne, J.; Armengaud, J. Taxonomical and Functional Changes in COVID -19 Faecal Microbiome Could Be Related to SARS-COV -2 Faecal Load. Environ. Microbiol. 2022, 1462-2920.16028. https://doi.org/10.1111/1462-2920.16028.
(29) Bache, N.; Geyer, P. E.; Bekker-Jensen, D. B.; Hoerning, O.; Falkenby, L.; Treit, P. V.; Doll, S.; Paron, I.; Müller, J. B.; Meier, F.; Olsen, J. V.; Vorm, O.; Mann, M. A Novel LC System Embeds Analytes in Pre-Formed Gradients for Rapid, Ultra-Robust Proteomics. Mol. Cell. Proteomics 2018, 17 (11), 2284–2296. https://doi.org/10.1074/mcp.TIR118.000853.
(30) Hosp, F.; Scheltema, R. A.; Eberl, H. C.; Kulak, N. A.; Keilhauer, E. C.; Mayr, K.; Mann, M. A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day*. Mol. Cell. Proteomics 2015, 14 (7), 2030–2041. https://doi.org/10.1074/mcp.O115.049460.
(31) Kapinusova, G.; Jani, K.; Smrhova, T.; Pajer, P.; Jarosova, I.; Suman, J.; Strejcek, M.; Uhlik, O. Culturomics of Bacteria from Radon-Saturated Water of the World's Oldest Radium Mine. Microbiol. Spectr. 2022, 10 (5), e01995-22. https://doi.org/10.1128/spectrum.01995-22.
(32) Suarez, S. Ribosomal Proteins as Biomarkers for Bacterial Identification by Mass Spectrometry in the Clinical Microbiology Laboratory. J. Microbiol. Methods 2013, 7.

Supplementary data disponibles dans la partie Annexe.

II-3 : Protéotypage flash par spectrométrie de masse en tandem : 36 secondes de signal MS/MS suffisent à permettre l'identification d'isolats microbiens

Toujours dans le but de l'amélioration du débit du protéotypage par phylopeptidomique, nous proposons dans cette partie une nouvelle méthode nommée protéotypage flash par spectrométrie de masse en tandem. Celle-ci se différencie de la méthode présentée dans les parties II-1 et II-2 par l'approche du haut-débit. Pour le protéotypage flash, l'idée est de réduire le temps d'analyse par spectrométrie de masse pour augmenter le débit. Pour cela, l'injection par infusion directe est choisie. Ce mode élimine l'étape de chromatographie liquide qui dans l'analyse LC-MS/MS injecte les peptides dans le spectromètre de masse au fur et à mesure de leur séparation chromatographique. Cela représente un réel challenge puisque la séparation permet de dé-complexifier le mélange extrêmement complexe de peptides résultant de la digestion du protéome et par ce fait d'obtenir des spectres MS/MS de qualité et donc interprétable pour l'attribution des peptides par les moteurs de recherche. Pour répondre à ce défi, un fractionnement du mélange de peptides est réalisé sur colonne SPE C18 s'éluant par centrifugation en amont de l'analyse pour simplifier le mélange de peptide. Une fraction spécifique correspondant à un pourcentage d'acétonitrile est injectée dans le spectromètre de masse. L'injection de la fraction spécifique de chaque isolat est réalisée par le système d'injection de la chromatographie liquide qui envoie directement les peptides à la source nanospray du spectromètre de masse. Une unique acquisition MS est faite pour l'enregistrement des injections. Pour le traitement des données de masse, l'analyse par phylopeptidomique a été appliquée.

Dans un premier temps la preuve de concept et l'efficacité de la méthode ont été démontrées sur des souches de référence. Les quatre souches de référence incluent *Deinococcus proteolyticus*, *Pseudomonas putida*, *Klebsiella aerogenes* ainsi que *Ruegeria pomeroyi*. Le fractionnement du mélange de peptides résultant de la digestion du protéome pour chaque souche a été réalisé sur colonne SPE C18 s'éluant par centrifugation et les fractions 21 (correspondant à 21 % d'acétonitrile) ont été injectées successivement. La souche *Pseudomonas putida* a été injectée à deux reprises. L'acquisition MS de ces 5 injections donne un profil représentant l'intensité en fonction du temps et permet l'observation de 5 pics qui correspondent chacun à une injection et donc aux 5 isolats. Chaque pic est composé de deux parties, la première correspondant à une contamination due à l'injection précédente et la seconde partie correspondant au signal associé à l'isolat injecté. La contamination résiduelle a été facilement évitée lors du traitement des données, en ne sélectionnant que le signal MS/MS enregistré à

partir de 36 secondes après l'injection. Les résultats de l'interprétation du protéotypage sur ces fenêtres de signal sont en parfait accord avec les organismes injectés, dans le bon ordre et ont donné une identification au niveau souche. Une fenêtre de signal de 36 secondes permet donc d'enregistrer suffisamment de spectres MS/MS pour permettre l'identification des souches.

Le domaine d'application ciblé pour la méthode étant le criblage de microorganismes, il était important d'évaluer cette méthode sur des échantillons d'isolats inconnus. Dix pics ont été obtenus sur le profil de l'acquisition MS unique de la séquence d'injections successives de dix isolats inconnus. Comme précédemment des fenêtres de signal MS/MS de 36 secondes ont été appliquées pour le traitement des données. Les dix identifications obtenues au niveau souche avec un nombre de TSMs et de peptides spécifiques suffisants permettent de montrer la validation de l'identification et la répétabilité de la méthode.

Cette méthode est très prometteuse puisqu'elle permet l'analyse d'un isolat en moins de 2 minutes comprenant l'injection et l'analyse de l'isolat. Pour faire suite à cette faisabilité il serait intéressant de mettre en œuvre la méthode sur un grand nombre d'isolats pour illustrer la capacité de criblage haut-débit. Cet article a été publié dans le journal « Proteomics » sous forme de « *Technical Brief* » (https://doi.org/10.1002/pmic.202300372) .

Technical Brief

Flash MS/MS proteotyping allows identifying microbial isolates in 36 seconds of mass spectrometry signal

Madisson Chabas[1,2], Jean-Charles Gaillard[1], Béatrice Alpha-Bazin[1*#], Jean Armengaud[1*#,]

[1]Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, 30200 Bagnols-sur-Cèze, France. [2]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Université de Montpellier, F-30207 Bagnols sur Cèze, France. [#]These co-authors should be considered as co-last authors. *Authors to whom correspondence should be addressed.

Keywords: Identification; Mass spectrometry; Microorganism; Proteotyping; Taxonomy

Abstract

Rapid identification of microorganisms is essential for medical diagnostics, sanitary controls, and food safety. High-throughput analytical platforms currently rely on whole-cell MALDI-TOF mass spectrometry to process hundreds of samples per day. Although this technology has become a reference method for isolates, it is unable to process most environmental isolates and opportunistic pathogens due to an incomplete experimental spectrum database. Furthermore, in most cases, its resolution is limited to the taxonomical rank of species. By recording much more sequence information at the peptide level, proteotyping by tandem mass spectrometry is able to identify the taxonomic position of any microorganism in the tree of life, and can be highly discriminating at the subspecies level. We propose here a methodology for ultra-fast identification of microorganisms by tandem mass spectrometry based on direct sample infusion and an original, highly sensitive procedure for data processing and taxonomic identification. Results obtained on reference strains and hitherto uncharacterized bacterial isolates show identification to species level in 36 seconds of tandem mass spectrometry signal, 102 seconds when including the injection procedure. Flash proteotyping is highly discriminating, as it can provide information down to strain level. The methodology is amenable to very high throughput and opens up new perspectives.

Main text

Microbial systematics aims at exploring microbial diversity through standardized characterization, classification, and naming of microorganisms [1]. This task of organizing information across the tree of life is crucial not only to microbiology, but also for microbial diagnostics. Proteotyping microorganisms involves literally typing microorganisms according to their protein or peptide signature, as previously described [2], [3]. Whole-cell MALDI-TOF mass spectrometry proteotyping, which uses mass fingerprinting of low-molecular weight, basic, and abundant proteins, is now a routine diagnostic due to its rapid analysis (~ 6 min per sample) and low cost [4]. However, several drawbacks have been enumerated, such as i) the difficulty of discriminating between closely related specific species, ii) the impossibility of discriminating between strains belonging to the same species, iii) the need to use only pure material, iv) the requirement for standardized growth condition comparable to the condition chosen to populate the spectra database, and last but not least, v) the impossibility of identifying a species that has not been yet thoroughly characterized by mass spectrometry. This method is therefore inapplicable to most opportunistic pathogens or environmental isolates that are yet to characterize. Interestingly, proteotyping by tandem mass spectrometry

addresses all these limitations. Pioneered two decades ago [5], [6], this methodology is based on the analysis of thousands of peptide sequences instead of a hundred or so molecular weights, it can discriminate numerous microorganisms within a single sample [7]. As it relies on the interrogation of a generalist database comprising all sequenced genomes for peptide sequence identification, the methodology is able to recognize the position on the tree of life at the most valuable taxonomical rank. Sample preparation for this methodology has proved suitable for 96-well plates [8]. Proteotyping based on tandem mass spectrometry was shown to be highly sensitive, requiring only 4 × 104 colony-forming units from a sample volume of 1 mL [9]. This technique has enabled the characterization of closely related strains [3], environmental isolates from diverse origins [10], [11], and even diverse microorganisms from historical relics [12]. Recently, we have shown that the yield of tandem mass spectrometry proteotyping can be greatly improved by multiplexing samples without the need of expensive chemical reagents, achieving the identification at the species level of a total of 21 microorganisms in a single 60-min mass spectrometry analysis [13]. At this stage, improving further the performance of proteotyping by tandem mass spectrometry will be very beneficial in convincing users to invest in a more expensive high-resolution instrument compared to the standard equipment of control laboratories specialized in microbiology. Sequential analysis of samples by direct infusion of peptides into the tandem mass spectrometer is a hitherto unexplored methodology that would enable higher throughput to identify the taxonomy of any microbial isolate. As the peptides produced directly from each microorganism are far too diverse (*i.e.* over 105,000 entities), this strategy cannot be successfully applied to the entire proteolyzed proteome. Indeed, the current performance of tandem mass spectrometers is not sufficient to resolve such complex samples to a sufficient degree. Indeed, proteomic characterization of microorganisms relies on peptide separation by reverse phase chromatography prior to tandem mass spectrometry. However, the conventional shotgun proteomics analysis time for microorganisms is of the order of 30 to 60 min [14], [15]. Recent advances in chromatography and tandem mass spectrometry allow shorter gradients, but the procedure requires at least a dozen min to obtain a 5 min acetonitrile gradient [16]. Direct infusion for shotgun proteomics was recently shown to be possible when the mass spectrometer is coupled to an ion mobility device that could replace the chromatography function [17], [18]. Here, we have proposed a new method using direct infusion of a specific fraction of peptides produced from each microorganism to simplify the injected pool and avoid a large number of chimeric MS/MS spectra. In addition, the associated proteotyping interpretation must be efficient to specify the species based on a restricted set of peptide sequences acquired in a minimal window time. We have exploited the previously developed proteotyping pipeline [13], [19] that was shown recently to be sensitive even with low signal [9]. Briefly, MS/MS spectra are interpreted against a generalist database, the resulting peptide sequences are assigned to all possible taxa, assigning Taxon-to Spectrum Matches (TSMs) to all taxonomic ranks, and the most likely genus present is identified based on the basis of the number of species-specific peptides and TSMs. This search is followed by a second iteration of exploration of the dataset against a database containing all possible descendants of the identified genus, in order to establish the most likely species, or even strain.

To demonstrate how the method handles known bacterial strains, we prepared the proteome trypsin digest of four microorganisms for which the genome has been sequenced and annotated previously: *Deinococcus proteolyticus MRP*, *Pseudomonas putida KT2440*, *Klebsiella aerogenes KCTC2190* and *Ruegeria pomeroyi DSS-3*. Bacteria were grown in 5 mL of liquid culture of lysogeny broth (BD Bacto, Becton Dickinson) for the first 2 strains; trypticase soy broth (TSB, Biomerieux) and Marine Broth (MB) Medium, respectively, for the other strains. Cells were harvested by centrifugation at 4000 *g*. The resulting cell pellets (39 mg in average) were lysed according to the protocol described by Hayoun et al [20] after adding 10 µL of LDS 3X (Thermo) lysis buffer per 1.7 mg of pellets. After heating, ultrasonic and bead beating steps, proteins were digested with trypsin using SP3 protocol adapted to the 96-well format. Reduction and alkylation was done by adding 4 µL of 35 mM DTT and 4 µL of 105 mM iodoacetamide solution to 20 µL of each protein solution and incubating for 10 min at room temperature. Then, 4 µL of magnetic beads (*i.e.* 200 µg) were added to the solution, before supplementation of 200 µL of acetonitrile. After incubation for 2 min under agitation at 500 rpm, beads were retained with a magnet to remove the supernatant. Then, successive washes were performed with 70% ethanol in water. A volume of 30 µL of 0.1 µg/µL trypsin gold in 50 mM NH_4HCO_3 (50mM) was added on the beads. After an incubation for 30 min at 50°C, beads were retained and the

digest peptide solution was collected and supplemented with 0.5% trifluoroacetic acid (TFA) final concentration. Peptides were fractionated using an AttractSPE Disks 96 plate C18 for microelution (Affinisep). First, SPE disks into tip shape wells were activated and conditioned using acetonitrile solution and then with water solution acidified with 0.5 % of TFA. Peptides of low hydrophobicity were washed with 20% acetonitrile in 0.1% formic acid (FA). Peptides of interest for infusion were eluted with 21 % acetonitrile, 79% water with 0.1% formic acid. An Orbitrap Exploris 480 tandem mass spectrometer (Thermo Fisher) equipped with an EasySpray ion source and connected to a Vanquish Neo UHPLC (Thermo Fisher) was used without chromatographic column or precolumn. The capillary tube was fed with 30 % acetonitrile supplemented with 0.1 % formic acid at 1 µL per min. The mass spectrometer was operated with the following parameters: dynamic exclusion of 5 seconds, Top 20 method for fragmentation retaining only precursors with charge state 2+ or 3+, and a mass tolerance of 10 ppm. In preliminary experiments, we found that the threshold for triggering precursor ion fragmentation and MS/MS measurement was an important parameter. It was set at 8.0E3 to avoid low-value MS/MS spectra. After manual start of the acquisition, small volumes (2µL) of each of the five samples were injected sequentially each 1.70 min. After 8 min, the acquisition was stopped. The raw file was converted to a mgf file with standard conversion parameters. Specific windows of this mgf file were defined corresponding to the signal from each isolate. They were used to query a NCBInr-derived database (NCBInrS) as described [21] with Mascot version 2.6.1. A second query was performed on a database comprising all known annotated genomes from NCBInr associated with the genus identified in the first-round search. A third query was performed on an even more reduced database comprising only the annotated genomes of the strains belonging to the species identified in the second query. For MS/MS spectrum-to-peptide assignment, the following parameters were applied in the first search: mass tolerance of 3 ppm on parent ion and 2 or 3 possible positive charges, mass tolerance of 0.02 Da on MS/MS signals, a maximum of one missed cleavage, carbamidomethylation of cysteine as fixed modification, oxidation of methionine as variable modification, and trypsin as proteolytic enzyme. In the second and third searches, the same parameters were used, except that mass tolerance was 5 ppm on parent ions and a maximum of two missed cleavages was allowed. Peptide sequences were mapped to taxa at the species, genus, family, order, class, phylum, and superkingdom taxonomical ranks, as previously described, resulting in Taxon-to-Spectrum Matches (TSMs) as described in Chabas et al, 2023 [13].

Figure 1 shows the total ion current profile recorded by the tandem mass spectrometer over the course of 8.5 min, together with the chronology of five injections, the injection of *P. putida* peptidome being repeated but interspaced with two other isolates. Five distinct elution phases were observed, each separated by 1.7 min. Each elution lasted 1 min and was followed by a signal-free window for 36 seconds. However, we noted that the signal during each elution did not correspond solely to the expected peptidome, since an additional signal corresponding to residual contamination from the previously injected sample was systematically found before the main pure signal corresponding to the isolate peptidome injected. This phenomenon was systematically observed and could not be circumvented due to the injection rotor, which could not be washed extensively without losing time. The signal-free window of 1 min corresponds to the minimum time required for the Vanquish Neo injection needle to be washed, loaded with a new sample, and positioned for a new injection. Importantly, if this procedure could be carried out more quickly, more samples could be processed per unit of time. Residual contamination was easily avoided during data processing, by selecting only the MS/MS signal recorded during the 36 last seconds of each eluted peak, as shown by the red rectangles in **Figure 1**. **Table 1** shows the results of proteotyping interpretation for these five samples. The first signal recorded between 78 and 114 seconds, comprising 702 MS/MS spectra, was attributed to the species *Deinococcus proteolyticus* with 356 TSMs and 104 species-specific peptides. The second signal, extending from 186 to 222 seconds and corresponding to 699 MS/MS spectra, was unambiguously assigned to *Pseudomonas putida* with 195 TSMs and 40 species-specific peptides. The third signal observed from 282 to 318 seconds with 716 MS/MS spectra was attributed to *Klebsiella aerogenes* with 407 TSMs and 104 species-specific peptides. Next, *Ruegeria pomeroyi* could be identified with 295 TSMs and 100 species-specific peptides with the 384 to 420 seconds signal comprising 708 MS/MS spectra. Finally, the last eluted peak, injections are in perfect agreement with the expected samples and their order of injection. The number of MS/MS spectra observed for each of the five interpreted windows was highly reproducible, with an average of 706 ± 7 MS/MS

spectra. The high number of species-specific peptides for each reference microorganisms, in the range 40-104, shows the high confidence obtained if the corresponding annotated genomes are present in the database used for the search. As expected, this parameter is variable due to the density of genome sequences available per species, and phylogenetic relationships between closely related species can vary considerably, influencing the number of theoretical species-specific peptides. In some cases, several strains of the same species have been genome sequenced, as for *Klebsiella aerogenes* and *Pseudomonas putida*. Here, our pipeline is able to specify which strains were used. Strains *Klebsiella aerogenes* KCTC 2190 and *Pseudomonas putida* KT2440 received the highest number of specific peptides and TSMs of all possible strains at the most resolved taxonomical rank. This shows the discriminating power of the tandem mass spectrometry proteotyping used here, even though a short window of acquisition was allowed and peptides of a given hydrophobicity were selected.

Next, flash proteotyping was applied to unknown isolates obtained after microbiota sampling from the skin of a human volunteer's arm and culture on agar plates. Ten isolates were grown in 3 mL of liquid culture of Reasoner's 2A or Tryptic Soy Broth. Cells were harvested by centrifugation. Their proteins were extracted and proteolyzed with trypsin as described above. Peptides from these ten samples were fractionated as described earlier and the eluted peptides obtained with 21 % of acetonitrile in 0.1 % FA were selected as fractions for the injection. Tandem mass spectrometer acquisition was initiated and ten fractions were sequentially injected by direct infusion. The resulting file showed a total ion current profile with ten peaks. The corresponding windows for the ten isolates were selected and interpreted individually. The first window selected, corresponding to the first isolate, starts at 78 seconds and comprised 663 MS/MS spectra. Proteotyping indicated the presence of a microorganism closely related to the *Micrococcus luteus* species with 44 TSMs and 19 species-specific peptides, and more precisely to strain KR (41 TSMs). The identification of the ten isolates (I1 to I10) is presented in Table 1 from acquisition windows starting at 78 seconds, 186 sec, 282 sec, 384 sec, 486 sec, 594 sec, 690 sec, 798 sec, 894 sec and 990 sec. Identification of the closest species was obtained for all isolates. In all, nine different species belonging to seven different genera were identified. The average number of MS/MS spectra for these ten windows was 687 ± 31, showing once again a high degree of reproducibility in terms of peptide signal. Remarkably, several isolates, such as I7 (*Staphylococcus warneri),* I5 (*Pseudomonas psychotolerans)*, I3 (*Kocuria rhizophilia)*, and I4 (*Roseomonas mucosa),* showed high numbers of TSMs and specific peptides like the reference strains used above. These high numbers indicate a close relationship between each of these isolates and the strains whose genomes are annotated and present in the database. These high numbers indicated high confidence in identification at the strain level. Other isolates, such as I1 (*Micrococcus luteus)*, I2 (*Staphylococcus equorum)*, I9 (*Microvirga ossetica)* and I10 (*Panaebacillus catalpae),* have fewer TSMs, but have at least two species-specific peptides enabling species validation. Noteworthy, isolates I7 and I8 belong to the same species (i.e. *Staphylococcus warneri*), but proteotyping allows us to be more precise at the strain level, showing that both isolates are most likely duplicates of the same strain (*Staphylococcus warneri* Lyso 1 2011). Thus, the methodology presented in this technical brief shows its potential for dereplicating identical clones that are often obtained in culturomics.

We present here a new method for identifying microbial isolates based on the direct infusion on the tandem mass spectrometer of peptides generated by trypsin proteolysis of proteins extracted from microbial isolates. The success of this approach relies on specific fractionation of the peptide pool to decomplexify the injected sample and avoid recording too many chimeric MS/MS spectra of low value. It also relies on our proteotyping pipeline, which has proved robust in several studies [10], [22], [23]. The current configuration, which uses an Orbitrap Exploris 480 tandem mass spectrometer coupled to a Vanquish neo UHPLC, records sufficient signals for species-level identification of any isolate in 36 seconds, enabling identification in 1.7 min including the time required for sample injection. As the method is simple and applicable to larger sample sequences, the theoretical identification flow is 850 isolates per day with the current set-up. It should be underlined that the Orbitrap Exploris 480 tandem mass spectrometer has a high scanning speed (20 Hz) enabling at least 600 MS/MS spectra to be recorded with our parameters, 21 of which can be assigned to the correct species in the worst-case scenario. As previously indicated [7], proteotyping of environmental isolates will be further improved over time by the inclusion of new sequenced genomes in the protein sequence database. This increase in annotated genomes will lead to greater discrimination power between strains. As demonstrated here on a limited number of

samples, the flash proteotyping method is highly reproducible. Firstly, the same time window was obtained for both reference bacteria and isolates, enabling full automation of on-the-fly interpretation. As data processing is rapid due to the low number of MS/MS spectra to be interrogated, the proteotyping response can be obtained within a couple of minutes after mass spectrometry recording. Interestingly, the methodology could be much faster than 1.7 min per isolate if using the latest generation of tandem mass spectrometers such as the ASTRAL instrument which can operate at 200 Hz [24], thus recording many more MS/MS spectra in much less time. The main bottleneck of the methodology at present is the injection procedure which, based on the Vanquish neo injection system, slows down the injection cascade. Faster washing and needle movement will definitively reduce the time required per sample. If this injection procedure cannot be optimized, further gains could be made by combining isolate multiplexing with the flash proteotyping approach. In this case, each injected peptidome would not come from a single isolate, but from an assemblage of several isolates that could be discriminated by labeling the peptides with isobaric reagents. In principle, the performance of the ASTRAL instrument enables it to record a signal sufficient to identify 16 mixed isolates per 1.7 min window, theoretically giving a flow of 13,600 identified isolates per day. This figure should be compared with previous bacterial identification works based on shotgun proteomics data, where hours of mass spectrometry per isolate were required [25]. While impressive progress has been made over the past five years, the next generation of high-resolution tandem mass spectrometers coming to market from several suppliers in 2023 should further enhance the power of our methodology.

In conclusion, we have demonstrated that the direct infusion mode can enable flash proteotyping of microorganisms, since an isolate can be identified to the taxonomical rank of species or even strain level in 36 seconds of tandem mass spectrometry signal, 1.7 min when including the injection procedure, i.e. a possible flow of 850 isolates per day. Fractionation performed on the peptide pool obtained by trypsin proteolysis prior to injection on the tandem mass spectrometer is necessary to inject a less diverse peptide pool, but this sample preparation could be fully automatized. Flash proteotyping has great potential for high-throughput analysis of isolates, making culturomics much faster and less costly.

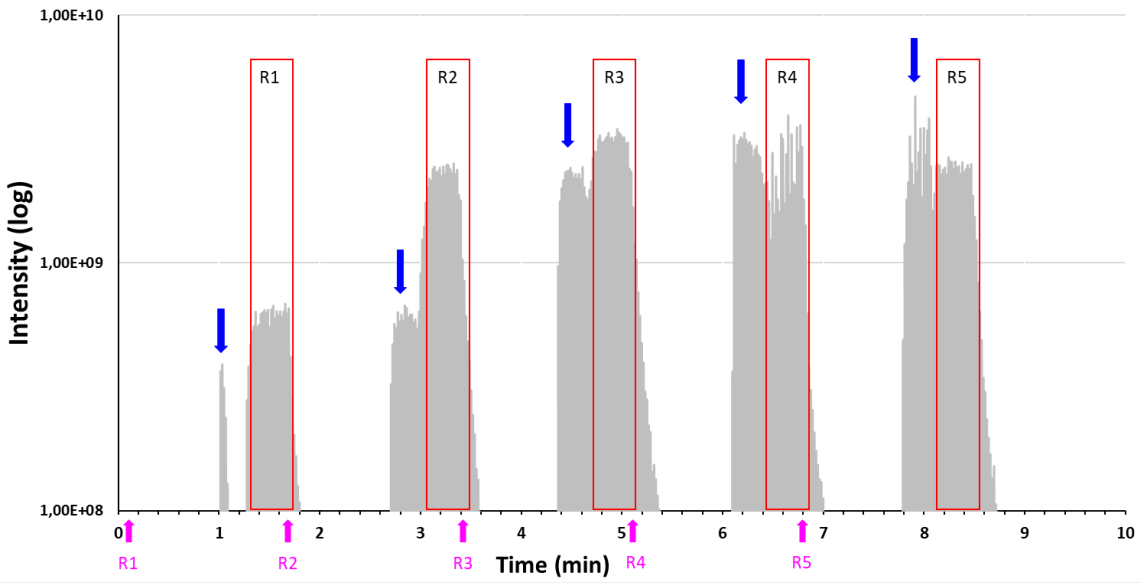


Figure 1. Total ion current profile of a sequence of five successive injections of references bacteria. Pink arrows indicate the time when each isolate was injected. Blue arrows indicate the residual contamination from the previously injected sample. Red boxes show the signal that was selected for interpretation.

Table 1. Phylopeptidomic results obtained for the ten successive injections of uncharacterized of microbial isolates

Isolate number	# MS/MS	Species identification	# TSMs	# Specific peptides	Most closely strain identification
R1	702	*Deinococcus proteolyticus*	356	104	MRP
R2	699	*Pseudomonas putida*	195	40	KT2440
R3	716	*Klebsiella aerogenes*	407	104	KCTC 2190
R4	708	*Ruegeria pomeroyi*	295	100	DSS-3
R5	705	*Pseudomonas putida*	237	83	KT2440
I1	663	*Micrococcus luteus*	44	19	KR
I2	608	*Staphylococcus equorum*	21	8	947_12
I3	702	*Kocuria rhizophila*	311	93	RF
I4	703	*Roseomonas mucosa*	242	63	AU37
I5	696	*Pseudomonas psychrotolerans*	380	121	L19
I6	690	*Pseudomonas zeshuii*	178	26	KACC 15471
I7	711	*Staphylococcus warneri*	413	106	Lyso 1 2011
I8	700	*Staphylococcus warneri*	129	32	Lyso 1 2011
I9	705	*Microvirga ossetica*	60	6	V5/3m
I10	693	*Paenibacillus catalpae*	39	2	CGMCC 1.10784

Acknowledgements

MC thanks the Région Occitanie and the CEA for supporting part of her PhD fellowship. The authors also acknowledge support from the Région Occitanie (DeepMicro grant) that contributed to the development of metaproteomics and proteotyping expertise in the research team.

Data Availability statement

The data that support the findings of this study are openly available in PRIDE at http://doi.org/10.6019/PXD045510. All mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository under the dataset identifiers PXD045510 and 10.6019/PXD04551.

References

[1] E. R. B. Moore, S. A. Mihaylova, P. Vandamme, M. I. Krichevsky, et L. Dijkshoorn, « Microbial systematics and taxonomy: relevance for a microbial commons », Res. Microbiol., vol. 161, no 6, p. 430-438, juill. 2010, doi: 10.1016/j.resmic.2010.05.007.

[2] L. Grenga, O. Pible, et J. Armengaud, « Pathogen proteotyping: A rapidly developing application of mass spectrometry to address clinical concerns », Clin. Mass Spectrom., vol. 14, p. 9-17, sept. 2019, doi: 10.1016/j.clinms.2019.04.004.

[3] R. Karlsson, « Proteotyping: Proteomic characterization, classification and identification of microorganisms – A prospectus », Syst. Appl. Microbiol., p. 12, 2015.

[4] S. Suarez et al., « Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory », J. Microbiol. Methods, vol. 94, no 3, p. 390-396, sept. 2013, doi: 10.1016/j.mimet.2013.07.021.

[5] P. A. Demirev et C. Fenselau, « Mass Spectrometry for Rapid Characterization of Microorganisms », Annu. Rev. Anal. Chem., vol. 1, no 1, p. 71-93, juill. 2008, doi: 10.1146/annurev.anchem.1.031207.112838.

[6] T. R. Sandrin et P. A. Demirev, « Characterization of microbial mixtures by mass spectrometry », Mass Spectrom. Rev., vol. 37, no 3, p. 321-349, mai 2018, doi: 10.1002/mas.21534.

[7] J. Armengaud, « Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future », Environ. Microbiol., vol. 25, no 1, p. 115-125, janv. 2023, doi: 10.1111/1462-2920.16238.

[8] K. Hayoun, J.-C. Gaillard, O. Pible, B. Alpha-Bazin, et J. Armengaud, « High-throughput proteotyping of bacterial isolates by double barrel chromatography-tandem mass spectrometry based on microplate paramagnetic beads and phylopeptidomics », J. Proteomics, vol. 226, p. 103887, août 2020, doi: 10.1016/j.jprot.2020.103887.

[9] C. Mappa, B. Alpha-Bazin, O. Pible, et J. Armengaud, « Evaluation of the Limit of Detection of Bacteria by Tandem Mass Spectrometry Proteotyping and Phylopeptidomics », Microorganisms, vol. 11, no 5, p. 1170, avr. 2023, doi: 10.3390/microorganisms11051170.

[10] K. Hayoun et al., « Proteotyping Environmental Microorganisms by Phylopeptidomics: Case Study Screening Water from a Radioactive Material Storage Pool », Microorganisms, vol. 8, no 10, p. 1525, oct. 2020, doi: 10.3390/microorganisms8101525.
[11] C. Lozano, M. Kielbasa, J.-C. Gaillard, G. Miotello, O. Pible, et J. Armengaud, « Identification and Characterization of Marine Microorganisms by Tandem Mass Spectrometry Proteotyping », Microorganisms, vol. 10, no 4, p. 719, mars 2022, doi: 10.3390/microorganisms10040719.
[12] V. Bourdin et al., « Deep Paleoproteotyping and Microtomography Revealed No Heart Defect nor Traces of Embalming in the Cardiac Relics of Blessed Pauline Jaricot », Int. J. Mol. Sci., vol. 24, no 3, p. 3011, févr. 2023, doi: 10.3390/ijms24033011.
[13] M. Chabas, O. Pible, J. Armengaud, et B. Alpha-Bazin, « Label-Free Multiplex Proteotyping of Microbial Isolates », Anal. Chem., p. acs.analchem.3c01975, août 2023, doi: 10.1021/acs.analchem.3c01975.
[14] M. Abele et al., « Unified Workflow for the Rapid and In-Depth Characterization of Bacterial Proteomes », Mol. Cell. Proteomics, vol. 22, no 8, p. 100612, août 2023, doi: 10.1016/j.mcpro.2023.100612.
[15] C. Rubiano-Labrador et al., « Proteogenomic insights into salt tolerance by a halotolerant alpha-proteobacterium isolated from an Andean saline spring », J. Proteomics, vol. 97, p. 36-47, janv. 2014, doi: 10.1016/j.jprot.2013.05.020.
[16] D. B. Bekker-Jensen et al., « A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients », Mol. Cell. Proteomics, vol. 19, no 4, p. 716-729, avr. 2020, doi: 10.1074/mcp.TIR119.001906.
[17] Y. Jiang, A. Hutton, C. W. Cranney, et J. G. Meyer, « Label-Free Quantification from Direct Infusion Shotgun Proteome Analysis (DISPA-LFQ) with CsoDIAq Software », Anal. Chem., p. acs.analchem.2c02249, déc. 2022, doi: 10.1021/acs.analchem.2c02249.
[18] J. G. Meyer, N. M. Niemi, D. J. Pagliarini, et J. J. Coon, « Quantitative shotgun proteome analysis by direct infusion », Nat. Methods, vol. 17, no 12, p. 1222-1228, déc. 2020, doi: 10.1038/s41592-020-00999-z.
[19] O. Pible, F. Allain, V. Jouffret, K. Culotta, G. Miotello, et J. Armengaud, « Estimating relative biomasses of organisms in microbiota using "phylopeptidomics" », Microbiome, vol. 8, no 1, p. 30, déc. 2020, doi: 10.1186/s40168-020-00797-x.
[20] K. Hayoun, D. Gouveia, L. Grenga, O. Pible, J. Armengaud, et B. Alpha-Bazin, « Evaluation of Sample Preparation Methods for Fast Proteotyping of Microorganisms by Tandem Mass Spectrometry », Front. Microbiol., vol. 10, p. 1985, sept. 2019, doi: 10.3389/fmicb.2019.01985.
[21] L. Grenga et al., « Taxonomical and functional changes in COVID -19 faecal microbiome could be related to SARS-COV -2 faecal load », Environ. Microbiol., vol. 24, no 9, p. 4299-4316, sept. 2022, doi: 10.1111/1462-2920.16028.
[22] H. Oumarou Hama, T. Chenal, O. Pible, G. Miotello, J. Armengaud, et M. Drancourt, « An ancient coronavirus from individuals in France, circa 16th century », Int. J. Infect. Dis., vol. 131, p. 7-12, juin 2023, doi: 10.1016/j.ijid.2023.03.019.
[23] O. Pible, P. Petit, G. Steinmetz, C. Rivasseau, et J. Armengaud, « Taxonomical composition and functional analysis of biofilms sampled from a nuclear storage pool », Front. Microbiol., vol. 14, p. 1148976, avr. 2023, doi: 10.3389/fmicb.2023.1148976.
[24] H. Stewart et al., « Parallelized Acquisition of Orbitrap and Astral Analyzers Enables High-Throughput Quantitative Analysis », Cell Biology, preprint, juin 2023. doi: 10.1101/2023.06.02.543408.
[25] L. Wöhlbrand, R. Rabus, B. Blasius, et C. Feenders, « Influence of NanoLC Column and Gradient Length as well as MS/MS Frequency and Sample Complexity on Shotgun Protein Identification of Marine Bacteria », Microb. Physiol., vol. 27, no 3, p. 199-212, 2017, doi: 10.1159/000478907.

Supplementary data disponibles sous :
https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.202300372

List of identified species and strains (Table S1); List of identified proteins and peptides_ (Table S2); and Total ion current profile of the flash proteotyping sequence of ten successive injections of isolates (Figure S1).

III. Discussions et perspectives

Certains domaines du diagnostic clinique ou encore des biotechnologies microbiennes sont en plein essor. Tous deux passent par l'identification des microorganismes. En diagnostic clinique, l'enjeu est d'identifier le microorganisme à l'origine d'une infection ou d'une pathologie afin de traiter le patient avec la thérapie la plus adaptée. Les méthodes que nous avons développées sont des méthodes qui peuvent permettre un criblage rapide, et ont donc pour vocation de révolutionner le diagnostic des agents pathogènes et des microorganismes en général. Dans cette partie Discussion et Perspectives, je propose de définir comment le protéotypage par spectrométrie de masse en tandem peut devenir une méthode de référence pour le criblage de microorganismes et être adopté par les laboratoires d'analyse.

III-1. Identifier plus rapidement les microorganismes par protéotypage par spectrométrie de masse en tandem

Le protéotypage par MALDI-TOF MS ainsi que le séquençage du gène de l'ARNr 16S sont les méthodes de référence en terme d'identification de microorganismes notamment le MALDI-TOF MS dans le domaine clinique pour le séquençage du gène de l'ARNr 16S pour la caractérisation de nouveaux organismes. Ces approches présentent cependant des limites notamment l'utilisation d'une empreinte spectrale de référence obligatoire pour permettre l'identification pour le MALDI-TOF MS rendant cette méthode utilisable principalement pour les souches pathogènes qui sont déjà bien caractérisées (Grenga et al., 2019), mais beaucoup plus difficiles et hasardeuses pour les souches pathogènes opportunistes et les souches environnementales. De plus, certaines souches proches sont difficiles à distinguer par cette méthode. Il en est de même pour le séquençage de l'amplicon d'ARNr 16S où la présence de séquences chimériques peut provoquer des erreurs d'identification (Jo et al., 2016). Enfin, ces opérations sont longues et peu pratiques à mettre en œuvre pour un diagnostic rapide. Ainsi, utiliser le protéotypage par MS/MS qui se base sur l'utilisation des peptides comme marqueurs d'identification est une alternative intéressante car plutôt rapide à mettre en œuvre et très riche en information. Cette méthode est très sensible du fait de l'utilisation des peptides et non des protéines. Typiquement, seulement quelques miliers de cellules bactériennes sont suffisantes pour une identification au niveau espèce (Mappa et al., 2023). La méthode utilise un principe de comparaison aux bases de données permettant de ne pas être limité à une empreinte mais de donner le meilleur score d'identification. Ainsi, elle permet de pouvoir discriminer des espèces très proches (Hayoun et al., 2020b; Mesuere et al., 2016a; Pible et al., 2020).

L'une des méthodologies pour augmenter le ratio d'organismes par unité de temps de spectrométrie de masse que j'ai exploré au cours de mes travaux de thèse est le multiplexage, qui consiste à analyser plusieurs échantillons mélangés en une seule étape analytique. Le marquage différentiel des échantillons est très souvent associé à ce concept de multiplexage. Toutefois, ce marquage des échantillons par l'utilisation d'isotopes non radioactifs, tel que le marquage par TMT qui peut permettre d'analyser jusqu'à seize échantillons en une analyse unique, est très coûteux : en comptant 1570 € pour un tube de 0.5 mg de marqueurs correspondant à une expérience utilisant le 16 plex et 8350 € pour un kit TMT 16 plex pour 5 mg de marqueurs correspondant à dix expériences utilisant le 16 plex. Dans le but d'augmenter le débit de la phylopeptidomique, nous avons souhaité développer une méthode de multiplexage sans marquage. Dans la partie II-1, nous avons décrit cette approche basée sur l'utilisation de fractions d'hydrophobicité différente et sur la puissance de la phylopeptidomique pour permettre l'identification de 21 organismes en une analyse unique nanoLC-MS/MS de 60 minutes. Cette méthode permet d'utiliser une fraction spécifique à un organisme et de pouvoir mélanger ces diverses fractions d'hydrophobicité différentes en un unique mélange. Ainsi lors de la séparation du mélange par nanoLC, la séparation par phase inverse permet de retrouver la fraction associée à un isolat et ainsi identifier 21 organismes en une analyse unique. L'utilisation de l'HPLC en phase inverse comme moyen de séparation au préalable de l'analyse nLC-MS/MS permet d'être dans des conditions relativement similaires à ceux de la nanoLC couplée au spectromètre de masse. Le détecteur UV présent sur l'équipement HPLC permet de suivre en temps réel le profil de quantité de chaque organisme fractionné montrant ainsi la reproductibilité des profils de chaque organisme fractionné. De plus la résolution de l'HPLC permet d'obtenir des fractions très spécifiques. Afin de rendre la méthode plus simple d'utilisation ne nécessitant pas l'équipement HPLC et l'expertise, un protocole a été mis au point utilisant des colonnes SPE C18 s'éluant par centrifugation. Du fait de la résolution diminuée des spins colonnes par rapport à l'HPLC, le nombre d'organismes par mélange a été réduit à six. Cependant le temps d'analyse nanoLC-MS/MS a également été réduit à 20 minutes car le mélange est plus simple, donnant ainsi le même ordre de temps d'analyse MS/MS par échantillon que pour l'HPLC soit 3.3 minutes au lieu de 3.0 minutes. A travers le mélange M21 obtenu par le fractionnement HPLC de 20 isolats, nous avons démontré que nous étions capables de distinguer à partir d'une simple fraction deux organismes d'un même genre notamment le genre *Pseudomonas*, *Bacillus* ou encore *Deinococcus* mais aussi la présence d'un même organisme dans deux fractions différentes avec l'espèce *Ralstonia pickettii*. Il est important de noter que dans le mélange M21 les organismes du même genre et du même organisme n'étaient pas contenus dans des fractions successives, la séparation étant d'au moins

1 fraction d'écart. Dans le cas d'organismes très proches génétiquement et qui sont très séquencés comme *Bacillus cereus* et *Bacillus thuringiensis* (une identité de séquence de 99 %) ou encore le genre *Pseudomonas*, cela peut avoir un impact sur l'identification provoquant une ambiguïté. Dans ce cas-là, la fraction pourrait être analysée seule pour confirmer l'identification et lever toute ambiguïté. Dans le second chapitre de résultats (II-2), la présence d'un même organisme dans des fractions successives a été illustrée à travers trois mélanges de 6 isolats. L'élargissement du pic après analyse du profil du SPi des mélanges donne une indication sur la présence d'un même organisme dans des fractions successives.

Les mélanges M11 et M21 présentés dans le premier article contiennent uniquement des espèces bactériennes. Dans les mélanges présentés dans le second article, les mélanges sont essentiellement constitués d'espèces bactériennes avec une levure *Saccharomyces cerevisiae* qui a été ajoutée. Cela montre qu'il est possible d'appliquer cette méthode a d'autres types d'organismes que les bactéries mais il serait intéressant de développer des mélanges avec une panoplie d'autres organismes, telles que des archées et des champignons. D'après les profils chromatographiques des digestats peptidiques obtenus avec des champignons, le profil est similaire au profil bactérien. On peut donc penser que la méthode peut s'appliquer à ce type d'organisme. Toutefois il est important de prendre en compte que la culture des champignons ou encore des archées ainsi que l'extraction des protéines et peptides sont plus compliquées que celles des bactéries.

Enfin la création d'un grand nombre de mélange réalisés à partir du protocole décrit dans le chapitre II-2 démontre la robustesse de notre méthode qui permet une identification en parfait accord avec la composition. Il est important de prendre en compte le fait que les cultures liquides ont été faites dans des volumes importants (5 mL) pour obtenir la quantité de matériel nécessaire pour le fractionnement peptidique, notamment avoir un concentré des extraits protéiques après l'étape de lyse. Il serait donc intéressant de tester ces méthodes sur des volumes de culture plus faibles et évaluer la sensibilité de ces deux méthodes et de leur gamme d'utilisation.

Les méthodes de multiplexage sans marquage développées lors de ces travaux de thèse permettent donc de pouvoir réduire le coût des analyses par nLC-MS/MS en diminuant le nombre d'analyses, la rendant plus accessible pour l'application de criblage de milliers de microorganismes. Pour une analyse par nanoLC-MS/MS d'un échantillon de 60 min, le prix est de l'ordre de 100 €. Il faut savoir que lors des analyses les pré-colonnes et les colonnes coûtent très chers et s'abîment lors de l'analyse d'un grand nombre d'échantillons. L'investissement en matériel est également particulièrement élevé, car un spectromètre de masse de type Exploris

480 par exemple coûte plus de 650 k€ à l'achat neuf, et son amortissement sur 5 ans impacte le prix de chaque échantillon. Enfin, les frais de maintenance d'un tel appareil sont également particulièrement élevés (26 k€ par an). Ainsi réduire le nombre d'analyse en ajoutant une étape de fractionnement puis de constitution de mélange permet de réduire le nombre d'analyse de 21 runs analytique à un unique run analytique. Le fractionnement ajoute forcément un coût en plus mais qui est moindre par rapport aux coûts engendrés par 21 analyses qui nécessitent un temps conséquent du banc de mesure (42 heures incluant le passage des échantillons et des blancs associés contre 2h). Si le fractionnement utilisé est réalisé par des spins colonnes, le coût de fractionnement par échantillon revient à 3,5 € par échantillon soit 21 € pour 6 échantillons. Si l'on veut par exemple identifier 100 échantillons, cela revient à faire 17 analyses MS/MS de 20 minutes, pour un coût de 2400 € contre 10000 € pour 100 analyses nanoLC-MS/MS. Notre proposition méthodologique permet donc de diminuer par 4 le coût de 100 analyses MS/MS. Pour le fractionnement HPLC, le coût de la colonne et des consommables tels les microplaques de récupération des fractions sont à prendre en compte mais est une nouvelle fois moins coûteuse que des analyses individuelles. Ce coût pourrait être réduit en automatisant ces méthodes et en négociant auprès des fournisseurs des quantités importantes de consommables.

La seconde méthodologie proposée pour augmenter le ratio d'organismes par unité de temps de spectrométrie de masse est une autre façon d'envisager le haut-débit. En effet, cette méthode a pour but d'aller vers le haut-débit en diminuant l'étape d'analyse par spectrométrie de masse. De plus en plus d'instruments et de solutions sont proposés pour diminuer cette étape comme l'utilisation d'une séparation utilisant des pompes à haute et basse pression type Evosep. Ici nous présentons une méthode qui s'affranchit de l'étape de chromatographie liquide et qui permet une injection directe de l'échantillon dans le spectromètre de masse sans séparation. L'enjeu ici était de voir si le signal généré par le mode infusion directe était interprétable par la phylopeptidomique pour donner une identification. La performance de l'Orbitrap Exploris 480 utilisé ici pour cette méthode a permis de donner des spectres MS/MS de qualité donnant suffisament de signal interprétable pour valider les espèces et les souches.

L'apport du fractionnement est primordial pour apporter une décomplexification de l'échantillon et ainsi donner une confiance concernant le résultat d'identification. Injecter un protéome entier par infusion directe dans le but d'identification de microorganismes peut non seulement endommager le spectromètre de masse, générer des spectres MS/MS non interprétables mais aussi engendrer de trop grosses contaminations entre l'injection des isolats successifs. Une nouvelle fois les côuts de cette méthode en comparaison à une analyse classique de protéotypage par spectrométrie de masse en tandem sont réduits. Le fait de supprimer l'étape

de chromatographie liquide permet de pas acheter de colonnes de séparation qui sont très coûteuses et de gagner du temps de machine puisque aucun lavage n'est nécessaire. Le côut du fractionnement est le même que pour la première méthode et permet donc de largement réduire les coûts et les temps d'analyse puisque la puissance d'analyse de 1.7 minutes par échantillon permet une analyse de 840 échantillons sur 24 heures par rapport à 24 échantillons en analyse classique où les gradients sont de 30 à 60 minutes pour le protéotypage comprenant également les étapes de lavages et réequilibration.

De plus l'injection automatisée que nous avons pu mettre en place a permis d'adapter ce format au haut-débit, automatisable et ainsi réduire le temps d'analyse pour chaque échantillon.

La sélection de fenêtres MS/MS de 36 secondes pour palier aux contaminations des échantillons précédents est une réelle prouesse sur cette analyse puisqu'elle permet d'être spécifique et sensible. De plus l'avantage d'avoir une petite fenêtre de spectres MS/MS réduit le temps d'analyse des données MS/MS par le logiciel µorgID en permettant une analyse de chaque fichier en 10 minutes. Ainsi en moins de 12 minutes nous sommes capable d'acquérir le signal et d'avoir un résultat d'identification.

Il est important de prendre en compte dans cette méthode que des culots de l'ordre de quelques dizaines de mg ont été utilisés et lysés de manière à avoir une plus grande quantité de protéines puis de peptides après digestion. La concentration obtenue au niveau des digestats avant fractionnement a un effet sur le fractionnement qui est ensuite observé lors de l'interprétation des données MS/MS. En effet, il y a une corrélation entre la diminution du nombre de TSMs et de peptides spécifiques et la faible concentration d'un digestat utilisé pour le fractionnement. Il est donc important d'optimiser ce facteur puisque dans le cas d'un criblage, les organismes ne poussent pas tous à la même vitesse, peuvent avoir des concentrations plus faibles. Ainsi, tester les limites de détection de la quantité de peptides nécessaires à l'utilisation de la méthode est nécessaire.

Les deux méthodes développées dans le cadre de ma thèse présentent chacune des avantages permettant de choisir la plus adaptée à chaque sujet.

III-2- Perspectives d'améliorations de la phylopeptidomique pour le criblage haut-débit

A travers les méthodes développées durant mes travaux de thèse qui reposent sur deux principes différents, nous avons démontré qu'il était possible d'améliorer le débit d'une analyse classique de phylopeptidomique (lyse, digestion, analyse nLC-MS/MS) pour l'adapter à des applications d'identification de microorganismes lors de criblages. Il reste néanmoins un certain nombre de

points qui permettrait d'améliorer encore le débit des analyses pour les rendre encore plus compétitives en terme d'identification par rapport à des méthodes telles que le MALDI-TOF MS. L'automatisation de la préparation des échantillons, l'acquisition des données de masse, les bases de données ou encore les performances de nouveaux appareils de masse sont autant de points dont l'amélioration peut grandement augmenter le débit des analyses et leurs performances en terme de sensibilité et de pouvoir discriminant.

Dans les travaux précédent ma thèse, les étapes de lyse par broyage mécanique ainsi que la digestion SP3 ont contribuées à réduire le temps de préparation d'échantillons mais aussi de rendre les méthodes de lyse adaptées à divers organismes tels les bactéries (Gram – et Gram +), champignons, levures. L'utilisation de la méthode SP3 sur plaques 96 puits ainsi que l'utilisation de deux colonnes en parallèle pour l'analyse nLC-MS/MS (Hayoun et al., 2020a) visaient à tendre vers des formats haut-débit pour réduire les temps de préparation des échantillons et ainsi adapter ces analyses à l'analyse de multiples échantillons. Les travaux que j'ai pu effectuer durant ma thèse étaient une suite logique après l'optimisation des protocoles de préparation des échantillons puisque l'analyse nLC-MS/MS est une étape coûteuse en temps et en argent. Ces travaux ont donc permis de démontrer les preuves de concepts des diverses méthodes. Il est à présent indispensable de pouvoir automatiser toutes ces méthodes de la préparation des échantillons à l'analyse nLC-MS/MS pour pouvoir prétendre à être compétitif et à devenir une méthode de référence pour le protéotypage par MS/MS. En effet, les préparations des échantillons ont été réalisées au format tube pour l'obtention des culots cellulaires et les étapes de lyse. Ce format présente une certaine contrainte puisqu'il rajoute une manipulation par l'expérimentateur, peut provoquer une perte d'échantillons lors des transferts de tubes, peut être une source d'erreurs de manipulations, mais est surtout limitant quant au nombre de tubes possibles insérables dans certains appareils (centrifugeuse, Precellys (24 tubes), …). Afin d'analyser des milliers d'échantillons, le format tube individuel n'est pas adapté. L'utilisation de plaques 96 puits pour la culture liquide de microorganismes ou le recueil de colonies permet une automatisation des étapes de centrifugation et d'élimination du surnageant par exemple pour la récolte et l'obtention des culots microbiens. L'utilisation d'un automate peut réduire drastiquement les efforts au niveau de cette étape, en limitant non seulement la manipulation et par conséquent les contaminations mais aussi en permettant la centrifugation de 96 échantillons en même temps en comparaison avec 24 échantillons en format tube. Un appareil qui permettrait la centrifugation de 4 plaques 96 puits en une fois permettrait l'obtention de 384 culots en une dizaine de minutes. Il en est de même pour la lyse. Nous utilisons actuellement le protocole développé par Hayoun et al., 2019 (Hayoun et al., 2019) qui a montré ses capacités en terme d'efficacité sur tous types d'organismes. Cependant

ce format a été réalisé en tube empêchant ainsi de tendre vers le haut-débit. L'étape de broyage mécanique réalisé par l'appareil Precellys n'était possible qu'en tube. En revanche depuis peu de temps, la possibilité d'utiliser des barrettes de tubes (type PCR) ont été créé pour s'adapter à la lyse sur Precellys permettant ainsi l'analyse de 96 échantillons en même temps. L'utilisation d'une lyse chimique par l'acide trifluoroacétique (TFA) peut être utilisée et permet l'utilisation d'un format 96 puits donc possiblement automatisable (Doellinger et al., 2020). La digestion protéolytique par billes magnétiques SP3 adaptée au format 96 puits a déjà été réalisée par Hayoun *et al* et appliquée durant tous mes travaux de thèse (Hayoun et al., 2020a). Cette méthode est donc très rapide et montre une très bonne efficacité puisqu'elle affiche de bons rendements d'obtention de peptides. En revanche, utiliser un automate qui réaliserait toutes les étapes de digestion limiterait une nouvelle fois les contaminations et réduirait le temps de manipulation. De plus il permet un traitement uniforme puisque les 96 puits sont traités en parallèle en comparaison avec un expérimentateur qui effectue les étapes puits par puits. L'appareil BRAVO développé par Agilent est une plateforme de manipulation de liquides permettant un traitement des échantillons de l'extraction des protéines à l'obtention des peptides par SP3 sous format microplaque (Müller et al., 2018). L'ajout de la méthode de fractionnement peut également être automatisée. Le fractionnement HPLC lors de mes travaux de thèse a été réalisé avec un format microplaque à l'aide d'un collecteur de fractions pour HPLC, permettant d'automatiser le fractionnement. Le fractionnement par spin colonne, quant à lui, se déroulait au format tube. Cependant il existe des formats de fractionnement en microplaque 96 puits développés par Affinisep qui ont été testés au laboratoire et utilisées notamment pour les travaux sur le protéotypage flash et montrent des performances similaires à celles obtenues en format tube. Créer un automate permettant de regrouper toutes ces étapes pourrait grandement améliorer le débit de préparation des échantillons et ainsi être couplé à notre méthode d'analyse MS/MS développées au cours de mes travaux de thèse. Cette automate qui comprendrait différents compartiments pourrait être comme décrit sur la figure 22.

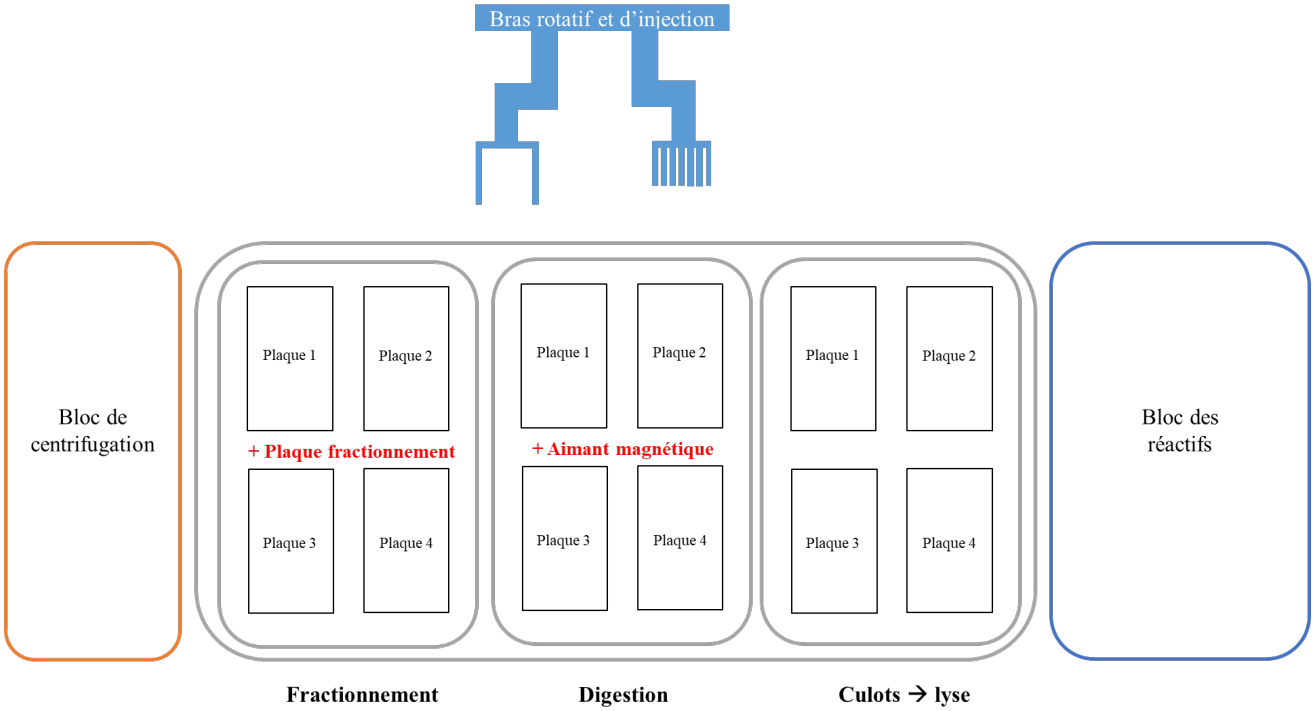


Figure 22 : Schéma de la composition des modules pour l'automatisation complète de la méthode de phylopeptidomique haut-débit

Un automate permettrait de réaliser toutes les étapes de préparation des échantillons de l'obtention des culots au fractionnement. L'automate serait constitué de plusieurs blocs, un contenant les différents réactifs nécessaires à chaque étape de préparation ainsi qu'un bloc de centrifugation. 4 plaques 96 puits pourraient être préparés en parralèle sur un bloc pour chaque étape.

Un des points importants à prendre en compte est l'actualisation des bases de données. En effet, les bases de données (cf. paragraphe 4.1.4) sont sans cesse enrichies au fil du temps. Il est donc important de réactualiser fréquemment ces bases de données pour limiter toute ambiguïté. Toutefois, le nombre de génomes séquencés augmentant, de moins en moins de séquences peptidiques sont directement spécifiques d'un microorganisme donné. Cet ajout de nouveaux organismes séquencés dans les bases de données peut générer parfois des erreurs d'identifications selon la méthode. Toutefois, la phylopeptidomique prennant en compte les nombres de peptides partagés et de peptides spécifiques pour l'identification finale permet d'être insensible à ce problème. L'essentiel est désormais d'améliorer la phylogénie des microorganismes et leur dénomination. Par exemple, certaines bactéries du genre *Pseudomonas* et du genre *Raoultella* sont trop proches pour être facilement distinguées, et une analyse profonde de la taxonomie microbienne est alors nécessaire.

Les performances des spectromètres de masse en tandem peuvent différer considérablement et ont donc un réel impact sur la rapidité et la sensibilité d'une analyse de protéotypage. Cet aspect

est abordé ici à travers les travaux réalisés sur deux appareils différents. La méthode de multiplexage a été réalisé en utilisant un spectromètre de masse haute résolution appelé Orbitrap Q-exactive HF. Cet appareil hybride très haute résolution est basée sur un analyseur Orbitrap à haut-champ et est suffisamment sensible pour générer des spectres interprétables par le pipeline utilisé. Les paramètres ont été fixés avec une méthode dite « Top 20 » sélectionnant suite à un scan MS les 20 ions précurseurs les plus abondants ainsi qu'un seuil d'intensité de 8^e10^3 et une dynamique d'exclusion de 60 ms afin de réaliser la fragmentation et les scans MS/MS. Ces paramètres peuvent être optimisés pour essayer d'avoir plus de signal interprétable et augmenter encore la rapidité des analyses. Il serait intéressant de tester les mélanges produits lors de mes travaux de thèse en les utilisant comme standards sur différents bancs de mesure et explorer les capacités des différents spectromètres de masse disponibles actuellement sur le marché. L'Orbitrap Exploris 480, qui est un spectromètre de masse plus récent que le Q-exactive HF, offre une vitesse de balayage plus importante que le Q-exactive HF engendrant ainsi l'acquisition d'un plus grand nombre de spectres MS/MS en un temps réduit par rapport au Q-exactive HF. Ainsi le nombre de spectres MS/MS acquis en 60 minutes par le Q-exactive HF est le même que celui acquis par l'Exploris 480 en 30 minutes. Le choix du spectromètre de masse est donc primordial en fonction du type d'information recherché. Par exemple, pour la méthode dite protéotypage flash, l'utilisation de l'Exploris était indispensable du fait de sa plus haute sensibilité et de sa rapidité pour traiter des échantillons n'étant pas séparés au préalable par chromatographie liquide. Récemment, Thermo Fisher Scientific a développé un nouvel appareil de masse nommé Orbitrap Astral permettant une rapidité d'analyse inégalée par rapport aux spectromètres précédents de la marque tel l'Orbitrap Exploris 480 ou encore l'Orbitrap Q-exactive HF. En effet il possède trois analyseurs de masse, un quadripole pour une grande sélectivité et une transmission élevée des ions, un Orbitrap pour une gamme dynamique élevée et des mesures à haute résolution, et le nouvel analyseur Astral pour des mesures rapides et sensibles. Ce nouvel appareil permet l'analyse de protéomes sur des gradients de 8 minutes sans pour autant perdre d'informations puisque le nombre de protéines identifiées est de l'ordre du nombre obtenus sur l'Orbitrap Exploris 480 ou encore sur l'Orbitrap Q-exactive HF sur des gradients plus longs. Ceci est dû aux différents paramètres améliorés notamment la résolution en masse, les vitesses de balayage et surtout le fait d'avoir un Top 60 permettant ainsi de sélectionner 60 peptides abondants contre 20 ou 40 pour les autres appareils dans un temps d'analyse équivalent.

Il semble raisonnable de penser qu'avec une méthode dite « Top 60 » au lieu d'un « Top20 », nous puissions obtenir en 8 min l'identification de 21 organismes, au lieu de 60 min, soit un débit d'analyse de 20 sec par isolat. Dans le cas du protéotypage flash, le pouvoir résolutif de

l'Orbitrap Astral permettrait d'identifier des échantillons mutiplexés, soit un flot d'analyse supérieur à celui obtenu ici (13 sec par isolat). Un tel spectromètre de masse combiné avec les méthodes développées lors de mes travaux de thèse pourrait donc révolutionner le diagnostic microbiologique, en proposant un débit incroyable d'analyse.

Utiliser un spectromètre de masse plus performant peut donc avoir un réel impact sur la sensibilité et la rapidité d'analyse. Le traitement des données générés par ces appareils peut également être amélioré en intervenant notamment sur le mode d'acquisition des données en passant du mode DDA au mode DIA qui est de plus en plus utilisé dans la littérature et qui est d'ailleurs utilisé pour démontrer les performances de l'Orbitrap Astral sur des gradients de 8 min. En effet, du fait que le mode DIA considère tous les peptides y compris les moins abondants, peut apporter des informations supplémentaires en terme d'identification. De plus, des logiciels tel que Chimerys sont aussi en développement (Thermo Fisher Scientific). Ces algorithmes basés sur l'intelligence artificielle sont notamment capables de déconvoluer les spectres très complexes obtenus avec des peptides coélués. Leur utilisation pourrait énormément aider à l'interprétation des spectres chimères comme par exemple avec la méthode Protéotypage Flash et permettre d'augmenter les identifications.

IV. Conclusion

Ce travail de thèse s'est concentré sur l'élaboration de méthodes ayant pour but d'augmentet le débit du protéotypage par spectrométrie de masse en tandem par la phylopeptidomique.

Deux méthodes ont été développées intégrant deux approches d'augmentation du débit de l'analyse de spectrométrie de masse en tandem : i) le multiplexage des échantillons et ii) la diminution du temps d'analyse. Le fractionnement des mélanges de peptides par hydrophobicité a été central dans ces travaux.

Ainsi la première méthode décrit une méthode de multiplexage sans marquage pour le protéotypage de plusieurs échantillons en une analyse unique de spectrométrie de masse en tandem. L'étape de fractionnement permet l'obtention de fractions de peptides d'hydrophobicité différente. C'est le mélange unique de plusieurs fractions correspondant chacune à un isolat différent qui est analysé par nLC-MS/MS. La première version de cette méthode (partie II-1) utilise le fractionnement par hydrophicité réalisé à l'aide d'une HPLC. Avec cette méthode, 21 isolats ont été analysés en une analyse unique MS de 60 minutes correspondant à une analyse MS de 3 minutes par échantillon. La méthode de multiplexage sans marquage a été simplifiée (partie II-2) de façon à être utilisable par quiconque et nécéssitant peu de matériel. Cette simplication repose sur un fractionnement par colonne SPE C18 s'éluant par centrifugation. Différents mélanges de six fractions analysés chacun en une unique analyse MS de 20 minutes illustrent la robustesse, la reproductibilité et la sensibilité de la méthode.

La seconde approche, le protéotypage flash par phylopeptidomique, se base sur la diminution du temps d'analyse en utilisant le mode d'injection directe, c'est-à-dire en supprimant l'étape de séparation par nLC. Le fractionnement est cette fois utilisé pour dé-complexifier le mélange très complexe de peptides résultant de la digestion du protéome de l'échantillon pour n'injecter qu'une fraction du protéome. Cette méthode implique une configuration pour le mode par infusion directe où l'injection est directement relié à la source ESI spray et non à une colonne. Nous avons démontré que la méthode était suffisamment sensible pour l'identification d'isolats microbiens par infusion nécéssitant seulement l'utilisation de 36 secondes de signal de spectrométrie de masse.

Ainsi mes travaux ont contribué à démontrer que la phylopeptidomique est adaptable à des analyses de protéotypage haut-débit pour des projets de criblage microbiens par exemple et pouvant être applicable à tout type de microorganismes.

Références

Akarsu, H., Bordes, P., Mansour, M., Bigot, D.-J., Genevaux, P., Falquet, L., 2019. TASmania: A bacterial Toxin-Antitoxin Systems database. PLOS Comput. Biol. 15, e1006946. https://doi.org/10.1371/journal.pcbi.1006946

Alauzet, C., n.d. Taxonomie des bacteries anaerobies : De la reclassification à la decouverte de nouveaux pathogenes. Nacy-Université, 2009.

Al-blooshi, S.Y., Latif, M.A.A., Sabaneh, N.K., Mgaogao, M., Hossain, A., 2021. Development of a novel selective medium for culture of Gram-negative bacteria. BMC Res. Notes 14, 211. https://doi.org/10.1186/s13104-021-05628-2

Altschul, S., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Alves, G., Ogurtsov, A., Karlsson, R., Jaén-Luchoro, D., Piñeiro-Iglesias, B., Salvà-Serra, F., Andersson, B., Moore, E.R.B., Yu, Y.-K., 2022. Identification of Antibiotic Resistance Proteins via MiCId's Augmented Workflow. A Mass Spectrometry-Based Proteomics Approach. J. Am. Soc. Mass Spectrom. 33, 917–931. https://doi.org/10.1021/jasms.1c00347

Alves, G., Wang, G., Ogurtsov, A.Y., Drake, S.K., Gucek, M., Sacks, D.B., Yu, Y.-K., 2018. Rapid Classification and Identification of Multiple Microorganisms with Accurate Statistical Significance via High-Resolution Tandem Mass Spectrometry. J. Am. Soc. Mass Spectrom. 29, 1721–1737. https://doi.org/10.1007/s13361-018-1986-y

Antonios, K., Croxatto, A., Culbreath, K., 2021. Current State of Laboratory Automation in Clinical Microbiology Laboratory. Clin. Chem. 68, 99–114. https://doi.org/10.1093/clinchem/hvab242

Armengaud, J., 2023. Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future. Environ. Microbiol. 25, 115–125. https://doi.org/10.1111/1462-2920.16238

Armengaud, J., 2022. Metaproteomics to understand how microbiota function: the crystal ball predicts a promising future. Environ. Microbiol. 1462-2920.16238. https://doi.org/10.1111/1462-2920.16238

Arul, A.B., Robinson, R.A.S., 2019. Sample Multiplexing Strategies in Quantitative Proteomics. Anal. Chem. 91, 178–189. https://doi.org/10.1021/acs.analchem.8b05626

Arushothy, R., Amran, F., Samsuddin, N., Ahmad, N., Nathan, S., 2020. Multi locus sequence typing of clinical Burkholderia pseudomallei isolates from Malaysia. PLoS Negl. Trop. Dis. 14, e0008979. https://doi.org/10.1371/journal.pntd.0008979

Austin, B., 2017. The value of cultures to modern microbiology. Antonie Van Leeuwenhoek 110, 1247–1256. https://doi.org/10.1007/s10482-017-0840-8

Bache, N., Geyer, P.E., Bekker-Jensen, D.B., Hoerning, O., Falkenby, L., Treit, P.V., Doll, S., Paron, I., Müller, J.B., Meier, F., Olsen, J.V., Vorm, O., Mann, M., 2018a. A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. Mol. Cell. Proteomics 17, 2284–2296. https://doi.org/10.1074/mcp.TIR118.000853

Bache, N., Geyer, P.E., Bekker-Jensen, D.B., Hoerning, O., Falkenby, L., Treit, P.V., Doll, S., Paron, I., Müller, J.B., Meier, F., Olsen, J.V., Vorm, O., Mann, M., 2018b. A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. Mol. Cell. Proteomics 17, 2284–2296. https://doi.org/10.1074/mcp.TIR118.000853

Balvočiūtė, M., Huson, D.H., 2017. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics 18, 114. https://doi.org/10.1186/s12864-017-3501-4

Bekker-Jensen, D.B., Martínez-Val, A., Steigerwald, S., Rüther, P., Fort, K.L., Arrey, T.N., Harder, A., Makarov, A., Olsen, J.V., 2020. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. Mol. Cell. Proteomics 19, 716–729. https://doi.org/10.1074/mcp.TIR119.001906

Bilen, M., 2020. Strategies and advancements in human microbiome description and the importance of culturomics. Microb. Pathog. 149, 104460. https://doi.org/10.1016/j.micpath.2020.104460

Bilen, M., Dufour, J.-C., Lagier, J.-C., Cadoret, F., Daoud, Z., Dubourg, G., Raoult, D., 2018. The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. Microbiome 6, 94. https://doi.org/10.1186/s40168-018-0485-5

Bizzini, A., Greub, G., 2010. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. Clin. Microbiol. Infect. 16, 1614–1619. https://doi.org/10.1111/j.1469-0691.2010.03311.x

Bjornson, R.D., Carriero, N.J., Colangelo, C., Shifman, M., Cheung, K.-H., Miller, P.L., Williams, K., 2008. X!!Tandem, an Improved Method for Running X!Tandem in Parallel on Collections of Commodity Computers. J. Proteome Res. 7, 293–299. https://doi.org/10.1021/pr0701198

Blevins, S.M., Bronze, M.S., 2010. Robert Koch and the 'golden age' of bacteriology. Int. J. Infect. Dis. 14, e744–e751. https://doi.org/10.1016/j.ijid.2009.12.003
Boers, S.A., Hays, J.P., Jansen, R., 2015. Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. Sci. Rep. 5, 14181. https://doi.org/10.1038/srep14181
Bohlin, J., Eldholm, V., Pettersson, J.H.O., Brynildsrud, O., Snipen, L., 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genomics 18, 151. https://doi.org/10.1186/s12864-017-3543-7
Bonnet, M., Lagier, J.C., Raoult, D., Khelaifia, S., 2020. Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. New Microbes New Infect. 34, 100622. https://doi.org/10.1016/j.nmni.2019.100622
Boulund, F., Karlsson, R., Gonzales-Siles, L., Johnning, A., Karami, N., AL-Bayati, O., Åhrén, C., Moore, E.R.B., Kristiansson, E., 2017. Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics. Mol. Cell. Proteomics 16, 1052–1063. https://doi.org/10.1074/mcp.M116.061721
Bourdin, V., Charlier, P., Crevat, S., Slimani, L., Chaussain, C., Kielbasa, M., Pible, O., Armengaud, J., 2023. Deep Paleoproteotyping and Microtomography Revealed No Heart Defect nor Traces of Embalming in the Cardiac Relics of Blessed Pauline Jaricot. Int. J. Mol. Sci. 24, 3011. https://doi.org/10.3390/ijms24033011
Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B., Raoult, D., 2010. Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. PLoS ONE 5, e15530. https://doi.org/10.1371/journal.pone.0015530
Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., Soni, G.V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., Schloss, J.A., 2008. The potential and challenges of nanopore sequencing. Nat. Biotechnol. 26, 1146–1153. https://doi.org/10.1038/nbt.1495
Brodbelt, J.S., 2016. Ion Activation Methods for Peptides and Proteins. Anal. Chem. 88, 30–51. https://doi.org/10.1021/acs.analchem.5b04563
Carbo, E.C., Blankenspoor, I., Goeman, J.J., Kroes, A.C.M., Claas, E.C.J., De Vries, J.J.C., 2021. Viral metagenomic sequencing in the diagnosis of meningoencephalitis: a review of technical advances and diagnostic yield. Expert Rev. Mol. Diagn. 21, 1139–1146. https://doi.org/10.1080/14737159.2021.1985467
Carbonnelle, É., Nassif, X., 2011. Utilisation en routine du MALDI-TOF-MS pour l'identification des pathogènes en microbiologie médicale. médecine/sciences 27, 882–888. https://doi.org/10.1051/medsci/20112710017
Chabas, M., Pible, O., Armengaud, J., Alpha-Bazin, B., 2023. Label-Free Multiplex Proteotyping of Microbial Isolates. Anal. Chem. acs.analchem.3c01975. https://doi.org/10.1021/acs.analchem.3c01975
Chen, W., Adhikari, S., Chen, L., Lin, L., Li, H., Luo, S., Yang, P., Tian, R., 2017. 3D-SISPROT: A simple and integrated spintip-based protein digestion and three-dimensional peptide fractionation technology for deep proteome profiling. J. Chromatogr. A 1498, 207–214. https://doi.org/10.1016/j.chroma.2017.01.033
Chen, W., Wang, S., Adhikari, S., Deng, Z., Wang, L., Chen, L., Ke, M., Yang, P., Tian, R., 2016. Simple and Integrated Spintip-Based Technology Applied for Deep Proteome Profiling. Anal. Chem. 88, 4864–4871. https://doi.org/10.1021/acs.analchem.6b00631
Chen, X., Wei, S., Ji, Y., Guo, X., Yang, F., 2015. Quantitative proteomics using SILAC: Principles, applications, and developments. PROTEOMICS 15, 3175–3192. https://doi.org/10.1002/pmic.201500108
Chenau, J., Fenaille, F., Ezan, E., Morel, N., Lamourette, P., Goossens, P.L., Becher, F., 2011. Sensitive Detection of Bacillus anthracis Spores by Immunocapture and Liquid Chromatography–Tandem Mass Spectrometry. Anal. Chem. 83, 8675–8682. https://doi.org/10.1021/ac2020992
Christie-Oleza, J.A., Maria Piña-Villalonga, J., Guerin, P., Miotello, G., Bosch, R., Nogales, B., Armengaud, J., 2013. Shotgun nanoLC-MS/MS proteogenomics to document MALDI-TOF biomarkers for screening new members of the Ruegeria genus. Environ. Microbiol.
Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science 311, 1283–1287. https://doi.org/10.1126/science.1123061

Cody, R.B., McAlpin, C.R., Cox, C.R., Jensen, K.R., Voorhees, K.J., 2015. Identification of bacteria by fatty acid profiling with direct analysis in real time mass spectrometry: Bacteria identification by DART fatty acid profiling. Rapid Commun. Mass Spectrom. 29, 2007–2012. https://doi.org/10.1002/rcm.7309

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., Tiedje, J.M., 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141–D145. https://doi.org/10.1093/nar/gkn879

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., Mann, M., 2011. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J. Proteome Res. 10, 1794–1805. https://doi.org/10.1021/pr101065j

Dagley, L.F., Infusini, G., Larsen, R.H., Sandow, J.J., Webb, A.I., 2019. Universal Solid-Phase Protein Preparation (USP ^{3}) for Bottom-up and Top-down Proteomics. J. Proteome Res. 18, 2915–2924. https://doi.org/10.1021/acs.jproteome.9b00217

Darwin, C., 1869. L'origine des espèces.

De Candolle, A., 1815. Théorie élémentaire de la botanique.

de Carvalho, C., Caramujo, M., 2018. The Various Roles of Fatty Acids. Molecules 23, 2583. https://doi.org/10.3390/molecules23102583

Demichev, V., Szyrwiel, L., Yu, F., Teo, G.C., Rosenberger, G., Niewienda, A., Ludwig, D., Decker, J., Kaspar-Schoenefeld, S., Lilley, K.S., Mülleder, M., Nesvizhskii, A.I., Ralser, M., 2022. dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. Nat. Commun. 13, 3944. https://doi.org/10.1038/s41467-022-31492-0

Deng, W., Sha, J., Plath, K., Wohlschlegel, J.A., 2021. Carboxylate-Modified Magnetic Bead (CMMB)-Based Isopropanol Gradient Peptide Fractionation (CIF) Enables Rapid and Robust Off-Line Peptide Mixture Fractionation in Bottom-Up Proteomics. Mol. Cell. Proteomics 20, 100039. https://doi.org/10.1074/mcp.RA120.002411

Dhiman, N., Hall, L., Wohlfiel, S.L., Buckwalter, S.P., Wengenack, N.L., 2011. Performance and Cost Analysis of Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry for Routine Identification of Yeast. J. Clin. Microbiol. 49, 1614–1616. https://doi.org/10.1128/JCM.02381-10

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F., Notredame, C., 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 39, W13–W17. https://doi.org/10.1093/nar/gkr245

Diakite, A., Dubourg, G., Dione, N., Afouda, P., Bellali, S., Ngom, I.I., Valles, C., Tall, M. lamine, Lagier, J.-C., Raoult, D., 2020. Optimization and standardization of the culturomics technique for human microbiome exploration. Sci. Rep. 10, 9674. https://doi.org/10.1038/s41598-020-66738-8

Diament, B.J., Noble, W.S., 2011. Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra. J. Proteome Res. 10, 3871–3879. https://doi.org/10.1021/pr101196n

Dimayacyac-Esleta, B.R.T., Tsai, C.-F., Kitata, R.B., Lin, P.-Y., Choong, W.-K., Lin, T.-D., Wang, Y.-T., Weng, S.-H., Yang, P.-C., Arco, S.D., Sung, T.-Y., Chen, Y.-J., 2015. Rapid High-pH Reverse Phase StageTip for Sensitive Small-Scale Membrane Proteomic Profiling. Anal. Chem. 87, 12016–12023. https://doi.org/10.1021/acs.analchem.5b03639

Do, C.B., Mahabhashyam, M.S.P., Brudno, M., Batzoglou, S., 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15, 330–340. https://doi.org/10.1101/gr.2821705

Doellinger, J., Schneider, A., Hoeller, M., Lasch, P., 2020. Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-free Protocol for Proteomics Based on Acid Extraction. Mol. Cell. Proteomics 19, 209–222. https://doi.org/10.1074/mcp.TIR119.001616

Doern, G.V., n.d. Detection of Selected Fastidious Bacteria.

Dollman, N.L., Griffin, J.H., Downard, K.M., 2020. Detection, Mapping, and Proteotyping of SARS-CoV-2 Coronavirus with High Resolution Mass Spectrometry. ACS Infect. Dis. 6, 3269–3276. https://doi.org/10.1021/acsinfecdis.0c00664

Dupré, M., Duchateau, M., Malosse, C., Borges-Lima, D., Calvaresi, V., Podglajen, I., Clermont, D., Rey, M., Chamot-Rooke, J., 2021. Optimization of a Top-Down Proteomics Platform for

Closely Related Pathogenic Bacterial Discrimination. J. Proteome Res. 20, 202–211. https://doi.org/10.1021/acs.jproteome.0c00351
Dupree, E.J., Jayathirtha, M., Yorkey, H., Mihasan, M., Petre, B.A., Darie, C.C., 2020. A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. Proteomes 8, 14. https://doi.org/10.3390/proteomes8030014
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340
El-Aneed, A., Cohen, A., Banoub, J., 2009. Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. Appl. Spectrosc. Rev. 44, 210–230. https://doi.org/10.1080/05704920902717872
Elias, J.E., Gygi, S.P., 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214. https://doi.org/10.1038/nmeth1019
Eliuk, S., Makarov, A., 2015. Evolution of Orbitrap Mass Spectrometry Instrumentation. Annu. Rev. Anal. Chem. 8, 61–80. https://doi.org/10.1146/annurev-anchem-071114-040325
Elschenbroich, S., Kislinger, T., 2011. Targeted proteomics by selected reaction monitoring mass spectrometry: applications to systems biology and biomarker discovery. Mol BioSyst 7, 292–303. https://doi.org/10.1039/C0MB00159G
Emadali, A., Gallagher-Gambarelli, M., 2009. La protéomique quantitative par la méthode SILAC: Technique et perspectives. médecine/sciences 25, 835–842. https://doi.org/10.1051/medsci/20092510835
Emele, M.F., Možina, S.S., Lugert, R., Bohne, W., Masanta, W.O., Riedel, T., Groß, U., Bader, O., Zautner, A.E., 2019. Proteotyping as alternate typing method to differentiate Campylobacter coli clades. Sci. Rep. 9, 4244. https://doi.org/10.1038/s41598-019-40842-w
Fedurco, M., 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 34, e22–e22. https://doi.org/10.1093/nar/gnj023
Fernández-Costa, C., Martínez-Bartolomé, S., McClatchy, D.B., Saviola, A.J., Yu, N.-K., Yates, J.R., 2020. Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. J. Proteome Res. 19, 3153–3161. https://doi.org/10.1021/acs.jproteome.0c00153
Florio, W., Baldeschi, L., Rizzato, C., Tavanti, A., Ghelardi, E., Lupetti, A., 2020. Detection of Antibiotic-Resistance by MALDI-TOF Mass Spectrometry: An Expanding Area. Front. Cell. Infect. Microbiol. 10, 572909. https://doi.org/10.3389/fcimb.2020.572909
Franco-Duarte, R., Černáková, L., Kadam, S., S. Kaushik, K., Salehi, B., Bevilacqua, A., Corbo, M.R., Antolak, H., Dybka-Stępień, K., Leszczewicz, M., Relison Tintino, S., Alexandrino de Souza, V.C., Sharifi-Rad, J., Melo Coutinho, H.D., Martins, N., Rodrigues, C.F., 2019. Advances in Chemical and Biological Methods to Identify Microorganisms—From Past to Present. Microorganisms 7, 130. https://doi.org/10.3390/microorganisms7050130
Fuerst, J., 2014. Microorganisms—A Journal and a Unifying Concept for the Science of Microbiology. Microorganisms 2, 140–146. https://doi.org/10.3390/microorganisms2040140
Gans, J., Wolinsky, M., Dunbar, J., 2005. Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil. Science 309, 1387–1390. https://doi.org/10.1126/science.1112665
Gao, B., Gupta, R.S., 2012. Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. Microbiol. Mol. Biol. Rev. 76, 66–112. https://doi.org/10.1128/MMBR.05011-11
Geojith, G., Dhanasekaran, S., Chandran, S.P., Kenneth, J., 2011. Efficacy of loop mediated isothermal amplification (LAMP) assay for the laboratory identification of Mycobacterium tuberculosis isolates in a resource limited setting. J. Microbiol. Methods 84, 71–73. https://doi.org/10.1016/j.mimet.2010.10.015
Gest, H., n.d. The Discovery of Microorganisms Revisited.
Ghurye, J.S., n.d. Metagenomic Assembly: Overview, Challenges and Applications.
Gilbert, W., Maxam, A., 1973. The Nucleotide Sequence of the *lac* Operator. Proc. Natl. Acad. Sci. 70, 3581–3584. https://doi.org/10.1073/pnas.70.12.3581
Gillet, L.C., Navarro, P., Tate, S., Reiter, L., Bonner, R., Aebersold, R., n.d. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis*□S.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91. https://doi.org/10.1099/ijs.0.64483-0

Gouveia, D., Miotello, G., Gallais, F., Gaillard, J.-C., Debroas, S., Bellanger, L., Lavigne, J.-P., Sotto, A., Grenga, L., Pible, O., Armengaud, J., 2020. Proteotyping SARS-CoV-2 Virus from Nasopharyngeal Swabs: A Proof-of-Concept Focused on a 3 Min Mass Spectrometry Window. J. Proteome Res. 19, 4407–4416. https://doi.org/10.1021/acs.jproteome.0c00535

Graham, R.L., Graham, C., McMullan, G., 2007. Microbial proteomics: a mass spectrometry primer for biologists. Microb. Cell Factories 6, 26. https://doi.org/10.1186/1475-2859-6-26

Granholm, V., Käll, L., 2011. Quality assessments of peptide-spectrum matches in shotgun proteomics. PROTEOMICS 11, 1086–1093. https://doi.org/10.1002/pmic.201000432

Grenga, L., Pible, O., Armengaud, J., 2019. Pathogen proteotyping: A rapidly developing application of mass spectrometry to address clinical concerns. Clin. Mass Spectrom. 14, 9–17. https://doi.org/10.1016/j.clinms.2019.04.004

Grenga, L., Pible, O., Miotello, G., Culotta, K., Ruat, S., Roncato, M., Gas, F., Bellanger, L., Claret, P., Dunyach-Remy, C., Laureillard, D., Sotto, A., Lavigne, J., Armengaud, J., 2022. Taxonomical and functional changes in COVID -19 faecal microbiome could be related to SARS-CoV -2 faecal load. Environ. Microbiol. 1462-2920.16028. https://doi.org/10.1111/1462-2920.16028

Hagen, J.B., 2012. Five Kingdoms, More or Less: Robert Whittaker and the Broad Classification of Organisms. BioScience 62, 67–74. https://doi.org/10.1525/bio.2012.62.1.11

Hall, B.G., 2013. Building Phylogenetic Trees from Molecular Data with MEGA. Mol. Biol. Evol. 30, 1229–1235. https://doi.org/10.1093/molbev/mst012

Hardinge, P., Murray, J.A.H., 2019. Reduced False Positives and Improved Reporting of Loop-Mediated Isothermal Amplification using Quenched Fluorescent Primers. Sci. Rep. 9, 7400. https://doi.org/10.1038/s41598-019-43817-z

Hardouin, P., Pible, O., Marchandin, H., Culotta, K., Armengaud, J., Chiron, R., Grenga, L., 2022. Quick and wide-range taxonomical repertoire establishment of the cystic fibrosis lung microbiota by tandem mass spectrometry on sputum samples. Front. Microbiol. 13, 975883. https://doi.org/10.3389/fmicb.2022.975883

Hayoun, K., 2020. Protéotypage de micro-organismes par spectrométrie de masse en tandem. Montpellier.

Hayoun, K., Gaillard, J.-C., Pible, O., Alpha-Bazin, B., Armengaud, J., 2020a. High-throughput proteotyping of bacterial isolates by double barrel chromatography-tandem mass spectrometry based on microplate paramagnetic beads and phylopeptidomics. J. Proteomics 226, 103887. https://doi.org/10.1016/j.jprot.2020.103887

Hayoun, K., Gouveia, D., Grenga, L., Pible, O., Armengaud, J., Alpha-Bazin, B., 2019. Evaluation of Sample Preparation Methods for Fast Proteotyping of Microorganisms by Tandem Mass Spectrometry. Front. Microbiol. 10, 1985. https://doi.org/10.3389/fmicb.2019.01985

Hayoun, K., Pible, O., Petit, P., Allain, F., Jouffret, V., Culotta, K., Rivasseau, C., Armengaud, J., Alpha-Bazin, B., 2020b. Proteotyping Environmental Microorganisms by Phylopeptidomics: Case Study Screening Water from a Radioactive Material Storage Pool. Microorganisms 8, 1525. https://doi.org/10.3390/microorganisms8101525

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Heil LR, Damoc E, Arrey TN, Pashkova A, Denisov E, Petzoldt J, Peterson AC, Hsu C, Searle BC, Shulman N, Riffle M, Connolly B, MacLean BX, Remes PM, Senko MW, Stewart HI, Hock C, Makarov AA, Hermanson D, Zabrouskov V, Wu CC, MacCoss MJ. Evaluating the Performance of the Astral Mass Analyzer for Quantitative Proteomics Using Data-Independent Acquisition. J Proteome Res. 2023 Oct 6;22(10):3290-3300. doi:
10.1021/acs.jproteome.3c00357. Epub 2023 Sep 8. PMID: 37683181; PMCID: PMC10563156.

Herraiz, T., Casal, V., 1995. Evaluation of solid-phase extraction procedures in peptide analysis. J. Chromatogr. A 708, 209–221. https://doi.org/10.1016/0021-9673(95)00388-4

Heyer, R., Benndorf, D., Kohrs, F., De Vrieze, J., Boon, N., Hoffmann, M., Rapp, E., Schlüter, A., Sczyrba, A., Reichl, U., 2016. Proteotyping of biogas plant microbiomes separates biogas plants according to process temperature and reactor type. Biotechnol. Biofuels 9, 155. https://doi.org/10.1186/s13068-016-0572-4

Hilt, E.E., Ferrieri, P., 2022. Next Generation and Other Sequencing Technologies in Diagnostic Microbiology and Infectious Diseases. Genes 13, 1566. https://doi.org/10.3390/genes13091566
Hirtz, C., Mannaa, A.M., Moulis, E., Pible, O., O'Flynn, R., Armengaud, J., Jouffret, V., Lemaistre, C., Dominici, G., Martinez, A.Y., Dunyach-Remy, C., Tiers, L., Lavigne, J.-P., Tramini, P., Goldsmith, M., Lehmann, S., Deville de Périère, D., Vialaret, J., 2022. Deciphering Black Extrinsic Tooth Stain Composition in Children Using Metaproteomics. ACS Omega 7, 8258–8267. https://doi.org/10.1021/acsomega.1c04770
Hosp, F., Scheltema, R.A., Eberl, H.C., Kulak, N.A., Keilhauer, E.C., Mayr, K., Mann, M., 2015. A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day*. Mol. Cell. Proteomics 14, 2030–2041. https://doi.org/10.1074/mcp.O115.049460
Hu, E.-Z., Lan, X.-R., Liu, Z.-L., Gao, J., Niu, D.-K., 2022. A positive correlation between GC content and growth temperature in prokaryotes. BMC Genomics 23, 110. https://doi.org/10.1186/s12864-022-08353-7
Huang, D., Yu, C., Shao, Z., Cai, M., Li, G., Zheng, L., Yu, Z., Zhang, J., 2020. Identification and Characterization of Nematicidal Volatile Organic Compounds from Deep-Sea Virgibacillus dokdonensis MCCC 1A00493. Molecules 25, 744. https://doi.org/10.3390/molecules25030744
Huang, M., Hull, C.M., 2017. Sporulation: how to survive on planet Earth (and beyond). Curr. Genet. 63, 831–838. https://doi.org/10.1007/s00294-017-0694-7
Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D.H., Soo, R.M., 2021. Prokaryotic taxonomy and nomenclature in the age of big sequence data. ISME J. 15, 1879–1892. https://doi.org/10.1038/s41396-021-00941-x
Ibrahim, A., Colson, P., Merhej, V., Zgheib, R., Maatouk, M., Naud, S., Bittar, F., Raoult, D., 2021. Rhizomal Reclassification of Living Organisms. Int. J. Mol. Sci. 22, 5643. https://doi.org/10.3390/ijms22115643
Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9, 5114. https://doi.org/10.1038/s41467-018-07641-9
Jang, K.-S., Kim, Y.H., 2018. Rapid and robust MALDI-TOF MS techniques for microbial identification: a brief overview of their diverse applications. J. Microbiol. 56, 209–216. https://doi.org/10.1007/s12275-018-7457-0
Jiang, Y., Hutton, A., Cranney, C.W., Meyer, J.G., 2022. Label-Free Quantification from Direct Infusion Shotgun Proteome Analysis (DISPA-LFQ) with CsoDIAq Software. Anal. Chem. acs.analchem.2c02249. https://doi.org/10.1021/acs.analchem.2c02249
Jin, W.-Y., Jang, S.-J., Lee, M.-J., Park, G., Kim, M.-J., Kook, J.-K., Kim, D.-M., Moon, D.-S., Park, Y.-J., 2011. Evaluation of VITEK 2, MicroScan, and Phoenix for identification of clinical isolates and reference strains. Diagn. Microbiol. Infect. Dis. 70, 442–447. https://doi.org/10.1016/j.diagmicrobio.2011.04.013
Jo, J.-H., Kennedy, E.A., Kong, H.H., 2016. Research Techniques Made Simple: Bacterial 16S Ribosomal RNA Gene Sequencing in Cutaneous Research. J. Invest. Dermatol. 136, e23–e27. https://doi.org/10.1016/j.jid.2016.01.005
Johnson, A.R., Carlson, E.E., 2015. Collision-Induced Dissociation Mass Spectrometry: A Powerful Tool for Natural Product Structure Elucidation. Anal. Chem. 87, 10668–10678. https://doi.org/10.1021/acs.analchem.5b01543
Jouffret, V., Miotello, G., Culotta, K., Ayrault, S., Pible, O., Armengaud, J., 2021. Increasing the power of interpretation for soil metaproteomics data. Microbiome 9, 195. https://doi.org/10.1186/s40168-021-01139-1
Kapinusova, G., Jani, K., Smrhova, T., Pajer, P., Jarosova, I., Suman, J., Strejcek, M., Uhlik, O., 2022. Culturomics of Bacteria from Radon-Saturated Water of the World's Oldest Radium Mine. Microbiol. Spectr. 10, e01995-22. https://doi.org/10.1128/spectrum.01995-22
Kapli, P., Yang, Z., Telford, M.J., 2020. Phylogenetic tree building in the genomic age. Nat. Rev. Genet. 21, 428–444. https://doi.org/10.1038/s41576-020-0233-0
Karlsson, R., 2015. Proteotyping: Proteomic characterization, classification and identification of microorganisms – A prospectus. Syst. Appl. Microbiol. 12.
Karlsson, R., Gonzales-Siles, L., Gomila, M., Busquets, A., Salvà-Serra, F., Jaén-Luchoro, D., Jakobsson, H.E., Karlsson, A., Boulund, F., Kristiansson, E., Moore, E.R.B., 2018a. Proteotyping bacteria: Characterization, differentiation and identification of pneumococcus and

other species within the Mitis Group of the genus Streptococcus by tandem mass spectrometry proteomics. PLOS ONE 13, e0208804. https://doi.org/10.1371/journal.pone.0208804
Karlsson, R., Gonzales-Siles, L., Gomila, M., Busquets, A., Salvà-Serra, F., Jaén-Luchoro, D., Jakobsson, H.E., Karlsson, A., Boulund, F., Kristiansson, E., Moore, E.R.B., 2018b. Proteotyping bacteria: Characterization, differentiation and identification of pneumococcus and other species within the Mitis Group of the genus Streptococcus by tandem mass spectrometry proteomics. PLOS ONE 13, e0208804. https://doi.org/10.1371/journal.pone.0208804
Karlsson, R., Thorsell, A., Gomila, M., Salvà-Serra, F., Jakobsson, H.E., Gonzales-Siles, L., Jaén-Luchoro, D., Skovbjerg, S., Fuchs, J., Karlsson, A., Boulund, F., Johnning, A., Kristiansson, E., Moore, E.R.B., 2020. Discovery of Species-unique Peptide Biomarkers of Bacterial Pathogens by Tandem Mass Spectrometry-based Proteotyping. Mol. Cell. Proteomics 19, 518–528. https://doi.org/10.1074/mcp.RA119.001667
Katoh, K., Rozewicki, J., Yamada, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166. https://doi.org/10.1093/bib/bbx108
Kim, J., Hur, J.I., Ryu, S., Jeon, B., 2021. Bacteriophage-Mediated Modulation of Bacterial Competition during Selective Enrichment of Campylobacter. Microbiol. Spectr. 9, e01703-21. https://doi.org/10.1128/Spectrum.01703-21
Kirwan, J.A., Weber, R.J.M., Broadhurst, D.I., Viant, M.R., 2014. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. Sci. Data 1, 140012. https://doi.org/10.1038/sdata.2014.12
Konstantinidis, K.T., Tiedje, J.M., 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. 102, 2567–2572. https://doi.org/10.1073/pnas.0409727102
Kozlova, A., Shkrigunov, T., Gusev, S., Guseva, M., Ponomarenko, E., Lisitsa, A., 2022. An Open-Source Pipeline for Processing Direct Infusion Mass Spectrometry Data of the Human Plasma Metabolome. Metabolites 12, 768. https://doi.org/10.3390/metabo12080768
Kralik, P., Ricchi, M., 2017. A Basic Guide to Real Time PCR in Microbial Diagnostics: Definitions, Parameters, and Everything. Front. Microbiol. 8. https://doi.org/10.3389/fmicb.2017.00108
Krasny, L., Huang, P.H., 2021. Data-independent acquisition mass spectrometry (DIA-MS) for proteomic applications in oncology. Mol. Omics 17, 29–42. https://doi.org/10.1039/D0MO00072H
Kreimer, S., Belov, M.E., Danielson, W.F., Levitsky, L.I., Gorshkov, M.V., Karger, B.L., Ivanov, A.R., 2016. Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling. J. Proteome Res. 15, 3563–3573. https://doi.org/10.1021/acs.jproteome.6b00312
Krieger, J.R., Wybenga-Groot, L.E., Tong, J., Bache, N., Tsao, M.S., Moran, M.F., 2019. Evosep One Enables Robust Deep Proteome Coverage Using Tandem Mass Tags while Significantly Reducing Instrument Time. J. Proteome Res. 18, 2346–2353. https://doi.org/10.1021/acs.jproteome.9b00082
Kuhring, M., Doellinger, J., Nitsche, A., Muth, T., Renard, B.Y., 2020. TaxIt: An Iterative Computational Pipeline for Untargeted Strain-Level Identification Using MS/MS Spectra from Pathogenic Single-Organism Samples. J. Proteome Res. 19, 2501–2510. https://doi.org/10.1021/acs.jproteome.9b00714
Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., Mann, M., 2014. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods 11, 319–324. https://doi.org/10.1038/nmeth.2834
Kunze-Szikszay, N., Euler, M., Perl, T., 2021. Identification of volatile compounds from bacteria by spectrometric methods in medicine diagnostic and other areas: current state and perspectives. Appl. Microbiol. Biotechnol. 105, 6245–6255. https://doi.org/10.1007/s00253-021-11469-7
Kurokawa, M., Ying, B.-W., 2017. Precise, High-throughput Analysis of Bacterial Growth. J. Vis. Exp. 56197. https://doi.org/10.3791/56197
Kutschera, U., 2011. From the scala naturae to the symbiogenetic and dynamic tree of life. Biol. Direct 6, 33. https://doi.org/10.1186/1745-6150-6-33
Lagier, J.-C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., Trape, J.-F., Koonin, E.V., La Scola, B., Raoult, D., 2012. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin. Microbiol. Infect. 18, 1185–1193. https://doi.org/10.1111/1469-0691.12023

Lagier, J.-C., Bilen, M., Cadoret, F., Drancourt, M., Fournier, P.-E., La Scola, B., Raoult, D., 2018. Naming microorganisms: the contribution of the IHU Méditerranée Infection, Marseille, France. New Microbes New Infect. 26, S89–S95. https://doi.org/10.1016/j.nmni.2018.08.006

Larsbrink, J., Sara McKee, L., 2020. Chapter Two - Bacteroidetes bacteria in the soil: Glycan acquisition, enzyme secretion, and gliding motility contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. Adv. Appl. Microbiol. https://doi.org/10.1016/bs.aambs.2019.11.001

Lenkowski, M., Nijakowski, K., Kaczmarek, M., Surdacka, A., 2021. The Loop-Mediated Isothermal Amplification Technique in Periodontal Diagnostics: A Systematic Review. J. Clin. Med. 10, 1189. https://doi.org/10.3390/jcm10061189

Li, J., Van Vranken, J.G., Pontano Vaites, L., Schweppe, D.K., Huttlin, E.L., Etienne, C., Nandhikonda, P., Viner, R., Robitaille, A.M., Thompson, A.H., Kuhn, K., Pike, I., Bomgarden, R.D., Rogers, J.C., Gygi, S.P., Paulo, J.A., 2020. TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. Nat. Methods 17, 399–404. https://doi.org/10.1038/s41592-020-0781-4

Li, Yanyan, Wu, S., Wang, L., Li, Ye, Shi, F., Wang, X., 2010. Differentiation of bacteria using fatty acid profiles from gas chromatography-tandem mass spectrometry: Differentiation of bacteria using fatty acid profiles. J. Sci. Food Agric. 90, 1380–1383. https://doi.org/10.1002/jsfa.3931

Linné, C. von, 1759. Tomus II: Vegetabilia. Facsimile. Systema Naturae Ed. 10., 1759.

Locey, K.J., Lennon, J.T., 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. 113, 5970–5975. https://doi.org/10.1073/pnas.1521291113

Loroch, S., Kopczynski, D., Schneider, A.C., Schumbrutzki, C., Feldmann, I., Panagiotidis, E., Reinders, Y., Sakson, R., Solari, F.A., Vening, A., Swieringa, F., Heemskerk, J.W.M., Grandoch, M., Dandekar, T., Sickmann, A., 2022. Toward Zero Variance in Proteomics Sample Preparation: Positive-Pressure FASP in 96-Well Format (PF96) Enables Highly Reproducible, Time- and Cost-Efficient Analysis of Sample Cohorts. J. Proteome Res. 21, 1181–1188. https://doi.org/10.1021/acs.jproteome.1c00706

Lozano, C., Kielbasa, M., Gaillard, J.-C., Miotello, G., Pible, O., Armengaud, J., 2022. Identification and Characterization of Marine Microorganisms by Tandem Mass Spectrometry Proteotyping. Microorganisms 10, 719. https://doi.org/10.3390/microorganisms10040719

Manadas, B., Mendes, V.M., English, J., Dunn, M.J., 2010. Peptide fractionation in proteomics approaches. Expert Rev. Proteomics 7, 655–663. https://doi.org/10.1586/epr.10.46

Mann, S., Chen, Y.-P.P., 2010. Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics 95, 7–15. https://doi.org/10.1016/j.ygeno.2009.09.002

Mappa, C., Alpha-Bazin, B., Pible, O., Armengaud, J., 2023a. Evaluation of the Limit of Detection of Bacteria by Tandem Mass Spectrometry Proteotyping and Phylopeptidomics. Microorganisms 11, 1170. https://doi.org/10.3390/microorganisms11051170

Mappa, C., Alpha-Bazin, B., Pible, O., Armengaud, J., 2023b. Mix24X, a Lab-Assembled Reference to Evaluate Interpretation Procedures for Tandem Mass Spectrometry Proteotyping of Complex Samples. Int. J. Mol. Sci. 24, 8634. https://doi.org/10.3390/ijms24108634

Marques, C., Liu, L., Duncan, K.D., Lanekoff, I., 2022. A Direct Infusion Probe for Rapid Metabolomics of Low-Volume Samples. Anal. Chem. 94, 12875–12883. https://doi.org/10.1021/acs.analchem.2c02918

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 6, 610–618. https://doi.org/10.1038/ismej.2011.139

Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., Dawyndt, P., 2012. Unipept: Tryptic Peptide-Based Biodiversity Analysis of Metaproteome Samples. J. Proteome Res. 11, 5773–5780. https://doi.org/10.1021/pr300576s

Mesuere, B., Van der Jeugt, F., Devreese, B., Vandamme, P., Dawyndt, P., 2016a. The unique peptidome: Taxon-specific tryptic peptides as biomarkers for targeted metaproteomics. PROTEOMICS 16, 2313–2318. https://doi.org/10.1002/pmic.201600023

Mesuere, B., Willems, T., Van der Jeugt, F., Devreese, B., Vandamme, P., Dawyndt, P., 2016b. Unipept web services for metaproteomics analysis. Bioinformatics 32, 1746–1748. https://doi.org/10.1093/bioinformatics/btw039

Meyer, J.G., 2021. Qualitative and Quantitative Shotgun Proteomics Data Analysis from Data-Dependent Acquisition Mass Spectrometry, in: Carrera, M., Mateos, J. (Eds.), Shotgun Proteomics, Methods in Molecular Biology. Springer US, New York, NY, pp. 297–308. https://doi.org/10.1007/978-1-0716-1178-4_19

Meyer, J.G., Niemi, N.M., Pagliarini, D.J., Coon, J.J., 2020. Quantitative shotgun proteome analysis by direct infusion. Nat. Methods 17, 1222–1228. https://doi.org/10.1038/s41592-020-00999-z

Mező, E., Hartmann-Balogh, F., Madarászné Horváth, I., Bufa, A., Marosvölgyi, T., Kocsis, B., Makszin, L., 2022. Effect of Culture Conditions on Fatty Acid Profiles of Bacteria and Lipopolysaccharides of the Genus Pseudomonas—GC-MS Analysis on Ionic Liquid-Based Column. Molecules 27, 6930. https://doi.org/10.3390/molecules27206930

Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77. https://doi.org/10.1016/j.coviro.2011.12.004

Montes-Osuna, N., Cernava, T., Gómez-Lama Cabanás, C., Berg, G., Mercado-Blanco, J., 2022. Identification of Volatile Organic Compounds Emitted by Two Beneficial Endophytic Pseudomonas Strains from Olive Roots. Plants 11, 318. https://doi.org/10.3390/plants11030318

Mooradian, A.D., van der Post, S., Naegle, K.M., Held, J.M., 2020. ProteoClade: A taxonomic toolkit for multi-species and metaproteomic analysis. PLOS Comput. Biol. 16, e1007741. https://doi.org/10.1371/journal.pcbi.1007741

Moreira, D., Brochier-Armanet, C., 2008. Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. BMC Evol. Biol. 8, 12. https://doi.org/10.1186/1471-2148-8-12

Mount, D.W., 2008. Maximum parsimony method for phylogenetic prediction. CSH Protoc. https://doi.org/doi: 10.1101/pdb.top32.

Müller, V., Sousa, J.M., Ceylan Koydemir, H., Veli, M., Tseng, D., Cerqueira, L., Ozcan, A., Azevedo, N.F., Westerlund, F., 2018. Identification of pathogenic bacteria in complex samples using a smartphone based fluorescence microscope. RSC Adv. 8, 36493–36502. https://doi.org/10.1039/C8RA06473C

Murray, A.E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kämpfer, P., Konstantinidis, K.T., Lane, C.E., Papke, R.T., Parks, D.H., Rossello-Mora, R., Stott, M.B., Sutcliffe, I.C., Thrash, J.C., Venter, S.N., Whitman, W.B., Acinas, S.G., Amann, R.I., Anantharaman, K., Armengaud, J., Baker, B.J., Barco, R.A., Bode, H.B., Boyd, E.S., Brady, C.L., Carini, P., Chain, P.S.G., Colman, D.R., DeAngelis, K.M., de los Rios, M.A., Estrada-de los Santos, P., Dunlap, C.A., Eisen, J.A., Emerson, D., Ettema, T.J.G., Eveillard, D., Girguis, P.R., Hentschel, U., Hollibaugh, J.T., Hug, L.A., Inskeep, W.P., Ivanova, E.P., Klenk, H.-P., Li, W.-J., Lloyd, K.G., Löffler, F.E., Makhalanyane, T.P., Moser, D.P., Nunoura, T., Palmer, M., Parro, V., Pedrós-Alió, C., Probst, A.J., Smits, T.H.M., Steen, A.D., Steenkamp, E.T., Spang, A., Stewart, F.J., Tiedje, J.M., Vandamme, P., Wagner, M., Wang, F.-P., Yarza, P., Hedlund, B.P., Reysenbach, A.-L., 2020a. Roadmap for naming uncultivated Archaea and Bacteria. Nat. Microbiol. 5, 987–994. https://doi.org/10.1038/s41564-020-0733-x

Murray, A.E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kämpfer, P., Konstantinidis, K.T., Lane, C.E., Papke, R.T., Parks, D.H., Rossello-Mora, R., Stott, M.B., Sutcliffe, I.C., Thrash, J.C., Venter, S.N., Whitman, W.B., Acinas, S.G., Amann, R.I., Anantharaman, K., Armengaud, J., Baker, B.J., Barco, R.A., Bode, H.B., Boyd, E.S., Brady, C.L., Carini, P., Chain, P.S.G., Colman, D.R., DeAngelis, K.M., de los Rios, M.A., Estrada-de los Santos, P., Dunlap, C.A., Eisen, J.A., Emerson, D., Ettema, T.J.G., Eveillard, D., Girguis, P.R., Hentschel, U., Hollibaugh, J.T., Hug, L.A., Inskeep, W.P., Ivanova, E.P., Klenk, H.-P., Li, W.-J., Lloyd, K.G., Löffler, F.E., Makhalanyane, T.P., Moser, D.P., Nunoura, T., Palmer, M., Parro, V., Pedrós-Alió, C., Probst, A.J., Smits, T.H.M., Steen, A.D., Steenkamp, E.T., Spang, A., Stewart, F.J., Tiedje, J.M., Vandamme, P., Wagner, M., Wang, F.-P., Yarza, P., Hedlund, B.P., Reysenbach, A.-L., 2020b. Roadmap for naming uncultivated Archaea and Bacteria. Nat. Microbiol. 5, 987–994. https://doi.org/10.1038/s41564-020-0733-x

Nasir, A., Caetano-Anollés, G., 2015. A phylogenomic data-driven exploration of viral origins and evolution. Sci. Adv. 1, e1500527. https://doi.org/10.1126/sciadv.1500527

Nasseri, B., Soleimani, N., Rabiee, N., Kalbasi, A., Karimi, M., Hamblin, M.R., 2018. Point-of-care microfluidic devices for pathogen detection. Biosens. Bioelectron. 117, 112–128. https://doi.org/10.1016/j.bios.2018.05.050

Notomi, T., 2000. Loop-mediated isothermal amplification of DNA. Nucleic Acids Res. 28, 63e–663. https://doi.org/10.1093/nar/28.12.e63

Ochoa, S., Martínez, O.A., Fernández, H., Collado, L., 2019. Comparison of media and growth conditions for culturing enterohepatic *Helicobacter* species. Lett. Appl. Microbiol. lam.13192. https://doi.org/10.1111/lam.13192

O'Dwyer, D.N., Ashley, S.L., Gurczynski, S.J., Xia, M., Wilke, C., Falkowski, N.R., Norman, K.C., Arnold, K.B., Huffnagle, G.B., Salisbury, M.L., Han, M.K., Flaherty, K.R., White, E.S., Martinez, F.J., Erb-Downward, J.R., Murray, S., Moore, B.B., Dickson, R.P., 2019. Lung Microbiota Contribute to Pulmonary Inflammation and Disease Progression in Pulmonary Fibrosis. Am. J. Respir. Crit. Care Med. 199, 1127–1138. https://doi.org/10.1164/rccm.201809-1650OC

Ojima-Kato, T., Nagai, S., Fujita, A., Sakata, J., Tamura, H., 2023. Proteotyping of Campylobacter jejuni by MALDI-TOF MS and Strain Solution Version 2 Software. Microorganisms 11, 202. https://doi.org/10.3390/microorganisms11010202

Opal, S.M., 2010. A Brief History of Microbiology and Immunology, in: Artenstein, A.W. (Ed.), Vaccines: A Biography. Springer New York, New York, NY, pp. 31–56. https://doi.org/10.1007/978-1-4419-1108-7_3

Oren, A., Arahal, D.R., Göker, M., Moore, E.R.B., Rossello-Mora, R., Sutcliffe, I.C., 2022. Proposals to emend Rules 8, 15, 22, 25a, 30(3)(b), 30(4), 34a, and Appendix 7 of the International Code of Nomenclature of Prokaryotes. Int. J. Syst. Evol. Microbiol. 72. https://doi.org/10.1099/ijsem.0.005630

Oren, A., Garrity, G.M., 2021. Valid publication of the names of forty-two phyla of prokaryotes. Int. J. Syst. Evol. Microbiol. 71. https://doi.org/10.1099/ijsem.0.005056

Oviaño, M., Rodríguez-Sánchez, B., Gómara, M., Alcalá, L., Zvezdanova, E., Ruíz, A., Velasco, D., Gude, M.J., Bouza, E., Bou, G., 2018. Direct identification of clinical pathogens from liquid culture media by MALDI-TOF MS analysis. Clin. Microbiol. Infect. 24, 624–629. https://doi.org/10.1016/j.cmi.2017.09.010

Pan Q, Zhuo X, He C, Zhang Y, Shi Q. Validation and Evaluation of High-Resolution Orbitrap Mass Spectrometry on Molecular Characterization of Dissolved Organic Matter. ACS Omega. 2020 Mar 9;5(10):5372-5379. doi: 10.1021/acsomega.9b04411. PMID: 32201827; PMCID: PMC7081437

Park, S.T., Kim, J., 2016. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. Int. Neurourol. J. 20, S76-83. https://doi.org/10.5213/inj.1632742.371

Penzlin, A., Lindner, M.S., Doellinger, J., Dabrowski, P.W., Nitsche, A., Renard, B.Y., 2014. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. Bioinformatics 30, i149–i156. https://doi.org/10.1093/bioinformatics/btu267

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20, 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Petit, P.C.M., Pible, O., Eesbeeck, V.V., Alban, C., Steinmetz, G., Mysara, M., Monsieurs, P., Armengaud, J., Rivasseau, C., 2020. Direct Meta-Analyses Reveal Unexpected Microbial Life in the Highly Radioactive Water of an Operating Nuclear Reactor Core. Microorganisms 8, 1857. https://doi.org/10.3390/microorganisms8121857

Peyroux, J., 2022. Etude et application de nouvelles techniques d'imagerie et d'intelligence artificielle pour l'identification bactérienne. Médecine humaine et pathologie. Grenoble Alpes.

Pible, O., Allain, F., Jouffret, V., Culotta, K., Miotello, G., Armengaud, J., 2020. Estimating relative biomasses of organisms in microbiota using "phylopeptidomics." Microbiome 8, 30. https://doi.org/10.1186/s40168-020-00797-x

Pino, L.K., Just, S.C., MacCoss, M.J., Searle, B.C., 2020. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. Mol. Cell. Proteomics 19, 1088–1103. https://doi.org/10.1074/mcp.P119.001913

Potriquet, J., Laohaviroj, M., Bethony, J.M., Mulvenna, J., 2017. A modified FASP protocol for high-throughput preparation of protein samples for mass spectrometry. PLOS ONE 12, e0175967. https://doi.org/10.1371/journal.pone.0175967

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596. https://doi.org/10.1093/nar/gks1219

Quezada, M., BuitrÃ³n, G., Moreno-Andrade, I., Moreno, G., LÃ³pez-MarÃ-n, L.M., 2007. The use of fatty acid methyl esters as biomarkers to determine aerobic, facultatively aerobic and anaerobic

communities in wastewater treatment systems. FEMS Microbiol. Lett. 266, 75–82. https://doi.org/10.1111/j.1574-6968.2006.00509.x
Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.-M., 2004. The 1.2-Megabase Genome Sequence of Mimivirus. Science 306, 1344–1350. https://doi.org/10.1126/science.1101485
Rappsilber, J., Mann, M., Ishihama, Y., 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protoc. 2, 1896–1906. https://doi.org/10.1038/nprot.2007.261
Rebrosova, K., Samek, O., Kizovsky, M., Bernatova, S., Hola, V., Ruzicka, F., 2022. Raman Spectroscopy—A Novel Method for Identification and Characterization of Microbes on a Single-Cell Level in Clinical Settings. Front. Cell. Infect. Microbiol. 12, 866463. https://doi.org/10.3389/fcimb.2022.866463
Renvoisé, A., Brossier, F., Sougakoff, W., Jarlier, V., Aubry, A., 2013. Broad-range PCR: Past, present, or future of bacteriology? Médecine Mal. Infect. 43, 322–330. https://doi.org/10.1016/j.medmal.2013.06.003
Révész, Á., Hevér, H., Steckel, A., Schlosser, G., Szabó, D., Vékey, K., Drahos, L., 2023. Collision energies: Optimization strategies for bottom-up proteomics. Mass Spectrom. Rev. 42, 1261–1299. https://doi.org/10.1002/mas.21763
Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. 106, 19126–19131. https://doi.org/10.1073/pnas.0906412106
Rinas, A., Jenkins, C., Orsburn, B., 2019. Assessing a commercial capillary electrophoresis interface (ZipChip) for shotgun proteomic applications (preprint). Biochemistry. https://doi.org/10.1101/559591
Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., Pappin, D.J., 2004. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. Mol. Cell. Proteomics 3, 1154–1169. https://doi.org/10.1074/mcp.M400129-MCP200
Saleh, S., Staes, A., Deborggraeve, S., Gevaert, K., 2019. Targeted Proteomics for Studying Pathogenic Bacteria. PROTEOMICS 19, 1800435. https://doi.org/10.1002/pmic.201800435
Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463
Savaryn, J.P., Toby, T.K., Kelleher, N.L., 2016. A researcher's guide to mass spectrometry-based proteomics. PROTEOMICS 16, 2435–2443. https://doi.org/10.1002/pmic.201600113
Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D., Karsch-Mizrachi, I., 2019. GenBank. Nucleic Acids Res. 47, D94–D99. https://doi.org/10.1093/nar/gky989
Sédillot, C.-E., 1878. De l'inflence des découvertes de M. Pasteur sur les progrès de la Chirurgie », in: Comptes Rendus Hebdomadaires Des Séances de l'Académie Des Sciences, t. 86, (1878), p. 634.
Selosse, M.-A., 2022. Les virus sont-ils vivants ? Leçon d'interdépendance. médecine/sciences 38, 1061–1063. https://doi.org/10.1051/medsci/2022167
Sentausa, E., Fournier, P.-E., 2013. Advantages and limitations of genomics in prokaryotic taxonomy. Clin. Microbiol. Infect. 19, 790–795. https://doi.org/10.1111/1469-0691.12181
Shi, Y., Wang, G., Lau, H.C.-H., Yu, J., 2022. Metagenomic Sequencing for Microbial DNA in Human Samples: Emerging Technological Advances. Int. J. Mol. Sci. 23, 2181. https://doi.org/10.3390/ijms23042181
Spatola Rossi, C., Coulon, F., Ma, S., Zhang, Y.S., Yang, Z., 2023. Microfluidics for Rapid Detection of Live Pathogens. Adv. Funct. Mater. 33, 2212081. https://doi.org/10.1002/adfm.202212081
Steppert, I., Schönfelder, J., Schultz, C., Kuhlmeier, D., 2021. Rapid in vitro differentiation of bacteria by ion mobility spectrometry. Appl. Microbiol. Biotechnol. 105, 4297–4307. https://doi.org/10.1007/s00253-021-11315-w
Suarez, S., 2013. Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory. J. Microbiol. Methods 7.
Suarez, S., Nassif, X., Ferroni, A., 2015. Applications de la technologie MALDI-TOF en microbiologie clinique. Pathol. Biol. 63, 43–52. https://doi.org/10.1016/j.patbio.2014.10.002

Sune, D., Rydberg, H., Augustinsson, Å.N., Serrander, L., Jungeström, M.B., 2020. Optimization of 16S rRNA gene analysis for use in the diagnostic clinical microbiology service. J. Microbiol. Methods 170, 105854. https://doi.org/10.1016/j.mimet.2020.105854

Taş, N., de Jong, A.E., Li, Y., Trubl, G., Xue, Y., Dove, N.C., 2021. Metagenomic tools in microbial ecology research. Curr. Opin. Biotechnol. 67, 184–191. https://doi.org/10.1016/j.copbio.2021.01.019

The neighbor-joining method: a new method for reconstructing phylogenetic trees., 1987. . Mol. Biol. Evol. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Thomas, S.N., 2019. Mass spectrometry, in: Contemporary Practice in Clinical Chemistry. Elsevier, pp. 171–185. https://doi.org/10.1016/B978-0-12-815499-1.00010-7

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Hamon, C., 2003. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. Anal. Chem. 75, 1895–1904. https://doi.org/10.1021/ac0262560

Thompson, J.D., Higgins, D.G., Gibson, T.J., n.d. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.

Trapp, J., Almunia, C., Gaillard, J.-C., Pible, O., Chaumot, A., Geffard, O., Armengaud, J., 2016. Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. J. Proteomics 135, 51–61. https://doi.org/10.1016/j.jprot.2015.06.017

Tsuchida, S., Umemura, H., Nakayama, T., 2020. Current Status of Matrix-Assisted Laser Desorption/Ionization–Time-of-Flight Mass Spectrometry (MALDI-TOF MS) in Clinical Diagnostic Microbiology. Molecules 25, 4775. https://doi.org/10.3390/molecules25204775

Uchiyama, J., Matsui, H., Murakami, H., Kato, S., Watanabe, N., Nasukawa, T., Mizukami, K., Ogata, M., Sakaguchi, M., Matsuzaki, S., Hanaki, H., 2018. Potential Application of Bacteriophages in Enrichment Culture for Improved Prenatal Streptococcus agalactiae Screening. Viruses 10, 552. https://doi.org/10.3390/v10100552

Urwin, R., Maiden, M.C.J., 2003. Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol. 11, 479–487. https://doi.org/10.1016/j.tim.2003.08.006

van de Velde, C.C., Joseph, C., Biclot, A., Huys, G.R.B., Pinheiro, V.B., Bernaerts, K., Raes, J., Faust, K., 2022. Fast quantification of gut bacterial species in cocultures using flow cytometry and supervised classification. ISME Commun. 2, 40. https://doi.org/10.1038/s43705-022-00123-6

Van Den Bossche, T., Arntzen, M.Ø., Becher, D., Benndorf, D., Eijsink, V.G.H., Henry, C., Jagtap, P.D., Jehmlich, N., Juste, C., Kunath, B.J., Mesuere, B., Muth, T., Pope, P.B., Seifert, J., Tanca, A., Uzzau, S., Wilmes, P., Hettich, R.L., Armengaud, J., 2021. The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. Microbiome 9, 243. https://doi.org/10.1186/s40168-021-01176-w

van der Laan, T., Dubbelman, A.-C., Duisters, K., Kindt, A., Harms, A.C., Hankemeier, T., 2020. High-Throughput Fractionation Coupled to Mass Spectrometry for Improved Quantitation in Metabolomics. Anal. Chem. 92, 14330–14338. https://doi.org/10.1021/acs.analchem.0c01375

Vanstokstraeten, R., Mackens, S., Callewaert, E., Blotwijk, S., Emmerechts, K., Crombé, F., Soetens, O., Wybo, I., Vandoorslaer, K., Mostert, L., De Geyter, D., Muyldermans, A., Blockeel, C., Piérard, D., Demuyser, T., 2022. Culturomics to Investigate the Endometrial Microbiome: Proof-of-Concept. Int. J. Mol. Sci. 23, 12212. https://doi.org/10.3390/ijms232012212

Vitorino, L., Bessa, L., 2018. Microbial Diversity: The Gap between the Estimated and the Known. Diversity 10, 46. https://doi.org/10.3390/d10020046

Voelkerding, K.V., Dames, S.A., Durtschi, J.D., 2009. Next-Generation Sequencing: From Basic Research to Diagnostics. Clin. Chem. 55, 641–658. https://doi.org/10.1373/clinchem.2008.112789

Wang, L., Liu, W., Tang, J.-W., Wang, J.-J., Liu, Q.-H., Wen, P.-B., Wang, M.-M., Pan, Y.-C., Gu, B., Zhang, X., 2021. Applications of Raman Spectroscopy in Bacterial Infections: Principles, Advantages, and Shortcomings. Front. Microbiol. 12, 683580. https://doi.org/10.3389/fmicb.2021.683580

Wensel, C.R., Pluznick, J.L., Salzberg, S.L., Sears, C.L., 2022. Next-generation sequencing: insights to advance clinical investigations of the microbiome. J. Clin. Invest. 132, e154944. https://doi.org/10.1172/JCI154944

Witt, N., Andreotti, S., Busch, A., Neubert, K., Reinert, K., Tomaso, H., Meierhofer, D., 2020. Rapid and Culture Free Identification of Francisella in Hare Carcasses by High-Resolution Tandem Mass Spectrometry Proteotyping. Front. Microbiol. 11, 636. https://doi.org/10.3389/fmicb.2020.00636

Woese, C.R., Fox, G.E., 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. 74, 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol. Biol. Evol. 14, 717–724. https://doi.org/10.1093/oxfordjournals.molbev.a025811

Yu, F., Teo, G.C., Kong, A.T., Fröhlich, K., Li, G.X., Demichev, V., Nesvizhskii, A.I., 2023. Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. Nat. Commun. 14, 4154. https://doi.org/10.1038/s41467-023-39869-5

Zarnowiec, P., Lechowicz, L., Czerwonka, G., Kaca, W., 2015. Fourier Transform Infrared Spectroscopy (FTIR) as a Tool for the Identification and Differentiation of Pathogenic Bacteria. Curr. Med. Chem. 22, 1710–1718. https://doi.org/10.2174/0929867322666150311152800

Zhang, D., Bi, H., Liu, B., Qiao, L., 2018. Detection of Pathogenic Microorganisms by Microfluidics Based Analytical Methods. Anal. Chem. 90, 5512–5520. https://doi.org/10.1021/acs.analchem.8b00399

Zhao, X., Li, M., Liu, Y., 2019. Microfluidic-Based Approaches for Foodborne Pathogen Detection. Microorganisms 7, 381. https://doi.org/10.3390/microorganisms7100381

Zheng, R., n.d. Fast, sensitive, and reproducible nano- and capillary-flow LCMS methods for high−throughput proteome profiling using the Vanquish Neo UHPLC system hyphenated with the Orbitrap Exploris 480 MS.

Zielinski, A.T., Kourtchev, I., Bortolini, C., Fuller, S.J., Giorio, C., Popoola, O.A.M., Bogialli, S., Tapparo, A., Jones, R.L., Kalberer, M., 2018. A new processing scheme for ultra-high resolution direct infusion mass spectrometry data. Atmos. Environ. 178, 129–139. https://doi.org/10.1016/j.atmosenv.2018.01.034

Zivanovic, Y., Armengaud, J., Lagorce, A., Leplat, C., Guérin, P., Dutertre, M., Anthouard, V., Forterre, P., Wincker, P., Confalonieri, F., 2009. Genome analysis and genome-wide proteomics of Thermococcus gammatolerans, the most radioresistant organism known amongst the Archaea. Genome Biol. 10, R70. https://doi.org/10.1186/gb-2009-10-6-r70

Zubarev, R.A., Makarov, A., 2013. Orbitrap Mass Spectrometry. Anal. Chem. 85, 5288–5296. https://doi.org/10.1021/ac4001223

Annexes

- **Liste des publications scientifiques parues et soumises dans des revues scientifiques internationales**

- **Label-free multiplex proteotyping of microbial isolates**
  Chabas, M., Pible, O., Armengaud, J., Alpha-Bazin, B., 2023. Label-Free Multiplex Proteotyping of Microbial Isolates. Anal. Chem. acs. analchem.3c01975.

- **A simplified label-free method for proteotyping sets of six isolates in a single tandem mass spectrometry analysis**
  Madisson Chabas, Jean Armengaud, Béatrice Alpha-Bazin ; soumis à le 30 août 2023 à Journal of Proteome Research : numéro de manuscript pr-2023-005353

- **Flash MS/MS proteotyping allows identifying microbial isolates in 36 seconds of mass spectrometry signal**
  Madisson Chabas, Jean-Charles Gaillard, Béatrice Alpha-Bazin, Jean Armengaud; Proteomics. 2024 Jan 2: e2300372

- **Label-free multiplex proteotyping of microbial isolates**
  Demande de brevet déposée par le CEA pour cette méthode innovante en Avril 2023 (Demande n° EP23305557.3).

- **Liste des formations**

- Journée de Rentrée des doctorants du Collège Doctoral CBS2 (2 heures)
- Éthique de la recherche (15 heures)
- Introduction à la statistique par R (25 heures)
- Formation école chercheurs-protéomique : "De la préparation des échantillons à l'interprétation des résultats. Stratégies et aspects pratiques" ; FPS, Sète (25 heures)
- Unlock your english (25 heures)
- Rédiger et publier un article scientifique (20 heures)

- **Liste des communications orales réalisées en thèse**

  - **Présentations orales :**

- **« Phylopeptidomique : Identification haut-débit de microorganismes par protéotypage en spectrométrie de masse en tandem »**

  Madisson Chabas, Olivier Pible, Jean Armengaud, Béatrice Alpha-Bazin

Journée des doctorants DMTS, 12 janvier 2021. Saclay, France

- **« Identification à haut-débit de microorganismes par protéotypage en spectrométrie de masse en tandem »**

  Madisson Chabas, Olivier Pible, Jean Armengaud, Béatrice Alpha-Bazin
  Journée des doctorants DMTS, 5 janvier 2022. Saclay, France

- **« Protéotypage haut-débit de microorganismes par phylopeptidomique »**

  Madisson Chabas, Olivier Pible, Jean Armengaud, Béatrice Alpha-Bazin

  Journée scientifique du Li2D, 31 mars 2022. Marcoule, France

- **« Innovative multiplex proteotyping of microbial isolates »**

  Madisson Chabas, Olivier Pible, Jean Armengaud, Béatrice Alpha-Bazin

  Congrès international ProtéoVilamoura FPS/SPS/PPS, 11-13 mai 2022. Villamoura, Portugal

- **« Le protéotypage de microorganismes par spectrométrie de masse en tandem : amélioration du débit d'analyse »**

  Madisson Chabas, Olivier Pible, Jean Armengaud, Béatrice Alpha-Bazin
  Journées ISEC Marcoule, 22-23 septembre 2022. Marcoule, France

- **« High-throughput identification of microorganisms by mass spectrometry proteotyping »**

  Madisson Chabas, Olivier Pible, Jean Armengaud, Béatrice Alpha-Bazin
  Journée des doctorants DMTS, 24 janvier 2023. Saclay, France

- **« Identification d'isolats microbiens par protéotypage »**

  Madisson Chabas, Jean-Charles Gaillard, Jean Armengaud, Béatrice Alpha-Bazin
  Journée scientifique du Li2D, 9 juin 2023. Marcoule, France

- **« Flash MS/MS identification of microbial isolates »**

  Madisson Chabas, Jean-Charles Gaillard, Jean Armengaud, Béatrice Alpha-Bazin
  Congrès internationale Protéoaix FPS/SPS/PPS, 20-23 juin 2023. Aix en Provence, France

  - **Présentations posters:**

- **« Protéotypage rapide d'isolats microbiens »**

  Madisson Chabas, Olivier Pible, Jean Armengaud, Béatrice Alpha-Bazin
  Congrès Microbes, SFM, 3-5 octobre juin 2022. Montpellier, France

- **« Flash MS/MS identification of microbial isolates »**

Madisson Chabas, Jean-Charles Gaillard, Jean Armengaud, Béatrice Alpha-Bazin
Congrès internationale Protéoaix FPS/SPS/PPS, 20-23 juin 2023. Aix en Provence, France

➢ Supplementary Data des articles

- **Article 1 :**

https://pubs.acs.org/doi/10.1021/acs.analchem.3c01975 : Strains used for the composition of the assemblages (Table S1); correlation between SPi maximum and fraction origin for the 11 identified species in M11 mix (Table S2); and Lorentzian curve fitting of each organisms corresponding to each 11 fractions of M11 (Figure S1) (PDF)

- **Article 2 :**

Table S1. List of assemblages and species for each of their six fractions.

Assemblage of five fractions	First fraction	Second fraction	Third fraction	Fourth fraction	Fith fraction	Sixth fraction	.RAW file*	.mgf file*	.dat file (Control Quality interpretation)*	.mzid file (Control Quality interpretation)*
M6_01	*O.indolifex*	*K.aerogenes*	*R. pomeroyi*	*S. stellata*	*S. cerevisae*	*M. tractuosa*	Q14778_M6_01	Q14778_M6_01	F230539.dat	F230539.mzid
M6_02	*M. tractuosa*	*O.indolifex*	*K.aerogenes*	*R. pomeroyi*	*S. stellata*	*S. cerevisae*	Q14779_M6_02	Q14779_M6_02	F230540.dat	F230540.mzid
M6_03	*S. cerevisae*	*M. tractuosa*	*O.indolifex*	*K.aerogenes*	*R. pomeroyi*	*S. stellata*	Q14780_M6_03	Q14780_M6_03	F230541.dat	F230541.mzid
M6_04	*S. stellata*	*S. cerevisae*	*M. tractuosa*	*O.indolifex*	*K.aerogenes*	*R. pomeroyi*	Q14781_M6_04	Q14781_M6_04	F230542.dat	F230542.mzid
M6_05	*R. pomeroyi*	*S. stellata*	*S. cerevisae*	*M. tractuosa*	*O.indolifex*	*K.aerogenes*	Q14782_M6_05	Q14782_M6_05	F230543.dat	F230543.mzid
M6_06	*K.aerogenes*	*R. pomeroyi*	*S. stellata*	*S. cerevisae*	*M. tractuosa*	*O.indolifex*	Q14783_M6_06	Q14783_M6_06	F230544.dat	F230544.mzid
M6_07	*K.aerogenes*	*O.indolifex*	*R. pomeroyi*	*S. stellata*	*M. tractuosa*	*S. cerevisae*	Q14784_M6_07	Q14784_M6_07	F230545.dat	F230545.mzid
M6_08	*O.indolifex*	*M. tractuosa*	*S. cerevisae*	*K.aerogenes*	*R. pomeroyi*	*S. stellata*	Q14785_M6_08	Q14785_M6_08	F230546.dat	F230546.mzid
M6_09	*R. pomeroyi*	*S. cerevisae*	*S. stellata*	*M. tractuosa*	*O.indolifex*	*K.aerogenes*	Q14786_M6_09	Q14786_M6_09	F230547.dat	F230547.mzid
M6_10	*S. stellata*	*R. pomeroyi*	*O.indolifex*	*S. cerevisae*	*K.aerogenes*	*M. tractuosa*	Q14787_M6_10	Q14787_M6_10	F230548.dat	F230548.mzid
M6_11	*S. cerevisae*	*K.aerogenes*	*M. tractuosa*	*R. pomeroyi*	*S. stellata*	*O.indolifex*	Q14788_M6_11	Q14788_M6_11	F230549.dat	F230549.mzid
M6_12	*M. tractuosa*	*S. stellata*	*K.aerogenes*	*O.indolifex*	*S. cerevisae*	*R. pomeroyi*	Q14789_M6_12	Q14789_M6_12	F230550.dat	F230550.mzid
M6_13	*M. tractuosa*	*S. cerevisae*	*S. stellata*	*R. pomeroyi*	*K.aerogenes*	*O.indolifex*	Q14790_M6_13	Q14790_M6_13	F230551.dat	F230551.mzid
M6_14	*S. cerevisae*	*S. stellata*	*R. pomeroyi*	*K.aerogenes*	*O.indolifex*	*M. tractuosa*	Q14791_M6_14	Q14791_M6_14	F230552.dat	F230552.mzid
M6_15	*S. stellata*	*R. pomeroyi*	*K.aerogenes*	*O.indolifex*	*M. tractuosa*	*S. cerevisae*	Q14792_M6_15	Q14792_M6_15	F230553.dat	F230553.mzid
M6_16	*R. pomeroyi*	*K.aerogenes*	*O.indolifex*	*M. tractuosa*	*S. cerevisae*	*S. stellata*	Q14793_M6_16	Q14793_M6_16	F230554.dat	F230554.mzid
M6_17	*K.aerogenes*	*O.indolifex*	*M. tractuosa*	*S. cerevisae*	*S. stellata*	*R. pomeroyi*	Q14794_M6_17	Q14794_M6_17	F230555.dat	F230555.mzid
M6_18	*O.indolifex*	*S. cerevisae*	*R. pomeroyi*	*M.tractuosa*	*S.stellata*	*K.aerogenes*	Q14795_M6_18	Q14795_M6_18	F230556.dat	F230556.mzid
M6_19	*S. stellata*	*O.indolifex*	*R. pomeroyi*	*M. tractuosa*	*S. cerevisae*	*K.aerogenes*	Q14796_M6_19	Q14796_M6_19	F230557.dat	F230557.mzid
M6_20	*M. tractuosa*	*R. pomeroyi*	*O.indolifex*	*S. stellata*	*K.aerogenes*	*S. cerevisae*	Q14797_M6_20	Q14797_M6_20	F230558.dat	F230558.mzid
M6_21	*O.indolifex*	*S. cerevisae*	*S. stellata*	*K.aerogenes*	*R. pomeroyi*	*M. tractuosa*	Q14806_M6_21	Q14806_M6_21	F230578.dat	F230578.mzid
M6_22	*K.aerogenes*	*S. stellata*	*S. cerevisae*	*O.indolifex*	*M. tractuosa*	*R. pomeroyi*	Q14807_M6_22	Q14807_M6_22	F230579.dat	F230579.mzid
M6_23	*R. pomeroyi*	*K.aerogenes*	*S. cerevisae*	*M. tractuosa*	*O.indolifex*	*S. stellata*	Q14808_M6_23	Q14808_M6_23	F230580.dat	F230580.mzid
M6_24	*K. aerogenes*	*M. tractuosa*	*O. indolifex*	*R. pomeroyi*	*S. cerevisae*	*S. stellata*	Q14798_M6_24	Q14798_M6_24	F230571.dat	F230571.mzid
M6_25	*S. stellata*	*R. pomeroyi*	*M. tractuosa*	*S. cerevisae*	*K. aerogenes*	*O. indolifex*	Q14799_M6_25	Q14799_M6_25	F230572.dat	F230572.mzid
M6_26	*O. indolifex*	*O. indolifex*	*O. indolifex*	*S. cerevisae*	*R. pomeroyi*	*K. aerogenes*	Q14801_M6_26	Q14801_M6_26	F230573.dat	F230573.mzid
M6_27	*S. cerevisae*	*K. aerogenes*	*R. pomeroyi*	*R. pomeroyi*	*R. pomeroyi*	*S. cerevisae*	Q14802_M6_27	Q14802_M6_27	F230574.dat	F230574.mzid
M6_28	*K. aerogenes*	*S. stellata*	*R. pomeroyi*	*O. indolifex*	*S. cerevisae*	*S. cerevisae*	Q14803_M6_28	Q14803_M6_28	F230575.dat	F230575.mzid
M6_29	*R. pomeroyi*	*K. aerogenes*	*S. cerevisae*	*R. pomeroyi*	*K. aerogenes*	*S. cerevisae*	Q14804_M6_29	Q14804_M6_29	F230576.dat	F230576.mzid
M6_30	*S. stellata*	*R. pomeroyi*	*O. indolifex*	*K. aerogenes*	*O. indolifex*	*R. pomeroyi*	Q14805_M6_30	Q14805_M6_30	F230577.dat	F230577.mzid
M6_31	*S. cerevisae*	*R. pomeroyi*	*K.aerogenes*	*S. stellata*	*K.aerogenes*	*R. pomeroyi*	Q14809_M6_31	Q14809_M6_31	F230581.dat	F230581.mzid
M6_32	*R.pomeroyi*	*K.aerogenes*	*S. cerevisae*	*S. stellata*	*S. cerevisae*	*K.aerogenes*	Q14810_M6_32	Q14810_M6_32	F230582.dat	F230582.mzid

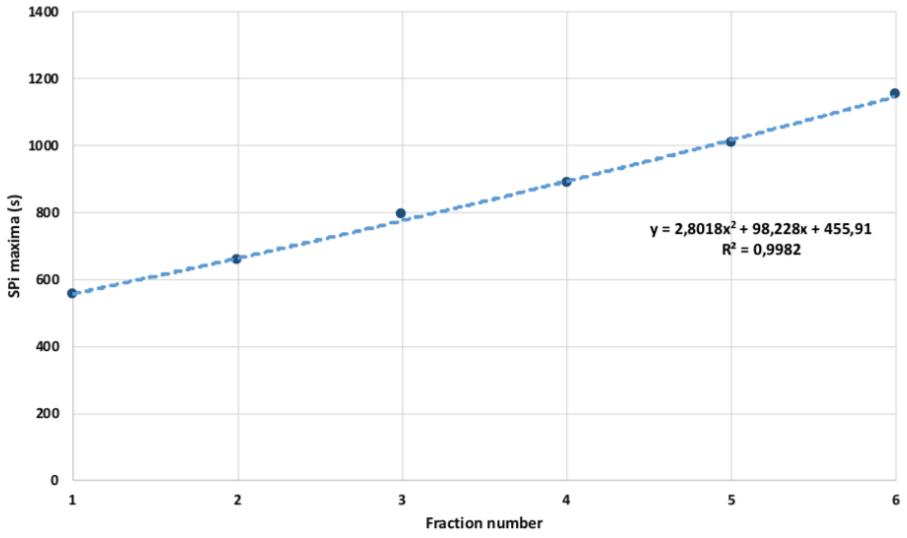


Figure S1. Correlation of SPi maxima obtained for the six microorganisms from the M610 assemblage. The six microorganisms, their fractions and their SPi maxima expressed in seconds are: *Sagittula stellata* (F1, 555.7), *Ruegeria pomeroyi* (F2, 658.8), *Oceanibulbus indoliflex* (F3, 791.7), *Saccharomyces cerevisiae* (F4, 887.0), *Klebsiella aerogenes* (F5, 10008.0), and *Marivirga tractuosa* (F6, 1152.0). The correlation is shown with dotted line.


M601
Oceanibulbus indolifex
Klebsiella aerogenes
Ruegeria pomeroyi
Sagittula stellata
Saccharomyces cerevisiae
Marivirga tractuosa
M602
Marivirga tractuosa
Oceanibulbus indolifex
Klebsiella aerogenes
Ruegeria pomeroyi
Sagittula stellata
Saccharomyces cerevisiae
M603
Klebsiella aerogenes
Ruegeria pomeroyi
Oceanibulbus indolifex
Saccharomyces cerevisiae
Sagittula stellata
Marivirga tractuosa
M604
Sagittula stellata
Ruegeria pomeroyi
Klebsiella aerogenes
Oceanibulbus indolifex
Marivirga tractuosa
Saccharomyces cerevisiae
M605
Ruegeria pomeroyi
Sagittula stellata
Saccharomyces cerevisiae
Marivirga tractuosa
Oceanibulbus indolifex
Klebsiella aerogenes
M606
Klebsiella aerogenes
Ruegeria pomeroyi
Sagittula stellata
Saccharomyces cerevisiae
Marivirga tractuosa
Oceanibulbus indolifex
M607
Klebsiella aerogenes
Oceanibulbus indolifex
Ruegeria pomeroyi
Sagittula stellata
Marivirga tractuosa
Saccharomyces cerevisiae
M608
Oceanibulbus indolifex
Marivirga tractuosa
Saccharomyces cerevisiae
Klebsiella aerogenes
Ruegeria pomeroyi
Sagittula stellata
M609
Ruegeria pomeroyi
Saccharomyces cerevisiae
Sagittula stellata
Marivirga tractuosa
Oceanibulbus indolifex
Klebsiella aerogenes
M611
Saccharomyces cerevisiae
Klebsiella aerogenes
Marivirga tractuosa
Ruegeria pomeroyi
Sagittula stellata
Oceanibulbus indolifex
M612
Marivirga tractuosa
Sagittula stellata
Klebsiella aerogenes
Saccharomyces cerevisiae
Oceanibulbus indolifex
Ruegeria pomeroyi
M613
Marivirga tractuosa
Saccharomyces cerevisiae
Sagittula stellata
Ruegeria pomeroyi
Klebsiella aerogenes
Oceanibulbus indolifex
M614
Saccharomyces cerevisiae
Sagittula stellata
Ruegeria pomeroyi
Klebsiella aerogenes
Oceanibulbus indolifex
Marivirga tractuosa
M615
Sagittula stellata
Ruegeria pomeroyi
Klebsiella aerogenes
Oceanibulbus indolifex
Marivirga tractuosa
Saccharomyces cerevisiae
M616
Ruegeria pomeroyi
Klebsiella aerogenes
Oceanibulbus indolifex
Marivirga tractuosa
Saccharomyces cerevisiae
Sagittula stellata
M617
Klebsiella aerogenes
Oceanibulbus indolifex
Marivirga tractuosa
Saccharomyces cerevisiae
Sagittula stellata
Ruegeria pomeroyi
M618
Oceanibulbus indolifex
Saccharomyces cerevisiae
Ruegeria pomeroyi
Marivirga tractuosa
Sagittula stellata
Klebsiella aerogenes
M619
Sagittula stellata
Oceanibulbus indolifex
Ruegeria pomeroyi
Marivirga tractuosa
Saccharomyces cerevisiae
Klebsiella aerogenes
SPI
Time(s)

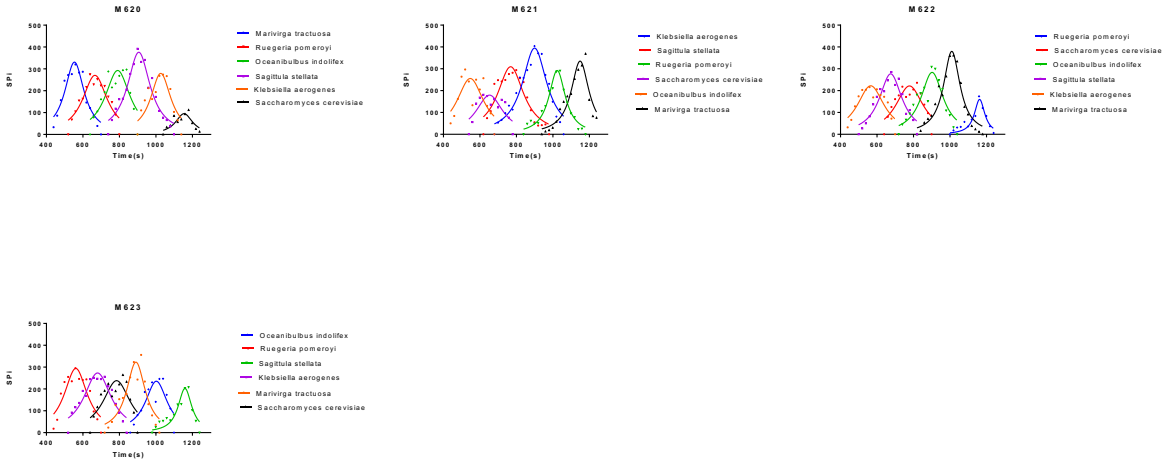


Figure S2. Scatter-plot of SPi values of each of the six organisms in 20-sec acquisition time windows for the assemblages M601 to M623. SPi values of each organisms are represented by colored dots for each assemblage as function of 20 sec acquisition time windows. Fitting curves are represented by colored lines for the different microorganisms.

Article 3 :

https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.202300372

List of identified species and strains (Table S1); List of identified proteins and peptides_ (Table S2); and Total ion current profile of the flash proteotyping sequence of ten successive injections of isolates (Figure S1).

Protéotypage haut-débit de microorganismes par spectrométrie de masse en tandem

Identifier rapidement des microorganismes est essentiel dans le domaine du diagnostic clinique, des contrôles sanitaires et alimentaires, et du criblage pour applications biotechnologiques. Améliorer les méthodes d'identification afin qu'elles soient plus rapides et sensibles est un enjeu de taille. Actuellement, le protéotypage par spectrométrie de masse MALDI-TOF est la méthode de référence pour les isolats bactériens dans les laboratoires de microbiologie clinique. Toutefois, cette méthodologie n'est pas en mesure de traiter la plupart des isolats environnementaux et des agents pathogènes opportunistes en raison d'une base de données de spectres expérimentaux incomplète. En enregistrant beaucoup plus d'informations sur les séquences au niveau des peptides, le protéotypage par spectrométrie de masse en tandem est capable d'identifier la position taxonomique de n'importe quel micro-organisme dans l'arbre de la vie, et peut s'avérer hautement discriminant au niveau des sous-espèces. Cette thèse a pour objectif d'adapter et rendre haut-débit une approche d'identification de microorganismes sans a priori par protéotypage de microorganismes, développée au laboratoire. La phylopeptidomique repose sur l'enregistrement de données de séquences peptidiques par spectrométrie de masse en tandem et l'association de ces données taxonomiques, permettant d'identifier l'organisme. Afin de réduire les coûts et le temps d'analyse, deux méthodes ont été inventées et testées durant ma thèse pour identifier tout type de microorganismes en quelques minutes d'analyse. La première méthode permet de multiplexer sans marquage plusieurs échantillons en créant un mélange contenant des fractions de chaque isolat qui différent en hydrophobicité. La robustesse, la reproductibilité et les limites de cette méthode ont été évaluées. La seconde méthode utilise les capacités de rapidité d'une analyse par infusion directe, permettant de réduire le temps d'analyse à 36 secondes de spectrométrie.

Mots clés : microorganismes, identification, protéotypage, spectrométrie de masse en tandem, taxonomie

High-throughput proteotyping of microorganisms by tandem mass spectrometry proteotyping

Rapid identification of microorganisms is essential in clinical diagnostics, health and food quality controls, and screening for biotechnology applications. Improving identification methods to make them faster and more sensitive is a major challenge. Currently, proteotyping by MALDI-TOF mass spectrometry is the reference method for isolates. However, this methodology is unable to handle most environmental isolates and opportunistic pathogens due to an incomplete database of experimental spectra. By recording much more sequence information at the peptide level, proteotyping by tandem mass spectrometry is able to identify the taxonomic position of any microorganism in the tree of life, and can be highly discriminating at the subspecies level. The aim of this thesis is to adapt and render high-throughput an approach without any a priori for microorganism identification by proteotyping developed in the laboratory. Phylopeptidomics is based on the recording of peptide sequence data by tandem mass spectrometry and their association with taxonomic information, enabling the organism to be identified. In order to reduce costs and analysis time, two methods were invented and tested during my thesis's work to identify any microorganism in just a few minutes of analysis. The first method enables sample multiplexing without labeling, by creating a mixture containing fractions for each isolate differing in hydrophobicity. The robustness, reproducibility and limitations of this method were evaluated. The second method takes advantage of the speed of direct infusion analysis, reducing analysis time to 36 seconds of spectrometry measurement.

Keywords: microorganisms, identification, high-throughput, proteotyping, tandem mass spectrometry