Université de Lille

Université de Lille – Sciences et Technologies

École doctorale Sciences de la Matière, du Rayonnement et de l'Environnement

Institut Charles Viollette

Thèse de Doctorat

Spécialité : Ingénierie des Fonctions Biologiques

Présentée pour l'obtention du grade de docteur de l'université de Lille

par

Mickaël Chevalier

Mise en place d'un *workflow* d'identification de microorganismes et de leurs métabolites secondaires d'origine non ribosomique par spectrométrie de masse

Préparée à l'institut Charles Viollette-ICV-EA 7394

Soutenue le 18 Décembre 2018, devant le jury composé de :

Mme Monique Royer	Chercheur (HDR), Montpellier, Cirad	Rapporteur
M. David Touboul	Chercheur (HDR), Gif sur Yvette, CNRS-ICSN	Rapporteur
M. Philippe Jacques	Professeur Univ. Liège, MiPI	Examinateur
M. Cédric Bertrand	Professeur Univ. Perpignan, CRIOBE	Examinateur
M. Edwin De Pauw	Professeur Univ. Liège, *MS Laboratory*	Examinateur
Mme Valérie Leclère	Professeur Univ. Lille, ICV	Directrice de thèse
M. Christophe Flahaut	Maître de conférences (HDR) Univ. Artois, ICV	Co-directeur de thèse
Mme Maude Pupin	Maître de conférences (HDR) Univ. Lille, CRIStAL	Invitée

Je remercie très sincèrement les membres du jury,

Madame le docteur Monique ROYER et Monsieur le docteur David TOUBOUL,

C'est un immense honneur pour moi de vous compter parmi les membres de ce jury en tant que rapporteur. Soyez assurés de mon estime et de mon plus grand respect pour avoir accepté cette lourde tâche.

Messieurs les Professeurs Edwin DE PAUW, Cédric BERTRAND et Philippe JACQUES,

Je suis très honoré que vous ayez accepté d'être les examinateurs de ce travail. Vos examens de cette thèse et des travaux qui y sont présentés seront un privilège. Veuillez agréer toute ma gratitude et mon plus profond respect.

Monsieur le Professeur Pascal DHULSTER,

Je vous remercie sincèrement de m'avoir accueilli au sein de votre laboratoire afin d'y réaliser mes travaux de recherche. Vous m'avez permis de vivre une expérience très enrichissante. Soyez assuré de ma plus grande reconnaissance.

Madame le Professeur Valérie LECLERE,

Il m'est impossible de te remercier à la hauteur de ton engagement tout au long de ma thèse. Je te remercie pour ton soutien sans faille dans toutes mes initiatives. Merci pour la clarté et la richesse de tes conseils. Ta grande disponibilité, ta sympathie et tes précieux conseils m'ont permis de travailler dans les meilleures conditions. Ton soutien et ta présence se sont avérés déterminants pour mener ce travail à terme. Merci pour nos nombreuses discussions qui ont à chaque fois permis de redéfinir les priorités de nos études et ont largement contribué à faire mûrir mes travaux de thèse. Un profond merci pour tout !

Monsieur le Docteur Christophe FLAHAUT,

J'ai tout d'abord su apprécier ton talent d'enseignant et de pédagogue durant ma dernière année de licence à la faculté Jean Perrin. Puis tu m'as accompagné dans la fameuse épreuve du Master Recherche et pour finir dans la réalisation de cette thèse. Je tiens à t'exprimer toute ma reconnaissance pour ce long suivi et pour cet apprentissage du métier de chercheur. Je te remercie également pour ta disponibilité et ton suivi constant de mon travail ainsi que pour la confiance accordée. Merci pour les sacrifices familiaux, je suis

conscient des nombreuses soirées et week-end sacrifiés pour moi. Merci de m'avoir transmis ta passion de la spectrométrie de masse. Un profond merci pour tout !

Je souhaite exprimer de sincères remerciements envers le **Docteur Maude PUPIN**, qui a participé à mon encadrement et m'a donné de précieux conseils tout au long de ma thèse sans jamais rien attendre en retour, merci beaucoup pour cette implication. J'en profite pour remercier également Yoann Dufresne et Areski Flissi membres du laboratoire CRIStAL pour leur support et leur aide technique durant ma thèse.

Madame le Docteur Frédérique LISACEK et Emma RICART,

Merci pour cette collaboration qui s'est établie maintenant depuis plus de trois ans. Merci pour votre accueil lors de ma venue dans votre belle ville de Genève. Merci également de m'avoir suivi dans mes choix et mes initiatives. C'est un grand privilège de pouvoir collaborer avec des chercheurs tels que vous. Emma, je te souhaite bon courage pour la suite de ta thèse, je te remercie également pour toutes les discussions qui ont permis de faire émerger des idées et m'ont permis d'aboutir à ce modeste travail. Je te remercie également pour ta sympathie et ta bonne humeur à chacune de nos réunions.

Monsieur le Professeur Gérard HOPFGARTNER,

Je vous remercie dans un premier temps de votre accueil, vous m'avez non seulement donné accès à vos équipements scientifiques et votre laboratoire, mais vous m'avez également formé sur l'appareillage et les subtilités qui en découlent. Je souhaite également remercier les membres du laboratoire LSMS de Genève qui m'ont accueilli les bras ouverts, je pense à Éliane, Chantal, mais aussi Sophie, Laura, Thomas, Piotr, David, Michel et Anita. Un grand merci pour votre accueil !

J'aimerais adresser de profonds remerciements au laboratoire MSLab de Liège dirigé par monsieur le professeur De Pauw. Merci de votre accueil, mais également de m'avoir accordé l'accès à un équipement de haute performance et de m'y avoir initié. Je remercie le très sympathique Nicolas Smargiasso, mais aussi Emeline Hanozin sans oublier les autres membres du laboratoire avec qui j'ai passé de très bons moments.

J'exprime également tous mes remerciements aux membres de l'UFR de biologie, Brigitte Delrue, Olivier Vidal, Virginie Cogez et Edwige Madec avec qui j'ai pu apprendre à enseigner

durant ses deux dernières années, je vous remercie vivement pour vos conseils, votre bienveillance et votre pédagogie. Je remercie également Estelle Hamon et Peggy Gruau pour leur dévouement et leur gentillesse, sans oublier Jean-Marie Lacroix qui m'a confié cette mission d'enseignement.

Je tiens à remercier l'équipe de l'Xperium mais notamment Sophie Picart et Jean Cosleou qui m'ont permis d'exercer un rôle de vulgarisation scientifique au large public pendant ma première année de thèse.

Je tiens à remercier certains enseignants de l'éducation nationale qui ont participé à mon développement intellectuel depuis mon enfance, une pensée pour Monsieur Deleu, Madame Pignard, Madame Zurawski et Monsieur Sauvage du collège Jules Verne, mais aussi Madame Martin qui m'a donné le goût pour la biochimie depuis le lycée et Monsieur Finet du lycée Henri Darras. Il y a des enseignants que l'on n'oublie jamais.

J'adresse de vifs remerciements à tous les chercheurs, les enseignants et le personnel, notamment du groupe « NRPS », que j'ai croisés au cours de ces années et avec qui j'ai pu discuter et échanger des points de vue.

Je souhaite à présent exprimer toute mon amitié envers l'ensemble des post-doctorants, doctorants et stagiaires que j'ai côtoyés. En commençant par les anciens, je citerais Luiz, Kalim, Adil, Amirouch, Juliette, Qassim, Yazin, Alaa, Debarun, Ameen, Marwane, Thibaut, Alexandra, Alexandre C, Sabrine, Oumaima, Menouar, Mahammed, Laetitia, Delphine, Leandro, Abdul, Roxane, Yannath et Rémi.

Je souhaite une bonne continuation aux futurs docteurs, je pense à Audrey, Hannah, Sandy, Justine, Elodie, Alice, Nathalie, Ludivine, Mégane, Nuria, Barbara F, Dhalia, Léa, Mouna, Alexandre Be, Alexandre Br, Quentin, Antoine, Cyril et Maxime.

Merci aux nombreux ingénieurs et techniciens que j'ai pu solliciter au cours de mes travaux, je citerais Gabrielle Chataigné, Barbara Deracinois avec qui j'ai partagé ma paillasse en biochimie et nos soucis en MS. Matthieu Duban, Max Béchet, Delphine Caly, Camille Dugardin, Florie Desriac, Hugues Mandavid et Corinne Boistel. J'en profite aussi pour remercier l'indispensable Cathy Oublion.

Je remercie l'équipe de Deinobiotics de m'avoir donné un accès à leur équipement, mais aussi pour leur sympathie durant ces trois années, je remercie Dominique Le Beller, Jozef Azodi, Denis Carniato, Prema Levasseur, Guillaume Lesquame, Marjorie Valenduc, Yannick Eveno, Paul Duhamel et Louis Guitard.

Je remercie très sincèrement les membres de la plateforme Realcat, j'ai passé la plupart de mon temps sur la plateforme et je vous suis sincèrement reconnaissant pour votre aide, vos conseils et vos encouragements, je remercie Sébastien Paul, Svetlana Heyet, Joëlle, mais surtout Egon Heuson (merci pour tes conseils et ta joie de vivre).

Je remercie chaque membre de l'institut Charles Viollette et du personnel de l'Université de Lille, je remercie Rezak, Jonathan et Adrien et tous les autres. C'est toujours un plaisir de croiser l'un de vous dans un laboratoire, en salle de convivialité ou encore dans un couloir que ce soit pour partager des moments scientifiques ou tout simplement bavarder. Bonne continuation à tous !

Je remercie toutes les personnes, qui de près ou de loin, ont contribué, de n'importe quelle façon, à la réalisation de ce manuscrit. Trouvez ici l'expression de ma gratitude.

J'aimerais remercier mes amis avec qui j'ai grandi, je pense, à la famille Willbaux au complet, Sylvie, Jean-Pierre, Noémie, Élodie et plus particulièrement Christopher, mais aussi Nico et Théo vous êtes une partie de moi-même merci pour tous les bons moments passés ensemble. Je remercie également ceux qui partagent ma vie depuis moins longtemps, je pense à Sophia, Lucie, Angelina, Maxime, Xavier et Matthieu, merci pour votre bienveillance, merci pour les nombreux fous rires et pour tous ces moments de joie.

J'ai une pensée énorme pour mes proches qui m'ont soutenu depuis le commencement. Je tenais à vous dire que vous êtes la source d'énergie m'ayant permis d'aboutir à ce travail.

Je remercie ma belle famille, j'ai la chance d'avoir rejoint une très grande famille, mais cette chance n'est rien face à toutes les personnalités qui la composent. Je remercie tout d'abord Évelyne et André, vous êtes les premiers à m'avoir accueilli dans votre famille, je vous remercie pour votre bienveillance, mais aussi pour toutes les soirées partagées avec vous qui m'ont permis d'apprendre à vous connaître. Merci pour tout !

Je remercie mes beaux parents Sabrina et Daniel pour leurs conseils de la vie, mais aussi Noa et Jules, c'est toujours un plaisir de partager du temps avec vous, surtout en bord de piscine durant les vacances d'été.

Je remercie Isabelle et Laurence pour leur joie de vivre (leur grain de folie) indéfectible, je remercie Virginie, Sébastien pour leur gentillesse. Un grand merci aussi à Daniel et Laurent avec qui je passe beaucoup de temps durant les longs repas de famille, merci pour toutes les discussions de tout et de rien. Merci à tous leurs enfants, merci à Pauline, Sarah-Laure, Lucas, Flora, Solène, Louise et Jeanne. Une spéciale dédicace à Laurence qui a eu la gentillesse et la patiente de relire mon manuscrit, merci beaucoup !

Je remercie également mes beaux parents Sylvie et Arnaud pour leur soutien inconditionnel, leur présence, merci pour tous les bons moments passés ensemble depuis maintenant plusieurs années, merci de m'avoir intégré à vos vies. Merci à Annabelle, Bastien et Laura pour leur soutien et leur présence, merci pour tous les bons moments, merci d'être vous !

Merci à Christophe et Laurence pour leur gentillesse, leur soutien vous êtes également des exemples pour nous. Merci à Adrien, Théo et Inès pour leur simplicité et leur joie de vivre simplement. Merci également à Adrien et Nicole vous êtes sans aucun doute les gens les plus simples et gentils que j'ai rencontrés jusqu'ici.

Finalement, je tiens à remercier ma très chère famille pour la joie et le bonheur permanent qu'elle m'apporte, mais aussi pour le soutien constant de mes sœurs Nathalie et Léa et de mon frère Arnaud. Jean-Marie merci de m'avoir accueilli et élevé comme un fils. Nathalie, merci pour tous les souvenirs d'enfance que nous partageons ensemble, « Le passé comme un petit musé, je m'y promène pour faire du bien », je n'oublie rien. Léa merci d'être toi en toute simplicité, tu es lumineuse, je t'aime ma bichette. Merci à Mathieu, prends soin de ma sœur et de mon magnifique neveu Raphaël.

Il me reste encore a remercier quelqu'un qui m'a permis de réaliser tous mes objectifs en m'inculquant tout ce qu'il y a de meilleur en moi. Maman, un simple merci serait une insulte, mais je peux toujours te dire que je t'aime.

Je remercie plus particulièrement la femme qui partage ma vie, Hélène, ma nébuleuse, tu es la plus belle rencontre que j'ai pu faire, tu partages ma vie depuis maintenant plus de dix ans, tu es devenue ma confidente, ma meilleure amie et mon amour. Tu es à présent ma plus grande source d'inspiration, mon garde-fou. Tu es lumineuse, ta joie de vivre et ton amour pour moi me permettent de m'épanouir pleinement. Il n'existe pas de mot pour décrire les sentiments que j'ai pour toi alors à défaut je te dis : « je t'aime ».

Je dédie ce modeste travail à

Ma mère, c'est ma reine, je lui dois tout, c'est mon exemple de force.

Sommaire

Liste des figures

Liste des tableaux

Liste des Equations

Abréviations

ACN : acétonitrile

CHCA : acide α-Cyano-4-hydroxycinnamique

CID : dissociation induite par collision

Da : Dalton

DHB : acide 2,5-Dihydroxybenzoïque

ESI : ionisation par électronébuliseur

FT-ICR : transformée de Fourier – résonance cyclotronique ionique

HPLC : chromatographie liquide haute performance

Hz : Hertz

IMS : spectrométrie de masse à mobilité ionique

KM : masse de Kendrick

KMD : défaut de masse de Kendrick

KNM : masse nominale de Kendrick

LC : chromatographie liquide

m/z : rapport masse sur charge

MALDI : ionisation/désorption laser assistée par matrice

MPTs : modifications post-traductionnelles

MS : spectrométrie de masse

MS/MS : spectrométrie de masse en tandem

NRPs : peptide non ribosomiques

NRPS : synthétase peptidique non ribosomique

PKs : polycétides

RKMD : défaut de masse de Kendrick régulier

TFA : Acide trifluoroacétique

TOF : temps de vol

UFC : unité formant colonie

UV : ultraviolet

Résumé

Les peptides non ribosomiques (NRPs) sont des métabolites secondaires microbiens qui présentent un intérêt important pour plusieurs domaines d'activités tels que l'environnement (biopesticides), la pharmacologie (antibiotiques, antitumeurs …), la cosmétique ou l'agro-alimentaire. Leurs structures diverses et variées et leurs activités sont intrinsèquement liées aux monomères qui les composent. Cette impressionnante diversité structurale pose des problèmes d'identification simple et rapide de leur structure moléculaire.

Alors qu'il existe un nombre important de méthodes et d'outils à notre disposition pour caractériser diverses molécules telles que l'ADN (génomique), l'ARN (transcriptomique), les protéines (protéomique) ou encore les métabolites (métabolomique), peu de méthodes et d'outils existants ne semblent réellement adaptés et transposables aux peptides qui ne sont pas issus de la synthèse classique ribosomique. À ce titre, il nous a semblé essentiel d'ajouter un maillon à cette cascade « omique » pour étudier des métabolites secondaires comme les NRPs. En effet, les spécificités des NRPs justifient la définition de «**NRPomics**» comme l'étude systématique des NRPs qui implique la caractérisation complète, dynamique, qualitative et quantitative des NRPs présents dans les échantillons biologiques.

Afin de répondre aux objectifs fixés, nous avons choisi de réaliser une série d'analyses séquentielles appelée « *Workflow* » ou flux de travail permettant l'incrimination, la discrimination et l'identification de peptides non ribosomiques connus ou inconnus dans une philosophie tournée vers le criblage de nouveaux composés actifs. Reposant sur les informations contenues dans les bases de données, cette combinaison de méthodes analytiques commence par **l'identification du micro-organisme par une méthode de profilage phénotypique par spectrométrie de masse (MS).** Ensuite, afin de valoriser au mieux les données de MS souvent peu exploitées en métabolomique à l'instar et au profit de la fragmentation par spectrométrie de masse (MS/MS). Dès lors, nous avons choisi de combiner pour la première fois une méthode de calcul itérative élaborée dans les années mille neuf cent soixante par un chimiste anglais, Edward Kendrick et les informations contenues dans la base de données, dénommée Norine dédiée aux peptides non ribosomiques. Cette combinaison nous a permis de **déterminer la composition élémentaire**

d'un peptide non ribosomique à partir de données de spectrométrie de masse haute résolution combinées à un maillage vectoriel reliant les différents NRPs de la base de données Norine et les formules chimiques. Nous illustrons également que d'une part cette méthode démontrée à partir des données MS peut-être extrapolée aux données de fragmentation de spectrométrie de masse en tandem (MS/MS) et que d'autre part elle présente **un intérêt pour la déréplication des NRPs mais aussi la caractérisation structurale de nouveaux composés actifs** pour des applications en particulier dans les secteurs de la santé et phytosanitaires. La performance du workflow sera illustrée par l'identification de lipopeptides produit par des souches de *Pseudomonas*. Ces lipopeptides sont particulièrement intéressants car ce sont des composés ayant des applications potentielles en biocontrôle.

Abstract

Nonribosomal peptides (NRPs) are microbial secondary metabolites that are of major interest for several fields of activity such as the environment (biopesticides), pharmacology (antibiotics, antitumors ...), cosmetics or agrifood. Their diverse and varied structures and their activities are intrinsically linked to the monomers that compose them. This impressive structural diversity poses problems regarding the simple and rapid identification of their molecular structure. While there are a number of methods and tools at our disposal to characterize various molecules such as DNA (genomic), RNA (transcriptomic), proteins (proteomics) or metabolites (metabolomics), Few existing methods and tools seem to be truly adapted and transposable to peptides that are not derived from classical ribosomal synthesis. As such, it seemed essential to add a link to this "omic" cascade to study secondary metabolites such as NRPs. Indeed, the specificities of NRPs justify the definition of "NRPomics" as the systematic study of NRPs that involves the complete, dynamic, qualitative and quantitative characterization of NRPs present in biological samples.
In order to meet the set objectives, we chose to carry out a series of sequential analyzes called "Workflow" allowing the incrimination, the discrimination and the identification of known or unknown NRPs in a philosophy turned towards the screening of new active compounds. Based on the information contained in the databases, this combination of analytical methods begins with **the identification of the microorganism by a phenotypic profiling method by mass spectrometry (MS)**. Then, in order to make the best use of MS data, which are often less used in metabolomics, for the benefit of fragmentation by mass spectrometry (MS / MS). Therefore, we chose to combine for the first time an iterative calculation method developed in nineteen sixties by an English chemist, Edward Kendrick, and the information contained in the database, called Norine dedicated to NRPs. This combination allowed us to **determine the elemental composition of a NRP from high resolution mass spectrometry data combined with a vector mesh linking the different NRPs of the Norine database and the chemical formulas**. We also illustrate that on the one hand that this method, demonstrated from the MS data, can be extrapolated to MS / MS fragmentation data and that, on the other hand, it is of **interest for the dereplication of the NRPs but also the structural characterization of new active compounds** for applications in particular in the health and phytosanitary sectors. The workflow performance will be illustrated by the identification of lipopeptides produced by Pseudomonas strains. These

lipopeptides are particularly interesting because they are compounds with potential applications in biocontrol.

Avant-propos

Dans le cadre de cette thèse, je vais tenter d'allier le monde de la microbiologie et celui de la spectrométrie de masse (*mass spectrometry*, MS) qui constitue un champ multidisciplinaire très varié (chimie, physique, mathématiques, microbiologie, bon sens …) faisant appel à de nombreux axiomes (bases) plus ou moins connus.

Même si cette thèse ne traite pas d'ingénierie ou de développement sur l'appareillage que sont les spectromètres de masse, il me semble intéressant au vu de mes travaux de thèse de décrire a minima les équipements disponibles afin, je l'espère, d'aider à la compréhension des choix pris au cours de ces travaux.

État de l'art

Partie I : La caractérisation moléculaire, la naissance des études « Omiques »

1. Génomique

La génomique constitue l'étude du génome dans tous ses états avec pour alphabet quatre bases nucléotidique, adénine (A), thymine (T), guanine (G) et cytosine (C) qui constituent le matériel génétique constitutif de l'acide désoxyribonucléique (ADN), excepté pour certains virus où le matériel génétique est de l'acide ribonucléique (ARN). La manière dont ces quatre bases sont organisées constitue le code génétique.

1.1. Le séquençage

Vers la fin des années 1970, une première méthode analytique permettant de déterminer l'enchaînement (séquençage) des bases nucléotidiques fut mise au point par Allan M. Maxam et Walter Gilbert (Maxam & Gilbert, 1977) suivie la même année d'une seconde : la méthode de Frederik Sanger (Sanger *et al.*, 1977). Ces deux méthodes ont révolutionné la biologie. Cependant, de par ses avantages (plus pratique, moins coûteuse et sans utilisation de la radioactivité…), la méthode de Sanger est devenue la méthode de choix et a ensuite connu plusieurs améliorations comme l'utilisation de fluorophores (Smith *et al.*, 1986) et l'automatisation (séquenceur d'ADN). C'est ce qui a conduit au séquençage à haut débit permettant de séquencer plusieurs milliards de bases en quelques jours contre plusieurs années il y a quarante ans.

1.2. Début de l'ère omique

Les avancées technologiques de ces dernières années ont créé un éventail d'appareils de plus en plus performants, capables d'analyser non seulement de manière précises et sensibles mais également avec des débits analytiques de plus en plus importants. Les approches « omiques » issues de cette expansion technologique corrélée à la

démocratisation de l'informatique ont permis l'étude de façon très fine de la biologie au sens large du terme. Ce sont les bioinformaticiens et les biologistes moléculaires qui sont les premiers à utiliser le suffixe « omiques » (-*omics*).

1.3. Banque de données génomiques

La démocratisation du séquençage haut débit a généré une quantité très importante de séquences nucléotidiques que les progrès parallèles de l'informatique ont permis de stocker sous forme de banques de données.

Il existe 3 banques de données nucléotidiques à accès libre réparties sur 3 continents : **i**) *Genbank* (Benson *et al.*, 2013) aux USA au centre national pour l'information en biotechnologie (*national center for biotechnology information*, NCBI), **ii**) la banque de données Européenne (*european molecular biology laboratory*, EMBL)(Li *et al.*, 2015; McWilliam *et al.*, 2013) gérée par l'institut Européen de bioinformatique (*european bioinformatics institute*, EBI) et **iii**) la banque de données d'annotation et d'assemblage des séquences (*databases annotated/assembled sequences*, DDBJ)(Mashima *et al.*, 2016) en Asie administrée par l'institut national de génétique (*national institute of genetics*, NIG).

Les banques de données recensent de nombreuses informations génériques comme le nom de l'organisme, le nombre de paires de bases, plusieurs références, chaque gène et bien sûr les séquences nucléotidiques sous divers formats informatiques comme le format FASTA. Ce fichier possède une organisation simple et facilement manipulable utilisant le code à une lettre des nucléotides : adénine (A), thymine (T), guanine (G) et cytosine (C) défini par l'union internationale de chimie pure et appliquée (*international union of pure and applied chemistry*, IUPAC). Ceci permet un stockage aisé des séquences.

1.4. La bioinformatique (le biologiste 2.0)

La bioinformatique est une discipline qui prend naissance naturellement vers la fin du 20^{éme} siècle. La bioinformatique est un domaine de recherches à part entière. Elle associe un nombre important de disciplines telles que la biologie moléculaire, l'informatique, les mathématiques afin de résoudre un problème biologique. La bioinformatique est composée

d'un ensemble de concepts et de techniques nécessaires à l'interprétation informatique de l'information biologique (Gerritsen *et al.*, 2016).

La bioinformatique va de l'analyse du génome à la modélisation de l'évolution de tout système vivant normal ou pathologique dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'images et l'assemblage de génomes. Au départ, la bioinformatique est utilisée pour stocker et manipuler la quantité importante de données générées par les biologistes. Le développement rapide de cette discipline permet de nos jours d'analyser et d'assembler le génome, de modéliser des systèmes vivants sains ou pathologiques dans divers environnements ou encore de modéliser des molécules.

Ainsi des outils d'alignement et de comparaison de séquences ont été mis en place comme le célèbre et encore très utilisé outil d'alignement (*basic local alignment search tool,* BLAST). Ce dernier a également permis en association avec le code génétique une traduction haut débit de toutes les séquences nucléotidiques en séquences protéiques potentielles.

En raison de la réussite des projets de séquençage à grande échelle, le suffixe omique s'est démocratisé et s'est étendu à une foule d'autres disciplines, parmi elles, la protéomique.

2. Le protéome

2.1. La voie de synthèse des protéines

La synthèse ribosomique est la seule voie de synthèse des protéines. L'ADN est dans un premier temps transcrit en ARN grâce à une polymérase puis comme son nom l'indique, cette synthèse fait appel à un organite de nature ribonucléoprotéique appelée le ribosome. Celui-ci est capable à partir de l'ARN messager d'agencer les acides aminés les uns après les autres afin de donner naissance à une séquence protéique primaire.

Les acides aminés incorporés dans la synthèse protéique sont appelés les acides aminés protéogéniques (Ambrogelly *et al.*, 2007). Ils sont classiquement au nombre de vingt et deux supplémentaires ont été plus récemment décrits chez les microorganismes uniquement (la sélénocysteine et la pyrrolysine). Tous les acides aminés, à l'exception de la glycine sont des molécules chirales (leurs images dans un miroir ne sont pas superposables) constituées d'un

carbone asymétrique. Par conséquent, les acides aminés présentent une stéréoisomérie dextrogyre (D) ou lévogyre (L), seuls les acides aminés L entrant dans la composition des protéines. Après synthèse, les protéines peuvent être modifiées de différentes manières le plus souvent par réaction enzymatique (méthylation, oxydation, phosphorylation …). Toutes ces modifications sont regroupées sous le terme de modifications post-traductionnelles (MPTs) et ne sont par conséquent pas codées directement par le génome.

2.2. La protéomique

La protéomique est l'étude de l'ensemble des protéines d'une cellule, d'un organite, d'un tissu ou d'un organe à un moment donné et sous des conditions données. Dans la pratique, la protéomique s'attache à identifier et/ou quantifier de manière globale les protéines extraites d'une culture cellulaire, d'un tissu ou fluide biologique, leur localisation dans les compartiments cellulaires, leurs éventuelles modifications post-traductionnelles, leur interaction avec d'autres protéines ou molécules biologiques ou non… Elle permet de quantifier les variations de taux d'expression en fonction du temps, de l'environnement, de l'état de développement, de l'état physiologique et/ou pathologique, de l'espèce d'origine. Elle étudie aussi les interactions que les protéines ont avec d'autres protéines, avec l'ADN ou l'ARN, ou d'autres substances (Wilkins *et al.*, 1996).

Depuis ces vingt dernières années, la protéomique n'a cessé de croitre comme en témoigne le nombre croissant de publications dans le domaine et la taille des données générées par la protéomique (Figure 1).

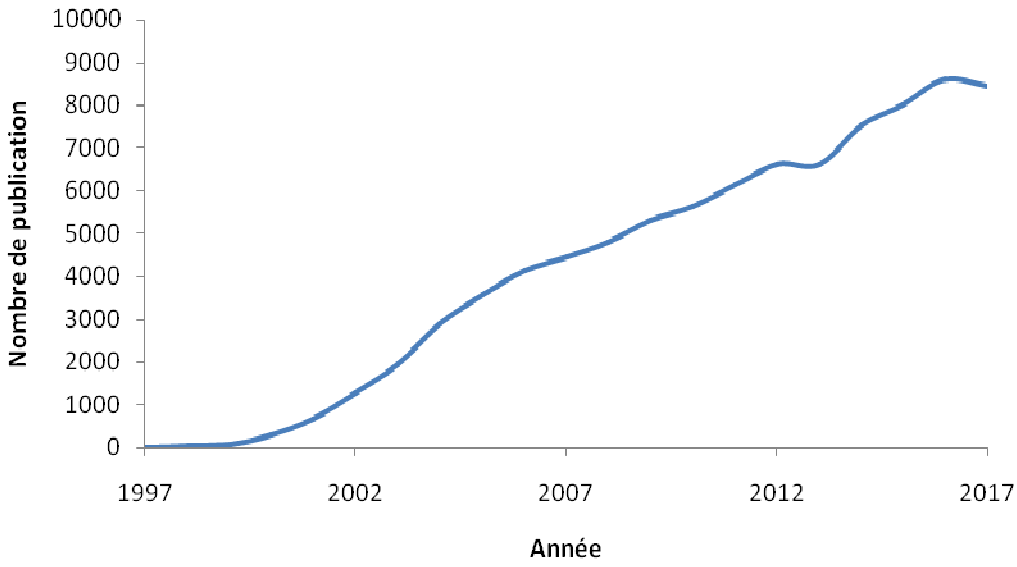


Figure 1. Représentation du nombre de publications annuelles de 1997 à 2017 contenant le mot « *proteomic*», recensées sur la plateforme NCBI.

2.3. Banque de données protéiques

Ce sont les programmes de séquençage qui nous ont légué en héritage de grandes banques de données de séquences nucléotidiques (voir section 1.3). La plupart des séquences protéiques des banques de données correspondantes sont issues de séquences nucléotidiques qui ont été assemblées et annotées automatiquement. Ainsi les séquences protéiques présentes dans les bases de données peuvent malgré tout contenir des erreurs issues de ces deux procédés. Dès lors, on distingue deux types de banques de données protéiques : **i**) les banques de données brutes (souvent annotées automatiquement) et **ii**) les banques de données vérifiées (annotées manuellement et contrôlées par le gestionnaire de l'unité de stockage) et appelées dans ce cas simplement bases de données.

Les banques de données protéiques sont organisées selon le même modèle que les banques de données génomiques. Chaque entrée de la banque de données correspond à une protéine (ou polypeptide) et est divisée en champs d'informations (numéro d'accession, nom de la protéine, espèce etc.). L'un des champs correspond à la séquence en acides aminés de la protéine au format FASTA.

Un consortium mondial (*worldwide protein data bank*, wwPDB) a permis la centralisation et la création d'une banque de données protéiques avec des données issues d'Amérique (*research collaboratory for structural bioinformatics*, RCSB) (Berman, 2000), d'Europe (PDBe/ UniProt) et d'Asie (PDBj) (Liu *et al.*, 2015). De nos jours, les bases de données renferment une richesse de données cruciales qui constituent un patrimoine scientifique inestimable. Pour exemple, le recensement des MPTs, au travers de multiples études dont les premières rapportées par les travaux de Fischer *et al.* et Phillips *et al.* (Fischer *et al.*, 1959; Phillips, 1963) ont permis de découvrir un bon nombre de ces dernières. Ces modifications post-traductionnelles ne peuvent se déduire de l'information génétique. Leur découverte, leur caractérisation moléculaire sont totalement dépendantes des méthodes d'analyse structurale des molécules. Il est important de noter que les banques de données génomiques et protéiques sont en accès libre et totalement gratuit pour n'importe quel chercheur.

2.4. La bioinformatique

La manipulation de données protéiques jouit d'un espace plus vaste directement corrélé à la structure plus complexe des protéines. En effet, la protéomique dispose d'un alphabet plus grand (20 acides aminés contre 4 nucléotides) et des modifications post-traductionnelles très variées en nombre et en structure (Wani *et al.*, 2015). Ainsi en témoigne la plateforme de ressource bioinformatique de l'institut de bioinformatique Suisse appelé ExPASy (www.expasy.org) qui met à disposition 243 outils dédiés à la protéomique contre 63 pour la génomique. Parmi les outils développés pour l'analyse des séquences protéiques, on retrouve une version adaptée aux protéines du logiciel d'alignement de séquence local de base BLAST (BLASTp pour les protéines) et bien d'autres logiciels d'analyse structurale.

3. Le métabolome

3.1. Définition du métabolome

Deux types de métabolites, appelés métabolites primaires et secondaires sont produits par les cellules. Les **métabolites primaires** sont associés aux fonctions vitales de la cellule. Aucun organisme ne peut vivre sans les grandes voies métaboliques relatives aux acides aminés, aux sucres, aux lipides, aux acides nucléiques, à l'énergie cellulaire... Je ne développerai pas les aspects relatifs aux métabolites primaires parce qu'ils ne sont pas au cœur des travaux réalisés dans le cadre de cette thèse. Par contre, **les métabolites secondaires** (aussi appelés produits naturels) sont des molécules organiques qui ne sont pas directement impliquées dans le développement ou encore dans la reproduction d'un organisme. Leur absence n'est pas létale (tout du moins immédiatement), mais peut conditionner la survie, l'apparence ou encore la multiplication du microorganisme. Cette absence peut aussi n'avoir aucun effet. Les métabolites secondaires vont procurer un avantage à un microorganisme sur d'autres microorganismes. L'organisme qui les synthétise augmente alors sa compétitivité (LaRossa, 2015; Welker, 2011).

3.2. Mode de production des métabolites secondaires

3.2.1. Synthèse ribosomique

Les métabolites secondaires sont de différentes natures chimiques mais beaucoup sont de nature peptidique ou partiellement peptidique. Certains métabolites secondaires partagent la même voie de synthèse ribosomique que les protéines pour donner naissance à des métabolites secondaires de nature peptidique (les bactériocines par exemple). Dans ce cas précis, comme pour une protéine, la composition en acides aminés du métabolite peut être déduite directement de la séquence du gène codant. De la même manière, ces métabolites secondaires de nature ribosomique sont susceptibles de subir des MPTs.

3.2.2. Synthèse Non Ribosomique (NRPS)

3.2.2.1. Les domaines majeurs des NRPS

Conjointement à cette voie de synthèse ribosomique, il existe chez les microorganismes un autre mode de synthèse dans lequel le ribosome ne joue pas de rôle direct dans la synthèse (la voie ribosomique servant à synthétiser les protéines responsables de la synthèse des métabolites secondaires). Les peptides non ribosomiques (*nonribosomal peptides*, NRPs) sont synthétisés par de larges complexes enzymatiques appelés *nonribosomal peptide synthetase* (NRPS). Les NRPS sont des protéines issues de la voie de synthèse ribosomique des protéines (Tomino *et al*, 1967). Elles sont organisées en modules, chacun, responsable de l'incorporation d'un monomère spécifique (Winn *et al.*, 2016). Chaque module se divise en plusieurs domaines enzymatiques possédant chacun un rôle défini au sein de la synthèse. Chaque module comprend au minimum trois domaines catalytiques :

-Le domaine d'adénylation (domaine A) est responsable non seulement de l'incorporation plus ou moins spécifique du monomère mais aussi de son activation en aminoacyl-adénylate. Cette forme « activée », de l'acide aminé, lui apporte l'énergie nécessaire pour être utilisé par le domaine suivant (Figure 2).

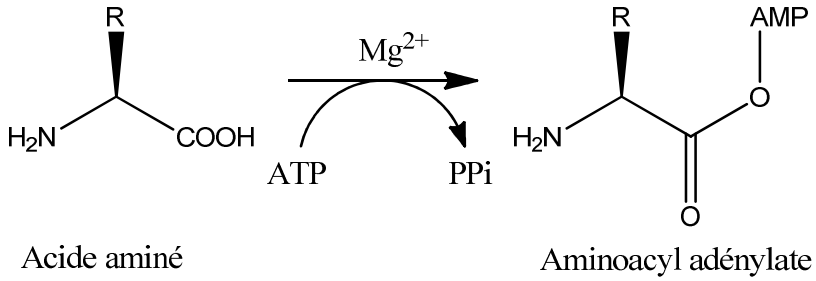


Figure 2. Réaction catalysée par le domaine d'adénylation des NRPS.

-Le domaine de thiolation (domaine T, ou *peptidyl carrier protein,* PCP) porte un groupement prosthétique dérivé de la pantéthéine (phosphopantéthéine) qui réalise la liaison entre la synthétase et le monomère afin d'assurer le transport au cours de l'élongation (Figure 3).

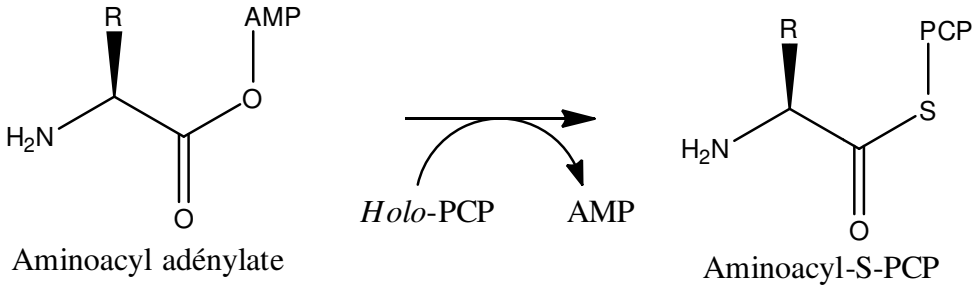


Figure 3. Réaction catalysée par le domaine de thiolation des NRPS.

-Le domaine de condensation (domaine C) forme la liaison peptidique entre le monomère porté par le domaine PCP d'un module et le monomère porté par le domaine PCP du module suivant (Figure 4).

Figure 4. Réaction catalysée par le domaine de condensation des NRPS

-Le domaine de thioestérase (domaine TE) libère le peptide du complexe enzymatique (Figure 5). De par sa fonction, il est le dernier domaine du dernier module de la synthétase. Il est également parfois responsable de la cyclisation du peptide.

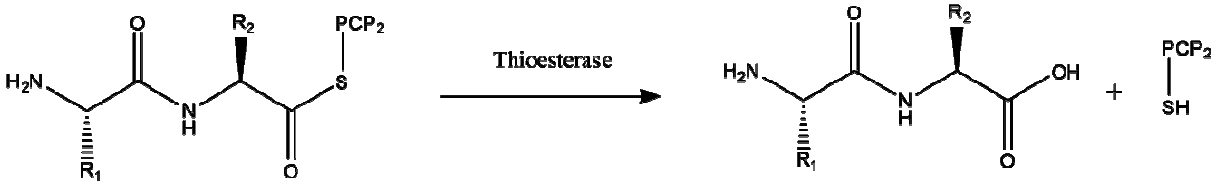


Figure 5. Réaction catalysée par le domaine de thioestérase des NRPS.

Pour résumer, les acides aminés ou monomères sont activés par réaction avec l'ATP pour former un intermédiaire aminoacyl-AMP, catalysé par un domaine d'adénylation (A). L'intermédiaire aminoacyl-AMP est ensuite capturé par le groupe thiol du bras flexible 4'-phospho-pantétheine lié à un domaine de thiolation (T). Les domaines de condensation (C) catalysent la formation de liaison peptidique successive entre les intermédiaires thioester chargés sur les domaines T adjacents. Le premier module est appelé module d'initiation (M1) et les modules suivants sont appelés modules d'élongation. Chaque module incorpore un

seul acide aminé, par conséquent, il y a, en général, autant de modules requis qu'il y a d'acides aminés dans le peptide final. Le dernier module contient un domaine thioestérase (TE) qui catalyse l'hydrolyse ou la cyclisation pour libérer le peptide de la NRPS (Figure 6). Les modules peuvent contenir des domaines supplémentaires, notamment des domaines d'épimérisation (E), de N-méthylation (NMT) et de cyclisation (Cy).

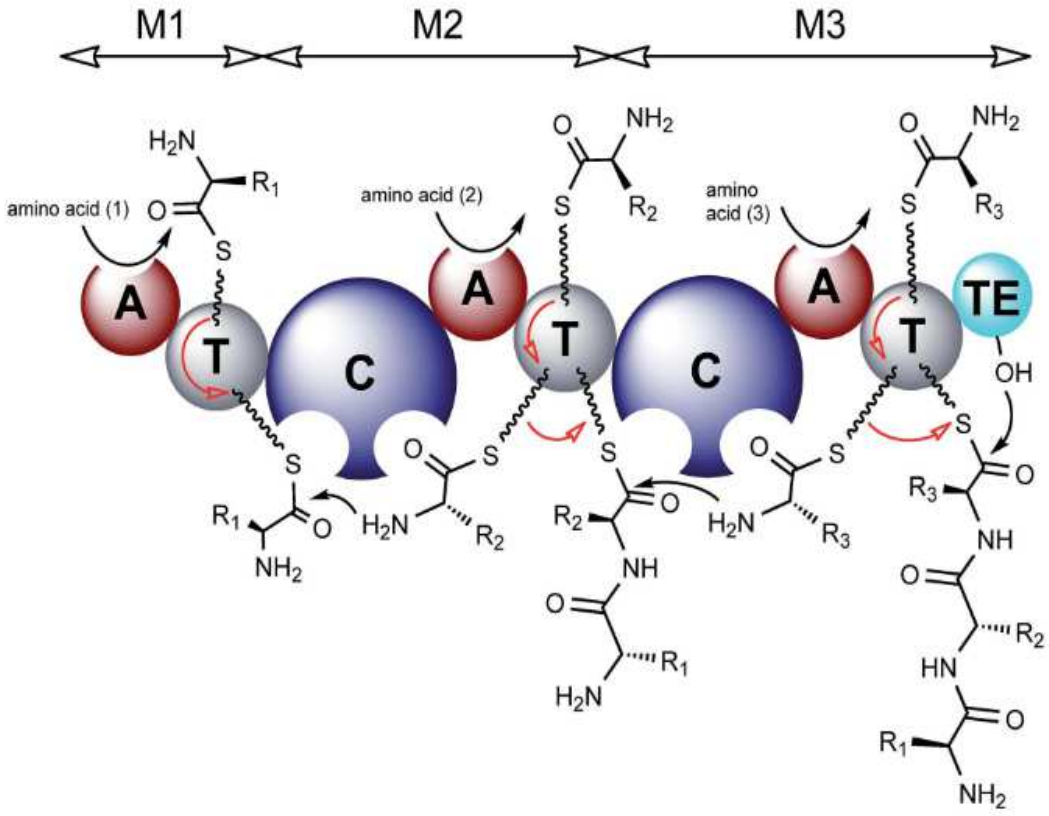


Figure 6. Modèle de biosynthèse des NRPS (Winn et al., 2016). Le domaine d'adénylation annoté A, le domaine de condensation annoté C, le domaine de thiolation annoté T et le domaine de thioestérase annoté TE.

En relation directe avec leur mode de synthèse, les peptides non ribosomiques présentent une biodiversité structurale importante. En effet, ils peuvent être structurellement linéaires, mais la plupart (plus des deux tiers) d'entre eux sont cycliques, multicycliques, partiellement cycliques ou encore branchés. Autre particularité et non des moindres, les monomères intégrés dans les NRPs sont issus de divers métabolismes. Ces monomères peuvent être des acides aminés protéogéniques ou non. Les NRPs peuvent intégrer une part non peptidique dans leur structure en incluant des chaînes lipidiques, des

glycannes ou encore des squelettes carbonés complexes comme des chromophores. Au final le nombre de monomères composant les NRPs s'élève à plus de cinq cents (Caboche *et al.*, 2007; Flissi *et al.*, 2016; Winn *et al.*, 2016).

3.2.2.2. Les domaines de modification des NRPS

Certains domaines optionnels sont capables de modifier les monomères au cours de la synthèse permettant d'obtenir des structures variées.

3.2.2.2.1. Le domaine d'épimérisation

Le domaine d'épimérisation est de loin le domaine secondaire le plus fréquemment rencontré. En effet, dans les NRPs, sont retrouvés des monomères sous forme L mais également sous forme D (environs 60% des NRPs contiennent au moins un D-monomère). La transformation d'une configuration L en configuration D est catalysée par le domaine d'épimérisation (domaine E) (Linne *et al.*, 2001). Le L-monomère lié au peptide passe du domaine C vers le domaine E. Chez quelques espèces bactériennes comme les *Pseudomonas*, il arrive que cette épimérisation soit catalysée par le domaine C lui-même. Le domaine C assure alors les deux fonctions de condensation et d'épimérisation, il est alors appelé domaine dual C/E (Balibar *et al.*, 2005).

3.2.2.2.2. Autres domaines secondaires

Il existe plusieurs domaines capables de modifier les monomères et contribuant à la grande diversité structurale des NRPs. Le domaine de N-méthylation (domaine NM) catalyse l'ajout d'un méthyle (CH_3) sur la fonction amine du monomère précédemment incorporé. Ce domaine est généralement situé entre deux modules. On peut également citer le domaine de formylation capable d'ajouter un groupement C(=O)H sur le groupement amine du premier monomère, le domaine d'hétérocyclisation et le domaine d'oxydation (Labby *et al.*, 2015; Süssmuth & Mainz, 2017).

Comme évoqué précédemment, la structure des protéines peut être directement déduite de la séquence génomique, mais ce n'est pas le cas pour les molécules en elles-mêmes que sont les NRPs. Cependant, il est possible de prédire partiellement et indirectement leur structure en analysant les séquences protéiques des synthétases correspondantes (Bode & Müller, 2005).

En définitive, les approches classiquement utilisées pour l'identification par séquençage en spectrométrie de masse des peptides ribosomiques ne sont pas directement applicables aux NRPs en raison de leur structure exotique et de la grande diversité de monomères intégrés dans leur structure. En fait, l'analyse réalisée sur des surnageants de culture bactérienne ou des purifications partielles de ces dernières ne s'avère pas plus complexe qu'en protéomique en terme de nombre de molécules détectées mais l'interprétation des données MS et MS/MS est plus compliquée du fait du grand nombre de fonctions chimiques et de squelettes carbonés rencontrés. De plus, les banques de données métabolomiques sont incomplètes et peu documentées.

3.2.3. Polycétides synthases (PKS)

Des métabolites secondaires autres que les NRPs sont abondamment produits par les microorganismes. Ce sont par exemple les polycétides synthétisés grâce à des polycétides synthases (*polyketides synthases*, PKS). Les PKS sont également de larges complexes enzymatiques modulaires et la synthèse PKS est très proche de celle des NRPS. Ainsi, à la place des domaines d'adénylation (A), de condensation (C) et de thiolation (T) des NRPS, les PKS possèdent des domaines AT, ACP et KS responsables respectivement de l'incorporation, de la liaison au complexe enzymatique et de la condensation des monomères entre eux. Parfois d'autres domaines responsables notamment de la diversité structurale des polycétides sont également présents. Les PKS se distinguent toutefois des NRPS car elles n'incorporent que quatre monomères différents (Shen, 2003).

3.2.4. Les hybrides PK-NRPS

Il existe également une voie de synthèse commune donnant des molécules hybrides : les PK-NRPs. En réalité, soit le NRP et le PK sont synthétisés de façon indépendante puis sont ensuite liés post-synthèse à l'aide d'enzymes. Soit c'est la NRPS et la PKS qui forment une chaine d'assemblage unique qui synthétise les molécules hybrides PK-NRPs.

3.3. L'implication des NRPs dans le biocontrôle

Les NRPs sont des molécules possédant un panel d'activités important, parmi eux les lipopeptides et les sidérophores produits par les microorganismes sont particulièrement intéressants pour des applications en biocontrôle. Le biocontrôle est défini comme un ensemble de méthodes de protection des cultures végétales basé sur l'utilisation « d'agents de biocontrôle ». Ces agents sont classés en 4 groupes : **i**) les macrooganismes (invertébrés, insectes ou acariens), **ii**) les médiateurs chimiques tels que les phéromones d'insectes et les kairomones qui contrôlent des populations d'insectes ravageurs par le piégeage et la méthode de confusion sexuelle, **iii**) les microorganismes comme les champignons, bactéries et virus mais également **iv**) les substances naturelles. Cet ensemble de stratégies a pour but de mettre en œuvre une alternative à la lutte chimique qui n'est pas sans conséquences néfastes pour l'environnement et la santé humaine. En effet, la mise en place de molécules naturelles biodégradables, l'équilibre des populations d'agresseurs plutôt que leur éradication, tout en en épargnant les polinisateurs semble être une solution plus raisonnée et respectueuse de l'environnement et du consommateur.

Cette notion de biocontrôle n'est pas empirique et on retrouve déjà des exemples capables de justifier cette approche. Le ministère de l'agriculture et de l'alimentation propose une liste de produits phytopharmaceutiques de biocontrôle possédant déjà une autorisation de mise sur le marché (https://info.agriculture.gouv.fr/gedei/site/bo-agri/instruction-2018-726). Le ministère définit également la méthodologie d'élaboration de la liste, et notamment les critères généraux de définition des produits concernés. Ainsi, on y retrouve des phéromones à chaîne linéaire de lépidoptères efficaces sur eudémis et cochylis sur vigne (Decoin, 2018), des microorganismes du genre *Bacillus* (Deravel *et al.*, 2014; Jacques *et al.*, 2014) et des substances actives telles que le 6-benzyladénine et l'acide gibbérellique. Certains composés naturels contribuent à améliorer l'état physiologique des

plantes, dans ce cas de figure, ils ne font pas partie des agents de biocontrôle mais des agents biostimulants, stimulateur de la vitalité des plantes. Les sidérophores et les lipopeptides présentent des activités particulièrement intéressantes pour des applications en biocontrôle. En effet, en privant les pathogènes en fer, les sidérophores jouent un rôle de compétition dans laquelle les champignons phytopathogènes peuvent perdre. L'utilisation des lipopeptides produits par des bactéries des genres *Pseudomonas* et *Bacillus* s'est révélée efficace à la fois comme agents de biocontrôle et inducteur de la défense des plantes (Ongena & Jacques, 2008). Ces lipopeptides produits par des microorganismes entrent dans la catégorie « microorganismes » avec les bactéries qui les produisent.

3.4. La métabolomique

La métabolomique a pour définition l'étude de l'ensemble des métabolites de petite masse moléculaire (en général, inférieure à 1500 Da) d'une cellule, d'un organite, d'un tissu ou d'un organe à un moment donné et sous des conditions données. La métabolomique s'attache donc à identifier et/ou quantifier de manière globale **i**) les métabolites extraits d'une culture cellulaire (d'un tissu ou fluide biologique), **ii**) leur localisation dans les compartiments cellulaires chez les eucaryotes) et **iii**) leurs éventuelles modifications post-traductionnelles ou post-synthèses. Elle permet d'identifier et/ou quantifier les variations de taux d'expression en fonction du temps, de l'environnement, de l'état de développement, de l'état physiologique et pathologique, de l'espèce d'origine (sauvage versus mutée). Elle étudie aussi les interactions que les métabolites ont avec d'autres métabolites mais aussi les protéines, l'ADN, l'ARN, ou d'autres substances endogènes ou exogènes (Welker, 2011).

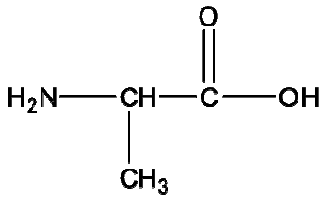
3.5. Banque de données métabolomiques

L'organisation des banques de données de métabolomique n'est pas aussi uniformisée que celle de génomique et de protéomique pour plusieurs raisons. La métabolomique est une discipline en plein essor et n'a vu le jour que très récemment. Par conséquent, le travail de recensement de données reste immense en raison de la jeunesse scientifique de cette discipline. D'autre part, les activités biologiques des métabolites sont très convoitées par les

grands groupes industriels, où chacun va réaliser sa propre recherche sans partage pour des raisons de propriété intellectuelle et d'exploitation.

Parmi les banques de données métabolomiques, la plus importante et la plus complète est sans nul doute la base de données PubChem du NCBI. Cette dernière recense actuellement presque 95 millions de composés. Plusieurs informations structurales sont recensées comme le nom IUPAC, la formule moléculaire et surtout les spécifications moléculaires simplifiées en une ligne, le « SMILES » (*Simplified Molecular Input Line Entry Specification)* (Figure 7).

A **B**



NC(C)C(O)=O

Figure 7. Formule semi-développée de l'alanine (A) et son format SMILES (B).

Le SMILES est une représentation en une dimension, sous forme de texte condensé, de la structure moléculaire (semi-développée). La Figure 7, illustre la formule semi-développée de l'alanine en A et son équivalent en représentation SMILES en B.

4. La classification des microorganismes

4.1. Phylogénie

La classification consiste à créer des groupes afin d'établir une certaine hiérarchie. Pour la classification d'individus, ces groupes sont appelés des taxons, les microorganismes n'échappent pas à cette appellation. L'ensemble de ces taxons constitue la taxonomie ; chez les microorganismes, les taxons sont organisés selon leur relation phylogénétique. L'espèce est l'unité fondamentale de la classification, une souche étant la sous-division d'une espèce. Actuellement, la banque de données du NCBI recense quatre grandes catégories de microorganismes : les archées (218 taxons, 178 genre, 668 espèces), les bactéries (1921

taxons, 3621 genre, 18415 espèces), les champignons filamenteux (1865 taxons, 6122 genre, 41923 espèces) et les virus (928 taxons, 793 genre, 4048 espèces)

(https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=statistics&uncultured=hide&unspecified=hide).

Ainsi, les espèces partagent de nombreux caractères. Cette classification peut être réalisée depuis des méthodes moléculaires. Pour les bactéries, l'ARN ribosomal 16S et 23S est utilisé pour distinguer les différents taxons c'est-à-dire espèces (Baron, 1996).

En fonction de certains champs disciplinaires d'autres classifications sont parfois mises en place. Par exemple, une classification liée au caractère pathogène des souches avec un intérêt plus important en bactériologie clinique (Baron, 1996).

4.2. L'identification des microorganismes

L'identification des microorganismes s'appuie sur des techniques classiques de microbiologie qui ont relativement bien changé depuis les travaux de Louis Pasteur dans les années 1800. La stratégie générale d'identification des microorganismes repose sur un principe universel qui consiste à déterminer un maximum de caractéristiques d'un micro-organisme pour en permettre sa distinction parmi une multitude d'autres microorganismes. Dès lors, la stratégie d'identification des bactéries consiste en plusieurs étapes comprenant **i**) des tests rapides et simples d'orientation tels que l'observation du phénotype macro-biologique, la coloration de Gram ou les tests de catalase et d'oxydase et **ii**) des tests basés sur les caractéristiques biochimiques tels que les galeries API, pour compléter l'identification. Ces techniques sont connues comme les « gold standards » de la microbiologie.

Au fil des années, les chercheurs ont développé des méthodes plus fiables et plus rapides pour identifier des microorganismes. Ces méthodes dites « méthodes alternatives » sont basées sur l'utilisation de composés brevetés (par exemple, des substrats chromogènes, les anticorps, les sondes ou les amorces d'ADN). On distingue **i**) les méthodes basées sur la croissance microbienne (milieu chromogène), **ii**) les méthodes basées sur des réactions immunologiques ou immunoenzymatiques et **iii**) les méthodes utilisant les progrès de la

biologie moléculaire (PCR, RT-PCR ou encore Multiplex-PCR), (Lombard & Leclercq, 2010). Ces tests permettent d'identifier des infectieux suspectés.

Plus récemment est apparue dans le domaine de la microbiologie, une approche d'identification différente. Cette technique est une approche par phénotypage moléculaire qui repose sur l'obtention de la signature majoritairement protéique par spectrométrie de masse de type MALDI-TOF d'un microorganisme (Buchan & Ledeboer, 2013; Gravet & Gessier, 2013; Marklein *et al.*, 2009). L'identification des microorganismes par spectrométrie de masse MALDI-TOF est basée sur l'empreinte spectrale (*Main Spectral Projection*, MSP) de ces derniers. Cette empreinte spectrale est une « sorte d'empreinte digitale spécifique » ou un phénotypage moléculaire partiel mais toujours identique. Certains composés (signaux) composant le MSP sont spécifiques du genre, de l'espèce et même dans certains cas de la sous-espèce. Les spectres obtenus sont reproductibles. L'identification est basée sur la comparaison de la position et de l'intensité des signaux de masse du spectre de masse inconnu avec ceux de tous les spectres enregistrés dans la banque de données de spectres. Les librairies de spectres sont fournies, validées et mises à jour par les entreprises commercialisant les systèmes MALDI-TOF. Actuellement, la banque de spectres de la compagnie Bruker Daltonics permet l'identification de 5298 bactéries (entérobactéries, bacilles Gram négatif non fermentant, Staphylocoques, Streptocoques, mycobactéries, bactéries anaérobies, …) et 691 champignons (Candida, champignons filamenteux, …). Cette banque de données, permettant l'identification de pathogènes d'intérêt clinique et de microorganismes de l'environnement, peut être enrichie par les utilisateurs.

Partie II : Les défis pour une nouvelle stratégie d'identification de peptides non ribosomiques d'origine microbienne

1. Approche guidée par le génome

Des dizaines de milliers de génomes microbiens séquencés ou en partie séquencés sont disponibles sur la base de données internationale « *International Nucleotide Sequence Database Collaboration* » (INSDC) et ce nombre devrait encore augmenter de façon exponentielle au cours des prochaines décennies (Cochrane *et al.*, 2016).

Quelques méthodes *in silico* d'automatisation de l'analyse des métabolismes secondaires dans les génomes bactériens et fongiques ont été publiées. Le site « *secondary metabolites* » (http://www.secondarymetabolites.org) recense de façon succincte tous les outils et bases de données disponibles pour l'exploitation des données génomiques.

ClustScan (Starcevic *et al.*, 2008), le premier de ces outils, est conçu pour une annotation rapide et semi-automatique de séquences d'ADN codant pour des PKS, des NRPS et des hybrides PKS/NRPS (http://csdb.bioserv.pbf.hr/csdb/ClustScanWeb.html).

Plus récemment, un pipeline plus complet appelé antiSMASH (Weber *et al.*, 2015) capable d'identifier des locus biosynthétiques couvrant toute la gamme des classes de métabolites secondaires connus (PKs, NRPs mais aussi terpènes, amino-glycosides, aminocoumarines, indolo carbazoles, lantibiotiques, bactériocines, nucléosides, bêta-lactamines, butyrolactones, siderophores, mélanines et autres) (https://antismash.secondarymetabolites.org).

Bien que le génome présente une vue d'ensemble des métabolites potentiellement produits, l'expression de la plupart des métabolites est conditionnée par l'environnement (par exemple la présence ou l'absence de certains nutriments), la température, la présence d'un autre microorganisme et bien d'autres paramètres (Hernandez-Eugenio *et al.*, 2015).

La connaissance de la séquence du génome reste insuffisante pour connaitre les structures élémentaires complètes des métabolites eux-mêmes. D'autre part, le génome

n'est pas toujours une information disponible et le séquençage reste encore coûteux et génère une quantité d'informations importantes et uniquement manipulables par des experts.

2. La résonance magnétique nucléaire (RMN)

La résonance magnétique nucléaire (RMN) fait appel à une propriété physique des particules élémentaires appelée le spin, que l'on retrouve dans tout noyau atomique (proton et neutron) ainsi que les électrons. Certains noyaux comme le proton ^{1}H, le carbone ^{13}C, le phosphore ^{31}P, l'azote ^{15}N ou le fluor ^{19}F, possèdent un spin nucléaire de 1/2 et sont par conséquent sensibles à un champ magnétique externe. D'autres noyaux comme le deutérium ^{2}H, l'azote ^{14}N, l'oxygène ^{17}O ou encore le soufre ^{33}S, possèdent un spin supérieur à 1/2 (en réalité ils possèdent un moment quadripolaire nucléaire et tous ces noyaux sont capables de résonner mais dans des conditions différentes).

Les molécules sont dans un premier temps placées dans un champ magnétique statique intense. Puis un champ magnétique différent qui provoque une perturbation des atomes est appliqué sur de courte durée de quelques microsecondes. Les noyaux génèrent à leur tour un micro-champ magnétique (ils résonnent) qui sera enregistré par une bobine réceptrice, c'est le signal RMN. Ce signal est alors analysé puis transformé en spectre grâce au calcul de transformée de Fourier.

Le couplage direct de la RMN avec des méthodes d'analyse par séparation reste encore une tâche difficile à réaliser. Par conséquent, cette technique reste encore peu utilisée pour le criblage ou la caractérisation directe de mélanges complexes tels que les surnageants de culture bactérienne. L'analyse RMN de fractions chromatographiques issues de la séparation d'un mélange complexe après leur collecte a néanmoins été précédemment rapportée (Brkljača & Urban, 2015). L'inconvénient de cette méthode réside dans le stockage temporaire des fractions afin d'éviter d'éventuels phénomènes de dégradation car les échantillons sont généralement séchés (évaporés sous vide ou lyophilisés) avant l'analyse par RMN (Balayssac *et al.*, 2009).

Malgré tout, la RMN est une technique précise et résolutive qui permet d'obtenir des informations détaillées de la structure moléculaire d'un composé. Elle reste la technique indispensable et de choix lors d'élucidation de structure moléculaire.

3. La spectrométrie de masse

La spectrométrie de masse est née avec les débuts de la physique atomique. Sir Joseph John Thomson en 1913 (Sir J.J. Thomson, 1906) et Francis William Aston en 1919 ont mis tour à tour au point une combinaison de champs électriques et magnétiques (produits dans des gaz lors de décharges électriques) pour caractériser les particules atomiques chargées. À cette époque, le seul moyen de détecter des particules était d'utiliser une plaque photographique comme nous le rappelle l'histoire des découvertes d'Henri Becquerel sur la radioactivité à partir de minerais radioactifs déposés par hasard sur une plaque photographique dans un tiroir. Les particules ionisantes noircissaient la plaque photographique comme la lumière. La trace laissée par les ions permettait leur détection et de déduire leur rapport masse sur charge (m/z).

L'histoire de la spectrométrie de masse commerciale couvre plus de 50 ans. Brunnée revient sur les principes d'analyseurs de masse communs dans une revue de 1987 (Brunnée, 1967). Gelpi traite plus de 130 spectromètres de masse différents construits depuis 1965 dans une série de deux revues (Gelpí, 2009). Dans les années 2000, de nouveaux types de spectromètres de masse ont été équipés d'analyseur en masse innovant comme l'analyseur Orbitrap (Makarov, 2000). De même, de nouveaux instruments hybrides, équipés de la technologie relative à la mobilité ionique couplée à un analyseur à temps de vol (*time of flight*, TOF) ont été introduits sur le marché récemment.

3.1. Architecture générale d'un spectromètre de masse

La spectrométrie de masse consiste en une mesure très précise du rapport masse sur charge (m/z) de molécules isolées ou en mélange. Quel que soit le spectromètre de masse, les molécules à analyser sont ionisées et transférées en phase gazeuse. Après accélération des ions dans une direction de l'espace, le rapport m/z des ions ainsi formés est alors

déterminé par soit **i**) la mesure du temps mis pour parcourir une distance donnée dans un espace vide de champ (temps de vol) ou **ii**) par la trajectoire des ions dans un champ électrique et/ou magnétique dynamique ou statique. Un spectromètre de masse est classiquement formé de quatre composantes (Figure 8):

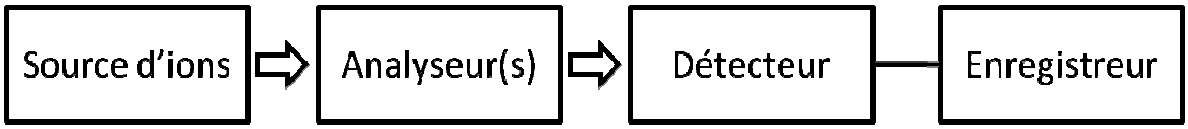


Figure 8. Schématisation simple de l'organisation générale d'un spectromètre de masse. Un spectromètre de masse est composé de trois parties, une source d'ions, un ou plusieurs analyseurs et un détecteur. Cependant un enregistreur tel qu'une unité informatique est toujours associé à un spectromètre de masse pour convertir les signaux électriques en signaux informatiques, ce qui en fait une partie indissociable du spectromètre de masse.

-La source d'ions permet dans un premier temps de produire des ions intacts ou fragmentés de molécules en phase gazeuse à partir d'une phase initiale (gazeuse, liquide ou solide). Les ions produits sont caractérisés par leur masse moléculaire « m » et par le nombre de charges élémentaires « z » qu'ils portent;

-L'analyseur de masse a quant à lui pour rôle de trier les différents ions transférés de la source. Ce tri repose **i**) soit sur le temps mis par les ions pour parcourir une distance donnée ou **ii**) la propriété d'un champ électrique ou magnétique à dévier les trajectoires des ions. Ce tri est lié au rapport m/z des ions;

-Les détecteurs sont des compteurs et amplificateurs des signaux d'ions. Ils permettent de transformer un courant ionique faible en un signal électrique mesurable. Le détecteur fournit des informations relatives au nombre d'ions atteignant le détecteur et par conséquent l'intensité du signal en masse reçu. Les ions sont détruits et se déchargent par collisions avec le détecteur.

-L'enregistrement par un système informatique permet de déterminer le rapport *m/z* en fonction des ions arrivant au détecteur et des paramètres de l'analyseur. L'unité informatique permet le stockage des informations d'acquisitions et ainsi permet la reconstruction du spectre de masse.

Pour être parfaitement rigoureux, il faut associer à ces trois composantes une quatrième composante qui correspond au vide élevé ou modéré via le système de pompage du spectromètre de masse, tout aussi nécessaire à l'analyse des ions et à la conservation de leur intégrité. Si la chronologie d'apparition des différents analyseurs est bien connue, du simple analyseur magnétique aux débuts de l'analyseur Orbitrap, celles concernant les technologies du vide, des détecteurs et des systèmes d'acquisition sont aussi spectaculaires. Les pompes turbomoléculaires utilisées sur tous les spectromètres de masse d'aujourd'hui permettent d'obtenir des vides très poussés (10^{-10} Torr soit 10^{-10} mbar) ne nécessitant pas des dégazages à très haute température (200°C) pendant plusieurs jours.

3.2. La source d'ions

La source d'ions est le dispositif qui permet d'ioniser les substances à analyser dans le spectromètre. Une des caractéristiques les plus importantes de ces différents types de sources réside dans l'énergie interne transférée pendant le processus d'ionisation. Certaines techniques d'ionisation sont très énergétiques et par conséquent entraînent une fragmentation en source de la molécule ionisée. Les techniques d'ionisation plus douces produisent seulement des espèces moléculaires sans fragmentation. Un autre facteur important de la source d'ions réside dans la nature physico-chimique de l'analyte (composé analysé). Par exemple, l'impact électronique (IE) et l'ionisation chimique (IC) sont seulement appropriés à l'ionisation en phase gazeuse. Par conséquent, l'analyte doit être suffisamment volatile et thermiquement stable pour pouvoir utiliser l'IE et l'IC. Les biomolécules sont d'avantages des molécules thermiquement labiles ou n'ont pas la tension de vapeur suffisante. D'autres sources sont plus couramment utilisées en biologie. L'introduction dans la source est réalisée en phase solide, liquide ou gazeuse (directement introduit dans la source d'ions). Dans les sources d'ions en phase liquide, l'analyte est en solution, celui-ci est introduit dans le spectromètre de masse par nébulisation, sous forme de gouttelettes par l'intermédiaire de plusieurs étapes de pompage pour maintenir un vide acceptable. Cette source est appelée l'électrospray (*electrospray ionization*, ESI). Dans les sources d'ions à l'état solide, l'analyte est dans un dépôt non volatile. L'analyte est alors irradié par des particules énergétiques ou des photons qui désorbent des ions présents à la surface du

dépôt. Ces ions peuvent être extraits par un champ électrique et être ensuite focalisés vers un analyseur.

Si la source d'ionisation est le dispositif qui permet la production d'ions en phase gazeuse, la méthode d'ionisation se rapporte plus particulièrement au mécanisme de l'ionisation. Quelle que soit la source, les ions sont produits principalement selon l'une des méthodes d'ionisation suivantes : soit en ionisant une molécule neutre par l'éjection d'un électron, par capture d'un électron, par la protonation, par la déprotonation ou par la formation d'adduits (cationisation), ou soit par le transfert en phase gazeuse d'une espèce chargée présente dans la phase.

3.2.1. Ionisation très énergétique, impact électronique (IE)

L'impact électronique (IE) à 70 eV inventé par A.J. Dempster (Dempster, 1918) et perfectionné par W. Bleakney (Bleakney, 1929) et Nier est historiquement considéré comme la plus ancienne technique d'ionisation pour l'analyse de petites molécules. Elle est constituée d'un filament chauffé qui émet des électrons. Ceux-ci sont accélérés vers une anode et entrent en interaction avec les molécules gazeuses de l'échantillon analysé introduit dans la source. La substance doit donc être à l'état gazeux pour être ionisée. Pour ce faire, certains composés peu volatils présents à l'état liquide ou solide sont habituellement chauffés pour permettre leur vaporisation.

En raison de l'énergie d'ionisation constante choisie, cette technique se traduit par des spectres de masse cohérents et riches en fragments. Ces spectres de masse peuvent être facilement utilisés pour la création et la recherche en bibliothèques spectrales. L'impact électronique est couramment utilisé pour les configurations GC-MS. Un inconvénient majeur des spectres de masse obtenus avec une source à IE est qu'elle provoque une fragmentation intense et par conséquent l'ion moléculaire (parent) n'est pas toujours observé.

3.2.2. Ionisation douce

3.2.2.1. La désorption / ionisation laser assistée par matrice (MALDI)

L'introduction en 1988 de la source de désorption et d'ionisation laser assistée par matrice (*Matrix Assisted Laser Desorption/Ionisation*, MALDI) est due principalement à Karas et Hillenkamp (Karas & Hillenkamp, 1988). Cette source permet d'ioniser une large gamme de molécules de hauts poids moléculaires avec une préparation rapide et une mise en œuvre simple. L'ionisation MALDI se déroule en deux étapes.

Dans la première étape, la substance à analyser est mélangée à une solution d'une petite molécule organique, appelée matrice, possédant une forte absorption à la longueur d'onde du laser. Cette solution est déposée sur une surface métallique, appelée Cible MALDI, et le solvant est alors évaporé avant l'analyse. Le résultat est un dépôt de matrice/analyte co-cristallisé où les cristaux de matrice sont en excès vis-à-vis des molécules d'analyte.

La deuxième étape se produit sous vide à l'intérieur de la source du spectromètre de masse (Figure 8). Cette étape comporte l'ablation de ce dépôt par des impulsions laser pendant une courte durée. En réalité, le mécanisme exact du processus MALDI n'est pas complètement connu. Cependant, l'énergie transmise par le laser est absorbée par la matrice, l'irradiation par le laser induit l'accumulation d'une grande quantité d'énergie sur le dépôt par excitation des molécules de la matrice. Ces excitations sont rapidement converties en mouvements thermiques et cet apport d'énergie très localisé cause la sublimation des cristaux de la matrice et provoque l'ablation d'une portion de la surface du cristal suivie de l'expansion de la matrice en phase gazeuse, entraînant l'analyte intact dans cette expansion. Les processus d'ionisation mis en jeu sont la photoionisation en phase gazeuse, le transfert de proton, les réactions ions-molécules et la désorption des ions préformés. Le mécanisme préférentiellement admis pour la formation d'ions est le transfert de proton soit avant désorption en phase solide, soit lors de l'expansion en phase gazeuse (entre la matrice photo-excitée et l'analyte). Les ions formés sont ensuite accélérés à l'aide d'un champ électrostatique vers l'analyseur (Figure 9).

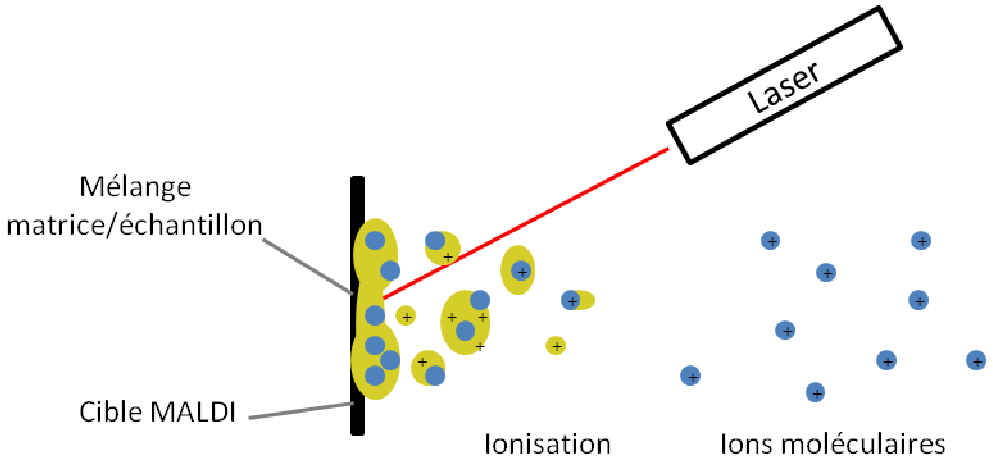


Figure 9. Principe de la désorption / ionisation laser assistée par matrice (MALDI).

Cette source générant des ions par paquets est bien appropriée aux analyseurs à temps de vol et à résonance cyclonique ionique à transformée de Fourier (*fourier-transformation ion cyclotron resonance*, FT-ICR).

Le MALDI est une source largement utilisée en biologie dans de multiples études de caractérisation moléculaire de petites molécules (Zidour *et al.*, 2017) ou encore de peptides (Caly *et al.*, 2017). Elle permet même aujourd'hui d'identifier des microorganismes par phénotypage moléculaire (Nacef *et al.*, 2017).

3.2.2.2. Electrospray

Cette technique de production des ions a été développée par Fenn, dont il publia la première description en 1984 (Yamashita & Fenn, 1984). L'ESI a été considéré à ses débuts comme une source d'ionisation dédiée à l'analyse de protéines. Très rapidement, son utilisation a été élargie aux polymères, et aussi aux petites molécules. Depuis les années 1990 et encore de nos jours, l'ESI est devenu la technique d'ionisation la plus répandue. En effet, cette technique permet d'atteindre des sensibilités très élevées, mais surtout elle est facile à coupler à la chromatographie liquide ou à l'électrophorèse capillaire (Wakayama *et al.*, 2015). L'ESI est produit par application à pression atmosphérique d'un fort champ électrique sur un liquide traversant un tube capillaire avec un faible débit. Le champ électrique est obtenu par application d'une différence de potentiel pouvant aller jusqu'à 6

kV entre ce capillaire et la contre-électrode. Ce champ provoque une accumulation de charges à la surface du liquide, située à l'extrémité du capillaire, qui va se rompre pour former des gouttelettes fortement chargées. Ces gouttelettes contiennent les composants (analytes, électrolytes, etc…) de la solution introduite dans le capillaire. À de faibles tensions, la gouttelette semble sphérique, puis elle s'allonge sous la pression des charges accumulées à l'extrémité dans le champ électrique. Lorsque la pression est suffisante, la force exercée est supérieure à la tension superficielle, la surface se rompt, la gouttelette se transforme alors en un « cône de Taylor » et le spray de gouttelettes fortement chargées apparaît (Figure 10).

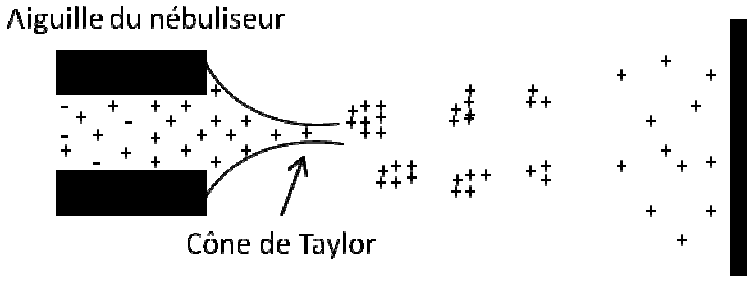


Figure 10. Schématisation de l'éléctronébulisation par électrospray. L'augmentation de la densité de charge jusqu'à la valeur du rayon critique de Rayleigh entraine des explosions coulombiennes hétérogènes (pour une ionisation positive des molécules dans cet exemple).

L'évaporation du solvant contenu dans ces gouttelettes va provoquer leur rétrécissement et donc encore augmenter la densité de charge jusqu'au moment où les forces coulombiennes répulsives vont approcher le niveau des forces de cohésion des gouttelettes et provoquer leur explosion. Gomez et Tang (Gomez & Tang, 1994) ont obtenu des photographies des gouttelettes formées dans une source ESI. De leurs observations, ils concluent que la décomposition des gouttelettes se produit avant la limite donnée par l'équation de Rayleigh parce que les gouttelettes sont déformées par le champ électrique intense auquel elles sont soumises. Cela réduit la répulsion nécessaire à leur rupture et induit la formation de nouveaux « cônes de Taylor ». Un jet d'une vingtaine de nouvelles petites gouttelettes ayant une densité de charge encore plus élevée se forme. Une gouttelette de première génération aura un diamètre de l'ordre de 1,5 micromètre et portera environ 50 000 charges élémentaires. Les gouttelettes résultant d'un jet auront un

diamètre de l'ordre de 0,1 micromètre et porteront environ 300 à 400 charges élémentaires, ce qui multiplie la charge par unité de volume d'un facteur sept.

Ce processus d'évaporation suivie de fission de la gouttelette mère se répète. La densité de charge augmente encore par évaporation ou par formation d'un nouveau jet de gouttelettes encore plus petites jusqu'au moment où leur densité de charge devient suffisante pour provoquer la désorption d'ions. Ces ions ainsi produits sont porteurs d'un grand nombre de charges s'il existe plusieurs sites ionisables suffisamment séparés sur la molécule. L'ESI est également applicable à des molécules ne possédant aucun site ionisable par protonation grâce à la formation d'adduits sodique, potassique, ammonium, etc… ou en ions négatifs d'adduits chlorure, acétate…La mise en place de l'ionisation par électrospray (Fenn, 2003; Gaskell, 1997) fut une percée majeure pour l'analyse de grandes biomolécules intactes. L'ESI est maintenant le procédé d'ionisation de choix pour la chromatographie liquide MS (LC-MS) dans de nombreux laboratoires dans le monde (Cech & Enke, 2002). En outre, la miniaturisation de l'ESI, la source nanoélectrospray (*nanoelectrospray ionization*, nanoESI) (Schmidt *et al.*, 2003) a fait son apparition au cours de ces dernières années (Almeida *et al.*, 2008; Koulman *et al.*, 2009; Lazar *et al.*, 2006; Lydic *et al.*, 2009; Wickremsinhe *et al.*, 2006) ; son efficacité supérieure d'ionisation et ses avantages en termes analytiques sont clairement démontrés. En effet, l'infusion de nano litres permet d'augmenter le temps d'analyse avec un minimum d'échantillons et permet d'augmenter grandement la sensibilité. Ces longs temps d'infusion sont souvent nécessaires pour les identifications structurales, car ils permettent de réaliser des fragmentations multiples (MS^n).

3.3. Les analyseurs

Après avoir produit les ions, il faut les séparer selon leur rapport m/z, qu'il faudra simultanément déterminer. Tout comme il existe une grande variété de sources d'ions, il existe de nombreux analyseurs différents avec des pouvoirs de résolution plus ou moins importants.

Le pouvoir de résolution est la capacité d'un instrument à distinguer entre eux deux signaux voisins. Si l'on considère le pouvoir séparatif, la résolution est définie comme le

rapport m/Δm où m est la masse d'ions considérés et Δm la différence minimale entre un pic et son voisin le plus proche dont il peut être distingué. Selon cette définition, un instrument qui est en mesure de distinguer des ions de masses 100 et 100,1 possède un pouvoir de résolution de 100/(100,1-100)=1000. Le pouvoir de résolution est calculé pour le rapport m/Δm avec Δm corespondant à la largeur du pic à mi-hauteur. Par conséquent, pour des ions de masse 100, la largeur à mi-hauteur doit être de 0,1 pour atteindre une résolution de 1000. Il est également possible d'utiliser le rapport Δm/m exprimé alors en ppm (parties par million). Un appareil avec un pouvoir de résolution important permet d'obtenir des mesures précises mais pas nécessairement de mesures exactes. L'exactitude dépend en effet de la qualité de la calibration (étalonnage).

3.3.1. Basse résolution

3.3.1.1. Analyseur quadripolaire (Q)

Un analyseur quadripolaire est constitué de quatre électrodes parallèles de section hyperbolique ou cylindrique. Les électrodes opposées distantes sont reliées entre elles et soumises au même potentiel. Les électrodes adjacentes sont portées à des potentiels de valeurs, avec une tension continue et une tension alternative. Un champ électrostatique quadripolaire est ainsi créé dans la région entre les quatre électrodes. En pénétrant dans le quadripôle, les ions conservent leur vitesse longitudinale et progressent alors dans l'analyseur. Parallèlement, la différence de potentiel entre l'entrée et la sortie du quadripôle guide les ions dans la direction voulue. Le mouvement d'un ion selon la direction x et y est défini par la solution de l'équation de Mathieu (Mathieu, 1868). Ainsi, pour une valeur de tension alternative et continue, seuls les ions de valeur m/z définie auront une trajectoire stable jusqu'au détecteur.

Cet analyseur seul ne permet pas d'obtenir des données très résolues ; en revanche il est très utilisé en quantification où plusieurs analyseurs de même type se suivent séquentiellement (triple quadripôle). Cet analyseur peut également être couplé à des analyseurs plus résolutifs comme le temps de vol par exemple. Ces spectromètres de masse sont alors qualifiés d'appareils hybrides car possédant une combinaison d'analyseurs de masse aux principes de fonctionnement différents.

3.3.2. Haute résolution

3.3.2.1. Analyseur à temps de vol

Le principe de l'analyseur à temps de vol, a été décrit par William E. Stephens en 1946 (Stephens, 1946; Wolff & Stephens, 1953). A.E. Cameron et D.F. Eggers ont ensuite publié en 1948 les plans instrumentaux et des spectres du premier modèle d'analyseur TOF linéaire (Cameron & Eggers, 1948). Il fallut attendre 1955 pour voir l'élaboration par W.C. Wiley et I.H. McLaren d'un spectromètre à analyseur TOF dans un but commercial (Wiley & McLaren, 1955). Celui-ci fut par la suite vendu par la société Bendix (Figure 11).

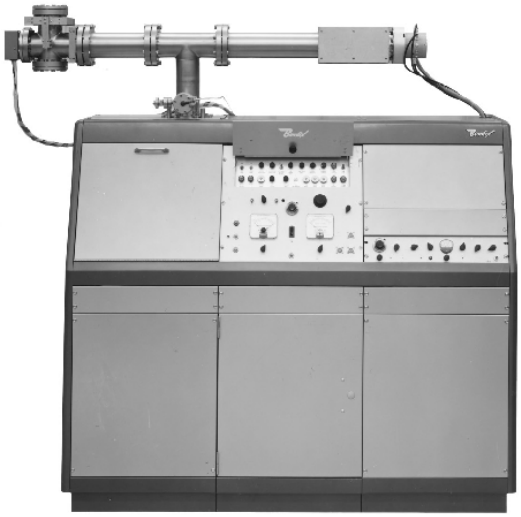


Figure 11. Spectromètre de masse à analyseur à temps de vol, Model Bendix 12-101A, fabriqué en Californie en 1961.

Un regain d'intérêt pour ces instruments est apparu depuis la fin des années 1980. Tout d'abord, les progrès de l'électronique ont permis de maîtriser des flux élevés de données. Ensuite, la désorption laser pulsée est parfaitement adaptée à l'analyseur à temps de vol car les ions sont produits par paquets pendant des intervalles de temps très courts. Le développement technologique du couplage de la source MALDI à l'analyseur TOF a ouvert la voie à de nouvelles applications non seulement pour les biomolécules mais également pour les polymères synthétiques.

<u>Analyseur de masse à temps de vol linéaire</u>

Le principe d'un instrument à temps de vol linéaire est illustré dans la figure 12.

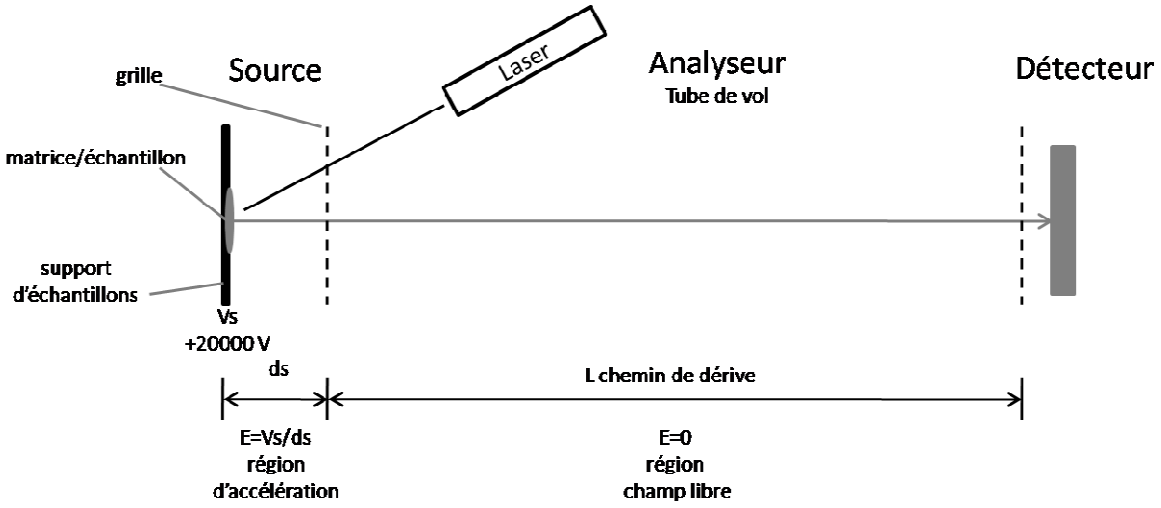


Figure 12. Illustration du principe d'un analyseur à temps de vol linéaire (illustration avec une source MALDI et un détecteur).

Cet analyseur sépare les ions en fonction de leur vitesse lorsqu'ils se déplacent dans une zone libre de champ appelée tube de vol. Les ions sont expulsés de la source par paquets, soit parce qu'ils sont produits par intermittence comme dans une source à désorption laser, soit encore parce qu'ils sont expulsés par une application très brève de potentiels voulus sur les électrodes de focalisation de la source. Ces ions sont alors accélérés vers le tube de vol par une différence de potentiel appliquée entre une électrode et la grille d'extraction. Comme tous les ions acquièrent la même énergie cinétique, des ions caractérisés par une distribution de leur masse présentent une distribution de leur vitesse. Une fois sortis de la zone d'accélération, ils entrent dans le tube de vol où ils sont séparés en fonction de leur vitesse acquise, avant d'atteindre le détecteur placé à l'autre extrémité du tube de vol.

Les rapports m/z sont déterminés en mesurant le temps nécessaire aux divers ions pour parcourir d'un mouvement rectiligne uniforme la distance de vol « L » correspondant à la distance dans la région libre de champ entre la zone d'accélération et le détecteur.

Mode Réflectron

Une manière d'améliorer la résolution en masse des analyseurs à temps de vol est d'employer un réflectron également appelé miroir électrostatique dont l'utilisation fut proposée pour la première fois par Mamyrin (Mamyrin, 2001). Le réflectron est constitué

d'une série de grilles et d'électrodes annulaires portées à des potentiels croissants et qui définissent un champ électrique homogène. Comme ce champ électrique est de même direction, mais de sens opposé au champ électrique présent dans la zone d'accélération, il s'oppose à la progression des ions et agit en tant que miroir d'ions réfléchissant les ions. En effet, les ions qui entrent dans le réflectron ralentissent jusqu'à s'arrêter et sont réaccélérés en sens inverse vers le tube de vol. L'emploi d'un réflectron permet d'allonger la distance de vol et donc d'augmenter la résolution sans agrandir les dimensions du spectromètre de masse. Les inconvénients sont une diminution de la sensibilité et une limitation dans la gamme de masse.

Le réflectron permet également la refocalisation temporelle au niveau du détecteur des ions quittant la source avec le même m/z et présentant une dispersion en énergie cinétique. En effet, les ions possédant plus d'énergie cinétique pénétreront plus profondément au sein du réflectron et passeront donc plus de temps dans le réflectron. De cette manière, ils atteindront le détecteur en même temps que les ions de même m/z possédant moins d'énergie cinétique du fait que ces derniers mettront moins de temps pour être réfléchis. Ceci permet de refocaliser tous les ions de même m/z sur un plan, quelle que soit leur énergie cinétique initiale (Figure 13).

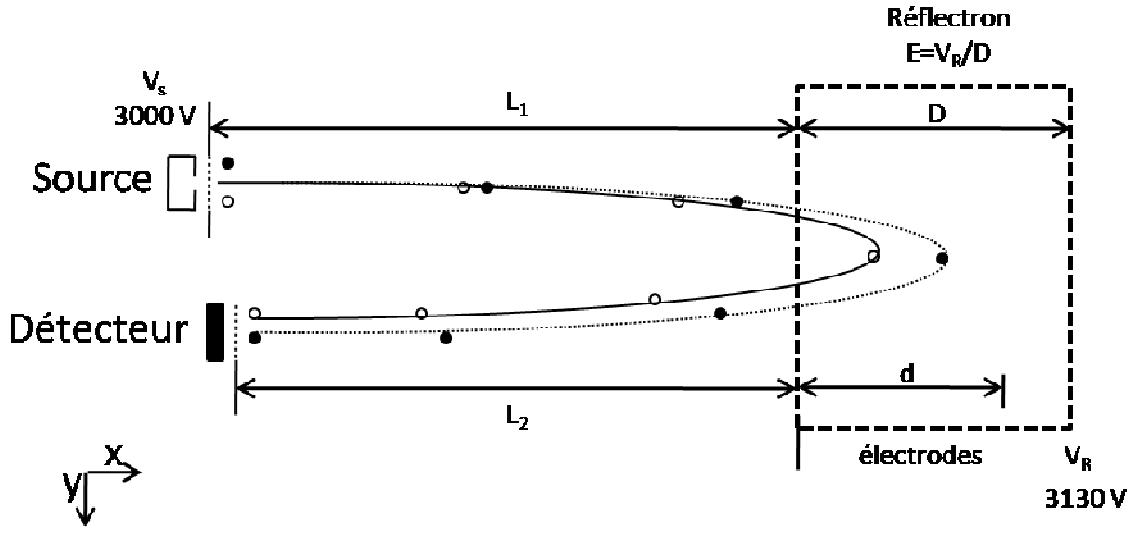


Figure 13. Illustration du principe du réflectron d'un analyseur à temps de vol.

3.3.2.2. Transformée de Fourrier – résonance cyclotronique ionique

Le développement de cet analyseur est attribué à E.O. Lawrence (Lawrence, 1931) qui l'a mis au point pour étudier les propriétés fondamentales des atomes sous la forme d'un premier accélérateur cyclotronique. Dans les années suivantes, F.M. Penning (Penning, 1936) conçoit un piège ionique capable de garder les ions dans un champ électrique et les électrons dans un champ magnétique. Ce principe a été utilisé dès 1949 en spectrométrie de masse par le groupe de recherche de J.A. Hipple (Hipple *et al.*, 1949).

La seconde révolution des analyseurs FT-ICR est venue de Melvin B. Comisarow et Alan G. Marshall qui en concomitance avec le développement de l'informatique ont amené la digitalisation du signal électrique grâce au développement des convertisseurs analogiques/numériques. Vers le milieu des années 1970, ces auteurs ont utilisé une opération mathématique, aujourd'hui célèbre, la transformation de Fourier dans le traitement du signal cyclotronique (Comisarow & Marshall, 1974). Plusieurs évolutions instrumentales majeures ont conduit les analyseurs FT-ICR à devenir les appareils les plus résolutifs de la spectrométrie de masse. Récemment, un spectromètre de masse FT-ICR de 21 teslas a été mis au point par les américains avec un pouvoir de résolution incroyable capable d'atteindre une résolution de 1 400 000 pour une masse de m/z 600 (Smith *et al.*, 2018). D'autres modifications ont ensuite permis la démocratisation de l'analyseur FT-ICR sur plusieurs champs disciplinaires.

Le principe de la cellule ICR est basé sur le piégeage des ions par un champ magnétique. Un ion de vitesse initiale dans un champ magnétique uniforme subit une force (force de Lorentz) qui induit un mouvement circulaire de l'ion, perpendiculaire à la direction du champ magnétique. Ce mouvement de rotation est appelé mouvement cyclotronique. La fréquence cyclotronique ne dépend que du champ magnétique et du rapport m/z de l'ion et est indépendante de la vitesse initiale. La mesure de la fréquence cyclotronique permet de déterminer le rapport m/z des ions dans la cellule ICR. Cette fréquence est inversement proportionnelle au rapport m/z. Par conséquent, un ion de haut rapport m/z aura donc une fréquence cyclotronique faible (et inversement). A l'aide de la combinaison d'un champ magnétique et d'un champ électrique, les ions sont piégés dans la cellule ICR (Penning, 1936). Deux plaques de piégeage sont disposées perpendiculairement au champ magnétique

pour créer un puits de potentiel électrique permettant le piégeage des ions suivant l'axe de la cellule. Les ions vont alors osciller entre ces deux plaques : ce phénomène est appelé mouvement de piégeage. Un troisième mouvement appelé mouvement magnétron est issu de la composante radiale du champ électrique car en effet, le champ électrique et magnétique ne peuvent pas être parallèles (les lignes de champs électrostatiques ne sont pas droites). Ce dernier mouvement tend à réduire la fréquence cyclotronique en entrainant les ions sur d'autres trajectoires. La trajectoire issue de la combinaison des trois mouvements est illustrée dans la Figure 14. En réalité, les mouvements de piégeage et de magnétron sont négligeables par rapport au mouvement cyclotronique.

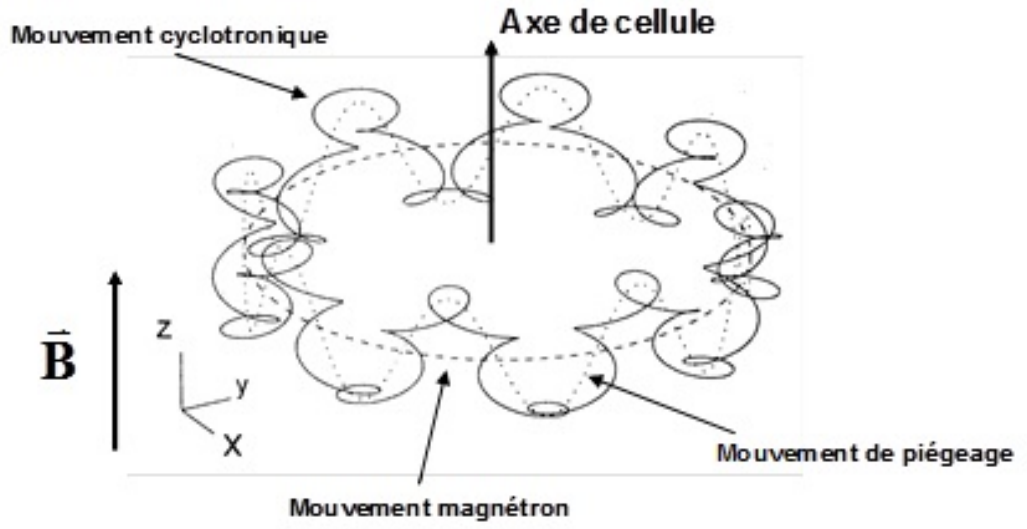


Figure 14. Trajectoire des ions dans une cellule à résonance cyclonique ionique (ICR), illustrant les trois mouvements, piégeage, magnétron et cyclotronique.

Aujourd'hui, l'analyseur FT-ICR équipe les spectromètres utilisés dans de nombreux champs disciplinaires tels que l'environnement (Ohno *et al.*, 2014), l'analyse de polymères (Fouquet *et al.*, 2018) ou encore dans des champs plus précis comme la pétroléomique (Roach *et al.*, 2011).

3.4. Le couplage chromatographique

Le couplage chromatographique est une méthode physico-chimique qui permet de séparer des composés présents dans un mélange complexe en vue de faciliter son analyse afin de déterminer sa composition et même de doser les éléments composant ce mélange.

Le principe repose sur des équilibres de concentration des composés présents entre deux phases non miscibles (à l'exception de la chromatographie d'exclusion stérique). La première est appelée phase stationnaire car elle est généralement emprisonnée dans une colonne ou fixée à un support et permet l'absorption des composés à séparer. La seconde est appelée la phase mobile car elle se déplace (effluent) au contact de la première dont les propriétés physico-chimiques varient au cours du gradient d'élution. Le changement des propriétés physico-chimiques de la phase mobile modifie l'équilibre de concentration des composés appelé coefficient de partage ou d'absorption et ainsi permet leur désorption de la phase stationnaire. La difficulté est de bien choisir ses deux phases afin de permettre aux constituants du mélange d'être retenus inégalement entre eux. Les constituants du mélange injecté progressent moins vite (ou à la même vitesse) que la phase mobile car leur vitesse de déplacement sont différentes en raison du coefficient de partage ou d'absorption. Ce phénomène est appelé la rétention.

Il existe de nombreuses configurations : la chromatographie en phase liquide à haute performance (*High Pressure Liquid Chromatography*, HPLC), la chromatographie liquide ultra performance (*Ultra High Pressure Chromatography,* UPLC), la chromatographie en phase gazeuse (*Gaz Chromatography,* GC), chromatographie sur couche mince (CCM) faisant appel à différents principes chromatographiques. Elles sont généralement décrites selon les types d'interactions : la chromatographie d'adsorption (liaisons de faible énergie comme les liaisons hydrogènes et de Van-der-Waals), la chromatographie de partage (affinité plus ou moins forte pour l'une des deux phases, les composés seront repartis différentiellement dans les phases), la chromatographie d'échange d'ions (les molécules seront séparées selon leur charge), la chromatographie d'affinité ou encore la chromatographie d'exclusion stérique (séparation selon leur volume hydrodynamique qui se définit comme leur taille en solution). Les champs d'application sont très vastes et il existe également des champs très spécialisés comme la chromatographie énantiosélective adaptée à la séparation d'énantiomères.

On retrouve la plus part des phases à l'état liquide, solide ou gazeux mais il existe également une chromatographie utilisant les fluides super critiques appelée simplement la chromatographie en phase super critique.

Malheureusement, cette diversité ne contribue pas à une harmonisation simple des méthodes de séparations et par conséquent il n'existe pas de technique universelle, applicable à toutes les classes de molécules, permettant de les séparer. Cependant, la force de cette méthode de séparation réside essentiellement dans sa facilité à être couplée à divers appareils de mesure physique comme l'absorbance ultra-violet (U.V., 100-400 nm) l'absorbance infrarouge (400-750 nm) et surtout la spectrométrie de masse. De plus, de par cette dimension séparatrice, cette technique a également révolutionné et facilité la quantification de molécules.

En spectrométrie de masse la présence d'une quantité importante de sels dans les échantillons est souvent indésirable car leur présence peut provoquer une ionisation moins efficace des molécules d'intérêt. Cette interférence qualifiée « d'effet de suppression d'ionisation » conduit à l'absence de signaux de masse. De plus, leur présence conduit à la formation d'adduits (processus chimique qui correspond à l'interaction d'un ion avec une molécule neutre) avec les molécules d'intérêt, décalant leur rapport m/z par rapport à la valeur attendue.

3.5. L'acquisition de données et la mesure de masse

La mesure de masse exacte dépend essentiellement du pouvoir de résolution pouvant être atteint par l'analyseur de l'appareil, mais également des protocoles de calibration machine indispensables. La calibration, également appelée l'étalonnage est une procédure permettant de régler, caler, référencer cet analyseur par rapport à des données de références (contenues dans un calibrant ou étalon) afin d'obtenir un comportement fidèle lors de l'enregistrement de la mesure des m/z et obtenir des masses exactes. Une masse est exacte si elle donne un résultat très proche de ce qu'on attend. L'exactitude d'une mesure repose sur deux concepts : la fidélité et la justesse. Si l'on répète plusieurs fois une même mesure, il est possible de tracer les résultats sur un repère orthonormé avec pour origine la masse exacte (ME), (Figure 15).

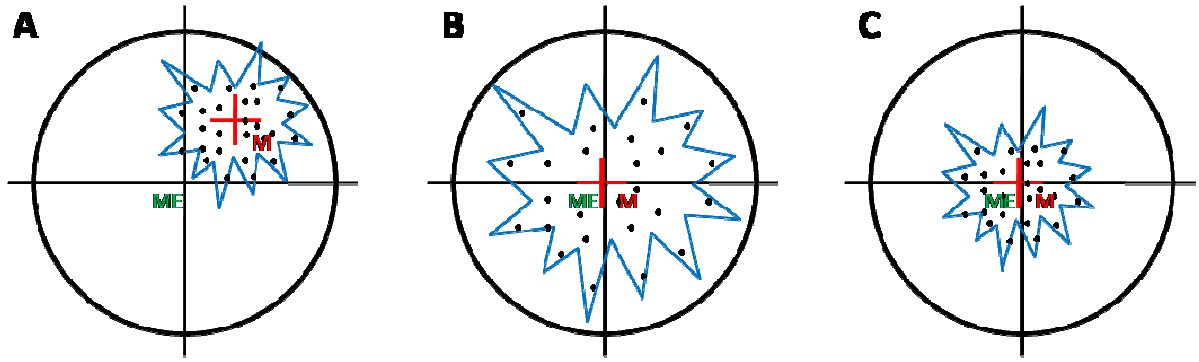


Figure 15. Notion d'exactitude de mesure. Représentation schématique de la mesure avec en A, la notion de fidélité, en B, la notion de justesse et en C, la notion d'exactitude. ME en vert pour la masse exacte et M en rouge pour la moyenne des mesures.

La Figure 15 A montre un ensemble de points fortement centrés autour de leur moyenne (l'aire couverte est réduite, l'écart-type est faible). Cependant cette moyenne est assez éloignée de la masse exacte. Les mesures sont fidèles mais pas justes. Dans ce cas l'erreur commise peut être corrigée par re-calibration. Le deuxième graphe (Figure 15 B) montre un ensemble de points dont la moyenne est confondue avec la masse exacte. Cependant ils sont fortement dispersés (l'aire couverte est étendue, l'écart-type est élevé). Dans ce cas, les mesures sont justes mais pas fidèles. La dispersion statistique du résultat rend la correction de la mesure compliquée voire impossible. Le troisième graphe (Figure 15 C) montre un ensemble de points peu dispersés dont la moyenne est confondue avec la valeur vraie. Les mesures sont à la fois justes et fidèles.

A l'inverse, l'inexactitude est exprimée par l'incertitude de mesure. Cette incertitude est la probabilité d'erreur sur une mesure isolée. Sur une série de mesures effectuées sur une valeur connue, il est facile de mesurer les erreurs de chaque mesure et de calculer leur moyenne, qu'on appelle l'écart-type. La calibration d'un appareil permet d'ajuster au mieux les paramètres de justesse et de fidélité. Il est important de ne pas confondre toutes ces notions avec la notion de précision de mesure. La précision d'une valeur numérique mesurée est donnée par le nombre de décimales. La précision au dixième se définit pour le premier chiffre après la virgule, au centième pour le second….

3.6. L'exploitation des données MS

Il est important de ne pas négliger la masse de l'électron au cours des mesures effectuées en masse exacte. Rappelons que la masse de l'électron est de 0,00054857990924 u (Beier *et al.*, 2002). Cette masse peut entrainer des erreurs allant jusqu'à 3 ppm pour un rapport m/z de 100 (Ferrer & Thurman, 2005). Plusieurs études comparatives montrent la pertinence des mesures en spectrométrie de masse réalisées sur divers types d'appareils au sein de plusieurs laboratoires. En 2003, une étude comparative a été menée dans 45 laboratoires (Bristow & Webb, 2003). Les auteurs ont montré que les mesures en masse faites sur des spectromètres de masse de type FT-ICR permettent d'obtenir au quotidien des exactitudes en masse de moins de 1 ppm. D'autres analyseurs ont montré des performances similaires comme l'analyseur à temps de vol (Bereman *et al.*, 2008; Stroh *et al.*, 2007) et plus récemment la technologie Orbitrap (Makarov & Scigelova, 2010).

3.6.1. Détermination de la composition élémentaire

3.6.1.1. Les générateurs de formules moléculaires

La détermination de la formule moléculaire (formule brute ou composition élémentaire) à partir de logiciels qualifiés de générateurs de formules moléculaires nécessite une mesure m/z obtenue avec une grande exactitude et précision de mesure.

Les générateurs de formules moléculaires utilisent les masses théoriques des atomes pour calculer la formule moléculaire dont la valeur est la plus proche de la masse moléculaire déduite de la mesure de m/z. Ce calcul est similaire à la méthode informatique « d'attaque par force brute », qui signifie que toutes les combinaisons d'atomes vont être testées une à une. On parle d'algorithme « *Find-all* » car toutes les possibilités de formules moléculaires sont calculées (Patiny & Borel, 2013). Il est alors possible de réduire le nombre de formules moléculaires candidates en réduisant le nombre d'atomes à intégrer dans la formule recherchée et en réduisant l'erreur de mesure pour diminuer l'espace de recherche.

Plusieurs logiciels existent en accès libre sur internet comme « *molecular weight calculator* » ou encore « *ChemCalc* » (Patiny & Borel, 2013) et la plupart des constructeurs de spectromètres de masse possèdent leur propre logiciel de génération de formule brute

comme Bruker avec « *Smart Formula Manually* » ou encore Waters avec une fonction directe dans « *MassLynx* ».

En conclusion, ces logiciels de calcul permettent de générer un nombre de formules moléculaires en fonction de l'exactitude de la mesure. En effet, plus la masse moléculaire est importante et plus le nombre de formules moléculaires candidates est statistiquement important.

3.6.1.2. les profils isotopiques

On appelle isotopes des atomes possédant au sein de leur noyau le même nombre de protons et d'électrons (donc la même charge, z), mais un nombre de neutrons différents (donc une masse moléculaire différente). Les isotopes d'un même élément ont le même comportement chimique (mais pas le même comportement physique, voir le chapitre : La résonance magnétique nucléaire). Sur notre planète, l'abondance relative (la proportion) de chaque isotope pour un atome donné est connue, par exemple pour le carbone son isotope le plus stable est l'isotope ^{12}C et son abondance relative est de 98,93%, en revanche pour son isotope ^{13}C son abondance est de 1,07%. Cette abondance est différente selon les atomes et définit l'abondance polyisotopique comme pour l'hydrogène (H), le carbone (C), l'azote (N), l'oxygène (O), le soufre (S), le chlore (Cl) ou encore le brome (Br) mais aussi l'abondance monoisotopique comme le fluor (F), le sodium (Na), le phosphore (P) et l'iode (I) (Loss, 2001). Afin d'éviter la confusion, nous préciserons la masse moléculaire de façon à les caractériser sans aucune équivoque, nous parlerons donc de l'hydrogène ^{1}H, du carbone ^{12}C, ou encore de l'oxygène ^{16}O.

Lors d'une mesure en spectrométrie de masse, tous les isotopes sont mesurés et une molécule est par conséquent représentée par sa distribution isotopique. L'abondance isotopique de chaque atome étant connue (Meija *et al.*, 2016), il est possible de calculer la distribution isotopique théorique d'une molécule. Ainsi, certains algorithmes de modélisation des signaux de masse prennent en compte la distribution isotopique des molécules détectées afin de pouvoir proposer la formule moléculaire candidate ayant le profil isotopique théorique le plus proche de celui mesuré. Les algorithmes mettent en

œuvre soit des méthodes basées sur les polynômes, soit des méthodes basées sur les calculs de transformée de Fourier (IsoDalton, MWTWIN, Mercure, Isotope Calculator, IsoPro, EMASS/qmass, libmercury ++, ISOMABS et Decon2Ls) (Li *et al.*, 2008; Olson & Yergey, 2009; Rockwood & Haimi, 2006; Snider, 2007). L'abondance isotopique fournit donc une information supplémentaire (en plus de la mesure exacte) utile au cours du processus d'élucidation de la formule moléculaire mais aussi de la structure chimique (Alon & Amirav, 2009; Ramaley & Herrera, 2008; Rockwood *et al.*, 2003).

Un autre logiciel de l'Institut National des Standards et Technologies (*National Institut of Standards and Technology,* NIST) traitent de l'utilisation des rapports d'abondances isotopiques pour confirmer ou rejeter des résultats de la recherche d'identité moléculaire en bibliothèque spectrale (Alon & Amirav, 2006). Bon nombre de publications traitent du processus de recherche de profils isotopiques pour la détermination de la formule élémentaire en chimie environnementale (Grange *et al.*, 2002; Grange & Sovocool, 2008), suite à des expériences de profilage métabolique (Rogers *et al.*, 2009; Werner *et al.*, 2008), et en géochimie (Koch *et al.*, 2008; Reemtsma, 2009). Le logiciel SIRIUS (https://bio.informatik.uni-jena.de/software/sirius/) disponible gratuitement analyse également les profils isotopiques (Böcker *et al.*, 2009). Il possède une interface graphique conviviale et peut être utilisé sur les plateformes LINUX, MAC et Windows.

Il est important de noter qu'il faut être très vigilant sur la génération des massifs isotopiques. En effet, des phénomènes de saturation du signal (dus à des concentrations trop importantes) peuvent engendrer des erreurs de mesure sur le massif isotopique (notamment sur l'intensité relative entre chaque isotope). Les analyses réalisées sans dimension séparative entrainent parfois des chevauchements de massif isotopiques et par conséquent des erreurs dans la génération des formules moléculaires. Malgré cela cette méthode permet de réduire de manière importante le nombre de formules moléculaires possibles éliminant même jusqu'à 90% des formules moléculaires incorrectes pour des masses de plus de 1000 Da (Böcker *et al.*, 2009).

Les sept règles d'or proposées par Kind et Fiehn en 2007 (Kind & Fiehn, 2007) peuvent également être prises en compte dans les calculs de formules moléculaires candidates. Ces 7 règles sont un ensemble de règles heuristiques pour le calcul de la

composition élémentaire. Elles comprennent les règles de Senior et Lewis, les règles de rapport élémentaire hydrogène/carbone et hydrogène/oxygène (respectivement, rapport H/C et O/C) et un filtre de prise en compte de la valence atomique, le tout codé par un langage « *Virtual Basic for Applications*» (VBA) disponible sur Excel. Le logiciel est accessible librement et utilisable pour calculer la formule moléculaire avec les atomes C, H, N, O, S, P. Il couvre plus de deux milliards de compositions élémentaires et il a été déduit que seules 623 millions de compositions élémentaires sont hautement probables (Kind & Fiehn, 2010).

D'autres approches utilisant la combinaison d'un algorithme pour le calcul des motifs isotopiques théoriques et le marquage des molécules d'intérêt avec les isotopes stables ^{13}C et ^{15}N a également été développé (Hegeman *et al.*, 2007).

3.6.1.3. Les défauts de masse (Kendrick)

Bien avant l'ère de la puissance du calcul informatique, Edward Kendrick, un chimiste anglais propose, en 1963, une méthode mathématique élégante basée sur la détermination d'un défaut de masse (maintenant appelé défaut de masse de Kendrick) pour faciliter la discrimination entre des composés homologues ayant des nombres différents de mêmes unités de base (Kendrick, 1963).

Brièvement, le défaut de masse d'un élément ou composé chimique est calculé comme la différence entre la masse exacte d'un isotope et sa masse nominale qui est la simple addition du nombre de protons et de neutrons dans une formule donnée. En vertu de la convention éditée par l'union internationale de chimie pure et appliquée (IUPAC), le carbone 12 (^{12}C) a été défini comme l'élément présentant un défaut de masse nul. Sa masse atomique est donc de 12 Da très précisément tandis que l'hydrogène (^{1}H) a une masse atomique de 1,00783 pour une masse nominale de 1. Le défaut de masse (différence entre la masse nominale et la masse théorique) de l'hydrogène est par conséquent de -0,00783 (= 1-1,00783).

La masse de Kendrick (*Kendrick mass,* KM) utilise un groupe d'atomes comme unité de base (blocs de construction) à la manière de la définition IUPAC pour le carbone 12. Dès lors, la masse nominale de Kendrick (*nominal Kendrick mass*, NKM) se réfère à la même

définition que précédemment : l'arrondi à l'entier le plus proche. Typiquement, pour le bloc de construction $^{12}C_1{}^1H_2$, la NKM est de 14.

Par conséquent, la masse de Kendrick (KM) d'un composé se calcule selon l'équation ci-dessous :

$$KM = masse\ protonée\ IUPAC \times \frac{bloc\ de\ construction\ NKM\ (si\ {}^{12}C_1 \quad {}^1H_2 = 14)}{bloc\ de\ masse\ exact\ IUPAC\ (si\ {}^{12}C_1 \quad {}^1H_2 = 14.01565)}$$

La KM peut être extrapolée à d'autres unités de base et à leurs isotopes (par exemple $^{12}C_1{}^1H_1{}^2H_1$; $^{12}C_1{}^2H_2$; $^{13}C_1{}^1H_2$, $^{13}C_1{}^1H_1{}^2H_1$; ... $^{12}C_2{}^1H_1{}^{16}O_1$). Ainsi, par définition, le défaut de masse de Kendrick (*Kendrick mass defect*, KMD) est défini comme la soustraction de la KM à la NKM et est compris entre : - 0,5 <KMD <+ 0,5.

Pour plus de clarté, au-delà de ce point, les informations relatives à l'approche de Kendrick se réfèreront à la masse monoisotopique.

3.6.1.3.1. Représentation graphique en deux dimensions

Il est possible d'illustrer graphiquement les défauts de masse de Kendrick en représentant la valeur de KMD (sur l'axe des ordonnées) en fonction de NKM (sur l'axe des abscisses). Chaque point des diagrammes 2D représente une formule moléculaire monoisotopique unique et les molécules qui diffèrent d'un bloc de construction sont corrélées horizontalement (Figure 16).

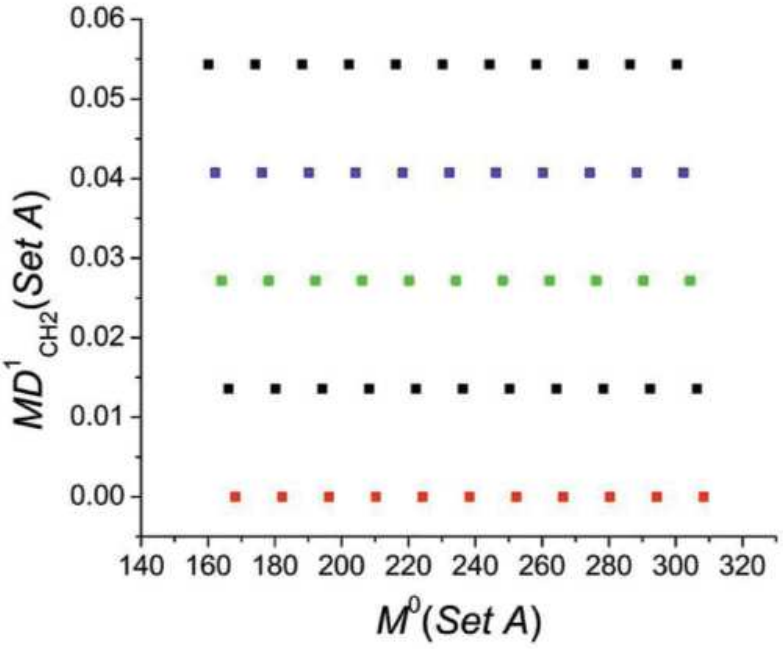


Figure 16. Tracé du défaut de masse de Kendrick calculé selon une unité de base CH2 (MD1CH2 =KMD) en fonction de la masse nominale (M0=NKM) d'une huile brute *(Roach et al.*, 2011).

Cette illustration extraite d'une étude de Patrick J. Roach (Roach *et al.*, 2011) montre un incrément croissant et périodique de CH_2 en fonction de la masse nominale. Les différentes couleurs illustrent des équivalentes doubles liaisons (DBE, z) différents (z = 1, 2, 3, 4, et 5 ici respectivement en rouge, noir, vert, bleu, et noir).

Toute variation de masse et donc de structure, même minime, entrainera une variation du défaut de masse. Cette méthode graphique permet donc de distinguer la moindre modification atomique, même pour un simple atome d'hydrogène mais également des états de charge différents. Par conséquent, il est possible de distinguer et donc de trier facilement les signaux de masse mesurés lors d'une analyse en masse.

Dans le cas de mélanges complexes, l'utilisation des défauts de masses de Kendrick s'avère utile lorsque la fragmentation (MS/MS) des composés n'est pas possible. Elle permet alors d'obtenir des informations relatives à la structure sur des composés analogues constitués d'une répétition successive d'un bloc de construction tels que les polymères chimiques, les huiles ou encore les hydrocarbures.

Cette représentation graphique et géométrique KMD= f(NKM) permet de créer une seconde dimension mais l'introduction d'un axe engendre une symétrie par rapport à celui-ci. Ce phénomène appelé « symétrie axiale » en géométrie entraine une « réflexion en miroir» et par conséquent une image de chaque point vis-à-vis de chaque axe.

3.6.1.3.2. Le repliement spectral (*Dealiasing*)

Par définition, dans l'ensemble du plan (d'un axe orthonormé) il existe une réflexion d'un même point suivant chaque axe. En bref, il existe une image de chaque point par réflexion sur l'axe des abscisses (x), des ordonnées(y) et même par rapport à l'origine. La réflexion est donc une fonction bijective. Nous pouvons donc définir une application de transformation réciproque.

Dans notre cas, il n'existe pas de réflexion en miroir par rapport à l'axe des ordonnées car il est impossible de mesurer une masse moléculaire inférieure à zéro (NKM fait partie des entiers naturels : N). En revanche, la réflexion par rapport à l'axe des abscisses est possible, car une molécule peut avoir un défaut de masse négatif par rapport à l'unité de base choisie (-KMD existe). Enfin, la réflexion par rapport à l'origine est une image intéressante dans le cas de l'utilisation de delta de NKM (ΔNKM). Pour cela, l'origine de l'axe orthonormé ne peut être utilisée (car NKM n'existe pas en valeur négative, –NKM n'existe pas) mais il est possible de choisir une origine différente, par exemple une autre masse mesurée (Figure 17).

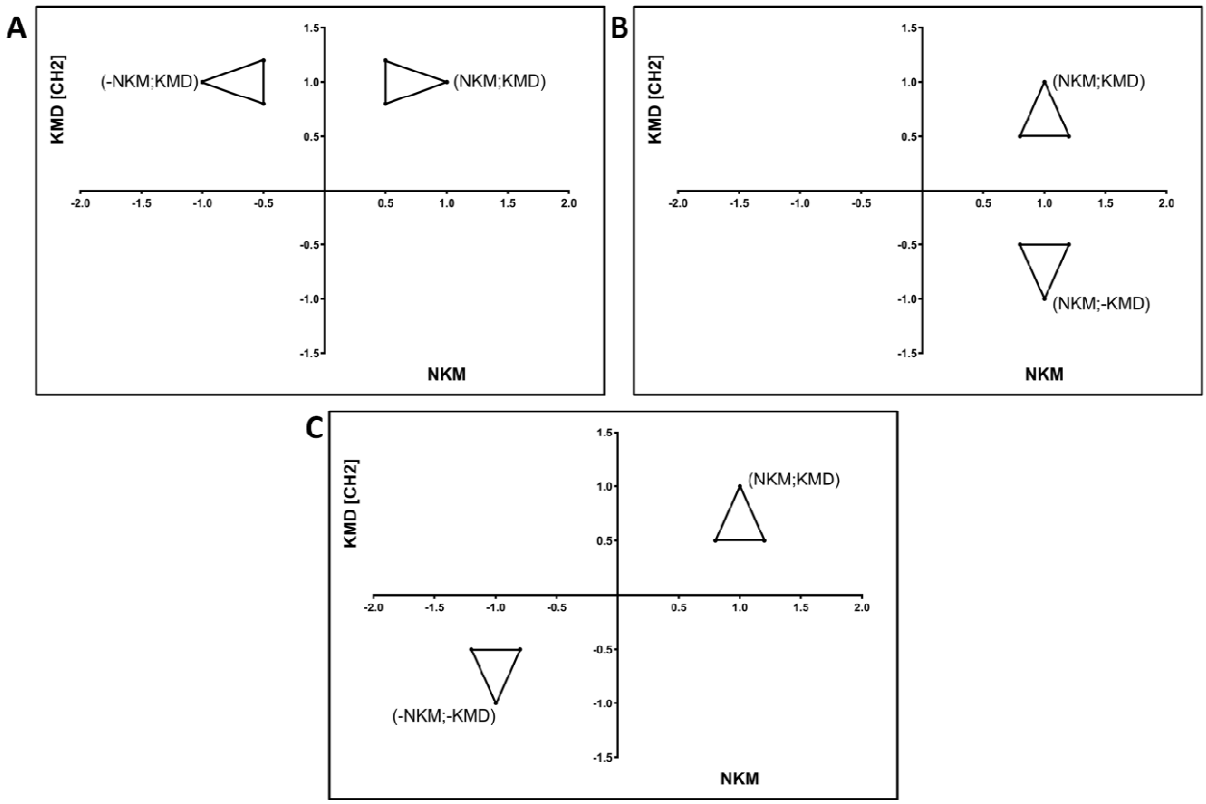


Figure 17. Principe de symétrie axiale. Réflexion en miroir par rapport à l'axe des ordonnées (KMD [CH2]) en A, réflexion en miroir par rapport à l'axe des abscisses (NKM) en B, réflexion en miroir par rapport à l'origine en C.

Afin d'éviter le repliement spectral (*aliasing*) sur un tracé 2D, il est possible de soustraire une partie du tracé à la masse de Kendrick pour contrer cet effet et ainsi obtenir une valeur corrigée du défaut de masse de Kendrick appelé RKMD pour *Regular Kendrick Mass Defect*. Cette méthode de filtrage de masse a été appliquée avec succès, en particulier pour l'analyse de mélanges complexes en pétroléomique, en chimie des polymères ou lors d'analyses de traitements de l'eau (Zhang *et al.*, 2012).

3.6.1.3.3. Amélioration de la résolution spectrale

Les défauts de masse sont employés uniquement pour le retraitement des données obtenues sur des spectromètres de masse de type FT-ICR car le concept repose sur une mesure de masse extrêmement résolutive et précise avec une tolérance de plus ou moins 1 ppm. Récemment, le concept d'amélioration spectrale de la représentation graphique en

deux dimensions a été proposé. Ce concept permet de mieux séparer les différentes séries d'ions (Fouquet & Sato, 2017b). L'amélioration spectrale est une transformation mathématique qui repose sur l'utilisation d'une unité de base fractionnaire différente. Ainsi, lors du calcul de la masse de Kendrick, le numérateur et le dénominateur sont inversés et le dénominateur n'est pas simplement un arrondi de la masse exacte de l'unité de masse choisie mais un entier plus ou moins proche, x dans l'équation ci-dessous. L'équation se présente alors ainsi :

$$KM = masse\ protonée\ IUPAC \times \frac{bloc\ de\ masse\ exact\ IUPAC\ (si\ {}^{12}C_1 \quad {}^{1}H_2 = 14.01565)}{x}$$

Cet entier est choisi arbitrairement jusqu'à obtention d'une meilleure résolution spectrale sur la représentation graphique (KMD=f(NKM)). Par conséquent, elle conduit à améliorer la résolution graphique et à un alignement corrigé de données de masse obtenues avec un analyseur moins résolutif et surtout moins juste (précis et exact) comme dans cette étude de Thierry Fouquet et al., où les auteurs ont utilisé un spectromètre de masse avec un analyseur à temps de vol (Fouquet & Sato, 2017a)(Figure 18).

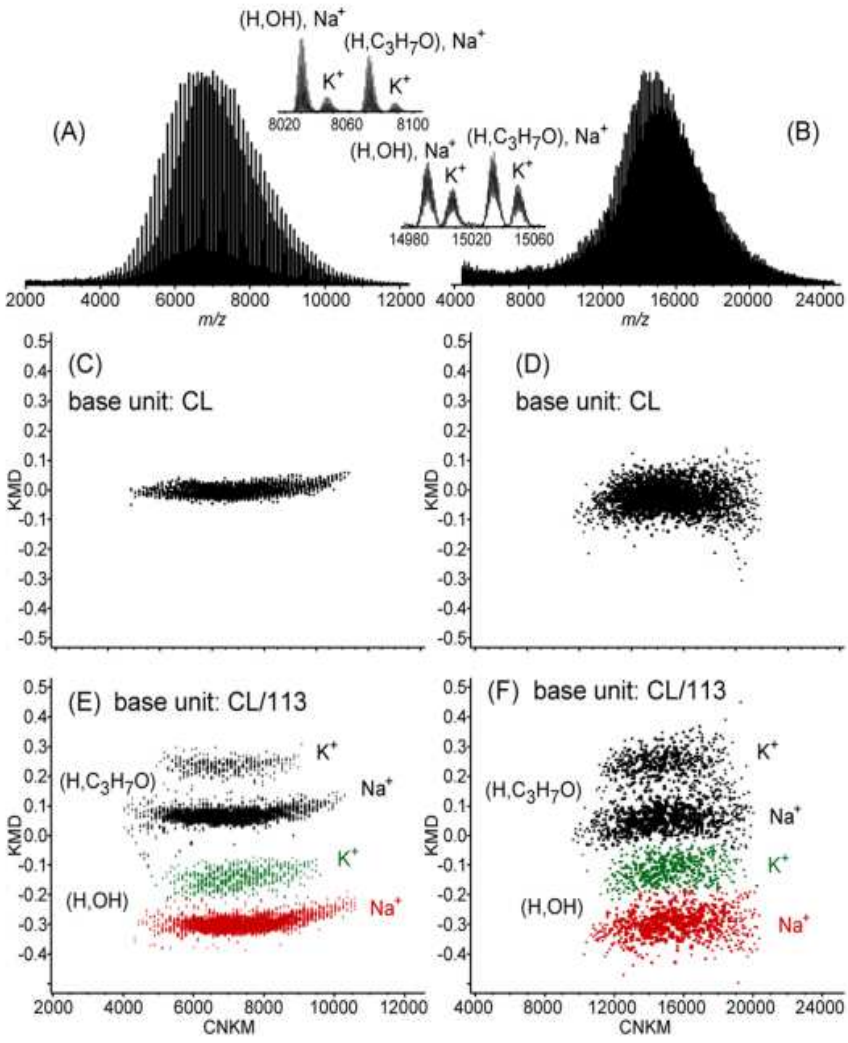


Figure 18. Illustration de l'amélioration spectrale sur deux analyses à des tailles différentes d'un polymère de polyéthylène glycol (polymère de masse moléculaire centré sur 7000 g/mol en A et 15000 g/mol en B). Représentation graphique avec le calcul des défauts de masse pour un dénominateur de 14 pour les deux tailles 7000 g/mol et 15000 g/mol, respectivement en C et D. Représentations graphiques avec dénominateur de 113 pour les deux tailles 7000 g/mol en E et 15000 g/mol en F.

La Figure 18 montre une analyse complète d'un polymère de polyéthylène glycol dont les mesures en masse ont été obtenues sur un appareil possédant un analyseur TOF. Les panneaux A et B illustrent les profils très complexes obtenus lors des mesures en masse. Les défauts de masse de cette analyse sont calculés avec une unité de base d'un CH_2 mais un dénominateur différent améliorant la résolution spectrale. Les panneaux C, D, E et F correspondent aux représentations graphiques de KMD=f(NKM) selon un dénominateur de

14 (panneaux C et D) et 113 (panneaux E et F). L'amélioration spectrale obtenue par le second calcul permet ainsi de distinguer les adduits potassique et sodique (panneaux E et F).

3.6.2. Diagramme de Van Krevelen

Le diagramme de Van Krevelen est une représentation graphique illustrant le ratio du nombre d'hydrogènes par rapport au nombre de carbones (H/C) en fonction du ratio du nombre d'oxygènes par rapport au nombre de carbones (O/C), Figure 19.

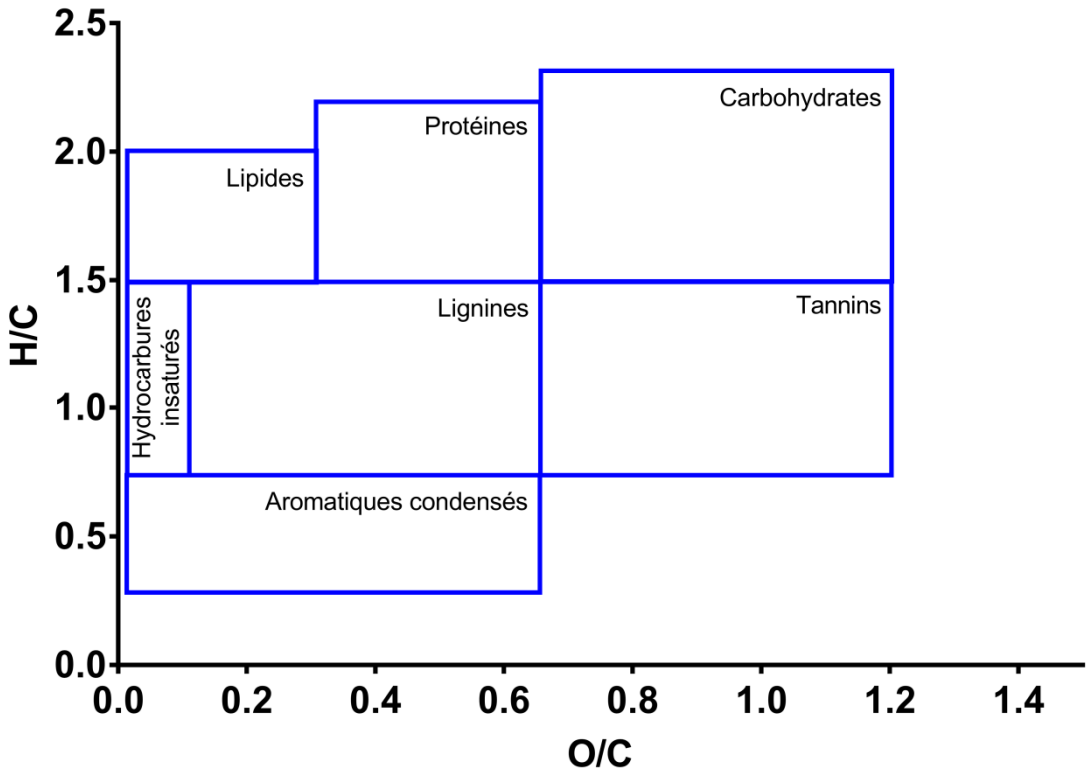


Figure 19. Diagramme de Van Krevelen et masques (rectangles bleus) catégorisant diverses classes de molécules.

Chaque catégorie moléculaire occupe une localisation plus ou moins précise sur le diagramme. En effet, la classe des lipides possède un rapport H/C compris entre 1,50 et 2,00 pour un rapport O/C de 0,00 et 0,30 compte tenu de leur formule moléculaire générale. Les protéines ont un rapport H/C de 1,50 à 2,15 et un rapport O/C de 0,30 à 0,65. Les carbohydrates possèdent un rapport H/C proche des deux catégories précédentes car il est de 1,50 à 2,25 mais un rapport O/C plus important compris entre 0,65 et 1,20. Les

hydrocarbures insaturés, les lignines et les tannins possèdent un rapport H/C commun compris entre 0,75 et 1,50 mais des rapports O/C différents, respectivement de 0,00 à 0,10, 0,10 à 0,65 et 0,65 à 1,20. Enfin, les composés aromatiques condensés ont un rapport H/C compris entre 0,25 à 0,75 et un rapport O/C compris entre 0,00 et 0,65.

Le diagramme de Van Krevelen permet en premier lieu une visualisation rapide de la catégorie moléculaire. Ainsi la formule brute permet d'obtenir une information sur la nature du composé. Bien évidemment, les rapports H/C et O/C permettent de définir des zones (masques) par classe de molécules mais il n'est pas toujours aussi simple de catégoriser parfaitement un échantillon car des chevauchements existent entre les zones. Cette représentation peut cependant être très informative lors d'analyses différentielles par comparaison de cohorte de divers échantillons.

L'article de Ohno et *al.* utilise notamment cette représentation pour comparer la composition d'échantillons de terre à différentes profondeurs et différentes localisations géographiques (Ohno *et al.*, 2014).

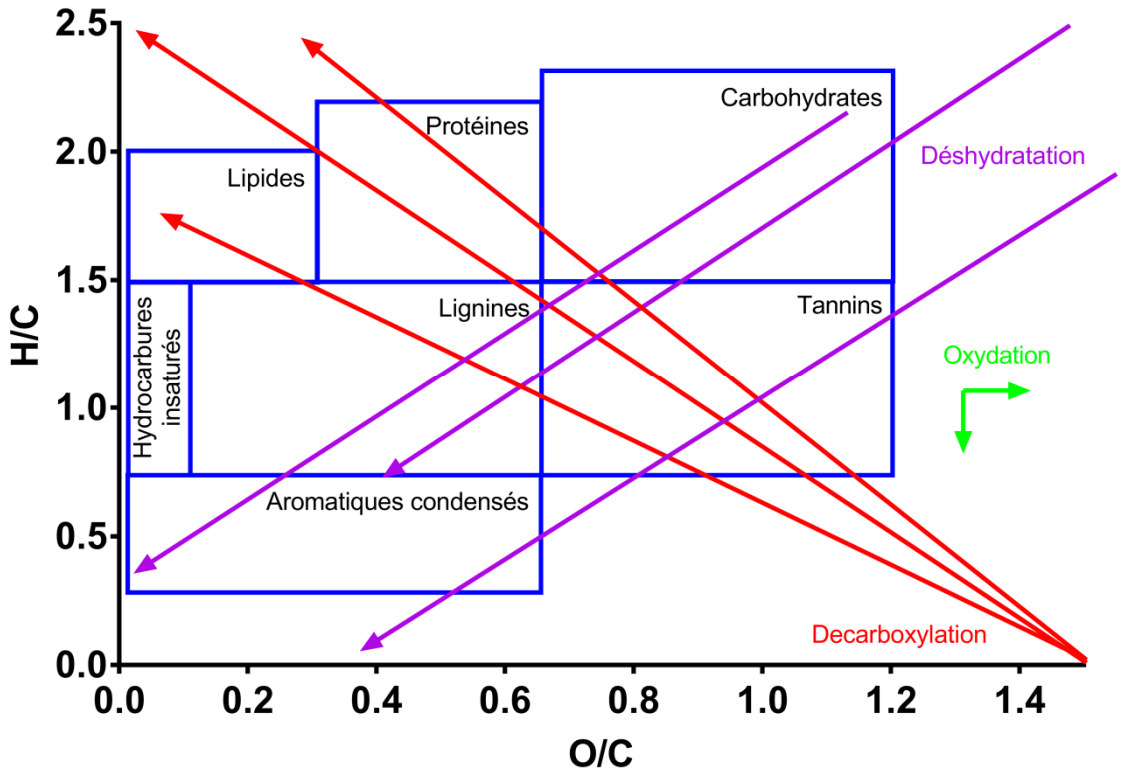


Figure 20. Diagramme de Van Krevelen avec masques et fléchages montrant divers sens de lecture et d'interprétations.

Le diagramme de Van Krevelen permet également de visualiser rapidement d'éventuelles carboxylation, déshydratation ou encore oxydation (Figure 20).

Plus la position d'un composé sur le diagramme se déplace de gauche à droite, plus le rapport O/C de ce composé est grand. Par conséquent i) soit son nombre d'atomes d'oxygène augmente (oxydation, flèche verte) ii) soit son nombre de carbones diminue, dans ce cas le rapport H/C augmente également (ce phénomène est symbolisé par les flèches rouges et est représentatif d'une décarboxylation). De la même manière, plus la position d'un composé se déplace de haut en bas, plus le rapport H/C de ce composé est petit. Par conséquent **i**) soit le nombre d'hydrogènes de ce composé diminue **ii**) soit le nombre de carbones de ce composé augmente. Ainsi, les flèches violettes représentent une déshydratation. *In fine*, l'analyse par diagramme de Van Krevelen permet le suivi d'un échantillon à des temps différents ou ayant subi divers traitements physique, chimique ou enzymatique.

3.7. L'exploitation des données MS/MS

Le couplage de plusieurs analyseurs identiques (double Q, triple Q) ou grâce à des instruments hybrides (Q-TOF) (Gelpí, 2009) permet d'obtenir des spectres de masse de fragmentation (MS/MS) voire même de réaliser de multiples étapes de fragmentation (MS^n) (Payne & Glish, 2005). Les informations collectées au cours de ce type d'analyse, à savoir le m/z des ions pseudo-moléculaires appelés aussi les ions parents ou les ions précurseurs et le m/z des ions moléculaires fragments aussi appelés les ions fils permettent d'obtenir des informations structurales.

Il existe plusieurs méthodes de fragmentation moléculaire en spectrométrie de masse. La dissociation induite par collision ou « *Collision induced dissociation* » (CID) est la technique la plus couramment utilisée pour obtenir des spectres de fragmentation de biomolécules. Les ions à analyser sont introduits dans une cellule de collision contenant un gaz de collision neutre (un gaz rare). Les ions s'entrechoquent avec les molécules du gaz de collision et se fragmentent. La quantité de gaz de collision dans cette cellule est toujours la même, la vitesse d'introduction plus ou moins importante des ions à analyser dans celle-ci permet la fragmentation. La fragmentation des ions est due à l'augmentation de l'énergie interne des ions à la suite des collisions. Ainsi, des ions d'énergie interne excessive sont instables et retrouvent leur stabilité par fragmentation.

Une nomenclature d'annotation (Roepstorff & Fohlman, 1984) a été établie pour définir les ions fragments générés. (Figure 21). En ce qui concerne les peptides, les collisions effectuées en CID génèrent majoritairement des ions fragments « b » et « y ».

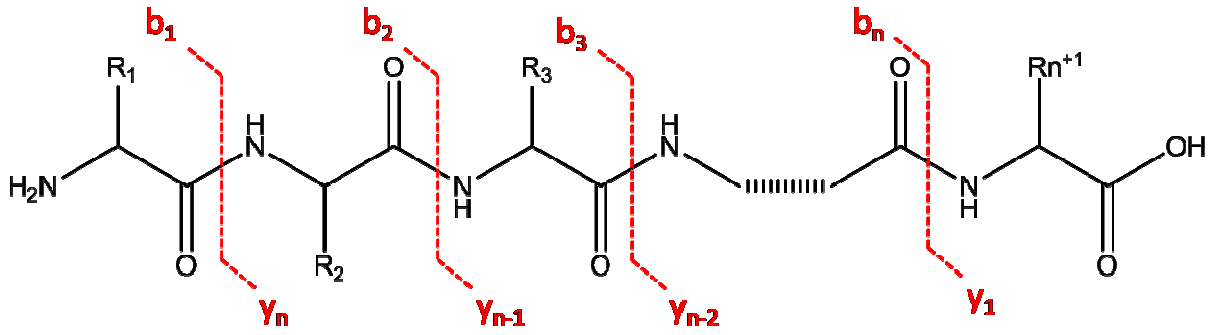


Figure 21. Illustration schématique de la génération des ions fragments pour un peptide issu d'une fragmentation par CID (Roepstorff & Fohlman, 1984).

D'autres modes de fragmentation moléculaire existent comme la dissociation par transfert d'électrons (*electron transfer dissociation*, ETD)(Good *et al.*, 2007; Syka *et al.*, 2004; Zubarev *et al.*, 2008), la dissociation par capture d'électrons (*electron capture dissociation*, ECD)(Cooper *et al.*, 2005; Syrstad & Turecek, 2005; Zubarev, 2004) et la dissociation multiphotonique infrarouge (*infrared multiphoton dissociation*, IRMPD) (Sleno & Volmer, 2004). Ces modes de fragmentation sont souvent utilisés de façon plus spécifique et ne sont pas encore pleinement exploités pour des applications de petites molécules en dehors de la protéomique.

Il est également possible de transmettre plus d'énergie que nécessaire aux molécules pour leur ionisation en source et par conséquent entrainer une fragmentation de celles-ci dès la source, ce phénomène est appelé « fragmentation en source ». Cette énergie transmise lors de l'ionisation conduit également à d'autres phénomènes de fragmentation précoce qui ont lieu tout au long du déplacement des ions au travers des différentes parties du spectromètre de masse.

3.7.1. Détermination des formules moléculaires

Un spectromètre de masse est également capable de mesurer les masses exactes des ions fils produits lors de la fragmentation. Cette information peut aussi être utilisée afin d'obtenir la composition élémentaire. La formule moléculaire candidate de la molécule non fragmentée est ainsi déduite par combinaison de celles de plusieurs ions fragments (Jarussophon *et al.*, 2009; Konishi *et al.*, 2007). La société Bruker a développé un algorithme au travers du logiciel « *SmartFormula*3D » capable d'exclure les formules moléculaires incorrectes à partir d'ions fragments. Le logiciel SIRIUS (Böcker & Rasche, 2008) précédemment évoqué possède également cette fonctionnalité.

Récemment, Hill *et al.* ont développé une approche utilisant des calculs de formule élémentaire couplés à une stratégie de recherche en base de données pour générer *in silico* les spectres de masse CID du composé (Hill *et al.*, 2008). Les spectres de fragmentation obtenus *in silico* sont générés par le logiciel *MassFrontier* puis comparés avec les spectres expérimentaux obtenus en fragmentation CID. Cette approche combinée avec des

contraintes supplémentaires de filtration peut être utilisée pour annoter les spectres de fragmentation, à condition que la structure moléculaire du composé soit référencée dans les banques de données.

3.7.2. Base de données spectrales

La recherche en bibliothèque spectrale peut être réalisée à partir d'un spectre de masse à haute résolution mais aussi avec les spectres de fragmentation de type MS^n. Le but de cette recherche est soit d'obtenir la structure des composés (déjà présents dans la bibliothèque) ou d'obtenir une structure partielle à partir des spectres de fragmentation de composés similaires. À cet effet, un spectre de masse expérimental est recherché dans une grande collection de spectres de masse déjà enregistrés et stockés dans une banque de données. Une revue générale inventorie les bibliothèques de spectre de masse (Halket *et al.*, 2005) et des algorithmes de recherche (McLafferty *et al.*, 1974; Stein & Scott, 1994).

Les algorithmes de recherche ont été développés en premier lieu pour faciliter et fiabiliser l'interprétation de spectres de masse issus d'une ionisation par IE (Sparkman, 1996), ils comprennent l'algorithme INCOS et l'appariement probabiliste (*Probability-Based matching*, PBM) (McLafferty *et al.*, 1974; Stein & Scott, 1994). La taille des bibliothèques MS/MS publiquement disponibles est faible en comparaison de celle des bibliothèques commerciales qui couvrent plusieurs centaines de milliers de spectres de masse essentiellement générés par IE (Wiley et NIST). Actuellement, la collection MS/MS NIST08 est une grande base de données commerciale de 14 802 spectres MS/MS issus de 5308 ions précurseurs fragmentés à différentes énergies de collision. Ainsi, une grande variété de bibliothèques commerciales ont été générées mais sont souvent réservées à certains types et configurations d'appareils.

Des banques de données publiques comme *Massbank* (Horai *et al.*, 2010; Horai *et al.*, 2008) et *ReSpect* (RIKEN)(Akiyama *et al.*, 2008; Matsuda *et al.*, 2010; Sawada *et al.*, 2009) couvrent actuellement 24772 spectres de masse et spectres de fragmentation issus de 13200 composés. Une bibliothèque de spectres de masse en tandem, effectuée sur un spectromètre de masse de type ESI-Trap-MS/MS (spectromètre de masse doté d'une source électrospray et d'un analyseur hybride couplé à un piège à ions linéaire (Dresen *et al.* 2009))

a été créée pour des applications médico-légales. Cette bibliothèque est composée de 5600 spectres MS/MS de 1253 composés (obtenus avec des énergies de collisions différentes). Plusieurs bibliothèques spécialisées comme celle-ci sont recensées pour le dépistage toxicologique mais également l'analyse des drogues (Mueller *et al.*, 2005; Schreiber *et al.*, 2000) mais elles sont généralement plus petites. Une bibliothèque interne de spectres MS/MS réalisés à partir de 1200 produits naturels composés en majorité de molécules pouvant s'ioniser en mode positif a été générée dans la référence (Fredenhagen *et al.*, 2005).

Les spectres MS/MS générés à partir de sources d'ionisation douces suivis d'un mode de fragmentation de type CID et ETD ne sont pas aussi reproductibles, surtout s'ils sont obtenus à partir d'instruments différents que les spectres de masse MS/MS issus d'une source IE. Ainsi, la création d'une bibliothèque spectrale MS/MS reproductible et transférable à de multiples types d'instruments reste difficilement possible quelles que soient les molécules d'intérêt (Bristow *et al.*, 2004; Hopley *et al.*, 2008; Milman, 2005). La raison majeure réside dans le manque de normalisation de l'énergie de collision à utiliser lors de fragmentations MS/MS. Un indice de l'énergie de fragmentation a été proposé pour la LC-MS (Palit & Mallard, 2009) afin de normaliser les énergies de collision et créer des spectres reproductibles comparables à ceux générés par l'énergie normalisée de 70 eV utilisée pour les spectres de masse issus d'une source IE. Une autre étude a été réalisée dans le but de comparer les spectres MS/MS obtenus à partir de spectromètres de masse i) à analyseur quadripolaire couplé à un analyseur TOF (Q-TOF), ii) équipé d'un triple quadripôle (QqQ) et d'un analyseur FT-ICR. La conclusion de cet article relate qu'une plateforme de spectres MS/MS peut être obtenue avec plusieurs réglages d'énergies de fragmentation (Oberacher *et al.*, 2009; Pavlic *et al.*, 2006) mais la tâche reste malgré tout longue et fastidieuse.

Pour la spectrométrie de masse couplée à la chromatographie en phase gaz (GC-MS), d'énormes bibliothèques de références spectrales sont couramment utilisées. En revanche pour la spectrométrie de masse couplée à la chromatographie liquide (LC-MS/MS), les bibliothèques contiennent moins de composés et sont limitées dans leur disponibilité.

3.7.3. Fragmentation *in silico*

La banque de données PubChem contient actuellement environ 94 millions de composés, tandis que même les plus grandes bibliothèques spectrales commerciales telles que « l'Institut National des Standards et Technologies » (version 11) et « le Registre Wiley » (9e édition) contiennent des spectres de masse pour seulement 200 000 et 600 000 composés, respectivement (Figure 22).

Figure 22. Proportion du nombre de molécules dans les banques de données spectrales NIST V11 (n=200 000) et Wiley V9 (n=600 000) en rouge et dans la banque de données de structure moléculaire PubChem (n= 94 000 000), en bleu.

Les bibliothèques spectrales expérimentales contiennent (et contiendront toujours) moins d'informations que les bases de données de structure moléculaire. Cette lacune peut être comblée par la création de spectres de masse MS/MS générés *in sillico* à partir de grandes bases de données de structure moléculaire telle que PubChem. Un algorithme de fragmentation *in silico* doit prévoir précisément les fragments de masse, leurs profils isotopiques et leurs abondances isotopiques. Un tel outil de modélisation de spectres de masse théorique *in silico* pourrait être utile car les spectres de masse expérimentalement obtenus pourraient alors être comparés par l'intermédiaire de cet outil aux données de bases de données de structure moléculaire. Cette approche est utilisée avec succès en protéomique depuis de nombreuses années car les règles de fragmentation des peptides sont relativement simples et constantes comparées à celles de composés plus complexes tels que les métabolites secondaires microbiens.

Les premières tentatives de prédiction de structure chimique potentielle et de prédiction des spectres de fragmentation, en utilisant des structures modèles, et des règles

de fragmentation spécifiques, ont été faites dans le cadre du projet DENDRAL dès 1965 (Lederberg, 1964). Cependant, ce projet a échoué dans son objectif majeur d'élucidation de structure automatique à partir de données de spectrométrie de masse, et les recherches ont été interrompues.

Plusieurs algorithmes de simulation de spectres de masse MS/MS ont été publiés dans la littérature. Cependant, bon nombre de ces programmes ne furent jamais réalisés commercialement ou disponibles publiquement. Le problème de la plupart des algorithmes est de simuler et de calculer les intensités des ions fragments (Clark & Jurs, 1981) mesurées expérimentalement (Gay *et al.*, 2002; Timm *et al.*, 2008; Zhou *et al.*, 2008). Ce problème n'est pas encore résolu pour la plupart des petites molécules sous différents modes d'ionisation.

Le taux de réussite de tout algorithme doit être déterminé par une étude de validation en confrontant des spectres de masse expérimentaux de molécules connues à une recherche d'identité à partir d'une bibliothèque spectrale générée *in silico*. Notons que pour être cohérente, cette étude de validation doit être réalisée à partir d'une bibliothèque dont la diversité structurale et le nombre de composés doivent être importants pour permettre une analyse statistique.

Plusieurs exemples de réussite en termes de production de spectres MS/MS *in silico* ont été mis en place pour des molécules composées de monomères structuraux connus et répondant à des règles de fragmentation définies, comme les lipides, les oligosaccharides (Zhang *et al.*, 2005), les glycanes (Kameyama *et al.*, 2006), et les peptides (Chen *et al.*, 2001). MASSIS/MASSIMO est un système de simulation spectrale disposant de règles de fragmentation découlant des spectres issus d'une source IE qui comprend les réarrangements de McLafferty, les rétro-Diels-Alder et les pertes de neutre (Chen *et al.*, 2003; Chen *et al.*, 2003; Fan *et al.*, 2005). L'algorithme public *MetFrag* (Wolf *et al.*, 2010) compare *in silico* des spectres de masse, obtenus par une approche de dissociation de liaison, à des spectres de masse expérimentaux et attribue un score à tous les résultats potentiels. Le logiciel *Metfrag* gère uniquement des entrées présentes dans les banques de données PubChem, KEGG, Chemspider, MetaCyc, FOR-IDENT, LipidsMaps, ChEBI et HMBD.

La plupart des métabolites secondaires qui nous intéressent ne sont pas répertoriés dans ces banques de données, par conséquent l'utilité de ce dernier reste limité.

Aujourd'hui, il existe trois grands logiciels commerciaux capables de générer des spectres MS/MS *in silico* à partir de banques de données de structure moléculaire : MOLGEN-MS (Université de Bayreuth), ACD/MS fragmenter (Advanced Chemistry development Inc., (Pelander *et al.*, 2009)), et MassFrontier (HighChem Ltd.). Une étude de validation (Schymanski *et al.*, 2009) a comparé le taux de réussite de ces trois programmes et a conclu que la simulation de spectres de masse MS/MS est encore très loin de pouvoir être utilisée quotidiennement en routine. Depuis 2012, Il existe une solution gratuite appelée *iSNAP fragmenter* sur une plateforme dédiée à l'analyse de données de spectrométrie de masse appelée iSNAP Analogue (Ibrahim *et al.*, 2012). La première version de 2012 permet de fragmenter *in silico* une molécule structurellement linéaire à partir de son SMILES. Depuis l'algorithme a subi une mise à jour permettant la fragmentation de molécules cycliques et notamment de NRPs

3.7.4. Combinaison de fragments

Contrairement aux approches précédentes, la fragmentation combinatoire vise à interpréter les signaux de masse d'un spectre MS/MS expérimental. Ceci est basé sur l'hypothèse que la plupart des signaux résultent de sous structures du composé sans réarrangement majeur. Les premières méthodes de fragmentation combinatoire telles que « *Elucidation of Product Ion Connectivity* » (EPIC) (Hill & Mortishire-Smith, 2005) et *Fragment IDentificator* (FID) (Heinonen *et al.*, 2008) n'avaient pas pour but de trouver la structure moléculaire mais d'expliquer chaque signal dans un spectre de fragmentation d'une molécule connue. Ces premières approches énumèrent tous les fragments en appliquant toutes les combinaisons de clivages. La liste des signaux potentiels est ensuite comparée à celle des signaux mesurés expérimentalement. Cette énumération exhaustive est très lente et, par conséquent, ne peut être appliquée pour des grandes structures moléculaires.

Plus tard, Wolf et al. (Wolf *et al.*, 2010) ont introduit la méthode heuristique pour ce problème au logiciel *MetFrag*. Celui-ci est plus rapide que les approches mentionnées ci-dessus ; une base de données de structures complètes peut être utilisée pour identifier le

composé dont la structure moléculaire correspond le mieux au spectre MS/MS expérimental (à condition que la structure soit présente dans la base de données). Avec la fragmentation combinatoire, il est important de prendre en compte l'énergie utilisée lors de la fragmentation, car certaines liaisons sont plus ou moins faibles, et par conséquent l'apport énergétique pour cliver telles ou telles liaisons peut être différent. L'un des problèmes majeurs est de choisir l'énergie de collision appropriée à chaque molécule afin d'obtenir un maximum d'informations. Par exemple, le type de liaisons chimiques (simple, multiple ou aromatique) (Ridder *et al.*, 2012) peut être utilisé pour calculer l'énergie nécessaire au clivage d'une liaison. Heinonen *et al.* (Heinonen *et al.*, 2008) ont proposé un programme appelé « *Mixed Integer Linear Program* » (MILP) pour résoudre ce problème, mais en raison de la complexité des calculs, les temps d'analyse trop importants posent problème même pour de petites molécules. Hill et Mortishire-Smith (Hill & Mortishire-Smith, 2005) ainsi que Wolf et al. (Wolf *et al.*, 2010) ont proposé de diminuer l'espace de recherche en limitant le nombre de clivages permis.

Plus récemment, Gerlich et Neumann (Gerlich & Neumann, 2013) ont créé le logiciel *MetFusion*, qui combine l'algorithme de *MetFrag* et la recherche en bibliothèque spectrale depuis la banque de données *MassBank* pour améliorer l'identification des composés.

Une préoccupation majeure en ce qui concerne la fragmentation combinatoire est que les fragments résultant de réarrangements structuraux moléculaires issus de la chimie des molécules en phase gazeuse ne sont pas couverts par cette approche. Un autre problème est de trouver une bonne fonction de coût, c'est-à-dire l'importance accordée aux différents clivages potentiels. Par exemple, la notation de Wolf qui considère les énergies obligatoires de dissociation aboutit à une précision de la prédiction des fragments qui diminue lorsque l'on augmente le nombre autorisé de clivages obligatoires. Il est possible que les clivages obligatoires génèrent des fragments improbables sans expliquer plus de signaux (Ridder *et al.*, 2012).

3.7.5. Arbre de fragmentation

L'approche des arbres de fragmentation part du postulat simple que tous les ions fragments présents dans un spectre MS/MS sont issus d'un précurseur ; par conséquent la

formule moléculaire des fragments fait partie de la formule moléculaire du précurseur. En résumé, les formules moléculaires des fragments doivent être des sous-formules de la formule du précurseur. Compte tenu de la structure moléculaire du composé et du spectre MS/MS mesuré, l'attribution des signaux de fragments au composé peut être réalisée afin d'en tirer un « diagramme de fragmentation ». Si la structure moléculaire n'est pas connue, ceci reste très difficile. Les arbres de fragmentation sont similaires aux « schémas de fragmentation », mais sont extraits directement à partir des données, sans connaissance sur la structure d'un composé.

Ainsi, un arbre de fragmentation est constitué de nœuds, correspondant aux précurseurs et aux fragments et ces derniers sont reliés entre eux par des arêtes. Chaque nœud est annoté avec la formule moléculaire du fragment. Les nœuds suivants sont implicitement annotés avec des formules moléculaires déduites des pertes issues du fragment précurseur.

Pour l'arbre résultant, chaque nœud « explique » un signal dans le spectre de fragmentation mesuré. Autrement dit, la différence de masse entre la formule moléculaire du nœud et la masse maximale observée est inférieure à la masse supposée (Figure 23).

La Figure 23 illustre l'élaboration d'un arbre de fragmentation issue du spectre MS/MS d'un ion précurseur de m/z 166,1. Huit ions fragment sont générés par fragmentation de celui-ci. Après l'attribution d'une formule moléculaire de chaque signal, un arbre de fragmentation probable est alors créé, capable de hiérarchiser les masses des fragments entre eux (Figure 23 B).

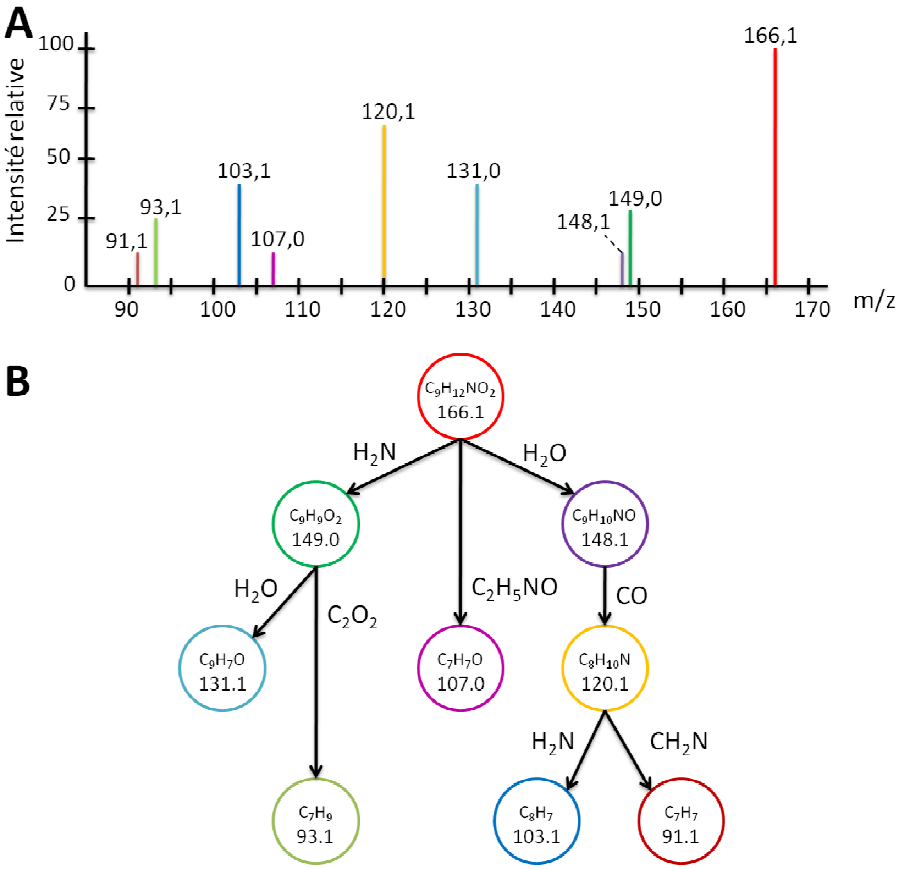


Figure 23. Illustration schématique de l'annotation d'un spectre de masse (A) depuis un arbre de fragmentation (B).

Les arbres de fragmentation ont été introduits initialement pour cette tâche (Böcker & Rasche, 2008). En 2012, Hufsky *et al.* ont montré que les arbres de fragmentation sont des descriptions raisonnables du processus de fragmentation et, par conséquent, peuvent également être utilisés pour obtenir des informations complémentaires sur des composés inconnus (Hufsky *et al.*, 2012).

Malheureusement, le spectre de fragmentation d'un composé peut être expliqué par de nombreux arbres de fragmentation : plus le composé est grand et plus les possibilités augmentent ainsi que les erreurs. Bien sûr le phénomène s'aggrave avec des spectres de fragmentation possédant de nombreux fragments et si l'on considère des éléments autres que les atomes de carbone (C), hydrogène (H), azote (N), oxygène (O), phosphore (P) et soufre (S).

Pour trouver le meilleur (et espérons correct) arbre de fragmentation, l'optimisation combinatoire peut être utilisée. Malheureusement, la recherche d'un arbre de fragmentation optimale s'avère informatiquement difficile et ralentit considérablement la recherche. Le nombre d'arbres à fragmentation peut facilement atteindre des proportions importantes, même pour les petits composés, parfois même un nombre beaucoup plus grand que le nombre d'atomes observables dans l'univers (Hufsky *et al.*, 2012). Néanmoins, des algorithmes qui garantissent de trouver la solution optimale ont été développés et sont également rapides dans la pratique, l'arbre de fragmentation optimale peut être trouvé en quelques secondes (Hufsky *et al.*, 2015).

Notons que la comparaison de deux composés inconnus sur la base de leurs spectres de fragmentation est possible. Les arbres de fragmentation correspondants sont alors « alignés » et les cascades de fragmentations similaires dans les deux arbres sont identifiées et marquées. Même pour des composés qui ne peuvent être identifiés, il est alors possible de tirer des informations utiles, comme de visualiser un rassemblement de famille moléculaire.

Les arbres de fragmentation ne doivent pas être confondus avec les « arbres spectraux » (Sheldon *et al.*, 2009) qui décrivent les relations entre spectres de fragmentation séquentielle d'un composé unique. Les "arbres de fragmentation" ne contiennent aucune information en ce qui concerne les mécanismes de fragmentation.

3.7.6. Réseaux moléculaires ou réseaux de similarité spectrale

L'approche des réseaux moléculaires a été initiée par des chercheurs américains de San Diego (Californie, USA) du laboratoire de Pieter Dorrestein. Ils ont mis en ligne une plateforme appelée « *Global Natural Product Social molecular networking* » (GNPS) regroupant plusieurs outils pour la déréplication de molécule (http://gnps.ucsd.edu). GNPS possède également un serveur qui héberge des données spectrales (*Mass spectrometry Interactive Virtual Environment, MassIVE*) dans lequel il est possible non seulement de déposer des données mais également d'annoter et d'analyser les données de la communauté. Mais le principal engouement autour de cette plateforme résulte de la

possibilité de générer des réseaux de similarité spectrale grâce à nos propres données de fragmentation.

Le groupe de chercheurs part du principe que des voies de fragmentation similaires vont donner des fragments ou des pertes de neutre communes pour des molécules structurellement proches. Ainsi, ils élaborent une méthode qui permet de rassembler un ensemble d'analogues structuraux à partir de spectre de fragmentation afin d'interpréter plus rapidement et dans son intégralité des données de fragmentation (MS/MS). Les réseaux moléculaires ou plutôt réseaux de similarité spectrale ne sont, ni plus ni moins, qu'une manière de classer des spectres de masse de fragmentation par homologie de spectre de fragmentation, sans décrire réellement les voies de fragmentation.

Cet algorithme a été initialement conçu pour l'analyse de protéines. Les spectres de fragmentation sont comparés deux à deux par alignement (superposition) spectral et un coefficient de similarité est attribué. Ce coefficient est appelé « score cosinus » (*cosinus score,* CS). Dans un premier temps, un énorme tri des spectres MS/MS est effectué afin de simplifier et réduire les données en fusionnant les spectres identiques (*MS cluster*). Ensuite, les spectres de fragmentation sont appariés en fonction de leur intensité relative et les différences entre les valeurs de m/z. Par conséquent, deux spectres de fragmentation parfaitement superposables posséderont un CS de 1 et à l'inverse deux spectres de fragmentation sans fragments ou perte de neutre en commun auront un CS de 0. Puis afin de créer de plus grands groupes ou des familles avec des molécules apparentées chimiquement, un seuil de similarité est attribué en se basant sur le CS. En général, une valeur supérieure à 0,7 reflète une ressemblance. Puis, les résultats sont mis en page graphiquement sous forme de réseaux avec le logiciel « Cytoscape », ainsi des groupes se dessinent (Figure 24).

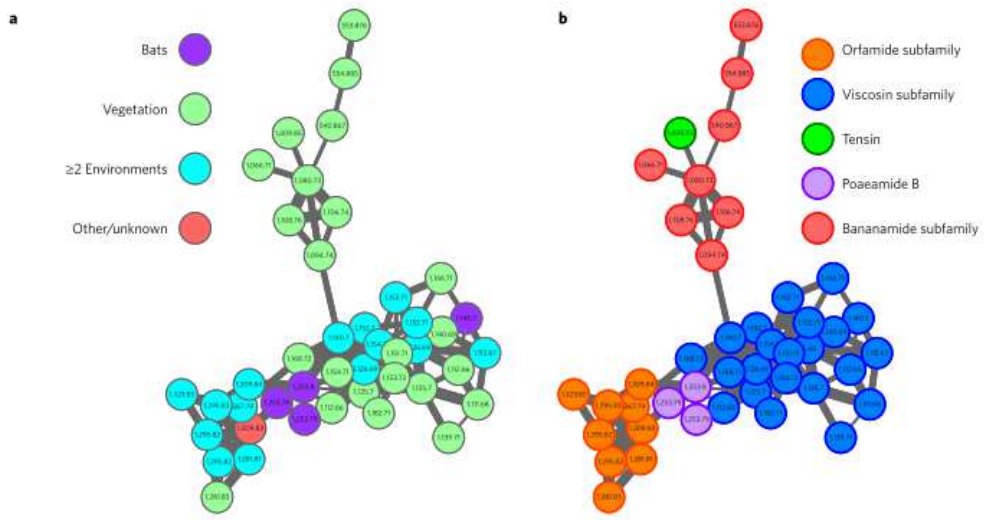


Figure 24. Réseaux moléculaires issus d'une analyse par spectrométrie de masse. Les couleurs de chaque nœud délimitent les environnements d'où proviennent les molécules (a). Familles de différents lipopeptides cycliques avec différentes couleurs de nœud (b).

Les réseaux moléculaires illustrés dans la Figure 24 présentent à la fois une interprétation selon la localisation géographique des échantillons et selon des ressemblances spectrales coïncidentes avec des spectres de fragmentations présents dans la base de données *MassIVE* (Nguyen *et al.*, 2016).

La puissance de cette approche réside essentiellement dans sa capacité à trier et organiser les données sans connaissance préalable de l'échantillon.

Des biais peuvent malgré tout apparaitre de par la banque de données elle-même dans laquelle tout le monde peut déposer n'importe quels spectres de masse de fragmentation car il n'existe aucune règle de soumission. La plateforme limite également la taille des fichiers d'interrogations par conséquent le retraitement de données trop volumineuses issues d'une analyse chromatographique contenant beaucoup d'informations (absorbance U.V., données de mobilité ionique…) et des spectres contenant toutes les informations de distribution isotopique.

Matériels et méthodes

1. Solvants et molécules

Les solvants utilisés pour les analyses en spectrométrie de masse tels que l'eau ultra pure (Fisher scientific, Illkirch, France), l'acétonitrile (ACN, Fisher scientific, Illkirch, France) mais aussi les acides comme l'acide formique (FA, Sigma-Aldrich, France), l'acide trifloroacétique (TFA, Sigma-Aldrich, France) et l'acide borique (BA, Sigma-Aldrich, France) sont de la plus haute qualité possible pour garantir la qualité des analyses par spectrométrie de masse.

Plusieurs molécules de type NRPs de pureté supérieure à 98% ont été achetées chez deux fournisseurs majeurs : l'actinomycine (référence : A1410, Sigma-Aldrich, France), l'entérobactine (référence : E3910, Sigma-Aldrich, France), la gramicidine (référence : 1298004, Sigma-Aldrich, France), la vancomycine (référence : V0045000, Sigma-Aldrich, France), un standard de surfactine (référence : S3523, Sigma-Aldrich, France), un standard de fengycine(référence : LP/F1, Lipofabrik, France) et la mycosubtiline (référence : LP/M1, Lipofabrik, France).

2. Souches

Les souches utilisées lors de ce travail de thèse ont été généreusement fournies par le Dr Sandra Matthijs de l'institut de recherche LABIRIS (anciennement Institut de Recherches microbiologiques Wiame (IRMW)) au sein de l'Université Libre de Bruxelles (Tableau 1).

Tableau 1. Liste des souches de *Pseudomonas* utilisées

#	Espèce	Souche	Numéro d'accession du génome
1	*P. aeruginosa*	PA7	CP000744
2	***P. agarici***	**NCPPB 2289T**	**AKBQ00000000**
3	*P. alcaligenes*	NBRC 14159T	BATI00000000.1
4	***P. balearica***	**DSM 6083T**	**CP007511**
5	*P. putida*	GB-1	CP000926
6	***P. caeni***	**DSM 24390T**	**NZ_ATXQ00000000**
7	*P. corrugata*	NCPPB 2445	LIGR01000000
8	***P. deceptionensis****	**DSM 26521T**	**NZ_JYKX00000000**
9	*P. entomophila*	L48T	NC_008027
10	***P. fluorescens***	**ATCC 17400**	**JENC01000000**
11	*P. gingeri*	NCPPB 3146T	AKBP00000000.1
12	***P. mediterranea***	**CFBP 5447T**	**AUPB01000000**
13	*P. monteilii*	DSM 14164T	JHYV01000000
14	***P. putida***	**F1**	**NC_009512**
15	*P. putida*	W15Oct28	NZ_JENB01000001
16	***P. putida***	**W619**	**CP000949**
17	*P. thivervalensis*	DSM 13194T	LHVE00000000.1
18	***P. protegens***	**Pf-5**	**NC_004129**
19	*Pseudomonas sp.*	W2Aug9	-
20	***Pseudomonas sp.***	**AF76**	**-**

La température optimale d'incubation utilisée pour la croissance des microorganismes est 30°C à l'exception de l'espèce *P.deceptionensis* pour laquelle l'incubation est réalisée à 20°C.

3. Milieux de culture

Les composants organiques et minéraux des milieux de culture sont produits industriellement. Par conséquent, la composition chimique précise des composants organiques varie en fonction du fournisseur mais également en fonction du lot de production. La totalité des composants des milieux de cultures utilisés au cours de ce travail de thèse ont été achetés chez les fournisseurs ou distributeurs français usuels. Tous les numéros de lots sont mentionnés. Les caractéristiques principales et composition précise des milieux de culture sont détaillées ci-dessous pour rappel.

En fonction de l'objectif (croissance cellulaire, production de lipopeptides ou de sidérophore), différents milieux peuvent être utilisés. De l'agar à 1,7% (p/v) peut être ajouté

aux milieux de culture qui n'en contiennent pas afin d'obtenir un milieu gélosé. Les milieux de culture sont stérilisés par l'autoclavage avec une température de 121°C durant 20 minutes.

4. Production de lipopeptides

4.1. Bouillon de lysogénie (LB)

Le milieu LB est un milieu de croissance bactérien empirique couramment utilisé en microbiologie (Bertani, 1951). Ce milieu est préparé avec 1% (p/v) de tryptone (FlukaAnalytical, Sigma-Aldrich, Lesquin, France, lot n°BCBN7379V), 0,5% (p/v) d'extrait de levure (Sigma-Aldrich, Lesquin, France, lot n°BCBQ9331V), 1% (p/v) de chlorure de sodium (NaCl, Sigma-Aldrich, Lesquin, France) puis ajusté à pH 7.

4.2. Milieu de LANDY

Le milieu de Landy est un milieu complexe très riche en cofacteurs enzymatiques. Il a été optimisé pour la production de lipopeptides (Coutte *et al.*, 2010). Le milieu LANDY contient 5% de glucose à 400 g/L, 2,5% d'acide glutamique à 100g/L, 2,5% d'extrait de levure à 50 g/L, 2,5% de sulfate de magnésium ($MgSO_4$, Sigma-Aldrich, Lesquin, France) à 10g/L, 2,5% de chlorure de potassium (KCl, Sigma-Aldrich, Lesquin, France) à 20 g/L, 2,5% d'hydrogénophosphate de potassium (K_2HPO_4, Sigma-Aldrich, Lesquin, France) à 20 g/L, 8% de sulfate de cuivre ($CuSO_4$, Sigma-Aldrich, Lesquin, France) à 3,2 mg/L, 8% de sulfate de manganèse ($MnSO_4$, Sigma-Aldrich, Lesquin, France) à 100 mg/L, 8% de sulfate de fer ($FeSO_4$, Sigma-Aldrich, Lesquin, France) à 3 mg/L, 5% de tampon MOPS (Sigma-Aldrich, Lesquin, France) à 420 g/L et 5% de sulfate d'ammonium ($(NH_4)_2SO_4$, Sigma-Aldrich, Lesquin, France) à 46 g/L. Le pH est ajusté à 7 à l'aide d'hydroxyde de potassium (KOH) à 0,1 % (Sigma-Aldrich, Lesquin, France).

5. Production de sidérophores

La production de sidérophores est conditionnée par l'absence de fer dans le milieu. Pour cette raison, toute la verrerie est préalablement rincée avec de l'acide nitrique à 10% (v/v) et l'eau utilisée pour l'élaboration des milieux de culture est préalablement traitée par une résine Chelex® 100 (Biorad, Steenvoorde, France) afin de la carencer en fer (Leclère *et al.*, 2009).

5.1. Milieu King B

Historiquement, ce milieu a été décrit pour la première fois par King, Ward et Raney en 1954 (King *et al.*, 1954). Le milieu King B (Sigma-Aldrich, France, lot n° BCBS0028V) est préparé selon les recommandations du fournisseur. Ce milieu renferme des cations comme le magnésium (issu du $MgS0_4$ à 0,15% (p/v)) nécessaires à la production de pyoverdines.

5.2. Le milieu casamino-acids (CAA)

Le milieu CAA Bacto™ Difco™ (Fisher scientific, France, lot n°12L729A) est issu d'une hydrolyse acide de caséine contenant de faibles concentrations en chlorure de sodium et en fer. L'hydrolyse acide est réalisée en présence de 6 M d'HCl chauffé à 110°C pendant 24h. Durant ce processus, les protéines sont dégradées de manière aléatoire et non spécifique. La condition acide induit une transformation des fonctions amides en fonctions acides (glutamine (Gln/Q) en glutamate (Glu/E) et asparagine (Asn/N) en aspartate (Asp/D)). La température importante prolongée sur 24h induit la dégradation de certains composés comme le tryptophane par exemple. Il est préparé selon les recommandations du fournisseur et le pH est ajusté à 7.

6. Identification de souches par MALDI-TOF-MS

6.1. Méthode de dépôt sur cible MALDI

Une à deux colonies sont prélevées à partir des boîtes de Pétri à l'aide d'un cône de pipette de 100 µl puis sont étalées, à la manière d'un frottis, directement sur un des 384 emplacements d'une cible MALDI préférentiellement de type Ground Steel (MTP 384 target plate ground steel BC, Bruker Daltonics, Allemagne). Après séchage, le dépôt est ensuite recouvert d'1µl d'acide formique 70% (v/v) puis séché à température ambiante. Un microlitre de solution de lavage, composée d'eau ultra pure contenant 0,1% de TFA, est déposé directement sur le frottis d'échantillon bactérien. Après quelques secondes, la goutte est ré-aspirée. L'opération est réalisée deux fois et le dépôt est de nouveau séché à température ambiante. Le dépôt est alors recouvert d'1µl de solution de matrice, l'acide α-Cyano-4-hydroxycinnamique (HCCA, *α-Cyano-4-hydroxycinnamic acid*) à 10 mg/mL fraîchement solubilisée en ACN/eau ultra-pure/TFA, 50/47,5/2,5 (v/v/v) et laissé à température ambiante jusqu'au séchage complet. Le calibrant (BTS, de l'anglais *bacterial test standard*) est repris dans 50 µL d'une solution standard composée d'eau à 47,5 %, d'ACN à 50 % et de TFA à 2,5 % (v/v/v). Après incubation, cinq minutes à température ambiante, 1 µL de celui-ci est déposé au centre de huit échantillons afin de réaliser leur calibration. Puis le dépôt est recouvert d'1 µL de la même solution de matrice préalablement décrite.

6.2. Acquisition d'empreinte spectrale

A chaque session d'identification par spectrométrie de masse, l'Autoflex Speed™ (Bruker Daltonics) sous FlexControl 3.0 (Bruker Daltonics) est calibré sur la gamme de masse d'intérêt à l'aide d'un standard de calibration dénommé *bacterial test standard* (BTS, Bruker Daltonics, référence 8255343). Pour rappel, ce standard de calibration BTS est une extraction éthanolique des protéines d'*Escherichia coli DH5 α* supplémentée avec deux protéines purifiées non apparentées à l'espèce (la RNase A (Mw = 13683,2 Da) et la myoglobine (Mw = 16952,3 Da)) afin de couvrir la gamme de masse 2-20 kDa. La méthode de mesure en masse pour la calibration est sensiblement la même que la méthode MBT_FC.par décrite ci-dessous exceptée que la plage d'intensité laser se situait de 20 à 50% de la puissance laser totale.

Tous les échantillons sont déposés en triplicat sur la cible MALDI. Les spectres de masse MALDI-TOF sont obtenus en utilisant la méthode de mesure en masse préconisée par le constructeur et dénommée « MBT_FC.par ». Brièvement, les spectres sont enregistrés en mode d'ions linéaire positif sur la gamme de masse de 2 à 20 kDa, avec des paramètres de voltage de la source d'ion 1 de 25,00 kilovolts (kV) et de 23,65 kV pour la seconde et sous une tension de lentille à 9 kV. L'extraction d'ions pulsés est fixée à 120 ns et la plage d'intensité laser est comprise entre 40 et 70% de la puissance laser totale pour une fréquence de tirs laser de 1 000 Hz. Pour chaque échantillon, les signaux de masse résultant de 1000 tirs laser sont accumulés pour obtenir un spectre de masse MALDI-TOF représentatif moyen (ou plus simplement dénommé profil MALDI-TOF d'une souche bactérienne).

6.3. Retraitement des spectres de masse

Les spectres sont ensuite retraités via le logiciel MALDI BioTyper® (version 3.0; Bruker Daltonics). Les spectres de masse subissent en premier lieu une soustraction de la ligne de base et un lissage spectral selon l'algorithme de Savistky-Golay. Pour permettre l'appariement plus aisé et plus rapide entre les profils MALDI expérimentaux des souches et les profils MALDI des souches présents dans la base de données, le logiciel MALDI BioTyper® transforme les signaux de masse des profils MALDI expérimentaux en signaux « bâton ». Les profils MALDI « bâton » ainsi obtenus sont ensuite comparés à ceux de la base de données BioTyper® DB-5989, contenant 5989 profils MALDI de souches de référence dont 5298 profils MALDI de bactéries, 626 profils MALDI de levures et 65 profils MALDI de champignons filamenteux.

La correspondance des profils MALDI entre les profils expérimentaux et théoriques est retranscrite selon le logarithme du score défini par le logiciel MALDI BioTyper® : le log (score) associé à un code couleur de pertinence-certitude d'identification. Brièvement, un log (score) supérieur à 2,3 (associé à la couleur vert foncé) indique une identification hautement probable au niveau du genre et de l'espèce. Un log (score) compris entre 2,0 et 2,3 (couleur vert clair) signifie une identification hautement probable au niveau du genre et probable au niveau de l'espèce. Un log (score) entre 1,7 et 2,0 (couleur jaune) implique

seulement une identification probable du genre. Enfin une valeur de log (score) inférieur à 1,7 (couleur rouge) signifie que le profil « bâton » expérimental ne possède que très peu de similitudes avec un profil MALDI stocké dans la base de données. L'identification de genre et d'espèce est alors faiblement probable.

7. Tests d'activités

7.1. Test du potentiel surfactant : l'effondrement de la goutte (*drop collapse*)

Pour rappel, ce test permet d'évaluer la présence de molécules aux propriétés surfactantes produites par un microorganisme dans le milieu de culture (Jain *et al.*, 1991). Une goutte de 10 µL de surnageant de culture est placée sur une surface hydrophobe, en général du ParafilmM® (BemysCompagny, Inc, Neenah, WI, USA). La présence de molécules tensioactives engendre l'effondrement ou l'étalement visible à l'œil, qui est alors évalué par rapport à deux témoins : l'un négatif, constitué d'eau ultrapure et l'autre positif, constitué d'une solution de surfactine à 1 mol/L. Cette méthode qualitative permet d'avoir une évaluation rapide de la production de surfactant par les microorganismes en culture (Figure 25).

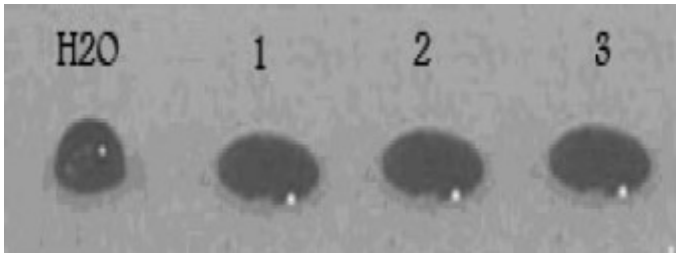


Figure 25. Illustration de l'effondrement d'une goutte (*Drop collapse*), de gauche à droite, une goutte d'eau, en 1, 2 et 3 une goutte de surnageant de culture d'un *Lactobacillus acidophilus* effondrée par la présence d'un surfactant (Arezoo *et al.* 2010).

7.2. Activité sidérophore : Dosage au chrome-azurol-S (CAS)

La concentration en sidérophores peut être déterminée grâce à un dosage chimique utilisant le réactif chrome-azurol-S (CAS) (Schwyn and Neilands, 1987). Pour rappel,

lorsqu'un chélateur capte le fer du principe actif (complexe CAS/Fer III et HDTMA) du dosage, sa couleur passe du bleu à l'orange. La méthode est assez sensible car le complexe CAS/Fer III et HDTMA possède un coefficient d'extinction de 100 000 $M^{-1}xcm^{-1}$ à 630 nm. L'élaboration du réactif CAS est réalisée avec de l'eau ultrapure carencée en fer afin de contrôler précisément la concentration de fer du réactif CAS. Pour cela, l'eau ultrapure est passée sur une résine Chelex® 100 (Biorad, Steenvoorde, France). Le réactif CAS est préparé avec 6 % de bromure d'hexadecyltrimethylammonium (HDTMA-Br) à 10 mM, 0,15 % de chlorure de fer III ($FeCl_3$) à 10 mM, 1,35 % d'acide chlorhydrique à 10 mM, 7,5 % de CAS à 2 mM puis 26% de pipérazine anhydre à 2 mM. La solution est stockée à l'abri de la lumière et de l'acide 5-sulfosalicylique est ajouté extemporanément à une concentration de 4 mM finale.

Le dosage s'effectue par mélange de 100µl de la solution à doser d'EDTA pour la gamme étalon ou surnageant, 40% d'eau appauvrie en fer et 50% de réactif CAS. La réaction est suivie par spectrophotométrie optique à une longueur d'onde de 630nm. Les sidérophores sont quantifiés en référence avec une gamme étalon d'EDTA (acide éthylène diamine tétraacétique) réalisée dans les mêmes conditions.

Il est possible de vérifier simultanément la présence de surfactant et celle de sidérophores en combinant les deux méthodes. La goutte déposée sur le Parafilm est alors constituée d'un µL de surnageant de culture et 4 µL de réactif CAS.

8. Préparation des surnageants de culture en vue de leur analyse en MS

8.1. Méthode de purification et de dessalage rapide

Les lipopeptides des milieux de culture sont purifiés et dessalés par ZipTip® Hypersep™ C18 10-200 µl (Thermo Ficher Scientific, Waltham, MA, États-Unis) tandis que les sidérophores sont purifiés et dessalés par ZipTip® Hypercarb™ 10-200 µl (Thermo Ficher Scientific). Brièvement, quel que soit leur nature, Les ZipTip® sont tout d'abord conditionnés par une étape d'activation de la phase chromatographique par cinq aspirations /refoulements successives d'environ 200 µl d'ACN suivie d'une étape de lavage de cinq aspirations/refoulements d'environ 200 µl d'eau ultrapure contenant 0,1% de TFA. Les

échantillons de surnageant de culture sont chargés sur les colonnes par plus ou moins vingt aspirations/refoulements d'environ 200 µl de surnageant de culture. Le matériel non fixé sur les phases chromatographiques est ensuite éliminé par dix aspirations/refoulements d'eau ultrapure contenant 0,1% de TFA. Enfin, l'élution et la concentration des composés retenus sur les phases chromatographiques sont réalisées par une seule aspiration/refoulement de 20µL d'ACN (ici, le TFA n'est pas ajouté en vue de l'analyse de l'élution en spectrométrie de masse).

8.2. Chromatographie liquide haute performance (HPLC, *high performance liquid chromatography*)

Les surnageants de culture sont séparés par chromatographie en phase inverse en utilisant un système HPLC ACQUITY biocompatible (Waters, Manchester, Royaume-Uni) équipé d'une colonne Coreshell Phenomenex C18 Kinetex (ID 3,0 mm × 150 mm, 2,6 µm, 100Å). L'élution chromatographique séparative est réalisée en utilisant un gradient linéaire d'ACN contenant 0,1% de TFA d'environ 1,27 % par minute (de 25 à 95% d'ACN en 55min) à un débit de 500 µL/min.

9. Spectrométrie de masse (MS)

9.1. Désorption ionisation laser assistée par matrice (MALDI) - temps de vol (TOF)

Les surnageants de culture et les fractions purifiées par ZipTip® sont dans un premier temps analysés par MALDI-TOF à l'aide de l'AutoflexSpeed™ (Bruker Daltonics) piloté par le logiciel Flexcontrol (version 3.0). A chaque session de mesure en spectrométrie de masse, l'AutoFlex Speed™ est préalablement calibré sur la gamme de rapports de m/z d'intérêt (700 – 3500 Da) à l'aide d'un standard de calibration le « *peptide calibration standard I* » (BrukerDaltonics). Le Tableau 2 reprend la liste exhaustive des peptides de calibration utilisés.

Tableau 2. Composition du mélange des peptides du « *peptide calibration standard I* » utilisé pour la calibration du MALDI TOF sur la gamme de masse.

Peptide	Masse monoistopique $[M+H]^+$	Masse moyenne $[M+H]^+$
Bradykinin 1-7	757,3992	757,86
Angiotensin II	1046,5418	1047,19
Angiotensin I	1296,6848	1297,49
Substance P	1347,7354	1348,64
Bombesin	1619,8223	1620,86
ACTH clip 1-17	2093,0862	2094,43
ACTH clip 18-39	2465,1983	2466,68
Somatostatin 28	3147,471	3149,57

La solution de calibration (0,5 µl) et les échantillons d'intérêt (1 µL) sont co-cristallisés sur une cible MALDI en acier poli (Polishsteel™ 384 MTP, Bruker Daltonics) avec 1 µL d'une solution de matrice d'HCCA préparée à 10 mg/mL dans un mélange ACN/eau ultra-pure/TFA (50/49,9/0,1 ; (v/v/v)) et sont laissés à température ambiante jusqu'au séchage complet.

Les mesures en masse, pour la calibration comme pour l'analyse des échantillons d'intérêt, se font en mode reflectron positif sur une gamme de masse allant de 700 à 3 500 de rapport masse/charge (m/z) selon la méthode constructeur « RP_700-3500.par ». Brièvement, l'acquisition est réalisée avec les paramètres de voltage de la source d'ions 1 et 2 de, respectivement 19 kV et 16,7 kV et des paramètres de reflectron 1 et 2 de, respectivement 21 kV et 9,5 kV pour une tension de lentille de 8 kV. L'extraction pulsée est quant à elle fixée à 120 ns. La plage d'intensité laser est comprise entre 50 et 80% de la puissance laser maximale pour une fréquence de tirs laser de 1 000 Hz. En moyenne, les signaux de masse issus de 1 000 tirs laser sont sommés pour constituer le spectre de masse de l'échantillon.

9.2. Retraitement informatique des spectres de masse MALDI-TOF

Le retraitement informatique des spectres de masse est réalisé à l'aide du logiciel FlexAnalysis 3.4 (Bruker Daltonics). Une diminution de la ligne de base, un lissage selon l'algorithme Savitzky-Golay par cycle de 2 et une recherche des masses suivant une intensité minimale de 100 et un signal sur bruit de 5 sont appliqués.

9.3. Désorption ionisation laser assistée par matrice (MALDI) - résonance cyclotronique d'ions à transformer de Fourier (FT-ICR, *Fourier transform ion cyclotron resonance*)

Les analyses en masse par spectrométrie de masse sur analyseur FT-ICR ont été réalisées à l'aide d'un spectromètre de masse Bruker 9,4 Tesla SolariX équipé d'une double source ESI/MALDI d'ions dont la dernière source utilise un laser Smartbeam™ II (Bruker Daltonics). Avant chaque analyse, l'appareil est calibré en mode positif en utilisant une solution d'acide phosphorique (Sigma-Aldrich) à 10 µg/mL dans 50 % d'ACN/50 % H_20 (v/v). Les échantillons purs et/ou dessalés (1 µL) sont co-cristallisés, comme pour les analyses MALDI-TOF-MS, sur une cible MALDI en acier poli (Bruker Daltonics) avec 1 µL de solution de matrice HCCA et séchés à température ambiante. Les spectres sont acquis en mode ions positifs à partir de 100 tirs laser. La fréquence de tirs laser était de 1 000 Hz et la puissance du laser était réglée à façon pour chaque échantillon. Les mesures en masse ont été réalisées soit sur une large gamme de m/z allant de 72,2 à 3500 avec une valeur de temps de vol de 2 ms et sur gamme plus restreinte de m/z dont la masse du centre a été fixée à m/z 1046 ± 13,9 et une valeur du temps de vol de 0,002 ms. En moyenne, les signaux de masse issus de 1 000 tirs laser sont sommés pour constituer le spectre de masse de l'échantillon.

9.4. Retraitement informatique des spectres de masse MALDI-FT-ICR

Les masses monoisotopiques des spectres de masse MALDI-FT-ICR ont été assignées à l'aide du logiciel DataAnalysis 4.0. (Bruker Daltonics) selon les paramètres par défaut de l'algorithme de recherche de masse FTMS.

9.5. Electrospray-quadruple – temps de vol (ESI-Q-TOF)

Les analyses ESI-Q-TOF des échantillons sont réalisées sur un spectromètre de masse SYNAPT-G2-Si-IMS (Waters). Avant chaque session d'analyse, le spectromètre de masse est calibré à l'aide d'une solution de formiate de sodium (Waters, Manchester, UK) contenant 0,1 mol/L d'hydroxyde de sodium (NaOH) et 200 µL d'acide formique à 10% dans une solution de 80% d'acétonitrile et 20% d'eau (v/v) sur une gamme de masse allant de 50 à 2000 Da. Les analyses ESI-Q-TOF en infusion (introduction directe des échantillons) sont conduites, en moyenne à un débit de 10 µl/min, à l'aide d'une seringue Hamilton de 250 µl (Dominique Dutscher, Brumath, France) et un pousse seringue (NE-1000, New EraPumpSyringeinc., Farmingdale, États-Unis).

Notre système de chromatographie Acquity Biocompatible est directement couplé à un spectromètre de masse SYNAPT-G2-Si-IMS. Les échantillons chromatographiés sur colonne de C18 sont par conséquent directement électronébulisés à partir de l'extrémité finale de la colonne à une tension de 3 kV, en utilisant un débit de gaz de désolvatation (N2) de 500 L/h, un débit de gaz nébuleux de 6,5 bars et des températures de désolvatation et de source, respectivement de 300°C et 150°C. Les mesures en masse sont réalisées en mode dépendant des données.

Les mesures MS sont réalisées à un pouvoir de résolution minimal de 20 000 (FWHM) sur la gamme de m/z de 300-3000. Les ions précurseurs détectés avec une intensité d'au minimum 1000 coups durant 200 ms sont sélectionnés, avec une fenêtre d'isolation de 3 unités de masse atomique (*atomic mass unit,* amu) pour fragmentation en dissociation induite par collision (*collision induced dissociation,* CID). Les rampes d'énergie de collision pour les composés de faible masse et pour les composés de masse élevée sont programmées, de respectivement 10 à 20 V et de 40 à 120 V. Les spectres MS/MS sont enregistrés sur la plage de 50 à 3000 m/z. Un temps d'injection maximal de 100 ms et un maximum de cinq ions les plus intenses a été utilisé pour générer les spectres CID-MS/MS.

9.6. Retraitement informatique des spectres de masse ESI-Q-TOF

La visualisation et l'extraction des données de masse est accessible depuis le logiciel Mass Lynx (version 4.1, Waters).

10. Retraitement mathématique des données de spectrométrie de masse

10.1. Calcul de la masse de Kendrick (KM); masse nominale de Kendrick (NKM) et défaut de masse de Kendrick (KMD)

La masse de Kendrick (KM, *Kendrick mass*) liée au motif CH_2, est calculée à partir de la formule moléculaire des composés référencés dans NORINE et des masses mesurées expérimentalement en utilisant l'Équation 1 suivante :

Équation 1. Calcul de la masse de Kendrick avec un motif CH_2.

$$KM = masse\ monoisotopique\ protoné\ de\ NORINE\ ou\ experimentalement\ mesuré \times \frac{masse\ nominale\ d'un\ CH_2}{masse\ exacte\ d'un\ CH_2}$$

La valeur KM est ensuite arrondie au nombre entier le plus proche (Équation 2) et définit la masse nominale de Kendrick (NKM, de l'anglais *nominal Kendrick mass*) :

Équation 2. Calcul de la masse nominale de Kendrick.

$$NKM = KM\ arrondi\ au\ nombre\ entier\ le\ plus\ proche$$

La NKM est ensuite soustraite de la masse de Kendrick pour obtenir le défaut de masse de Kendrick, Équation 3 (KMD, de l'anglais *Kendrick mass defect*).

Équation 3. Calcul du défaut de masse de Kendrick.

$$KMD = KNM - KM$$

10.2. Création de la carte de Kendrick relative aux composés de la base de données NORINE.

Les masses de toutes les NRPs compris dans la base de données NORINE sont calculées depuis la formule moléculaire en additionnant la masse de chaque atome donnée par l'union internationale de chimie pure et appliquée de 2016 (Meija *et al.*, 2016). Ensuite la masse exacte d'un proton (H^+) est ajoutée à chaque molécule, à savoir : 1, 00727646677, c'est à dire la masse d'un hydrogène auquel est soustraite la masse exacte d'un électron.

10.3. Correction de la réflexion *(aliasing*) de la représentation graphique de KMD=f(NKM).

Afin d'éviter le phénomène de réflexion (voir chapitre État de l'art) de la représentation graphique de KMD = f(NKM), la constante de 0,28, déterminée à partir de la carte de NORINE basée sur le calcul de Kendrick établi selon la manière développée ci-dessus est soustraite au calcul du KM pour donner une valeur corrigée appelée RKM, Équation 4 :

Équation 4. Calcul du défaut de masse régulier de Kendrick avec un motif CH2.

$$RKM = masse\ monoisotopique\ protoné\ de\ NORINE\ ou\ experimentalement\ mesuré \times \frac{masse\ nominale\ d'un\ CH_2}{masse\ exacte\ d'un\ CH_2} - 0{,}28$$

La valeur RKM est ensuite arrondie au nombre entier le plus proche (Équation 5) et définit la masse nominale de Kendrick (NKM) :

Équation 5. Calcul de la masse nominale de Kendrick corrigée.

$$NKM = RKM\ arrondi\ au\ nombre\ entier\ le\ plus\ proche$$

La NKM est ensuite soustraite de la masse de Kendrick (RKM) pour obtenir le défaut de masse de Kendrick régulier (RKMD), Équation 6.

Équation 6. Calcul du défaut de masse régulier de Kendrick.

$$RKMD = KNM - RKM$$

La représentation graphique sans réflexion spectrale de tous les composés de NORINE correspond alors à RKMD = f(NKM) que nous dénommerons pour la suite : le graphe 2D-RKMD/NKM ou encore la carte de NORINE. Cette représentation est générée à l'aide du logiciel Prism 7.

10.4. Augmentation de la résolution spectrale

Dans cette approche récemment développée (Fouquet & Sato, 2017c), la masse de Kendrick (KM liée au motif CH_2), est calculée à partir de la formule moléculaire des composés référencés dans NORINE et des masses mesurées expérimentalement en utilisant l'Équation 7 :

Équation 7. Calcul du défaut de masse de Kendrick augmentant la résolution.

$$KM = masse\ monoisotopique\ protoné\ de\ NORINE\ ou\ experimentalement\ mesuré \times \frac{13}{masse\ exacte\ d'un\ CH_2} - 0{,}28$$

10.5. Variation de RKMD (ΔRKMD), variation de NKM (ΔNKM) et maillage trigonométrique de Kendrick

Les variations (Δ) du RKMD et de la NKM entre deux points du graphe 2D RKMD/NKM 2D sont définies comme la différence entre le RKMD et les valeurs NKM de chaque point. L'addition ou la soustraction d'un atome ou d'un groupe d'atomes génèrent toujours la même valeur de variation de masse de Kendrick (ΔRKMD) et la même variation de masse nominale de Kendrick (ΔNKM). Par conséquent, à l'exception d'une variation de formule d'un CH_2entre deux composés qui sont par définition alignés sur une ligne horizontale, deux points proches (2 composés distincts) sont reliés entre eux par un triangle rectangle où la valeur de la ΔNKM forme un côté du triangle rectangle, la ΔRKMD forme l'autre côté et, finalement la ligne reliant les deux points forme l'hypoténuse de ce triangle rectangle. La valeur de l'hypoténuse est donc calculée à partir du théorème de Pythagore en utilisant l'Équation 8 suivante:

Équation 8. Calcul d'hypoténuse pour le défaut de masse.

$$Hypothenuse = \sqrt{(\Delta RKMD^2 + \Delta NKM^2)}$$

Parallèlement, les valeurs d'angle au point de référence sont calculées selon l'Équation 9 et Équation 10 ci-dessous :

Équation 9. Cosinus du Δ atomique des défauts de masse.

$$Cos\ (\Delta\ atome) = \frac{\Delta NKM\ (atome)}{Hypothenuse\ (atome)}$$

Puis

Équation 10. Angle θ du Δ atomique des défauts de masse.

$$\theta\ (\Delta\ atome) = \cos^{-1}(\Delta\ atome)$$

In fine, le maillage trigonométrique de Kendrick se définit, au niveau du graphe 2D-RKMD/NKM, comme l'ensemble des portions de droites (définissant un vecteur), définies par l'angle et la valeur de l'hypoténuse, reliant un point central à tout point de fermeture entourant ce point central.

11. Logiciels

11.1. Chemcalc®

Les formules moléculaires ont été déterminées à l'aide de l'outil en ligne "*molecular formula finder*" (http://www.chemcalc.org/mf_finder, Lausanne, Suisse) du site Chemcalc en utilisant les paramètres suivants: le nombre de carbones est compris entre 0 et 100, d'hydrogènes entre 0 et 100, d'azotes entre 0 et 20 puis d'oxygènes entre 0 et 20 (C0-100 ; H0-100 ; N0-20 ; O0-20), et en limitant les résultats par insaturation, les insaturations autorisées sont de 0 à 999 et erreur de masse de 0,1 Da.

11.2. Peaks Studio 9®

PEAKS Studio® (version 9) est un logiciel dédié à l'analyse protéomique. Ce logiciel analyse les données obtenues en LC MS/MS pour permettre l'identification de peptides et de protéines. Le logiciel intègre directement les données depuis l'extension « .raw ». Ensuite, les données sont retraitées selon plusieurs paramètres. L'erreur autorisée sur l'ion précurseur et sur les ions fragments est de 20 ppm. Aucune enzyme n'est utilisée pour la recherche de site de coupure. Les oxydations, les pertes d'ammonium, la recherche d'adduits sodium sont autorisées dans le volet modification post-traductionnelle. La base de

données Uniprot (TrEMBL) est utilisée pour l'interrogation et l'identification des protéines. Les données retraitées peuvent ensuite être exportées en fichier PDF ou XML.

11.3. Prism® 7

Les figures nécessitant un repère orthonormé ont été créées à l'aide du logiciel Prism® 7 (GraphPad software Inc, version 7.0, San Diego, CA, USA) en mode XY, en utilisant les paramètres par défaut et les entrées numéraires en X et Y pour tracer les fonctions souhaitées.

11.4. ChemDraw Ultra®

Les structures moléculaires semi-développées ont été dessinées en utilisant le logiciel ChemDraw Ultra® (version 12.0.2.1076, Cambridge, MA, États-Unis).

11.5. Molinspiration

La prédiction de l'hydrophobie des molécules sur la base de leur formule semi-développée est obtenue grâce à un outil en ligne sur le site : http://www.molinspiration.com/ depuis l'onglet « Calcul des propriétés moléculaires et prédiction de la bioactivité ».

11.6. Proteowizard

Pour rappel, Le logiciel ProteoWizard a été utilisé comme convertisseur de fichiers de données de masse issues de différents constructeurs en fichiers de sortie plus universels comme le format générique Mascot (*mascot generic format,* mgf), le format mzML ou encore le format mzXML.

11.7. Cyclobranch

La détermination de manière *de novo* des séquences pour les peptides non ribosomiques est menée à l'aide du logiciel Cyclobranch (http://ms.biomed.cas.cz/cyclobranch/docs/html/). Les fichiers sont convertis au format mzXML avant soumission. Celle-ci est réalisée en type de peptide cyclique, de charge 1 avec une tolérance d'erreur sur le précurseur de 3 ppm. La base de données des monomères utilisés est issue de NORINE. Le nombre maximum de blocks combinés lors de la fragmentation est de 5 pour le début, 3 pour le milieu et 5 pour la fin. La génération de permutation de monomères est autorisée. Les types d'ion choisis pour le spectre théorique sont les ions b et y. Les résultats sont exportables sous Excel dans un format « CSV ».

11.8. *iSNAP Fragmenter*

La fragmentation in silico des peptides ribosomiques et non ribosomiques est réalisée par le logiciel *iSNAPfragmenter* (https://magarveylab.ca/analogue/#!/fragmenter). La soumission de chaque molécule est réalisée depuis un fichier SMILES.

Le logiciel est utilisé avec plusieurs règles de fragmentation théorique autorisées ou non et un nombre de sites pouvant être clivés simultanément. Les paramètres utilisés par défaut sont i) clivage amide autorisé de 0 à 2, ii) le clivage ester autorisé de 0 à 1, iii) le clivage inversé d'ester, de thioéther et sucre non autorisé et iv) les pertes d'eau et d'ammonium autorisées de 0 à 10. Les fonctionnalités sont généralement adaptées en fonction de la molécule à fragmenter ; dans le cas de glycopeptides par exemple, la fonction de clivage de sucre est bien évidemment autorisée à hauteur du nombre de sucres présents. Le résultat est donné sous forme d'un tableau récapitulatif avec les données de m/z, de charge, de masse et le SMILES associées à chaque fragment.

Résultats

Inspiré des informations contenu dans les bases de données, cette série analytique commence par ***l'identification du microorganisme par une méthode de profilage phénotypique obtenu par spectrométrie de masse****. Dans cette partie nous utilisons la spectrométrie de masse MALDI TOF afin de permettre une identification simple et rapide d'un nombre important de souches.*

1. Identification de souches par MALDI TOF

L'identification des microorganismes par spectrométrie de masse MALDI-TOF est basée sur l'empreinte spectrale (*Main Spectral Projection,* MSP) de ces derniers. Les signaux de masse détectés composant le spectre sont majoritairement ceux correspondant aux rapports masse sur charge (m/z) des protéines du ribosome et dans une moindre mesure des protéines membranaires issues du métabolisme primaire du microorganisme. L'identification est basée sur la comparaison des empreintes spectrales (=profils en masse) (valeurs de m/z et intensité des signaux de masse) d'une souche bactérienne inconnue avec les empreintes spectrales, issues de souches bactériennes connues, enregistrés dans la banque de données de spectres (Figure 26).

La Figure 26 illustre les empreintes spectrales caractéristiques et différentes en fonction des souches. Certains signaux de masse sont communs dans les six spectres caractéristiques du genre, comme les ions de m/z 3582, 4430 ou encore 7167. Cependant, les empreintes comportent également des signaux spécifiques à chaque espèce voire même sous-espèce. Ce sont ces signaux de masse différents et communs qui permette l'identification des souches. La gamme de m/z utilisée pour l'identification MALDI-TOF va de 2 000 à 20 000 Da. Cette gamme de masse est cependant dans la pratique plus réduite (m/z de 2 000 à 12 000).

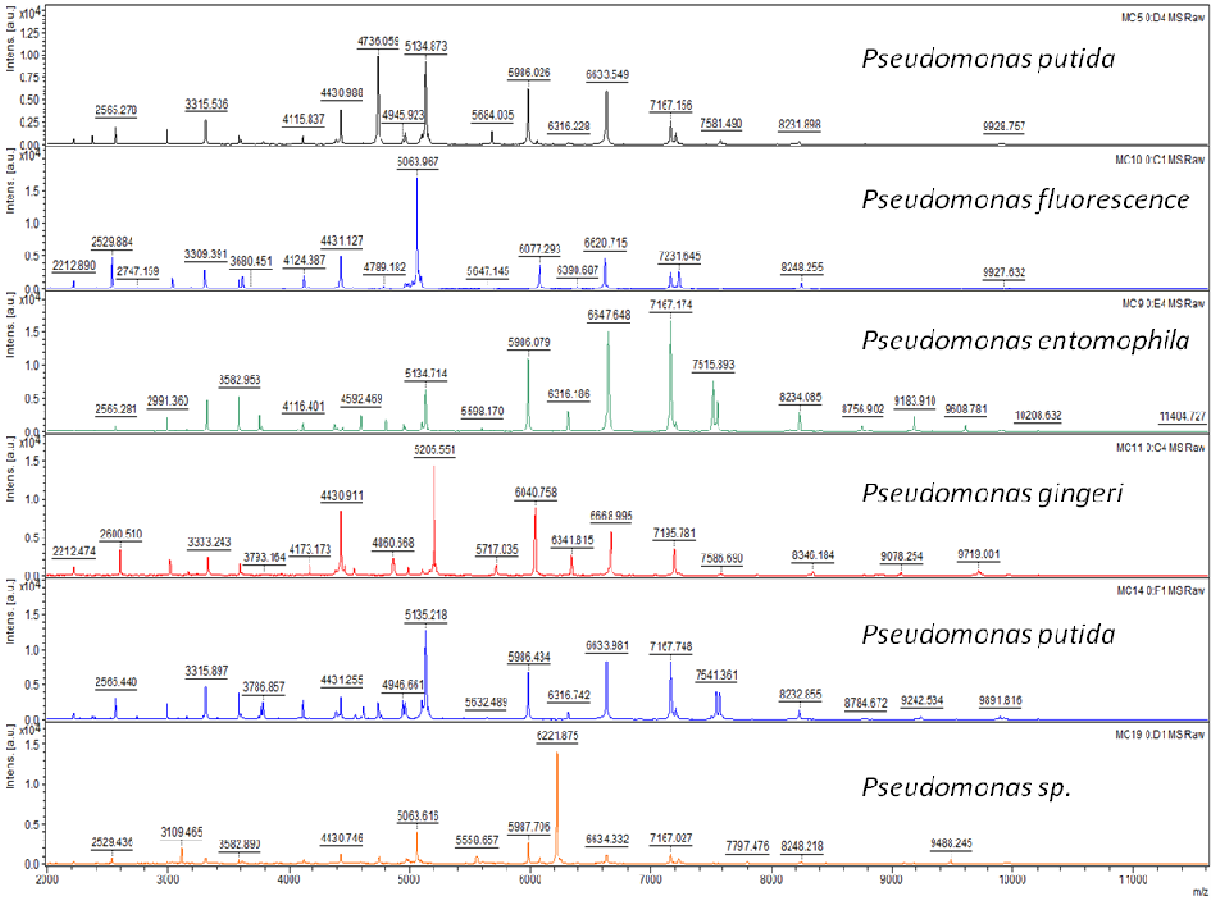


Figure 26. Empreintes spectrales de six souches du genre *Pseudomonas* dans la gamme de masse de m/z 2000-12000. Les spectres résultent de 1000 tirs laser accumulés pour obtenir un spectre de masse MALDI-TOF représentatif moyen.

La production de métabolites secondaires qui sont des molécules de faible masse moléculaire, étant en général une réponse au stress des microorganismes, elle est fortement influencée par les conditions environnementales ou de culture de ces derniers. Les souches sont donc cultivées sur ou dans plusieurs milieux de culture différents afin de favoriser la production de différents métabolites secondaires. Bien qu'il soit de nos jours parfaitement démontré et rapporté par la littérature que le milieu de culture (Valentine *et al.*, 2005) n'a que peu d'influence voire aucune sur l'identification des microorganismes par MALDI-TOF, nous avons néanmoins expérimentalement et systématiquement vérifié cet état de fait à partir de nos microorganismes et nos conditions de culture (Figure 27).

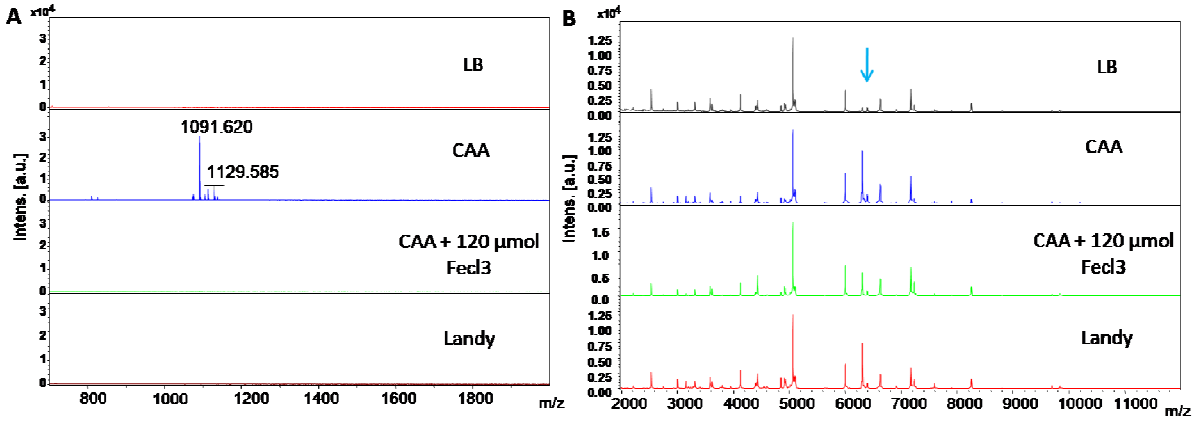


Figure 27. Spectres de masse MALDI-TOF obtenus dans différentes conditions de culture. A) Mode reflectron positif sur la gamme de masse de m/z 850-2000 et B) Mode linéaire positif sur la gamme de masse de m/z 2000-20000 pour la souche *Pseudomonas agarici LMG 2112T*.

Au cours de notre étude, nous utiliserons à plusieurs reprises le genre *Pseudomonas* connu pour être un microorganisme capable de moduler son métabolisme en fonction de son environnement notamment grâce à des mécanismes d'adaptation tels que la production de métabolites secondaires par synthèse non ribosomique.

Ainsi, les bactéries sont cultivées dans un milieu carencé en fer capable d'induire la production de sidérophores (milieu CAA) ou encore dans des conditions favorisant la production de lipopeptides (milieu LB ou milieu de Landy). Les souches sont cultivées pendant 24h à 30°C en boîte de Pétri sur le milieu LB, CAA, CAA additionné de fer (Témoin négatif) et le milieu de Landy. Une colonie isolée est prélevée et déposée sur une cible MALDI. Sur un même dépôt, une mesure en masse MS est réalisée en mode réflectron positifs la gamme de masse de m/z 850-2000 afin de détecter la présence éventuellement de molécules de faible masse moléculaire susceptibles d'être des NRPs puis dans un second temps en mode linéaire positif dans la gamme de masse de m/z 2000-12000 pour l'identification du microorganisme. La Figure 27 illustre les spectres obtenus sur la souche *Pseudomonas agarici* LMG 2112T.

Comme le montre clairement la Figure 27, aucune molécule de faible masse moléculaire n'est détectée dans la gamme de masse de m/z 850-2000 (Figure 27 A) et dans les conditions LB, CAA additionnée de chlorure de fer III et Landy. En revanche, un signal très intense avec un m/z de 1091,6 est présent pour la condition CAA. Il n'y a pas de différences

importantes au niveau des signaux de masse entre les quatre conditions de culture pour la gamme de masse d'identification, à savoir de m/z 2000-12000. Notons néanmoins, qu'un seul signal en masse (flèche bleue) situé à un m/z proche de 6300 présente une intensité variable notable selon les quatre conditions de culture.

Cette expérience illustre l'impact des conditions de culture sur la production des métabolites secondaires puisque les sidérophores ne sont détectés que lorsque les bactéries sont cultivées en absence de fer. Elle illustre également que la spectrométrie de masse MALDI-TOF peut être utilisée comme un outil rapide de contrôle de production de NRPs et de mesure précise de leur rapport m/z. De manière non surprenante, les conditions de culture n'impactent que faiblement les empreintes spectrales MALDI-TOF de la souche utilisée et n'ont donc que peu d'influence sur l'identification des microorganismes par MALDI-TOF. Des résultats similaires ont été obtenus avec les vingt souches de *Pseudomonas* utilisées au cours de mes travaux.

*Le chapitre suivant traite d'analyse permettant l'orientation vers différentes catégories moléculaire à travers des tests d'**activités biologiques**, ou d'analyses physico chimiques telles que l'**absorbance U.V.** et des méthodes séparatives.*

2. Criblage d'activités ayant une application dans le biocontrôle

Les NRPs ont des activités qui peuvent trouver des applications dans divers secteurs tels que la santé (beaucoup sont des antibiotiques) ou le secteur phytosanitaire, c'est-à-dire qu'ils peuvent être utilisés dans le biocontrôle. C'est notamment le cas des lipopeptides (Al-Ali *et al.*, 2017) ou encore des sidérophores (Esmaeel *et al.*, 2016). Les tests d'activités permettent de savoir si un microorganisme est capable de synthétiser ou non, soit un ou des lipopeptides, soit un ou des sidérophores, soit les deux. Ces deux classes de NRPs ayant une application potentielle en biocontrôle, des tests d'activités, complémentaires des mesures en masse précédentes, permettent de s'assurer de la production de NRPs actifs.

2.1. Le potentiel surfactant par test d'effondrement de la goutte

Il existe plusieurs moyens physiques (mesure d'arrachements, mesure de pression…) pour visualiser la présence de surfactant et par extension la production de lipopeptides, mais dans un contexte de dépistage, rien n'est plus simple, rapide et économique que de visualiser l'effondrement d'une goutte sur elle-même (Jain *et al.*, 1991). Une goutte de 10 µL d'échantillon (témoins et surnageants de culture) est placée sur une surface hydrophobe. La présence de molécules tensioactives engendre l'effondrement ou l'étalement visible à l'œil de la goutte. Cette observation est évaluée par rapport à des témoins présents (Figure 28).

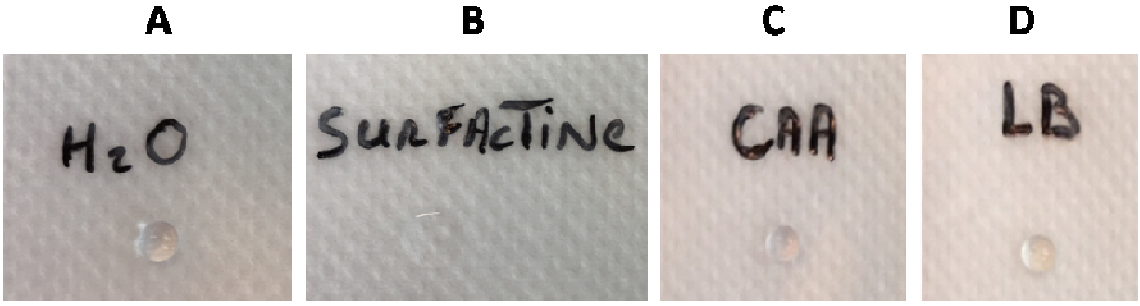


Figure 28. Illustration du test de l'effondrement de la goutte sur une goutte d'eau (H_2O, A), d'une goutte de surfactine à 1 mmol/L (B), d'une goutte de milieux de culture CAA (C) et LB (D) non ensemencés.

De l'eau (H_2O) est utilisée en tant que témoin négatif d'effondrement sur notre support hydrophobe, nous permettant d'avoir un standard de non effondrement. Ici dans la Figure 28 A, la goutte d'eau conserve une intégrité parfaite. Une goutte de surfactine à une concentration de 1 mol/L (reprise dans de l'eau) est utilisée en tant que témoin positif (Figure 28 B) afin d'observer un standard d'effondrement de l'intégrité d'une goutte. Les milieux de culture seuls sont toujours testés en parallèle des échantillons en qualité de témoins négatifs. Les milieux CAA et LB, respectivement présents sur la Figure 28 C et D conservent une bonne intégrité formant une demi-sphère parfaite pour les deux milieux.

2.2. Production de sidérophores

La production de sidérophores est évaluée en ajoutant le réactif de Schwyn et Neilands dans une goutte de surnageant analysé. La couleur orange traduit la présence d'un chélateur de fer. Les milieux de culture CAA et LB non métabolisés par un microorganisme et l'EDTA sont testés et servent de référence (Figure 29).

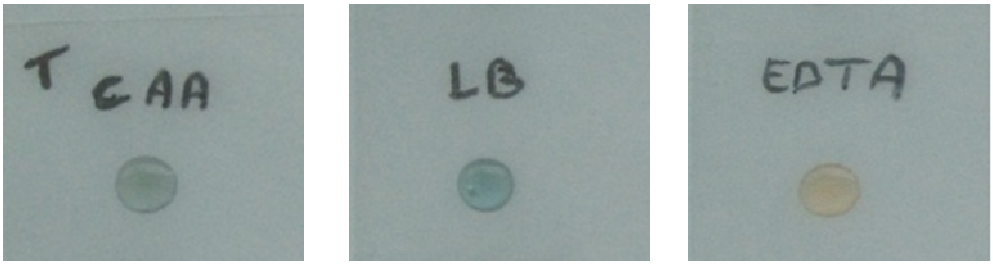


Figure 29. Test de la goutte de Schwyn et Neilands réalisé sur les milieux de culture non ensemencés CAA et LB utilisés en tant que témoins négatifs, et une solution d'EDTA utilisée en tant que témoin positif.

La réaction réalisée sur les milieux de culture CAA et LB non métabolisés n'a pas permis de mettre en évidence un changement colorimétrique du réactif. Ainsi, les milieux de culture seuls ne contiennent pas initialement de composés capables de chélater le fer. Les milieux de culture CAA et LB pourront être utilisés en tant que témoins négatifs. En revanche, un virage du réactif du bleu à l'orange est observé pour la solution d'EDTA. L'EDTA est un chélateur d'ions bivalent capable de chélater le fer. L'EDTA sera utilisé en tant que témoin positif de réaction.

3. Caractérisation moléculaire des surnageants bactériens

La recherche de métabolites spécifiquement produits par les souches en culture nécessite d'être capable de les détecter parmi l'ensemble des composés du milieu présents dans le milieu de culture ou issus du métabolisme des microorganismes suite à leur croissance.

3.1. Création d'une liste d'exclusion peptidique

Le profil peptidique du milieu avant et après culture des souches de *Pseudomonas* est réalisé par HPLC sur colonne C18 (Figure 30, A et B).

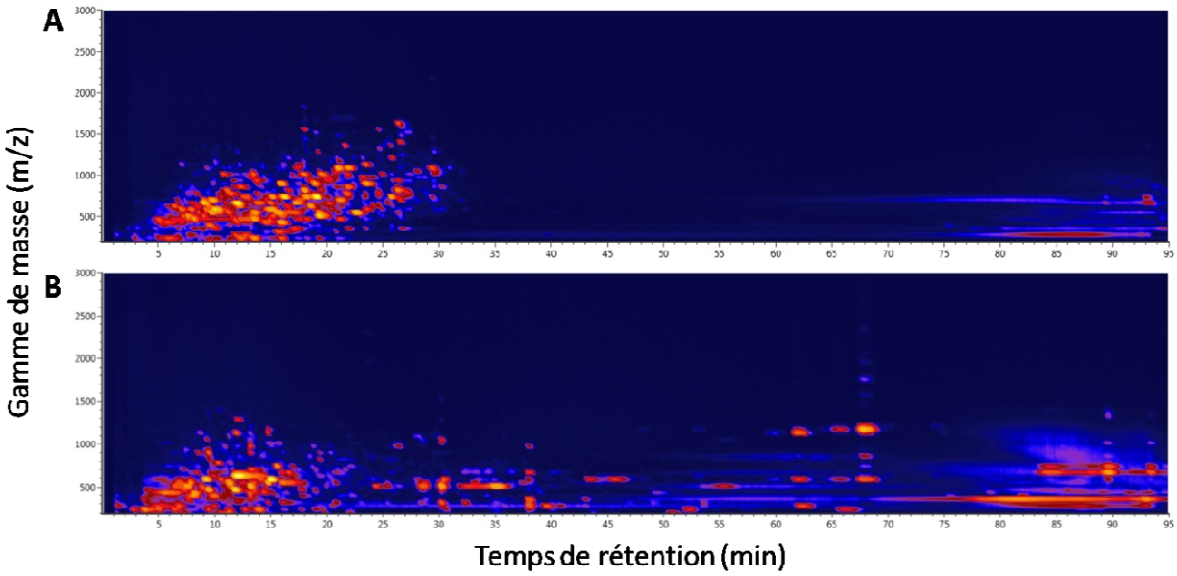


Figure 30. Chromatogramme du milieu LB (A) et du milieu LB métabolisé par la souche *Pseudomonas Sp. W2Aug9* (B) représentant la gamme de masse (m/z) en fonction du temps de rétention. L'intensité est représentée par un code couleur allant du bleu foncé pour des intensités nulles au rouge pour des intensités maximales obtenu après injection du milieu de culture ou du surnageant analysé en HPLC ESI Q-TOF.

La majorité des composés du milieu LB sont élués avant 35 min et forment un nuage dans la zone d'élution 2 % - 45 % d'ACN, sur une gamme de masse allant de m/z 200 à 1800 (Figure 30 A). Les molécules chromatographiées du milieu LB métabolisé parsèment davantage l'ensemble du chromatogramme sur sa totalité (Figure 30 B). Ainsi, une partie des composés formant le « nuage » présent sur la Figure 30 A est consommé par le microorganisme. Cela se traduit par une réduction des temps de rétention résultant de composés moins hydrophobes et de composés de masse moléculaire plus petite. Le nuage couvre alors une zone ne dépassant pas des temps de rétention de 25 minutes, pour des composés de masse moléculaire allant au maximum à m/z 1500 dans cet espace. Par ailleurs, il existe potentiellement des composés produits moins hydrophiles confondus dans la zone d'élution 0-25 minutes plus difficiles à discriminer. En revanche, certaines molécules sont éluées tout au long de la chromatographie. Ceci suggère que les molécules présentes après 35 minutes de rétention sont des composés produits par le microorganisme lors de sa culture. Les molécules présentes entre 35 et 95 minutes de temps de rétention sont des molécules présentant une hydrophobie plus élevée par rapport aux composés du milieu avec

pour certaines, des masses comprises entre des m/z de 1000 et 1500 (62,5; 65,5 et 67,5 minutes).

Comme le suggère le diagramme de Venn (Figure 31) réalisé à partir des peptides identifiés via la banque de données Uniprot (TrEMBL) à partir du milieu LB métabolisé (rouge) et du milieu LB non métabolisé (bleu), peu de peptides (n=10) sont communs aux deux conditions. Il apparaît également que le nombre de peptides est plus important dans le milieu LB seul (n=298) que dans le milieu LB métabolisé (60). Ce phénomène peut s'expliquer par la consommation des peptides du milieu LB par le microorganisme.

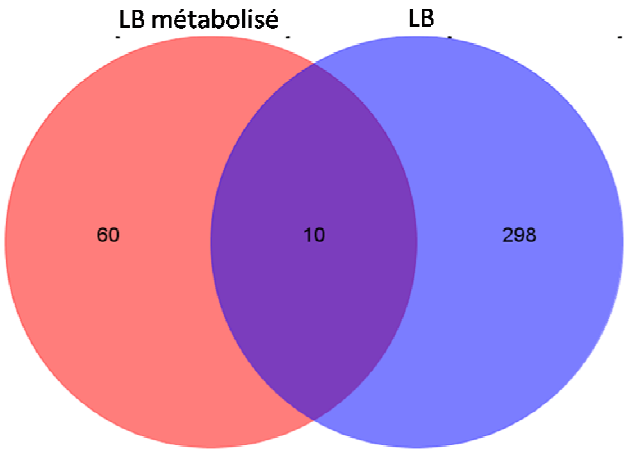


Figure 31. Diagramme de Venn regroupant les peptides identifié en MS à partir d'un surnageant de culture LB métabolisé par la souche *Pseudomonas Sp.* en rouge et les peptides du milieu de culture LB seul en bleu.

Notons que l'analyse est probablement biaisée par le mode d'identification des peptides. En effet le logiciel Peaks ne permet pas l'identification de NRPs (potentiellement produit dans le milieu LB métabolisé). Ce diagramme de Venn montre que la création d'une liste de peptides à exclure pour nous permettre de différencier les peptides du milieu et les peptides spécifiquement issus du métabolisme microbien n'est pas adapté à notre problématique (dans notre cas, dix peptides communs). La raison principale est sans doute liée au fait que les milieux sont empiriques et leur composition suivant les lots est très variable et que les microorganismes ont des métabolismes également variables.

4. Analyse de surnageant par approche directe

L'approche directe désigne l'analyse de mélanges complexes de peptides pour leur identification dans un échantillon. L'échantillon ne subit que des méthodes de préparation simples et rapides de purification voire même aucun traitement lorsqu'il s'agit de réaliser un dépistage.

4.1. Dessalage rapide sur Zip-Tip® phase C18 et graphite poreux (GPC)

Cette approche utilisant des microparticules fonctionnalisées et pacquées dans une pointe de cône formant une colonne permet une purification et un enrichissement rapides. Cette purification dépend de la fonctionnalisation des particules. Tout comme en HPLC, il n'existe pas de phase chromatographique « miracle » permettant la purification en une seule étape de tous les métabolites. Par conséquent nous utiliserons des colonnes C18 pour les composés hydrophobes comme les lipopeptides et des colonnes de graphite poreux (*graphitized porous carbon, GPC*) pour les sidérophores. Cette séparation différentielle permet non seulement le dessalage de l'échantillon, mais également de mettre en évidence le caractère hydrophobe ou polaire d'une molécule.

La Figure 32 illustre l'utilisation d'une de ces colonnes sur le surnageant de culture de la souche *Pseudomonas putida W15Oct28*, ici la détection par MALDI-TOF MS permet de voir des composés non-retenus élués au volume mort en A (spectre noir), des composés non-retenus élués par lavage en B (spectre vert) et des composés retenus et élués en solvant organique en C (spectre rouge, Figure 32).

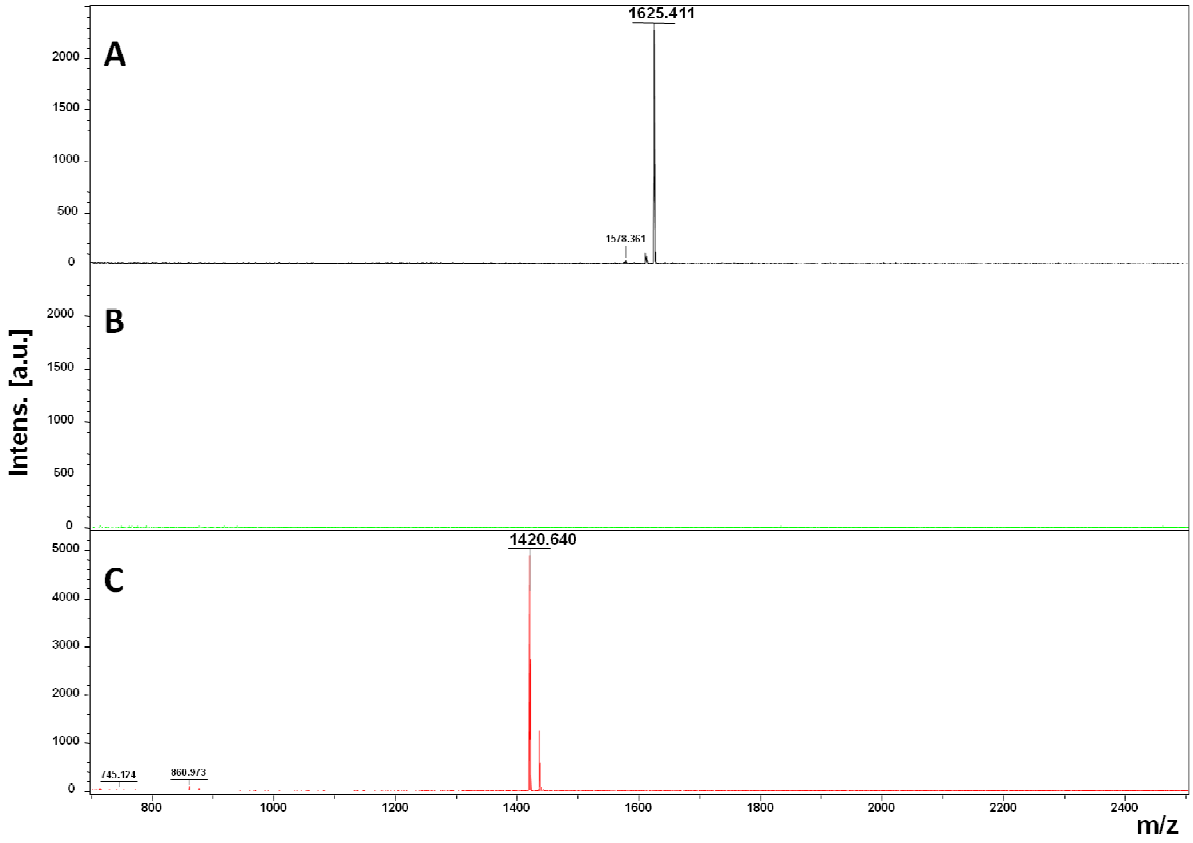


Figure 32. Spectres MALDI-TOF MS en réflectron positif de 1000 tirs laser dans la gamme de masse de m/z 850-2450 de l'analyse de la préparation sur ZipTip® C18 du surnageant de culture de la souche *Pseudomonas putida W15Oct28*, A) spectre du volume mort B) spectre du lavage et C) spectre de l'élution.

Très peu de signaux sont présent dans la gamme de masse m/z 900-2400. La majorité des composés des milieux de culture ont une taille inférieure à 900 Da et par conséquent ne sont pas visibles dans la Figure 32. L'analyse des composés non retenu (spectre noir, Figure 32 A), présente une masse à un m/z de 1625,4 pour une intensité de 2500. L'analyse de la fraction permettant le lavage de la phase stationnaire ne semble pas contenir de composés détectables dans la gamme de masse de m/z 850-2500. Enfin, l'analyse de la fraction éluée permet la détection d'une molécule à m/z de 1420,6 et d'une intensité de 5000.

4.2. Analyse par introduction d'un échantillon non prépurifié dans le spectromètre de masse

Certaines préparations d'échantillons comme l'extraction (liquide-liquide, liquide-solide (Agar)) ou encore la précipitation avec différents solvants permettent d'enrichir d'un facteur non négligeable certains composés en fonction des propriétés physico-chimiques des molécules et du solvant d'extraction utilisé (Figure 33).

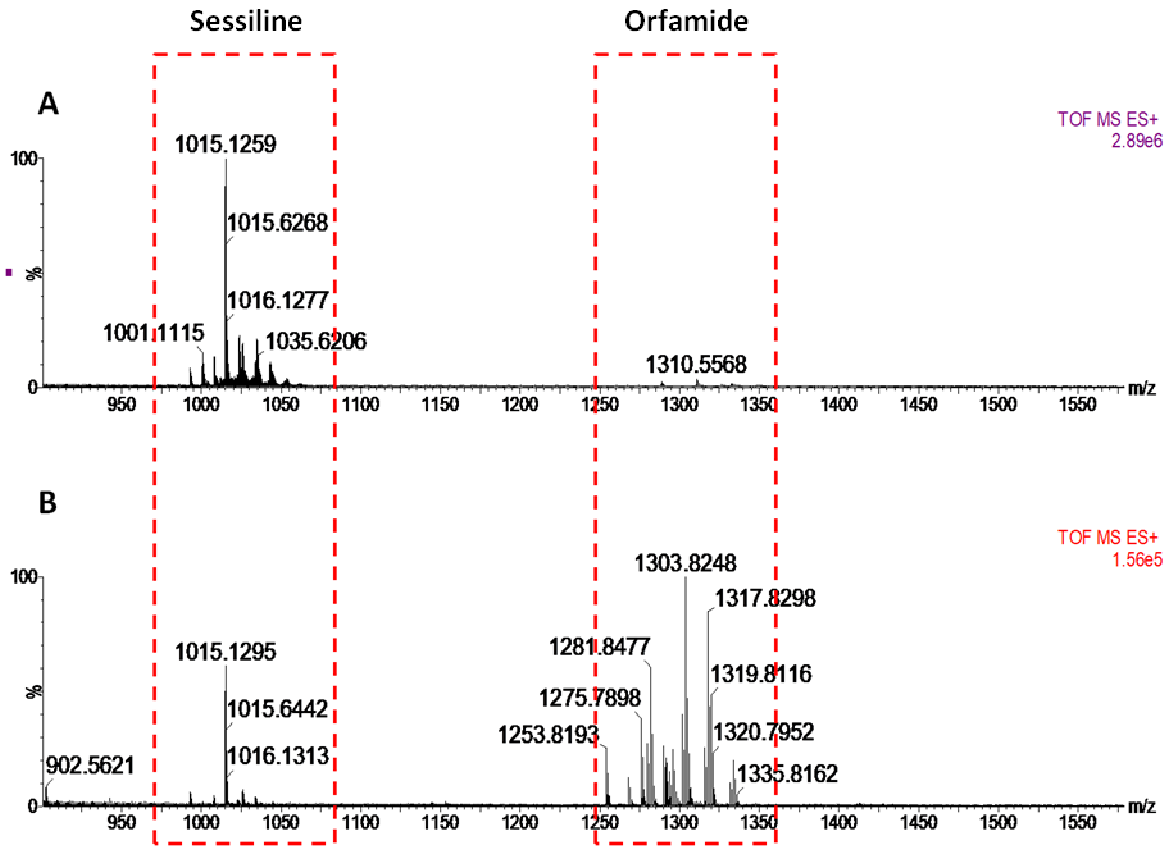


Figure 33. Spectres ESI QTOF MS du surnageant de la souche *Pseudomonas protegens Pf5* après précipitation acide en A et sans précipitation acide mais dilué 10 fois en B.

La souche de *Pseudomonas* protegens *Pf5* est connue pour produire deux familles de lipopeptides, les sessilines et les orfamides (Ma *et al.*, 2016). Après une précipitation acide (acide chlorhydrique 6M) réalisée à partir du surnageant de cette souche, le précipité est repris dans un volume de méthanol. Comme le montre la Figure 33, l'analyse ESI-Q-TOF-MS en mode positif, révèle la présence d'un massif intense de signaux de masse centré sur un m/z d'environ 1015 et un second, d'intensité très faible aux environs de m/z 1310. La famille des sessilines sont des molécules de plus faibles masses moléculaires que les orfamides. Des

molécules de la famille des sessilines semblent donc être détectées sur le spectre de masse ESI-Q-TOF dans la gamme de masse de m/z 1000-1050 (Figure 33, A). En revanche, l'analyse ESI-Q-TOF-MS dans les mêmes conditions de ce même surnageant dilué dix fois révèle des signaux de masse pouvant correspondre à des molécules des deux familles à la fois puisque le premier (zone des m/z 975 à 1050) et le second (des m/z 1250 à 1350) massifs de signaux de masse sont clairement détectés (Figure 33, B).

Cette expérience montre l'importance de la préparation de l'échantillon et bien souvent le biais potentiellement introduit par celle-ci. Bien que les deux familles fassent partie de la même classe moléculaire des lipopeptides, seules les sessilines ont été enrichies lors de la préparation de l'échantillon aux dépends de la famille des orfamides. Malgré tout, lors d'une analyse en infusion directe l'effet suppression d'ion causé par les petits composés et une quantité de sels importante présents dans le milieu de culture permet une moins bonne ionisation de certaines molécules de plus haut poids moléculaire. De plus la majorité des ions détectés sont présents sous forme d'adduits protonés mais également sous forme d'adduits sodium et potassium. La dilution permet de réduire bien évidement la quantité des composés d'intérêts mais permet de réduire également la quantité de sel et par conséquent l'effet de suppression d'ions.

Afin de valoriser au mieux les données de spectrométrie de masse (MS) souvent peu exploitées à l'instar et au profit de la fragmentation par spectrométrie de masse (MS/MS) nous avons choisi de combiner pour la première fois une méthode de calcul itérative inspirée des travaux d'un chimiste anglais dans les années 60, Edward Kendrick et de la seule base de données dédiée aux peptides non ribosomiques NORINE, historiquement créée par une collaboration entre l'Institut Charles Viollette (ICV) et le laboratoire d'informatique fondamentale de Lille (LIFL) devenu CRIStAL aujourd'hui. Cette combinaison nous permet de déterminer la composition élémentaire d'un peptide non ribosomique à partir des données de spectrométrie de masse haute résolution et d'un maillage vectoriel.

5. Création d'un kit de calibration pour l'analyse de composés NRPs

5.1. Le NRPmix pour la calibration MS mais surtout MS/MS

La masse moléculaire d'un composé est la première information récoltée lors d'une analyse par spectrométrie de masse. La précision en masse mesurée dépend intrinsèquement du fonctionnement physique de l'analyseur (voir section 3.3), mais également de la calibration de ce dernier. Afin d'obtenir des données les plus exactes possibles, nous avons réfléchi au choix d'un ensemble de molécules de type NRPs connues qui nous permettrait d'harmoniser la calibration des spectromètres de masse d'une part mais surtout d'harmoniser les réglages des spectromètres de masse de configuration et géométrie différentes en vue de l'analyse de NRPs et ce afin d'obtenir des spectres de fragmentation MS/MS similaires à un spectre de référence. Ce kit permettra de tendre vers une harmonisation mondiale de mesures MS/MS dans le but de créer une base de données homogène de spectres MS/MS de NRPs.

J'ai alors choisi plusieurs NRPs de poids moléculaires différents afin de couvrir une gamme de masse allant de 500 à 2000 Da. Pour cela, j'ai réalisé la courbe du nombre de NRPs en fonction de la masse moléculaire à partir des données de la base de données NORINE. Cette courbe est rapportée dans la Figure 34.

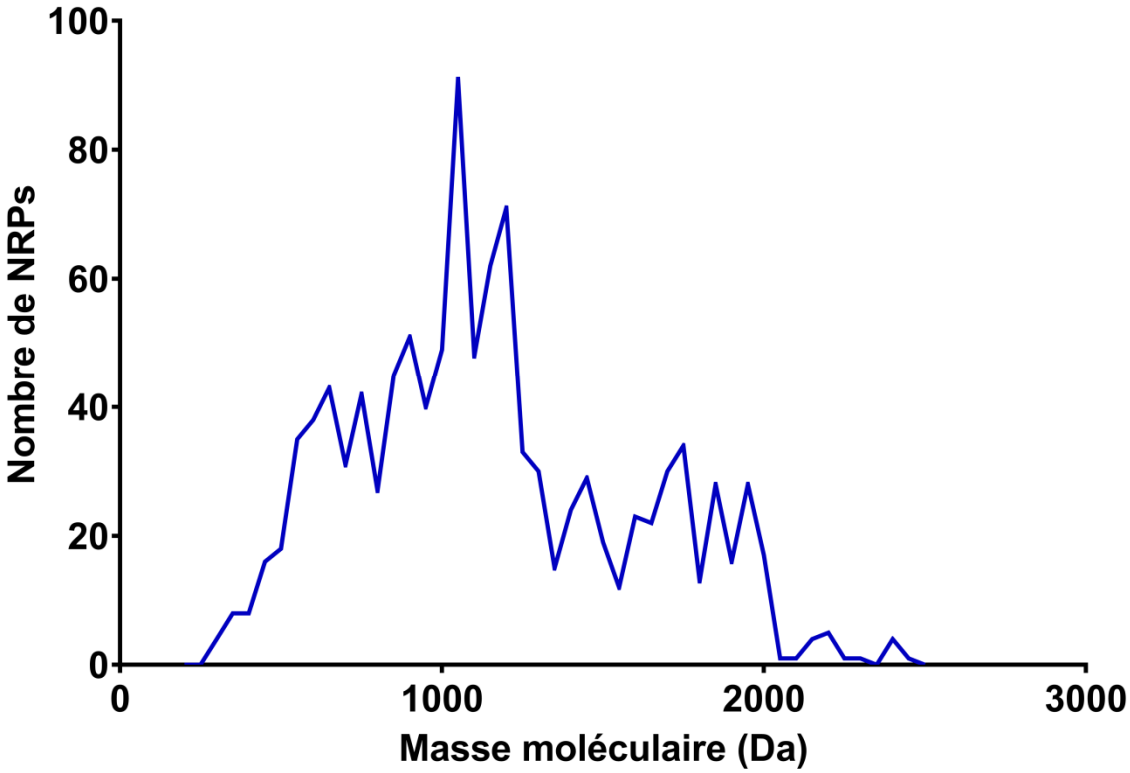


Figure 34. Nombre de NRPs (n=1187) en fonction de la masse moléculaire. Informations extraites de la base de données NORINE.

Les NRPs présents dans la base de données NORINE couvrent une gamme de masse allant de 300 à 3000 Da. La majorité des NRPs ont une masse moléculaire comprise entre 900 et 1300 Da avec un nombre maximum de NRPs entre 900 et 1050 Da. Au-delà de 1300 Da et jusqu'à 2000 Da, le nombre moyen de NRPs est de 30. Ainsi six molécules ont été judicieusement choisies pour couvrir cette gamme de masse, ce mélange est appelé le NRPmix (Figure 35).

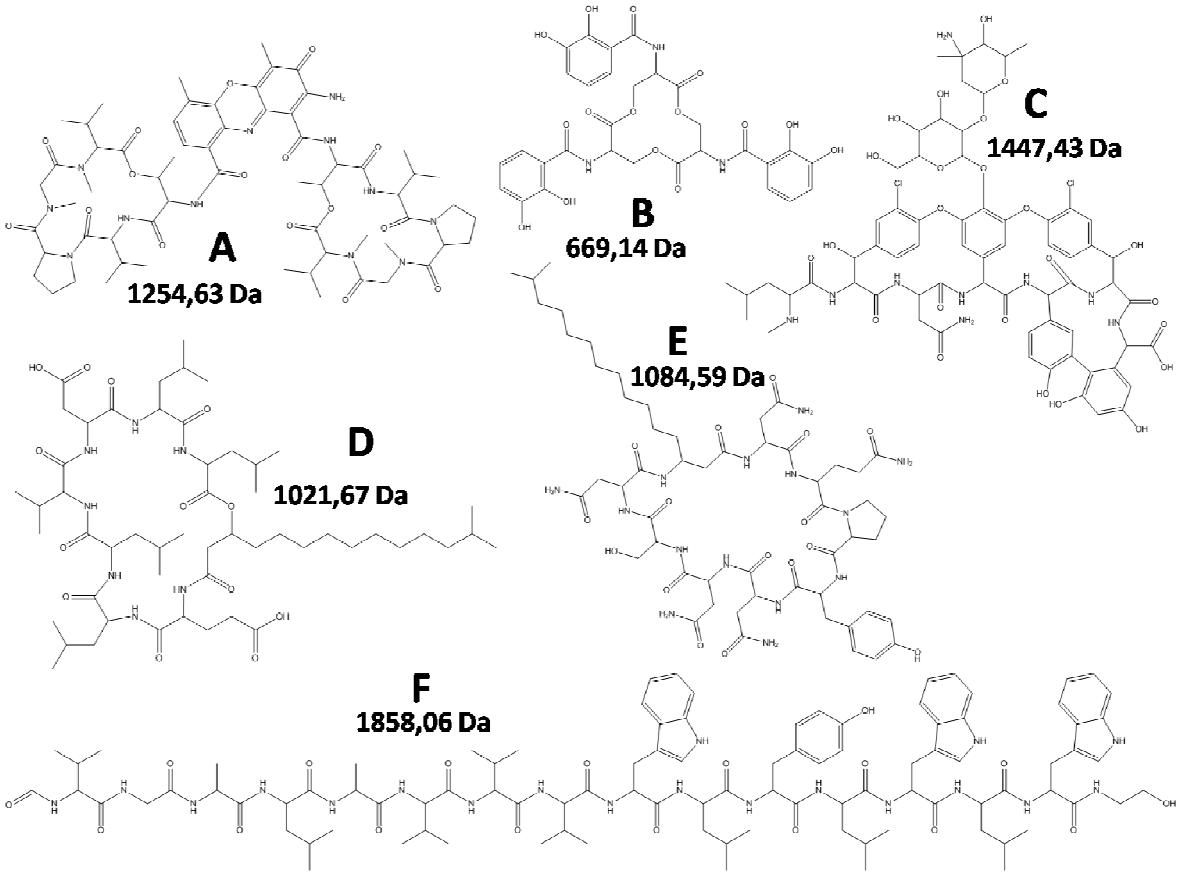


Figure 35. Structure des molécules présentent dans le NRPmix : (A) actinomycine D, (B) entérobactine, (C) vancomycine, (D) surfactine, (E) mycosubtiline, (F) gramicidine.

Le NRPmix est composé d'un chromopeptide, l'actinomycine D (**A**) qui possède une structure particulière composée de deux cycliques identiques reliés par un corps lui-même cyclique. Le second composé, l'entérobactine (**B**), est le plus petit, mais il n'en reste pas moins de structure complexe. L'entérobactine est le sidérophore le plus puissant connu sur terre. C'est un chélateur d'ions Fe^{3+} avec une affinité de K= 10^{52} M^{-1}. L'entérobactine est un NRP qui possède une structure comparable à un palindrome moléculaire tripartite. La troisième molécule (**C**) est la vancomycine qui est un glycopeptide possédant une activité antibiotique dirigée contre les bactéries à Gram positif. Notre intérêt est multiple (notamment pour la fragmentation) pour cette structure à l'image des cycles qui la composent et de la partie glucidique. Ensuite, la quatrième et la cinquième molécule (**D** et **E**) sont deux lipopeptides cycliques, respectivement la surfactine et la mycosubtiline, composés chacun de deux à trois variants (par le cycle peptidique ou par la chaîne aliphatique). Enfin la sixième et dernière molécule (**F**), la gramicidine, est un peptide linéaire uniquement composé d'acides aminés protéogéniques.

Les molécules ont également été choisies selon leur structure (linéaire, cyclique, multicyclique, branché), leur disponibilité sur le marché, leur pureté disponible, mais aussi selon leur prix (Tableau 3).

Tableau 3. Récapitulatif de la composition du NRPmix mentionnant le prix en euros par mg, la classe, le type, la formule moléculaire et le m/z théorique et expérimental.

Molécule	Prix* (€/mg)	Classe	Type	Formule moléculaire	Masse monoisotopique	Théorique		
						z	m	[M+Na]+
Actinomycin	40,00	Chromopeptide	Double cyclique	C62H86N12O16	1254,62847	1	1255,63575	1277,61824
						2	628,32151	
Enterobactin	246,00	Siderophore	Cyclique / Branché	C30H27N3O15	669,14422	1	670,15149	692,13399
Gramicidin	1,72	Peptide	Lineaire	C97H139N19O18	1858,05455	1	1859,06183	1881,04432
						2	930,03455	
Vancomycin	1,22	Glycopeptide	Multicyclique	C66H75Cl2N9O24	1447,43020	1	1448,43748	1470,41997
						2	724,72238	
Surfactin	25,4	Lipopeptide	Cyclique	C51H89N7O13	1007,65184	1	1008,65911	1030,64161
						2	504,83319	
Surfactin		Lipopeptide	Cyclique	C52H91N7O13	1021,66749	1	1022,67476	1044,65726
						2	511,84100	
Surfactin		Lipopeptide	Cyclique	C53H93N7O13	1035,68314	1	1036,69041	1058,67291
						2	518,84884	
Mycosubtilin		Lipopeptide	Cyclique	C51H88N12O14	1084,59170	1	1085,59897	1107,58146
						2	543,30312	
Mycosubtilin		Lipopeptide	Cyclique	C50H78N12O14	1070,57605	1	1071,58332	1093,56581
*variable en fonction du fournisseur, de la disponibilité ainsi que de la pureté						2	536,29530	

Les plus anciens d'entre eux, la vancomycine et la gramicidine, sont deux antibiotiques historiquement connus depuis plus de 70 ans et étudiés bien avant de connaître leur mode de synthèse. Aujourd'hui, il est possible d'obtenir ces molécules par synthèse chimique ce qui réduit considérablement leur coût. Les quatre autres NRPs sont toujours exclusivement produits par des microorganismes. Par conséquent, le coût de production reste, pour le moment, plus élevé que par synthèse chimique. La purification est également un facteur postproduction important et représentant une part importante du coût du produit fini (Walsh *et al.*, 1990) et explique les coûts plus importants notamment pour l'entérobactine. Les lipopeptides étant synthétisé par les microorganismes sous plusieurs variant (famille), le mélange final du NRPmix est composé de trois variant de surfactine (m/z 1008, 1022, 1036) et deux variants de mycosubtiline (m/z 1071 et 1085).

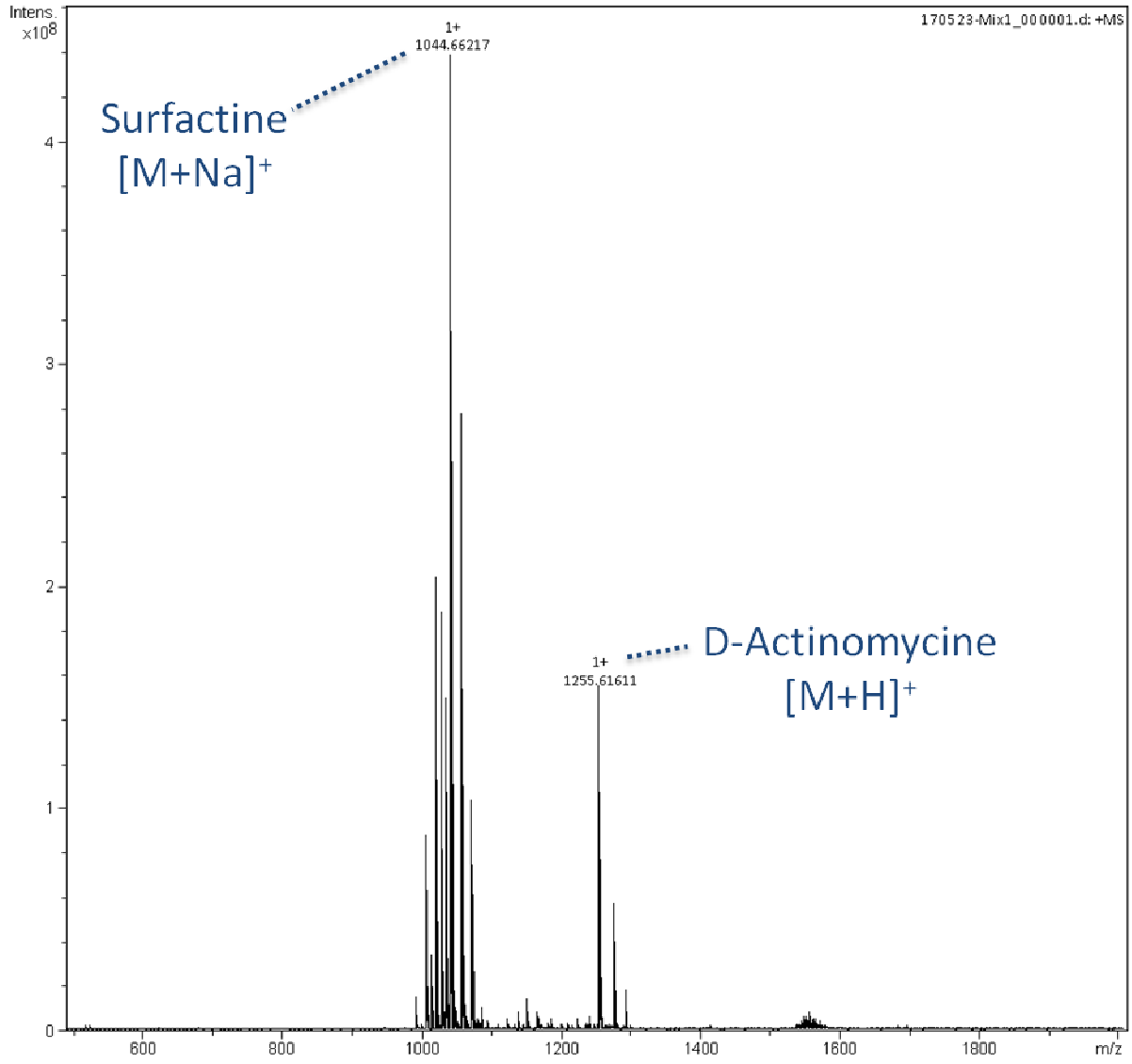


Figure 36. Spectre MALDI FT-ICR MS du NRPmix avec un mélange équimolaire dans la gamme de masse de m/z 500 à 2000.

Un premier mélange équimolaire des cinq molécules est réalisé puis analysé par MALDI FT-ICR (Figure 36). Le spectre de masse de la Figure 36 du mélange équimolaire du NRPmix présente des intensités différentes selon les molécules présentes dans le NRPmix. Ici l'intensité la plus importante est portée par les différents variants de surfactine (m/z de 1022,6747, 1036,6904, 1044,6622) et dans une moindre mesure par l'actinomycine D (m/z de 1255,6161).

Le MALDI FT-ICR n'est pas une méthode de mesure directement quantitative, en effet l'ionisation d'une molécule ne dépend pas uniquement de sa concentration. Certaines molécules comme les surfactants dont fait partie la surfactine sont de par leur nature des

molécules plus faciles à ioniser. Par conséquent dans un mélange équimolaire le signal de la surfactine prend le dessus sur les autres espèces moléculaires présentes et implique un effet de suppression ionique.

Afin d'illustrer nos propos, un nouveau mélange est réalisé sans la surfactine (Figure 37). L'absence de surfactine permet une meilleure ionisation des autres molécules. Le signal en masse de l'actinomycine D (m/z 1255,63655) est à présent le plus intense (3,8.10^8 intensité arbitraire, I.A.), puis vient l'entérobactine à un m/z de 670,15198 avec une intensité à 1,6.10^8 I.A.. Ensuite on retrouve la mycosubtiline à un m/z de 1085,59845 pour une intensité de 0,6.10^8 I.A.. La vancomycine est présente avec une intensité inférieure à 0,25.10^8 I.A. et pour une charge z de deux c'est-à-dire à 724,72278.

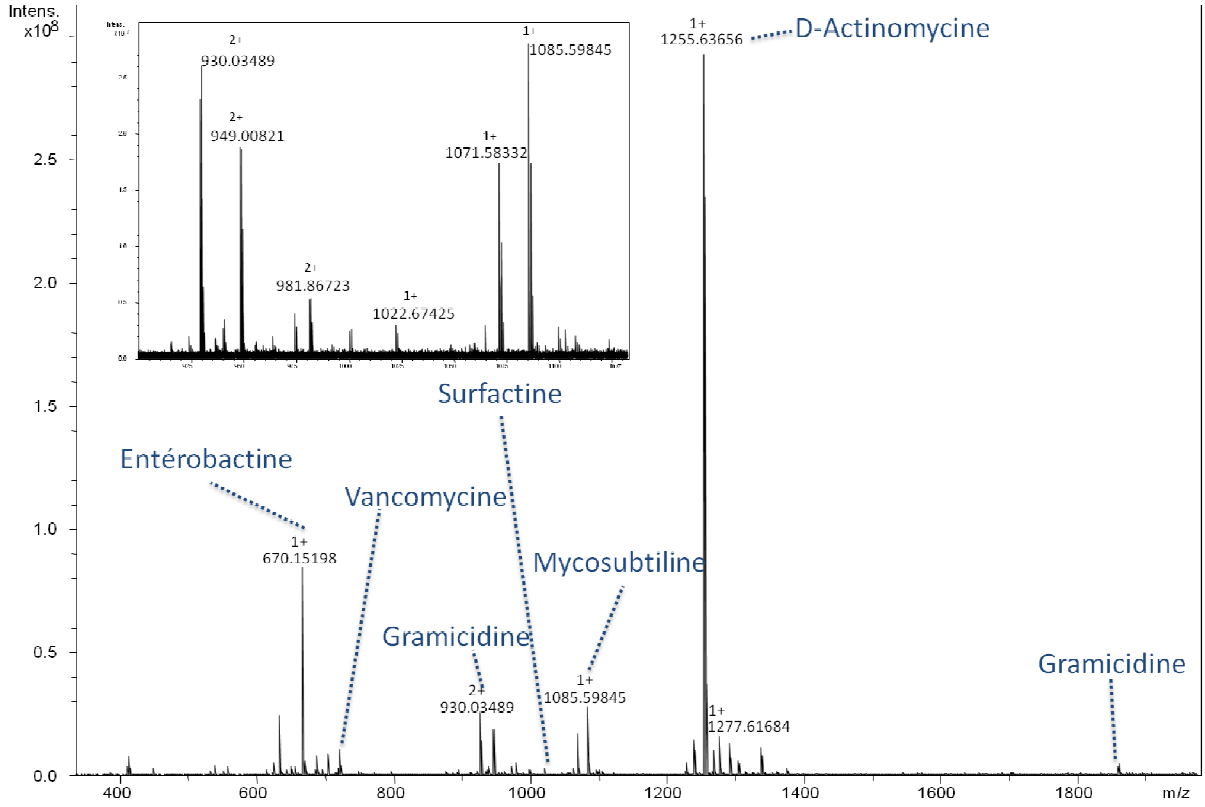


Figure 37. Spectre MALDI FT-ICR MS du mélange NRPmix à des concentrations équimolaires sans la surfactine.

Les lipopeptides sont des molécules plus facilement ionisables dans ses conditions d'analyse mais après plusieurs mélanges réalisés avec différentes concentrations, nous avons trouvé les proportions à respecter pour analyser toutes les molécules en même temps. Les molécules de vancomycine, mycosubtiline, entérobactine et actinomycine D

peuvent être analysées en concentration équimolaire. La surfactine peut être ajoutée en concentration moindre. Le mélange final du NRPmix est composé des cinq molécules, vancomycine, mycosubtiline, entérobactine, actinomycine D et surfactine en proportions respectives de 1/1/1/1/0,3. L'effet de suppression d'ions n'est pas observé avec la mycosubtiline (lipopeptide) mais un biais de concentration a été constaté par la suite car la pureté de départ n'était pas aussi importante que celle des autres molécules. Par conséquent, lorsque les mélanges équimolaires ont été réalisés, la mycosubtilline était en concentration moindre par rapport aux autres molécules.

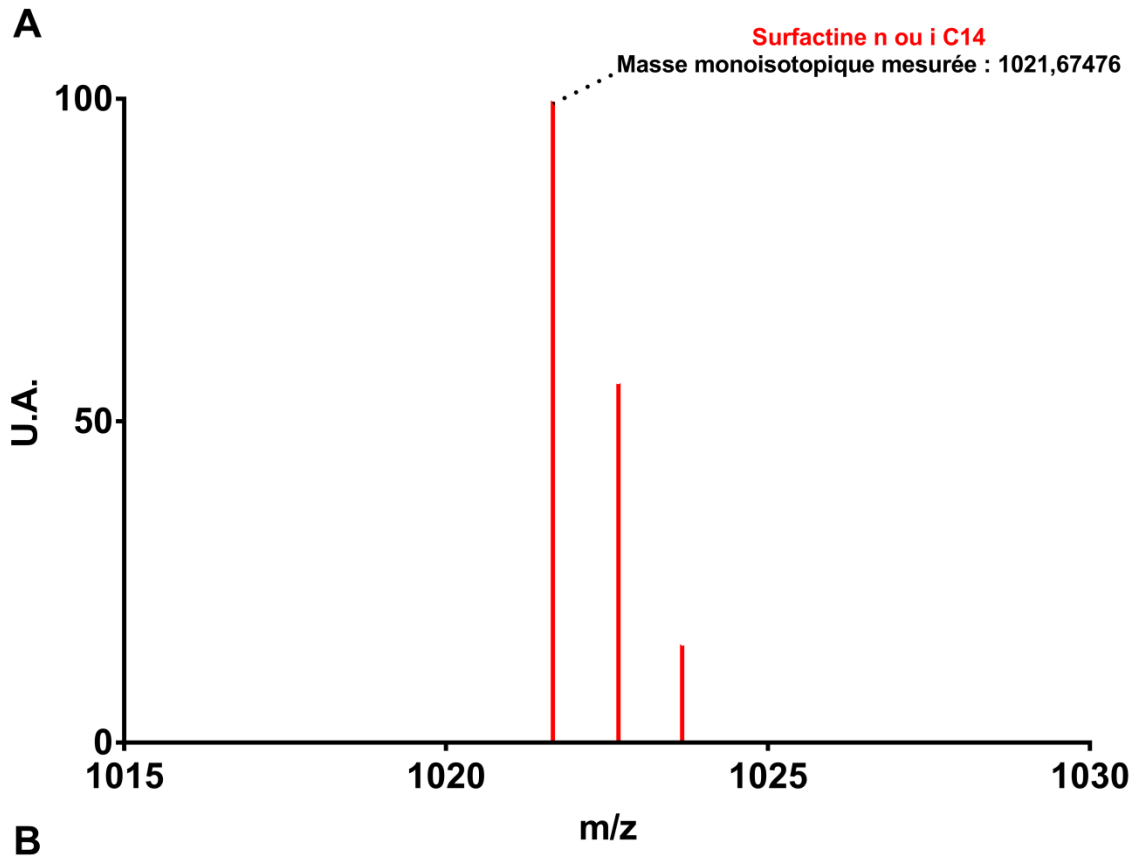
6. L'exploitation des données MS issues de l'analyse de NRPs en vue de déterminer leur formule moléculaire

6.1. Détermination de formule moléculaire à l'aide de générateur de formule brute

Après avoir calibré l'appareil et les spectres à l'aide du NRPmix et afin de réduire au maximum l'erreur de mesure, il est possible de déterminer la formule moléculaire du composé suivant plusieurs algorithmes de génération de formules moléculaires (voir section 3.6.1.1 de l'État de l'Art). Cette étape constitue la première information importante lors de l'élucidation et la caractérisation de structure moléculaire.

Comme l'illustre la Figure 38A, la masse monoisotopique théorique exacte des surfactines nC14 et iC14 est la même: 1021,667491263. Ces surfactines présentent également la même distribution isotopique théorique. A un tel rapport de m/z, la précision de mesure en spectrométrie de masse à haute résolution fournit une masse monoisotopique protonée ($[M+H]^+$) de m/z 1022,67476. Utilisant la masse monoisotopique correspondante (non chargée) (1021,66749) et une précision de masse inférieure à 1 ppm, un logiciel comme Chemcalc (Patiny & Borel, 2013) qui est dédié à la corrélation d'une masse monoisotopique et d'une formule moléculaire propose sept candidats moléculaires distincts (Figure 38B). Le tableau de la Figure 38B rapporte pour chaque formule moléculaire potentielle (candidate) sa masse monoisotopique théorique et l'erreur (exprimée en ppm) calculée depuis la masse monoisotopique mesurée.

A



B

#	Formule moléculaire	Masse monoisotopique	Δ (ppm)
1	C65H83N9O2	1021,666973	0,506
2	C66H89N2O7	1021,666978	0,501
3	C51H85N14O8	1021,667481	0,009
4	C52H91N7O13	1021,667486	0,004
5	C53H97O18	1021,667491	0,001
6	C37H87N19O14	1021,667989	0,488
7	C67H85N6O3	1021,668316	0,808

Figure 38. Représentation schématique du spectre de masse montrant la distribution isotopique des surfactines nC14 et iC14 en A. Tableau illustrant les 7 formules moléculaires obtenues en utilisant la fonction «trouver une formule moléculaire à partir d'une masse précise» du logiciel Chemcalc (http://www.chemcalc.org/) en B.

Si l'on exclut des informations de préparations d'échantillons ou encore de fermentation dans un milieu de culture favorisant une catégorie moléculaire en particulier, il est impossible à ce stade de faire un choix parmi les sept formules moléculaires proposées par un générateur de formule brute. D'autres informations peuvent dès lors s'avérer utiles pour réduire encore le nombre de formules moléculaires.

6.2. Utilisation de La distribution isotopique

La distribution isotopique permet d'obtenir plusieurs informations structurales. Pour reprendre notre exemple et les sept formules moléculaires issues du générateur de formules brutes, l'analyse de la distribution isotopique nous permet de réduire significativement le nombre de possibilités. En effet, la Figure 39 est une simulation des massifs isotopiques attendus issue des sept formules moléculaires précédemment obtenues à partir du générateur de formules moléculaires Chemcalc (Figure 39).

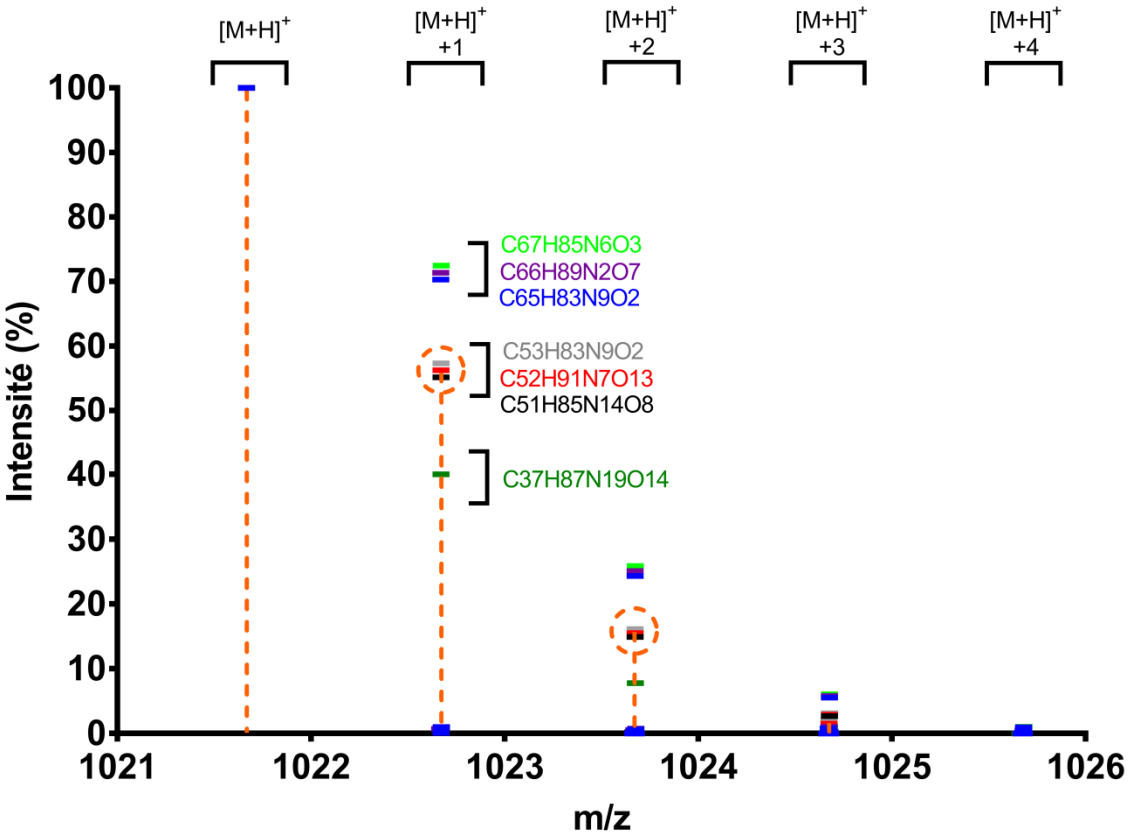


Figure 39. Représentation de la distribution isotopique théorique des sept formules moléculaires et la distribution isotopique mesurée en pointillé orange.

Les signaux sont normalisés sur l'espèce $[M+H]^+$ afin d'obtenir une intensité en pourcentage et une représentation relative de chaque isotope des 7 formules moléculaires. La Figure 39, présente les cinq premiers isotopes composant le massif isotopique de la surfactine (nC14 et iC14). Le premier signal $[M+H]^+$ ou signal monoisotopique est issu de la détection de molécules de surfactine nC14 et iC14 pour des atomes (^{12}C,^{1}H,^{14}N et ^{16}O) les plus stables et abondants. Le signal $[M+H]^+$+1 représente quant à lui, la détection d'une molécule de surfactine (nC14 et iC14) ayant intégré un isotope de carbone 13 (^{13}C, car plus

abondant que les autres isotope tels que ^{2}H, ^{15}N ou ^{17}O) et des atomes stables de ^{1}H,^{14}N et ^{16}O (Voir section 3.6.1.2). Ainsi, ce sont pour les signaux $[M+H]^{+}+1$ et $[M+H]^{+}+2$ que les différences de massif isotopique sont les plus remarquables. Comme l'illustre la Figure 39, pour les signaux de masse de ces isotopes, trois groupes sont distincts : le premier, le plus intense, est situé au-dessus du massif isotopique mesuré avec une intensité aux alentours de 70% soit environ 15 % plus intense pour les composés suivants : $C_{67}H_{85}N_6O_3$, $C_{66}H_{89}N_2O_7$ et $C_{65}H_{83}N_9O_2$. Un second groupe, plus proche du $[M+H]^{+}+$ 1 mesuré, est constitué de $C_{53}H_{83}N_9O_2$, $C_{52}H_{91}N_7O_{13}$ et $C_{51}H_{85}N_{14}O_8$ situé entre 50 et 60 % d'intensité. Enfin, la formule $C_{37}H_{87}N_{19}O_{14}$ proche de 40 % d'intensité est située à une valeur d'environ 15 % inférieure à la valeur mesurée.

Cette approche nous permet d'éliminer 4 des 7 formules moléculaires potentielles générées préalablement en considérant l'abondance isotopique théorique face aux valeurs mesurées. La mesure précise de l'abondance des isotopes n'est cependant pas toujours facile et dépend non seulement de la résolution de l'analyseur mais également du détecteur ainsi que du traitement du signal. La saturation d'intensité des signaux de masse est souvent la première cause de mesures erronées de l'abondance des massifs isotopiques.

6.3. Les défauts de masse Kendrick

Dans cette partie, nous démontrons, à partir des données HRMS de NRPs modèles disponibles dans le commerce, que l'approche de Kendrick en combinaison avec la base de données NORINE est un outil intelligent, facile à utiliser, rapide et utile pour l'attribution d'une masse moléculaire. Dans un premier temps, nous avons utilisé les formules moléculaires référencées dans la base de données NORINE pour calculer les masses monoisotopiques théoriques de tous les NRPs de NORINE. À partir de ces dernières et des calculs des valeurs de NKM et KMD, nous avons construit une carte de NORINE basée sur le calcul de Kendrick. Enfin, nous avons réalisé un preuve de concept en comparant les masses expérimentales mesurées avec précision issues de l'analyse FT-ICR-MS des NRPs disponibles dans le commerce sur cette carte afin d'identifier leur formule moléculaire démontrant que l'approche peut être appliquée pour identifier de nouveaux composés.

6.3.1. Attribution des formules moléculaires de pour les NRPs de NORINE

La base de données NORINE est composée de 1187 entrées (Juillet 2018). Chaque entrée contient des informations collectées manuellement (structure, activité, famille, organismes producteurs) à partir de la littérature scientifique et liées à un seul NRP. Cependant, les informations relatives à la masse moléculaire des composés référencés dans NORINE sous souvent soit absentes, soit erronées (uniquement 25% sont présentes et exactes).

Nous avons donc implémenté la base de données en formules moléculaires à partir des informations collectées dans d'autres bases de données comme PubChem (Kim *et al.*, 2016) ou encore la plateforme de données génomique « *Minimum Information about a Biosynthetic Gene cluster* » (MIBIG)(Epstein *et al.*, 2018). Nous avons également dû redessiner certaines molécules présentes dans la littérature pour déterminer la formule moléculaire grâce au logiciel Chemdraw®. Ainsi 72% des formules moléculaires manquantes ont été retrouvées ou recalculées. Ceci a permis d'accroitre à 97% le nombre d'entrées possédant une formule moléculaire.

Dès lors, les formules moléculaires de tous les NRPs ont été extraites de la base de données NORINE et utilisées pour calculer, en utilisant le logiciel en ligne Chemcalc, la masse monoisotopique théorique selon IUPAC de 2013 (Meija *et al.* 2016). Ces dernières sont maintenant incorporées dans NORINE. Les masses monoisotopiques théoriques protonée ont alors été utilisées pour créer la carte de NORINE basée sur le calcul de Kendrick, correspondant au tracé 2D KMD/NKM.

6.3.2. Le principe graphique expliqué d'un point de vue théorique

Les variants de surfactines [Ala4] iC15, [Val7] iC15, [Ala4] iC14 et [Ile7] iC15 servent ici de molécules modèles pour expliquer le principe et les avantages de l'approche par défaut de masse de Kendrick (KMD) (Figure 40).

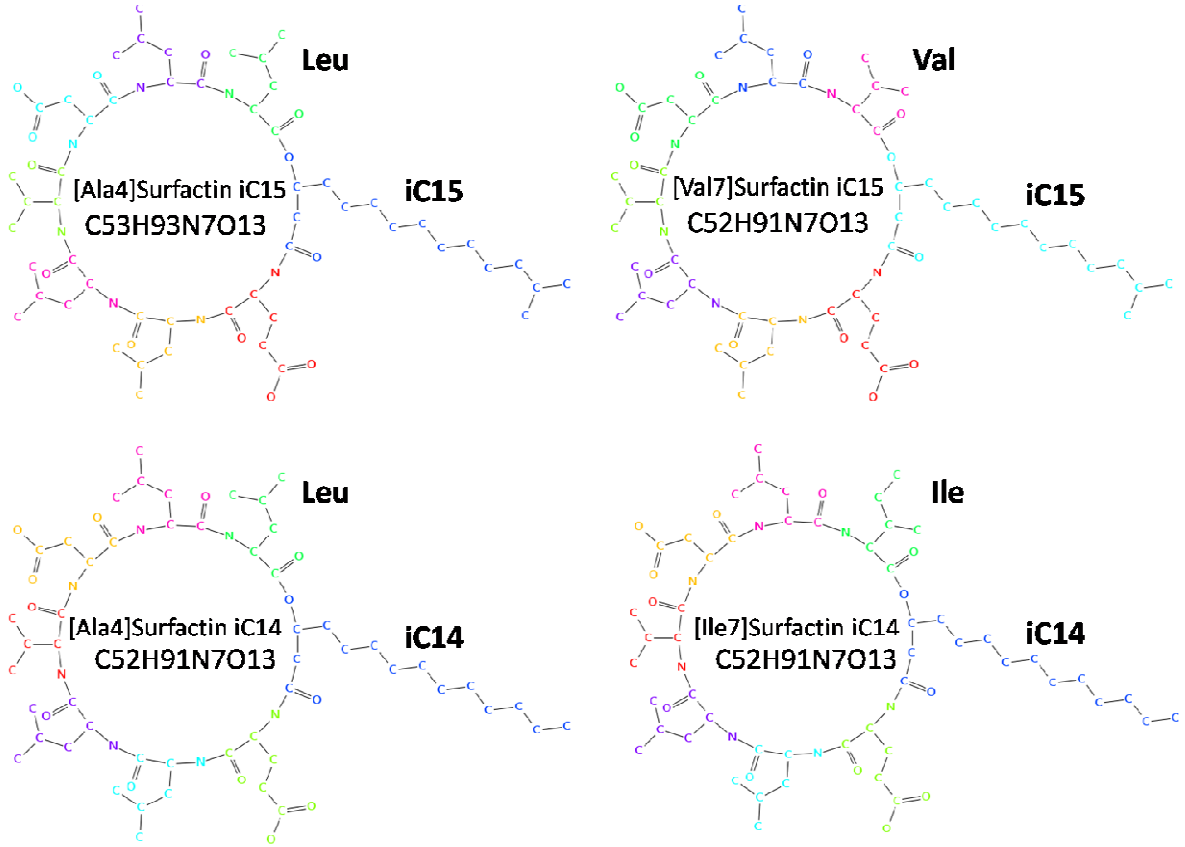


Figure 40. Formules semi-développées de variants de surfactine obtenues grâce au SMILES traités par le logiciel S2M (http://bioinfo.lifl.fr/norine/smiles2monomers.jsp).

Les trois variants [Val7] iC15, [Ala4] iC14 et [Ile7] iC14 ont la même formule moléculaire ($C_{52}H_{91}N_7O_{13}$) et par conséquent la même masse (1021,667491263 Da). La quatrième forme [Ala] iC15 a comme formule brute $C_{53}H_{93}N_7O_{13}$ pour une masse exacte de 1035,683136071 Da. Ces quatre formes de surfactine possèdent des différences structurales dans leur cycle peptidique ([Ala4] iC15 / [Val7] iC15) ou sur leur chaîne lipidique ([Ala4] iC15 / [Ala4] iC14). Le graphique 2D (KMD/NKM 2D-plot) représentant la valeur de défaut de masse de Kendrick (KMD) en fonction de la masse nominale de Kendrick (NKM) calculé pour une unité de base CH_2 des sept formules moléculaires possibles est représenté sur la Figure 40. Il montre que les sept points correspondant aux 7 variants précédemment identifiés à partir de la masse sont alignés verticalement sur la même valeur nominale (même NKM).

Sur cet alignement on retrouve tout d'abord deux premiers composés (points bleu et violet), respectivement $C_{65}H_{83}N_9O_2$ et $C_{66}H_{89}N_2O_7$, proches l'un de l'autre et de valeurs de KMD de 0,46767 et 0,46768; Ensuite, un triplet (points noir ($C_{51}H_{85}N_{14}O_8$), rouge

($C_{52}H_{91}N_7O_{13}$) et gris ($C_{53}H_{97}O_{18}$)) de composés de valeurs de KMD comprises entre 0,46716 et 0,46718; Enfin, les deux derniers composés (points vert foncé ($C_{37}H_{87}N_{19}O_{14}$) et vert clair ($C_{67}H_{85}N_6O_3$)) de valeurs de KMD de 0,46620 et 0,46680 (Figure 41).

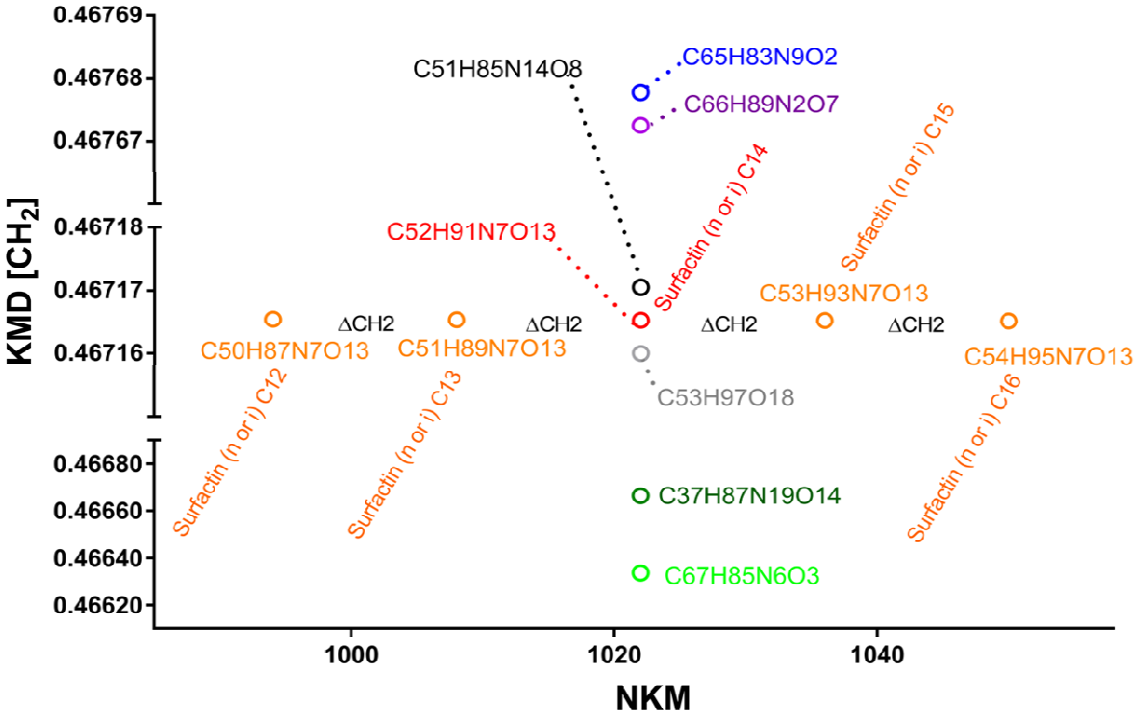


Figure 41. Tracé 2D représentant le KMD en fonction de NKM (KMD/NKM 2D-plot) pour une unité de base CH_2 des 7 formules moléculaires possibles et de 4 autres membres de la famille des surfactines (points oranges).

Certaines classes de NRPs sont des molécules produites par des micro-organismes sous la forme d'un ensemble de composés de même famille. Cette particularité est un atout pour l'identification de composés inconnus. En effet, comme le montre la Figure 41, les molécules de la même famille affichent une corrélation structurale basée sur un CH_2 et sont par conséquent alignées sur la même valeur de KMD (ligne horizontale dans le graphe 2D KMD/NKM). Ainsi, cinq composés (points rouge et orange) apparaissant sur le graphe 2D KMD/NKM comme cinq points distincts alignés sur la même ligne horizontale sont structurellement liés et ne diffèrent que par un groupe CH_2 (Figure 41).

6.3.3. Création d'un maillage vectoriel de Kendrick pour les NRPs

Plus généralement, le tracé bidimensionnel KMD/NKM peut être extrapolé à tous les NRPs (= toutes les formules moléculaires) extraits de NORINE. En effet, quel que soit le composé d'intérêt, l'addition ou la soustraction d'un atome ou d'un groupe d'atomes (par exemple ± CH_2, ± O, ± N ...) génère toujours la même variation (Δ) de la valeur de KMD (ΔKMD) et la même variation de la masse nominale de Kendrick (ΔNKM). En effet, comme l'illustre la Figure 42, quelle que soit la molécule, l'addition ou la perte (dans la formule moléculaire) d'un atome d'azote (ligne noire) provoque un changement de masse nominale de Kendrick (ΔNKM) de 14 mais aussi un changement de défaut de masse de Kendrick (ΔKMD) de 0,01256. De plus, le mode de calcul NKM (pour rappel : NKM = valeur de KM arrondi à l'entier le plus proche) et le tracé 2D basé sur la valeur de KMD et de NKM créent une symétrie par rapport à un point d'origine donné, en l'occurrence le point 1021 ; -0,25283 dans la Figure 42.

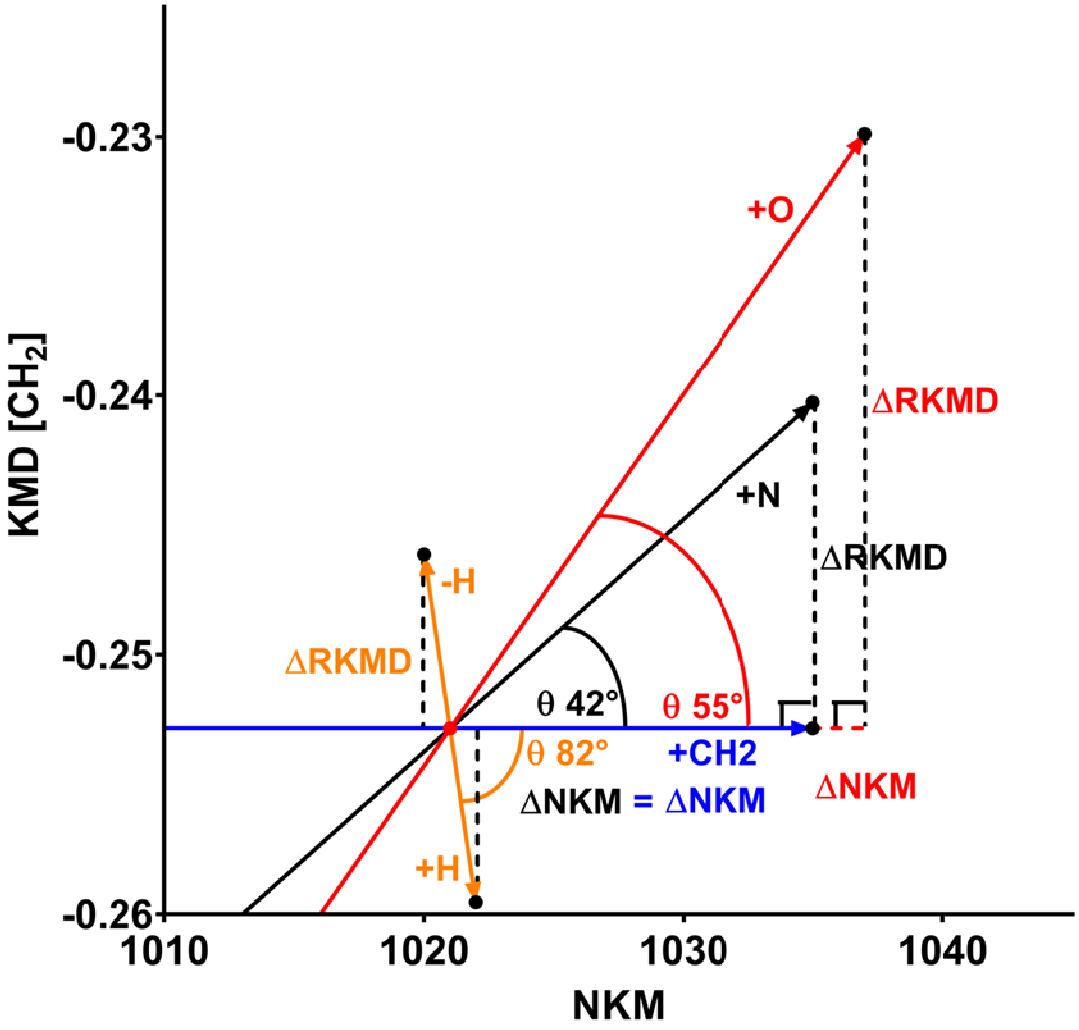


Figure 42. Tracé 2D KMD/NKM corrélé au bloc de construction CH_2 illustrant le vecteur (défini par l'angle thêta et la longueur du vecteur) obtenu pour chaque incrément d'un atome donné.

Comme illustré sur la Figure 42, deux points proches du graphe 2D KMD/NKM sont donc liés entre eux par un triangle rectangle (sauf pour une variation d'un CH_2 qui forme, par définition, seulement une ligne horizontale (ligne bleue) puisque ΔKMD = zéro). Ce triangle rectangle est défini par les lignes de valeur ΔNKM et ΔKMD et la ligne joignant les deux points qui forment l'hypoténuse du triangle rectangle. Par exemple, quel que soit le composé, l'addition d'un atome d'oxygène (droite rouge, Figure 42) à une formule moléculaire correspond trigonométriquement à deux points, formant l'hypoténuse (valeur = 27,9730) d'un triangle rectangle et par définition un même angle thêta (θ) (valeur = 55°) par rapport à l'axe horizontal (x ou NKM). En revanche, la soustraction d'un atome d'oxygène

(droite rouge, Figure 42) au sein d'une formule moléculaire correspond trigonométriquement à deux points formant l'hypoténuse (même valeur = 27,9730) du triangle rectangle et le même angle θ (valeur = 55°) par rapport à la ligne horizontale. *In fine*, en raison du mode de calcul du KMD et du ΔNKM, les formules moléculaires différant par un atome ou un groupe d'atomes (autre que CH_2) sont diagonalement corrélées, en valeurs positives de KMD et de NKM ou en valeurs négatives de KMD et de NKM, par rapport au point d'origine que constitue la position de l'une des molécules. Ce sens diagonal de la lecture est illustré à travers la Figure 43.

Ainsi, quelle que soit la valeur du KMD (au-dessus de zéro ou en dessous de zéro, Figure 43), une variation positive du NKM par ajout d'un atome ou d'un groupe d'atomes se traduit par un vecteur vers la droite et inversement une variation négative du NKM se traduit par un vecteur vers la gauche. Par conséquent, le positionnement des NRPs connus, sur le graphique 2D KMD/NKM basé sur des formules moléculaires, fournit un moyen nouveau et rapide d'identifier de nouveaux NRPs.

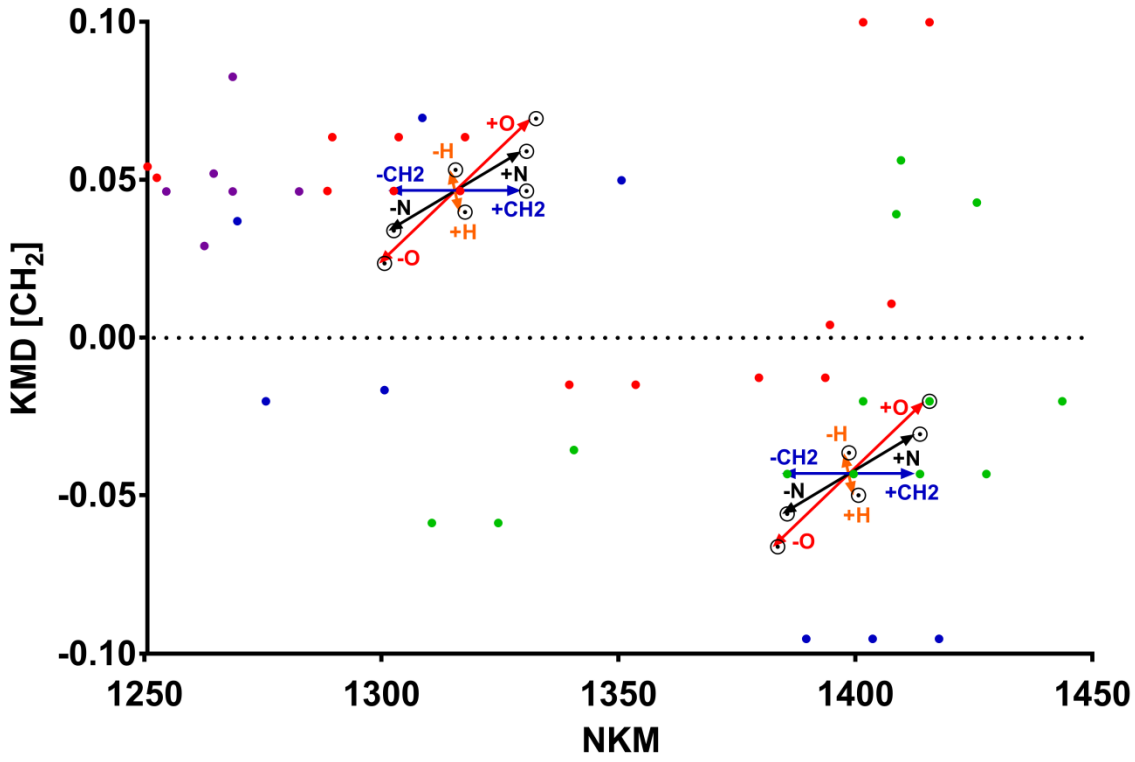


Figure 43. Tracé 2D KMD/NKM pour le bloc de construction CH_2 illustrant pour chaque incrément d'un atome ou groupe d'atomes donné leur corrélation par rapport à un point d'origine pour des valeurs positives ou négatives de KMD.

6.3.4. Création de la carte NORINE basée sur le calcul de défaut de masse de Kendrick

Depuis plusieurs années, NORINE est reconnue comme la base de données de référence unique avec des informations vérifiées relatives aux NRPs et, à ce titre, elle fournit une bonne couverture des formules moléculaires possibles des NRPs. À partir de ces formules, la masse monoisotopique protonée théorique a été calculée et reportée sur le graphique 2D KMD/NKM (Figure 44) par rapport au défaut de masse du CH_2. Ce tracé 2D a été appelé la carte de NORINE basée sur le calcul du défaut de masse de Kendrick. Les classes de molécules (lipopeptides, peptides, peptaibols, glycopeptides, chromopeptides et hybrides PK/NRPs) sont représentées en utilisant différentes couleurs. La masse moléculaire des NRPs est comprise entre 200 et 3000 Da. La grande majorité se situe entre 500 et 1500 Da et forme deux groupes. Le premier groupe rassemble des molécules affichant une valeur

NKM comprise entre 200 et 1000 Da et une valeur de KMD élevée de l'ordre de 0,15 à 0,5 tandis que les points les plus dispersés constituent le second groupe. Les peptides (points bleus) et les lipopeptides (points rouges) sont les composés principaux du premier groupe qui contient également quelques hybrides PK/NRPs (points oranges). Le second groupe est plus hétérogène et englobe toutes les classes de molécules (par exemple, tous les chromopeptides (points violets) et glycopeptides (points noirs), tous les peptaibols (points verts) et le reste des peptides et hybrids PK/NRPs. Les classes de peptides et de lipopeptides représentent respectivement 42,2% et 25,1% de l'ensemble des NRPs annotés de NORINE. Ces classes de composés se retrouvent en partie dans le premier groupe. Dès lors, la représentativité plus importante des peptides et des lipopeptides explique qu'ils apparaissent plus hétérogènes et éparses sur cette carte (Figure 44).

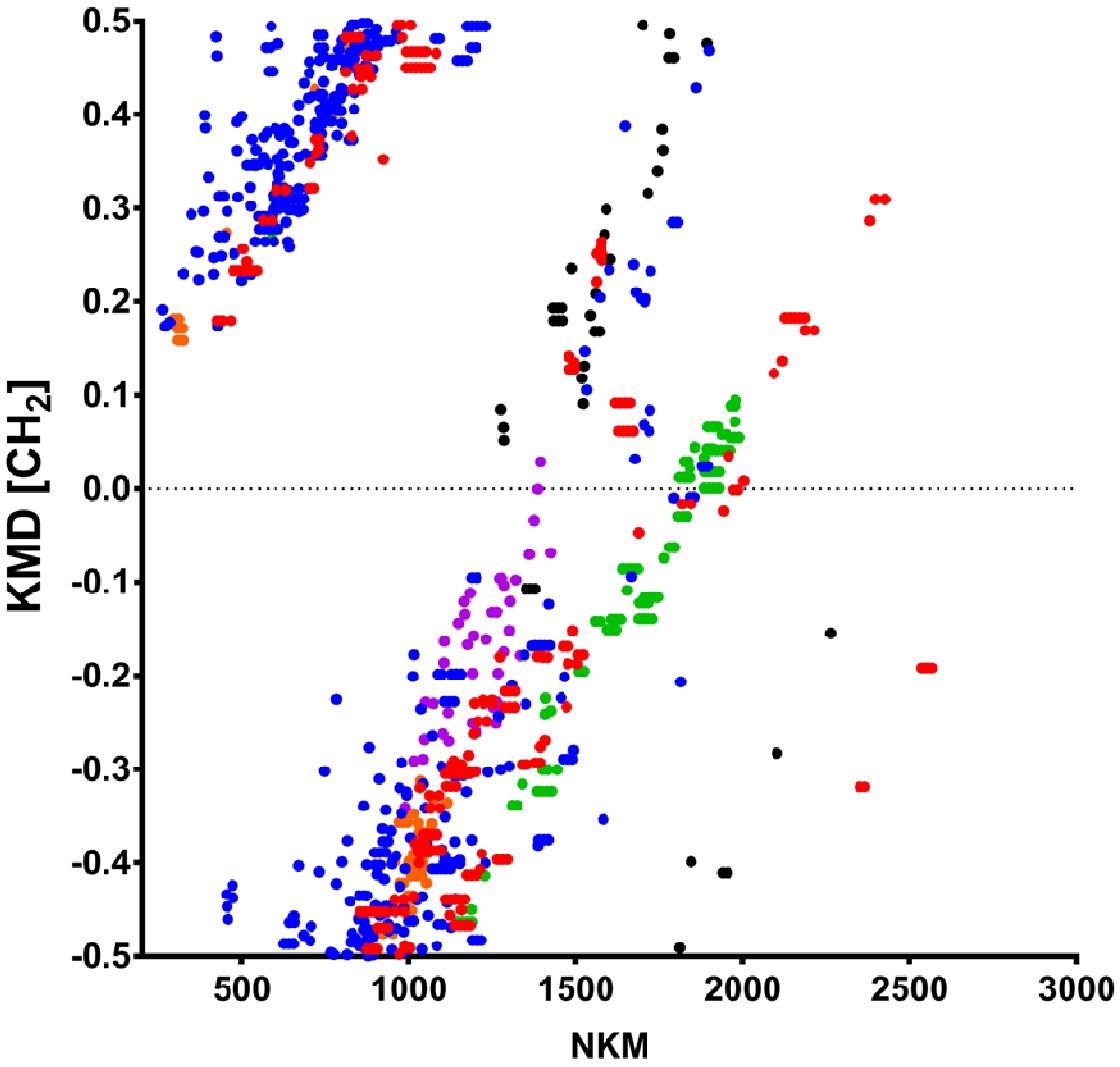


Figure 44. Graphique 2D du KMD/NKM (carte de NORINE basée sur le calcul de défaut de masse de Kendrick) de tous les NRPs référencés dans la base de données NORINE. Les classes de molécules sont représentées en utilisant des couleurs: lipopeptides (points rouges), peptides (points bleus), polycétides (points oranges), chromopeptides (points violets), glycopeptides (points noirs) et peptaibols (points verts).

La formation des deux groupes de molécules sur cette carte peut s'expliquer par le phénomène de réflexion spectrale par rapport à l'axe des x (NKM). Afin d'éviter d'éventuels repliements de formule brute, il est possible de soustraire une partie du tracé à la masse de Kendrick pour retracer une carte de Kendrick « corrigée » appelé tracé « régulier » du défaut de masse de Kendrick (RKMD). Ici, nous avons choisi une valeur de KMD comprise entre 0,2 et 0,3 suivant la gamme de masse (masse nominale) respective de 200 ou 500 Da. Ces

valeurs nous permettent de corriger les valeurs présentes dans le groupe en haut à gauche de la Figure 44, c'est-à-dire pour des valeurs de KMD comprises entre 0,28 et 0,5.

6.3.5. La carte de NORINE corrigée par dé-repliement spectral

Toutes les valeurs de KMD [CH_2] sont corrigées par une valeur de 0,28. Les nouvelles valeurs maintenant appelé RKMD permettent de tracer une nouvelle carte de NORINE corrigée (RKMD/NKM 2D-plot), Figure 45. Cette carte ne présente plus de symétrie engendrant deux groupes de molécules (Figure 45).

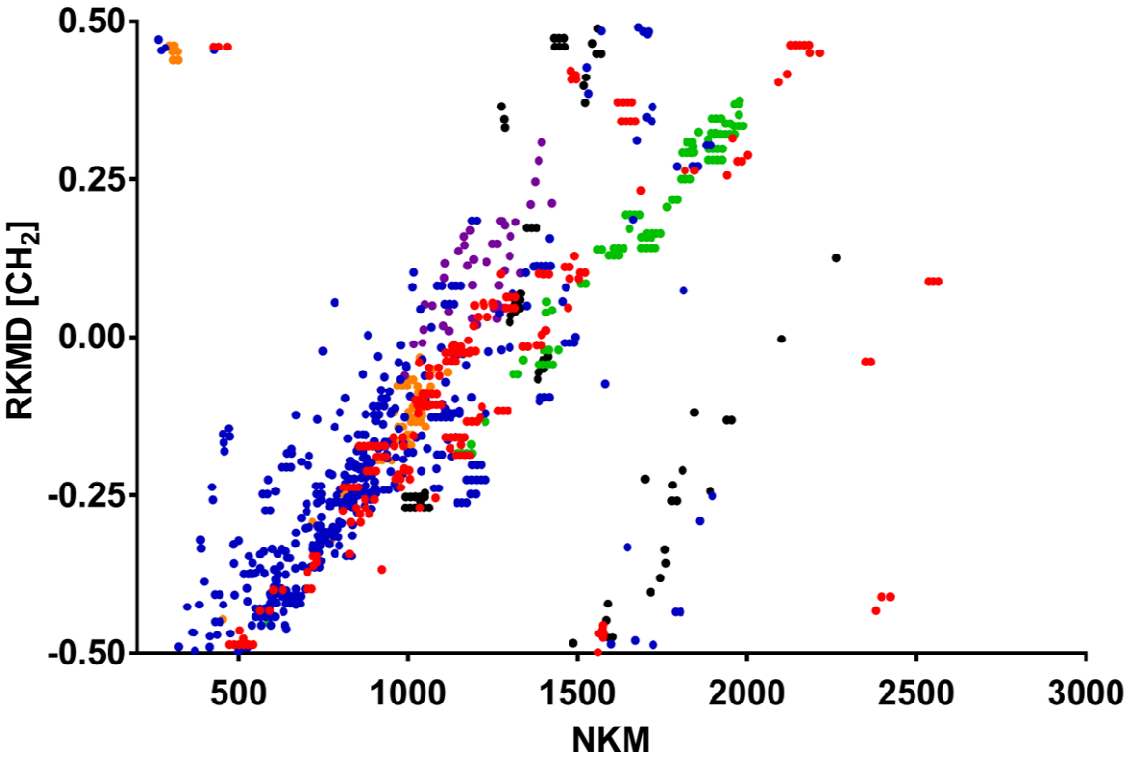


Figure 45. Graphique 2D du RKMD en fonction de NKM (carte de NORINE basée sur le calcul de Kendrick) de tous les NRPs référencés. Les classes de molécules sont représentées en utilisant des couleurs: lipopeptides (points rouge), peptides (points bleu), polycétides (points orange), chromopeptides (points violets), glycopeptides (points noirs), peptaibols (points verts).

Ici on retrouve une répartition organisée en un bloc en diagonale représenté par toutes les classes de NRPs et quelques points satellites. Cette correction des valeurs de KMD en RKMD nous permet de palier au phénomène de réflexion spectrale et par conséquent des

erreurs dans la détermination des formules moléculaires. Ainsi, le groupe de molécules en haut à gauche de la Figure 44 avec des KMD compris entre 0,28 et 0,50 issus de ce phénomène est alors déplacé dans les valeurs de KMD de -0.22 et -0.50 (Figure 45). Les nouvelles valeurs (RKMD) réintègrent donc le groupe central en diagonale.

6.3.6. Preuve de concept en utilisant les NRPs de la famille des surfactines

Des surfactines disponibles dans le commerce ont été co-cristalisées avec la matrice (HCCA) et analysées par MALDI-FT-ICR-MS (Figure 46). Le spectre présente quatre signaux d'une intensité supérieure à $0{,}5.10^{7}$ I.A. présentant des masses monoisotopique de m/z 994,64374 ; 1008,65934 ; 1022,67464 et 1036,69009.

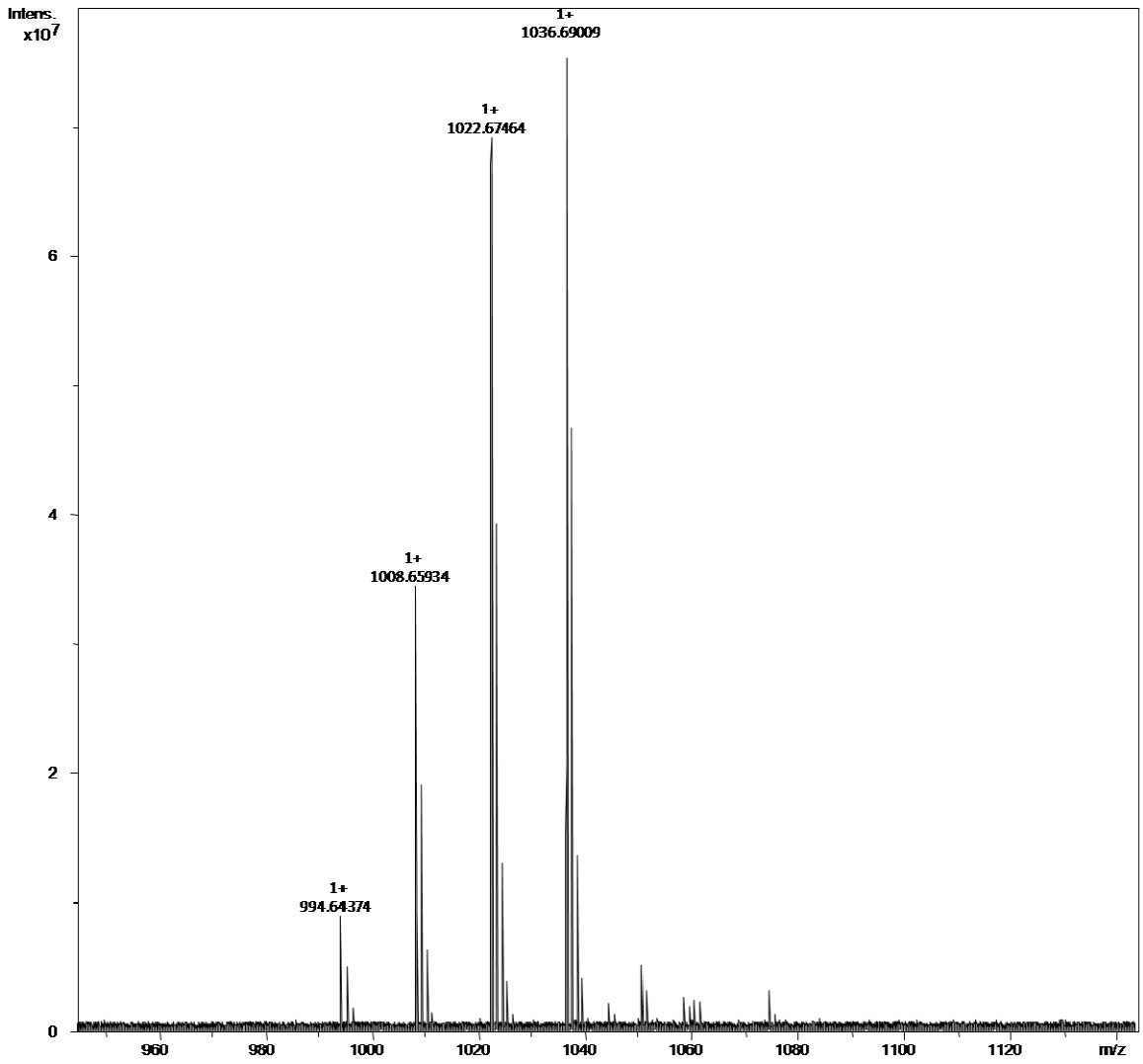


Figure 46. Spectre MALDI FT-ICR d'un échantillon commercial de surfactine.

Les masses monoisotopiques provenant des spectres de masse acquis ont servi pour les calculs de RKMD et NKM et les couples de valeurs RKMD/NKM ont été reportés sur la carte de NORINE (complète) non tronquée (Figure 47) puis sur une carte de NORINE tronquée dans laquelle toutes les molécules de la famille des surfactines d'intérêt ont été préalablement retirées (Figure 48).

Les quatre masses monoisotopiques les plus intenses (994,64374 ; 1008,65934 ; 1022,67464, 1036,69009) ont été extraites du spectre HRMS (Figure 46) et utilisées pour les calculs de RKMD et NKM. La surfactine de masse moléculaire (994,64374) correspond à la valeur de NKM 994 et à la valeur de RKMD de -0,2531. Par analogie, les surfactines de masse moléculaire 1008,65934 et 1022,67464 et 1036,69009 fournissent des couples de valeur

NKM/RKMD de respectivement 1008 et -0,2531 ; 1022 et -0,2527 et 1036 et -0,2534, tracé sur les graphiques 2D RKMD/NKM des Figure 47 et Figure 48 (points rouges).

Comme attendu, les mesures issues des quatre variants de surfactines correspondent parfaitement aux composés de NORINE possédant les formules moléculaires suivantes: $C_{50}H_{87}N_7O_{13}$, $C_{51}H_{89}N_7O_{13}$, $C_{52}H_{91}N_7O_{13}$ et $C_{53}H_{93}N_7O_{13}$, respectivement (Figure 47, cercles rouges autour points bleus). Évidemment, ces formules moléculaires correspondent à celles des variants de surfactines où les variations expriment une différence de CH_2 au niveau de leur formule brute qui peut correspondre à différentes longueurs de chaînes d'acides gras (ou des acides aminés différent du corps peptidique).

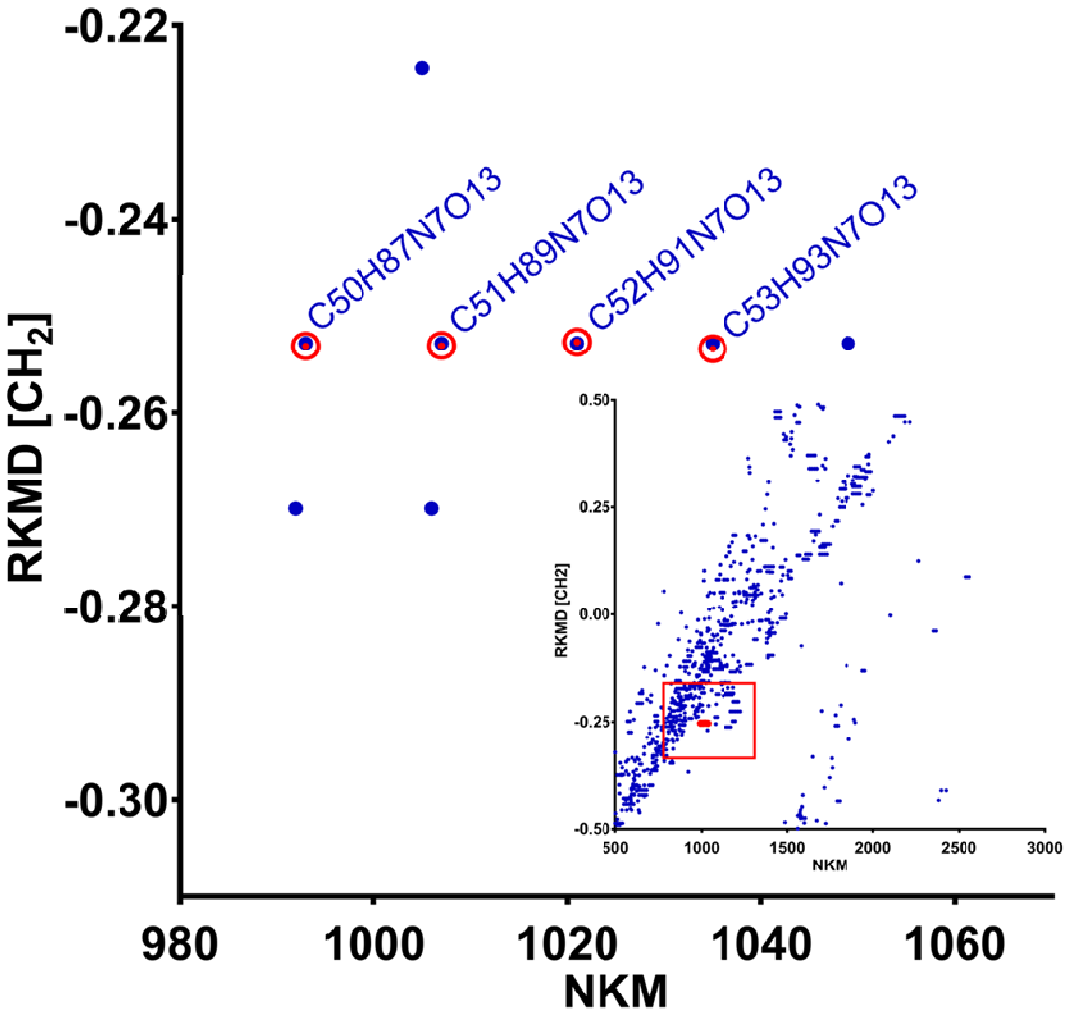


Figure 47. Graphique 2D de RKMD/NKM montrant la position des points correspondant aux 4 surfactines disponibles dans le commerce (A) sur la carte NORINE complète (non

tronquée). Le positionnement des 4 surfactines sur la carte correspond parfaitement aux formules moléculaires des composés connus de la base de données NORINE.

Fort logiquement, en utilisant la base de données NORINE tronquée des membres de la famille des surfactines, aucun des points de coordonnées RKMD/NKM, correspondant aux masses des surfactines dont nous avons précisément mesuré le m/z, ne correspond aux points de la base de données NORINE tronquée (ne contenant pas les surfactines référencées).

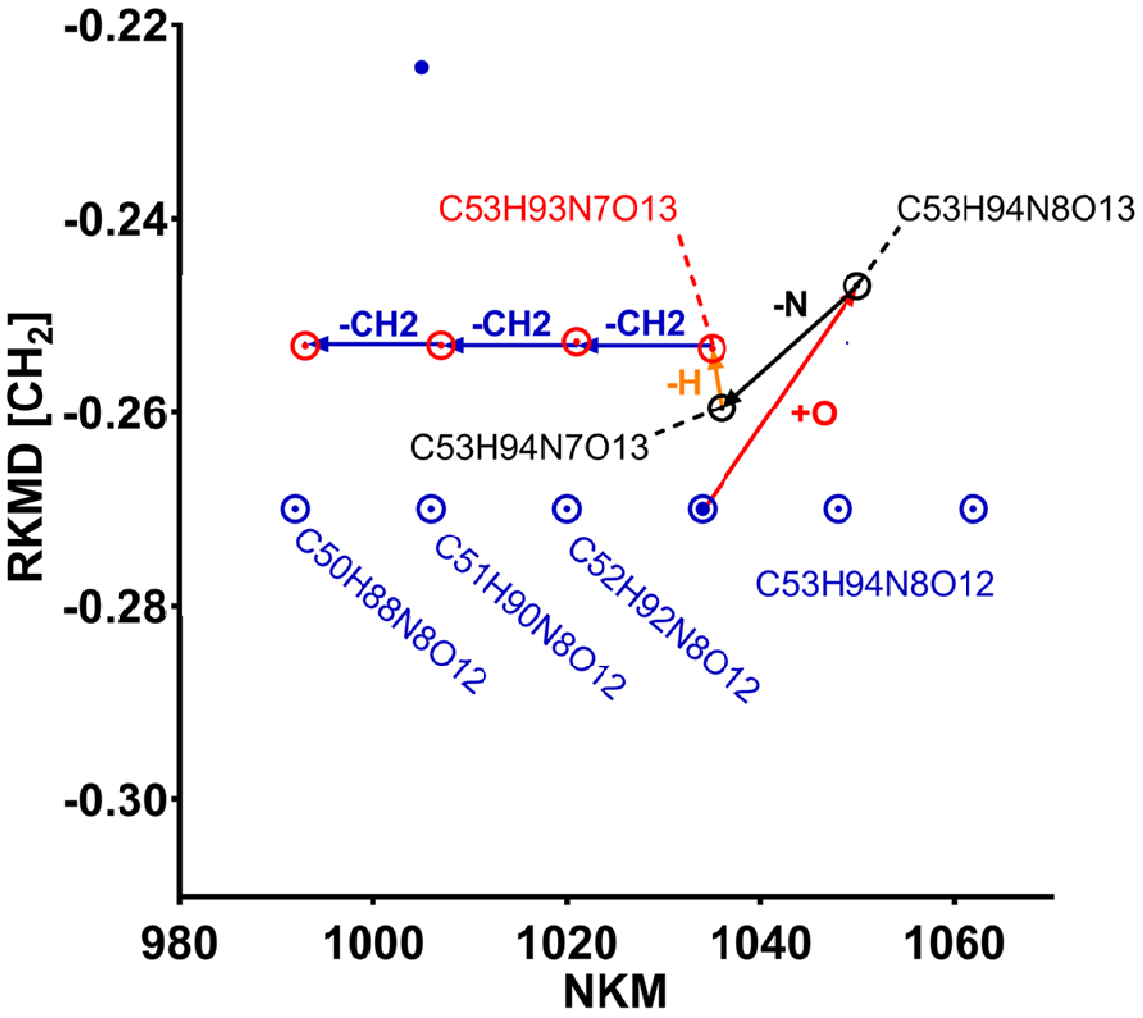


Figure 48. Graphique RKMD/NKM 2D montrant la carte de NORINE où les surfactines d'intérêt ont été retirés de la base de données. En bleu les composés connus et en rouge les composés inconnus.

La déduction des formules moléculaires s'effectue alors étape par étape en utilisant le maillage vectoriel de Kendrick. Comme illustré, un point de la carte, correspondant à une molécule inconnue, peut être lié à au moins un point connu de la carte de NORINE déficiente en surfactines par l'intermédiaire d'un trajet vectoriel. Les vecteurs définis par les couples hypoténuse et angle θ sont utilisés pour relier des points connus et inconnus afin de déterminer la différence en termes d'atomes entre molécules à identifier et molécules de référence. Par exemple, à partir de la molécule de formule brute $C_{53}H_{94}N_8O_{12}$ (cercle bleu entourant un point bleu), trois étapes sont nécessaires (addition (flèche rouge) d'un atome d'oxygène (^{16}O), soustraction (flèche noire) d'un atome d'azote (^{14}N) et soustraction (flèche orange) d'un atome d'hydrogène (^{1}H)) pour atteindre le point du composé inconnu marqué [1036; -0,2534] (cercle rouge autour d'un point rouge). Par conséquent, la formule moléculaire putative du composé inconnu de coordonnées [1036; -0,2534] sur la carte de NORINE basée sur le défaut de masse de Kendrick est $C_{53}H_{93}N_7O_{13}$. *In fine*, la corrélation entre les membres de cette famille de molécules que sont les surfactines apparaît, comme illustré sur la Figure 48 (flèches bleues), par la différence atomique facilement reconnaissable d'un groupement $^{12}C_1{}^{1}H_2$. Notons que plusieurs chemins sont possibles pour relier deux points distincts mais ces chemins mènent tous à la même formule moléculaire.

6.3.7. Conclusion

À l'heure actuelle, l'identification des NRPs connus (= déréplication) ou inconnus produits par un microorganisme reste encore une tâche longue, coûteuse et difficile (Yang *et al.*, 2013). Même si les mélanges analysés en NRPomique ne sont pas plus complexes que ceux analysés en protéomique ou en métabolomique, la diversité structurale des NRPs pose de nombreux problèmes dans l'interprétation des données expérimentales (Nguyen *et al.*, 2016).

En effet, la nature chimique composite des NRPs entrave significativement l'élucidation de leur structure chimique (Ibrahim et al. 2012; Niedermeyer and Strohalm 2012). Dans ce contexte, la RMN est une méthode analytique, cruciale et inestimable pour l'élucidation des structures moléculaires nouvellement découvertes. Cependant, la première étape du processus d'élucidation structurale consiste souvent à mesurer aussi exactement que possible la masse moléculaire des composés pour déduire leurs formules moléculaires

de la masse monoisotopique et du massif isotopique mesurés. Évidemment, la formule moléculaire ne conduit pas à la résolution de la structure dans son ensemble, mais c'est le point de départ qui guide l'élucidation progressive de la structure du composé. Malheureusement, pour les composés de masse moléculaire supérieure à 1000 Da, l'imprécision très relative des spectromètres de masse à haute résolution empêche la déduction assistée par ordinateur d'une formule moléculaire unique. Cependant, la réduction du nombre de formules moléculaires candidates apparaît comme un avantage certain pour la déréplication moléculaire et l'accélération de l'élucidation des structures.

Dans l'approche de Kendrick, tous les composés qui diffèrent dans leurs formules moléculaires par un ou plusieurs groupements $^{12}C_1{}^{1}H_2$ se distinguent par une même valeur de RKMD égale à zéro (les composés sont alignés sur une ligne horizontale). Nous montrons au travers de ce travail de thèse que tous les composés qui diffèrent par le même incrément de formule moléculaire (par exemple une addition ou délétion dans une formule moléculaire donnée d'un atome d'azote, d'un atome d'oxygène ou d'un groupe hydroxyle) possèdent la même valeur de variation de RKMD (ΔRKMD) et la même valeur de variation de NKM (ΔNKM). Rappelons que lorsqu'un atome ou un groupe d'atomes est ajouté à une formule moléculaire donnée, la valeur NKM augmente et inversement. Dès lors, l'addition ou la soustraction d'un atome donné correspond toujours, sur la carte de NORINE basée sur le défaut de masse de Kendrick, à un même vecteur défini (à partir du plan horizontal) par **i**) la valeur de l'angle θ qu'il forme avec l'horizontale et la valeur de la distance qui sépare les deux points de cette carte. À partir d'un point donné (formule moléculaire), il existe donc un maillage vectoriel qui relie un point donné aux points environnants proches. Ce maillage vectoriel permet de connecter une molécule inconnue à une formule moléculaire connue, permettant ainsi son identification. Par conséquent, le graphe représentant le RKMD en fonction NKM (RKMD/NKM 2D-plot), présente deux avantages: premièrement, il met en évidence l'alignement horizontal de composés apparentés au groupe atomique de référence (par exemple $^{12}C_1{}^{1}H_2$), tels que des homologues ou des membres d'une même famille lipopeptidique; d'autre part, la superposition du maillage vectoriel sur un point donné de la courbe 2D RKMD/NKM relie deux formules moléculaires proches. Dans notre exemple, il a conduit à l'identification de la formule moléculaire de variants de surfactine comme [Ala4] iC14, [Ala4] iC15, [Val7] iC15 et [Ile7] iC14. En fin de compte, le tracé 2D RKMD/NKM conduit

à la corrélation rapide des composés proches sur la base de leur différence de composition atomique.

L'approche KMD est largement utilisée pour l'étude et la caractérisation structurale de polymères chimiques structurellement corrélés entre eux comme le pétrole (Ohno et al. 2014; Roach, Laskin, and Laskin 2011; Sato 2017). Dans ces études, la détermination des formules moléculaires est basée sur l'identification, à partir du graphique 2D KMD/NKM, de séries de composés en grand nombre corrélés horizontalement entre eux, suivis d'un processus de régression moléculaire vers les faibles masses moléculaires jusqu'à l'obtention de la formule brute, à l'aide d'un générateur de formule brute, en s'appuyant sur les profils isotopiques et la mesure exacte de composés apparentés mais de plus faible masse. En effet, plus la masse moléculaire diminue et plus le nombre de possibilités de formules moléculaires candidates diminue. Ainsi, connaissant la formule brute du composé de plus faible masse moléculaire et le lien de parenté entre les molécules d'une même famille, il est possible de réattribuer la formule brute des toutes les molécules de la famille.

Malheureusement, un tel processus n'est pas directement applicable à la détermination des formules moléculaires des NRPs qui présentent des structures plus complexes car, construites à partir de plus de 500 monomères (Caboche et al. 2007). De plus, les séries de molécules NRPs de la même famille sont toujours composées d'un petit nombre d'individus et ne permettent pas d'utiliser la distribution isotopique et la masse exacte des composés de faible masse de cette famille pour établir la formule moléculaire. Par conséquent, la combinaison d'une base de données chimiques ou biochimiques (telle que PubChem, ChemSpider (Online 2015)) et le graphe 2D RKMD/NKM apparaît comme un moyen simple et efficace d'annoter cette représentation graphique. Les formules moléculaires des composés connus peuvent être positionnés sur le graphe et ensuite servir de points de référence pour déduire les formules moléculaires des points proches. À notre connaissance, une telle combinaison n'a jamais été utilisée et reportée dans la littérature.

Dans le domaine des NRPs, NORINE est la base de données unique qui rassemble près de 1200 composés recueillis dans la littérature et vérifiés manuellement. NORINE contient plusieurs types d'informations sur les NRPs, y compris leur formule moléculaire. Toutes les formules moléculaires de NRPs dans NORINE ont été extraites et ont servi à créer

la carte de NORINE basée sur le défaut de masse de Kendrick (graphe 2D RKMD/NKM) pour annoter cette représentation graphique.

La déréplication des NRPs compte pour une large part dans le processus de découverte de nouveaux NRPs. Les composés produits par les microorganismes sont principalement soumis à une analyse HRMS (avec ou sans séparation préalable) afin de mesurer aussi exactement que possible la masse moléculaire et la distribution isotopique des composés. Ainsi, la formule moléculaire des composés produits peut être déduite des m/z mesurés en utilisant une approche combinant le RKMD et la base de données NORINE où les composés connus de la base de données NORINE servent à annoter le graphique: RKMD/NKM. *In fine*, par transformation de la valeur du m/z mesuré d'un composé comme point du tracé 2D RKMD/NKM, sa formule moléculaire peut être déduite, soit en faisant correspondre avec un point d'un composé connu de NORINE, soit en caractérisant le chemin vectoriel pour connecter un point inconnu à un point connu ou inversement.

Cette approche est dépendante de données obtenues sur des appareils de très haute résolution comme les appareils équipés d'analyseur FT-ICR. Cependant, le groupe de recherche du Pr Sato (Tsukuba, Japon) ont récemment, rapporté qu'une transformation mathématique différente de l'approche du défaut de masse de Kendrick permet d'améliorer la résolution spectrale des graphiques RKMD/NKM et permet ainsi l'utilisation de données de masse issues d'appareils de spectrométrie de masse moins résolutifs mais juste (Fouquet & Sato, 2017c). Cette transformation mathématique nouvelle peut également être utilisée pour la caractérisation de la formule moléculaire des NRPs (Figure 48). Le principe de cette transformation mathématique repose sur l'utilisation d'une unité de base fractionnaire différente.

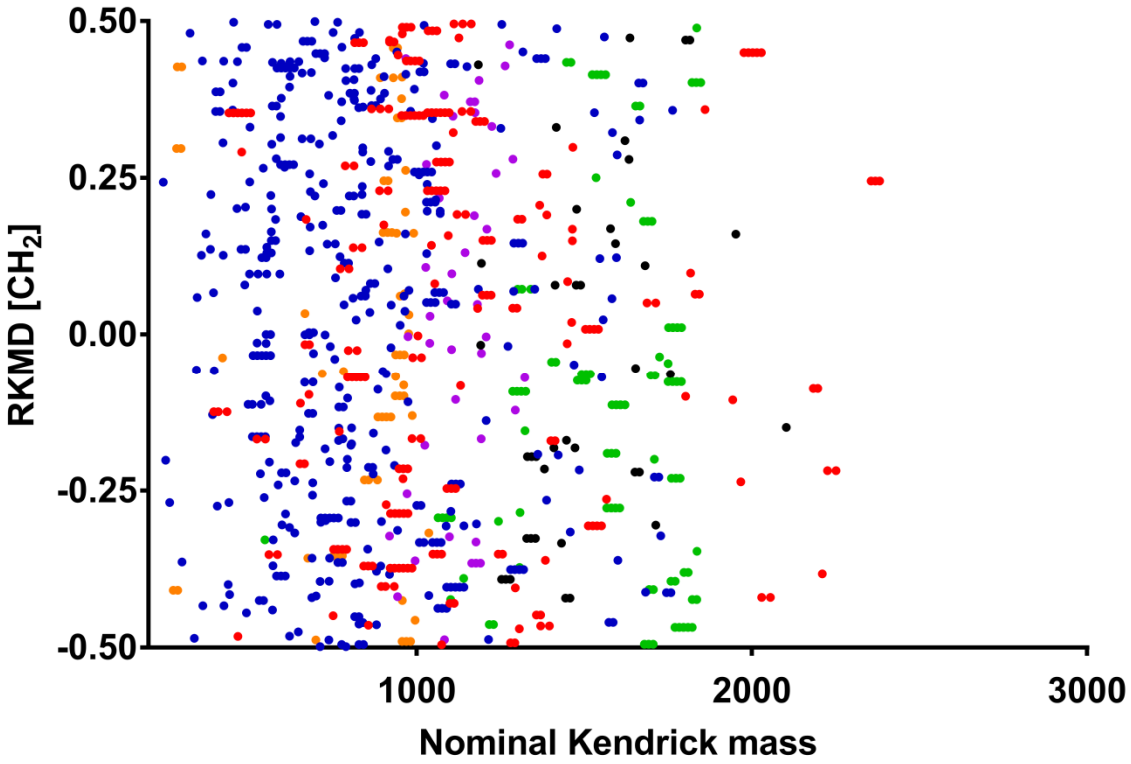


Figure 49. Graphique 2D RKMD/NKM sur une valeur de 13 permettant l'augmentation de la résolution graphique pour une exploitation de données de masse expérimentales moins résolutives. Les classes de molécules sont représentées en utilisant des couleurs: lipopeptides (points rouge), peptides (points bleu), polycétides (points orange), chromopeptides (points violets), glycopeptides (points noirs), peptaibols (points verts).

Le graphique, Figure 49, montre une distribution homogène de toutes les classes de NRPs sur la gamme des NKM mais surtout sur la gamme des RKMD. En effet, contrairement à la Figure 45 (carte de NORINE 2D RKMD/NKM avec unité de base fractionnaire de 14) qui montre une répartition diagonale des NRPs, la Figure 49 (carte de NORINE 2D RKMD/NKM avec une unité de base fractionnaire de 13) présente une répartition plus éclatée des valeurs de RKMD des NRPs. D'autre part, la visualisation de séries d'ions analogues aux blocks de construction CH_2 (séries d'ions horizontales) est plus aisée. Nous n'avons pas encore complètement exploré cette approche mais elle nous permettrait d'étendre notre approche pour des données obtenues sur des spectromètres de masse de type Q-TOF par exemple.

6.3.8. Automatisation du maillage de Kendrick

Ce raisonnement simple et ces calculs ont permis le développement d'un logiciel dédié à l'assignement de formules moléculaires à partir de données de masse de haute résolution. Le logiciel a été développé en Java (back-end) et Java script, HTML et CSS (front-end) en collaboration avec Maude Pupin du laboratoire CRIStAL et Emma Ricart, doctorante au SIB (Genève, Suisse). Nous avons appelé cet outil le « *Kendrick formula Predictor* », il est hébergé par un serveur de l'Université de Lille et est disponible à cette adresse :

http://bioinfo.cristal.univ-lille.fr/kendrick-webapp/

L'interface utilisateur du logiciel est accessible depuis les différents moteurs de recherche tels que Chrome, Firefox ou encore safari. L'interface est simple d'utilisation, il suffit d'entrer la masse protonée ($[M+H]^+$) obtenue par spectrométrie de masse dans le premier encadré (Figure 50).

L'interface graphique comprend plusieurs champs à remplir sur le panneau de gauche et à droite une visualisation graphique interactive du tracé RKMD en fonction du NKM est obtenue. La recherche d'une formule moléculaire peut être réalisée avec les données présentes dans la base de données NORINE et/ou PubChem. En effet, une partie des formules moléculaires de PubChem ont été extraites afin de déterminer les masses monoisotopiques et pouvoir recalculer tous les RKMD et NKM suivant la même démarche que précédemment décrite avec la base de données NORINE.

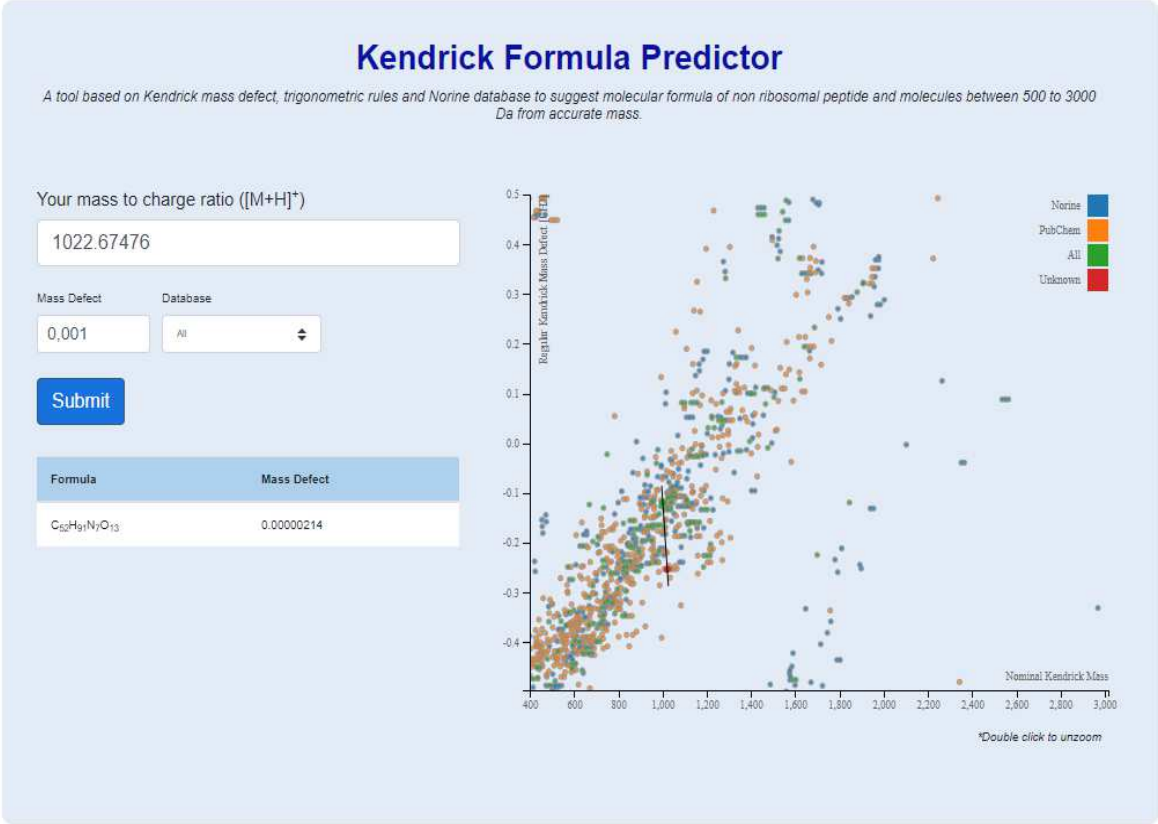


Figure 50. Copie écran illustrant l'interface utilisateur du « *Kendrick Formula Predictor* » après soumission du m/z 1022,67476.

Compte tenu du rapport masse sur charge, un ensemble de formules chimiques prédites à l'aide de l'approche de Kendrick sont suggérées (Figure 50). Les paramètres d'entrée supplémentaires incluent le défaut de masse et la base de données choisie. Le défaut de masse détermine la précision de la recherche. Enfin, l'option de base de données augmente l'espace de recherche avec un ensemble supplémentaire de masses extraites de PubChem. Cet ensemble de données a été soigneusement sélectionné à l'aide de la recherche d'ontologies fournies par le navigateur de classification PubChem dans le but d'obtenir des composés de type NRPs. Les formules prédites sont présentées dans un tableau spécifiant la formule moléculaire candidate et le défaut de masse de chaque formule lié à une représentation graphique du tracé 2D RKMD/NKM, montrant à la fois le point de la carte et le vecteur utilisés pour chaque prédiction ainsi que l'ensemble des données qui lui sont associées. Un code couleur est utilisé pour la représentation graphique des formules moléculaires référencées dans NORINE (en bleu), dans PubChem (en orange) et éventuellement pour l'affichage des deux bases de données (en vert). Enfin le point rouge représente le composé inconnu soumis. Cet outil est une première version qui nécessite une

soumission manuelle des masses monoisotopique protonées ($[M+H]^+$). De plus, seuls les vecteurs pour les atomes d'azote, d'oxygène, d'hydrogène et pour une différence de CH_2 sont disponible. Par conséquent, seules les molécules possèdent une formule brute composée d'atomes d'azote, d'oxygène, d'hydrogène et de carbone peuvent être attribuées. Une prochaine version intégrera l'interrogation à partir de fichier MGF ou mzXML directement récupérés des spectres.

L'utilisation manuelle des défauts de masse de Kendrick demande du temps et une compréhension approfondie du mode de calcul, l'automatisation permet de reproduire les résultats précédemment obtenus manuellement et avec une rapidité élevée car la recherche ne prend pas plus de 2 secondes contre au minimum 2 heures pour le faire manuellement. La rapidité du calcul permet à présent de retraiter un nombre plus important de données de masse.

6.4. Analyse de Van Krevelen

6.4.1. Création de masque de Van Krevelen dédiés aux NRPs

Le diagramme de Van Krevelen est largement utilisé en géochimie pour étudier l'évolution des échantillons de charbon ou d'huile (Kim*et al*., 2003). Il est construit en utilisant le rapport du nombre d'atomes d'hydrogène (H) sur le nombre d'atomes de carbone (C) en ordonnée (H/C) et le rapport du nombre d'atomes d'oxygène (O) sur le nombre d'atomes de carbone en abscisse (O/C). Les principales classes de composés tels que la lignine, les lipides, les hydrates de carbone ou encore les peptides ont des rapports H/C ou O/C caractéristiques. Par conséquent, chaque classe de composés possède un emplacement spécifique sur le diagramme et ainsi les types de composés peuvent être identifiés à partir de l'emplacement des points localisés dans les parcelles de Van Krevelen. Très peu voire non utilisé pour décrire des métabolites microbiens, ce diagramme semblait pouvoir apporter des informations intéressantes pour orienter l'identification de nouvelles molécules, particulièrement les NRPs.

Déjà évoqué dans le chapitre précédent, la mise à jour des formules moléculaires des composés référencés dans la base de données NORINE permet de calculer les rapports H/C et O/C pour chaque NRPs (Figure 51).

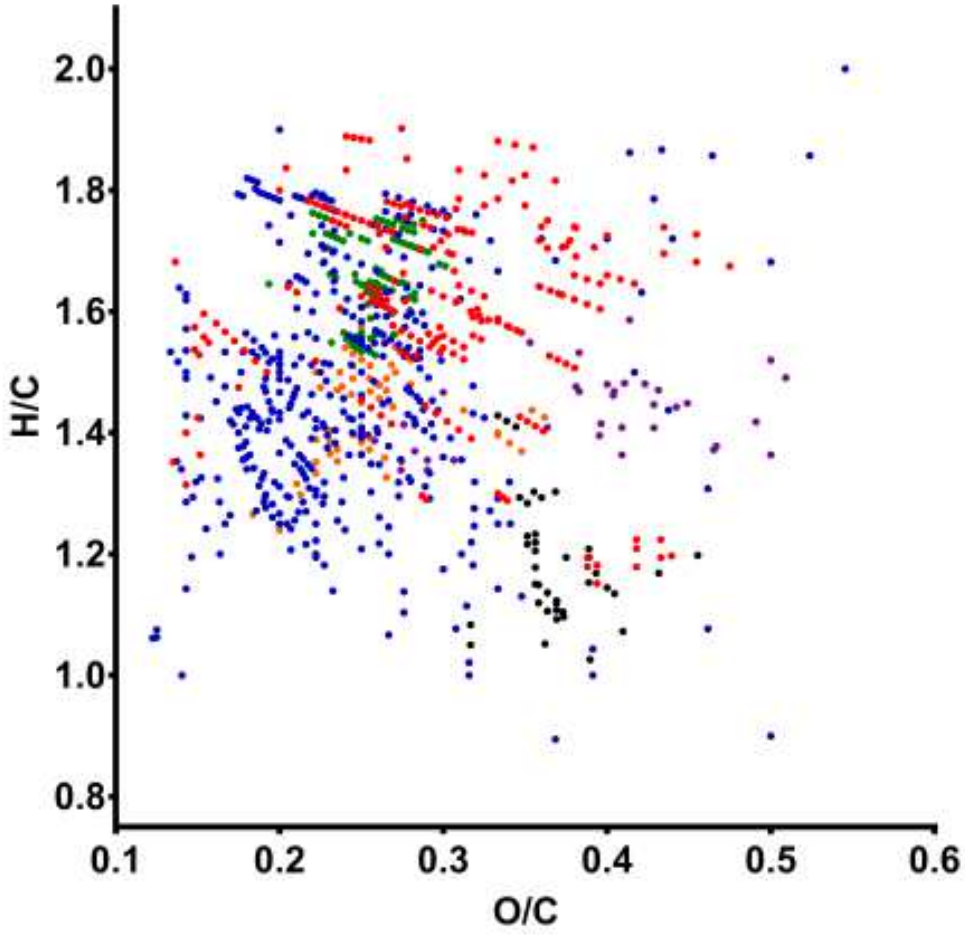


Figure 51. Diagramme de Van Krevelen de tous les NRPs référencés dans la base de données NORINE. Les classes de molécules sont représentées en utilisant des couleurs: lipopeptides (points rouges), peptides (points bleus), hybrides PK-NRPs (points oranges), chromopeptides (points violets), glycopeptides (points noirs), peptaibols (points verts).

Le diagramme de Van Krevelen a été réalisé avec les 1190 composés présents dans la base de données NORINE. Les ratios H/C de l'ensemble des molécules s'étendent de 1,0 à 1,9 et pour les ratios O/C de 0,12 à 0,50.

Les moyennes et écarts-types pour chaque classe moléculaire permettent de dessiner des zones caractéristiques et propres à chaque catégorie moléculaire. L'ensemble de ces zones forme un masque de Van Krevelen plus adapté aux NRPs correspondant à la classification proposée dans la base de données NORINE (Figure 52).

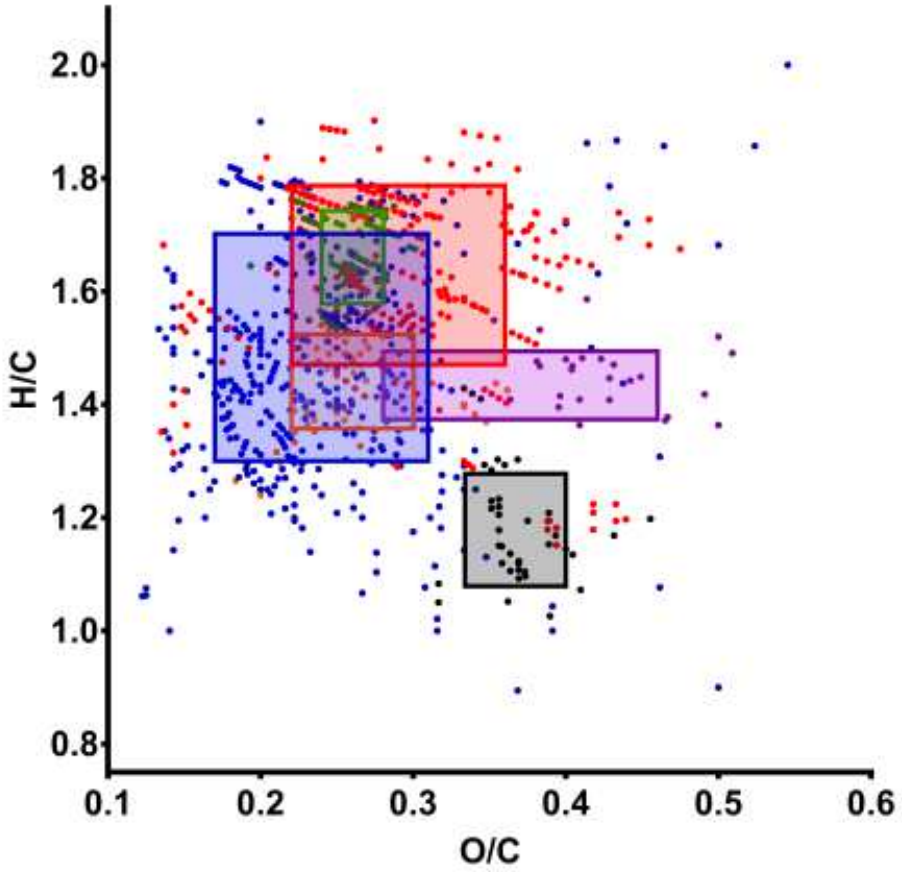


Figure 52. Diagramme de Van Krevelen de tous les NRPs référencés dans la base de données NORINE. La moyenne et l'écart-type de chaque classe de molécule a permis le tracé de rectangles appelés « masques ». Les masques de molécules sont représentés en utilisant des couleurs: lipopeptides (points rouge), peptides (points bleu), hybrides PKs-NRPs (points orange), chromopeptides (points violet), glycopeptides (points noir), peptaibols (points verts).

Les lipopeptides (n=293) et les peptides (n=492) sur la Figure 52, forment les deux groupes les plus importants sur le diagramme de Van Krevelen avec respectivement des ratios H/C moyens de 1,63 à 1,49 et des ratios O/C moyens, très proches, de 0,29 et 0,24. L'écart-type du ratio H/C des lipopeptides et des peptides sont respectivement de 0,16 et 0,20 et pour le ratio O/C de 0,07 et 0,06 ce qui montre également que ce sont les deux groupes les plus dispersés. Les peptaibols (n= 177) et les hybrides PK-NRPs (n=80) possèdent respectivement des ratios H/C moyen de 1,66 et 1,43 pour un ratio O/C moyen identique de 0,26. Ces derniers forment des groupes moins dispersés avec un écart-type pour le ratio H/C de 0,08 et des ratios O/C respectivement de 0,02 et 0,04. Les chromopeptides (n=51)

forment un groupe avec un ratio moyen en H/C de 1,43 très peu variable comme l'illustre l'écart-type de 0,06. À l'inverse les chromopeptides possèdent un ratio O/C moyen de 0,37 pour un écart-type de 0,08 montrant une disparité importante. Les glycopeptides (n = 44) possèdent un ratio moyen en H/C de 1,18 pour un ratio moyen de 0,37 et forme ainsi le groupe le plus dissociable des autres, mais également le plus petit avec les chromopeptides.

Les lipopeptides et peptides sont les groupes les plus étendus et possèdent une zone chevauchante importante de 1,45 à 1,70 environ pour le ratio H/C et de 0,22 à 0,32 pour le ratio O/C. Les peptaibols et les hybrides PK-NRPs chevauchent entièrement ou partiellement la classe des lipopeptides et peptides. Les chromopeptides chevauchent également les zones des lipopeptides, peptides et hybrides PKs-NRPs, mais dans une moindre mesure. Seule la classe des glycopeptides ne possède pas de chevauchement avec les cinq autres classes.

Comme démontré dans le paragraphe précédent, les données de spectrométrie de masse à haute résolution peuvent être utilisées pour déterminer la composition élémentaire de chaque signal de masse présent sur un spectre de masse. Ainsi, la composition élémentaire de chaque signal permet de calculer les rapports H/C et O/C puis de les positionner sur le diagramme de Van Krevelen et ainsi de prédire à quelle catégorie de NRPs appartiennent les composés potentiellement nouveaux.

7. Obtention d'informations structurale à partir des données MS/MS

7.1. L'interprétation manuelle des spectres MS/MS

L'analyse globale de la composition chimique d'extraits naturels complexes issus d'organismes vivants est une tâche laborieuse qui nécessite la mise en œuvre de techniques analytiques modernes. La multiplicité monomérique des NRPs engendre de réels problèmes d'identification et leur structure cyclique, multi-cyclique et/ou branchée vient compliquer l'interprétation des spectres de masse issus des analyses MS/MS. Ainsi, l'annotation des spectres de masse MS/MS et l'identification *de novo* de composés restent des tâches très compliquées en comparaison avec celles des peptides linéaires. Le spectre de fragmentation MS/MS de la surfactine [Leu5] iC14 est réalisé en mode CID sur un spectromètre de masse hybride (ESI-Q-TOF), son spectre de fragmentation est présenté dans la Figure 53.

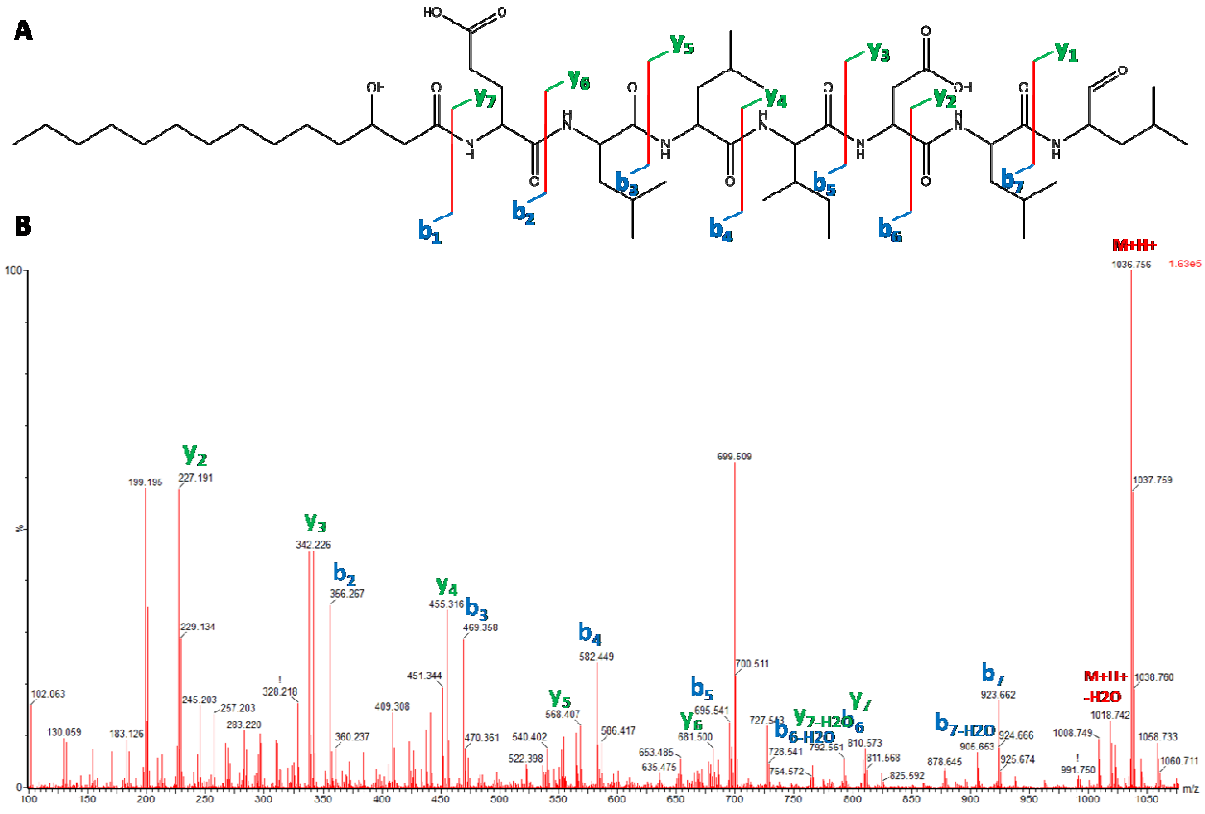


Figure 53. Structure linéaire (A) et spectre de fragmentation annoté (B) de la surfactine [Ile4] nC14 (Sigma-Aldrich).

Le spectre de fragmentation CID de la surfactine [Ile4] nC14 est composé d'un nombre important d'ions fragments. En mode de fragmentation CID, c'est la liaison peptidique qui sera rompue. Les incréments de masse permettent de déterminer la séquence en acides aminés du corps peptidique. Les ions parents de masse moléculaire 1036,7 possèdent une intensité maximale. Les ions 923,6 ; 810,5 ; 695,5 ; 582,4 ; 459,3 et 356,2 et plus particulièrement leur incrément permette de déterminer la séquence en acides aminés issue de la rupture des acides aminés du coté C-terminal (les ions b, bleu)). Les ions 810,5 ; 681,5 ; 568,4 ; 455,3 ; 342,2 et 227,2 sont quant à eux, générés lors de la fragmentation du peptide par le coté N-terminal (les ions y, vert). L'annotation des ions fragments y et b, obtenus en MS/MS à partir de surfactine, permet d'expliquer quinze fragments (avec les pertes d'eau), les ions y_1 et b_1 ne sont pas présents sur les spectres. Un certain nombre de pics sont mesurés sans pouvoir être tous expliqués réellement. Ceci est dû à la structure cyclique de la surfactine. Lors de la fragmentation le cycle peptidique va être ouvert. En effet la majorité des ions sont issus de l'ouverture de la surfactine au niveau de la liaison ester du cycle peptidique car celle-ci demande moins d'énergie pour être rompue. En revanche, une plus

faible quantité de surfactine peut également se rompre à d'autres endroits dans le cycle et ainsi générer des ions fragment de masses moléculaires encore différentes. Les différents « premier point de rupture » du cycle génère alors une quantité importante d'ions fragments. Par conséquent, il est très compliqué d'interpréter tous les signaux obtenus à partir de la fragmentation d'une molécule cyclique, multi-cyclique ou branchée. Dans la Figure 54, l'annotation d'un spectre de fragmentation de deux molécules, l'une linéaire et l'autre cyclique est présentée. Ici, le but est de montrer la différence dans l'interprétation manuelle de spectre MS/MS de l'une et de l'autre (Figure 54).

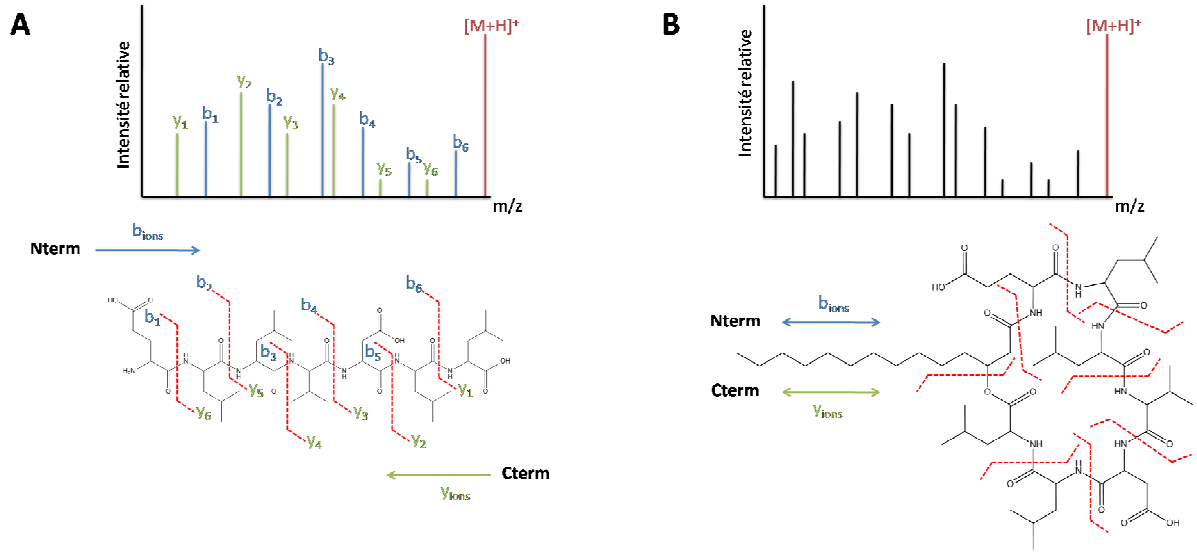


Figure 54. Illustration schématique de l'annotation d'un spectre de fragmentation d'un peptide linéaire (A) et d'un peptide cyclique NRP : la surfactine (nC14) (B).

La Figure 54, A montre un spectre de fragmentation issu d'un peptide linéaire pour lequel il est plus ou moins aisé de retrouver la séquence en acides aminés en analysant les incréments de masse entre deux signaux de masse selon la masse moléculaire des acides aminés. Les ions b (bleu) sont les ions fragments obtenus depuis le coté N-terminal, et les ions y (vert) sont les ions fragments obtenus depuis le coté C-terminal (Figure 54 A). Inversement, la structure cyclique d'une molécule ne permet pas d'établir et de distinguer un coté N-terminal d'un côté C-terminal. De plus, la surfactine d'intérêt ne possède pas une structure entièrement peptidique mais continent également une partie lipidique dont la fragmentation CID n'amène que peu d'informations structurales. *In fine*, l'interprétation du spectre de fragmentation CID de cette surfactine ne peut donc pas se résumer à une simple application de méthodes déjà existantes.

En conséquence, l'interprétation des données MS/MS issues de la fragmentation de composé cyclique est habituellement réalisée par un spécialiste qui, malgré son expertise et sa détermination, ne pourra analyser que partiellement les résultats. Très peu d'outils sont capables de prendre en charge les données MS/MS de fragmentation d'une molécule cyclique contrairement aux algorithmes utilisés en protéomique où des peptides linéaires issus de digestions enzymatiques sont séquencés en MS/MS et où les interrogations en base de données sont réalisées en routine depuis plusieurs années.

Afin de tendre vers des outils d'annotation automatique d'analyse des spectres MS/MS de molécules plus complexes, nous nous sommes intéressés à d'autres approches.

7.2. Fragmentation *in silico*

L'approche *in silico* consiste à fragmenter de manière théorique la molécule en utilisant des règles de fragmentation. Dans notre étude nous avons utilisé la fragmentation par dissociation induite par collision (CID). Cette approche est valable uniquement lorsque la molécule fragmentée est connue. Par conséquent son utilité reste limitée dans des approches de dépistage ou des approches non ciblées. Cette fragmentation théorique peut être réalisée manuellement ou à l'aide de logiciel dédié comme « iSNAP fragmenter » (Tableau 4).

Tableau 4. Information du logiciel *iSNAP fragmenter* après interrogation depuis le SMILES d'un variant linéaire de surfactine (nC14) avec un premier point de rupture sur la liaison ester (liaison entre l'acide carboxylique de la leucine et la chaine aliphatique). Les informations fournies sont le m/z monoisotopique, la masse moléculaire (m) correspondante, la charge (non présentée ici) et le SMILES. La colonne d'annotations des ions a ensuite été ajoutée manuellement.

Ion	m/z	Masse	SMILES
y1-H2O	98,0242	97,0164	O CCC(N)C O
	100,0762	99,0684	O=CC(N)C(C)C
	112,0398	111,0320	O CCCC(N)C O
	114,0918	113,0840	O=CC(N)CC(C)C
y1	116,0347	115,0269	O CC(N)CC(O)O
	116,1075	115,0997	O=CC(N)CC(C)C
	130,0504	129,0426	O CC(N)CCC(O)O
	197,0925	196,0847	O=CCC(C=O)NC(=O)C(N)C(C)C
b1-H2O	209,1904	208,1826	O CCCCCCCCCCCCC
y2-H2O	211,1082	210,1004	O=CCC(N)C(=O)NC(C=O)CC(C)C
	213,1602	212,1524	O CC(NC(O)C(N)CC(C)C)C(C)C
	215,1031	214,0953	O=CC(NC(=O)C(N)C(C)C)CC(=O)O
	225,1238	224,1160	O CCCC(N)C(O)NC(C O)CC(C)C
b1	227,1758	226,1680	O=CC(NC(=O)C(N)CC(C)C)CC(C)C
	227,2010	226,1932	O CCC(O)CCCCCCCCCC
y2	229,1187	228,1109	O=CC(NC(=O)C(N)CC(=O)O)CC(C)C
	229,1915	228,1837	O CC(NC(O)C(N)CC(C)C)CC(C)C
	243,1344	242,1266	O=CC(NC(=O)C(N)CCC(=O)O)CC(C)C
	310,1766	309,1688	O CCC(C O)NC(O)C(NC(O)C(N)CC(C)C)C(C)C
	310,1766	309,1688	O=CCC(NC(=O)C(N)C(C)C)C(=O)NC(C=O)CC(C)C
y3-H2O	326,2078	325,2000	O CCC(N)C(O)NC(C(O)NC(C O)CC(C)C)CC(C)C
	326,2442	325,2364	O=CC(NC(=O)C(NC(=O)C(N)CC(C)C)CC(C)C)C(C)C
	328,1871	327,1793	O CC(NC(O)C(NC(O)C(N)CC(C)C)C(C)C)CC(O)O
	328,1871	327,1793	O=CC(NC(=O)C(NC(=O)C(N)C(C)C)CC(=O)O)CC(C)C
b2-H2O	338,2079	337,2000	O CCCC(N)C(O)NC(C(O)NC(C O)CC(C)C)CC(C)C
	338,2329	337,2251	O=CCCC(C=O)NC(=O)CC(O)CCCCCCCCCCC
	338,2329	337,2251	O CC(NC(O)CCCCCCCCCCCCCC)CCC(O)O
y3	344,2184	343,2106	O=CC(NC(=O)C(NC(=O)C(N)CC(=O)O)CC(C)C)CC(C)C
b2	356,2184	355,2106	O CC(NC(O)C(NC(O)C(N)CCC(O)O)CC(C)C)CC(C)C
	356,2435	355,2357	O=CC(NC(=O)CC(O)CCCCCCCCCCCC)CCC(=O)O
	423,2606	422,2528	O CCC(NC(O)C(NC(O)C(N)CC(C)C)C(C)C)C(O)NC(C O)CC(C)C
	423,2606	422,2528	O=CCC(C=O)NC(=O)C(NC(=O)C(NC(=O)C(N)CC(C)C)CC(C)C)C(C)C
y4-H2O	425,2762	424,2684	O CCC(NC(O)C(N)C(C)C)C(O)NC(C(O)NC(C O)CC(C)C)CC(C)C
	437,2762	436,2684	O=CCCC(N)C(=O)NC(C(=O)NC(C(=O)NC(C=O)C(C)C)CC(C)C)CC(C)C
	441,2711	440,2633	O CC(NC(O)C(NC(O)C(NC(O)C(N)CC(C)C)C(C)C)CC(O)O)CC(C)C
	441,2711	440,2633	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(N)CC(C)C)CC(C)C)C(C)C)CC(=O)O
y4	443,2867	442,2789	O CC(NC(O)C(NC(O)C(NC(O)C(N)C(C)C)CC(O)O)CC(C)C)CC(C)C
b3-H2O	451,3170	450,3091	O=CCCC(NC(=O)CC(O)CCCCCCCCCC)C(=O)NC(C=O)CC(C)C
	451,3170	450,3091	O CC(NC(O)C(NC(O)CCCCCCCCCCCCC)CCC(O)O)CC(C)C
	455,2867	454,2789	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(N)CCC(=O)O)CC(C)C)CC(C)C)C(C)C
b3	469,3276	468,3197	O CC(NC(O)C(NC(O)CC(O)CCCCCCCCCCC)CCC(O)O)CC(C)C
	536,3446	535,3368	O=CCC(NC(=O)C(NC(=O)C(NC(=O)C(N)CC(C)C)CC(C)C)C(C)C)C(=O)NC(C=O)CC(C)C
y5-H2O	538,3602	537,3524	O CCC(NC(O)C(NC(O)C(N)CC(C)C)C(C)C)C(O)NC(C(O)NC(C O)CC(C)C)CC(C)C
	552,3031	551,2953	O=CCC(C=O)NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)CCC(=O)O)CC(C)C)CC(C)C)C(C)C
	552,3031	551,2953	O CCCC(N)C(O)NC(C(O)NC(C(O)NC(C(O)NC(C O)CC(O)O)C(C)C)CC(C)C)CC(C)C
	554,3551	553,3473	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)CC(C)C)CC(C)C)C(C)C)CC(=O)O)CC(C)C
y5	556,3707	555,3629	O CC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(N)CC(C)C)C(C)C)CC(O)O)CC(C)C)CC(C)C
b4-H2O	564,4010	563,3932	O=CCCC(NC(=O)CC(O)CCCCCCCCCC)C(=O)NC(C(=O)NC(C=O)CC(C)C)CC(C)C
	564,4010	563,3932	O CC(NC(O)C(NC(O)C(NC(O)CCCCCCCCCCCCC)CCC(O)O)CC(C)C)CC(C)C
	570,3137	569,3059	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)CCC(=O)O)CC(C)C)CC(C)C)C(C)C)C(C)C)CC(=O)O
b4	582,4116	581,4038	O CC(NC(O)C(NC(O)C(NC(O)CC(O)CCCCCCCCCCC)CCC(O)O)CC(C)C)CC(C)C
y6-H2O	651,4442	650,4364	O=CCC(NC(=O)C(NC(=O)C(NC(=O)C(N)CC(C)C)CC(C)C)C(C)C)C(=O)NC(C(=O)NC(C(=O)NC(C=O)CC(C)C)CC(C)C
b5-H2O	663,4694	662,4616	O CCCC(NC(O)CC(O)CCCCCCCCCC)C(O)NC(C(O)NC(C(O)NC(C O)C(C)C)CC(C)C)CC(C)C
	663,4694	662,4616	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)CCCCCCCCCCCC)CCC(=O)O)CC(C)C)CC(C)C)C(C)C
	665,3872	664,3793	O CCC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(N)CCC(O)O)CC(C)C)CC(C)C)C(C)C)C(O)NC(C O)CC(C)C
	665,3872	664,3793	O=CCCC(N)C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C=O)CC(C)C)CC(=O)O)C(C)C)CC(C)C)C(C)C
y6	669,4548	668,4470	O CC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(NC(O)C(N)CC(C)C)CC(C)C)C(C)C)CC(O)O)CC(C)C)CC(C)C
b5	681,4799	680,4721	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)CC(O)CCCCCCCCCCC)CCC(=O)O)CC(C)C)CC(C)C)C(C)C
	683,3977	682,3899	O CC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(NC(O)C(N)CCC(O)O)CC(C)C)CC(C)C)C(C)C)CC(O)O)CC(C)C
	778,4963	777,4885	O=CCC(C=O)NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)CC(O)CCCCCCCCCCC)CCC(=O)O)C(C)C)CC(C)C)C(C)C
b6-H2O	778,4963	777,4885	O CCCC(NC(O)CC(O)CCCCCCCCCC)C(O)NC(C(O)NC(C(O)NC(C(O)NC(C O)CC(O)O)C(C)C)CC(C)C)CC(C)C
	778,4963	777,4885	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)CCCCCCCCCCC)CCC(=O)O)CC(C)C)CC(C)C)C(C)C)CC(=O)O
y7-H2O	780,4868	779,4789	O CCC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(N)CCC(O)O)CC(C)C)CC(C)C)C(C)C)C(O)NC(C(O)NC(C O)CC(C)C)CC(C)C
	780,4868	779,4789	O=CCCC(N)C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C=O)CC(C)C)CC(=O)O)C(C)C)CC(C)C)CC(C)C
b6	796,5068	795,4990	O CC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(NC(O)CC(O)CCCCCCCCCCC)CCC(O)O)CC(C)C)CC(C)C)C(C)C)CC(O)O
y7	798,4973	797,4895	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)CCC(=O)O)CC(C)C)CC(C)C)C(C)C)CC(=O)O)CC(C)C)CC(C)C
b7-H2O	891,5803	890,5725	O CCC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(NC(O)CC(O)CCCCCCCCCCC)CCC(O)O)CC(C)C)CC(C)C)C(C)C)C(O)NC(C O)CC(C)C
	891,5803	890,5725	O=CCCC(NC(=O)CC(O)CCCCCCCCCC)C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C(=O)NC(C=O)CC(C)C)CC(=O)O)C(C)C)CC(C)C)CC(C)C
	891,5803	890,5725	O CC(NC(O)C(NC(O)C(NC(O)C(NC(O)C(NC(O)C(NC(O)CCCCCCCCCCCCC)CCC(O)O)CC(C)C)CC(C)C)C(C)C)CC(O)O)CC(C)C
b7	909,5908	908,5830	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)CC(O)CCCCCCCCCC)CCC(=O)O)CC(C)C)CC(C)C)C(C)C)CC(=O)O)CC(C)C

Le nombre de fragments retrouvés par iSNAP s'élève à 70. Le tableau présente plusieurs masses redondantes à cause de la structure cyclique et palindromique (LLDLL) d'une partie du corps peptidique de la surfactine. La plupart des ions annotés sont les ions b (n=7) et y (n=7) ainsi que les ions correspondant à ces mêmes ions moins une molécule d'eau : b-H_20 et y-H_20 (n=14, le logiciel recherche également des ions avec des pertes d'ammonium : b-NH_4 et y-NH_4). Ici, c'est une forme linéaire de la surfactine avec ses huit monomères (7 acides aminés + l'acide gras) qui est utilisé, ce qui implique sept fragmentations théoriques possibles (sept liaisons entre deux monomères) et prenant en compte les deux séries d'ions. Le nombre de masses théoriques s'élève alors à quatorze. La structure de la surfactine est cyclique. Depuis 2017, le logiciel iSNAP est capable de retraiter des SMILES de structure cyclique mais malheureusement le nombre de fragments est plus important avec parfois des aberrations car il prend en compte plusieurs premiers points de rupture et effectue ensuite la fragmentation pour chacun (soit n x (n-1) avec n = nombre de monomères). En bref, la surfactine sous forme cyclique étant composée de huit monomères d'acides aminés, il existe huit points de rupture (lorsqu'on cyclise, on reforme une liaison peptidique qui va se fragmenter en MS/MS) soit huit structures linéaires possibles que le logiciel va ensuite fragmenter de manière théorique. 54 fragments sont donc en théorie possibles (8 x 7=54). Ce nombre est alors doublé en raison de la nature y et b des ions pouvant être générés (54x2 séries d'ion (b et y)). *In fine*, c'est 108 masses théoriques qu'il faudra considérer. Cependant, sur ces 108 fragments, il y a un nombre important de redondances à cause de la fréquence d'apparition de la leucine et de l'isoleucine (même masse) dans la structure de la surfactine (4 monomères sur les 8). Le logiciel calcule également le nombre potentiel de pertes d'eau ou d'ammonium, le nombre de fragments théoriques s'envole alors rapidement.

Pour un lipopeptide comme la surfactine (nC14), le premier point de rupture de la structure parait plus évident car elle possède une liaison ester plus fragile qu'une liaison peptidique dans sa structure. Cependant, il existe d'autres lipopeptides ne possédant pas de liaison ester mais une liaison amide (par exemple la mycosubtiline), c'est-à-dire entre la fonction carboxylique portée par le carbone α de la chaine aliphatique et la fonction amide portée par le carbone α de l'acide aminé. Par conséquent, plusieurs points de rupture potentiels apparaissent. Evidemment, dans la pratique il existe des fragmentations préférentielles mais il est difficile de les anticiper et parfois même de les comprendre.

En conclusion cet outil bioinformatique peut être très utile pour l'annotation de spectres MS/MS mais ne peut pas réellement être utilisé dans le cas de composés inconnus. En effet, il existe une forte probabilité d'obtenir un nombre important de résultats de type faux positifs car le nombre de fragments théoriques est tellement élevé qu'il est statistiquement possible de le faire correspondre avec des signaux du spectre MS/MS et obtenir de bonnes identifications. Par conséquent, nous avons fait le choix d'explorer l'utilisation des défauts de masse de Kendrick pour l'interprétations des données MS/MS afin de nous permettre d'obtenir précisément au moins la formule brute des fragments à partir des données de MS/MS. Ces formules brutes des fragments seront ensuite une précieuse aide dans l'interprétation des spectres MS/MS de NRPs.

7.3. Logiciel dédiés au séquençage *de novo* : Cyclobranch

Un logiciel d'interprétation des spectres de masse MS/MS issus de la fragmentation de NRPs a été mis à disposition en 2015 par l'équipe de Vladimír Havlíček (Novák *et al.*, 2015). Il est dédié au séquençage *de novo* des NRPs. Il n'intègre que 344 monomères (contre 540 recensés dans la base de données NORINE) dans son architecture et recherche ainsi les incréments de masse entre les signaux pour les faire coïncider avec les masses des monomères en mémoire. Nous avons réalisé plusieurs interrogations sur ce logiciel, notamment avec un spectre MS/MS de surfactine nC14 (Figure 55).

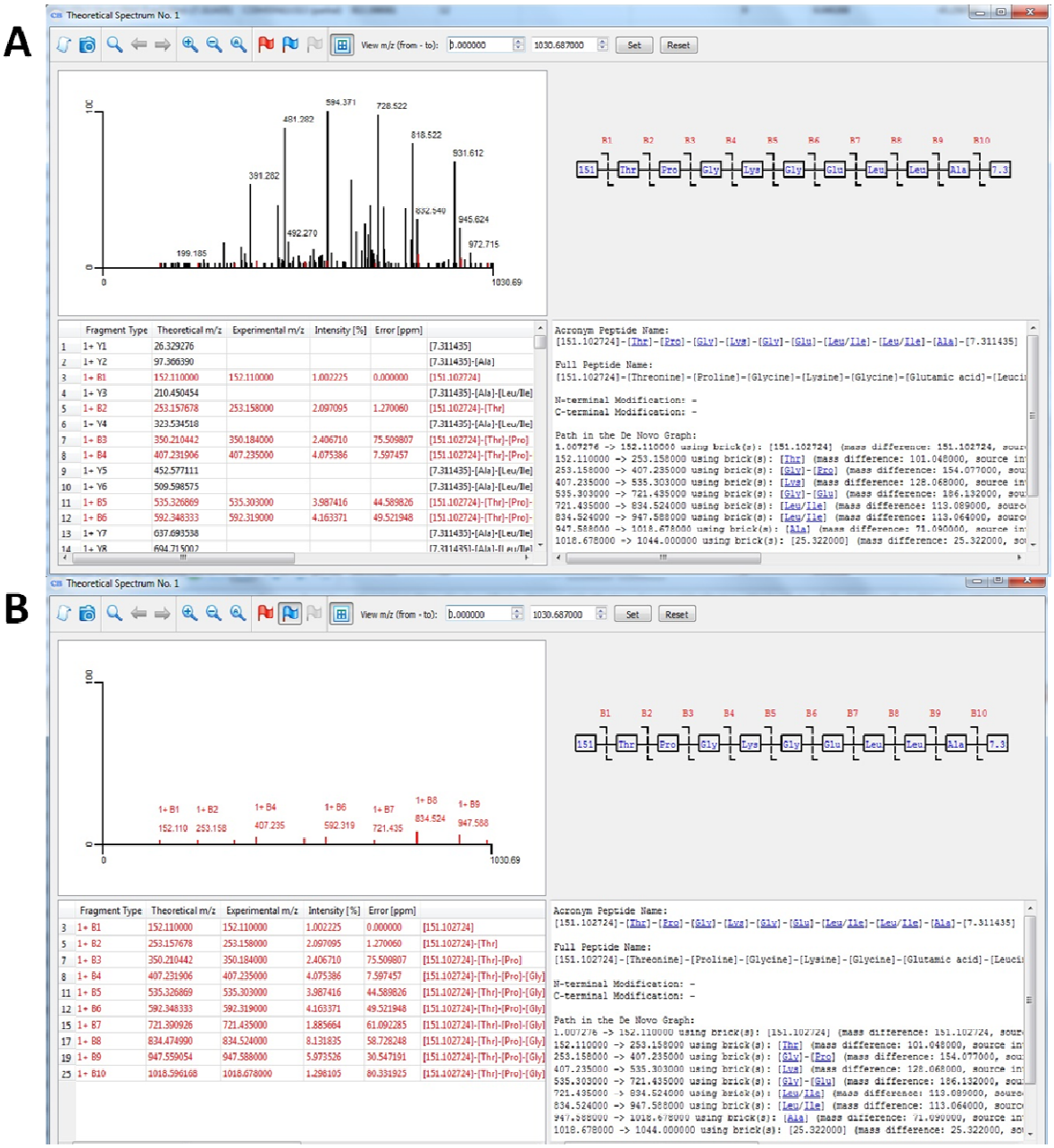


Figure 55. Impression écran des résultats obtenus sur Cyclobranch. A : résultat obtenu affichant tous les signaux du spectre de fragmentation ; B : résultat obtenu pour les ions capables d'expliquer la séquence monomérique proposé par Cyclobranch.

La Figure 55 A, montre le résultat obtenu pour la surfactine (nC14), avec sur le panneau de gauche, le spectre d'interrogation avec en rouge les signaux correspondant à la séquence proposée par le logiciel et en noir les pics non expliqués par le logiciel. Juste en dessous est présenté le tableau de fragmentation avec les annotations. Le panneau de droite présente la séquence monomérique déduite des ions b et y correspondant aux signaux de masse présents sur le spectre. La Figure 55 B, est issue de la même recherche mais les ions

non attribués ont été retirés. Cyclobranch propose une structure linéaire contenant 10 monomères correspondant aux dix ions b présents dans le spectre MS/MS. Le logiciel se contente d'expliquer au mieux le spectre avec les monomères à sa disposition. De plus, Il n'attribue pas de facteur permettant la prise en compte des signaux les plus intenses. De même, aucun score n'est proposé en fonction des signaux attribués ou non. Par conséquent, le nombre de résultats d'attribution révèle un nombre de faux positifs potentiellement important et il n'est pas possible de réaliser une recherche en prenant en compte l'intensité des ions. Par conséquent il est indispensable de préparer des données déjà filtrées en éliminant les ions les moins intenses.

L'analyse de la base de données de monomères du logiciel fait ressortir que tous les monomères possèdent la bonne masse d'incrémentation c'est-à-dire la masse du monomère moins une molécule d'eau mais en revanche aucune masse sous forme protonée ($[M+H]^+$). De plus le temps de calcul, même pour des données de petite taille, reste important et ne donne pas toujours de résultats.

7.4. Combinaison de la fragmentation *in silico* avec les spectres de fragmentation MS/MS : NRPro

Un logiciel, appelé NRPro, permettant de combiner une fragmentation *in silico* avec des spectres de fragmentation a été développé en Java (back-end) et Java script, HTML et CSS (front-end) en collaboration avec Maude Pupin du laboratoire CRIStAL et Emma Ricart, doctorante au SIB (Genève, Suisse) sous la supervision de Frédérique Lisacek, responsable de l'équipe de protéomique informatique au SIB. Le logiciel permet de visualiser et de comparer un spectre de masse MS/MS obtenu par fragmentation CID *in silico* de NRPs présents dans la base de données NORINE avec un spectre de fragmentation MS/MS mesuré obtenu expérimentalement. L'interrogation est accessible depuis un spectre de masse MS/MS au format mgf ou mzXML.

m/z	Intensity	Charge	Mass	ppm	Peptide	Neutral Losses
199.14419555664062	12050433	1	198.13682783	0.46006353606418393	Ala, NMe-Leu	NA
241.19114685058594	14157669	1	240.18377802199998	0.3844292081807129	NMe-Leu, NMe-Val	NA
251.17547607421875	7344716	1	250.168127958	0.2866861865328776	Abu, NMe-Bmt	H2O

Figure 56. Interface utilisateur du logiciel NRPro après soumission d'un spectre MS/MS de cyclosporine A.

L'interface est composée de trois fenêtres principales avec à gauche la représentation semi développée de la molécule interrogée, dans notre cas, la cyclosporine A (Figure 56). Sur cette fenêtre chaque monomère est affiché avec une couleur différente afin de mieux visualiser la composition monomérique des NRPs. La fenêtre de droite présente le spectre MS/MS obtenu expérimentalement avec en vert, les signaux indiquant une correspondance de m/z entre les spectres théoriques et expérimentaux et en bleu les signaux ne correspondant pas aux m/z théoriques obtenus par fragmentation *in silico* en mode CID. La fenêtre du bas est composée d'un tableau avec les m/z, les intensités, la charge (z), la masse, l'erreur en ppm, la séquence du NRP et les pertes de molécules neutres.

Le logiciel est encore en cours de développement, par conséquent il n'est pas accessible. Dans le cadre du programme européen Hubert Curien, Germaine de Staël, obtenu conjointement entre l'institut Charles Viollette de Lille et l'équipe de Frédérique Lisacek du SIB de Genève pour l'année 2018, nous avons pu générer une importante quantité de données de fragmentation de molécules de type NRPs qui constituera à terme le début d'une base de données spectrales accessible depuis NRPro.

7.5. Le défaut de masse pour l'interprétation des spectres de fragmentation MS/MS

Les défauts de masse sont utilisés pour retrouver la formule moléculaire des composées à partir d'une masse exacte obtenue sur un spectre HRMS. Il semble dès lors cohérent de pouvoir obtenir la même information à partir de données de masse MS/MS issues de la fragmentation d'ions précurseurs (Figure 57).

Le spectre de fragmentation MS/MS de la surfactine iC14 est réalisé en mode CID sur un spectromètre de masse FT-ICR (Figure 57 A). Les masses des ions fragments générés ont été extraits afin de recalculer leur RKMD et NKM avant de les tracer sur la carte de NORINE pour déterminer leur formule moléculaire (Figure 57).

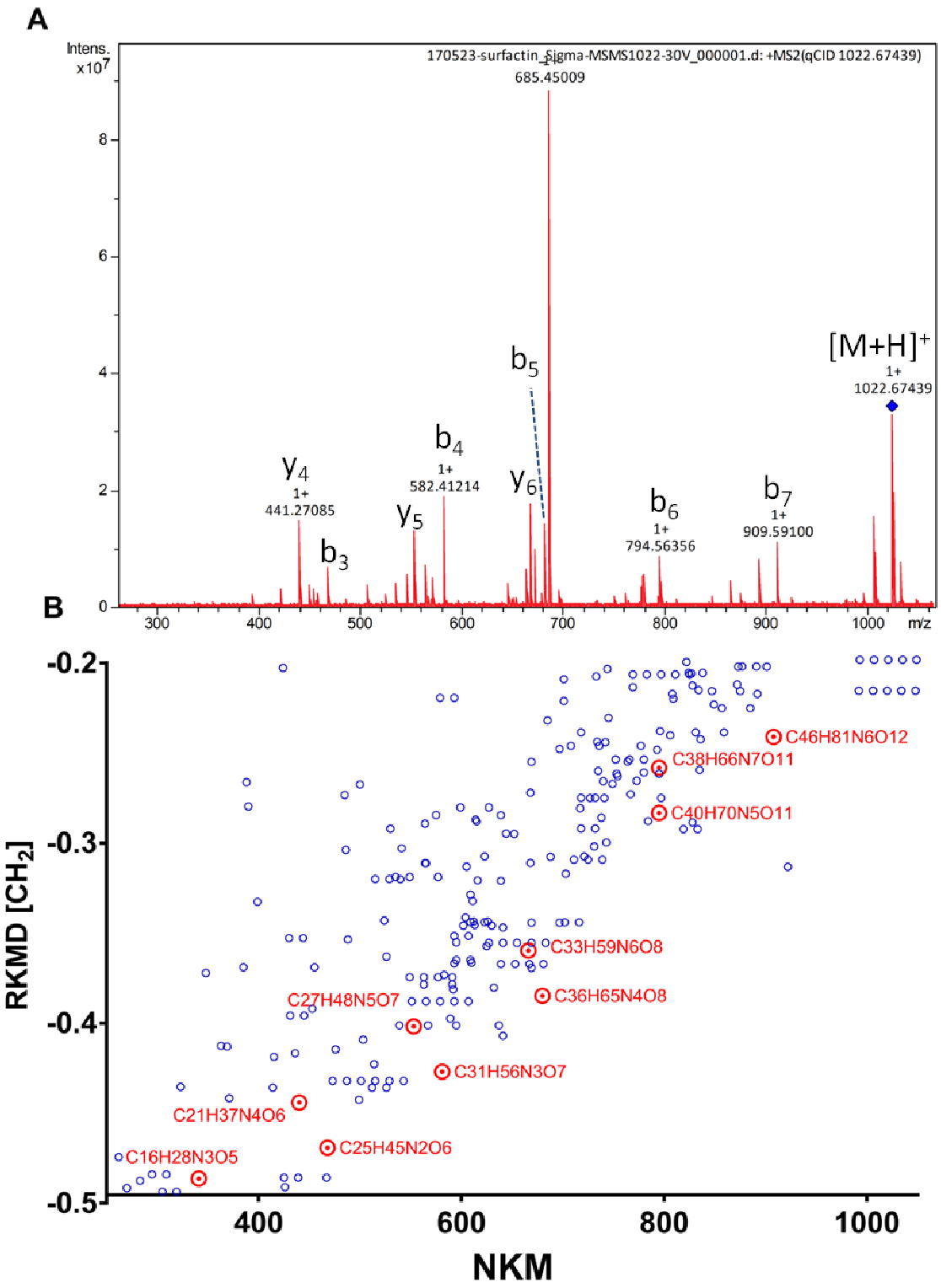


Figure 57. Détermination d'une formule moléculaire. En A : le spectre FT-ICR de fragmentation de la surfactine à 30V ; en B : La carte de Kendrick corrigée avec en bleu les molécules de la base de données NORINE et en rouge les ions issus de la fragmentation de la surfactine cyclique nC14 (Sigma-Aldrich).

Le spectre de fragmentation de la surfactine nC14 de masse moléculaire 1022,67439 Da a été réalisé sur un spectromètre de masse de type FT-ICR (Figure 57 A). Les ions fragments couvrent une gamme de m/z comprise entre 400 et 1022 (Figure 57 A). Après extraction des masses présentes sur ce spectre, le logiciel « *Kendrick formula predictor* » a permis l'attribution des formules moléculaires des signaux mesurés (Figure 57 B). Le report des données de défaut de masse issue de la fragmentation sur la carte de NORINE permet de constater qu'aucun des ions fragments (points rouge) n'est référencé dans la base de données (points bleu). En effet, ce fait est logique puisque la carte de NORINE est générée depuis les RKMD et NKM des NRPs intacts, non fragmentés.

Tableau 5. Récapitulatif des ions fragment présents dans le spectre de masse de fragmentation MS/MS avec leur formule moléculaire et le couple de valeur NKM et KMD.

Ions	m/z	NKM	RKMD	Formule moléculaire candidate
y_4	441.27085	440	-0,448582	C21H37N4O6
b_3	469.32749	468	-0,473707	C25H45N2O6
y_5	554.35567	553	-0,406376	C27H48N5O7
b_4	582.41214	581	-0,43150	C31H56N3O7
y_6	667.44011	666	-0,364169	C33H59N6O8
b_5	681.48043	680	-0,389293	C36H65N4O8
b_6	794.56356	795	-0,28780	C40H70N5O11
b_7	909.59100	908	-0,245589	C46H81N6O12

Les formules moléculaires ont permis de retrouver les ions fragments b et y correspondant à la fragmentation de le surfactine nC14 (Tableau 5). Huit signaux ont permis de retrouver la formule moléculaire à l'aide des défauts de masse de Kendrick et de les faire coïncider avec des ions fragments de la surfactine iC14. L'approche du défaut de masse de Kendrick appliquée à des ions fragments (MS/MS) semble donc, sur la base de la démonstration ci-dessus, être aussi efficace qu'à partir de données MS obtenues à partir d'une molécule non fragmentée. Ainsi, il est possible d'utiliser facilement la même stratégie afin d'annoter en termes de formules moléculaires les spectres de masse MS/MS de fragmentation.

7.6. Défaut de masse Kendrick et règles de fragmentation *in silico*

Il est possible d'aller plus loin encore dans l'utilisation des défauts de masse pour l'interprétation des données MS/MS. En effet, en appliquant les règles de fragmentation *in silico*, les séries d'ions b et y sont structurellement distinguables (Figure 58 A). Cette représentation en version linéaire de la surfactine d'intérêt permet de mettre en évidence une différence importante, du point de vue du défaut de masse de Kendrick, entre les deux séries d'ions b et y. En effet, la série d'ions b possède au départ (ions b_1) un défaut en azote à cause de la chaine aliphatique car les ions b_1 ne possèdent pas d'atome d'azote. Par contre, les ions b_2 en possèdent un, les ions b_3, en possèdent deux et ainsi de suite jusqu'aux ions b_7 qui eux possèdent six azotes. En revanche, la série d'ions y possède dès son premier fragment un atome d'azote. C'est à dire que les ions y_1 possèdent un atome d'azote et suivant le même raisonnement les ions y_7 possèdent quant à eux 7 atomes d'azote. Cette différence structurale est mise en exergue dans la Figure 58 B. Dans ce panneau B, les masses des ions b (rouge) et y (bleu) obtenues *in silico* ont été utilisées afin de recalculer leur KMD avec un groupement CH_2 comme unité de base et leur NKM afin de générer une représentation graphique plus explicite. Afin de simplifier la représentation graphique nous utilisons uniquement les valeurs de KMD et non de RKMD.

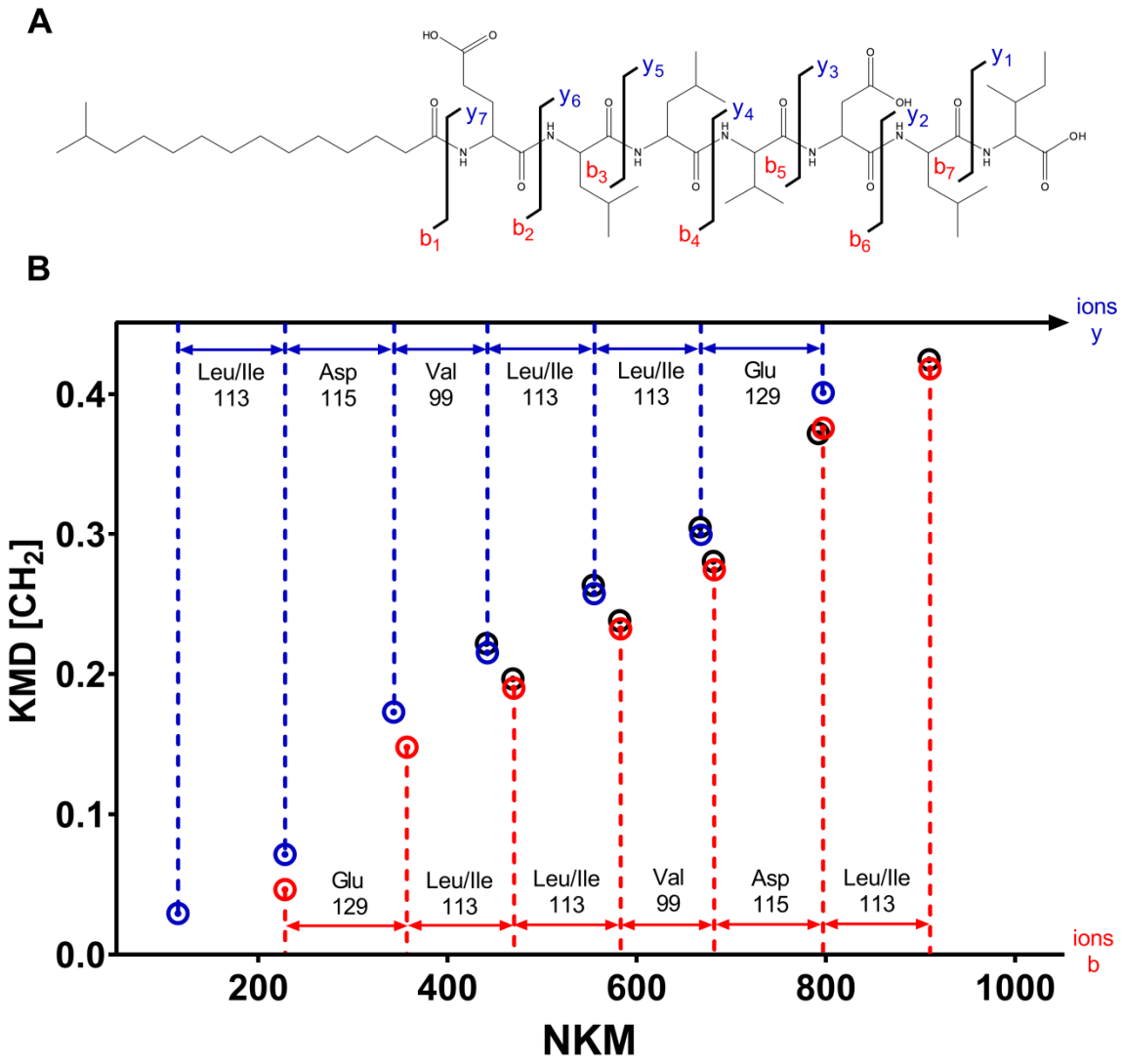


Figure 58. Détermination des ions b et y à l'aide des défauts de masse de Kendrick. En A) structure « ouverte » de la surfactine modèle annotée avec les ions b et y attendus après fragmentation CID. B) carte de Kendrick avec en bleu, les ions théoriques y et en rouge, les ions théoriques b après fragmentation *in silico* et en noir, les ions obtenus expérimentalement.

En effet, tous les ions y possèdent en plus un azote qui entraine un ΔKMD constant de 0,0126 sur toute la gamme de masse. Cette différence permet de distinguer les séries d'ions et donc les ions b des ions y. La Figure 58 B permet de comparer rapidement les ions b et y obtenus *in silico* et les ions mesurés lors de la fragmentation MS/MS de notre surfactine modèle.

In fine, comme nous l'avons évoqué à plusieurs reprises, les NRPs sont parfois composés en partie d'acides aminés mais également d'une autre partie plus variable qui est en général un monomère plus « exotique ». Dans cette partie variable, très peu de monomères possèdent un azote. C'est le cas par exemple pour la chaine aliphatique des lipopeptides, pour les glycannes des glycopeptides ou encore le chromophore des chromopeptides.

Actuellement, ce monomère « plus exotique » crée une réelle contrainte d'interprétation des spectres MS/MS et rend le travail d'identification plus complexe. De plus, il est un frein au transfert des outils bioinformatiques de protéomique au développement d'outils bioinformatiques dédiés à l'interprétation structurale des spectres MS/MS des NRPs. Nous proposons donc d'utiliser cette contrainte comme un atout nous permettant de distinguer plus surement les ions entre eux afin de faciliter l'interprétation des spectres MS/MS de NRPs et de caractériser ainsi avec plus de certitude la structure moléculaire des NRPs. Cette piste pourrait permettre la mise en place d'un outil (logiciel) d'annotations de spectres MS/MS.

8. Criblage d'une collection de *Pseudomonas*, application du workflow

Une collection de vingt souches du genre *Pseudomonas* a été utilisée pour réaliser une preuve de concept de nos différentes approches dans le but d'identifier rapidement des molécules d'intérêts de type NRPs.

8.1. Identification des souches de la collection de *Pseudomonas* par MALDI-TOF

Dans un premier temps les souches ont été analysées par MALDI-TOF afin de les identifier (Tableau 6). Un log (score) supérieur à 2,3 (associé à la couleur vert foncé) indique une identification hautement probable au niveau du genre et de l'espèce. Un log (score) compris entre 2,0 et 2,3 (couleur vert clair) signifie une identification hautement probable au niveau du genre et probable au niveau de l'espèce. Un log (score) entre 1,7 et 2,0 (couleur jaune) implique seulement une identification probable du genre.

L'analyse a permis d'identifier le genre *Pseudomonas* pour toutes les souches testées. Quatre souches sont identifiées avec un score supérieur à 2,3 signifiant une identification du genre et de l'espèce hautement probable. Quinze autres souche son identifiées avec un score supérieur à 2,0 permettant une corrélation importante du genre et de l'espèce. Une seule souche (9) possède un score compris entre 1,7 et 2,0. Ce score nous permet d'identifier le genre mais pas l'espèce.

Tableau 6. Identification des vingt souches de *Pseudomonas* par MALDI-TOF MS.

#	Genre espèce	Score	#	Genre espèce	Score
1	*Pseudomonas aeruginosa*	2,28	11	*Pseudomonas chlororaphis*	2,19
2	*Pseudomonas chlororaphis*	2,18	12	*Pseudomonas corrugata*	2,19
3	*Pseudomonas alcaligenes*	2,52	13	*Pseudomonas chlororaphis*	2,07
4	*Pseudomonas balearica*	2,25	14	*Pseudomonas putida*	2,27
5	*Pseudomonas putida*	2,46	15	*Pseudomonas putida*	2,09
6	*Bacillus subtilis*	2,19	16	*Serratia quinivorans*	2,27
7	*Pseudomonas corrugata*	2,47	17	*Pseudomonas thivervalensis*	2,28
8	*Pseudomonas fragi*	2,10	18	*Pseudomonas chlororaphis*	2,17
9	*Pseudomonas monteilii*	1,84	19	*Pseudomonas synxantha*	2,27
10	*Pseudomonas proteolytica*	2,29	20	*Pseudomonas monteilii*	2,51

Ce score reflète une similitude des profils MALDI-TOF de référence stockés dans la base de données et des profils MALDI-TOF obtenus expérimentalement. La base de données Bruker possède 171 spectres de référence de souche de *Pseudomonas*. L'analyse des souches présentes dans la base de données, révèle que la souche 9 ne possède pas de spectre de référence permettant son identification de manière optimale.

8.2. Recherche d'activités portées par des lipopeptides et des sidérophores

Les souches de *Pseudomonas* ont été cultivées en milieu LB ou CAA pendant 48 h à 30°C sous agitation avant analyse des surnageants. L'effondrement des gouttes de surnageant de culture de quelques souches est présenté dans la Figure 59.

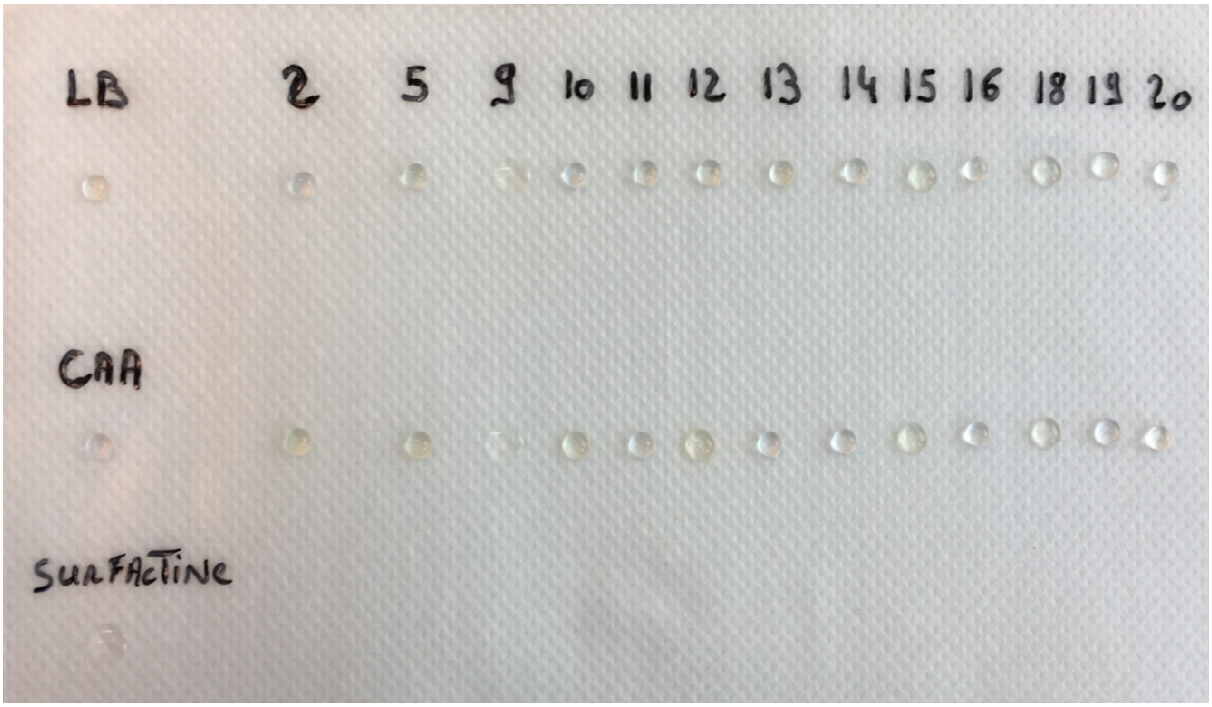


Figure 59. Effondrement de goutte de surnageant de culture après croissance des souches de *Pseudomonas*. A gauche, les témoins négatifs des milieux seuls (LB et CAA) et le témoin positif d'une solution de surfactine à 1 mol/L. En haut au centre, les souches *Pseudomonas* cultivées en milieu LB et en bas au centre, les souches cultivées en milieu CAA.

Cette notion d'effondrement est plus facilement observable à l'œil nu. La photo ne permet pas réellement d'apprécier dans le détail certaines différences. Les surnageants des souches 5, 10, 11, 16 et 20 ne révèlent pas d'effondrement de goutte. Pour les cultures réalisées en milieu LB et CAA, les tests d'effondrements de goutte sont positifs et importants pour les surnageants des souches 2, 9, 12, 13, 14, 15, 18, et 19. Le surnageant de la souche 9 présentant le plus fort effondrement.

Ces résultats indiquent une production potentielle de lipopeptides par les souches de *Pseudomonas* 2, 5, 9, 10, 11, 12, 13, 14, 15, 18, 19 et 20 qu'elles soient cultivées en milieu LB ou en milieu CAA.

La production de sidérophores est évaluée en ajoutant le réactif de Schwyn et Neilands dans une goutte de 10µL de surnageant de culture. La Figure 60 présente les résultats obtenus pour quelques surnageants de culture dans les milieux LB et CAA pour les souches 2, 10, 11, 13, 14, 16, 19 et 20.

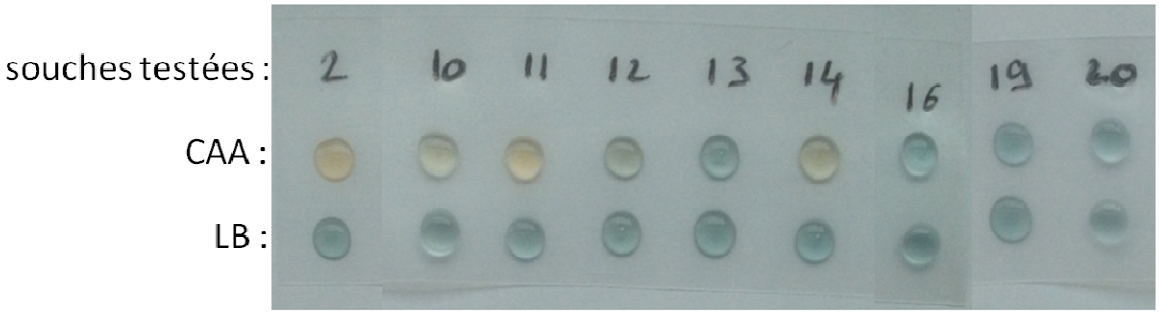


Figure 60. Production de sidérophores révélée par le réactif de Schwyn et Neilands. La ligne du haut correspond aux cultures en milieu CAA et la ligne du bas aux cultures en milieu LB.

Un changement colorimétrique est observé pour les surnageants de culture en milieu CAA obtenus à partir des souches 2, 10, 11, 12 et 14. Aucun changement n'est observé pour les surnageants de culture en milieu CAA des souches 13, 16, 19 et 20. Aucune production de sidérophore n'est observée lorsque les souches sont cultivées en milieu LB.

8.3. Détermination de la formule moléculaire des métabolites secondaires produits par des souches de *Pseudomonas*

8.3.1. Le maillage de Kendrick : application sur culture bactérienne

Les NRPs ne sont pas toujours produits en famille, mais nous montrons ici que le maillage permet d'établir un chemin entre les différents composés produits permettant ainsi d'obtenir une formule moléculaire candidate.

Un mélange complexe comme un surnageant de culture de bactéries contient de nombreux composés différents issus à la fois des milieux de culture eux-mêmes et de la croissance et du métabolisme des microorganismes qui rendent difficile l'identification des métabolites secondaires, en particulier des PKs et NRPs. La souche de *Pseudomonas entomophila* a été cultivée dans du milieu LB ou du milieu CAA à 30°C pendant 48 h. Les cultures ont été centrifugées et les surnageants ont été filtrés pour éliminer les cellules

avant d'être analysés par MALDI-FT-ICR-MS sur un appareil calibré (Figure 61). Les mesures en masse ont permis de déterminer avec précision les masses m/z de 6 composés de masse 1297,51330 ; 1314,53469 ; 1315,52245 ; 1332,58832 ; 1348,58138 et 1721,05720.

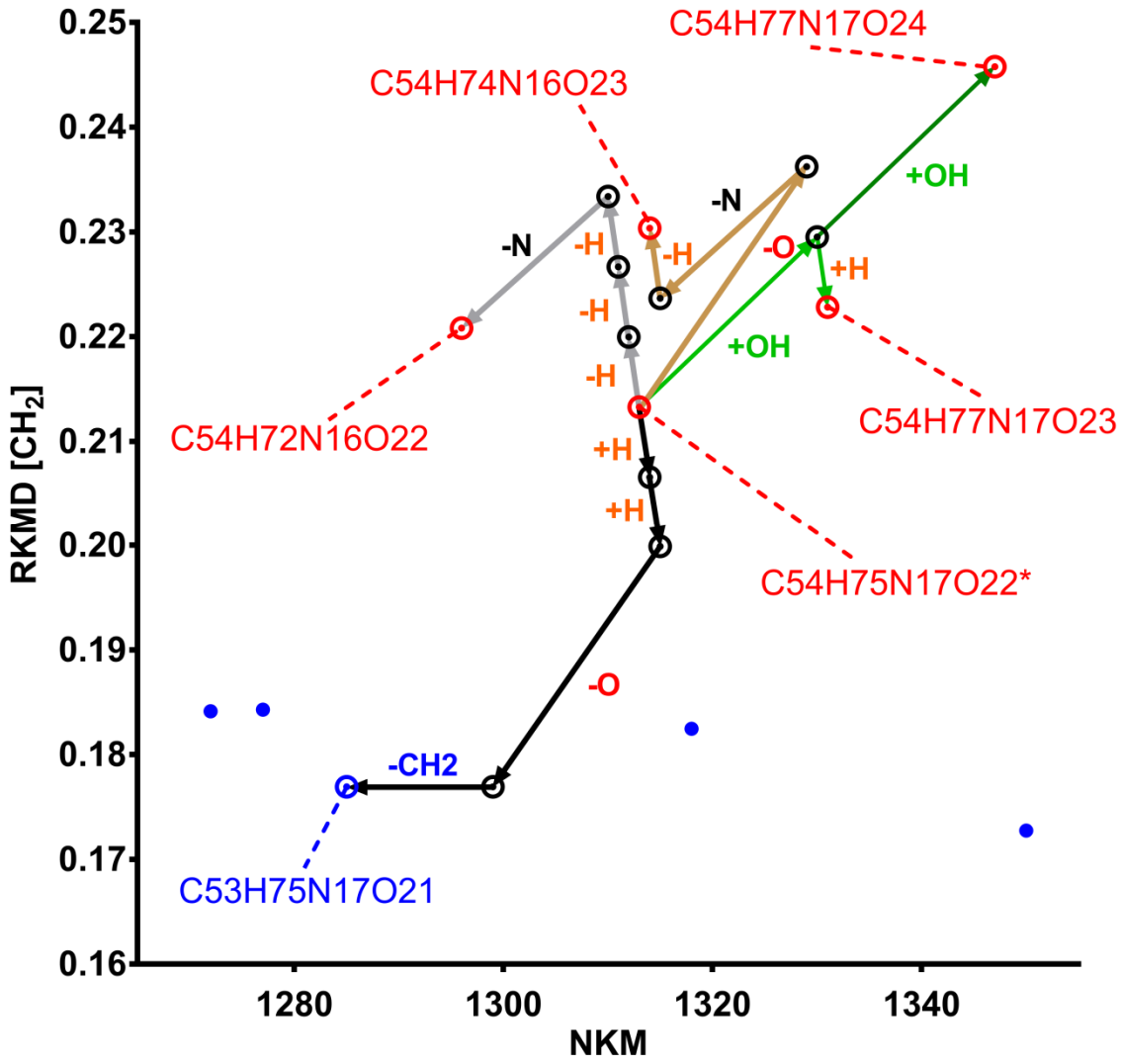


Figure 61. Graphique 2D RKMD/NKM des composés retrouvés dans le milieu de culture de *Pseudomonas entomophila L48* dans la gamme de masse de m/z compris entre 1260 et 1360. Les cercles bleus représentent les composés référencés dans NORINE, les cercles rouges représentent les composés inconnus produits par *P.entomophila* et les points dans les cercles noirs représentent les formules moléculaires intermédiaires du maillage de Kendrick.

Pour plus de clarté, seuls 5 composés inconnus (cercles rouges), ayant respectivement chacun les coordonnées de NKM ; RKMD suivantes : 1313; 0,213287, 1296;

0,220824, 1331; 0,222833, 1314; 0,230370 et 1347; 0,245778, sont représentés sur la Figure 61 qui correspond à la carte NORINE basée sur le calcul de Kendrick. Aucun de ces composés ne correspond à un composé connu (cercles bleus et points bleus) de NORINE. Cependant, il est possible de trouver un chemin vectoriel (flèches de couleur) reliant un point inconnu à un point connu de la carte de NORINE. Pour exemple, à partir du composé central de coordonnées 1313; 0,21287 (point rouge central) et en trois étapes (chemin de la flèche noire), l'ajout de deux atomes d'hydrogène (H), la soustraction d'un atome d'oxygène (O) et d'un groupement CH_2 permet de relier les coordonnées de ce composé inconnu aux coordonnées 1285; 0,176942 du composé de formule moléculaire : $C_{53}H_{75}N_{17}O_{21}$ (en bleu). Par conséquent, la formule moléculaire candidate du composé de coordonnées 1313; 0,21287 est $C_{54}H_{75}N_{17}O_{22}$*. De la même manière, le composé de coordonnées 1296; 0,220824 correspond à la formule moléculaire $C_{54}H_{72}N_{16}O_{22}$ de par la soustraction de trois atomes d'hydrogènes et d'un atome d'azote (N) pour correspondre à $C_{54}H_{75}N_{17}O_{22}$ (chemin de la flèche grise). Le composés de coordonnées 1331; 0,222833 correspond à la formule moléculaire $C_{54}H_{77}N_{17}O_{23}$ par l'ajout d'un groupement hydroxyle (OH) et d'un atome d'hydrogène depuis la formule $C_{54}H_{75}N_{17}O_{22}$. Le composé de coordonnées 1314; 0,230370est issu de la même formule par l'ajout d'un atome d'oxygène et la soustraction consécutive d'un atome d'azote et d'un atome d'hydrogène (chemin de la flèche brune). Enfin, pour retrouver la formule moléculaire du composé de coordonnées 1347; 0,245778, il suffit d'ajouter deux groupements hydroxyles à la formule : $C_{54}H_{75}N_{17}O_{22}$ (chemin de la flèche verte).

De la même manière et comme illustrée par la Figure 62, le point rouge de coordonnées 1719 ; 0,155834 correspondants à la formule brute $C_{81}H_{141}N_{17}O_{23}$ peut être lié, par le trajet de la flèche noire (Figure 62), au composé connu de NORINE de formule brute $C_{79}H_{137}N_{19}O_{22}$ mais également par le chemin de la flèche grise, a un autre composé connu de NORINE de formule brute $C_{81}H_{142}N_{20}0_{21}$. Ainsi, cette méthode nous a permis d'attribuer une seule formule moléculaire à une masse exacte.

D'autre part, cette méthode permet aussi de trier rapidement et de rassembler les composés et leurs adduits ioniques (par exemple Na^+, K^+ ....). Ceci est illustré sur la Figure 62 où le composé 1719; 0,155834 et les deux autres points rouges 1742; 0,192284 et 1758; 0,236183 sont liés entre eux et correspondent aux adduits, respectivement, de sodium

(flèche pourpre) et potassium (flèche brune) du composé de formule brute $C_{81}H_{141}N_{17}O_{23}$ (point 1719 ; 0,155834).

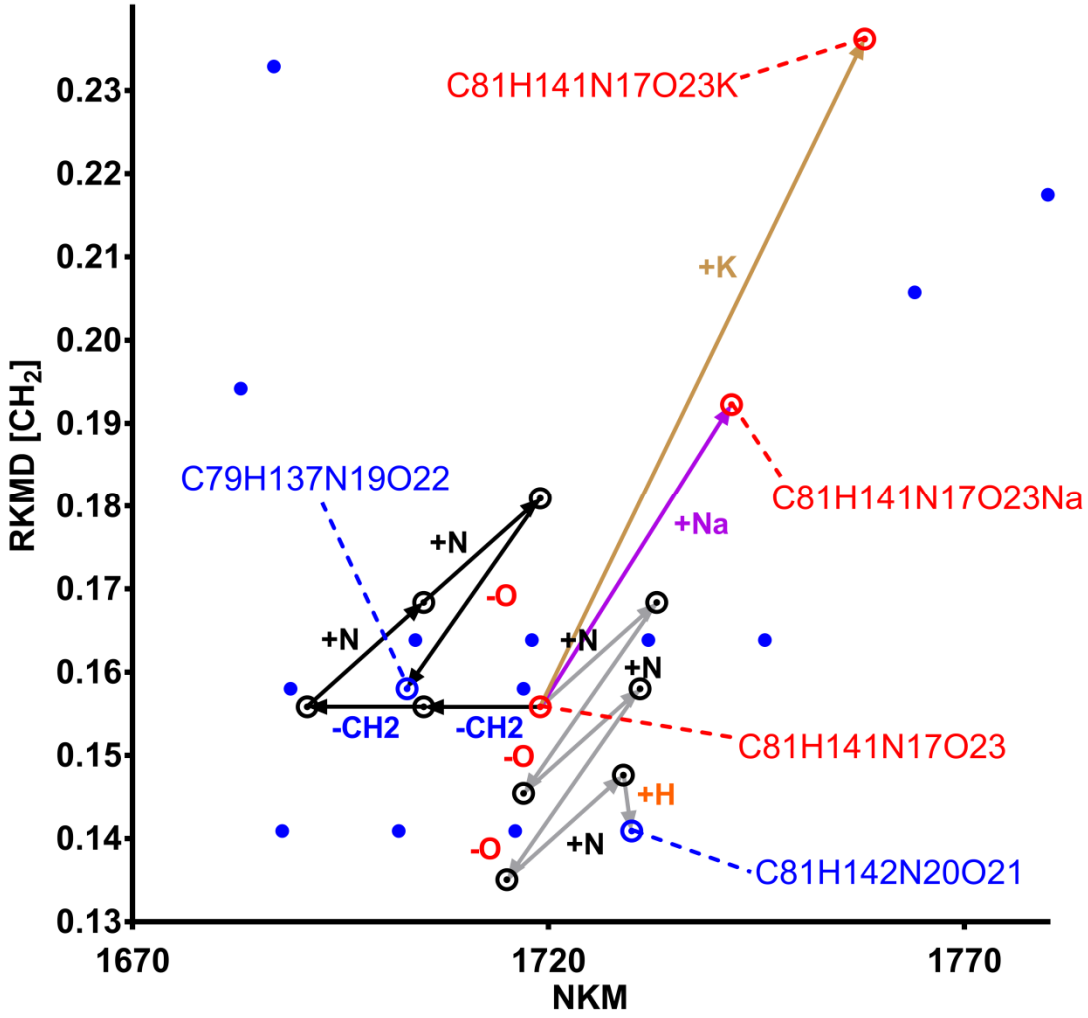


Figure 62. Graphique 2D RKMD/NKM des composés retrouvés dans le milieu de culture de *Pseudomonas entomophila* L48 dans la gamme de masse de m/z compris entre 1670 et 1780. Les cercles bleus représentent les composés référencés dans NORINE, les cercles rouges représentent les composés non référencés dans NORINE et produits par *P.entomophila* et les points dans les cercles noirs représentent les formules moléculaires intermédiaires du maillage de Kendrick.

L'analyse par spectrométrie de masse et l'utilisation des défauts de masse de Kendrick nous ont permis d'identifier la formule moléculaire de cinq composés dans la gamme de masse de m/z 1260-1360 et d'un composé dans la gamme de masse de m/z de 1670-1780 produits pas la souche *Pseudomonas entomophila*. Ces composés (1296,1314,

1331 et 1347) pourraient être des variants de la pyoverdine déjà connu 1313, sidérophore déjà décrit produit par cette même souche (Matthijs *et al.,* 2009). D'autre part, la formule moléculaire du composé 1719 correspond à la celle d'un lipopeptide appelé entolysine déjà retrouvé dans un surnageant de culture de *Pseudomonas* mais de l'espèce *putida* non répertorié dans la base de données NORINE (Li *et al*., 2013). Évidemment la formule moléculaire seule ne permet pas d'affirmer que ce composé est bien l'entolysine. Pour pouvoir l'affirmer il faudrait pourvoir comparer les spectres de fragmentations MS/MS.

8.4. Analyse de Van Krevelen appliquée aux NRPs produits par des souches de *Pseudomonas*

Le criblage à partir de la collection de *Pseudomonas* a été réalisé afin de montrer la pertinence, à la fois de notre maillage de Kendrick pour l'attribution de formules moléculaires mais également la cohérence de l'utilisation d'analyse de Van Krevelen pour la prédiction de classe de nouvelles molécules d'intérêts. Les souches 1, 2, 5, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20 de *Pseudomonas* ont été cultivées dans le milieu LB ou dans le milieu CAA. Les surnageants de culture des différentes espèces de *Pseudomonas* sont analysés par spectrométrie de masse haute résolution afin de déterminer, à l'aide du *Kendrick formula predictor,* les formules moléculaires candidates des deux composés présentant les intensités les plus élevées sur le spectre. Les ratios des molécules présentes dans les surnageants de culture des quatorze souches sont positionnées sur un diagramme de Van Krevelen et les masques précédemment établis sont appliqués (Figure 63).

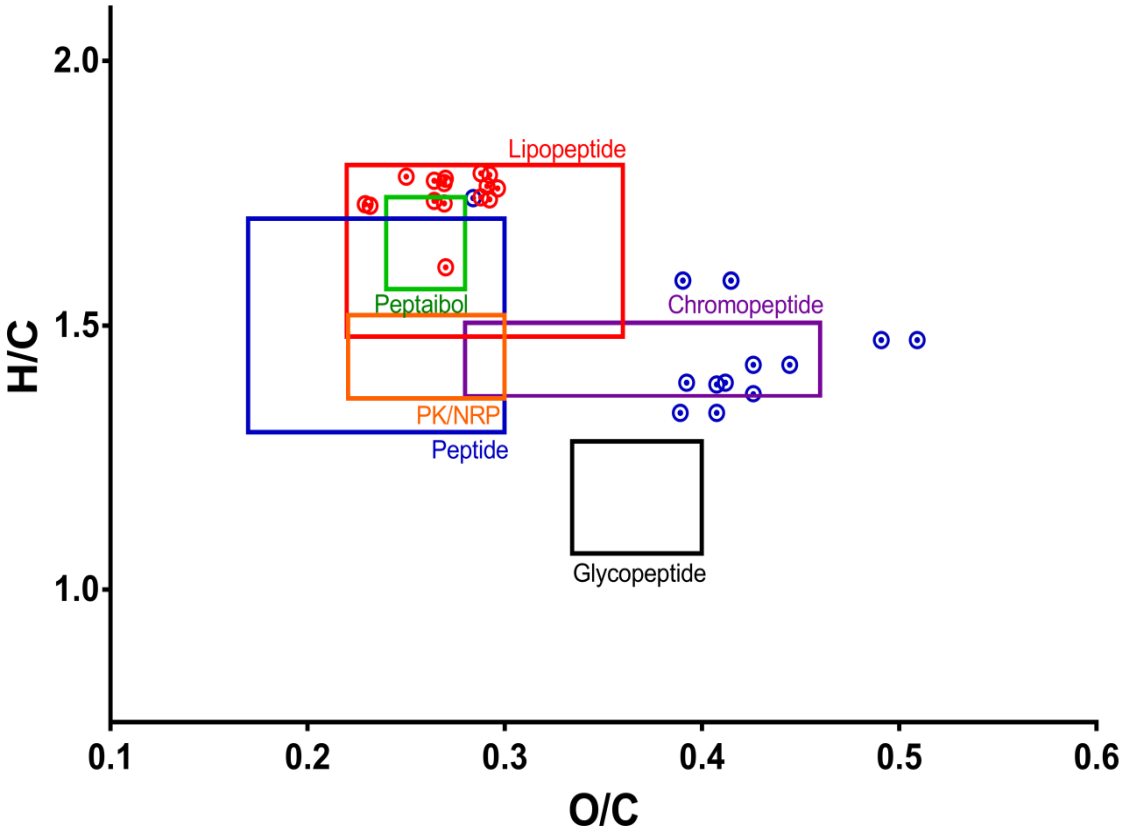


Figure 63. Diagramme de van Krevelen avec les masques de classe moléculaire de NRPs appliqué aux surnageants de culture des souches de *Pseudomonas*. Les ratios H/C et O/C résultant des formules moléculaires candidates issues de l'analyse en masse à haute résolution des surnageants de culture en milieu LB sont mentionnés par des points rouges. Les ratios de formule moléculaire retrouvés dans les surnageants de culture en milieu CAA sont mentionnés par des points bleus.

Douze molécules produites par des souches cultivées dans le milieu LB (souche 5, 13, 15, 16, 18, 19 et 20) et une molécule produite en milieu CAA (souche 9) sont retrouvées dans le masque des lipopeptides. Deux autres molécules retrouvées dans le milieu LB sont dans la zone d'intersection entre les masques lipopeptides et peptaibols (souche 13) et enfin un seul est chevauchant sur le masque lipopeptides, peptaibols et peptides (souche 16). Six molécules issues du milieu CAA sont retrouvées dans le masque des chromopeptides et six autres sont en dehors mais très proches de celui-ci (souche 1, 2, 10, 11, 12 et 14).

Les différences observées à partir des deux milieux de culture démontrent la capacité des *Pseudomonas* à faire varier leur métabolisme en fonction des ressources à disposition et ceci est bien illustré par notre analyse de Van Krevelen.

8.5. Conclusion du criblage

En résumé, le criblage d'activité des lipopeptides et des sidérophores nous ont permis de mettre en évidence la production de plusieurs lipopeptides dans les surnageant de culture de milieu LB pour les souches 5, 9, 13, 15, 16, 18, 19 et 20 et pour le surnageant de milieu de culture CAA de la souche 9 et ceux pour chaque culture réalisée. Nous avons également mis en évidence la production de sidérophores dans les surnageant de culture de milieu CAA des souches 1, 2, 10, 11, 12 et 14. La mesure du m/z, l'utilisation du défaut de masse de Kendrick et du maillage de Kendrick ont permis la détermination des formules moléculaire des composés majoritairement détectés. L'analyse de Van Krevelen coïncide parfaitement avec les tests d'activités réalisés. Les données de fragmentation ont été générées, elles sont essentielles pour le développement d'un logiciel d'interprétation des données MS/MS qui est encore en cours de création au SIB de Genève. Au final, pour les quatorze souches, le criblage à l'aide du workflow analytique a permis d'identifier la formule moléculaire de 16 lipopeptides et de 12 sidérophores potentiels soit 28 composés éventuellement intéressant pour le biocontrôle. Parmi 16 lipopeptides, on retrouve une correspondance en termes de formule moléculaire, pour des molécules de la famille des entolysines pour la souche 9, la famille des putisolvines pour la souche 15, des orfamide et des sessilines pour la souche 18 et des massetolides (viscosines) pour la souche 19, soit potentiellement 5 familles de lipopeptide connues. La formule moléculaire des 11 autres lipopeptides n'ont pas permis de les corréler avec des composés connus. Pour les 12 sidérophores, deux formules correspondent à des sidérophores connus, pour la souche 9 un sidérophores de *P.entomophila* L48 et une formule moléculaire correspondant à la pyocheline produit par la souche 14. Par conséquent, les 10 autres formules moléculaires ne semblent pas correspondent à des formules moléculaire connues. Au final, potentiellement 11 lipopeptides et 10 sidérophores pourrait correspondre à de nouveaux composés.

9. Conclusion Générale

L'objectif de ces travaux était de proposer un workflow d'identification de métabolites secondaires. Ce workflow contient des outils utilisant la spectrométrie de masse pour identifier les microorganismes criblés et l'identification des formules moléculaires des métabolites secondaires (Figure 64).

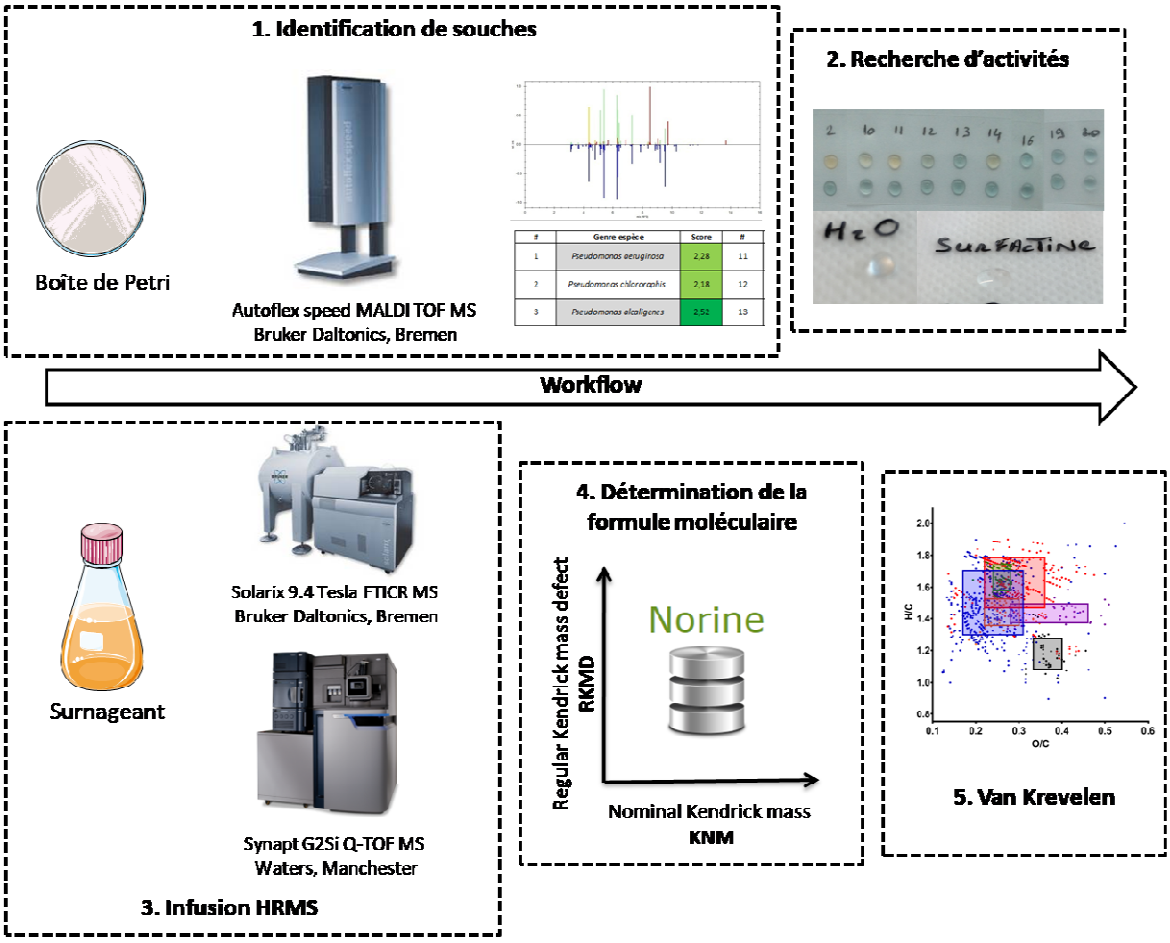


Figure 64. Schématisation des différentes étapes du *workflow* analytique.

Nous avons choisi d'utiliser l'approche des défauts de masse de Kendrick et de l'associer à la base de données NORINE pour le retraitement des informations de masse haute résolution. Cette approche combinatoire qui repose sur la création d'une carte 2D à partir des données de masse référencées dans NORINE ou toutes autres banques de données permet dès lors de réduire le nombre de formules moléculaires candidates à une seule. D'autre part, cette méthode est transposable aux données MS/MS de fragmentation (attribution de formules moléculaires d'ions fragments) et permet également de faciliter l'annotation des spectres de fragmentation en distinguant des séries d'ions spécifiques de

par leur défaut de masse de Kendrick. Ces différences sont alors utilisées dans le cadre du séquençage *de novo* des NRPs afin d'élucider leur structure moléculaire.

Des améliorations restent encore à effectuer comme par exemple des améliorations relatives au mode d'interrogation automatique d'une liste de masse à l'aide d'un fichier mgf ou mzXML via le logiciel *Kendrick formula predictor*. L'amélioration de la résolution spectrale permettra d'utiliser aussi des données issues d'appareil disposant d'un pouvoir de résolution moindre mais juste. Pour finir, Il serait également intéressant d'envisager l'utilisation automatique des défauts de masse de Kendrick pour l'exploitation optimale des données de fragmentation en vue de réaliser des séquençages *de novo* de NRPs.

10. Discussion générale

La deréplication et la recherche de nouvelles molécules au sein de surnageant de culture sont des processus lents et très couteux. Pieter C. Dorrestein a évalué le temps passé à« redécouvrir » et « recaractériser » des molécules déjà connues (Poaeamide B et 3 variants de Bananamide) par son équipe en incluant les coûts d'utilisation des appareils et le temps de travail de son personnel. L'équipe a mis 79 jours et dépensé 38 000 Dollars (soit environ 33000 euros) simplement pour s'orienter vers les 4 molécules présentant une activité intéressante. La détermination et la validation de la structure de la Poaeamide B a demandé 355 jours et couté 86 000 Dollars (74000 euros) contre 90 jours et 25 000 Dollars (21500 euros) pour la détermination des structures des Bananamides 1, 2 et 3 et ce par une équipe expérimentée (Nguyen *et al.*, 2016). Il est donc clair qu'il faut développer des outils permettant d'accélérer les processus de criblage et d'identification de nouveaux composés actifs pour des applications en santé par exemple.

L'identification rapide de souches par phénotypage moléculaire permet de diminuer l'espace de recherche et de s'orienter vers une famille ou une catégorie de NRPs potentiels. Les NRPs sont pour la plupart des molécules sécrétées dans le surnageant de culture. La plupart résulte en effet d'une stratégie mise en place par le microorganisme afin de lui procurer un avantage pour se défendre mais également pour survivre en allant chercher des nutriments manquant dans son environnement proche par l'intermédiaire de la sécrétion de ces molécules. La possibilité de travailler directement sur le sécrétome est un grand

avantage et nous permet d'utiliser des approches analytiques plus directes contrairement à d'autres sources (végétaux, insectes …) où l'obtention de molécules passe par des extractions physiques ou chimiques.

L'utilisation des défauts de masse de Kendrick nous a permis de mettre en place un outil puissant pour l'obtention de formules moléculaires à partir des données de masse HRMS. L'utilisation du concept n'est pas simple et très chronophage à réaliser manuellement. Par conséquent la réalisation d'une interface logicielle nous a permis d'accélérer le processus d'attribution des formules brutes mais également de le rendre accessible à tous les chercheurs, initiés ou non au concept de défaut de masse. Cependant, cette approche est perfectible et doit encore être améliorée en passant par une augmentation de la résolution spectrale (de la représentation graphique) pour rendre la méthode applicable à des données obtenues sur un spectromètre de masse avec un pouvoir de résolution moindre. Il est également envisageable d'introduire une dimension supplémentaire au vecteur en introduisant un second bloc de construction autre que le CH_2 ou encore en utilisant des données de mobilité ionique comme la section efficace de collision (*collision cross section*, CCS). La détermination de la formule moléculaire nous permet rapidement d'interroger les banques de donnée pour obtenir le caractère nouveau ou non du composé. Évidemment plusieurs composés peuvent partager la même formule moléculaire mais les multiples informations récoltées au cours des différentes analyses du workflow comme l'identification de souches permet souvent de conclure sur le caractère nouveau ou non du composé. Le diagramme de Van Krevelen est une représentation graphique très visuelle et rapide à mettre en œuvre mais non utilisée dans la recherche et le dépistage de métabolites secondaires. Nous illustrons ici, la simplicité de mise en œuvre d'une telle représentation pour des informations non négligeables lors de la recherche de nouveaux composés d'intérêts.

Afin d'obtenir d'avantage d'information de structure, les données MS/MS sont cruciales mais pas toujours faciles à interpréter. Les auteurs du logiciel Cyclobranch (Novák *et al.*, 2015), ont récemment réalisé une mise à jour importante. Le logiciel nous permet à présent de choisir un premier point de rupture possible d'une molécule cyclique fragmentée (Novák *et al.*, 2018). Ils ont également ajouté un outil d'attribution de formules brutes en utilisant les massifs isotopiques de la même manière que le logiciel SIRIUS (Böcker *et al.*,

2009). Le logiciel semble présenter des performances bien supérieures à la version de 2015, mais ce dernier étant très récent nous n'avons pas pu réaliser d'avantage de tests. Nous avons montré que l'utilisation des défauts de masse peut s'avérer utile pour l'interprétation et l'annotation de données de fragmentation MS/MS, premièrement par sa capacité à attribuer la formule moléculaire des ions fragments mais également grâce à son pouvoir de différenciation des ions fragments (ions b et y) de NRPs. Cependant, il reste encore un travail important pour rendre la méthode automatisable pour pouvoir l'inclure dans un logiciel. Les outils développés dans le but d'identifier de nouveaux NRPs pour des applications dans différents secteurs pourront être étendus à d'autres métabolites secondaires comme les PKS par exemple.

En aparté, une brève histoire de masse et de microorganismes

On parle souvent de l'incroyable pouvoir d'adaptation de l'être humain, Charles Darwin dans son célèbre ouvrage, l'origine des espèces de 1859 a écrit : *« Les espèces qui survivent ne sont pas les espèces les plus fortes, ni les plus intelligentes, mais celles qui s'adaptent le mieux aux changements*. » L'homme est souvent cité en exemple pour sa capacité d'adaptation, mais pourtant nous ne sommes pas les champions toutes catégories dans cette discipline. Le temps n'est pas universel, il est difficile de s'imaginer que les microorganismes (dans un tube à essai) vivent plus longtemps qu'un être humain en réalité. Il faut exclure le fait que notre reproduction n'est pas la même et que nous sommes des organismes multicellulaires. En tenant compte du temps de génération de chacun, 24 heures de notre précieux temps équivaut à environ un siècle pour une bactérie. Si l'on admet un temps de génération moyen de 40 minutes pour une bactérie, celle-ci est capable de se reproduire 36 fois en 24 heures soit autant d'occasions de muter et de s'adapter. Si l'on se cantonne à comparer uniquement avec l'*homo sapiens*, celui-ci étant présent depuis environ 200 000 ans et possède un temps de génération de 29 ans en moyenne (actuellement et pour la France car les premières générations ont sans doute été longtemps beaucoup plus courtes), nous somme le fruit de 7 000 générations. Incroyable, mais infiniment ridicule face aux bactéries capables de réaliser cette prouesse en moins de six mois. Aujourd'hui, les bactéries (et autres microorganismes) peuvent nous apporter des solutions concrètes dans le domaine de la cosmétique, de l'agroalimentaire, de la santé mais aussi dans le

biocontrôle. Rien n'est plus d'actualité que la menace croissante de la résistance aux antibiotiques. L'organisation mondiale de la santé prévoit 10 millions de décès par an d'ici à 2050, soit plus que les cancers aujourd'hui (8 millions). Autre impasse, les enjeux écologiques, nous sommes dans une société ou les niveaux de production ne cessent de s'accroitre pour répondre à une demande alimentaire toujours croissante. Malheureusement, les techniques de lutte chimique ne sont pas sans conséquences néfastes pour l'environnement et la santé humaine. Un effort majeur doit être engagé pour proposer des solutions de biocontrôle alternatives et efficaces. Le biocontrôle est défini comme un ensemble de méthodes de protection des cultures, basé sur l'utilisation de microorganismes ou de substances naturelles issues de microorganismes telles que les métabolites secondaires. Nous tentons modestement par ces travaux de tendre vers une voie capable d'éviter des situations d'impasse thérapeutique et de proposer des solutions alternatives aux luttes chimiques contre les parasites et ravageurs.

Références bibliographiques

Akiyama, K., Chikayama, E., Yuasa, H., Shimada, Y., Tohge, T., Shinozaki, K. and Saito, K. (2008). PRIMe: A web site that assembles tools for metabolomics and transcriptomics. *In Silico Biology*, *8*(3–4), 339–345.

http://www.scopus.com/inward/record.url?eid=2-s2.0-53249144283&partnerID=tZOtx3y1

Al-Ali, A., Deravel, J., Krier, F., Béchet, M., Ongena, M. and Jacques, P. (2017). Biofilm formation is determinant in tomato rhizosphere colonization by Bacillus velezensis FZB42. *Environmental Science and Pollution Research*, pp. 1–11. http://doi.org/10.1007/s11356-017-0469-1

Almeida, R., Mosoarca, C., Chirita, M., Udrescu, V., Dinca, N., Vukelić, Ž. and Zamfir, A. D. (2008). Coupling of fully automated chip-based electrospray ionization to high-capacity ion trap mass spectrometer for ganglioside analysis. *Analytical Biochemistry*, *378*(1), 43–52. http://doi.org/10.1016/j.ab.2008.03.039

Alon, T. and Amirav, A. (2006). Isotope abundance analysis methods and software for improved sample identification with supersonic gas chromatography/mass spectrometry. *Rapid Communications in Mass Spectrometry : RCM*, *20*(17), 2579–88. http://doi.org/10.1002/rcm.2637

Alon, T. and Amirav, A. (2009). Isotope abundance analysis for improved sample identification with tandem mass spectrometry. *Rapid Communications in Mass Spectrometry : RCM*, *23*(23), 3668–72. http://doi.org/10.1002/rcm.4306

Ambrogelly, A., Palioura, S. and Söll, D. (2007). Natural expansion of the genetic code. *Nature Chemical Biology*. http://doi.org/10.1038/nchembio847

Arezoo, T., Rasoul, S. and Rooha Kasra, K. (2010). *Lactobacillus acidophilus*-derived biosurfactant effect on GTFB and GTFC expression level in Streptococcus mutans biofilm cells. *Brazilian Journal of Microbiology*. http://doi.org/10.1590/S1517-83822011000100042

Balayssac, S., Trefi, S., Gilard, V., Malet-Martino, M., Martino, R. and Delsuc, M. A. (2009). 2D and 3D DOSY1H NMR, a useful tool for analysis of complex mixtures: Application to herbal drugs or dietary supplements for erectile dysfunction. *Journal of Pharmaceutical*

and Biomedical Analysis, *50*(4), 602–612. http://doi.org/10.1016/j.jpba.2008.10.034

Balibar, C. J., Vaillancourt, F. H. and Walsh, C. T. (2005). Generation of D amino acid residues in assembly of arthrofactin by dual condensation/epimerization domains. *Chemistry and Biology*, *12*(11), 1189–1200. http://doi.org/10.1016/j.chembiol.2005.08.010

Baron, S. (1996). *Medical Microbiology*. *Medical Microbiology*. University of Texas Medical Branch at Galveston. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21413252

Beier, T., Häffner, H., Hermanspahn, N., Karshenboim, S. G., Kluge, H.-J., Quint, W. and Werth, G. (2002). New determination of the electron's mass. *Physical Review Letters*, *88*(1), 011603. http://doi.org/10.1103/PhysRevLett.88.011603

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1), 36–42. http://doi.org/10.1093/nar/gks1195

Bereman, M. S., Lyndon, M. M., Dixon, R. B. and Muddiman, D. C. (2008). Mass measurement accuracy comparisons between a double-focusing magnetic sector and a time-of-flight mass analyzer. *Rapid Communications in Mass Spectrometry : RCM*, *22*(10), 1563–6. http://doi.org/10.1002/rcm.3544

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. http://doi.org/10.1093/nar/28.1.235

Bertani, G. (1951). Studies on lysogenesis the mode of phage liberation by lysogenic *Escherichia coli*, 293–300.

Bleakney, W. (1929). A New Method of Positive Ray Analysis and Its Application to the Measurement of Ionization Potentials in Mercury Vapor. *Physical Review*, *34*(1), 157–160. http://doi.org/10.1103/PhysRev.34.157

Böcker, S., Letzel, M. C., Lipták, Z. and Pervukhin, A. (2009). SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, *25*(2), 218–224. http://doi.org/10.1093/bioinformatics/btn603

Böcker, S. and Rasche, F. (2008). Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, *24*(16), 49–55. http://doi.org/10.1093/bioinformatics/btn270

Bode, H. B. and Müller, R. (2005). The impact of bacterial genomics on natural product research. *Angewandte Chemie - International Edition*, *44*(42), 6828–6846. http://doi.org/10.1002/anie.200501080

Bristow, A. W. T. and Webb, K. S. (2003). Intercomparison study on accurate mass measurement of small molecules in mass spectrometry. *Journal of the American Society for Mass Spectrometry*, *14*(10), 1086–1098. http://doi.org/10.1016/S1044-0305(03)00403-3

Bristow, A. W. T., Webb, K. S., Lubben, A. T. and Halket, J. (2004). Reproducible product-ion tandem mass spectra on various liquid chromatography/mass spectrometry instruments for the development of spectral libraries. *Rapid Communications in Mass Spectrometry : RCM*, *18*(13), 1447–54. http://doi.org/10.1002/rcm.1492

Brkljača, R. and Urban, S. (2015). HPLC-NMR and HPLC-MS investigation of antimicrobial constituents in Cystophora monilifera and Cystophora subfarcinata. *Phytochemistry*, *117*, 200–208. http://doi.org/10.1016/j.phytochem.2015.06.014

Brunnée, C. (1967). A Combined Field Ionisation/Electron Impact Ion Source for High Molecular Weight Samples of Low Volatility, *35*, 1966–1968.

Buchan, B. W. and Ledeboer, N. A. (2013). Advances in identification of clinical yeast isolates by use of matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Journal of Clinical Microbiology*, *51*(5), 1359–66. http://doi.org/10.1128/JCM.03105-12

Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P. and Kucherov, G. (2007). NORINE: a database of nonribosomal peptides. *Nucleic Acids Research*, *36*(Database), D326–D331. http://doi.org/10.1093/nar/gkm792

Caly, D. L., Chevalier, M., Flahaut, C., Cudennec, B., Al Atya, A. K., Chataigné, G. and Drider, D. (2017). The safe enterocin DD14 is a leaderless two-peptide bacteriocin with anti-Clostridium perfringens activity. *International Journal of Antimicrobial Agents*.

http://doi.org/10.1016/j.ijantimicag.2016.11.016

Cameron, A. E. and Eggers, D. F. (1948). An Ion ``Velocitron''. *Review of Scientific Instruments*, *19*(9), 605–607. http://doi.org/10.1063/1.1741336

Cech, N. B. and Enke, C. G. (2002). Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews*, *20*(6), 362–87. http://doi.org/10.1002/mas.10008

Chen, H., Fan, B., Petitjean, M., Panaye, A., Doucet, J.-P., Li, F. and Yuan, S. (2003). MASSIS: a mass spectrum simulation system. 2: Procedures and performance. *European Journal of Mass Spectrometry (Chichester, England)*, *9*(5), 445–57. http://doi.org/10.1255/ejms.577

Chen, H., Fan, B., Xia, H., Petitjean, M., Yuan, S., Panaye, A. and Doucet. (2003). MASSIS: a mass spectrum simulation system 1. Principle and method. *European Journal of Mass Spectrometry (Chichester, England)*, *9*(3), 175–86. http://doi.org/10.1255/ejms.549

Chen, T., Kao, M. Y., Tepel, M., Rush, J. and Church, G. M. (2001). A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *8*(3), 325–37. http://doi.org/10.1089/10665270152530872

Clark, H. A. and Jurs, P. C. (1981). Simulation of mass spectral intensities by regression analysis of calculated structural characteristics. *Analytica Chimica Acta*, *132*(C), 75–88. http://doi.org/10.1016/S0003-2670(01)93879-6

Cochrane, G., Karsch-Mizrachi, I. and Takagi, T. (2016). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, *44*(D1), D48–D50. http://doi.org/10.1093/nar/gkv1323

Comisarow, M. B. and Marshall, A. G. (1974). Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters*, *25*(2), 282–283. http://doi.org/10.1016/0009-2614(74)89137-2

Cooper, H. J., Hakansson, K. and Marshall, A. G. (2005). The role of electron capture dissociation in biomolecular analysis. *Mass Spectrometry Reviews*, *24*(2), 201–222.

http://doi.org/10.1002/mas.20014

Coutte, F., Leclère, V., Béchet, M., Guez, J.-S., Lecouturier, D., Chollet-Imbert, M. and Jacques, P. (2010). Effect of *pps* disruption and constitutive expression of *srfA* on surfactin productivity, spreading and antagonistic properties of *Bacillus subtilis* 168 derivatives. *Journal of Applied Microbiology*. http://doi.org/10.1111/j.1365-2672.2010.04683.x

Decoin, M. (2018). Produits phyto de biocontrôle : les innovations fleurissent. *Phytoma*, *713*, 26–32. Retrieved from http://www.phytoma-ldv.com/article-24396-Produits_phyto_de_biocontrole_les_innovations_fleurissent

Dempster, A. J. (1918). A new Method of Positive Ray Analysis. *Physical Review*, *11*(4), 316–325. http://doi.org/10.1103/PhysRev.11.316

Deravel, J., Krier, F. and Jacques, P. (2014). Les biopesticides, compléments et alternatives aux produits phytosanitaires chimiques (synthèse bibliographique). *Biotechnology, Agronomy and Society and Environment*, *18*(2), 220–232.

Dresen, S., Gergov, M., Politi, L., Halter, C. and Weinmann, W. (2009). ESI-MS/MS library of 1,253 compounds for application in forensic and clinical toxicology. *Analytical and Bioanalytical Chemistry*, *395*(8), 2521–6. http://doi.org/10.1007/s00216-009-3084-2

Epstein, S. C., Charkoudian, L. K. and Medema, M. H. (2018). A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. *Standards in Genomic Sciences*, *13*(1), 16. http://doi.org/10.1186/s40793-018-0318-y

Fan, B., Chen, H., Petitjean, M., Panaye, A., Doucet, J., Xia, H. and Yuan. (2005). New Strategy of Mass Spectrum Simulation Based on Reduced and Concentrated Knowledge Databases. *Spectroscopy Letters*, *38*(2), 145–170. http://doi.org/10.1081/SL-200049577

Fenn, J. B. (2003). Electrospray wings for molecular elephants (Nobel lecture). *Angewandte Chemie - International Edition*, *42*(33), 3871–3894. http://doi.org/10.1002/anie.200300605

Ferrer, I. and Thurman, E. M. (2005). Measuring the mass of an electron by LC/TOF-MS: A

study of “twin ions.” *Analytical Chemistry*, *77*(10), 3394–3400. http://doi.org/10.1021/ac0485942

Fischer, E. H., Graves, D. J., Snyder Crittenden, E. R. and Krebs, E. G. (1959). Structure of the Site Phosphorylated in the Phosphorylase b to a Reaction. *The Journal of Biological Chemistry*, *234*(7), 1698–1705.

Flissi, A., Dufresne, Y., Michalik, J., Tonon, L., Janot, S., Noé, L. and Pupin, M. (2016). Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. *Nucleic Acids Research*, *44*(D1), D1113–D1118. http://doi.org/10.1093/nar/gkv1143

Fouquet, T. N. J., Cody, R. B., Ozeki, Y., Kitagawa, S., Ohtani, H. and Sato, H. (2018). On the Kendrick Mass Defect Plots of Multiply Charged Polymer Ions: Splits, Misalignments, and How to Correct Them. *Journal of The American Society for Mass Spectrometry*, 1–16. http://doi.org/10.1007/s13361-018-1972-4

Fouquet, T. and Sato, H. (2017a). Extension of the Kendrick Mass Defect Analysis of Homopolymers to Low Resolution and High Mass Range Mass Spectra Using Fractional Base Units. *Analytical Chemistry*, *89*(5), 2682–2686. http://doi.org/10.1021/acs.analchem.6b05136

Fouquet, T. and Sato, H. (2017b). How to choose the best fractional base unit for a high-resolution Kendrick mass defect analysis of polymer ions. *Rapid Communications in Mass Spectrometry*, *31*(12), 1067–1072. http://doi.org/10.1002/rcm.7868

Fouquet, T. and Sato, H. (2017c). Improving the Resolution of Kendrick Mass Defect Analysis for Polymer Ions with Fractional Base Units. *Mass Spectrometry*, *6*(1), A0055–A0055. http://doi.org/10.5702/massspectrometry.A0055

Fredenhagen, A., Derrien, C. and Gassmann, E. (2005). An MS/MS library on an ion-trap instrument for efficient dereplication of natural products. Different fragmentation patterns for [M + H]+ and [M + Na]+ ions. *Journal of Natural Products*, *68*(3), 385–91. http://doi.org/10.1021/np049657e

Gaskell, S. J. (1997). Special feature : Electrospray : Principles and Practice, *32*(April), 677–

688.

Gay, S., Binz, P.-A., Hochstrasser, D. F. and Appel, R. D. (2002). Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics*, *2*(10), 1374–91. http://doi.org/10.1002/1615-9861(200210)2:10<1374::AID-PROT1374>3.0.CO;2-D

Gelpí, E. (2009). From large analogical instruments to small digital black boxes: 40 Years of progress in mass spectrometry and its role in proteomics. Part II 1985-2000. *Journal of Mass Spectrometry*, *44*(8), 1137–1161. http://doi.org/10.1002/jms.1621

Gerlich, M. and Neumann, S. (2013). MetFusion: integration of compound identification strategies. *Journal of Mass Spectrometry : JMS*, *48*(3), 291–8. http://doi.org/10.1002/jms.3123

Gerritsen, V. B., Barbié, V., Durinx, C. and Beckmann, J. S. (2016). [Bio-informatic and personalized medicine: Switzerland is pioneer]. *Revue Medicale Suisse*, *12*(507), 414–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27039608

Gomez, A. and Tang, K. (1994). Charge and fission of droplets in electrostatic sprays. *Physics of Fluids*, *6*(1), 404–414. http://doi.org/10.1063/1.868037

Good, D. M., Wirtala, M., McAlister, G. C. and Coon, J. J. (2007). Performance characteristics of electron transfer dissociation mass spectrometry. *Molecular & Cellular Proteomics : MCP*, *6*(11), 1942–51. http://doi.org/10.1074/mcp.M700073-MCP200

Grange, A. H., Genicola, F. A. and Sovocool, G. W. (2002). Utility of three types of mass spectrometers for determining elemental compositions of ions formed from chromatographically separated compounds. *Rapid Communications in Mass Spectrometry*, *16*(24), 2356–2369. http://doi.org/10.1002/rcm.842

Grange, A. H. and Sovocool, G. W. (2008). Automated determination of precursor ion, product ion, and neutral loss compositions and deconvolution of composite mass spectra using ion correlation based on exact masses and relative isotopic abundances. *Rapid Communications in Mass Spectrometry : RCM*, *22*(15), 2375–90. http://doi.org/10.1002/rcm.3619

Gravet, A. and Gessier, M. (2013). Spectrométrie de masse et microbiologie. *Immuno-*

Analyse & Biologie Spécialisée, *28*(5–6), 297–308. http://doi.org/10.1016/j.immbio.2013.09.003

Halket, J. M., Waterman, D., Przyborowska, A. M., Patel, R. K. P., Fraser, P. D. and Bramley, P. M. (2005). Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *Journal of Experimental Botany*, *56*(410), 219–243. http://doi.org/10.1093/jxb/eri069

Hegeman, A. D., Schulte, C. F., Cui, Q., Lewis, I. A., Huttlin, E. L., Eghbalnia, H. and Sussman, M. R. (2007). Stable isotope assisted assignment of elemental compositions for metabolomics. *Analytical Chemistry*, *79*(18), 6912–6921. http://doi.org/10.1021/ac070346t

Heinonen, M., Rantanen, A., Mielikäinen, T., Kokkonen, J., Kiuru, J., Ketola, R. A. and Rousu, J. (2008). FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry : RCM*, *22*(19), 3043–52. http://doi.org/10.1002/rcm.3701

Hernandez-Eugenio, G., Fardeau, M. L., Garcia, J. L. and Ollivier, B. (2015). *Bergey's Manual of Systematics of Archaea and Bacteria.* (Vol. 1–6).

Hill, A. W. and Mortishire-Smith, R. J. (2005). Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Communications in Mass Spectrometry*, *19*(21), 3111–3118. http://doi.org/10.1002/rcm.2177

Hill, D. W., Kertesz, T. M., Fontaine, D., Friedman, R. and Grant, D. F. (2008). Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Analytical Chemistry*, *80*(14), 5574–5582. http://doi.org/10.1021/ac800548g

Hipple, J. A., Sommer, H. and Thomas, H. A. (1949). A Precise Method of Determining the Faraday by Magnetic Resonance. *Physical Review*, *76*(12), 1877–1878. http://doi.org/10.1103/PhysRev.76.1877.2

Hopley, C., Bristow, T., Lubben, A., Simpson, A., Bull, E., Klagkou, K. and Langley, J. (2008).

Towards a universal product ion mass spectral library - reproducibility of product ion spectra across eleven different mass spectrometers. *Rapid Communications in Mass Spectrometry : RCM*, *22*(12), 1779–86. http://doi.org/10.1002/rcm.3545

Horai, H., Arita, M. and Nishioka, T. (2008). Comparison of ESI-MS Spectra in MassBank Database. In *2008 International Conference on BioMedical Engineering and Informatics* (Vol. 2, pp. 853–857). IEEE. http://doi.org/10.1109/BMEI.2008.339

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K. and Nishioka, T. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry : JMS*, *45*(7), 703–14. http://doi.org/10.1002/jms.1777

Hufsky, F., Rempt, M., Rasche, F., Pohnert, G. and Böcker, S. (2012). De novo analysis of electron impact mass spectra using fragmentation trees. *Analytica Chimica Acta*, *739*, 67–76. http://doi.org/10.1016/j.aca.2012.06.021

Hufsky, F., Scheubert, K. and Böcker, S. (2015). Mending the pieces : Computational mass spectrometry for small molecule fragmentation, (c).

Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B. and Magarvey, N. A. (2012). Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(47), 19196–201. http://doi.org/10.1073/pnas.1206376109

Krier, F., Boistel, C., Deravel, J., Coutte, F., Béchet, M., Leclère, V. and Jacques, P. (2014). Les lipopeptides d'origine microbienne. *Phytoma*, *672*, 38–42.

Jain, D. K., Collins-Thompson, D. L., Lee, H. and Trevors, J. T. (1991). A drop-collapsing test for screening surfactant-producing microorganisms. *Journal of Microbiological Methods*, *13*(4), 271–279. http://doi.org/10.1016/0167-7012(91)90064-W

Jarussophon, S., Acoca, S., Gao, J.-M., Deprez, C., Kiyota, T., Draghici, C. and Konishi, Y. (2009). Automated molecular formula determination by tandem mass spectrometry (MS/MS). *The Analyst*, *134*(4), 690–700. http://doi.org/10.1039/b818398h

Kameyama, A., Nakaya, S., Ito, H., Kikuchi, N., Angata, T., Nakamura, M. and Narimatsu, H. (2006). Strategy for simulation of CID spectra of N-linked oligosaccharides toward

glycomics. *Journal of Proteome Research*, *5*(4), 808–14. http://doi.org/10.1021/pr0503937

Karas, M. and Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, *60*(20), 2299–2301. http://doi.org/10.1021/ac00171a028

Kendrick, E. (1963). A Mass Scale Based on CH 2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. *Analytical Chemistry*, *35*(13), 2146–2154. http://doi.org/10.1021/ac60206a048

Kim, S., Kramer, R. W. and Hatcher, P. G. (2003). Graphical Method for Analysis of Ultrahigh-Resolution Braodband mass spectra of Natural Organic Matter, the Van Krevelen diagram. *Anal. Chem.*, *75*(20), 5336–5344. http://doi.org/10.1021/ac034415p CCC:

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A. and Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic Acids Research*, *44*(D1), D1202–D1213. http://doi.org/10.1093/nar/gkv951

Kind, T. and Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, *8*, 105. http://doi.org/10.1186/1471-2105-8-105

Kind, T. and Fiehn, O. (2010). Advances in structure elucidation of small molecules using mass spectrometry, 23–60. http://doi.org/10.1007/s12566-010-0015-9

King, E. O., Ward, M. K. and Raney, D. E. (1954). Two simple media for the demonstration of pyocyanin and fluorescin. *The Journal of Laboratory and Clinical Medicine*, *44*(2), 301–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13184240

Koch, B. P., Ludwichowski, K.-U., Kattner, G., Dittmar, T. and Witt, M. (2008). Advanced characterization of marine dissolved organic matter by combining reversed-phase liquid chromatography and FT-ICR-MS. *Marine Chemistry*, *111*(3–4), 233–241. http://doi.org/10.1016/j.marchem.2008.05.008

Konishi, Y., Kiyota, T., Draghici, C., Gao, J. M., Yeboah, F., Acoca, S. and Purisima, E. (2007). Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of

natural products. *Analytical Chemistry*, *79*(3), 1187–1197. http://doi.org/10.1021/ac061391o

Koulman, A., Cao, M., Faville, M., Lane, G., Mace, W. and Rasmussen, S. (2009). Semi-quantitative and structural metabolic phenotyping by direct infusion ion trap mass spectrometry and its application in genetical metabolomics. *Rapid Communications in Mass Spectrometry : RCM*, *23*(15), 2253–63. http://doi.org/10.1002/rcm.4142

Labby, K. J., Watsula, S. G. and Garneau-Tsodikova, S. (2015). Interrupted adenylation domains: Unique bifunctional enzymes involved in nonribosomal peptide biosynthesis. *Natural Product Reports*, *32*(5), 641–653. http://doi.org/10.1039/c4np00120f

LaRossa, R. A. (2015). Making metabolism accessible and meaningful: is the definition of a central metabolic dogma within reach? *Biotechnology Letters*, *37*(4), 741–751. http://doi.org/10.1007/s10529-014-1750-8

Lawrence, E. O. (1931). The Production of High Speed Protons Without the use of High Voltages, *38*, 1–2.

Lazar, I. M., Grym, J. and Foret, F. (2006). Microfabricated devices: A new sample introduction approach to mass spectrometry. *Mass Spectrometry Reviews*, *25*(4), 573–594. http://doi.org/10.1002/mas.20081

Leclère, V., Beaufort, S., Dessoy, S., Dehottay, P. and Jacques, P. (2009). Development of a biological test to evaluate the bioavailability of iron in culture media. *Journal of Applied Microbiology*, *107*(5), 1598–1605. http://doi.org/10.1111/j.1365-2672.2009.04345.x

Lederberg, J. (1964). Topological mapping of organic molecules.

Lederberg, J. (1987). How DENDRAL Was Conceived and Born. *ACM Symposium on the History of Medical Informatics National Library of Medicine*, *1945*(38). http://doi.org/10.1145/41526.41528

Li, L., Kresh, J. A., Karabacak, N. M., Cobb, J. S., Agar, J. N. and Hong, P. (2008). A Hierarchical Algorithm for Calculating the Isotopic Fine Structures of Molecules. *Journal of the American Society for Mass Spectrometry*, *19*(12), 1867–1874. http://doi.org/10.1016/j.jasms.2008.08.008

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S. and Lopez, R. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Research*, *43*(W1), W580–W584. http://doi.org/10.1093/nar/gkv279

Li, W., Rokni-Zadeh, H., De Vleeschouwer, M., Ghequire, M. G. K., Sinnaeve, D., Xie, G.-L. and De Mot, R. (2013). The Antimicrobial Compound Xantholysin Defines a New Group of Pseudomonas Cyclic Lipopeptides. *PLoS ONE*, *8*(5), e62946. http://doi.org/10.1371/journal.pone.0062946

Linne, U., Doekel, S. and Marahiel, M. A. (2001). Portability of Epimerization Domain and Role of Peptidyl Carrier Protein on Epimerization Activity in Nonribosomal Peptide Synthetases †. *Biochemistry*, *40*(51), 15824–15834. http://doi.org/10.1021/bi011595t

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z. and Wang, R. (2015). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, *31*(3), 405–412. http://doi.org/10.1093/bioinformatics/btu626

Lombard, B., & Leclercq, A. (2010). Validation of Innovative Food Microbiological Methods According to the EN ISO 16140 Standard. *Food Analytical Methods*, *4*(2), 163–172. http://doi.org/10.1007/s12161-010-9154-4

Loss, R. D. (2001). Atomic Weights of the Elements 2001. *Pure and Applied Chemistry*, *73*(4), 667–683. http://doi.org/10.1351/pac200173040667

Lydic, T. A., Busik, J. V., Esselman, W. J. and Reid, G. E. (2009). Complementary precursor ion and neutral loss scan mode tandem mass spectrometry for the analysis of glycerophosphatidylethanolamine lipids from whole rat retina. *Analytical and Bioanalytical Chemistry*, *394*(1), 267–275. http://doi.org/10.1007/s00216-009-2717-9

Ma, Z., Geudens, N., Kieu, N. P., Sinnaeve, D., Ongena, M., Martins, J. C. and Höfte, M. (2016). Biosynthesis, chemical structure, and structure-activity relationship of orfamide lipopeptides produced by Pseudomonas protegens and related species. *Frontiers in Microbiology*, *7*(MAR), 1–16. http://doi.org/10.3389/fmicb.2016.00382

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, *72*(6), 1156–1162.

http://doi.org/10.1021/ac991131p

Makarov, A. and Scigelova, M. (2010). Coupling liquid chromatography to Orbitrap mass spectrometry. *Journal of Chromatography A*, *1217*(25), 3938–3945. http://doi.org/10.1016/j.chroma.2010.02.022

Mamyrin, B. A. (2001). Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry*. http://doi.org/10.1016/S1387-3806(00)00392-4

Marklein, G., Josten, M., Klanke, U., Müller, E., Horré, R., Maier, T. and Sahl, H.-G. (2009). Matrix-assisted laser desorption ionization-time of flight mass spectrometry for fast and reliable identification of clinical yeast isolates. *Journal of Clinical Microbiology*, *47*(9), 2912–7. http://doi.org/10.1128/JCM.00389-09

Mashima, J., Kodama, Y., Kosuge, T., Fujisawa, T., Katayama, T., Nagasaki, H. and Takagi, T. (2016). DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Research*, *44*(D1), D51–D57. http://doi.org/10.1093/nar/gkv1105

Mathieu, E. (1868). Le mouvement vibratoire d'une membrane de forme elliptique. *Journal de Mathématiques Pures et Appliquées 2e Série*, *2*, 137–203.

Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J. and Saito, K. (2010). AtMetExpress development: a phytochemical atlas of Arabidopsis development. *Plant Physiology*, *152*(2), 566–78. http://doi.org/10.1104/pp.109.148031

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA, *74*(2), 560–564.

McLafferty, F. W., Hertel, R. H. and Villwock, R. D. (1974). Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures. *Organic Mass Spectrometry*, *9*(7), 690–702. http://doi.org/10.1002/oms.1210090710

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N. and Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research*, *41*(Web Server issue), 597–600. http://doi.org/10.1093/nar/gkt376

Meija, J., Coplen, T. B., Berglund, M., Brand, W. A., De Bièvre, P., Gröning, M. and

Prohaska, T. (2016). Atomic weights of the elements 2013 (IUPAC Technical Report). *Pure and Applied Chemistry*. http://doi.org/10.1515/pac-2015-0305

Milman, B. L. (2005). Towards a full reference library of MS(n) spectra. Testing of a library containing 3126 MS2 spectra of 1743 compounds. *Rapid Communications in Mass Spectrometry : RCM*, *19*(19), 2833–9. http://doi.org/10.1002/rcm.2131

Mueller, C. A., Weinmann, W., Dresen, S., Schreiber, A. and Gergov, M. (2005). Development of a multi-target screening analysis for 301 drugs using a QTrap liquid chromatography/tandem mass spectrometry system and automated library searching. *Rapid Communications in Mass Spectrometry : RCM*, *19*(10), 1332–8. http://doi.org/10.1002/rcm.1934

Nacef, M., Chevalier, M., Chollet, S., Drider, D. and Flahaut, C. (2017). MALDI-TOF mass spectrometry for the identification of lactic acid bacteria isolated from a French cheese: The Maroilles. *International Journal of Food Microbiology*, *247*, 2–8. http://doi.org/10.1016/j.ijfoodmicro.2016.07.005

Nguyen, D. D., Melnik, A. V., Koyama, N., Lu, X., Schorn, M., Fang, J. and Dorrestein, P. C. (2016). Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nature Microbiology*, *2*(October 2016). http://doi.org/10.1038/nmicrobiol.2016.197

Novák, J., Lemr, K., Schug, K. A. and Havlíček, V. (2015). CycloBranch: De Novo Sequencing of Nonribosomal Peptides from Accurate Product Ion Mass Spectra. *Journal of The American Society for Mass Spectrometry*, *26*(10), 1780–1786. http://doi.org/10.1007/s13361-015-1211-1

Novák, J., Škríba, A., Zápal, J., Kuzma, M. and Havlíček, V. (2018). CycloBranch: an Open Tool for Fine Isotope Structures in Conventional and Product Ion Mass Spectra. *Journal of Mass Spectrometry*. http://doi.org/10.1002/jms.4285

Oberacher, H., Pavlic, M., Libiseller, K., Schubert, B., Sulyok, M., Schuhmacher, R. and Köfeler, H. C. (2009). On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study. *Journal of Mass Spectrometry : JMS*, *44*(4), 485–93. http://doi.org/10.1002/jms.1545

Ohno, T., Parr, T. B., Gruselle, M.-C. I., Fernandez, I. J., Sleighter, R. L. and Hatcher, P. G. (2014). Molecular Composition and Biodegradability of Soil Organic Matter: A Case Study Comparing Two New England Forest Types. *Environmental Science & Technology*, *48*, 7729–7236. http://doi.org/10.1021/es405570c

Olson, M. T. and Yergey, A. L. (2009). Calculation of the isotope cluster for polypeptides by probability grouping. *Journal of the American Society for Mass Spectrometry*, *20*(2), 295–302. http://doi.org/10.1016/j.jasms.2008.10.007

Ongena, M. and Jacques, P. (2008). Bacillus lipopeptides: versatile weapons for plant disease biocontrol. *Trends in Microbiology*, *16*(3), 115–125. http://doi.org/10.1016/j.tim.2007.12.009

Palit, M. and Mallard, G. (2009). Fragmentation energy index for universalization of fragmentation energy in ion trap mass spectrometers for the analysis of chemical weapon convention related chemicals by atmospheric pressure ionization-tandem mass spectrometry analysis. *Analytical Chemistry*, *81*(7), 2477–85. http://doi.org/10.1021/ac802079w

Patiny, L. and Borel, A. (2013). ChemCalc: A Building Block for Tomorrow ' s Chemical Infrastructure. *Journal of Chemical Information and Modeling*, 1–21. http://doi.org/10.1021/ci300563h

Pavlic, M., Libiseller, K. and Oberacher, H. (2006). Combined use of ESI-QqTOF-MS and ESI-QqTOF-MS/MS with mass-spectral library search for qualitative analysis of drugs. *Analytical and Bioanalytical Chemistry*, *386*(1), 69–82. http://doi.org/10.1007/s00216-006-0634-8

Payne, A. H. and Glish, G. L. (2005). Tandem mass spectrometry in quadrupole ion trap and ion cyclotron resonance mass spectrometers. *Methods in Enzymology*, *402*(05), 109–148. http://doi.org/10.1016/S0076-6879(05)02004-5

Pelander, A., Tyrkkö, E. and Ojanperä, I. (2009). In silico methods for predicting metabolism and mass fragmentation applied to quetiapine in liquid chromatography/time-of-flight mass spectrometry urine drug screening. *Rapid Communications in Mass Spectrometry : RCM*, *23*(4), 506–14. http://doi.org/10.1002/rcm.3901

Penning, F. M. (1936). *The glow discharge at low pressure*. Retrieved from http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-trans-0175.pdf

Phillips, D. M. P. (1963). The presence of acetyl groups in Histones. *BBA - General Subjects*, *100*(2), 598–599. http://doi.org/10.1016/0304-4165(65)90032-2

Esmaeel, Q., Chevalier, M., Chataigne, G., Rathinasamy, S., Jacques, P., & Leclère, V. (2016). Nonribosomal peptide synthetase with a unique iterative-alternative-optional mechanism catalyzes amonabactin synthesis in Aeromonas. *Applied Microbiology and Biotechnology*. http://doi.org/10.1007/s00253-016-7773-4

Ramaley, L. and Herrera, L. C. (2008). Software for the calculation of isotope patterns in tandem mass spectrometry. *Rapid Communications in Mass Spectrometry : RCM*, *22*(17), 2707–14. http://doi.org/10.1002/rcm.3668

Reemtsma, T. (2009). Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry. Status and needs. *Journal of Chromatography A*, *1216*(18), 3687–3701. http://doi.org/10.1016/j.chroma.2009.02.033

Ridder, L., van der Hooft, J. J. J., Verhoeven, S., de Vos, R. C. H., van Schaik, R. and Vervoort, J. (2012). Substructure-based annotation of high-resolution multistage MS(n) spectral trees. *Rapid Communications in Mass Spectrometry : RCM*, *26*(20), 2461–71. http://doi.org/10.1002/rcm.6364

Roach, P. J., Laskin, J. and Laskin, A. (2011). Higher-order mass defect analysis for mass spectra of complex organic mixtures. *Analytical Chemistry*, *83*(12), 4924–4929. http://doi.org/10.1021/ac200654j

Rockwood, A. L. and Haimi, P. (2006). Efficient calculation of accurate masses of isotopic peaks. *Journal of the American Society for Mass Spectrometry*, *17*(3), 415–419. http://doi.org/10.1016/j.jasms.2005.12.001

Rockwood, A. L., Kushnir, M. M. and Nelson, G. J. (2003). Dissociation of individual isotopic peaks: Predicting isotopic distributions of product ions in MSn. *Journal of the American Society for Mass Spectrometry*, *14*(4), 311–322. http://doi.org/10.1016/S1044-

0305(03)00062-X

Roepstorff, P. and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, *11*(11), 601. http://doi.org/10.1002/bms.1200111109

Rogers, S., Scheltema, R. A., Girolami, M. and Breitling, R. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, *25*(4), 512–518. http://doi.org/10.1093/bioinformatics/btn642

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors, *74*(12), 5463–5467.

Sawada, Y., Akiyama, K., Sakata, A., Kuwahara, A., Otsuki, H., Sakurai, T. and Hirai, M. Y. (2009). Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant & Cell Physiology*, *50*(1), 37–47. http://doi.org/10.1093/pcp/pcn183

Schmidt, A., Karas, M. and Dülcks, T. (2003). Effect of different solution flow rates on analyte ion signals in nano-ESI MS, or: When does ESI turn into nano-ESI? *Journal of the American Society for Mass Spectrometry*, *14*(5), 492–500. http://doi.org/10.1016/S1044-0305(03)00128-4

Schreiber, A., Efer, J. and Engewald, W. (2000). Application of spectral libraries for high-performance liquid chromatography–atmospheric pressure ionisation mass spectrometry to the analysis of pesticide and explosive residues in environmental samples. *Journal of Chromatography A*, *869*(1–2), 411–425. http://doi.org/10.1016/S0021-9673(99)01271-6

Schwyn B. and Neilands. (1987). Universal Chemical Assay for the Detection Determination of Siderophores', *56*, 47–56.

Schymanski, E. L., Meringer, M. and Brack, W. (2009). Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? *Analytical Chemistry*, *81*(9), 3608–17. http://doi.org/10.1021/ac802715e

Sheldon, M. T., Mistrik, R. and Croley, T. R. (2009). Determination of Ion Structures in

Structurally Related Compounds Using Precursor Ion Fingerprinting. *Journal of the American Society for Mass Spectrometry*, *20*(3), 370–376. http://doi.org/10.1016/j.jasms.2008.10.017

Shen, B. (2003). Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology*, *7*(2), 285–295. http://doi.org/10.1016/S1367-5931(03)00020-6

Sir J.J. Thomson. (1906). The Nobel Prize in Physics 1906 (Physics 19).

Sleno, L. and Volmer, D. A. (2004). Ion activation methods for tandem mass spectrometry. *Journal of Mass Spectrometry*, *39*(10), 1091–1112. http://doi.org/10.1002/jms.703

Smith, D. F., Podgorski, D. C., Rodgers, R. P., Blakney, G. T. and Hendrickson, C. L. (2018). 21 Tesla FT-ICR Mass Spectrometer for Ultrahigh-Resolution Analysis of Complex Organic Mixtures. *Analytical Chemistry*, *90*(3), 2041–2047. http://doi.org/10.1021/acs.analchem.7b04159

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C. and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*(6071), 674–679. http://doi.org/10.1038/321674a0

Snider, R. K. (2007). NIH Public Access. *Journal of the American Society for Mass Spectrometry*, *18*(8), 1511–1515. http://doi.org/10.1016/j.jasms.2007.05.016

Sparkman, O. D. (1996). Evaluating electron ionization mass spectral library search results. *Journal of the American Society for Mass Spectrometry*, *7*(4), 313–8. http://doi.org/10.1016/1044-0305(95)00705-9

Starcevic, A., Zucko, J., Simunkovic, J., Long, P. F., Cullum, J. and Hranueli, D. (2008). ClustScan : an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures, *36*(21), 6882–6892. http://doi.org/10.1093/nar/gkn685

Stein, S. E. and Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, *5*(9), 859–866. http://doi.org/10.1016/1044-0305(94)87009-8

Stephens, W. E. (1946). A Pulsed Mass Spectrometer with Time Dispersion. *Physical Review*, *69*, 691. http://doi.org/10.1103/PhysRev.69.674.2

Stroh, J. G., Petucci, C. J., Brecker, S. J., Huang, N. and Lau, J. M. (2007). Automated Sub-ppm Mass Accuracy on an ESI-TOF for Use with Drug Discovery Compound Libraries. *Journal of the American Society for Mass Spectrometry*, *18*(9), 1612–1616. http://doi.org/10.1016/j.jasms.2007.06.001

Süssmuth, R. D. and Mainz, A. (2017). Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie - International Edition*, *56*(14), 3770–3821. http://doi.org/10.1002/anie.201609079

Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. and Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9528–33. http://doi.org/10.1073/pnas.0402700101

Syrstad, E. A. and Turecek, F. (2005). Toward a general mechanism of electron capture dissociation. *Journal of the American Society for Mass Spectrometry*, *16*(2), 208–224. http://doi.org/10.1016/j.jasms.2004.11.001

Timm, W., Scherbart, A., Böcker, S., Kohlbacher, O. and Nattkemper, T. W. (2008). Peak intensity prediction in MALDI-TOF mass spectrometry: a machine learning study to support quantitative proteomics. *BMC Bioinformatics*, *9*, 443. http://doi.org/10.1186/1471-2105-9-443

Valentine, N., Wunschel, S., Wunschel, D., Petersen, C. and Wahl, K. (2005). Effect of Culture Conditions on Microorganism Identification by Matrix-Assisted Laser Desorption Ionization Mass Spectrometry, *71*(1), 58–64. http://doi.org/10.1128/AEM.71.1.58

Wakayama, M., Hirayama, A. and Soga, T. (2015). Capillary Electrophoresis-Mass Spectrometry (pp. 113–122). Humana Press, New York, NY. http://doi.org/10.1007/978-1-4939-2377-9_9

Walsh, C. T., Liu, J., Rusnak, F. and Sakaitani, M. (1990). Molecular Studies on Enzymes in Chorismate Metabolism and the Enterobactin Biosynthetic Pathway. *Chemical Reviews*,

90(7), 1105–1129. http://doi.org/10.1021/cr00105a003

Wani, W. Y., Boyer-Guittaut, M., Dodson, M., Chatham, J., Darley-Usmar, V. and Zhang, J. (2015). Regulation of autophagy by protein post-translational modification. *Laboratory Investigation*, *95*(1), 14–25. http://doi.org/10.1038/labinvest.2014.131

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R. and Medema, M. H. (2015). AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, *43*(W1), W237–W243. http://doi.org/10.1093/nar/gkv437

Welker, M. (2011). Proteomics for routine identification of microorganisms. *Proteomics*, *11*(15), 3143–3153. http://doi.org/10.1002/pmic.201100049

Werner, E., Heilier, J.-F., Ducruix, C., Ezan, E., Junot, C. and Tabet, J.-C. (2008). Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, *871*(2), 143–63. http://doi.org/10.1016/j.jchromb.2008.07.004

Wickremsinhe, E., Singh, G., Ackermann, B., Gillespie, T. and Chaudhary, A. (2006). A Review of Nanoelectrospray Ionization Applications for Drug Metabolism and Pharmacokinetics. *Current Drug Metabolism*, *7*(8), 913–928. http://doi.org/10.2174/138920006779010610

Wiley, W. C. and McLaren, I. H. (1955). Time-of-Flight Mass Spectrometer with Improved Resolution. *Review of Scientific Instruments*, *26*(12), 1150–1157. http://doi.org/10.1063/1.1715212

Wilkins, M. R., Gasteiger, E., Sanchez, J. C., Appel, R. D. and Hochstrasser, D. F. (1996). Protein identification with sequence tags. *Current Biology : CB*, *6*(12), 1543–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8994807

Winn, M., Fyans, J. K., Zhuo, Y. and Micklefield, J. (2016). Recent advances in engineering nonribosomal peptide assembly lines. *Nat. Prod. Rep.*, 317–347. http://doi.org/10.1039/C5NP00099H

Wolf, S., Schmidt, S., Müller-Hannemann, M. and Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, *11*, 148. http://doi.org/10.1186/1471-2105-11-148

Wolff, M. M. and Stephens, W. E. (1953). A Pulsed Mass Spectrometer with Time Dispersion. *Review of Scientific Instruments*, *24*(8), 616–617. http://doi.org/10.1063/1.1770801

Yamashita, M. and Fenn, B. J. (1984). Negative ion production with the electrospray ion source. *J Phys.Chem.*, *88*(20), 4671–4675. http://doi.org/10.1021/j150664a046

Yang, J. Y., Sanchez, L. M., Rath, C. M., Liu, X., Boudreau, P. D., Bruns, N. and Dorrestein, P. C. (2013). Molecular networking as a dereplication strategy. *Journal of Natural Products*, *76*(9), 1686–1699. http://doi.org/10.1021/np400413s

Zhang, H., Singh, S. and Reinhold, V. N. (2005). Congruent strategies for carbohydrate sequencing. 2. FragLib: an MSn spectral library. *Analytical Chemistry*, *77*(19), 6263–70. http://doi.org/10.1021/ac050725r

Zhang, H., Zhang, Y., Shi, Q., Ren, S., Yu, J., Ji, F. and Yang, M. (2012). Characterization of low molecular weight dissolved natural organic matter along the treatment trait of a waterworks using Fourier transform ion cyclotron resonance mass spectrometry. *Water Research*, *46*(16), 5197–5204. http://doi.org/10.1016/j.watres.2012.07.004

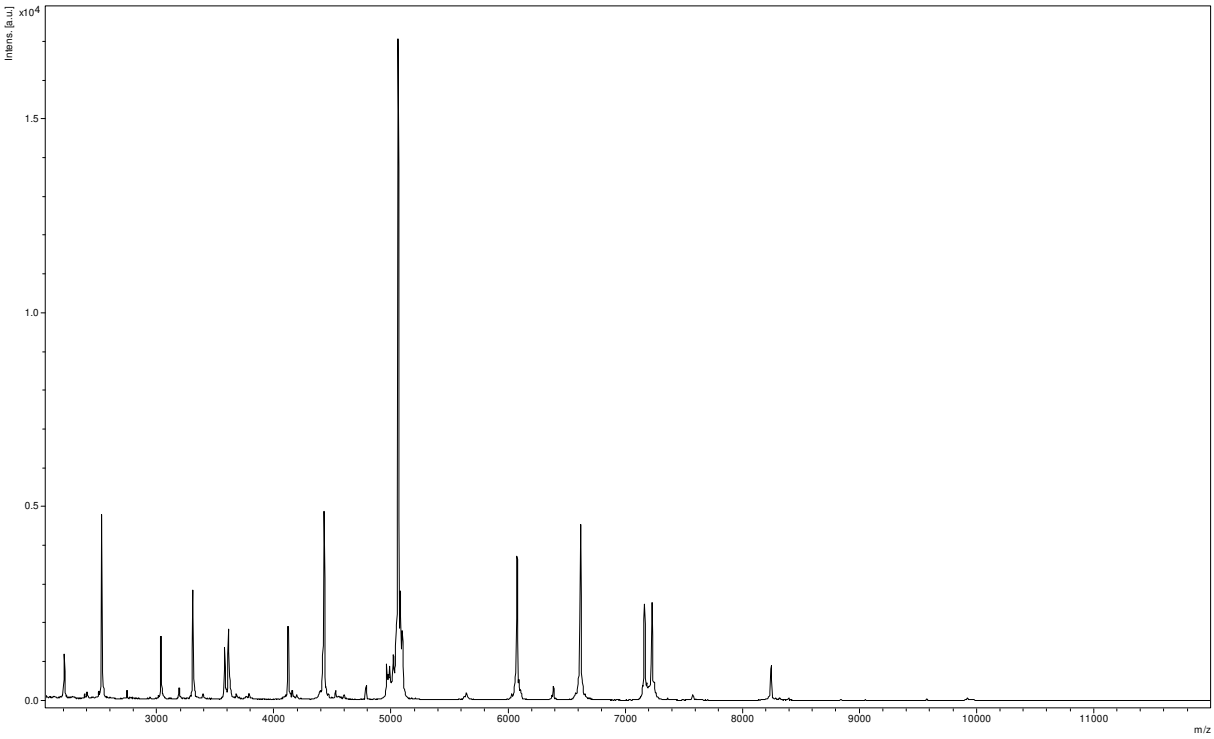
Zhou, C., Bowler, L. D. and Feng, J. (2008). A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC Bioinformatics*, *9*, 325. http://doi.org/10.1186/1471-2105-9-325

Zidour, M., Chevalier, M., Belguesmia, Y., Cudennec, B., Grard, T., Drider, D. and Flahaut, C. (2017). Isolation and characterization of bacteria colonizing Acartia tonsa copepod eggs and displaying antagonist effects against Vibrio anguillarum, Vibrio alginolyticus and Other pathogenic strains. *Frontiers in Microbiology*, *8*(OCT), 1919. http://doi.org/10.3389/fmicb.2017.01919

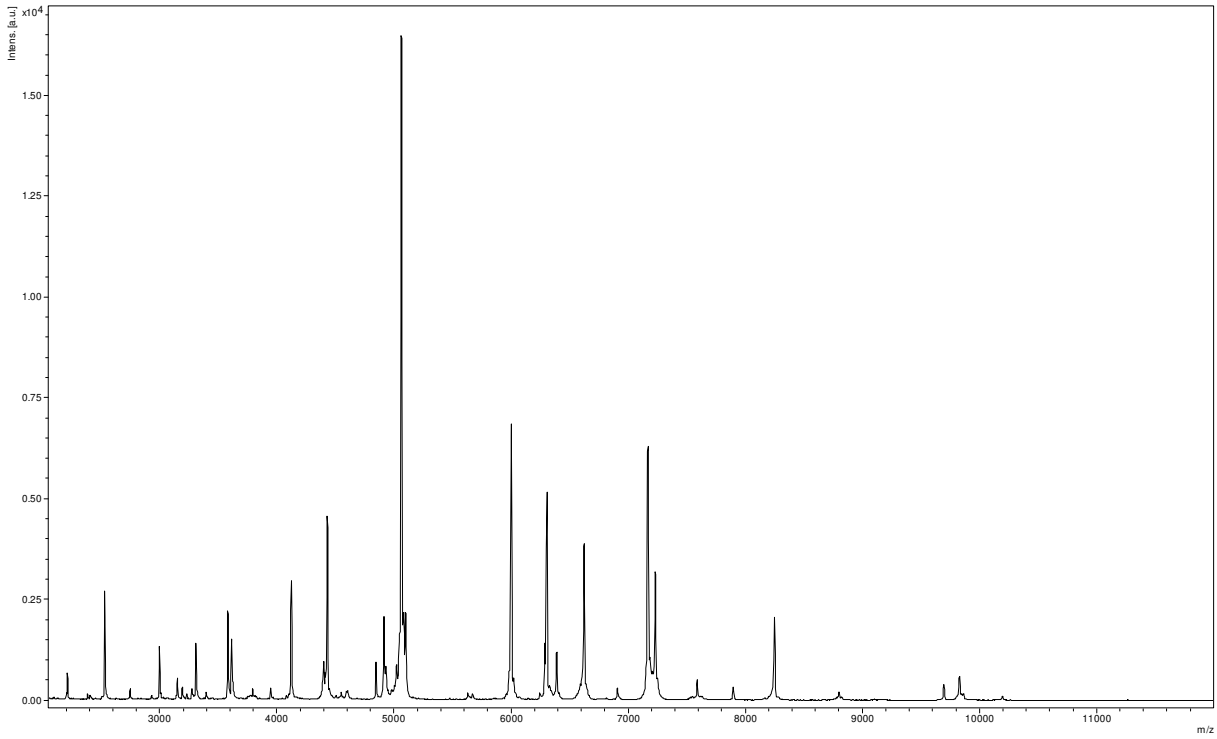
Zubarev, R. A. (2004). Electron-capture dissociation tandem mass spectrometry. *Current Opinion in Biotechnology*, *15*(1), 12–16. http://doi.org/10.1016/j.copbio.2003.12.002

Zubarev, R. A., Zubarev, A. R. and Savitski, M. M. (2008). Electron Capture/Transfer versus Collisionally Activated/Induced Dissociations: Solo or Duet? *Journal of the American Society for Mass Spectrometry*, *19*(6), 753–761. http://doi.org/10.1016/j.jasms.2008.03.007

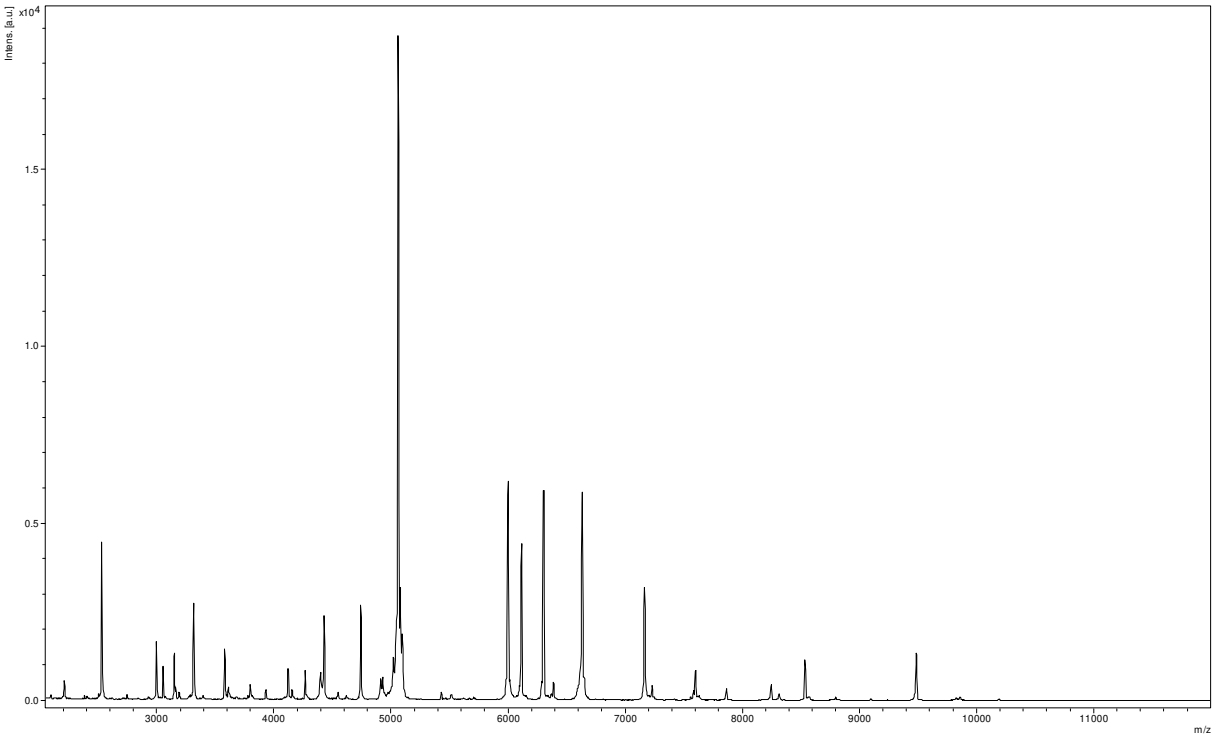
Annexe



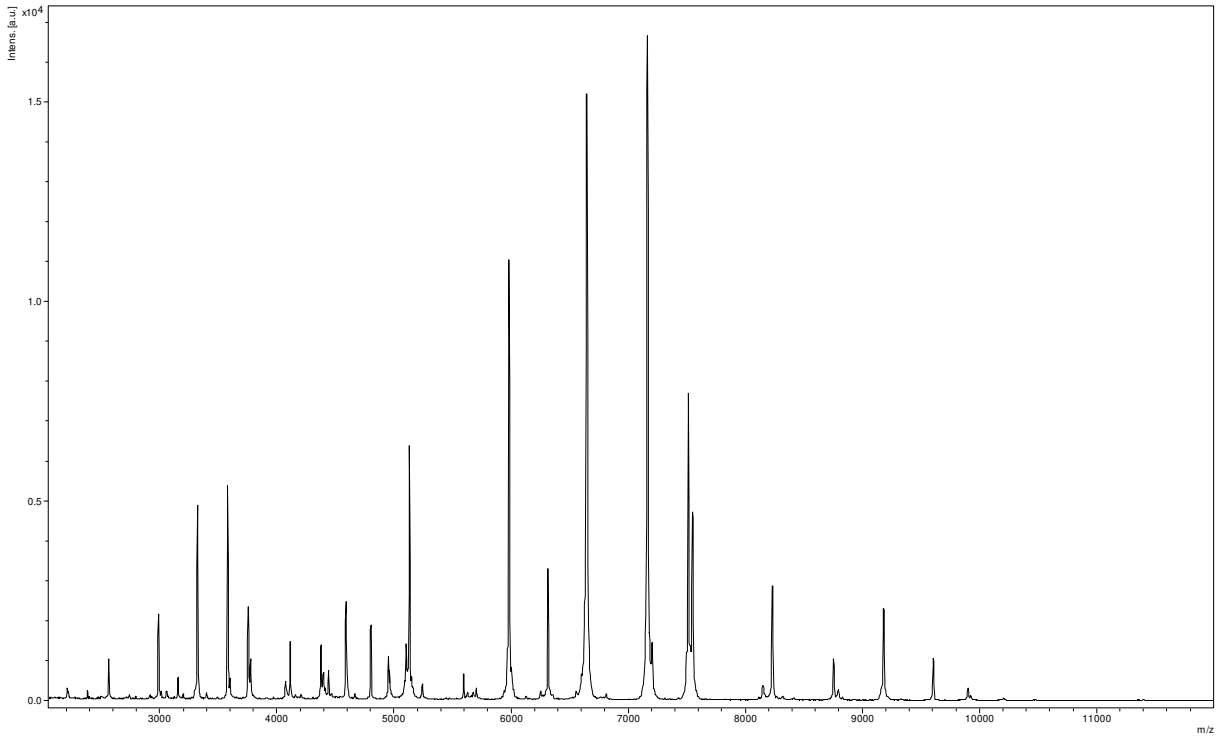
Annexe 1 : Souche 1, *Pseudomonas aeruginosa PA7*



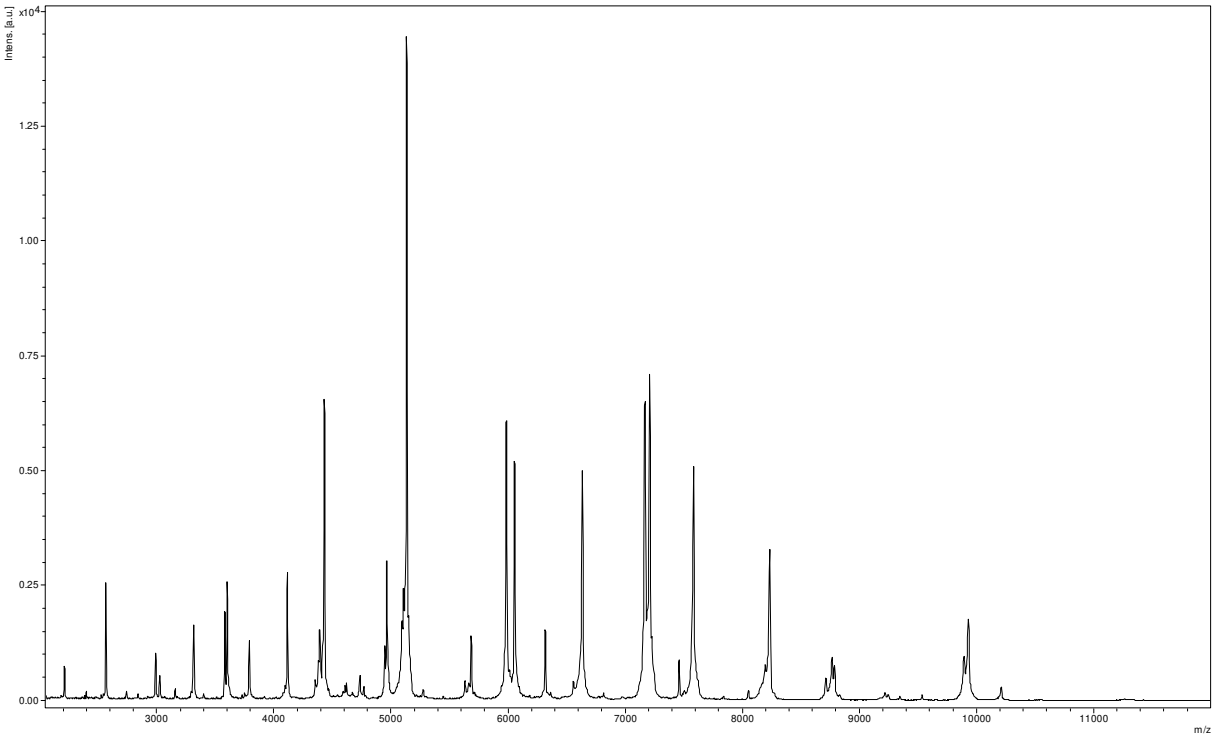
Annexe 2 : Souche 2, *Pseudomonas agarici NCPPB 2289T*



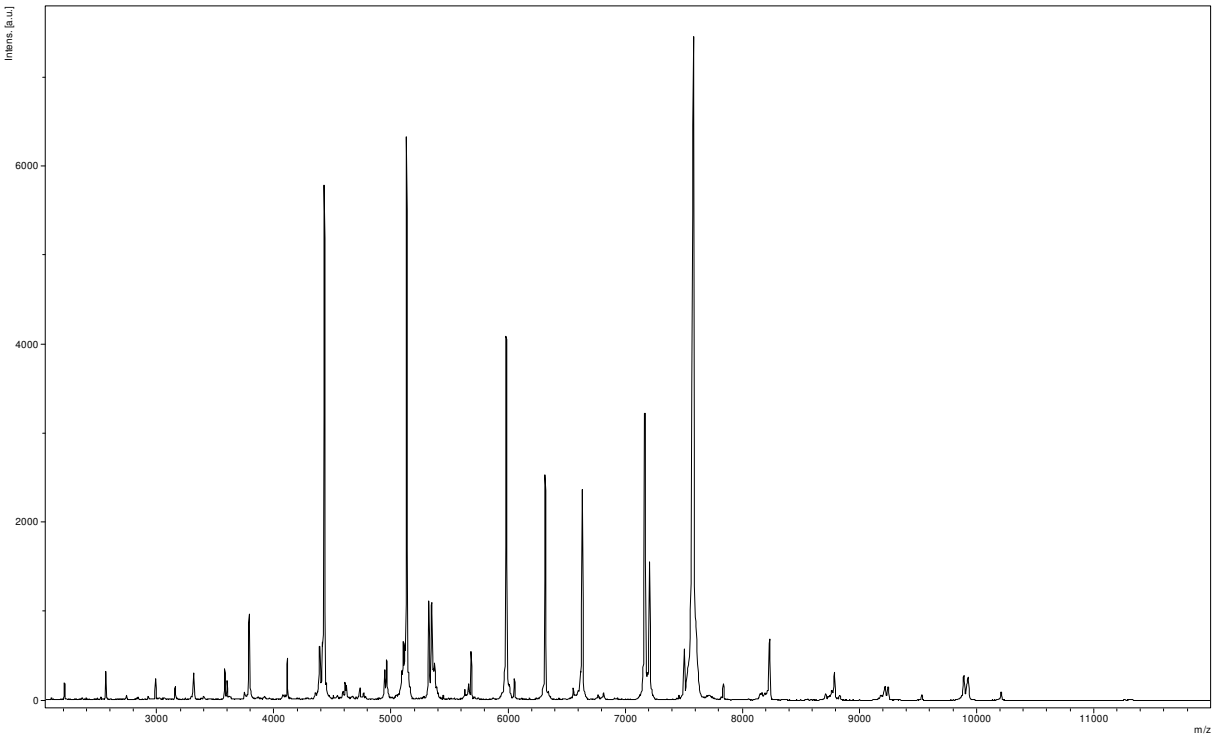
Annexe 3 : Souche 3, *Pseudomonas alcaligenes NBRC 14159T*



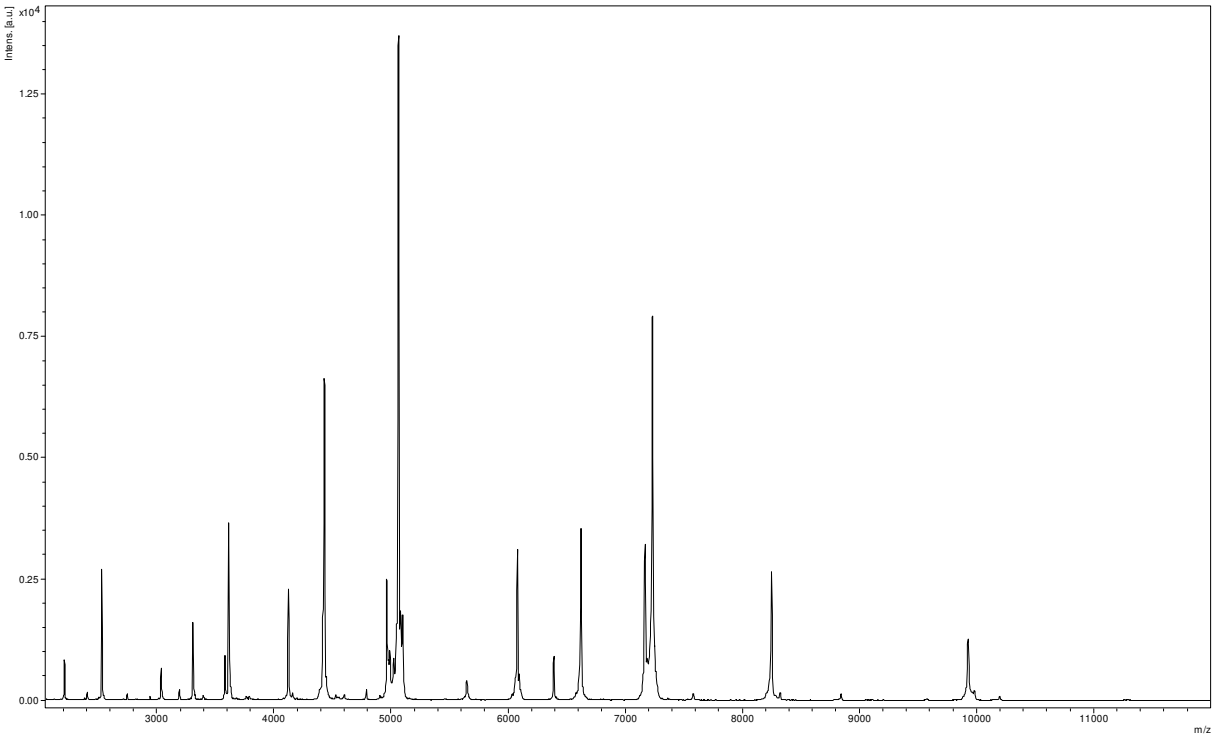
Annexe 4 : Souche 4, *Pseudomonas balearica DSM 6083T*



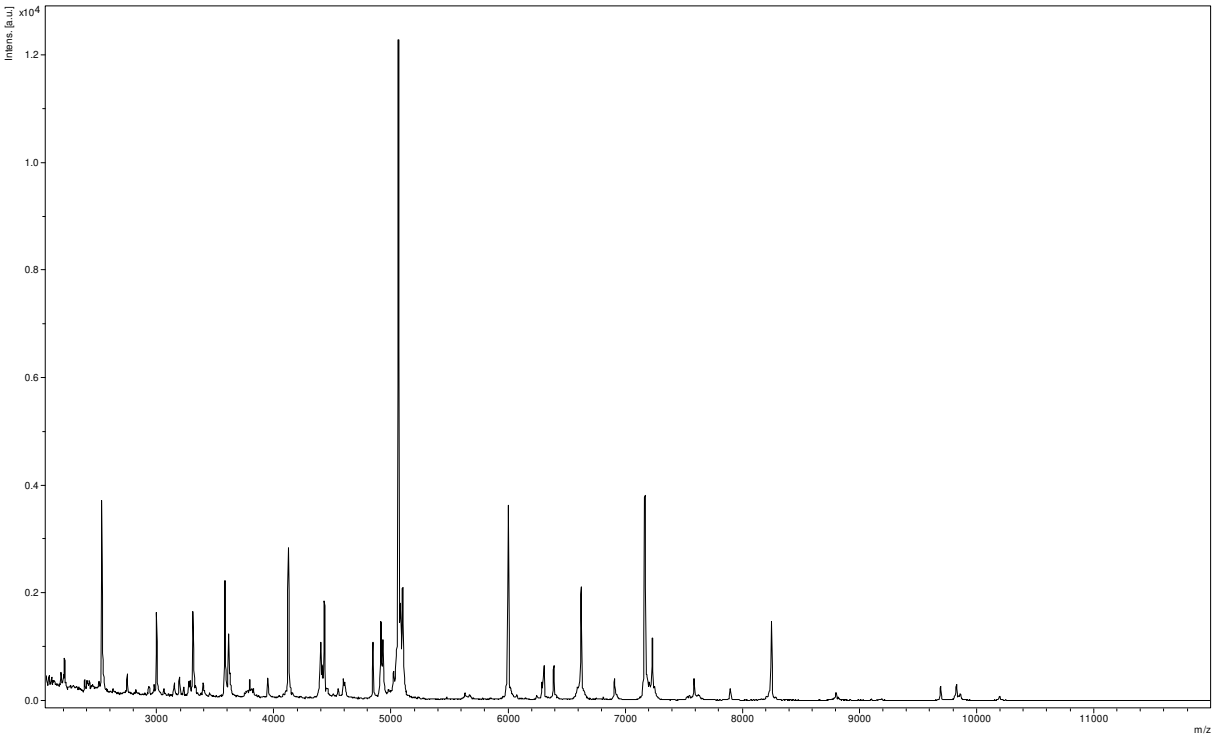
Annexe 5 : Souche 5, *Pseudomonas putida GB-1*



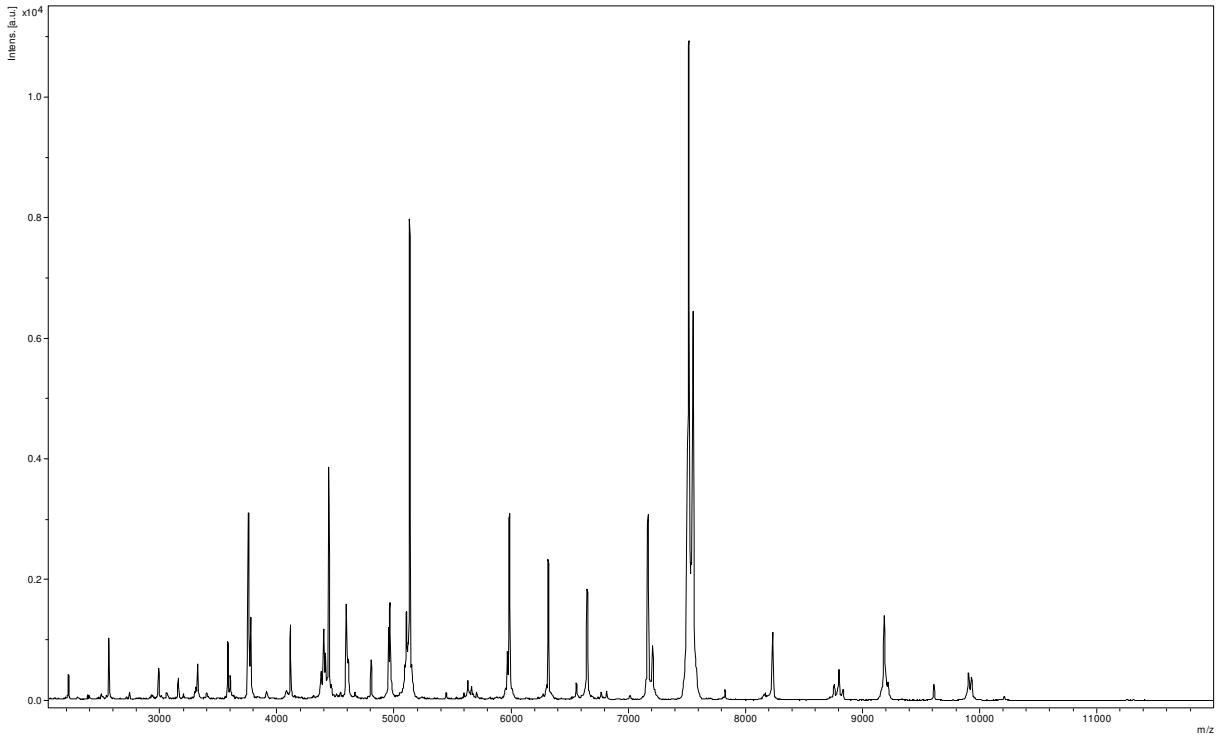
Annexe 6 : Souche 6, *Pseudomonas caeni DSM24390T*



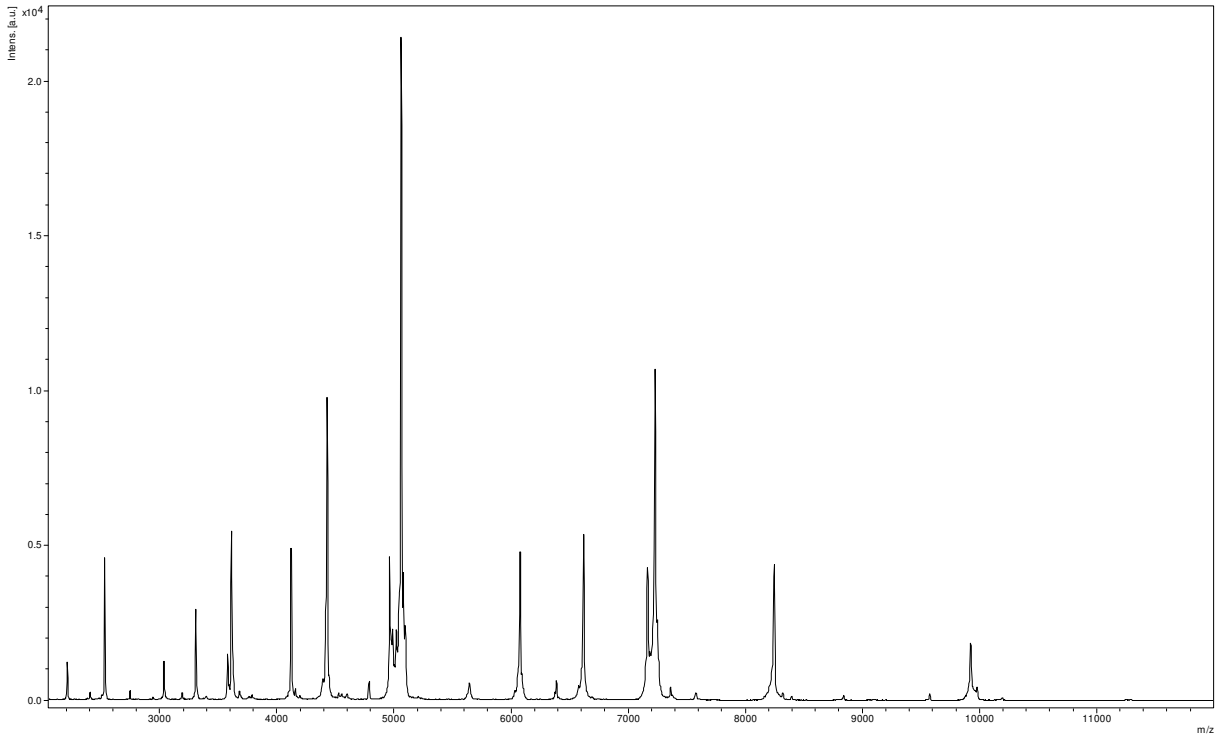
Annexe 7 : Souche 7, *Pseudomonas corrugata NCPPB 2445*



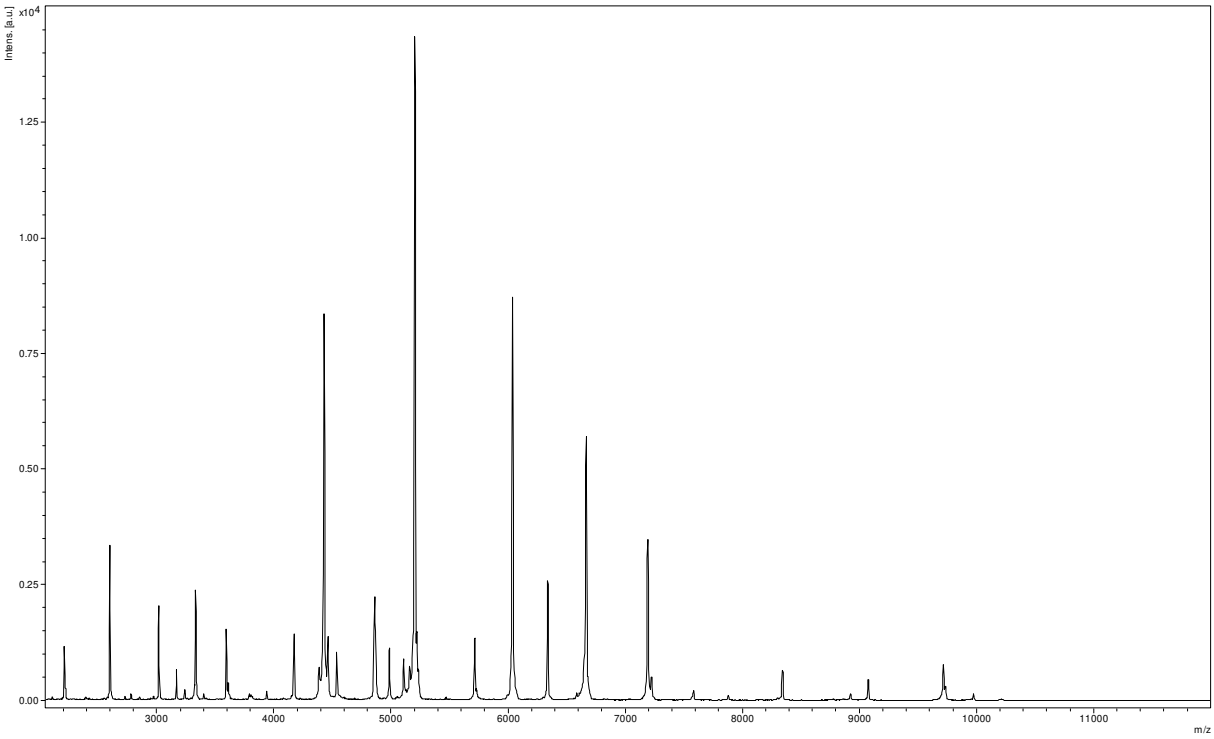
Annexe 8 : Souche 8, *Pseudomonas deceptionensis DSM 26521T*



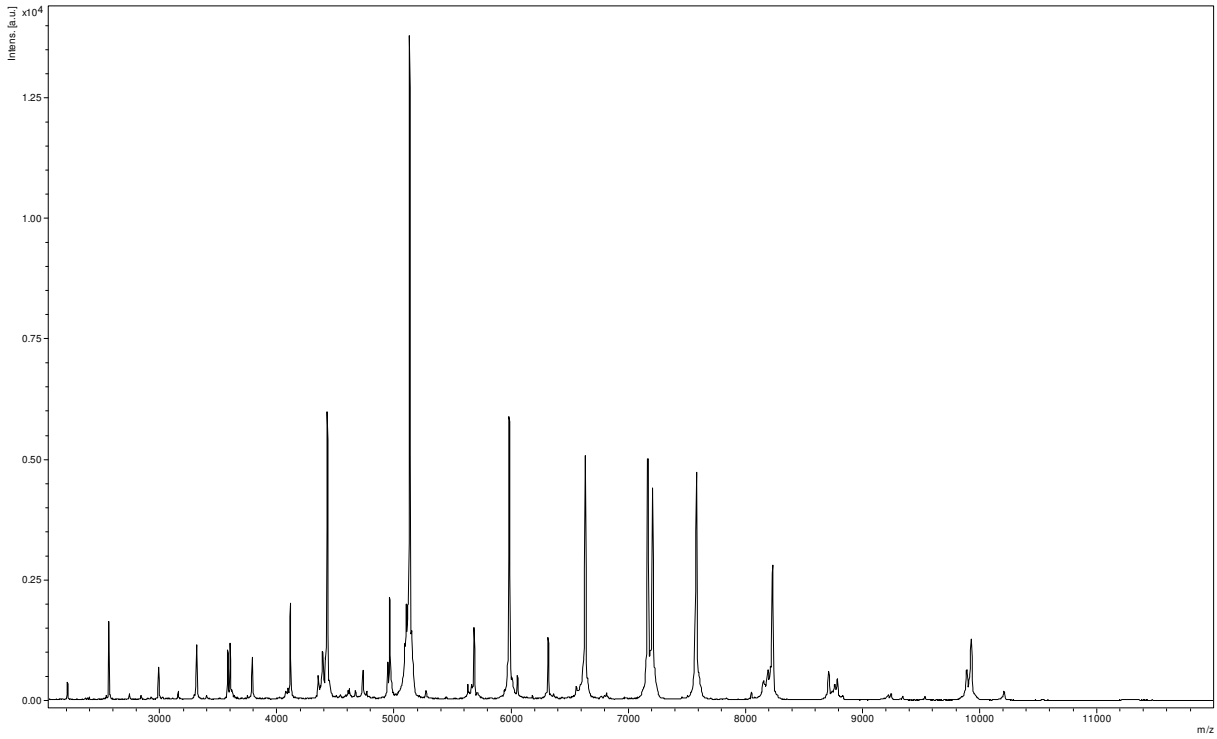
Annexe 9 : Souche 9, *Pseudomonas entomophila L48T*



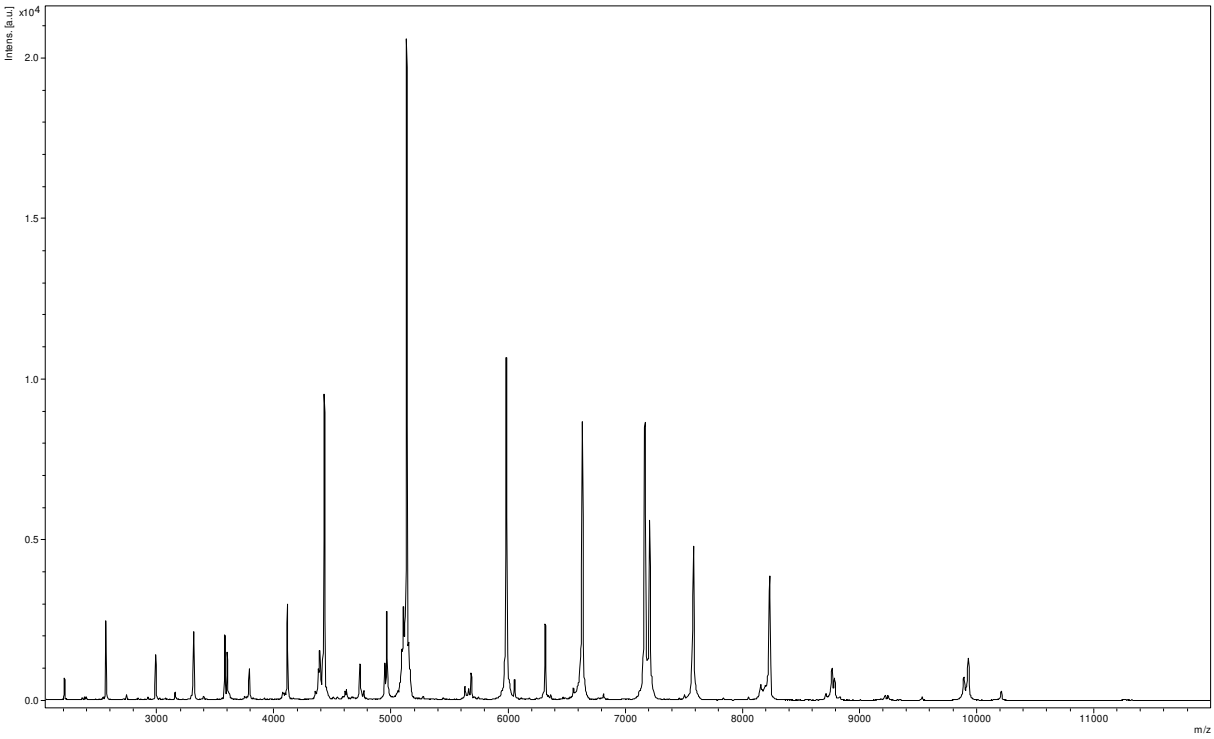
Annexe 10 : Souche 10, *Pseudomonas fluorescens ATCC 17400*



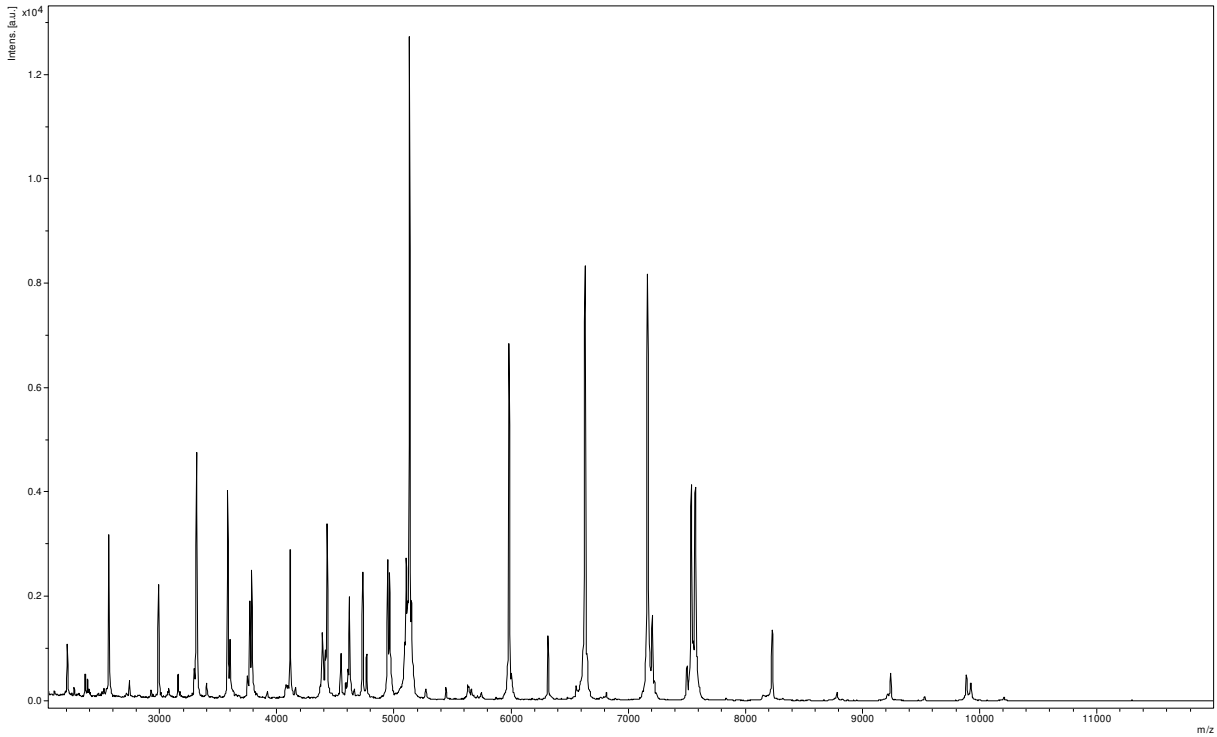
Annexe 11 : Souche 11, *Pseudomonas gingeri NCPPB 3146T*



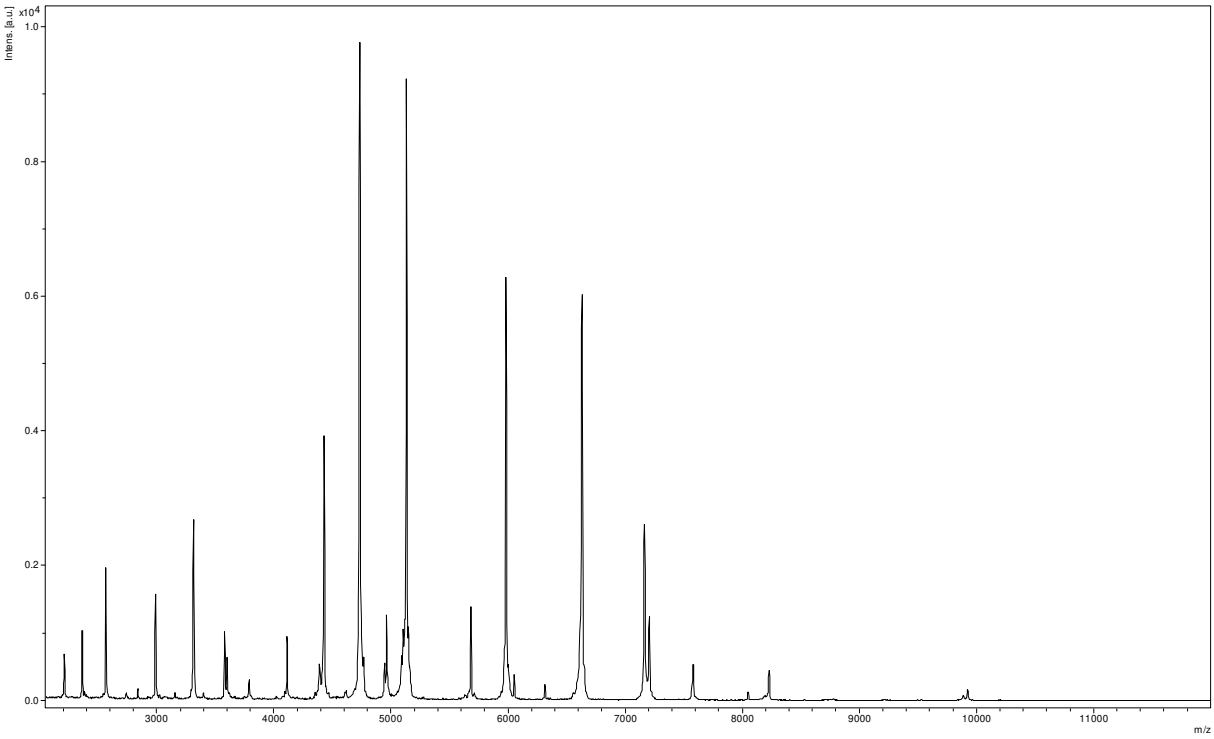
Annexe 12 : Souche 12, *Pseudomonas mediterranea CFBP 5447T*



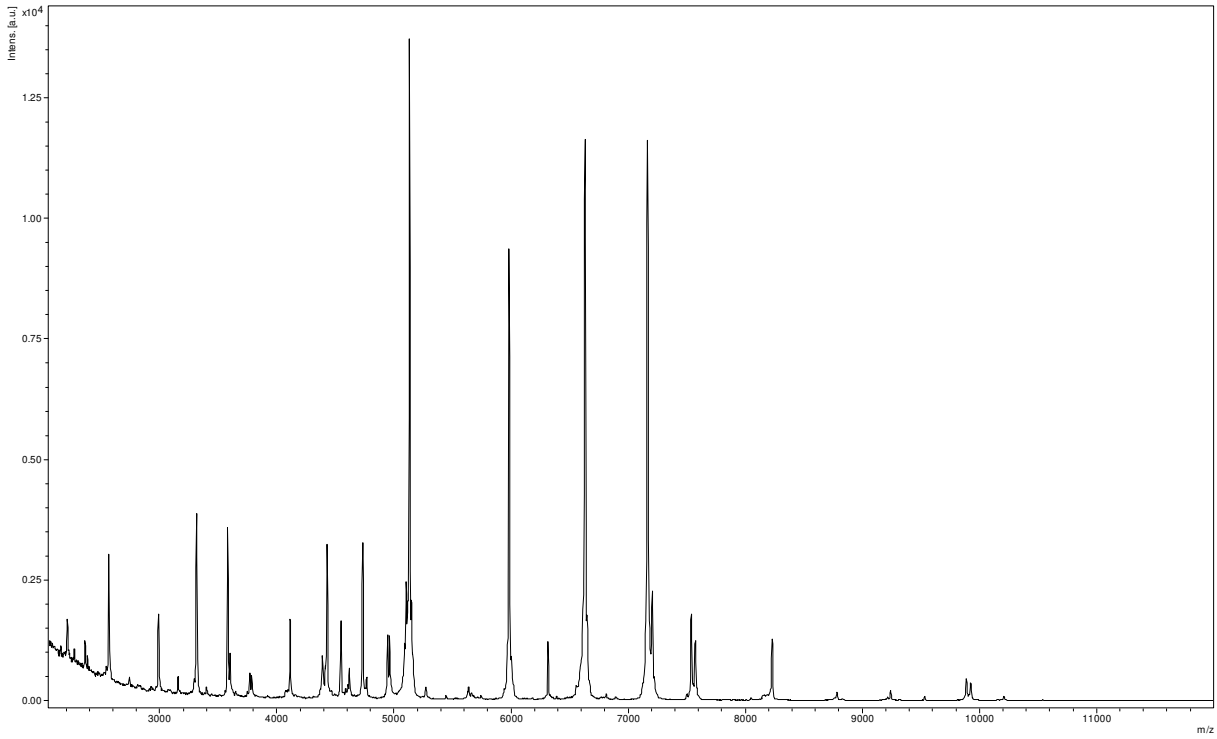
Annexe 13 : Souche 13, *Pseudomonas monteilii DSM14164T*



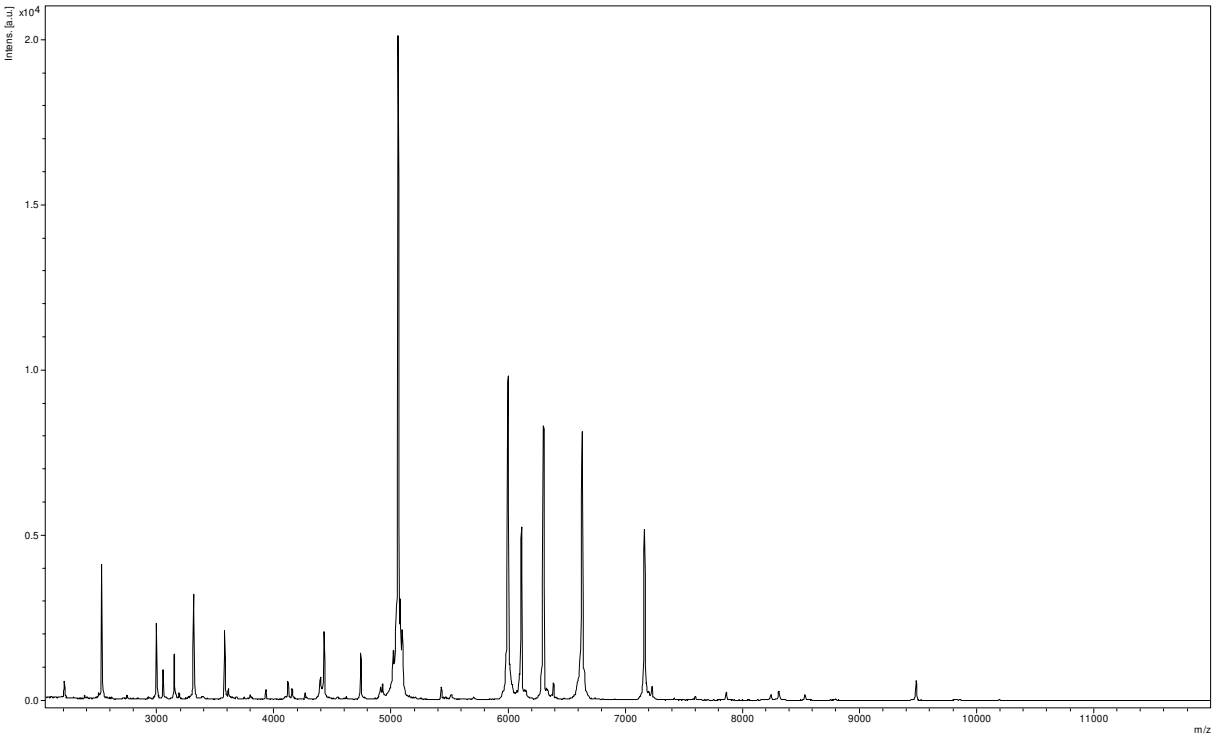
Annexe 14 : Souche 14, *Pseudomonas putida F1*



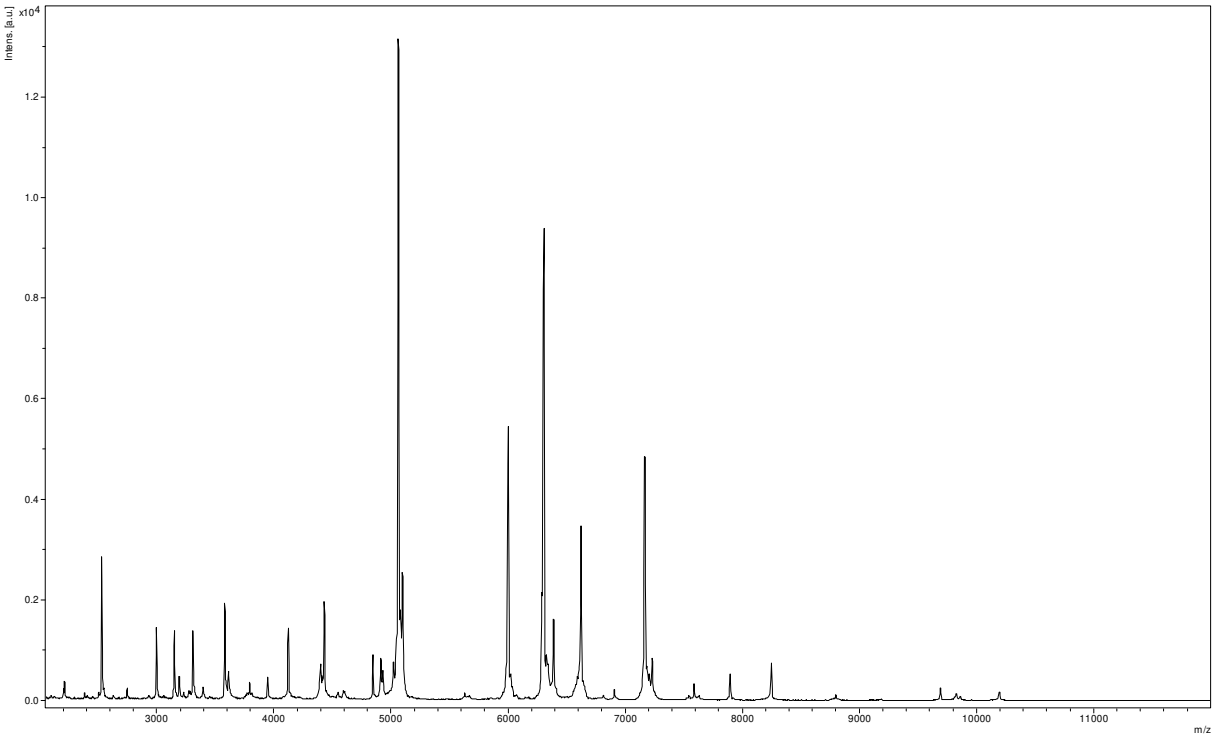
Annexe 15 : Souche 15, *Pseudomonas putida W15Oct28*



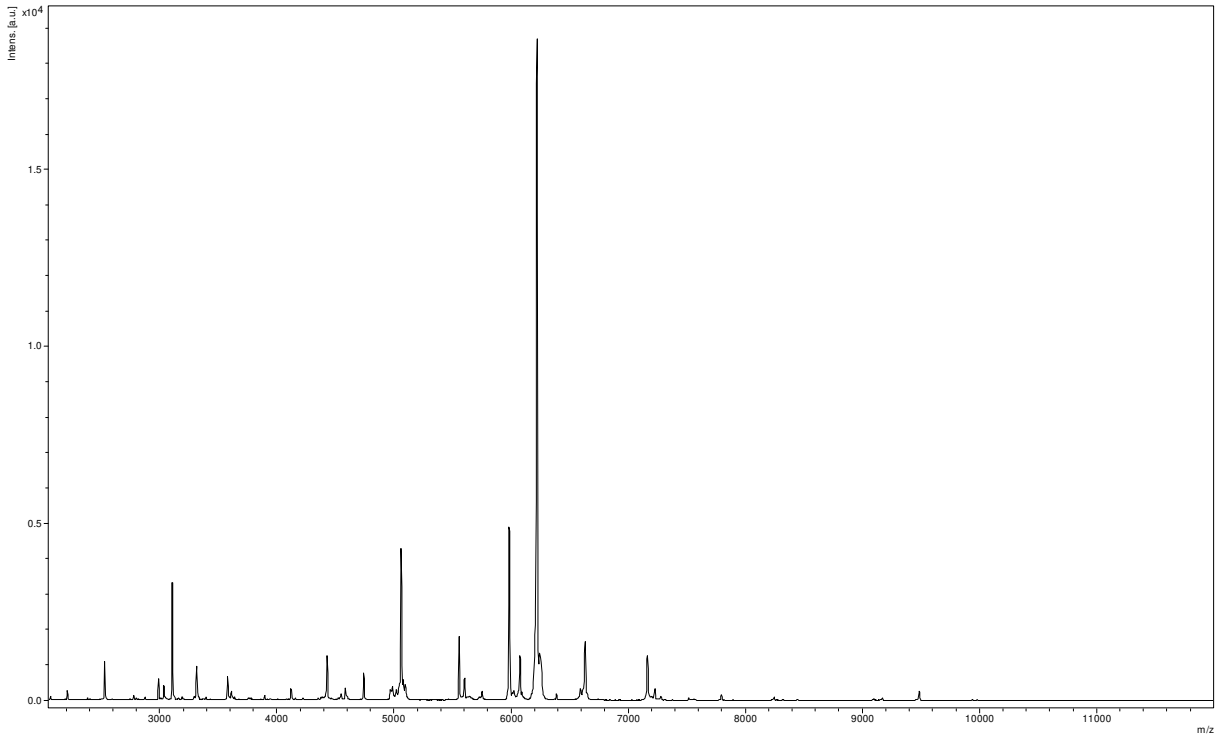
Annexe 16 : Souche 16, *Pseudomonas putida W619*



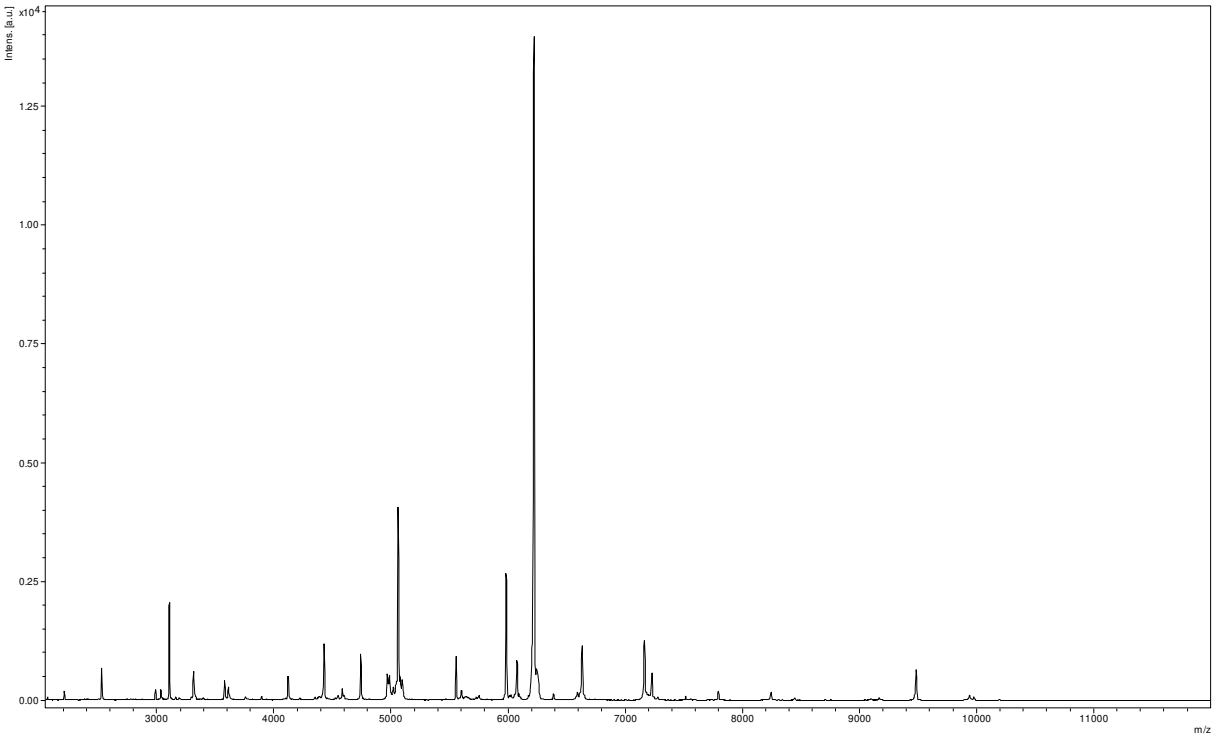
Annexe 17 : Souche 17, *Pseudomonas thivervalensis DSM 13194T*



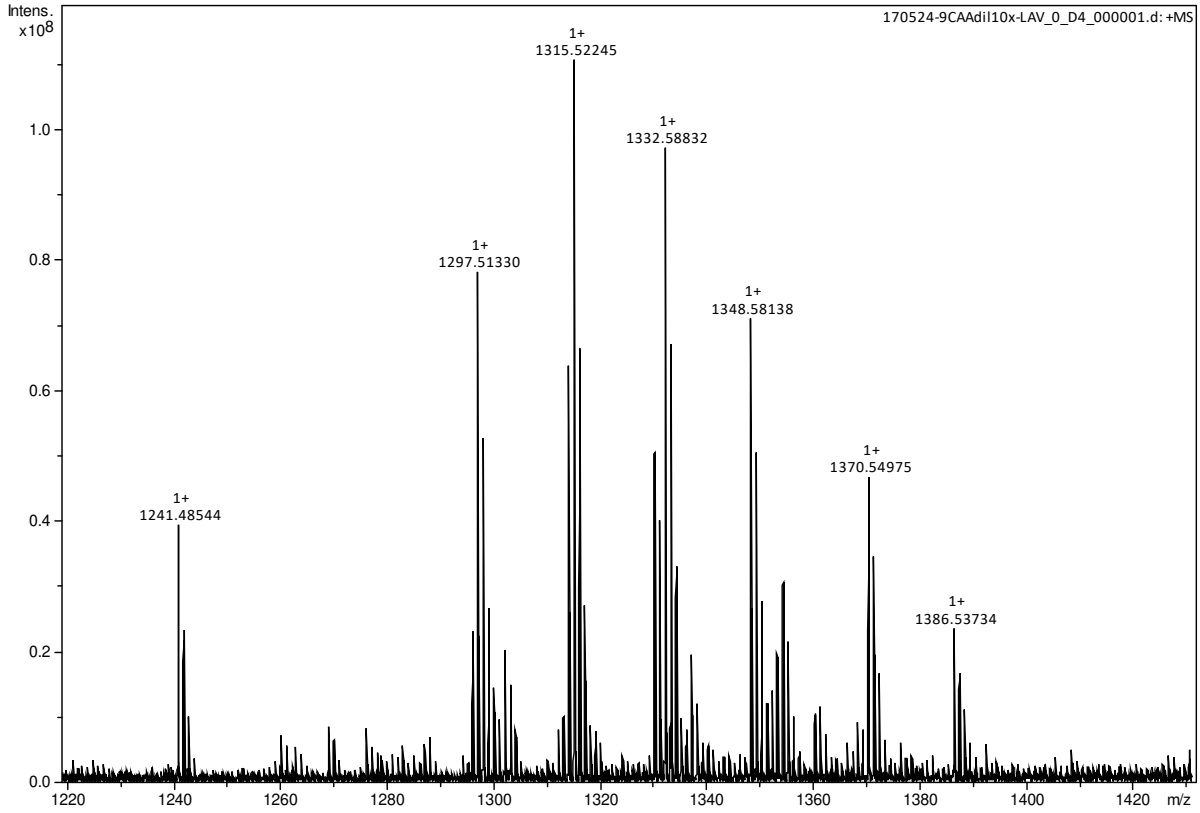
Annexe 18 : Souche 18, *Pseudomonas protegens Pf-5*



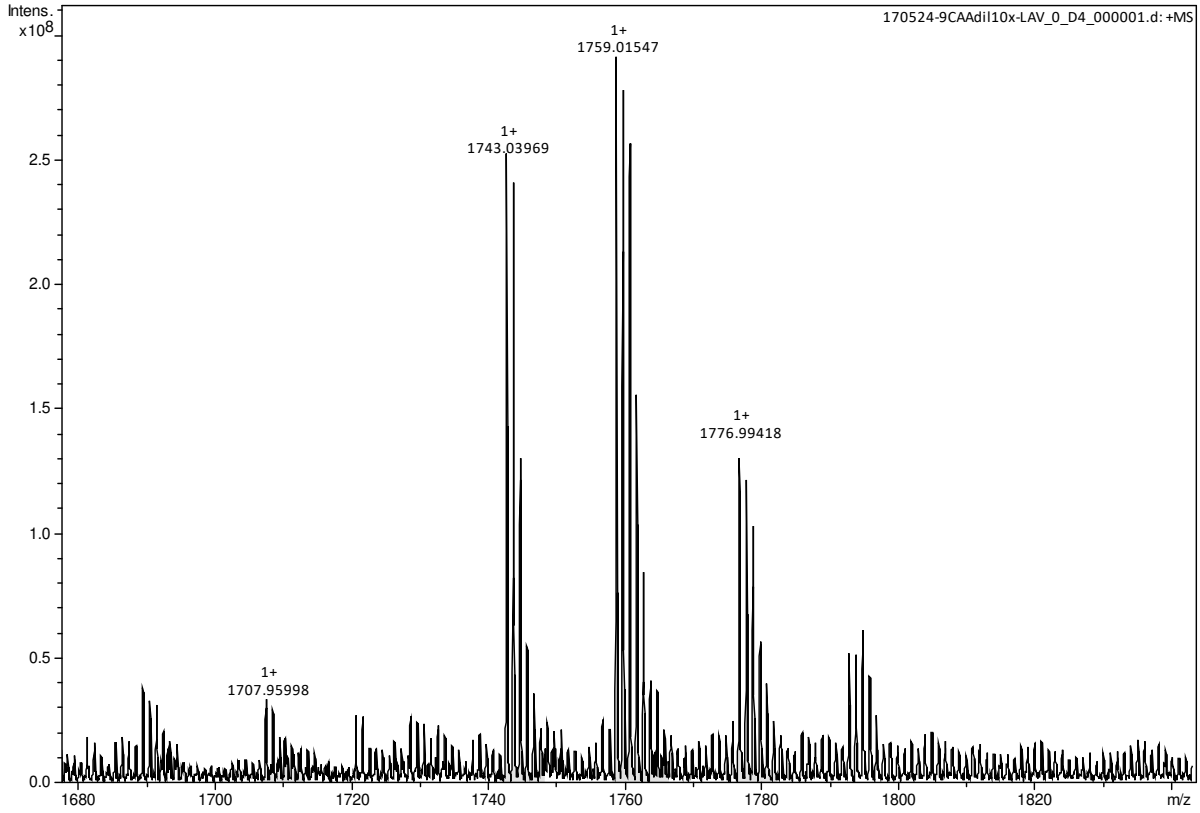
Annexe 19 : Souche 19, *Pseudomonas sp. W2Aug9*



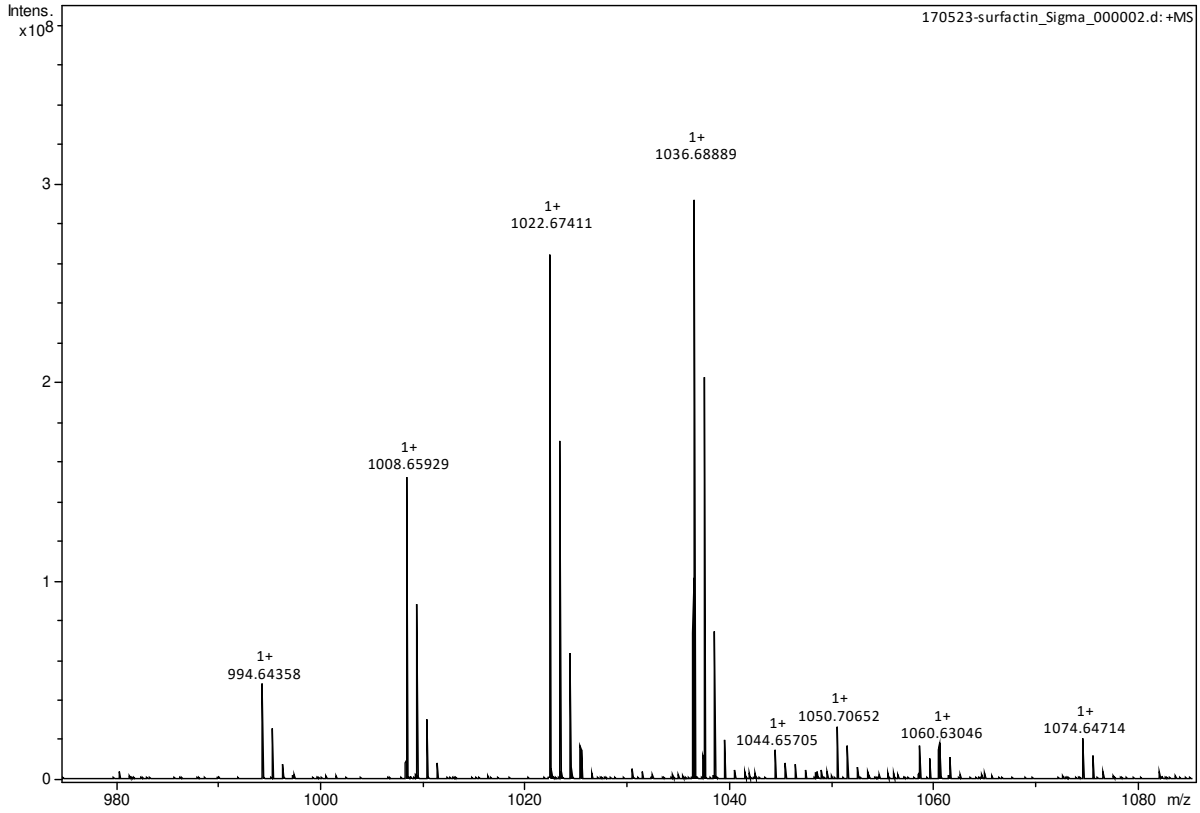
Annexe 20 : Souche 20, *Pseudomonas sp. AF76*



Annexe 21 : Spectre liée à la figure 61



Annexe 22 : Spectre liée à la figure 62



Annexe 23 : Spectre liée à la figure 48

Nous avons choisi de réaliser un *Workflow* permettant l'incrimination, la discrimination et l'identification de peptides non ribosomiques (NRPs) connus ou inconnus dans une philosophie tournée vers le criblage de nouveaux composés actifs. Cette combinaison de méthodes analytiques commence par **l'identification du microorganisme par une méthode de profilage phénotypique par spectrométrie de masse (MS).** Puis, nous avons choisi de combiner pour la première fois une méthode de calcul itérative et les informations contenues dans la base de données, Norine, dédiée aux NRPs. Cette combinaison nous a permis de **déterminer la composition élémentaire de peptide non ribosomique à partir de données HRMS combinées à un maillage vectoriel reliant les différents NRPs de Norine et les formules chimiques.** Nous illustrons également que d'une part cette méthode démontrée à partir des données MS peut-être extrapolée aux données de fragmentation MS/MS et que d'autre part elle présente **un intérêt pour la déréplication des NRPs mais aussi la caractérisation structurale de nouveaux composés actifs** pour des applications en particulier dans les secteurs de la santé et phytosanitaires. La performance du workflow sera illustrée par l'identification de lipopeptides produit par des souches de *Pseudomonas*. Ces lipopeptides sont particulièrement intéressants car se sont des composés ayant des applications potentielles en biocontrôle.

We have chosen to carry out a Workflow allowing the incrimination, the discrimination and the identification of known or unknown nonribosomal peptides (NRPs) in a philosophy oriented towards the screening of new active compounds. This combination of analytical methods begins with the **identification of the microorganism by a phenotypic profiling method by mass spectrometry (MS)**. Then, we chose to combine for the first time an iterative calculation method and the information contained in the database, Norine, dedicated to NRPs. This combination allowed us **to determine the nonribosomal peptide elemental composition from HRMS data combined with a vector mesh linking the different Norine NRPs and the chemical formulas**. We also illustrate that on the one hand this method demonstrated from the MS data can be extrapolated to the MS / MS fragmentation data and that on the other hand it is of **interest for the dereplication of NRPs but also the structural characterization of new compounds** for applications especially in the health and plant health sectors. The workflow performance will be illustrated by the identification of lipopeptides produced by Pseudomonas strains. These lipopeptides are particularly interesting because they are compounds with potential applications in biocontrol.