Développement de méthodes informatiques pour l'évaluation et l'amélioration de l'identification par spectrométrie de masse des peptides modifiés

Albane Lysiak

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Albane LYSIAK

Développement de méthodes informatiques pour l'évaluation et l'amélioration de l'identification par spectrométrie de masse des peptides modifiés

Thèse présentée et soutenue à Nantes, le 6 décembre 2022
Unité de recherche : LS2N (Université de Nantes), BIA (INRAE Nantes)

Rapporteurs avant soutenance :

Thierry LECROQ	Professeur des universités, LITIS/Université de Rouen Normandie
Christine CARAPITO	Directrice de recherche CNRS, IPHC/Université de Strasbourg

Composition du Jury :

Président :	Thomas BURGER	Directeur de recherche CNRS, EDyp/Université Grenoble-Alpes
Examinateurs :	Thierry LECROQ	Professeur des universités, LITIS/Université de Rouen Normandie
	Christine CARAPITO	Directrice de recherche CNRS, IPHC/Université de Strasbourg
	Martial REY	Chargé de recherche CNRS, Institut Pasteur (Paris)
Dir. de thèse :	Guillaume FERTIN	Professeur des universités, LS2N/Université de Nantes
Co-dir. de thèse :	Dominique TESSIER	Ingénieure de recherche, BIA/INRAE Nantes
Co-enc. de thèse :	Géraldine JEAN	Enseignant-chercheur, LS2N/Université de Nantes

Invité(s) :

Hélène ROGNIAUX	Ingénieure de recherche, BIA/INRAE Nantes

Remerciements

Je remercie d'abord mes deux rapporteurs, Christine Carapito et Thierry Lecroq, d'avoir relu mon manuscrit, de m'avoir fait part de leurs remarques constructives et d'avoir accepté de participer à mon jury de thèse. Merci également à Thomas Burger, Martial Rey et Hélène Rogniaux de participer à mon jury.

Je remercie ensuite mes trois encadrants ; mon directeur de thèse Guillaume Fertin pour son implication tout au long de la thèse ; Dominique Tessier pour m'avoir transmis une partie de ses connaissances sur ce domaine que je connaissais peu au début de ma thèse, pour les longues discussions toujours très intéressantes et constructives, et pour sa patience ; Géraldine Jean pour ses idées et remarques toujours très pertinentes. Merci à vous trois de m'avoir laissé beaucoup d'autonomie et de liberté dans mon travail tout en sachant me guider et m'aiguiller lorsque j'en avais besoin. Merci pour le temps que vous m'avez consacré au cours de réunions très régulières.

Merci aux équipes COMBI et BIA pour leur accueil.

Merci aux membres de l'ANR DeepProt pour m'avoir donné l'occasion de présenter mon travail à plusieurs reprises pendant ma thèse, pour leurs idées et les discussions pendant les réunions, pour leurs encouragements, leur énergie, leur gentillesse et leur bonne humeur. Merci à Hélène Rogniaux d'avoir pris le temps de me montrer comment fonctionnait un spectromètre de masse, ce qui a été précieux pour mieux comprendre certains aspects de cette discipline et son application.

Un énorme merci à ma famille pour m'avoir accueillie à bras ouverts pendant de longues périodes de confinement et de télétravail ; votre soutien m'a été extrêmement précieux dans ces moments où la quantité de travail à fournir était parfois très importante, et que je n'aurais sans doute pas eu la motivation d'accomplir sans vous. Merci à Papa et Maman pour leur soutien financier et moral tout au long de mes études de biologie et de bioinformatique, qui ont été un long fleuve, mais pas si tranquille.

Merci à Sophie pour les soirées séries, les irish coffees, pour tous les délicieux repas et les intoxications alimentaires qui leur étaient mystérieusement corrélées ; et surtout les promenades canines sans lesquelles mon pauvre cerveau aurait été bien moins efficacement aéré, et qui me permettaient d'atteindre mon quota de mignonitude hebdomadaire.

Merci à mon grand frère de thèse, Benjamin, pour ses discussions et son soutien ; tu auras ta thèse les doigts dans le nez.

Pour mon plus grand plaisir, les confinements ont resuscité le serveur Discord de la Pieuvre, qui permettait de rester en contact avec les collègues et amis rennais ; merci à vous pour tous les messages échangés et les drinkscords ; merci pour le soutien des autres doctorants du serveur, particulièrement à Loskann pour son aide en LaTeX.

Merci à Tchaïkovsky, Mozart, Dvořák, et surtout Jeremy Soule pour avoir été mes collègues de travail depuis plusieurs années.

TABLE DES MATIÈRES

INTRODUCTION

Les protéines dans le vivant

Les êtres vivants sont composés d'une grande variété de molécules, et les familles de molécules ont des fonctions spécifiques au sein d'un organisme. L'ADN (acide désoxyribonucléique) et l'ARN (acide ribonucléique) sont les supports de l'information génétique. Les protéines sont des enchaînements d'acides aminés (nommés résidus lorsqu'ils sont impliqués dans une séquence de protéine) produits selon l'information génétique, et sont les acteurs directs de fonctions très variées au sein du vivant. Le protéome, qui se définit par les protéines présentes dans un échantillon biologique à un instant donné, est donc représentatif des évènements qui s'y déroulent [Omenn 2012].
Un gène donné peut être à l'origine de la production de plusieurs protéines ; le protéome a donc une complexité plus importante que l'information génétique contenue dans l'ADN, appelée le génome ; il est donc difficile de savoir quelles protéines sont présentes dans un échantillon grâce à la seule information génétique. Ainsi, pour obtenir des informations précises sur le fonctionnement d'un échantillon biologique, le protéome doit être étudié directement.
L'un des éléments majeurs qui complexifient le protéome par rapport au génome est que les protéines peuvent porter des modifications chimiques. En effet, les protéines peuvent subir, lors de leur maturation, l'ajout d'une PTM (*Post-Translational Modification*), c'est-à-dire que des groupements chimiques peuvent être ajoutés ou supprimés de leurs résidus ; cette modification a un impact sur la fonction de la protéine, qui dépend de la modification et de son emplacement sur la protéine. Parmi de nombreux exemples, nous pouvons citer la phosphorylation, qui joue un rôle primordial dans la signalisation cellulaire, ou encore l'ubiquitination, qui est un signal indiquant que la protéine doit être dégradée (voir [Walsh, Garneau-Tsodikova et Gatto 2005] ou encore [Kamath, Vasavada et Srivastava 2011] pour plusieurs exemples de modifications chimiques subies par les protéines).
Lors de l'étude des protéines, il est donc nécessaire d'identifier les PTM (quelle est la nature des modifications ?) et de les localiser (à quel emplacement ?) pour connaître pré-

cisément la fonction d'une protéine donnée. Ce problème est crucial dans des thématiques aussi variées que la médecine, ou encore l'étude des protéines dans les aliments, appelée *foodomics* [Braconi *et al.* 2018]. La *foodomics* est l'une des thématiques de l'équipe BIA, de l'INRAE de Nantes, qui a encadré en partie le travail décrit dans ce manuscrit.

L'évolution de la protéomique, la spectrométrie de masse et ses défis

L'étude du protéome, ou protéomique, est une discipline qui a beaucoup évolué au cours du 20ème siècle. Les premières méthodes (dont certaines sont encore utilisées) détectaient la présence d'une protéine donnée dans un échantillon en se basant par exemple sur l'affinité naturelle entre la protéine d'intérêt et une autre protéine (comme un anticorps) qui peut être révélée par un signal lumineux ou un indicateur coloré. C'est le cas, par exemple, du test ELISA. Afin d'avoir une information supplémentaire, les protéines peuvent être au préalable séparées selon leur taille grâce à un gel qui agit comme un tamis, et la présence d'une protéine donnée peut être révélée (méthode du western blot). Cependant, ces méthodes sont restreintes à un petit nombre de protéines et ne permettent pas d'avoir une indication précise sur leur quantité [Aslam *et al.* 2017].
La spectrométrie de masse (MS) est une méthode d'analyse qui consiste à identifier les molécules (chargées, et donc sous forme d'ions) à partir de leurs masses. Elle trouve son origine au début du 20ème siècle. Cependant, ce n'est qu'à la fin des années 80 que des progrès dans le domaine de l'ionisation des molécules de taille importante ont rendu la MS applicable à l'étude des protéines. Aujourd'hui, la MS est l'une des méthodes les plus utilisées pour étudier à grande échelle les protéines et leurs modifications. C'est un outil puissant pour explorer le protéome [Aebersold et Mann 2016]. En effet, contrairement aux méthodes citées précédemment, elle permet d'identifier et de quantifier un grand nombre de protéines dans un échantillon.
La MS réunit maintenant une communauté importante et active de scientifiques à travers le monde. L'un des exemples qui montrent l'intérêt que la MS suscite est la convention annuelle de l'ASMS (*American Society of Mass Spectrometry*) qui, en 2022 à Minneapolis, a réuni plus de 6 500 personnes du monde entier ; environ 3 000 contributions au domaine y ont été présentées sous la forme de présentations orales et posters.
Cependant, le potentiel d'identification de la MS a un coût en terme de quantité de données ; en effet, une expérience de MS peut générer des centaines de milliers de spectres. De plus, les spectres sont des données complexes à interpréter pour connaître les protéines qui les ont générés. En plus de leur stockage, l'interprétation des spectres de masse nécessite

souvent une automatisation du processus, et implique parfois la gestion de bases de données pour les identifier. Ainsi, les compétences à mobiliser pour une étude de protéomique par MS sont devenues autant liées à l'informatique et au traitement des données qu'à la biochimie. En effet, depuis les années 90, de nombreux outils informatiques d'identification de spectres de masse ont vu le jour, avec des principes très différents. Les outils d'identification sont toujours aujourd'hui confrontés à de nombreux obstacles, en particulier lorsque les spectres de masse sont produits à partir de protéines portant des PTM ; celles-ci complexifient en effet leur analyse. Étant donné l'importance d'identifier les PTM, un effort doit être fourni afin d'améliorer l'identification de spectres de masse générés à partir de protéines portant des PTM. Cet effort peut notamment être fait par l'amélioration des outils informatiques d'identification et de leurs algorithmes sous-jacents.

Objectif de la thèse et organisation du manuscrit

Ce travail de thèse a été réalisé dans le cadre du projet DeepProt (ANR-18-CE45-004), qui a pour objectif une meilleure connaissance du protéome. Cela passe par l'étude des protéines et de leurs PTM.
L'objectif de ma thèse est d'apporter ma contribution aux méthodes informatiques qui visent à identifier des spectres de masse issus de protéines modifiées, en particulier lorsqu'elles portent des PTM à la fois multiples et non connues *a priori*. Pour ce travail de thèse, je me suis placée dans le contexte de la MS *bottom-up* ; en mode *bottom-up*, les protéines sont d'abord digérées en peptides (fragments de protéines). Les peptides sont ensuite eux-mêmes fragmentés dans l'appareil, et les ions produits permettent de générer des spectres. Un spectre de bonne qualité contient l'information de séquence en résidus du peptide qui l'a généré.
Pour obtenir cette information de séquence, les spectres expérimentaux (produits par l'appareil) peuvent être comparés, à l'aide d'un score de similarité, à une base de données de spectres de masse théoriques construits à partir d'une base de données de peptides. Puisque les mécanismes de fragmentation sont connus, il est possible de créer un spectre théorique "idéal" à partir d'un peptide donné. Cette étape de comparaison permet de produire un ensemble de PSM (*Peptide-Spectrum Matches*) où chaque spectre expérimental est associé à un spectre théorique, et donc à un peptide ; le peptide qui a produit le spectre peut ensuite être identifié à l'aide de la séquence du peptide candidat sélectionné. Les peptides identifiés grâce à ces spectres permettent d'inférer quelles protéines sont présentes dans l'échantillon.

Cependant, lorsque les peptides portent des PTM, l'identification de leurs spectres est plus difficile. Il faut non seulement sélectionner le bon peptide candidat, mais aussi être capable d'interpréter le PSM afin de déterminer la nature et l'emplacement des modifications éventuelles qui séparent le peptide candidat de celui qui a généré le spectre ; il est ensuite possible de reconstruire la séquence du peptide, et donc de dire que le spectre est identifié.

Pour relever ces défis, les méthodes dites OMS (*Open Mass Search*) ont été développées depuis les années 2000. Elles sont capables d'identifier un grand nombre de spectres modifiés, mais elles se basent souvent pour cela sur des PTM connues. L'une des raisons est notamment le temps de calcul requis pour considérer un grand nombre de PTM. Or, pour découvrir des fonctions biologiques originales, il est nécessaire de pouvoir identifier des spectres issus de peptides portant des PTM non connues *a priori*.

Pour atteindre cet objectif et ainsi améliorer l'identification de spectres de masse issus de tels peptides, les stratégies OMS doivent être étudiées en profondeur pour comprendre leur fonctionnement, et donc les améliorer ou en développer de nouvelles.

Pour avoir une vision claire du fonctionnement de ces stratégies, je me place dans un contexte maîtrisé où j'utilise des spectres de masses théoriques à la place de spectres expérimentaux avant d'appliquer les méthodes d'identification. Étant donné que les séquences en résidus correspondant à ces spectres sont connues, il est possible de mettre au point des critères pour juger de la qualité des identifications selon la méthode employée et ses paramètres. Pour ce faire, il faut se demander ce qui caractérise un PSM de qualité, et mettre en œuvre des méthodes informatiques pour mettre en évidence de tels PSM, qui peuvent notamment reposer sur le traitement des séquences des peptides d'un PSM.

Le premier chapitre de ce manuscrit est dédié à la présentation des protéines et des principes des méthodes de protéomique. Le Chapitre 2 introduit la MS, les spectres de masse et les méthodes d'identification de ces spectres par comparaison à une base de données de spectres théoriques. Le chapitre explique également les problématiques liées aux méthodes OMS, conçues pour identifier les spectres de protéines portant des PTM. Le Chapitre 3 décrit comment les spectres théoriques du protéome humain ont été identifiés par des stratégies OMS implémentées dans `SpecOMS` [David, Fertin, Rogniaux *et al.* 2017]. Ce logiciel est capable de comparer tous les spectres d'une base les uns aux autres grâce à un score appelé `SPC` (*Shared Peaks Count*), le nombre de pics avec des masses identiques entre deux spectres. Puisque ce calcul peut être fait rapidement et sans restriction de comparaison, `SpecOMS` est particulièrement adapté à l'analyse d'un grand nombre de

spectres dont les PSM peuvent contenir des modifications non prévues à l'avance. Le chapitre décrit également comment de nouveaux critères ont pu être mis au point pour juger de la qualité des identifications et comparer les stratégies d'identification.
L'utilisation de spectres théoriques a également permis de nous intéresser à la problématique des modifications multiples dans les PSM ; le Chapitre 4 décrit `SpecGlob`, un algorithme que nous avons mis au point afin d'interpréter les PSM dans lesquels plusieurs modifications séparent les peptides d'un PSM. Les performances de cet outil ont pu être évaluées.
Enfin, ce manuscrit se termine par une conclusion générale des travaux réalisés pendant cette thèse ; je donne également les perspectives qui me paraissent les plus intéressantes à étudier pour poursuivre ce projet de recherche.

Première partie

État de l'art

Chapitre 1

Les protéines

Sommaire

Préambule : Ce premier chapitre a plusieurs objectifs. D'abord, je présente les raisons pour lesquelles les protéines sont massivement étudiées. Pour cela, je définis ce qu'est une protéine, et décris le rôle majeur qu'elle joue au sein du vivant.
Ensuite, j'explique pourquoi la spectrométrie de masse est la méthode principale pour étudier les protéines. Pour ce faire, je décris les différentes méthodes qui existent pour analyser les protéines ainsi que les difficultés qu'elles doivent surmonter.
Enfin, ce chapitre définit un certain nombre de termes utiles à la compréhension de mon travail de thèse.

1.1 Les protéines et la protéomique

1.1.1 La synthèse des protéines

Les protéines, molécules indispensables du vivant

Le vivant est composé de molécules qui peuvent être classées en quatre catégories, avec des caractéristiques chimiques qui les différencient et qui expliquent les fonctions qu'elles ont dans un organisme. Les *oses* (appelés aussi glucides, ou simplement sucres) sont présents dans les parois cellulaires des végétaux, ou encore dans la matrice extracellulaire des animaux. Ils jouent donc un rôle de structure chez les êtres vivants, mais occupent également une place importante dans leur métabolisme énergétique. Les *lipides*, ou graisses, sont présents dans les membranes cellulaires et servent à stocker l'énergie dans le tissu adipeux. Les *acides nucléiques*, ADN (acide désoxyribonucléique) et ARN (acide ribonucléique), sont des molécules constituées d'une séquence en nucléotides qui constitue l'information génétique portée par un organisme, et qui détermine en partie la manière dont un être vivant se développe et fonctionne. Les **protéines** [1] sont des molécules de tailles et formes variées avec des fonctions qui le sont tout autant dans la cellule. Les protéines sont définies par un enchaînement de petites molécules appelées **acides aminés**.

Les acides aminés

Un acide aminé possède un groupe carboxylique (COOH), un groupe amine (NH_2) ainsi qu'un radical, ou chaîne latérale, qui est spécifique à chacun. L'extrémité du groupe carboxylique est appelée extrémité C-terminale, ou **C-ter**, et l'extrémité du groupe amine est appelée extrémité N-terminale, ou **N-ter**.
Il existe 20 acides aminés dits protéinogènes, c'est-à-dire qui rentrent dans la composition des protéines. Ils sont représentés, en plus de leur nom, par un code à trois lettres et un code à une lettre détaillés Figure 1.1, dans laquelle la formule chimique et la masse de chacun (sous forme de résidu, voir page 11) sont également indiquées. Les masses des acides aminés, et donc des protéines, sont exprimées en **Dalton** (Da). Cette unité, largement utilisée en physique et en chimie, correspond au douzième de la masse d'un atome de carbone 12 (^{12}C).

1. Les mots du manuscrit écrits en gras peuvent être retrouvés dans le glossaire, à la fin du document.

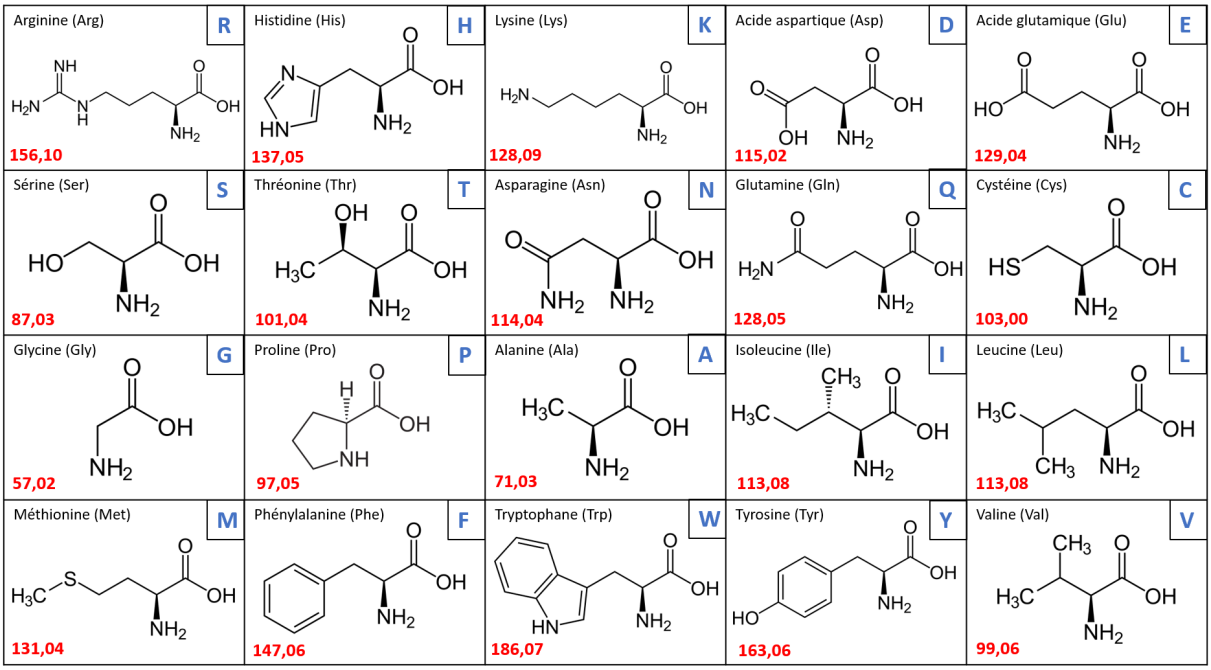


FIGURE 1.1 – **Les 20 acides aminés protéinogènes.** Pour chaque acide aminé sont indiqués le nom avec le code à trois lettres entre parenthèses, le code à une lettre en bleu, au-dessus à droite de la formule chimique, ainsi que la masse du résidu (Da) en rouge en bas à gauche de la case. Schéma de l'auteure à partir des formules de `wikipedia.org`.

Les protéines sont les produits de l'information génétique

Les protéines sont synthétisées selon l'information génétique de la cellule, donc selon son ADN. L'information portée par une portion d'ADN qui code pour une protéine (que l'on appelle un *gène*) détermine la taille d'une protéine et l'ordre des acides aminés dans celle-ci.

La synthèse des protéines à partir d'un gène se fait en plusieurs étapes (voir Figure 1.2). D'abord, à partir d'un gène, un brin d'ARN est produit dans un processus appelé *transcription*. Lors de cette étape, l'ADN double brin, constitué des nucléotides A (adénine), T (thymine), C (cytosine), G (guanine), est transformé en ARN simple brin constitué des nucléotides A, U (uracile), C, G. Ce brin d'ARN est nommé ARN messager (ARNm), car il porte le message de l'ADN qui permet l'expression d'une protéine.

En effet, des complexes moléculaires nommés ribosomes lisent le brin d'ARNm et synthétisent une protéine au cours d'un processus appelé *traduction*. Lors de la synthèse d'une protéine par la cellule, les acides aminés sont ajoutés les uns à la suite des autres selon l'information portée par un brin d'ARN ; trois nucléotides forment un codon, et chaque

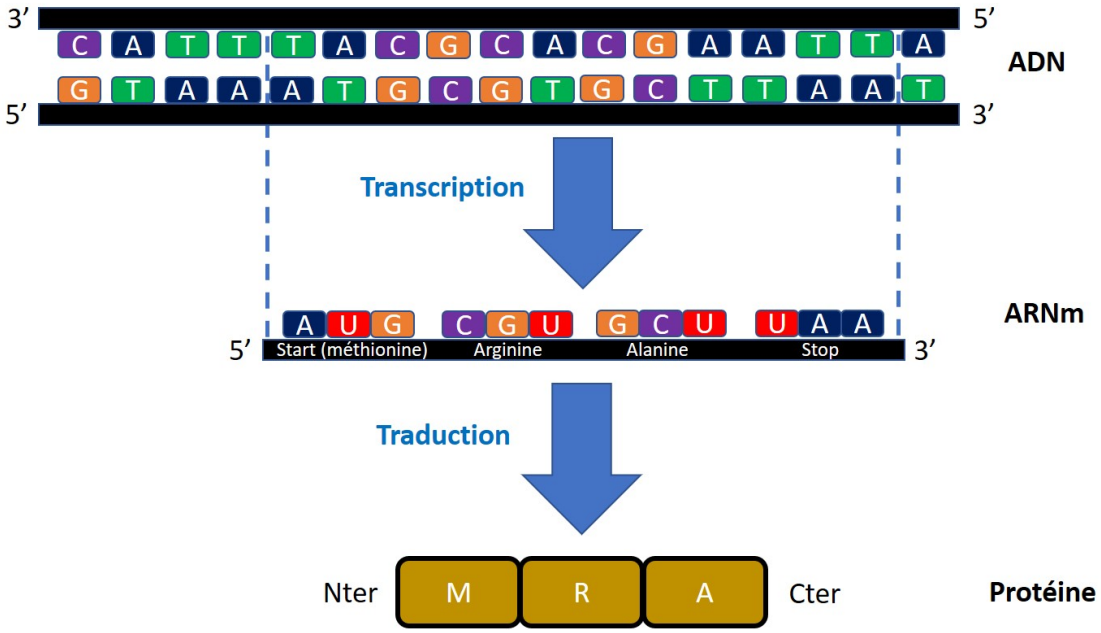


FIGURE 1.2 – **La transcription et la traduction.** Le brin d'ADN est composé de 4 types de nucléotides (A, T, C, G) associés en 2 brins. Il est traduit en ARN messager (ARNm) composé d'un seul brin et de 4 nucléotides différents, comme l'ADN, à la différence que l'uracile (U) remplace la thymine (T). Le brin d'ARNm est traduit en protéine selon ses codons et le code génétique (voir Figure 1.3). Dans cet exemple, un gène fictif de 12 nucléotides (4 codons) est transcrit (de 5' en 3') à partir du codon ATG (codon "start") jusqu'au codon TAA (codon "stop"). L'ARNm résultant fait également 12 nucléotides et 4 codons (3 codons produisant des résidus et un codon stop). L'ARNm est ensuite traduit en une protéine fictive de 3 résidus (AUG produit la méthionine, CGU produit l'arginine, et GCU produit l'alanine). Schéma de l'auteure.

codon correspond à un acide aminé spécifique selon le code génétique, décrit Figure 1.3. Il existe également un codon "start" auquel la traduction commence, et des codons "stop" où elle s'arrête. Pour chaque codon du brin d'ARNm, l'acide aminé correspondant - présent dans la cellule - est capturé par le ribosome qui avance sur le brin. Au fur et à mesure de son avancée, chaque acide aminé est ajouté à la protéine en formation.

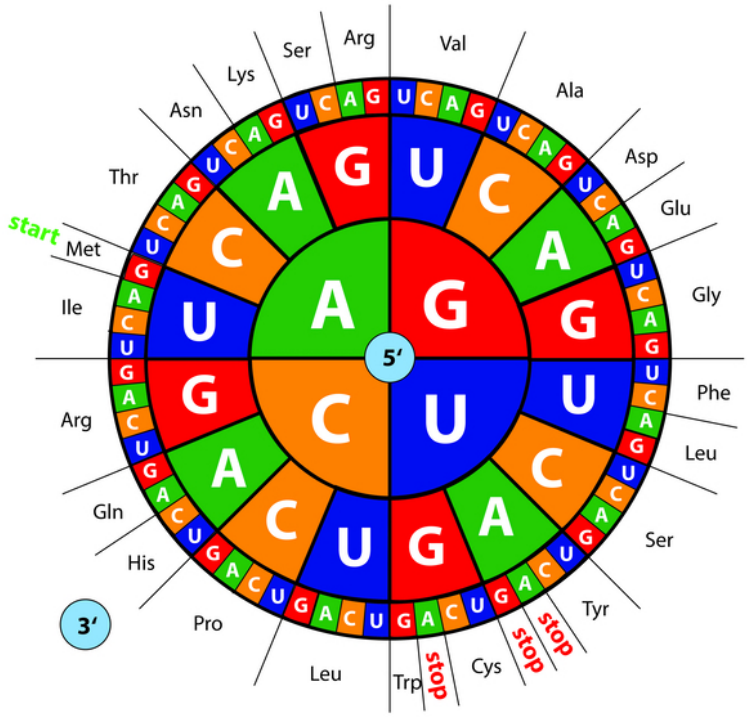


FIGURE 1.3 – **Le code génétique.** Lors de la synthèse des protéines, les acides aminés sont ajoutés les uns après les autres selon les codons contenus dans le brin d'ARN messager. La synthèse d'une protéine est souvent démarrée par un codon "start" AUG, qui code pour une méthionine. Ensuite, si on a par exemple le triplet GCU, une alanine sera ajoutée à la protéine en formation, puis une arginine si le ribosome lit le codon CGA, et ainsi de suite. La fin de la synthèse est indiquée sur le brin par un codon "stop". Schéma issu de `shutterstock.com`.

La liaison chimique entre deux acides aminés, formée par le ribosome, est appelée liaison peptidique (Figure 1.4). Elle se forme entre l'extrémité C-ter d'un acide aminé et l'extrémité N-ter d'un autre, et implique la perte d'une molécule d'eau dans la réaction. Un acide aminé est donc sous la forme d'un **résidu** lorsqu'il est impliqué dans une chaîne protéique.

Une protéine peut faire une longueur de quelques dizaines à plusieurs dizaines de milliers de résidus. On parle généralement de protéine au-delà d'une cinquantaine de résidus, alors qu'un enchaînement plus court (naturel ou issu du clivage d'une protéine) est qualifié de **peptide**. Les résidus et leur ordre dans une protéine ont une influence sur sa structure tridimensionnelle, et sa structure a un impact sur sa fonction au sein de la cellule. Les protéines ont ainsi des fonctions très variées au sein d'un organisme vivant.

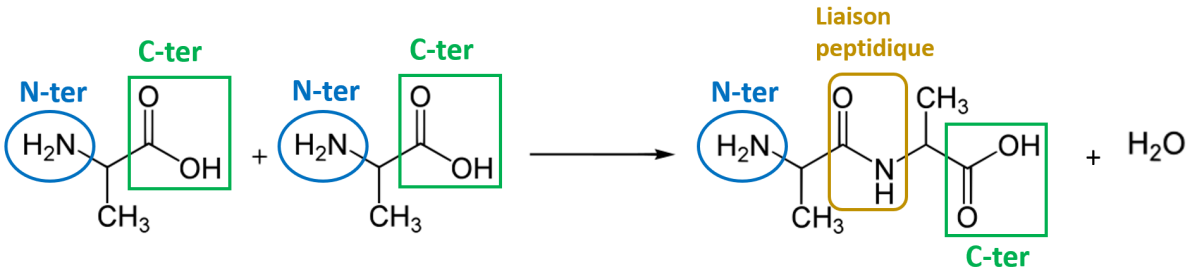
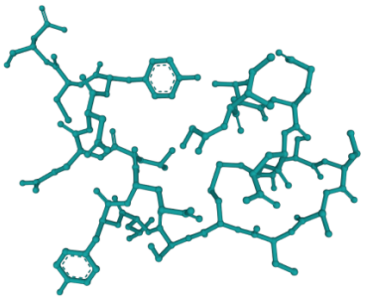
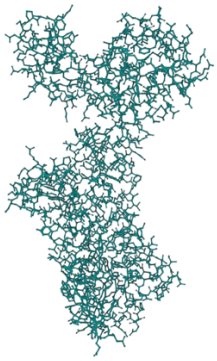
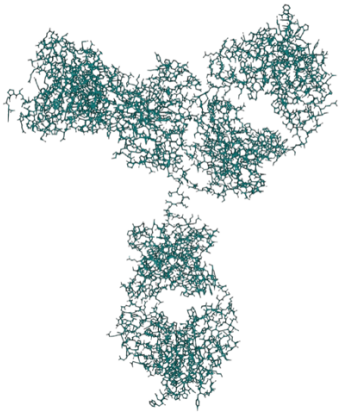


FIGURE 1.4 – **La liaison peptidique.** Lors de la synthèse des protéines, des liaisons sont formées entre chaque acide aminé. Elles impliquent la réaction du groupement C-ter du premier acide aminé avec le groupement N-ter du second, avec la perte d'une molécule d'eau au cours de la réaction. La liaison formée est appelée liaison peptidique. Sur cette figure est montrée la formation d'une liaison peptidique entre deux alanines. Schéma issu de `wikipedia.org`.

La diversité des protéines

Les protéines ont des tailles et des compositions en résidus extrêmement variées, et donc des tailles, formes et fonctions qui le sont tout autant. Cette diversité est illustrée Table 1.1. L'insuline (Table 1.1, première ligne), protéine bien connue pour son rôle important dans le métabolisme du sucre, ne fait que 51 résidus pour 5 808 Da (5,8 kDa). Au contraire, la titine, impliquée dans le fonctionnement des muscles, est une protéine de grande taille avec environ 30 000 résidus et 3 000 kDa.
La structure tridimensionnelle d'une protéine, en lien avec sa séquence, est reliée directement à sa fonction. Une **enzyme** prend la forme d'une grande molécule avec une poche qui peut catalyser certaines réactions chimiques indispensables au bon fonctionnement du métabolisme. Un récepteur cellulaire, par ses propriétés chimiques proches de celles de la membrane d'une cellule, peut s'y ancrer et permettre à la cellule de réagir à un événement extérieur capté par le récepteur en entraînant une cascade de réactions chimiques à l'intérieur de celle-ci ; c'est le cas du récepteur bêta-adrénergique (Table 1.1, deuxième ligne). La protéine peut prendre la forme en Y d'un anticorps, molécule clé du système immunitaire qui se lie à une molécule de façon spécifique, comme une IgG (Table 1.1, troisième ligne). Les protéines font partie intégrante des êtres vivants et de leur structure. Chez les mammifères, les muscles sont composés essentiellement de protéines. Chez les bactéries, les peptides sont essentiels au maintien de la paroi cellulaire (voir par exemple [PAZOS et PETERS 2019]).

TABLE 1.1 – **La diversité des protéines.**

Nom de la protéine	Représentation 3D + code PDB	Type et rôle
Insuline	 4EFX	Hormone protéique Métabolisme (régulation de la glycémie)
Récepteur bêta-adrénergique	 5D5B	Récepteur cellulaire Transmission du signal (apporté par l'adrénaline)
Immunoglobuline G (IgG)	 1HZH	Anticorps Système immunitaire (neutralisation des bactéries, virus, toxines)

Trois exemples de protéines avec des caractéristiques très différentes sont montrés ici afin d'illustrer la diversité des protéines. J'ai choisi de présenter l'insuline, le récepteur bêta-adrénergique et une IgG. Pour chaque protéine, son nom, sa représentation 3D issue de la PDB (*Protein Data Bank*, base de données qui recense les structures 3D connues de protéines) sont indiqués, ainsi que son rôle dans le vivant. Schéma de l'auteure à partir de structures de la PDB.

Les PTM et l'épissage alternatif complexifient les protéines

Les protéines sont les conséquences de l'expression du code génétique. Étudier l'ADN et l'ARN peut donc permettre d'obtenir une partie de l'information sur les protéines exprimées. Cependant, cela n'est pas suffisant pour plusieurs raisons (voir l'étude de [Wang *et al.* 2019]). En effet, de nombreux facteurs cellulaires interviennent dans la transcription de l'ADN en ARN, et la traduction de l'ARN en protéines ; ainsi il est possible de produire plusieurs protéines pour un même gène, et en différentes quantités.
La quantité de protéines est reliée en partie à la quantité d'ARN. Elle est cependant difficile à prévoir, car elle dépend aussi de l'activité de l'ARN polymérase, enzyme qui permet la transcription de l'ADN en ARN, ainsi que de la durée de vie des protéines.
Une fois que le brin d'ARN est produit à partir du brin d'ADN, le mécanisme d'épissage alternatif, qui dépend de l'activité enzymatique de la cellule, permet de combiner différentes parties du brin d'ARNm avant la traduction. Un grand nombre d'ARNm différents peuvent donc être produits à partir d'un même gène. Ces phénomènes augmentent le nombre de protéines possibles pour un seul gène.
Une autre raison de la complexité des protéines est la suivante : une protéine peut subir une ou plusieurs modifications, appelées modifications post-traductionnelles (**PTMs**, pour *Post-Translational Modifications*). Une PTM est une modification de la chaîne latérale d'un résidu par un groupement chimique - comme son nom l'indique - après la traduction de la protéine. La PTM peut avoir lieu juste après la synthèse de la protéine, ou plus tard dans la vie de celle-ci. Les protéines sont le plus souvent modifiées par d'autres protéines, des enzymes spécifiques. L'effet d'une PTM dépend de son type, de la protéine modifiée et du ou des résidu(s) impacté(s). Elle peut par exemple influencer le repliement d'une protéine, et donc sa structure, sa fonction, et même sa durée de vie et sa localisation cellulaire. Une PTM peut aussi jouer le rôle d'interrupteur pour la signalisation cellulaire ou dans des processus de dégradation de la protéine.
Je décris dans le paragraphe suivant cinq PTM très fréquentes ([Walsh, Garneau-Tsodikova et Gatto 2005]) et donc très étudiées, comme le montre la Figure 1.5 ; ces modifications sont décrites Table 1.2.
La PTM la plus courante et la plus couramment étudiée est la *phosphorylation* ; un ion phosphate est ajouté au résidu, la sérine ou la thréonine le plus souvent, par une enzyme appelée kinase. La *méthylation* consiste en l'ajout d'un groupement méthyle (CH_3) à la lysine ou l'arginine. Elle est notamment importante pour la régulation de la transcription de l'ADN, qu'elle va diminuer [Bedford et Clarke 2009]. On parle d'*acétylation* lors-

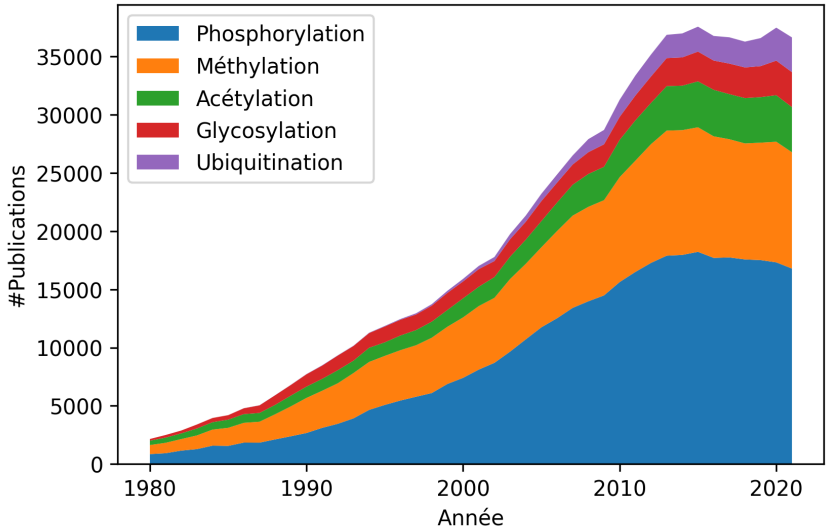
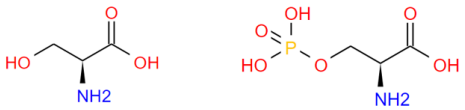
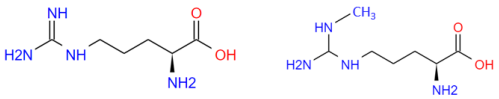
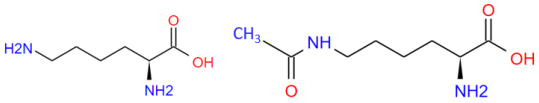
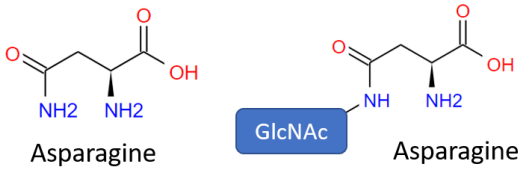
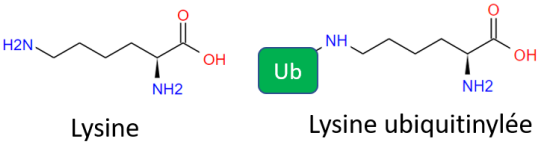


FIGURE 1.5 – **Statistiques des publications relatives à cinq PTM.** Le terme "protein "+modification a été cherché sur PubMed (par exemple "protein phosphorylation" pour la phosphorylation) et le nombre de résultats par recherche (nombre de publications) est représenté pour chaque année, de 1980 à 2021, sur ces courbes empilées ; parmi les cinq PTM présentées ici, "protein phosphorylation" renvoie le plus de résultats, "protein ubiquitination" en renvoie le moins. Schéma de l'auteure à partir des statistiques fournies par PubMed.

qu'un groupement acétyle ($CO\text{-}CH_3$) est ajouté à la lysine. Elle est également importante dans la régulation positive de la transcription de l'ADN. La *glycosylation* est l'ajout d'un sucre sur le résidu, le plus souvent sur l'asparagine ; elle est primordiale notamment dans le contrôle qualité des protéines lors de leur maturation (une protéine sera dégradée si elle est mal repliée [CHEREPANOVA, SHRIMAL et GILMORE 2016]). L'*ubiquitination* est une PTM particulière dans le sens où une protéine entière, l'ubiquitine, est ajoutée au résidu concerné, la lysine le plus souvent. L'ubiquitine est un signal qui montre que la protéine ubiquitinée doit être dégradée par la cellule.

À ce jour, plus de 1 500 PTM ont été répertoriées. Leur présence régule un grand nombre de processus biologiques [WALSH, GARNEAU-TSODIKOVA et GATTO 2005]. Les PTM peuvent interférer entre elles, c'est-à-dire s'influencer les unes les autres lorsqu'elles sont cooccurrentes ; on parle alors de "PTM crosstalk" [KHOURY, BALIBAN et FLOUDAS 2011 ; VENNE, KOLLIPARA et ZAHEDI 2014]. L'un des exemples les mieux documentés est l'interférence entre la phosphorylation et la GLcNAcylation [LAARSE, LENEY et HECK 2018]. La GLcNAcylation correspond à un type particulier de glycosylation où une N-

TABLE 1.2 – **Présentation de cinq PTM.**

Nom, description, masse (Da)	Principaux résidus affectés	Exemple de résidu sans et avec modification	Exemples de rôles
Phosphorylation + phosphate (H_3PO_4) 79,96	S, T, H		Métabolisme Signalisation cellulaire Réponse à l'environnement
Méthylation + méthyle (CH_3) 14,01	R, K		Régulation de la transcription Épissage Réparation de l'ADN
Acétylation + acétyle (CO-CH_3) 42,01	K		Régulation de la transcription Cycle cellulaire Métabolisme Virulence bactérienne
Glycosylation + oligosaccharide 220,17	N, S, T		Stabilité, repliement et contrôle qualité des protéines Compartimentalisation Interactions protéine – protéine et communication cellulaire
Ubiquitination + ubiquitine (protéine de 76 résidus) 8 564,84	K		Dégradation de la protéine

Chaque ligne est dédiée à une PTM avec son nom et sa description (molécule ajoutée au résidu concerné), sa masse, le(s) principal(ux) résidu(s) affecté(s), ainsi qu'un dessin représentant un résidu avec ou sans la PTM. Sur la dernière colonne sont indiqués pour chaque PTM ses effets et rôles les plus connus. Schéma de l'auteure, formules réalisées sur `https://chemoinfo.ipmc.cnrs.fr/LEA3D/drawonline.html`.

acétylglucosamine (molécule de glucose modifiée), ou GLcNAc, est ajoutée à un résidu (voir une illustration Table 1.2, exemple 4). La GLcNAcylation et la phosphorylation interviennent sur les deux mêmes résidus (sérine et thréonine) ; une augmentation ou une diminution de la phosphorylation de certains sites est corrélée avec une inhibition de O-GlcNAcase (enzyme responsable de la suppression de GLcNAc sur le résidu), et un changement dans les motifs de GLcNAcylation est corrélé avec une inhibition des phosphatases (enzymes dont la fonction est de supprimer un groupe phosphate d'un résidu, et donc d'"annuler" la phosphorylation).
L'importance des PTM dans le rôle des protéines, et donc dans le fonctionnement de la cellule, fait que certaines maladies (cancer, Alzheimer, diabète, etc.) peuvent se développer si les PTM sont affectées ; voir [Kamath, Vasavada et Srivastava 2011 ; Khoury, Baliban et Floudas 2011] pour des états de l'art à ce sujet.
Les PTM sont donc naturellement un enjeu crucial pour l'analyse des protéines, et certaines restent à découvrir.

1.1.2 Définition de la protéomique et de ses difficultés d'analyse

Le terme **protéomique** date des années 1990. Dans [Noor *et al.* 2021], la protéomique est définie comme "l'identification et la quantification des ensembles de protéines - ou protéomes - qui sont exprimés dans une dimension spatio-temporelle donnée, par une cellule, un tissu ou un organisme, dans des conditions données". Compte tenu du rôle primordial joué par les protéines dans les organismes, les analyser, c'est-à-dire les quantifier ou les identifier, est une étape inévitable lors de l'étude d'un échantillon biologique, et qui permet de le caractériser finement. Qu'est-ce qui différencie une cellule saine d'une cellule malade ? Quelles sont les caractéristiques d'un organisme donné ? La réponse à ces questions passe notamment par l'analyse des protéines.
L'épissage et les PTM expliquent que, pour environ 20 000 gènes qui sont prédits comme codant pour des protéines chez l'être humain, on estime que plusieurs centaines de milliers de protéines différentes sont produites [Aebersold, Agar *et al.* 2018]. Ces éléments sont difficilement prévisibles par des études génomiques ou transcriptomiques. Ainsi, pour examiner l'activité des protéines, il est nécessaire de les étudier directement. La protéomique serait donc le moyen le plus pertinent d'étudier un système biologique [Cox et Mann 2007].
Analyser les protéines directement soulève plusieurs difficultés. La première est qu'il n'existe aucune manière d'amplifier les protéines, alors que le matériel génétique peut

être amplifié par PCR. Toutes les analyses protéomiques doivent donc être faites avec le matériel de départ. Ensuite, la présence et la quantité des protéines varient selon le temps, les cellules, ou même au sein d'une cellule, contrairement à l'ADN qui reste stable. Ainsi, les études se focalisant sur les protéines apportent des informations précieuses, mais les expériences à réaliser pour les obtenir sont plus nombreuses et plus complexes.
L'une des principales difficultés de la protéomique est la complexité des mélanges. En effet, un échantillon biologique peut contenir de nombreuses protéines, à cause du nombre de protéines naturellement présentes dans un tissu vivant, mais il faut aussi tenir compte des protéines qui peuvent contaminer l'échantillon dans le laboratoire (par exemple la kératine humaine ou ovine, les protéines de gants de latex, etc.). Lors d'une analyse, il est donc possible de détecter une protéine à la place de la protéine d'intérêt (on a alors un faux-positif) si la méthode est trop sensible. Un autre problème est la gamme dynamique des protéines, c'est-à-dire que les protéines peuvent être présentes avec des abondances très différentes. La protéine à étudier peut donc être présente dans l'échantillon, mais en une quantité faible ou très faible, camouflée par les protéines les plus abondantes. Afin d'améliorer l'efficacité des techniques d'analyse, il va donc falloir mettre en œuvre des méthodes de séparation pour accéder aux protéines minoritaires, pour augmenter leur concentration, et ainsi les rendre détectables ou améliorer le signal de la protéine d'intérêt. Cette méthode peut être qualifiée de "purification" si on n'extrait qu'une seule protéine.
Une autre difficulté est de savoir où une protéine est exprimée dans la cellule, et quand. De plus, on peut souhaiter détecter les différentes formes d'une protéine en termes d'épissage alternatif, de variations en termes de résidus ou SAP (*Single Amino acid Polymorphism*). Les PTM complexifient également l'étude des protéines, pour des raisons qui sont abordées dans le Chapitre 2, Section 2.6.1.
L'étude des protéines soulève donc des difficultés d'ordre expérimental, car il faut séparer la protéine voulue, ou ses différentes formes, du reste des protéines par des méthodes de purification. Mais la difficulté est aussi d'ordre instrumental (la technique ou l'appareil employé pourra-t-il détecter la/les protéine(s) d'intérêt ?) et enfin, la phase d'analyse peut aussi être une étape exigeante à cause de la complexité des données à étudier.

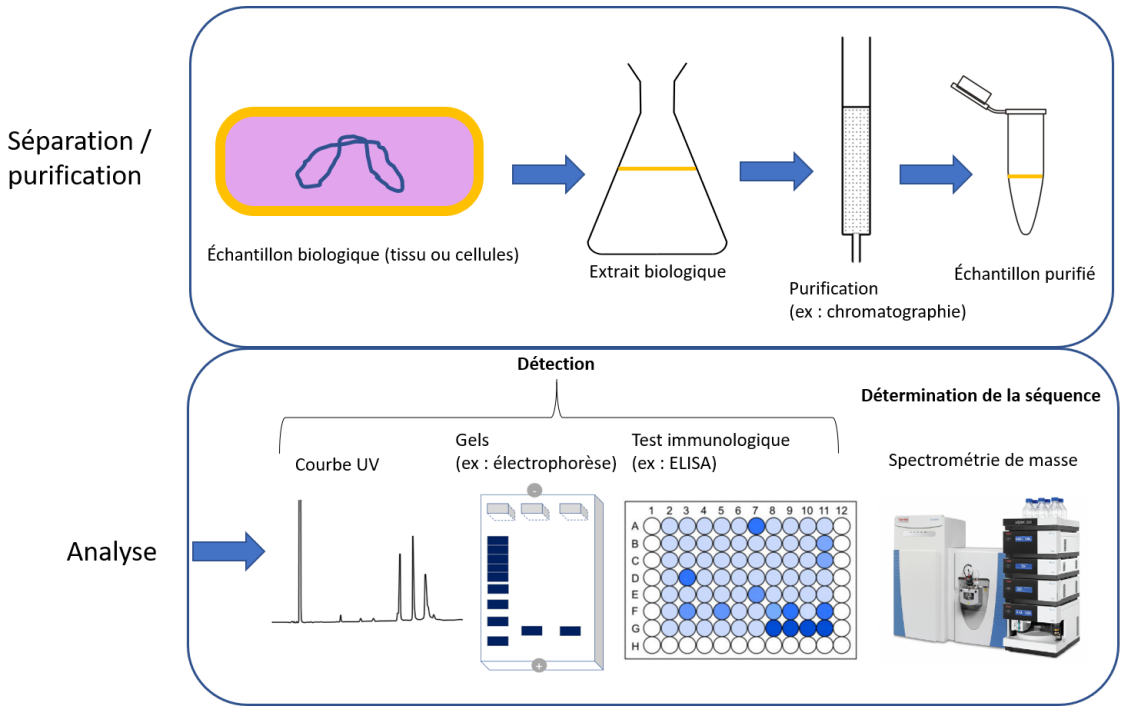


FIGURE 1.6 – **Les grandes étapes d'une étude de protéomique.** Une étude protéomique peut être divisée en deux étapes majeures : la séparation des protéines, ou bien la purification d'une protéine d'intérêt à partir d'un extrait biologique ; l'analyse de l'échantillon purifié par les méthodes présentées dans cette section, c'est-à-dire par électrophorèse, ou bien par des méthodes immunologiques (basées sur des anticorps). Schéma de l'auteure à partir des images issues de Pixabay (erlenmeyer), Wikipédia (colonne de chromatographie), FlyClipart (tupe Eppendorf), `www.hplc.eu/Downloads/ACE_Guide_Peptides.pdf` (courbe UV) et `www.jimmunol.org/content/jimmunol/early/2018/03/14/jimmunol.1701737.full.pdf` (test ELISA) et `https://iccf.uca.fr/services/spectroscopie-et-masse/spectrometrie-de-masse#/admin` (spectromètre de masse).

1.2 Les outils de protéomique

1.2.1 Les méthodes de séparation et d'analyse de protéines

La protéomique peut être divisée en deux grandes étapes, la séparation des protéines et leur analyse. Elles sont illustrées Figure 1.6. Pour un état de l'art sur l'éventail de méthodes de protéomique, voir [ASLAM *et al.* 2017].

La séparation des protéines

Plusieurs groupes de méthodes ont été développés afin de séparer les protéines.
L'une des techniques les plus utilisées pour séparer les protéines est la *chromatographie*. Elle consiste à faire passer la solution de protéines dans une colonne contenant une substance spécifique qui permet de séparer les protéines selon certaines caractéristiques.
La chromatographie d'*exclusion-diffusion* (voir [Burgess 2018] pour le principe ainsi qu'un bref historique) fait passer le mélange de protéines dans une colonne contenant des billes poreuses. Plus la protéine est petite, plus elle aura tendance à passer par les pores des billes, et donc plus elle mettra du temps à parcourir la colonne.
Pour la chromatographie d'*affinité* (voir l'état de l'art [Rodriguez *et al.* 2020]), la molécule à purifier est fixée de façon réversible à un ligand présent dans la colonne. Alors que les autres molécules passent dans la colonne sans s'y accrocher, la molécule à purifier s'y fixe. On ajoute ensuite à la colonne un produit qui a une affinité encore plus importante avec le ligand. Pour donner un exemple concret, il est possible de purifier une protéine produite par un organisme dont le génome est modifié par génie génétique (comme une bactérie) de façon à ce que la protéine soit exprimée avec une zone de plusieurs histidines (appelée étiquette histidine). Une fois la totalité des protéines extraites des cellules, on fait passer le mélange dans une colonne avec des billes recouvertes de nickel. L'étiquette histidine a une affinité importante avec le nickel, ainsi la protéine d'intérêt restera fixée à la colonne. On rince ensuite la colonne avec de l'imidazole, qui est un composé qui a une affinité encore plus importante avec le nickel que l'étiquette histidine ; la protéine d'intérêt est donc éluée (c'est-à-dire sort de la colonne) et peut être récupérée avec une concentration importante, voire presque pure.
Les protéines possèdent à leur surface des poches hydrophobes (c'est-à-dire qui repoussent l'eau). La chromatographie en *phase inverse* exploite cette propriété ; elle est réalisée avec une colonne composée de billes avec de longues chaînes de carbones hydrophobes. Selon leurs propriétés de surface, les protéines injectées auront une affinité plus ou moins importante avec la colonne et seront donc éluées avec un temps caractéristique appelé *temps de rétention*.
La séparation des protéines peut également se faire par *électrophorèse* (voir l'état de l'art [Zhu, Lu et Liu 2012]). Il s'agit alors de faire migrer les molécules dans un gel - pour les protéines, c'est un gel de polyacrylamide - au moyen d'une charge électrique. Le gel se comporte comme un tamis moléculaire dans lequel les protéines migrent plus ou moins vite selon leur masse.

Au cours de cette purification, il est important de conserver la stabilité des protéines, et donc les garder solubles (en solution) grâce à des mélanges dédiés. Cela permet de les conserver dans des conditions proches de leur environnement biologique (et donc d'obtenir des informations plus précises sur leur fonction), et de faire en sorte que les techniques telles que la chromatographie soient efficaces. En effet, si elles ne sont pas solubilisées, les protéines peuvent précipiter, c'est-à-dire former des agrégats qui seront inutilisables pour la suite de l'analyse.

La détection et l'identification des protéines

Une fois le mélange simplifié, voire purifié, de nombreuses méthodes existent pour détecter et/ou identifier une ou plusieurs protéine(s) d'intérêt. Ces méthodes utilisent des principes très variés, de la détection immunologique à l'utilisation de rayons X, en passant par la spectrométrie de masse.
Lorsque des protéines ont migré sur un gel d'électrophorèse, on peut révéler leur présence avec un colorant spécifique. En comparant leur progression sur le gel par rapport à des protéines étalons, il est possible de déterminer leur masse pour les identifier.
Les protéines ont également la propriété d'absorber les rayonnements ultraviolets (UV). En mesurant l'absorbance des UV d'un mélange - comme une solution qui sort d'une colonne de chromatographie - on peut détecter la présence des protéines.
Des anticorps se lient naturellement à des protéines spécifiques ; ils peuvent être produits à des fins de détection de protéines, afin de reconnaître une protéine d'intérêt. Plusieurs méthodes de détection de protéines impliquant des anticorps, dites méthodes immunologiques, ont été développées. Une fois une électrophorèse réalisée, il est possible de la combiner avec la technique du *western blot* (voir [Meftahi *et al.* 2021] pour un état de l'art sur le western blot et ses applications). Elle consiste à transférer les protéines du gel sur une membrane de nitrocellulose après la migration de celles-ci. On utilise ensuite un anticorps spécifique de la protéine cible, qui s'y fixe si celle-ci est présente. La présence de l'anticorps peut ensuite être révélée de différentes manières selon sa conception, par fluorescence par exemple. Le western blot s'est notamment montré utile pour étudier les protéines ayant subi des PTM, avec par exemple des anticorps anti-phosphotyrosine (voir un exemple dans [Sawasdikosol 2010]).
La méthode *ELISA* (plus d'information dans [Alhajj et Farhana 2022] disponible sur `https://www.ncbi.nlm.nih.gov/books/NBK555922/`) utilise un principe similaire : des anticorps spécifiques à la protéine cible sont fixés au fond d'une plaque. L'échantillon est

ajouté, puis un rinçage est effectué. On utilisera ensuite un second anticorps qui reconnaît la protéine d'intérêt, puis un troisième qui reconnaît le second anticorps. Ce troisième anticorps permet ensuite une révélation, par exemple par l'obtention d'un substrat coloré.
Les puces de protéines ("protein microarrays") impliquent l'utilisation d'une banque de protéines qui jouent le rôle de sondes, par exemple des anticorps, fixées au fond de très petits puits (100 à 300 µm). L'échantillon à analyser est ensuite déposé dans les puits ; les protéines de l'échantillon se lient aux sondes selon leur spécificité. La présence des protéines ayant interagi avec les sondes peut enfin être révélée avec des anticorps spécifiques.
L'avantage des puces de protéines par rapport à la méthode ELISA est la possibilité de détecter un grand nombre de protéines (caractéristique haut-débit) à l'aide d'une quantité d'échantillon nécessaire plus basse.
Ces techniques immunologiques permettent de détecter la présence de protéines connues, ou de découvrir des interactions entre protéines. Mais, pour avoir une connaissance fine d'une protéine à analyser, des méthodes plus avancées sont nécessaires.
Nous avons déjà évoqué que la structure 3D d'une protéine est directement reliée à sa fonction. Certaines études se focalisent donc sur la structure des molécules, pour développer des médicaments ou élucider le mécanisme de certaines enzymes.
Pour découvrir la structure 3D d'une protéine, on peut utiliser la *cristallographie aux rayons X*. Le principe est le suivant : la protéine est purifiée de façon à former un cristal. Cette structure ordonnée sera exposée à des rayons X qui ont la propriété de se diffracter selon la densité des électrons. En analysant la carte de densité électronique obtenue, il est possible d'élucider une partie de la structure de la protéine.
Une autre façon d'étudier finement une protéine est de déterminer sa séquence en résidus. Séquencer une protéine permet de l'identifier en comparant la séquence découverte à celles des protéines connues dans des bases de données dédiées (voir la Section 1.2.2) ou encore d'obtenir des informations sur la fonction de la protéine avec des liens séquence/structure/fonction déjà établis.
Le séquençage par la dégradation d'Edman a été développé en 1950 pour obtenir la séquence en résidus des protéines à partir de leurs peptides [Edman 1949]. Il consiste à utiliser des réactions chimiques pour éliminer et identifier, de façon séquentielle, le résidu en N-ter de la protéine d'intérêt.
Depuis les années 90, la dégradation d'Edman est moins utilisée avec l'avènement de la spectrométrie de masse. Cette dernière permet d'étudier des mélanges complexes de protéines, de déterminer leurs séquences et de découvrir les PTM qu'elles portent (voir [G.

Zhang *et al.* 2014] pour un état de l'art).
Pour ses applications multiples, la spectrométrie de masse est aujourd'hui la méthode principale pour étudier les protéines. En effet, elle permet d'identifier les protéines, mais aussi de les quantifier. La spectrométrie de masse peut être utilisée de façon exploratoire, c'est-à-dire pour découvrir de nouvelles protéines ou étudier un nouvel échantillon, ou au contraire de façon ciblée. Elle permet aussi de comparer plusieurs échantillons. La spectrométrie de masse est décrite en détails dans le Chapitre 2.

1.2.2 Les bases de données de protéomique

La description des protéines peut prendre des formes très variées [Chen, H. Huang et Wu 2017]. Elles peuvent représenter des séquences de protéines, mais aussi leurs structures 3D, les interactions entre protéines, ou encore les PTM qu'elles portent.
Nous avons pu voir que chaque résidu peut être représenté par une lettre ; une protéine peut donc être écrite sous forme de texte et déposée dans une base de données une fois que sa séquence est élucidée. Une protéine est souvent présentée au format FASTA. Dans un fichier au format FASTA (.fasta ou .fa), la première ligne ("header", ou en-tête) commence par un chevron ">" suivi par une description de la séquence du fichier. Ensuite, la séquence biologique est écrite avec 60 lettres par ligne. Étant donné que les nucléotides sont également souvent représentés sous forme de lettres, le format FASTA est également utilisé pour représenter les séquences génomiques. De par sa simplicité, ce format est devenu un standard dans le domaine de la bioinformatique. Les protéines peuvent ainsi être décrites et leurs séquences stockées dans des bases de données.
De nombreuses bases de données de référence de séquences de protéines existent. Elles sont majoritairement prédites à partir des séquences de nucléotides obtenues par les études de génomique. L'une des plus utilisées est UniProt [UniProt Consortium 2021]. Elle concentre les efforts d'équipes d'instituts de bioinformatique européen (EBI) et suisse (SIB) pour mettre à disposition des séquences de protéines et leurs annotations afin de soutenir la recherche en biologie et bioinformatique. Le cœur d'UniProt est la base de données UniProt Knowledgebase (UniProtKB) divisée en deux parties : UniProtKB/Swiss-Prot propose des séquences annotées manuellement (environ 568 000 séquences en septembre 2022), alors qu'UniProtKB/TrEMBL comporte des séquences annotées automatiquement (environ 227 millions d'entrées en septembre 2022).
Prenons l'exemple de l'albumine, la protéine la plus abondante du sérum humain. Cette protéine est composée de 585 résidus avec un poids total de 67 KDa. Elle peut être repré-

```
>sp|P02768|25-609
DAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAE
NCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEV
DVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAACLLP
KLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTK
VHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPA
DLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKC
CAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVST
PTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTES
LVNRRPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKAT
KEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAASQAALGL
```

FIGURE 1.7 – **Séquence de l'albumine.** La séquence de l'albumine, téléchargée à partir d'UniProt, est présentée au format FASTA. "sp" signifie que la séquence a été téléchargée à partir d'UniProtKB/Swiss-Prot, "P02768" est l'identifiant unique de l'albumine humaine chez UniProt. Dans l'en-tête, indiqué par ">", "25-609" indique que les résidus 25 à 609 de la protéine sont inscrits. En effet, les 24 premiers résidus sont supprimés lors de la maturation de la protéine. L'albumine finale comprend donc 585 résidus, et commence par un D (acide aspartique, voir le code à une lettre Figure 1.1, page 11).

sentée sous la forme de 585 lettres au format FASTA (voir Figure 1.7).

Ensembl [CUNNINGHAM *et al.* 2021] est une base de données de séquences génomiques, créée pour accéder facilement aux génomes de nombreuses espèces, pour parcourir les séquences d'ADN, les gènes et les annotations correspondantes (ARN et protéines).

Les séquences protéiques peuvent notamment être utilisées pour prédire la structure 3D des protéines, et donc leurs fonctions (voir [JISNA et JAYARAJ 2021] pour un état de l'art sur l'importance de la structure d'une protéine et les différentes méthodes pour la prédire). Ces informations, ajoutées aux structures 3D déterminées par des méthodes expérimentales (nous avons déjà évoqué la cristallographie aux rayons X) peuvent aussi être stockées dans des bases de données. La PDB (*Protein Data Bank*, voir `rcsb.org` et [BERMAN *et al.* 2000]) stocke les structures 3D des protéines au format .pdb dans lequel sont indiquées les positions des atomes des protéines. Les structures peuvent ensuite être analysées pour des études plus poussées, ou simplement visualisées (comme je l'ai fait par exemple pour la Table 1.1).

Nous avons évoqué précédemment l'importance des modifications (PTM) que subissent les résidus des protéines au cours de leur vie. Certaines bases de données se donnent pour objectif de les répertorier. Unimod [CREASY et COTTRELL 2004], orientée spectrométrie de masse, a été créée pour mettre à disposition de la communauté scientifique les caracté-

ristiques des PTM découvertes au cours d'études de protéomique, de façon standardisée. Sont indiqués pour chaque PTM le nom, le résidu modifié - ou bien une zone du peptide ou de la protéine (N-ter ou C-ter) - et le décalage de masse induit par la PTM.
Certaines bases de données placent les PTM dans leur contexte biologique. La base de données RESID [Garavelli 2004] compile les PTM annotées dans UniProt.
PSI-MOD [Montecchi-Palazzi *et al.* 2008] est le fruit du travail de la Proteomics Standards Initiative (PSI) fondée par HUPO (HUman Proteome Organization). La PSI a pour objectif de créer des standards pour représenter et échanger les données de spectrométrie de masse, dont les PTM, qui sont représentées dans PSI-MOD selon les critères de la PSI. Elle intègre notamment les données des bases Unimod et RESID. dbPTM [K.-Y. Huang *et al.* 2019] est entretenue depuis 2006 et elle repose sur plus de 30 outils pour fournir des données volumineuses sur les PTM et leur environnement dans une protéine ; elle donne également des indications sur le phénomène d'interférence de PTM.
Certaines bases de données sont spécifiques à certaines PTM [Kamath, Vasavada et Srivastava 2011]. La phosphorylation étant la PTM la plus courante et la plus étudiée, plusieurs bases de données de référence lui sont consacrées. Déjà en 1997 la création de la Protein Kinase Resource (PKR) [Smith *et al.* 1997] était motivée par les informations disponibles pour de nombreuses protéines, mais avec des informations limitées pour chacune ; cette base de données réunissait donc l'information de plusieurs autres bases de données afin de donner des informations détaillées sur les kinases et leurs cibles. PhosphoNet (`http://www.phosphonet.ca/`), encore régulièrement utilisée, répertorie les sites de phosphorylation chez les protéines de l'être humain.
Ces bases permettent de concentrer les données variées (séquences, structures, PTM, etc.) issues d'études diverses, ce qui permet à la fois d'y accéder facilement pour faire avancer une étude particulière, et de prendre en compte une grande quantité de données pour des analyses bioinformatiques. Une connaissance précise des PTM peut être cruciale notamment lors de l'analyse des protéines par spectrométrie de masse ; les bases de données évoquées dans cette section se montrent alors utiles. La problématique des PTM et la place des bases de données dans l'interprétation des spectres de masse seront présentées en détails dans le Chapitre 2.

1.3 Conclusion

Dans ce chapitre, nous avons pu voir que les protéines sont des molécules qui ont de grandes variabilités en termes de séquences, taille, structure, et donc en termes de fonction chez les êtres vivants.
Elles sont la conséquence de l'expression génétique d'un organisme, mais il n'est pas possible avec les moyens actuels de prévoir complètement le comportement des protéines à partir de l'ADN. En effet, à partir d'un même gène, de nombreuses protéines peuvent être produites, notamment avec les PTM subies par les protéines. Elles doivent donc être étudiées directement.
Pour ce faire, la protéomique déploie des ressources variées. Plusieurs étapes sont souvent nécessaires pour étudier les protéines, de leur séparation et purification à leur analyse par des techniques adaptées à ce que l'on cherche à faire : détecter, quantifier ou bien identifier les protéines. En parallèle de l'utilisation de ces méthodes, il est possible de s'aider de différents types de bases de données, dont quelques exemples sont fournis dans ce chapitre.
La méthode la plus utilisée pour étudier les protéines et leurs PTM est la spectrométrie de masse. Les avantages de la spectrométrie de masse expliquent qu'elle soit utilisée dans de nombreuses études pour identifier les protéines, et que de nombreux efforts soient dédiés à l'amélioration de cette technique. C'est dans ce cadre que se situe mon travail de thèse, qui a pour objectif d'améliorer l'identification des peptides modifiés par spectrométrie de masse. La spectrométrie de masse, les données qu'elle produit ainsi que les défis auxquels elle est confrontée sont décrits en détails dans le chapitre suivant.

Chapitre 2

La spectrométrie de masse en protéomique

Sommaire

Préambule : Ce chapitre fournit un bref historique de la spectrométrie de masse (MS) ainsi qu'une introduction aux principes que cette méthode utilise pour identifier les molécules en général. J'y présente ensuite l'utilisation de la MS pour la protéomique en particulier. Sont détaillées ici également les caractéristiques des données produites par MS - c'est-à-dire les spectres de masse - à partir des protéines, ainsi qu'un état de l'art des principales méthodes d'analyse qui existent pour les exploiter. Enfin, je discute de la problématique au cœur de mon travail de thèse : l'analyse par MS des protéines et peptides qui portent des modifications, chimiques ou de séquence. Celles-ci complexifient leur analyse et ont motivé la naissance des méthodes Open Mass Search (OMS), qui sont définies et discutées à la fin de ce chapitre.

2.1 Principe et historique de la spectrométrie de masse (MS)

2.1.1 Naissance et principe de la MS

La spectrométrie de masse (MS) est une méthode d'analyse dont les grands principes remontent à la fin du XIXème siècle, avec Joseph J. Thomson. Ce physicien s'est intéressé aux rayons cathodiques, qui sont des faisceaux d'électrons produits lorsqu'une tension est appliquée entre deux électrodes dans un environnement sous vide. Thomson a pu dévier les rayons cathodiques grâce à un champ électrique, ce qui l'a amené à étudier la déviation des ions (molécules chargées) dans ce même contexte, déviation qui dépend à la fois de la masse m de l'ion et de sa charge z. La déviation est d'autant plus importante que la masse est faible et que la charge est importante, et donc que le rapport masse sur charge m/z est faible. En plus de prouver expérimentalement l'existence de l'électron, ce qui a valu le prix Nobel de physique à Thomsom en 1906, ces expériences ont posé les bases qui ont permis la mise au point de la MS au début du XXème siècle. Le principe de la MS est le suivant : en mesurant la déviation d'un ion dans un champ électrique ou magnétique, on est capable de déterminer son ratio m/z, ce qui permet de formuler des hypothèses sur l'identité de l'ion étudié.

Le spectromètre de masse est un appareil qui contient trois parties principales (voir Fi-

gure 2.1). Pour qu'une molécule soit détectable en MS, elle doit être chargée (ionisée). Ainsi, pour analyser un échantillon par MS, celui-ci sera d'abord ionisé par une source d'ionisation. Ensuite, un analyseur va séparer les ions selon leur valeur de m/z. Enfin, le détecteur va produire un signal électrique dont l'intensité dépend du nombre d'ions rencontrés. Il existe de nombreux types de sources, d'analyseurs et de détecteurs, et leur choix dépend notamment du type des molécules que l'on souhaite étudier.

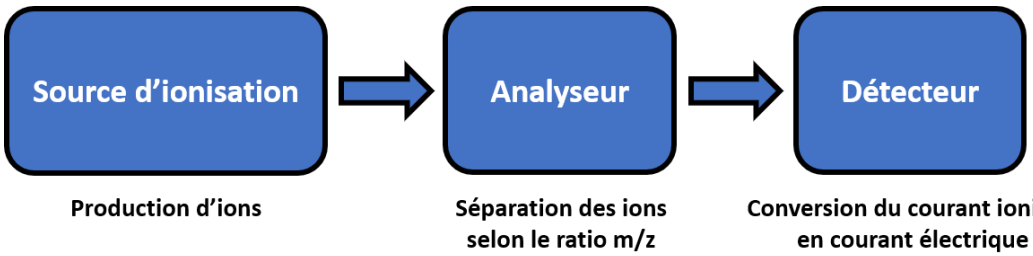


FIGURE 2.1 – **Les trois principaux composants d'un spectromètre de masse.** Adaptée de `wikipedia.org`.

2.1.2 Développement de la MS au cours du XXème siècle

Certaines découvertes importantes du XXème siècle ont été rendues possibles par la MS, notamment celles de certains isotopes, comme le deuthérium, un isotope de l'hydrogène découvert en 1931 par Harold Hurey, ce qui lui a valu le prix Nobel de chimie en 1934. Dans les années 1940, la MS a prouvé son utilité dans des domaines variés. Elle a par exemple été utilisée dans l'industrie du pétrole afin d'étudier les hydrocarbures. Elle a également mené à des progrès dans des sciences aussi diverses que la géologie (en aidant à estimer l'âge de la Terre par le calcul du ratio des isotopes de plomb dans la croûte terrestre) ou la physique nucléaire, en permettant la séparation des isotopes de l'uranium (voir l'état de l'art [GRIFFITHS 2008] pour plus d'informations). À la fin des années 1940, les spectromètres de masse sont construits et vendus par des entreprises. À cette époque, l'analyseur en *temps de vol* (TOF, *Time Of Flight*) est aussi développé. Il consiste à accélérer les ions avec un champ électrique d'une force connue. En mesurant le temps que met la particule à atteindre le détecteur, on peut calculer son ratio m/z. Deux ans plus tard, la *résonance cyclotronique ionique* (ICR) est appliquée à la MS. Elle consiste à entraîner la rotation des ions dans un champ magnétique afin de les séparer ; les ions avec le même m/z bougeront de façon synchrone.
Dans les années 1950, les analyseurs de type *quadrupôle* sont développés. Ils consistent à

faire passer des ions entre quatre cylindres électriques et à les analyser selon la stabilité de leur trajectoire. À ce jour, ce type d'analyseur est le plus populaire dans le monde.
La **résolution** et la **précision** d'un spectromètre de masse sont deux valeurs cruciales qui ont un impact important sur l'exploitabilité des résultats. La résolution indique la plus petite différence de m/z que l'on peut mesurer (peut-on faire la différence entre 1 et 2 Da, ou bien 0,1 et 0,2 Da ?). La précision indique à quelle point la valeur mesurée est proche de la valeur réelle. Ces valeurs se mesurent soit directement en Da, soit en **ppm** (partie par million, le ppm est une fraction valant un millionième) par rapport à la masse de l'ion. Dans les années 1970, une grande étape a été franchie du point de vue de la résolution lorsque la transformée de Fourier a été appliquée à des signaux d'ICR. Cette transformée permet l'acquisition du signal de tous les ions simultanément, ce qui augmente significativement la résolution (voir Section 2.1.3) tout en réduisant le temps d'acquisition.

2.1.3 La MS pour l'analyse des protéines

Malgré ces progrès, les molécules de taille importante, comme les grandes molécules biologiques (protéines, sucres, acides nucléiques), ne pouvaient pas encore être correctement étudiées par MS. En effet, les sources d'ionisation, qui reposaient sur une collision en phase gazeuse, étaient trop violentes pour ces molécules, qui se fragmentaient et se décomposaient avant l'analyse. À la fin des années 1980, le développement de deux méthodes d'ionisation douce, adaptées aux macromolécules, a permis l'application de la MS à la protéomique : le *MALDI* et l'*électrospray*.
Le MALDI (Matrix Assisted Laser Desorption Ionisation [Karas et Hillenkamp 1988]) consiste à mélanger l'échantillon à un solvant constitué notamment d'éléments acides pour charger les molécules (protonation), et une matrice solide est obtenue à partir de l'évaporation du mélange. Un laser va ensuite frapper la matrice, ce qui va permettre de charger les molécules et de les diffuser dans l'appareil avant de mesurer leur m/z.
L'électrospray [Fenn *et al.* 1989] consiste à faire passer l'échantillon en phase liquide (dans un solvant dédié qui permet une ionisation) par un cône, nommé cône de Taylor, muni d'une ouverture étroite et soumis à une tension électrique. La tension provoque la formation de gouttelettes d'échantillon de l'autre côté du cône, et l'évaporation du solvant des gouttelettes va y entraîner un surplus de charge, et donc ioniser les molécules de l'échantillon. À ce stade, on obtient les ions qui pourront être étudiés dans les autres parties de l'appareil.

Au moment de la naissance du MALDI, la précision des appareils était insuffisante, de l'ordre de 2 Da (plusieurs centaines de ppm). Des progrès ont permis de faire passer cette résolution à 10 ppm, puis de l'ordre de 1 ppm grâce aux analyseurs à transformée de Fourier [Clauser, Baker et Burlingame 1999].
Grâce à ses développements pour l'analyse des protéines, la MS est très utilisée. Elle reste prédominante pour analyser (quantifier ou identifier) les protéines, à la fois pour la qualité de sa résolution et la capacité de la méthode à analyser des échantillons volumineux en peu de temps (haut-débit, voir [K. Zhang *et al.* 2019] pour des ordres de grandeur de temps d'une expérience de MS). De nombreuses connaissances ont pu être apportées dans le domaine de la biologie par l'analyse des protéines par MS. Dans le paragraphe suivant, je donne quelques exemples d'études représentatives des possibilités variées de la MS.
Comme évoqué dans le Chapitre 1, les protéines doivent être étudiées directement car étudier le matériel génétique, et même l'ARN, ne permet pas d'obtenir toutes les informations sur l'expression des protéines, leurs différentes formes, leur quantité et localisation. L'étude de [Wang *et al.* 2019] s'est justement servie de la MS pour s'attaquer à cette problématique, et montrer les liens entre ARNs et protéines dans 29 tissus humains en combinant protéomique et transcriptomique. La protéomique par MS a également été utilisée afin d'étudier les caractéristiques de certaines bactéries pathogènes. L'étude du protéome des macrophages (cellules du système immunitaire) humains infectés par la version pathogène *Rickettsia conorii* ou non pathogène *Ricksettia montanensis* a en effet permis de mettre en évidence une différence d'abondance des protéines de l'hôte selon l'espèce qui l'infectait [Curto *et al.* 2019]. Les molécules de surface que la bactérie *Streptococcus pneumoniae* (responsable de la pneumonie et de certaines méningites) utilise pour passer la barrière hémato-encéphalique ont également été étudiées par MS [Jiménez-Munguía *et al.* 2018]. Mieux connaître les caractéristiques de l'infection d'une bactérie peut bien sûr permettre de mettre au point certains traitements.
Dans la section suivante, j'explique plus précisément comment les protéines sont étudiées par MS pour permettre de mieux les connaître, elles et les mécanismes biologiques dans lesquels elles sont impliquées. Malgré le fait que la MS soit massivement utilisée, elle soulève toujours un certain nombres d'obstacles ; ils sont évoqués dans le reste de ce chapitre afin de comprendre pourquoi la MS elle-même doit faire l'objet d'études pour être améliorée.

2.2 Protocole général de l'analyse des protéines par MS

Il existe une grande variété de façons d'appliquer la MS à l'analyse des protéines selon l'objectif visé. Elles peuvent être classées dans deux catégories principales selon ce qui est injecté dans l'appareil : l'approche *top-down* basée sur les protéines, et l'approche *bottom-up* basée sur les peptides.

Pour l'approche *top-down*, les protéines sont analysées entières. Puisqu'on a une vue d'ensemble d'une protéine, cette méthode est particulièrement intéressante pour analyser les différentes formes (variations de séquence ou différents motifs de PTM) d'une même protéine.

L'approche *bottom-up* (ou *shotgun*), elle, consiste à analyser un échantillon de protéines par les peptides issus de ces protéines ; les protéines seront alors reconstruites à partir des peptides identifiés. Le protocole habituel d'une analyse d'un échantillon de protéines par MS en mode bottom-up est résumé Figure 2.2.

Une fois l'échantillon de protéines purifié, elles seront digérées en peptides qui sont injectés dans le spectromètre de masse, le plus souvent après une étape de chromatographie afin d'avoir une première séparation des peptides avant l'analyse par l'appareil. La solubilisation dans la colonne de chromatographie, et donc la séparation des peptides en amont de l'injection dans l'appareil est en effet plus simple que celle des protéines entières (étudiées dans l'approche *top-down*), et l'ionisation des grandes protéines (avec une masse supérieure à 30 kDa) peut présenter des difficultés. Le mode *bottom-up* est donc en général privilégié ; c'est pourquoi je me suis intéressée aux données de type *bottom-up* pendant mon travail de thèse ; l'approche *top-down* ne sera pas discutée dans le reste du manuscrit.

Pour effectuer la digestion des protéines en peptides, une enzyme de type **protéase** est ajoutée à l'échantillon. Une protéase est une enzyme qui clive la chaîne de résidus (c'est-à-dire brise des liaisons peptidiques) de façon spécifique ou non. La protéase la plus utilisée est la trypsine, qui clive la protéine après ses lysines (K) et ses arginines (R). Les peptides obtenus sont appelés peptides tryptiques (voir Figure 2.3).

L'utilisation courante de la trypsine s'explique par plusieurs raisons, la première étant la bonne spécificité de la protéase aux résidus K et R ; la seconde est que, étant donné la distribution générale de ces résidus dans les protéines, couper la protéine à leur niveau permet d'obtenir des peptides dont la taille est compatible avec une analyse par MS ;

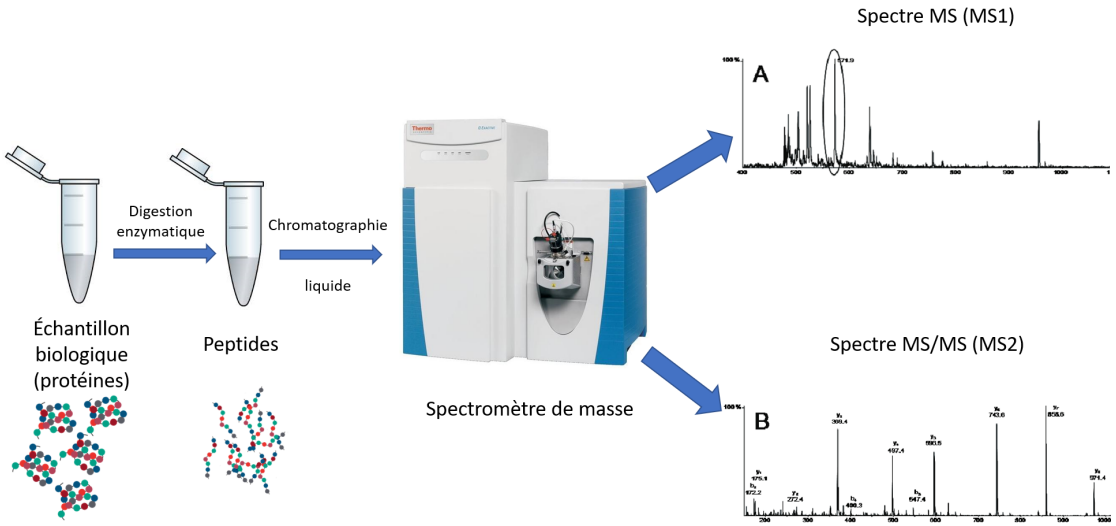


FIGURE 2.2 – **Analyse des protéines par spectrométrie de masse en mode bottom-up.** Un échantillon de protéines, plus ou moins purifié et donc plus ou moins complexe, est mis en contact avec une protéase qui va cliver les protéines. De cette digestion sont obtenus des peptides qui sont injectés dans l'appareil, habituellement après une étape de séparation par chromatographie. Le spectromètre peut renvoyer des spectres MS1 (spectre A en haut à droite) ou MS2 (spectre B en bas à droite) dont les caractéristiques sont détaillées dans la Section 2.3. Schéma de l'auteure à partir des images de `flyclipart.com` (tupes Eppendorf), `peptidessciences.com` (peptides et protéines), `thermofisher.com` (spectromètre de masse), [JOHNSON *et al.* 2011] (spectres de masse).

FIGURE 2.3 – **Exemple de digestion d'une protéine par la trypsine.** La trypsine coupe la chaîne peptidique après les résidus K et R. Dans cette protéine du protéome humain, la protéase va couper 4 fois, ce qui permet l'obtention de 5 peptides.

enfin, les peptides tryptiques ont K ou R à leur extrémité C-ter (sauf le peptide C-ter si la protéine ne se termine pas par K ou R), et ces résidus permettent une bonne ionisation. L'appareil permet d'obtenir, à partir des peptides qui y pénètrent, des **spectres de masse** qui peuvent être stockés avant leur analyse. Un tel protocole peut produire des dizaines de milliers de spectres à analyser. Les caractéristiques des spectres de masse sont détaillées dans la section suivante (Section 2.3).

2.3 Les ions et les spectres

Chaque peptide arrivant dans le spectromètre de masse (Figure 2.2) voit son ratio m/z calculé par l'analyseur ainsi qu'une intensité de signal enregistrée par le détecteur. L'intensité correspondant à un m/z donné dépend en partie du nombre de peptides ayant ce m/z, mais ne peut pas y être directement liée, car les peptides n'ont pas tous la même capacité à provoquer un signal. Grâce au système de traitement informatique relié à l'appareil, les données brutes produites par ce peptide peuvent-être représentées sous la forme d'un pic, avec un m/z et une intensité correspondante. Un échantillon de peptides dans le spectromètre de masse à un instant t correspond donc à un ensemble de tels pics. Cet ensemble sera représenté sur un spectre de masse, appelé spectre MS ou spectre MS1 (en haut à droite de la Figure 2.2), avec en abscisse le ratio m/z des ions entrés dans l'appareil, et en ordonnée une intensité correspondante.
Les spectres MS1 représentent une information déjà utile pour identifier les protéines d'un échantillon ; en effet l'ensemble des masses correspondant aux peptides issus de la digestion d'une protéine par une protéase donnée est une empreinte spécifique à la protéine, qui peut être comparée aux spectres MS1 issus de peptides tryptiques simulés à partir de protéines connues. Cette méthode est appelée l'empreinte peptidique, ou *Peptide Mass Fingerprinting* (PMF, [Henzel *et al.* 1993]). Un exemple de spectre MS1 analysé par PMF est représenté Figure 2.4.
Cependant, pour avoir une empreinte spécifique, le PMF doit être appliqué à des protéines qui sont hautement purifiées et ne peut pas être appliqué à des mélanges complexes. De plus, il ne peut pas être utilisé pour déterminer la séquence en résidus des peptides inconnus et donc des protéines inconnues, car une masse peut correspondre à de nombreux peptides, à cause du grand nombre de permutations possibles pour un même ensemble de résidus, mais aussi de séquences en résidus différentes qui ont des masses identiques. Pour déterminer la séquence des peptides, on utilise la MS en tandem, appelée aussi MS/MS

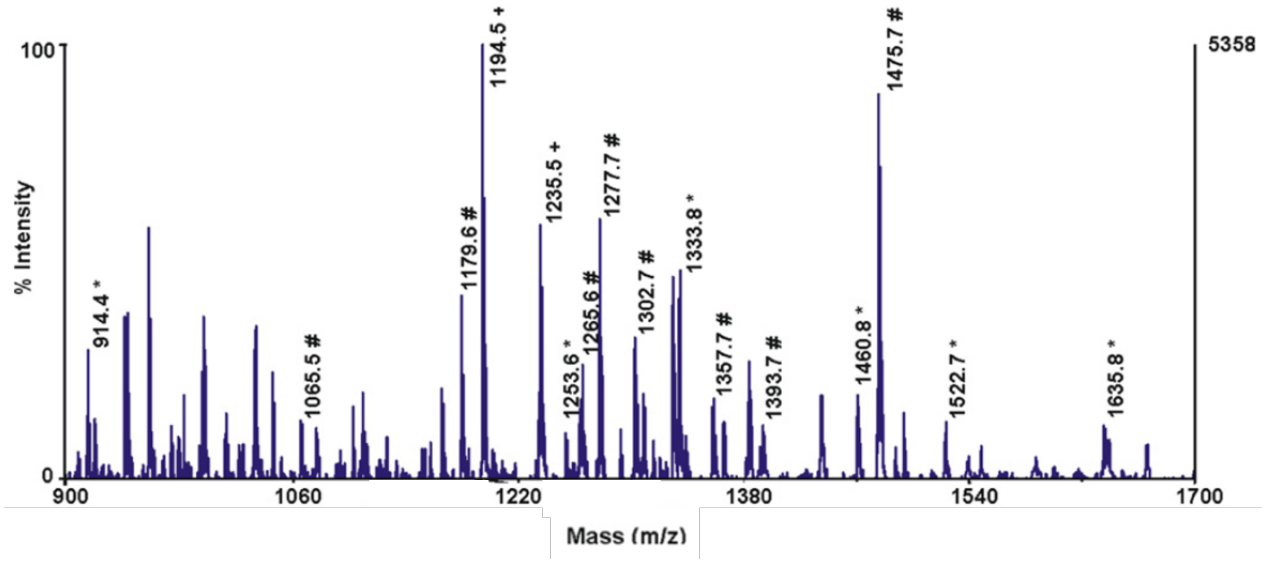


FIGURE 2.4 – **Exemple de spectre MS1 obtenu expérimentalement.** Spectre issu de [THIEDE *et al.* 2005]. Le spectre MS1 a été obtenu à partir d'un extrait cytosolique d'*Helicobacter pylori*. L'analyse par empreinte peptidique (PMF) permet de détecter la présence de contaminants (#), de peptides issus de la protéine régulatrice HP0703 (*) et de peptides issus de la protéine HP0163 (+).

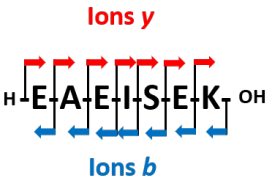
ou MS2 (en bas à droite de la Figure 2.2).

Pour obtenir une information de séquence par MS2, un peptide sera fragmenté (voir Figure 2.5a). Il est en effet possible de sélectionner une fenêtre de m/z du spectre MS1 (par exemple selon un seuil d'intensité), et les peptides arrivant dans l'appareil qui correspondent à cette fenêtre sont dirigés vers une partie de l'appareil nommée cellule de fragmentation.

La fragmentation brise certaines liaisons peptidiques (selon la séquence du peptide, la quantité d'énergie de la cellule de fragmentation, etc.) tout au long du peptide. Une fragmentation dans un peptide permet d'obtenir deux molécules qui sont les parties N-ter et C-ter sous forme d'ions. Par exemple, si la fragmentation brise un peptide de taille six entre le deuxième et le troisième résidu, deux ions complémentaires seront produits, l'ion N-ter comportant les deux premiers résidus, et l'ion C-ter comportant les quatre derniers. Les ions sont nommés selon une nomenclature qui dépend du côté du peptide qu'ils représentent (N-ter/C-ter), du type de fragmentation et de l'emplacement de celle-ci dans la séquence du peptide. La fragmentation produit principalement deux ions complémentaires de type b (partie N-ter) et de type y (partie C-ter) (voir Figure 2.5a).

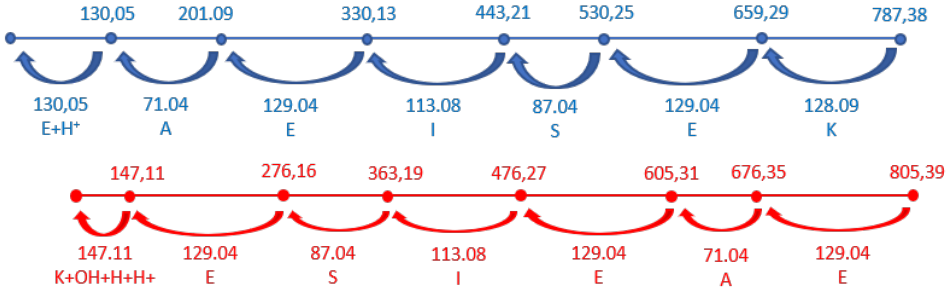
L'ion b est caractérisé par une ionisation positive en C-ter ; ainsi la masse de l'ion b numéro

Ion	Séquence	Masse (Da)
b_1	E-H$^+$	130,0499
b_2	EA-H$^+$	201,087
b_3	EAE-H$^+$	330,1296
b_4	EAEI-H$^+$	443,2136
b_5	EAEIS-H$^+$	530,2457
b_6	EAEISE-H$^+$	659,2883
b_7	EAEISEK-H$^+$	787,3832

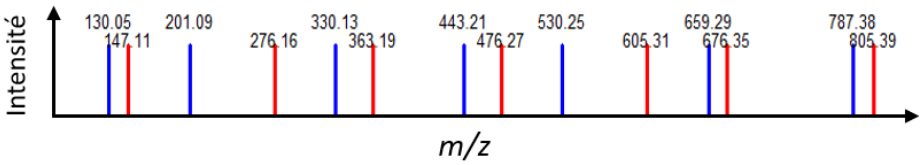


Ion	Séquence	Masse (Da)
y_1	H_3^+-K-OH	147,1128
y_2	H_3^+-KE-OH	276,1554
y_3	H_3^+-KES-OH	363,1874
y_4	H_3^+-KESI-OH	476,2715
y_5	H_3^+-KESIE-OH	605,3141
y_6	H_3^+-KESIEA-OH	676,3512
y_7	H_3^+-KESIEAE-OH	805,3938

(a) **Représentation des ions b et y produits lors de la fragmentation du peptide EAEISEK.** Lors de la fragmentation du peptide EAEISEK, deux principaux types d'ions sont produits : les ions de la série b (tableau bleu à gauche) et les ions de la série y (tableau rouge à droite). Un ion b contient la partie N-ter du peptide et est numéroté selon sa séquence, de gauche à droite. Un ion y contient la partie C-ter du peptide et est numéroté en sens inverse, de droite à gauche. Les masses sont données en Da.



(b) **Différence des masses des ions produits par la fragmentation du peptide EAEISEK et résidus correspondants.**



(c) **Spectre MS2 correspondant à la fragmentation du peptide EAEISEK.** Sur ce spectre MS2 théorique et idéal (en bas), la masse de chaque ion du peptide analysé ainsi qu'une intensité, arbitraire et égale pour tous les ions, sont représentées. La différence entre les masses de deux ions b (en bleu) consécutifs, et entre les masses de deux ions y (en rouge) consécutifs correspondent à la masse du résidu ajouté entre deux ions consécutifs (par exemple un A (alanine) est ajouté entre b_1 et b_2, et donc la différence de masse entre b_1 et b_2 est égale à la masse de A) ; ces différences de masses, représentées en (b), permettent d'identifier les résidus du peptide et leur ordre.

FIGURE 2.5 – **Ions issus de la fragmentation d'un peptide et spectre MS2 théorique correspondant.**

i (= issu de la $i$ème fragmentation), noté b_i, se calcule selon la formule 2.1.

$$masse(b_i) = \left(\sum_{k=1}^{i} masse(résidu_k)\right) + masse(H^+) \tag{2.1}$$

L'ion y est ionisé en N-ter ; pour un peptide de longueur T, la masse de l'ion y numéro i, noté y_i, se calcule donc selon la formule 2.2.

$$masse(y_i) = \left(\sum_{k=T-i+1}^{T} masse(résidu_k)\right) + masse(H) + masse(H^+) + masse(OH) \tag{2.2}$$

Les ions y sont numérotés dans le sens inverse de la séquence ; l'ion y_1 correspond donc à celui qui contient le dernier résidu du peptide.
Étant donné qu'un certain nombre de représentants du même peptide subissent le processus de fragmentation à des endroits différents du peptide, on obtient ainsi statistiquement des ions issus de la fragmentation à chaque liaison peptidique. Les ions b et y peuvent être à leur tour analysés ; leur m/z est mesuré et représenté sur un spectre MS/MS, noté aussi MS2. On obtient ainsi idéalement sur ce spectre l'ensemble des ions des deux séries produits par toutes les fragmentations du peptide ; un spectre MS2 théorique est représenté Figure 2.5c. À chaque spectre MS2 est associée la masse du peptide fragmenté, appelée **masse parente**.
L'intérêt des spectres MS2 réside dans le fait que les différences de masses entre les ions de masse croissante de la même série correspondent aux masses des résidus du peptide dans l'ordre de la séquence (de Nter en Cter pour les ions b et de Cter en Nter pour les ions y), comme représenté Figure 2.5b. Traiter ces spectres de la bonne façon peut ainsi donner une information de séquence, qui permettra d'identifier précisément le peptide. L'interprétation d'un spectre MS2 de qualité est représentée Figure 2.6. Différentes méthodes existent pour exploiter l'information de séquence présente dans un spectre MS2, et sont décrites dans la section suivante (Section 2.4).

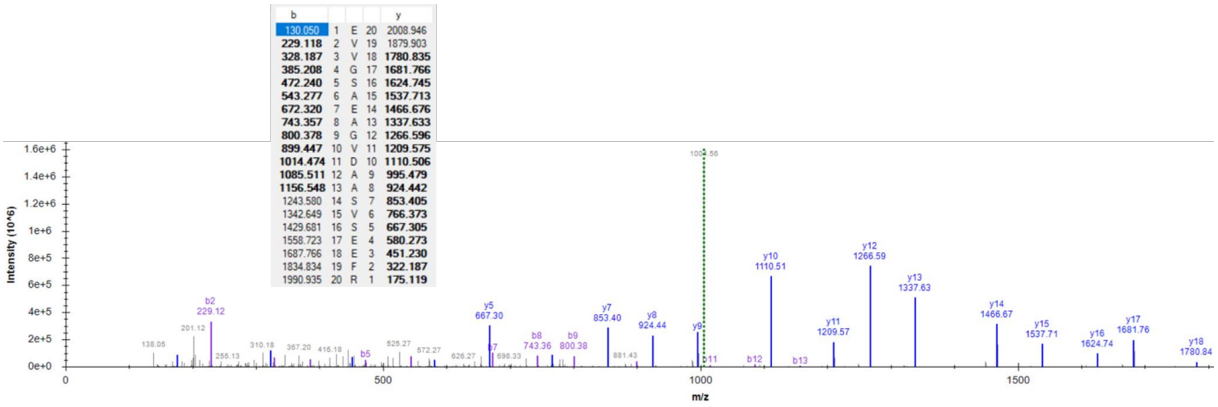


FIGURE 2.6 – **Exemple de spectre MS2 annoté.** Un peptide issu de l'albumine est fragmenté, et cette fragmentation produit des ions dont les masses (abscisses) et les intensités (ordonnées) peuvent être représentées par des pics sur un spectre MS2, ici visualisé sur le logiciel `SeeMS`. Sur ce spectre, les ions (b_2, b_5, y_5, etc.) et les masses correspondantes sont indiqués. Les ions b sont en violet, et les ions y en bleu. Les différences des masses des pics correspondent à des masses de résidus. Au dessus du spectre est montré un tableau avec les masses des ions b et y, où les ions permettant de déterminer la séquence des résidus sont en gras. Source : plateforme BIBS, unité BIA, INRAE Nantes.

2.4 Les premières méthodes d'identification des peptides à partir des spectres MS2

Pour identifier la séquence qui correspond à un spectre, une approche qui a été utilisée assez tôt (déjà dans les années 1980 avec [SAKURAI *et al.* 1984]) est la suivante : on produit toutes les permutations de résidus qui correspondent à la masse parente du spectre MS2. Pour chaque permutation, on calcule un score qui repose sur la correspondance entre les fragments attendus par la séquence théorique et ceux du spectre MS2 à identifier, et on suppose que celle qui correspond le mieux au spectre correspond au peptide à identifier. Les séquences - partielles ou complètes - obtenues par ces méthodes peuvent être ensuite recherchées dans une base de données de séquences de protéines. L'émergence de grandes bases de données décrites dans le Chapitre 1 permet en effet ce type d'analyse. En revanche, le nombre de possibilités pour une masse donnée fait que ce type d'algorithme est trop lent pour être utilisé à grande échelle. Connaître la composition en résidus du peptide correspondant au spectre à analyser peut réduire grandement le nombre de possibilités de séquences à considérer ; cependant cette information n'est pas disponible lors de l'analyse d'un mélange complexe de peptides.

Une méthode qui semble intuitive pour déterminer la séquence correspondant à un spectre MS2 est de calculer les différences de masse dans un spectre et de les annoter si celles-ci correspondent à des masses de résidus connues. Cette méthode est appelée le séquençage de peptides *de novo*, et n'utilise que l'information présente dans le spectre ainsi que les connaissances des masses des résidus.
L'étude de [Hunt *et al.* 1992] a consisté à séquencer par cette méthode des peptides du CMH (qui est un système de reconnaissance du soi essentiel au système immunitaire des vertébrés) chez l'être humain. Elle est l'un des exemples qui montrent que, grâce aux avancées technologiques comme le développement de l'électrospray (voir Section 2.1.3), l'analyse des spectres a permis de résoudre rapidement certaines questions dans des champs aussi importants que l'immunologie. `Sherenga`, développé par [Dancík *et al.* 1999], est un exemple d'algorithme d'interprétation *de novo* de spectre MS2.
Cependant, afin d'avoir une interprétation fine d'un spectre par le *de novo*, une proportion importante des masses doit être présente afin de pouvoir identifier les résidus par différences successives. Les méthodes *de novo* sont donc fortement dépendantes de la qualité des spectres et des modifications que portent les peptides correspondants (voir Section 2.6.1) ; elles sont complexes à automatiser et ainsi plutôt réservées à l'étude d'organismes pour lesquels des bases de données n'existent pas. En effet, les inconvénients de l'interprétation *de novo* rendent plus intéressantes les méthodes présentées dans la section suivante.

2.5 Identification des spectres MS2 par comparaison à une base de données de spectres

2.5.1 Utilisation de spectres théoriques

Principe d'un outil de comparaison à une base de données de spectres théoriques

La plupart des méthodes d'analyse de spectres MS2 reposent plutôt sur leur comparaison à des spectres théoriques correspondant à la digestion et la fragmentation *in silico* d'une base de données de protéines censée représenter l'échantillon étudié (voir Figure 2.7).
Chaque spectre MS2 expérimental est comparé aux spectres théoriques générés à par-

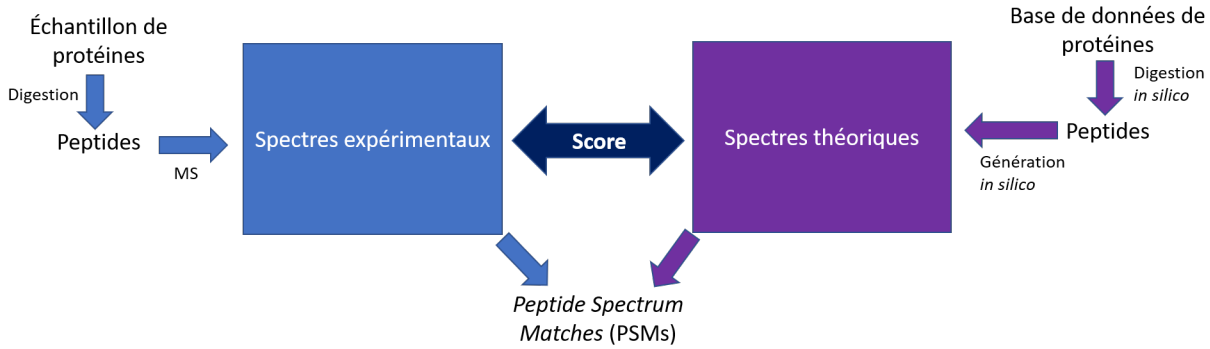


FIGURE 2.7 – **Identification des spectres de masse (MS2) par comparaison avec une base de données de spectres théoriques.** Cette identification comporte plusieurs étapes. Une fois que les spectres MS2 expérimentaux sont obtenus, on les confronte à des spectres théoriques issus de la digestion d'une base de données de séquences de protéines. Pour cela, on utilise un score de comparaison. Pour chaque spectre expérimental, on affecte un ou plusieurs spectres théoriques qui sont sélectionnés par le score choisi, par exemple selon un seuil. Ces ensembles spectre-peptide sont nommés PSM.

tir des peptides de la base de données. Cette comparaison est faite à l'aide d'un score de similarité qui va mesurer la ressemblance entre le spectre expérimental et le spectre théorique. Cette méthode a l'avantage de s'émanciper de la nécessité d'extraire une ou plusieurs séquences de résidus, mêmes courtes, du spectre, ce qui rend la comparaison possible même pour des spectres qui ne peuvent pas être interprétés par des approches *de novo*.

Après cette comparaison, à chaque spectre MS2 expérimental sera assigné un ou plusieurs spectre(s) théorique(s) - et donc peptide(s) - qui lui ressemblent d'après le score, et chaque couple spectre-peptide est communément nommé Peptide-Spectrum Match, ou **PSM**. Les PSM renvoyés par un outil d'identification permettent d'obtenir, grâce aux peptides candidats identifiés, des informations sur l'identité des peptides qui correspondent aux spectres analysés.

Ces méthodes de comparaison ont vu le jour dans les années 1990 et utilisaient déjà des stratégies variées pour comparer un spectre contre une base de données. Eng *et al* [ENG, MCCORMACK et J. R. YATES 1994] ont posé les bases de la recherche d'un spectre MS2 dans une base de données de spectres théoriques issus d'une base de données de protéines. Ces travaux ont permis le développement de `Sequest`, outil fondateur des méthodes d'identification de spectres MS2 par ce moyen [BRODBELT et RUSSELL 2015].

Les différentes étapes de `Sequest` sont très représentatives des différentes questions qui

doivent être résolues lors de l'identification des spectres MS2 par comparaison à des spectres théoriques issus d'une base de données de protéines. Elles constituent souvent le squelette des outils dédiés à ce type d'analyse et permettent de les caractériser.

Quels peptides et quels spectres ? À partir des séquences de protéines, comment produire les peptides et les spectres théoriques qui en résultent ?
Il faut d'abord choisir une base de protéines correspondant à l'échantillon à étudier. La manière dont les peptides théoriques sont créés dépend également de l'échantillon. En effet, si les protéines ont été digérées par la trypsine pour produire les peptides, la base de données sera digérée de la même manière (en coupant informatiquement les séquences après K et R). La trypsine est une protéase qui n'est pas fiable à 100%, il est donc courant de prendre en compte l'absence de clivage à certains sites et les peptides qui en résultent, appelés *missed cleavages*. Au contraire, au cours de leur manipulation, les protéines peuvent subir des clivages après des résidus autres que K et R ; les peptides qui en résultent sont qualifiés de "semi-tryptiques".
On obtient ainsi un certain nombre de séquences candidates qui peuvent être transformées en spectres théoriques, habituellement en calculant les masses des ions b et y de la séquence, comme expliqué Section 2.3. Ces spectres sont ensuite comparés au spectre expérimental étudié.

Quel filtre sur les spectres théoriques ? Étant donné le nombre de comparaisons à réaliser, celles-ci peuvent être très chronophages, voire irréalisables sur tous les spectres théoriques de la base de données. Comment filtrer les candidats les plus pertinents avant de calculer le score final ?
Un filtre peut être réalisé sur la masse parente des peptides, et donc des spectres théoriques, de la base de données. Si les masses parentes des spectres à comparer sont considérées comme égales, le score sera calculé. Sinon, on considérera qu'il s'agit de deux peptides différents, et la comparaison ne sera pas effectuée. Lors de cette étape, on considère que deux masses parentes sont égales à une certaine **tolérance** près ; les deux masses seront considérées égales si leur différence est inférieure à la tolérance. Celle-ci dépend de la précision de l'appareil, mais aussi du mode de recherche (voir la Section 2.6.3 sur les méthodes OMS). Elle peut s'exprimer en pourcentage, en ppm, ou simplement en Da. La tolérance lors de cette étape est donc une notion très importante, car elle décide en grande partie des candidats sélectionnés pour l'étape suivante.

Quelle comparaison ? Après un filtre éventuel, de nombreux spectres théoriques candidats peuvent être sélectionnés. Il faut comparer plus finement le spectre à identifier aux candidats afin de sélectionner le plus approprié. Comment comparer le spectre expérimental et le spectre théorique ? Cela est habituellement fait grâce à un score calculé entre deux spectres, qui va quantifier leur ressemblance. Il faudra choisir quel score de comparaison utiliser. Il existe de nombreux types de scores, mais la plupart reposent sur le nombre de pics dont les masses sont partagées entre deux spectres, appelé le **SPC** (Shared Peaks Count, voir par exemple la Figure 2.8, Section 2.6.1, page 50, qui montre deux spectres avec un SPC de 7). Étant donné le lien direct entre les masses des ions et la séquence du peptide dont ils sont issus, comparer les spectres à l'aide du SPC est pertinent pour capturer une similarité de séquence entre les peptides correspondants.

Exemples d'outils

Les outils d'identification de spectres contre une base de données de peptides gèrent ces étapes de différentes façons. Des exemples sont montrés Table 2.1.
`Sequest` a fait ses premiers tests avec la banque GenPept du NCBI (National Center of Biotechnology Information) contenant environ 75 000 protéines et 30 millions de résidus issus de la traduction de la base de nucléotides GenBank. Afin de ne pas être spécifique à une protéase donnée, `Sequest` choisit de créer les spectres théoriques à la volée pour chaque spectre expérimental ; il additionne les masses des résidus des protéines de la base de données jusqu'à ce que la somme soit égale à celle de la masse parente du spectre expérimental à identifier. Lors de cette étape, on a une tolérance de 5%, c'est-à-dire que l'on accepte des peptides qui ont jusqu'à ±5% de la masse parente du spectre.
Pour sélectionner les candidats, `Sequest` applique également un autre filtre : il se base sur le SPC entre les spectres pour calculer un score initial qui permet de sélectionner 500 peptides candidats. Ce score initial prend également en compte les intensités de ces masses. Plus tard, un pré-traitement des spectres (par factorisation du calcul du score) a été développé, permettant un calcul du score final (voir paragraphe suivant) bien plus rapide et donc un score initial rendu superflu [Eng, Fischer *et al.* 2008]. Ces améliorations ont été implémentées dans le logiciel `Comet` [Eng, Jahan et Hoopmann 2013].
Après ce score initial, `Sequest` évalue la similarité entre le spectre expérimental et le spectre théorique avec un score final de "cross-correlation". Les spectres sont considérés comme des vecteurs de valeurs dont le produit scalaire (multiplication des valeurs des

TABLE 2.1 – **Résumé du fonctionnement de 3 outils d'identifications de spectres MS2 selon 3 critères.**

	Peptides	Filtre	Score
`Sequest/Comet`	Tous selon masse parente	Score initial (SPC)	Score final ("cross-correlation")
Tags (Mann & Wilm)	Tryptiques	*Tags*	*Tags* + vérification (SPC)
`Mascot`	Tryptiques	Masse parente	Score probabiliste de `Mascot`

intensités des deux spectres) est calculé. Cette opération est répétée un certain nombre de fois avec un certain nombre de décalages de m/z entre les deux spectres, et la moyenne des produits est soustraite au produit sans décalage, dans l'objectif d'éliminer le bruit. Ce score est encore celui qui est utilisé dans la version actuelle de `Sequest`.

Mann et Wilm [MANN et WILM 1994] ont mis au point la recherche par tag, c'est-à-dire que l'on va extraire des petites séquences de 2 ou 3 résidus d'une zone du spectre qui en révèle, et les rechercher dans les peptides de la base de données. Cette méthode, à l'interface entre le *de novo* et la recherche d'une séquence dans une base de données, permet d'exploiter certains "îlots" de qualité dans le spectre afin de rechercher la séquence correspondante dans une base de données, avec des critères plus ou moins stringents par rapport aux peptides de la base de données, par exemple des critères de différence de masse. Cette recherche est donc un filtre qui est appliqué avant une étude plus approfondie des peptides candidats.

Pour illustrer la diversité de score, on peut citer `Mascot` [PERKINS *et al.* 1999], qui est un outil dédié à l'analyse de différents types de données de spectrométrie de masse. Il est utilisable en ligne (`https://www.matrixscience.com/search_form_select.html`) et est capable d'effectuer plusieurs types de recherche, en mode PMF (voir Section 2.3) ou en mode MS/MS. Il intègre également un module de recherche par *tags*. L'une des caractéristiques principales de `Mascot` est son score pour chaque match entre un spectre et une séquence de la base de données après un filtre sur la masse parente ; ce score est basé sur la similarité entre fragments, mais le score total repose sur la probabilité que le match soit dû au hasard, et va donc dépendre de la taille de la base de données.

2.5.2 Utilisation de données MS réelles : les bibliothèques spectrales

Le développement de la spectrométrie de masse a permis la génération d'une grande quantité de spectres de masse de bonne qualité et souvent identifiés qui peuvent être déposés dans des bases de données en ligne. Parmi les bases de données les plus utilisées, on trouve PRIDE [Vizcaíno *et al.* 2016], créée par l'EBI (*European Bioinformatics Institute*) dont l'objectif premier est l'accès facile et standardisé aux jeux de données MS décrits dans les publications scientifiques. La base de données contient notamment des outils de validation lors du dépôt, et de visualisation relatifs aux données déposées. PeptideAtlas [Desiere *et al.* 2004] a pour objectif l'annotation du génome humain et d'autres espèces eucaryotes par la protéomique. Les résultats déposés sont traités par le `Trans-Proteomic Pipeline` (voir Section 2.5.4).
Certains outils choisissent d'interpréter des spectres MS2 en tirant profit de ces données réelles. Ils fonctionnent en deux étapes : 1) les spectres des bases de données (comme par exemple PRIDE) sont traités afin de produire des spectres consensus (c'est-à-dire des spectres moyens) qui forment une base de données appelée **bibliothèque spectrale** ; 2) le spectre à identifier est comparé à ceux de la bibliothèque spectrale par des outils de comparaison. Voir [Lam 2011] pour un état de l'art détaillé sur les bibliothèques spectrales.
Il en résulte deux principaux avantages. Le premier est que les caractéristiques expérimentales d'un spectre, c'est-à-dire l'intensité des pics, ou encore les motifs de fragmentation qui sont caractéristiques d'un peptide, sont présentes dans le spectre de la bibliothèque, ce qui permet de les comparer avec une finesse particulière à des spectres expérimentaux qui possèdent eux aussi ces caractéristiques s'ils représentent le même peptide. Le second avantage est que l'espace de recherche est réduit à des peptides qui ont été observés expérimentalement, ce qui est un moyen pertinent de focaliser une recherche et de réduire le temps de calcul de façon stratégique, et donc de permettre à des groupes de recherche d'identifier des spectres sans forcément maintenir une infrastructure informatique trop coûteuse. Les bibliothèques spectrales sont un moyen d'utiliser les données de qualité déjà produites et sont donc particulièrement adaptées à l'étude d'un jeu de protéines prédéfinies, déjà connues.
Parmi les outils d'identification se basant sur les bibliothèques spectrales, on peut citer `SpectraST` [Lam *et al.* 2007]. Afin de classer les spectres candidats de la bibliothèque

pour un spectre à identifier, `SpectraST` calcule la similarité entre deux spectres par la somme du produit des intensités correspondant aux masses identiques. `X!Hunter` [Craig, Cortens *et al.* 2006], développé à partir de `X!Tandem` (présenté Section 2.6.2), sélectionne les spectres avec une masse proche de la masse parente du spectre à identifier et évalue la similarité grâce à un score qui prend en compte les intensités des pics ainsi que le nombre de pics au total entre les deux spectres à comparer. `Libquest` [J. R. Yates *et al.* 1998] utilise, comme `Sequest`, un score de corrélation pour comparer un spectre expérimental et un spectre d'une bibliothèque spectrale qui a la même masse parente. `BiblioSpec` [Frewen *et al.* 2006] fait également une sélection sur la masse avant de calculer le produit scalaire des spectres.
Les bibliothèques spectrales ont plusieurs inconvénients, comme la diversité des formats de données spectrales. De plus, les caractéristiques expérimentales des données spectrales - motifs de fragmentation et intensités - dépendent de l'instrument qui les a générées, ce qui signifie que, pour être efficaces, des bibliothèques spectrales doivent être disponibles pour les conditions expérimentales dans lesquelles le spectre à identifier a été obtenu. Enfin, le peptide, ou un peptide proche, doit être représenté dans la bibliothèque afin d'identifier un spectre, ce qui n'est pas forcément le cas du fait de la quantité de données disponibles, souvent moins importante que pour une base de données de spectres théoriques.
Les bibliothèques spectrales ont donc des avantages par rapport à une recherche par comparaison à une base de données de spectres théoriques, notamment le temps de calcul réduit et une meilleure identification si les données de la bibliothèques représentent correctement les échantillons étudiés. Néanmoins, pour la plupart des études, les inconvénients des bibliothèques spectrales (par exemple l'absence d'un peptide donné) sont moindres dans une base de données de spectres théoriques. Ainsi cette seconde option est en général privilégiée. C'est pour cela que mon travail de thèse s'est focalisé sur l'identification de spectres MS/MS par les méthodes se basant sur des spectres théoriques, présentées dans la Section 2.3.

2.5.3 La validation des identifications

Les méthodes de comparaison des peptides à des bases de données peuvent provoquer plusieurs types d'erreurs. Le premier type d'erreur est de ne pas retrouver dans la base de données le candidat idéal s'il y est présent. Cette identification manquée est un faux négatif (erreur de type II en statistiques), type d'erreur qui peut être réduit en étudiant la similarité entre deux spectres afin d'établir un score pertinent qui permet de trouver le

bon candidat quand il est présent.
Il arrive également qu'un PSM donné corresponde à une erreur, c'est-à-dire que le peptide assigné à un spectre ne lui corresponde pas. On a dans ce cas un faux positif (erreur de type I en statistiques). Une fois les PSM produits, une question cruciale est donc la pertinence d'un PSM donné. À quel point le match représente-t-il un "bon" résultat, c'est-à-dire un résultat souhaité ? Les PSM peuvent être vérifiés manuellement dans une certaine mesure, mais une telle inspection n'est bien évidemment pas pratiquable sur les centaines de milliers de PSM qui peuvent être renvoyés après une recherche contre une base de données. Une étape supplémentaire et automatique est donc nécessaire afin d'estimer et réduire ce taux de faux positifs et fiabiliser les résultats. L'une des plus utilisées est appelée **FDR** (*False Discovery Rate*). Il s'agit d'une valeur statistique calculée afin d'estimer le taux de faux positifs dans un jeu de résultats donné. La stratégie de FDR la plus utilisée afin d'écarter les faux positifs est la stratégie target/decoy [Elias et Gygi 2007 ; Elias et Gygi 2010]. Son principe est le suivant : lors d'une recherche pour identifier un jeu de spectres grâce à une base de données de protéines, appelée base **target**, on va créer à partir de cette dernière une base de données afin de représenter des peptides supposés non présents dans l'échantillon, et donc représenter de mauvaises identifications. Cette base, nommée base **decoy**, est produite en inversant les séquences des protéines target ou bien en mélangeant les résidus des peptides ou des protéines de façon aléatoire. On obtient ainsi une base de peptides "faux", mais pertinents au regard de la base target. Les peptides decoy issus de la digestion de telles protéines seront mêlés aux peptides target avant l'étape de recherche et joueront, comme leur nom l'indique, le rôle de "leurre". En effet, un PSM issu de cette recherche impliquant un peptide decoy sera considéré comme un faux positif puisqu'il serait dû à une similarité liée au hasard. L'hypothèse de cette démarche est que le nombre de faux positifs est le même dans les bases target et decoy au-dessus d'un certain seuil. Ce nombre est très élevé quand les spectres des PSM sont peu similaires, et de plus en plus faible quand la similarité des spectres augmente. Pour cette raison, les PSM sont classés en fonction du score de similarité supposé classer les PSM des plus fiables au moins fiables ; puis pour un seuil donné, le FDR peut être calculé en comparant le nombre de faux positifs (identifications dans la base decoy) au nombre total d'identifications. Le seuil de score (par exemple, le SPC) sera donc augmenté jusqu'à ne plus avoir qu'un certain FDR accepté fixé par l'utilisateur (moins de 1% habituellement).
La stratégie target/decoy n'est pas la seule manière de valider les PSM. En partant de l'observation que la validation des identifications était faite différemment au sein de la commu-

nauté scientifique, `PeptideProphet` [Keller 2002] a été développé dans l'objectif d'harmoniser cette étape afin de rendre la comparaison des résultats plus facile. `PeptideProphet` utilise du machine learning sur les scores de recherche et d'autres variables (comme le temps de rétention d'un peptide par chromatographie) afin de fournir une probabilité que l'identification soit correcte, et un moyen de filtrer les données selon ce critère. Cette validation ne nécessite pas l'emploi d'une base decoy, mais une telle base peut être utilisée pour améliorer la validation.
`iProphet` [D. Shteynberg *et al.* 2011] affine les résultats de PeptideProphet en combinant les résultats de plusieurs outils d'identification, et en prenant en compte le cas où un même peptide est identifié plusieurs fois. Ces outils peuvent être gérés simplement avec une interface web. Voir l'état de l'art [Nesvizhskii 2010] pour plus de détails sur les moyens variés de valider les PSM.

2.5.4 Les pipelines d'identification des peptides

Nous avons pu voir que l'identification des peptides par spectrométrie de masse est un travail qui peut nécessiter plusieurs étapes complexes, à réaliser soigneusement si on veut obtenir des PSM informatifs. Il faut pour cela utiliser une base de données représentative de l'échantillon étudié, des scores pertinents pour affecter un peptide à chaque spectre et donc produire les PSM. Il faut ensuite valider et filtrer les résultats afin de minimiser les faux positifs.
Même si un outil implémente un algorithme efficace, il peut être compliqué à utiliser à cause de son interface, ou encore à cause de son coût en termes de calculs. Ce problème se trouve multiplié lors d'une analyse qui demande d'utiliser plusieurs de ces outils. De plus, la manière dont les équipes de recherche utilisent les outils peut nuire à l'obtention de résultats comparables.
Afin d'utiliser plus aisément et de façon plus standardisée les outils qui correspondent à chacune de ces étapes, des pipelines ont été développés. Le `Trans-Proteomic Pipeline` [Deutsch *et al.* 2015] intègre une trentaine d'outils correspondant à différentes étapes d'une analyse de protéomique, depuis la conversion des fichiers bruts issus d'un spectromètre jusqu'à la validation statistique au niveau de la protéine en passant par la visualisation des données MS. Il intègre par exemple `X!Tandem` (présenté en Section 2.6.2) pour la recherche contre une base de données de séquences, `SpectraST` pour la recherche contre une bibliothèque spectrale, `PeptideProphet` pour la validation des résultats, ou encore `iProphet` pour affiner les probabilités de `PeptideProphet`.

`MaxQuant` [Tyanova, Temu et Cox 2016] est l'un des pipelines d'analyse de données MS les plus répandus, car il réalise automatiquement de nombreuses étapes de l'analyse des données MS, depuis le traitement des fichiers bruts sortis des appareils jusqu'à la validation et la recherche des spectres MS2 contre une base de données avec un moteur de recherche intégré.
En partant de l'observation que la gestion informatique de ces boîtes à outils est complexe (installation, gestion des bibliothèques, etc.), `Philosopher` [Veiga Leprevost *et al.* 2020] a été créé afin de simplifier la gestion de l'analyse par l'utilisation de conteneurs ; le logiciel peut donc être déployé à la fois sur un ordinateur de bureau et un serveur de calcul. `Philosopher` a des outils en commun avec le `Trans-Proteomic Pipeline`, notamment les outils de validation cités précédemment.

2.6 La problématique des modifications et les méthodes OMS

2.6.1 Le défi des peptides modifiés

L'importance des modifications portées par les résidus des protéines (PTM) a été exposée dans le Chapitre 1. Étant donné leur importance biologique, la MS doit être capable de les identifier. Cependant, en plus des modifications portées naturellement par les protéines, leur analyse par MS peut également modifier les protéines. En effet, les cystéines sont carbamidométhylées (+57 Da) à cause de l'iodoacétamide présent dans la solution de manipulation des peptides. Les méthionines peuvent également être oxydées, auquel cas leur masse est augmentée de 16 Da. Dans le reste du manuscrit, le terme PTM désignera également l'ensemble des artefacts expérimentaux.
Identifier et localiser les PTM sur les peptides sont un défi, car elles complexifient l'analyse pour différentes raisons.
Tout d'abord, d'un point de vue strictement technique, les PTM peuvent influencer l'ionisation des peptides ainsi que le motif de fragmentation. Ensuite, historiquement, les méthodes de comparaison évoquées auparavant vont souvent limiter la différence de masse, notée $\boldsymbol{\Delta m}$, entre la masse parente d'un spectre MS2 et celle du peptide du spectre théorique auquel il est comparé ; la tolérance sera en effet mise à 0, ou bien à des valeurs de masse qui correspondent à des valeurs attendues par l'utilisateur. Comme cela a été expliqué, cette limite a notamment pour objectif de réduire le nombre de peptides can-

didats pour un spectre, et donc le temps de comparaison et de production des PSM, en partant du principe que l'on va identifier le peptide qui correspond au spectre, et donc avec la même masse. Or, il est possible que le spectre expérimental étudié corresponde à un peptide modifié, que ce soit par une PTM ou bien une variation de séquence (insertion, délétion, substitution de résidus) qui le différencie du peptide qui lui est le plus proche dans la base de données utilisée. Cette différence aura un impact sur la masse parente du spectre, et donc sur le Δm lors de la comparaison. Les méthodes qui limitent le Δm écartent donc des PSM intéressants qui impliquent des modifications (PTM, ou modifications de séquences dans un PSM). De plus, les modifications portées par un peptide vont décaler la masse des ions qui les portent, et donc les masses des pics du spectre MS2 associé. En effet, avec une seule modification, la moitié des ions portent la modification, et voient donc leur masse décalée de la valeur de la masse de la modification. Ainsi, si on calcule le SPC entre deux spectres qui correspondent au même peptide, il sera divisé par deux si on ajoute une modification au peptide correspondant à l'un des deux spectres (voir Figure 2.8). Ne pas prendre en compte les modifications dans la comparaison des spectres peut donc avoir pour conséquence l'absence d'attribution d'un peptide pour un spectre MS2, ou bien l'attribution d'un candidat erroné qui ne permettra pas d'identifier correctement le spectre.
Chick *et al* [J. M. Chick *et al.* 2015] se sont demandés pourquoi moins de la moitié des spectres de masse étaient associés à un peptide avec une confiance importante. Leur étude a montré qu'avec une grande tolérance de masse (500 Da) lors de la comparaison des spectres, davantage de spectres correspondant à des peptides modifiés pouvaient être identifiés. En s'attaquant au même problème, [Griss *et al.* 2016] rappellent les trois hypothèses sur le peu de spectres identifiés : 1) la qualité des spectres (le bruit qu'ils comportent, ou bien les pics manquants) ; 2) l'absence du spectre dans la base de données utilisée ; 3) la présence de modifications non prévues dans la recherche. Grâce à leur combinaison de différentes méthodes de recherche, [Bogdanow, Zauber et Selbach 2016] estiment que les peptides modifiés peuvent être responsables de 20 à 30 % des faux positifs des jeux de données de protéomique, en plus de montrer également que les peptides modifiés sont sources de faux négatifs.

2.6.2 Ajout de modifications dans la base de protéines

Prendre en compte les modifications lors de l'identification des spectres est donc une nécessité ; pour cela, les méthodes de recherche contre une base de données ont mis au

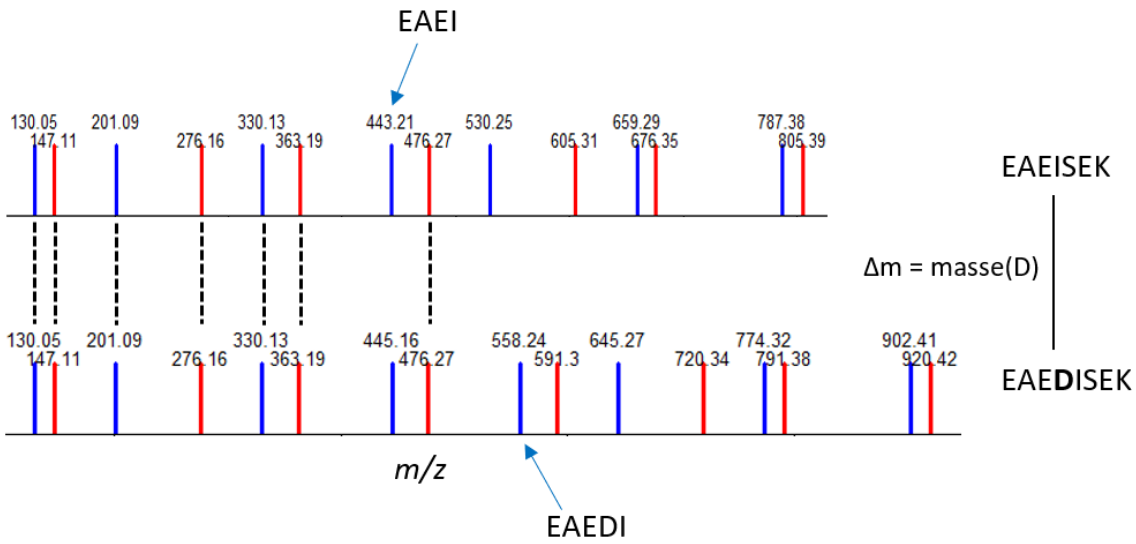


FIGURE 2.8 – **Comparaison des spectres de deux peptides séparés par une modification.** Dans ces deux spectres théoriques, les ions b sont représentés en bleu, les ions y en rouge et les identités de masse entre spectres en lignes pointillées. Les intensités des pics sont en unité arbitraire. Pour illustrer l'impact d'une modification, on crée un spectre théorique à partir du peptide EAEDISEK (en bas). On imagine que l'on souhaite identifier la séquence correspondant à ce spectre. Parmi ses candidats de la base de données se trouve le peptide EAEISEK, dont le spectre théorique est montré en haut ; les deux spectres ont 7 pics avec des masses identiques, et donc un SPC de 7. La seule différence entre les deux peptides est un acide aspartique (D) présent dans le peptide EAEDISEK. Cela entraîne donc un Δm correspondant à la masse de D (115,02 Da). De plus, tous les ions de EAEDISEK qui portent la modification (qui contiennent D), c'est-à-dire les ions b à partir de EAED (b_4) et les ions y à partir de KESID (y_5), ont leur masse décalée de la masse de D. Ils représentent la moitié des ions. Ainsi ce SPC de 7 est égal à la moitié du SPC que l'on aurait obtenu si les deux spectres correspondaient au même peptide EAEISEK.

point des stratégies variées. Pour `Sequest`, la recherche est faite dans la base de données telle quelle, mais il est également possible de prendre en compte certaines modifications fixes, c'est-à-dire qui sont portées par un résidu en particulier à chacune de ses occurrences, notamment à cause des expériences de spectrométrie de masse elles-mêmes ; on considérera par exemple que toutes les cystéines sont carbamidométhylées.

Certaines modifications de peptides sont variables, c'est-à-dire que tous les résidus ne les portent pas systématiquement. Elles peuvent être aussi liées à l'expérience ; l'oxydation de la méthionine entre dans cette catégorie. Les PTM, liées aux caractéristiques biologiques

d'une protéine donnée (comme par exemple la phosphorylation), entrent également dans cette catégorie. `Sequest` considère comme trop gourmande la recherche des modifications variables. En effet, si une modification est considérée comme possible pour un résidu particulier, il faut tester sa présence/absence pour chaque résidu, ce qui peut produire une grande combinatoire de peptides à tester. Cela n'était pas envisageable pour `Sequest`, qui mettait déjà plusieurs minutes pour effectuer une recherche contre une base de données d'environ 100 000 protéines.

La méthode nommée `Tandem` [Craig et Beavis 2003] propose de faire une recherche en plusieurs étapes afin de diminuer de façon stratégique le temps de calcul très important qui peut être induit par la prise en compte de tous les peptides possibles qui peuvent être produits par une même protéine, en termes de modifications comme en termes de sites de clivage. Au cours d'une première étape, on sélectionne les protéines considérées comme d'intérêt, c'est-à-dire qui contiennent des peptides tryptiques semblables au spectre MS/MS selon la masse et un score calculant un produit scalaire entre les spectres. Pour cette première étape, des modifications fixes et variables peuvent êtres ajoutées, mais en nombre limité. Lors d'une seconde étape, les protéines contenant au moins un peptide d'intérêt sont étudiées de plus près. Les données à traiter étant fortement réduites, on peut en effet se permettre d'augmenter l'espace de recherche par le nombre de modifications considérées, et donc d'ajouter des modifications sur certains résidus, et autoriser des sites de clivage non tryptiques, ou bien des *missed cleavages*. La méthode `Tandem` évoquée précédemment a vu son interface améliorée et une utilisation plus aisée des données d'entrée et de sortie sous le nom du logiciel `X!Tandem` [Craig et Beavis 2004]. `Mascot`, toujours couramment utilisé, permet de prendre en compte certaines modifications, fixes ou variables, lors des recherches.

L'ajout de certaines modifications dans la base de données peut être mis en œuvre lorsqu'on se limite à une ou quelques modifications d'intérêt (phosphorylation, ubiquitination, etc.). Cependant la combinatoire induite par le nombre très important de possibilités de modifications et leur emplacement implique un temps de calcul qui peut être quand même trop important si l'on souhaite considérer un grand nombre de modifications possibles. De plus, les méthodes présentées précédemment reposent sur l'introduction de modifications connues ; elles ne sont donc pas adaptées à la découverte de nouvelles modifications.

2.6.3 Les méthodes OMS

Introduire des modifications dans la base de données présente donc des limites. Celles-ci ont motivé le développement des méthodes dites OMS (Open Mass Search), qui sont apparues dans les années 2000 et apportent des éléments de réponses intéressants pour l'identification de peptides modifiés sans l'introduction de modifications dans la base de données, et donc sans *a priori* sur leur nature, leur nombre ou leur emplacement. Les méthodes OMS se définissent par une tolérance de masse importante entre les spectres lorsqu'ils sont comparés. En augmentant ainsi l'espace de recherche en termes de masse, et donc en augmentant le nombre de candidats pour un spectre, les méthodes OMS ont pour objectif de trouver le meilleur candidat pour un spectre possiblement modifié ou muté. Au contraire des méthodes classiques qui tentent d'identifier des paires de spectres que l'on pense correspondre au même composé chimique (c'est-à-dire qu'un spectre de référence idéal est associé à son équivalent expérimental imparfait), les méthodes OMS ont pour objectif d'associer des spectres similaires mais qui peuvent correspondre à des composés chimiques différents, avec des masses différentes. On présume ainsi que le Δm d'un PSM produit par l'une de ces méthodes correspond à une ou plusieurs modification(s) qui sépare(nt) le spectre du peptide.
Un autre avantage crucial des méthodes OMS est qu'elles permettent l'identification des spectres de masses issus de peptides d'espèces peu représentées dans les bases de données. En effet, étant donné que l'on part du principe que les spectres peuvent être modifiés au moment de la comparaison, on peut étudier à la fois les PTM et les modifications de séquence. Il devient donc possible d'identifier des spectres à partir d'une base de protéines proches/homologues aux protéines que l'on étudie, et ainsi se libérer en partie de la spécificité d'une base de données. Ce problème a motivé une partie du travail de Freddy Cliquet, qui a développé l'outil `PSA` [Cliquet *et al.* 2009], dans le cadre d'une thèse (soutenue en 2011) également encadrée par les équipes BIA (INRAE) et ComBi (LS2N). Dans la suite, nous aborderons la diversité des méthodes OMS, qui déploient des stratégies très variées pour relever les différents défis auxquels elles sont confrontées. Afin de montrer cette diversité, les caractéristiques de quatre d'entre elles, qui sont détaillées dans la suite de la section, sont présentées Table 2.2.

TABLE 2.2 – **Résumé du fonctionnement de quatre outils d'identifications de spectres MS2 en OMS selon 3 critères.**

	Filtre des PSM	Sélection du meilleur PSM	Interprétation du Δm
`Pedanta`	Convolution spectrale	Alignement spectral par programmation dynamique	Non
`SpecOMS`	SPC	SPC/shift SPC	Une modification sans *a priori*
`MSFragger`	SPC	SPC/shift SPC	Une modification sans *a priori*
`MODa/MODPlus`	*Tags*	*Tags* et programmation dynamique	Plusieurs modifications avec une composante *a priori*

Réduire la base de données pour réduire le temps de comparaison

Les méthodes OMS sont confrontées à différents obstacles. Elles ont notamment été au départ peu utilisées à cause de leur temps de calcul élevé, qui s'explique par le besoin de comparer un très grand nombre de peptides à un spectre expérimental donné, en raison de la tolérance bien plus importante qu'avec une méthode avec une tolérance de masse restreinte. Afin de faire sauter ce premier verrou, l'une des possibilités est de réduire la base de données de la façon la plus pertinente possible par rapport au spectre à analyser, c'est-à-dire sélectionner rapidement les candidats qui semblent les plus adaptés à une comparaison avec les spectres expérimentaux.
Avec l'amélioration des connaissances dans le domaine des peptides modifiés, des bases de données de modifications ont pu voir le jour (voir Chapitre 1, Section 1.2.2, page 25). La plus connue et utilisée se nomme Unimod [CREASY et COTTRELL 2004]. Certaines méthodes d'identification de peptides modifiés choisissent d'utiliser l'information présente dans ce type de bases de données afin de sélectionner les candidats sur la base de Δm connus (comme `Metamorpheus` [SOLNTSEV *et al.* 2018]) ou fréquents (comme `ModifiComb` [SAVITSKI, NIELSEN et ZUBAREV 2006] qui se base sur un histogramme de Δm pour élaguer les PSM).
Une autre façon de faire est d'exploiter des *tags* présents dans les spectres, en ne recherchant que les candidats qui les contiennent, comme le fait `OpenSea` [SEARLE *et al.* 2005] qui recherche des séquences obtenues *de novo* dans une base de données de protéines. `MODa`

[Na, Bandeira et Paek 2012 ; Na, Kim et Paek 2019] et `TagGraph` [Devabhaktuni *et al.* 2019] utilisent le même principe. Les *tags* sont en effet un puissant moyen de réduire l'espace de recherche pour des spectres modifiés contenant un nombre modéré de modifications.

Un score qui capture une similarité malgré les modifications

La sélection des PSM en OMS peut aussi se faire sur les masses en commun partagées entre les spectres (SPC). Cependant, le SPC diminue rapidement entre deux spectres séparés de k modifications au fur et à mesure que k augmente ; capturer cette similarité est donc un défi. Le second verrou auquel de nombreuses méthodes OMS ont été confrontées est donc le problème suivant : pour un spectre donné, quel est le peptide de la banque qui lui est le plus proche, en prenant en compte une ou plusieurs modifications non prévues, sans *a priori* ? Elles ont donc mis au point des stratégies variées pour sélectionner les PSM en capturant la similarité des spectres, tout en prenant en compte les modifications à l'aide de scores originaux.
Les travaux de Pevzner *et al* [Pevzner, Dančík et Tang 2000 ; Pevzner, Mulyukov *et al.* 2001] sont précurseurs dans le domaine de la recherche OMS. Ils avaient pour objectif de trouver des solutions pour évaluer la similarité entre deux spectres séparés de k modifications, solutions alternatives au SPC. Parmi ces outils, la méthode de *convolution spectrale* a été développée. Elle consiste à évaluer la similarité entre deux spectres en mesurant le nombre de pics dont la masse est séparée de x Da, où x est un paramètre que l'on fait varier. Cependant, il est possible qu'un certain nombre de pics soient séparés d'une masse de x entre les deux spectres, alors qu'il n'existe pas de solution cohérente du point de vue de la séquence du peptide, par exemple si les masses décalées de x sont intercalées avec celles qui ne le sont pas (en effet, si une modification est introduite dans un peptide, les masses des ions qui sont décalées doivent se suivre dans l'ordre croissant).

Des progrès informatiques pour une comparaison exhaustive

Le temps d'exécution important restait néanmoins un obstacle pour l'utilisation des méthodes OMS, mais récemment l'étude de [J. Chick *et al.* 2015] a relancé le développement de méthodes OMS avec des stratégies variées pour réduire le temps de calcul tout en améliorant la sélection des PSM. Depuis, des progrès algorithmiques permettent d'effectuer une comparaison entre chaque spectre expérimental et chaque spectre théorique en OMS. `MSFragger` [Kong *et al.* 2017] est capable de le faire grâce à une indexation

particulière des masses des spectres de la base de données sous forme de bins (intervalles de masse) ; la base de données peut être modifiée si l'utilisateur le souhaite. On a ensuite pour chaque spectre expérimental un certain nombre de candidats selon la tolérance de masse indiquée, qui peut être en OMS (500 Da de tolérance) ou bien en mode restreint (100 ppm de tolérance). Pour chaque fragment du spectre expérimental, on peut rapidement déterminer quels peptides candidats contiennent ce fragment, en accédant aux bins du peptide dans l'intervalle $[M - df,\ M + df]$, où M est la masse du fragment et df la tolérance de masse indiquée par l'utilisateur. Pour chaque peptide candidat, le score est incrémenté de 1 pour chaque pic en commun avec le spectre expérimental. Le fait de limiter ainsi le nombre de comparaisons entre un spectre et un peptide (avec les bins) explique en grande partie la différence de temps de calcul avec les autres algorithmes.
`SpecOMS` [David, Fertin, Rogniaux *et al.* 2017] effectue une comparaison entre chaque spectre en utilisant une structure de données particulière (`SpecTrees`, [David, Fertin et Tessier 2016]) qui permet un calcul rapide du SPC entre chaque couple de spectres. Pour sa rapidité et l'universalité de son score, `SpecOMS` est un logiciel qui a été central dans les analyses de mon travail de thèse ; il sera donc présenté en détail dans le chapitre suivant.

Des librairies spectrales en OMS

Certains outils misent sur la taille réduite des bibliothèques spectrales pour les rechercher en OMS. `MzMod` [Horlacher, Lisacek et Müller 2016] recherche dans des bibliothèques spectrales avec une tolérance de masse importante (réglée par l'utilisateur) en calculant le produit des spectres à identifier et de la bibliothèque, avec l'intensité de deux types de pics, ceux mesurés et ceux décalés de Δm ; cette étape est appelée *recherche hybride*. Pour un certain nombre de candidats, une sélection est faite afin que l'association des spectres soit cohérente du point de vue de la séquence, en positionnant le Δm sur chaque résidu du peptide du spectre de la bibliothèque. Ces calculs volumineux sont possibles grâce à une infrastructure informatique spécialisée (cluster et cloud) qui permet une parallélisation, ainsi que l'utilisation de frameworks pour y distribuer efficacement les tâches.
Cependant, dans [Bittremieux, Meysman *et al.* 2018 ; Bittremieux, Laukens et Noble 2019], les auteurs soulignent que ce type d'infrastructure est coûteux. De plus, ils remarquent que la taille des bibliothèques spectrales est en augmentation constante et que, pour les organismes les plus étudiés, le volume de données rivalise avec celui des bases

de données de peptides. Ainsi, rechercher dans les bibliothèques spectrales en OMS sans matériel excessivement coûteux est un défi qui a motivé le développement d'`ANN-SoLo`. Pour sélectionner les candidats pour un spectre, cet outil effectue une recherche en cascade qui consiste en deux étapes : 1) on extrait les SSMs (*spectrum-spectrum matches*) avec une faible tolérance afin d'identifier les spectres non modifiés ; 2) les spectres qui n'ont pas de match dans la première étape subissent une recherche avec une tolérance de masse importante afin d'identifier les spectres modifiés. Dans ce second cas, pour identifier le plus rapidement et le plus précisément possible l'équivalent non modifié d'un spectre, `ANN-SoLo` compare un spectre à identifier et le candidat d'une bibliothèque spectrale en calculant leur produit à partir des pics originaux, et des pics décalés de Δm pour calculer un produit scalaire en se basant sur la recherche hybride utilisée dans les travaux de [Horlacher, Lisacek et Müller 2016] ou encore de [Burke *et al.* 2017].

Interpréter les modifications

Une fois les candidats sélectionnés, une étape d'alignement est nécessaire pour sélectionner les plus appropriés et interpréter les modifications.
La convolution spectrale, déjà évoquée, est utilisée comme un filtre à candidats avant d'utiliser une autre méthode, la méthode d'**alignement spectral** qui prend en entrée deux spectres. Les masses des pics sont ensuite alignées dans l'objectif de calculer un score qui reflète la similarité entre deux spectres à k modifications près, tout en gardant une cohérence d'un point de vue de la séquence. Cet alignement est fait par programmation dynamique, méthode algorithmique dont les bases mathématiques ont été posées par [Bellman 1952] ; son principe est de résoudre un problème d'optimisation en résolvant les solutions de ses sous-problèmes, puis de les combiner. La programmation dynamique sera discutée plus en détails dans le Chapitre 4. Les deux étapes de convolution spectrale et alignement ont été regroupées dans un outil du nom de `Pedanta` [Pevzner, Dančík et Tang 2000].
Récemment, `MSFragger` a été modifié afin de prendre en compte à la fois les pics mesurés et les pics décalés de Δm lors de la comparaison entre les spectres [Yu *et al.* 2020]. Les pics décalés du spectre théorique sont comparés à ceux du spectre expérimental si Δm est suffisamment important - plus précisément en dehors de l'intervalle de masse [-1,5 ; 3,5 Da] - auquel cas on soupçonne au moins une modification entre les spectres.
`SpecOMS`, après avoir sélectionné les candidats sur la base du SPC, peut ensuite sélectionner le meilleur candidat pour un spectre en prenant en compte les pics décalés de Δm lors

de la comparaison entre les spectres. Le Δm est en effet considéré comme une modification présente sur chaque résidu du spectre théorique, et le SPC est recalculé à chaque emplacement. Cela permet de calculer un nouveau score (le **shift SPC**), mais permet aussi, si le Δm correspond à une seule modification, de la localiser.
L'identification des modifications après la production des PSM peut également se faire selon les modifications connues. `OpenSea`, après la sélection des candidats par *tags*, aligne les masses par programmation dynamique selon les modifications courantes / probables. Dans le même esprit, `MODa` [Na, Bandeira et Paek 2012 ; Na, Kim et Paek 2019] et `TagGraph` [Devabhaktuni *et al.* 2019] utilisent des *tags* alignés sur des spectres théoriques, à la fois pour réduire le nombre de candidats et localiser les modifications éventuelles. En effet, par programmation dynamique, l'alignement pourra révéler leur emplacement.

2.7 Conclusion

Nous avons pu voir que la MS était une méthode très utilisée depuis plusieurs décennies, et les développements dans ce domaine ont permis son application à la protéomique en particulier. Elle a été et est toujours un moteur important des avancées dans le domaine de la biologie.
Le spectre de masse est une donnée complexe, mais son étude peut permettre d'obtenir une information de séquence relative au peptide qui l'a produit. Différentes méthodes existent pour ce faire ; de nombreux outils ont en effet été développés pour analyser les spectres MS2, en particulier en les comparant à une base de données de spectres théoriques.
Les modifications portées par les peptides représentent un défi important. Elles sont essentielles à étudier, mais complexifient l'analyse. Afin de pouvoir découvrir des modifications sans *a priori*, ou identifier des peptides par des peptides de la base de données dont la séquence est proche, les méthodes OMS ont été développées. En autorisant la comparaison des spectres avec des écarts de masse importants, elles permettent d'améliorer l'identification des peptides modifiés.
Les méthodes OMS sont cependant aujourd'hui peu utilisées, notamment car la qualité de leurs résultats est toujours débattue. C'est pour cette raison qu'il est important de les étudier afin de déterminer leurs forces et leurs faiblesses, et donc les améliorer.
Afin de pouvoir dresser un bilan sur ce type de méthodes, durant ma thèse, j'ai étudié les identifications des peptides du protéome humain dans un contexte théorique, c'est-à-

dire que des spectres théoriques sont utilisés à la place de spectres expérimentaux pour réaliser les identifications. La connaissance de leurs séquences m'a permis de mettre en œuvre des méthodes originales à la fois pour déterminer la qualité des identifications et améliorer l'identification des PSM modifiés. Le FDR, introduit dans ce chapitre, peut être utilisé pour la validation des résultats des méthodes OMS ; il sera aussi discuté dans ce manuscrit.
Le point de départ de ce travail a été les résultats fournis par `SpecOMS`. Ce logiciel, ses résultats sur les spectres théoriques et leur analyse sont présentés dans le chapitre suivant.

Deuxième partie

Contributions au sujet de thèse

Chapitre 3

Évaluation de méthodes OMS basée sur des spectres théoriques

Sommaire

Préambule : Dans le chapitre précédent, nous avons pu voir que les méthodes OMS présentent des aspects intéressants pour identifier par MS les peptides qui portent des modifications. Cependant, ces méthodes sont aujourd'hui encore débattues ; leurs résultats méritent donc d'être étudiés en détails afin de statuer sur les avantages et limites de ces approches, dans le but de mieux les utiliser.
Pendant ma thèse, j'ai choisi de me placer dans un contexte où les identifications en mode OMS sont réalisées sur des spectres théoriques à la place de spectres expérimentaux, et cela dans l'objectif de répondre à certaines questions qui pourraient permettre d'améliorer la qualité des identifications des spectres produits à partir de peptides modifiés. Étant donné que les peptides correspondant aux spectres théoriques sont connus, j'ai pu mettre en place des méthodes originales pour évaluer les PSM en observant des phénomènes qui auraient probablement été très difficilement détectables lors de l'identification de spectres expérimentaux.
Pour réaliser les identifications entre spectres théoriques, j'ai utilisé `SpecOMS`, un logiciel d'identification de spectres en mode OMS développé dans le cadre d'une thèse précédente LS2N/INRAE. Ses caractéristiques, présentées dans ce chapitre, le rendent en effet adapté aux analyses que je souhaitais mener.
Ce chapitre présente deux études principales : tout d'abord une exploration des identifications entre spectres théoriques, que j'ai étudiées sous la forme d'un réseau, et ensuite une étude qui se fixait comme objectif de déterminer l'efficacité de deux stratégies OMS implémentées dans `SpecOMS`, et dont les résultats ont été publiés dans [Lysiak *et al.* 2021].

3.1 Évaluer la qualité des identifications de spectres par une méthode OMS

3.1.1 Utilisation de spectres simulés et théoriques

Dans le chapitre précédent, nous avons pu voir pourquoi les approches OMS ont des caractéristiques qui les rendent intéressantes pour étudier les spectres de masse issus de

peptides dont les protéines portent des modifications. Cependant, leur efficacité est encore débattue à l'heure actuelle ; l'étude récente de [Riffle *et al.* 2022] rappelle en effet que les méthodes OMS ont un potentiel de faux positifs plus important que les méthodes avec une faible tolérance de masse, car certaines divergences entre les spectres expérimentaux et théoriques sont interprétées comme des modifications, ce qui peut produire des fausses identifications qui n'auraient pas eu lieu dans une recherche avec tolérance restreinte. Il semble donc nécessaire d'étudier de façon approfondie les résultats issus de recherches en mode OMS, étude qui devrait permettre de mieux comprendre les forces et faiblesses des méthodes OMS, de proposer des pistes d'amélioration, et mieux cibler leur utilisation.
Afin d'obtenir des mesures sur la précision de l'identification des peptides par MS, les travaux réalisés dans le cadre de ma thèse ont été effectués avec des spectres MS2 simulés à partir de séquences de peptides connues, spectres qui représentent les spectres expérimentaux lors de la génération des PSM. Les imperfections inhérentes aux spectres expérimentaux (bruit, pics manquants, etc.) n'ont pas été prises en compte, ce qui permet de se focaliser sur les identifications indépendamment de ces éléments. De plus, étant donné que les spectres simulés sont, comme les spectres théoriques, produits à partir de peptides issus d'une base de données de protéines, les séquences des peptides qui correspondent à chaque spectre sont parfaitement connues. Il est donc possible de découvrir des comportements qui auraient été difficilement perceptibles dans un contexte expérimental, mais aussi de déterminer à quel point le peptide choisi pour un spectre donné est pertinent en comparant les deux séquences impliquées dans un PSM. Les identifications entre spectres simulés et théoriques, parfaitement contrôlées, peuvent être menées de façon à évaluer certains aspects des méthodes d'identification utilisées. Nous verrons notamment comment les spectres simulés issus du protéome humain ont été utilisés pour produire des PSM comportant des modifications de séquences en empêchant l'identification d'un peptide contre lui-même, et ce afin de mettre à l'épreuve les stratégies OMS. Ainsi, grâce à ces spectres simulés, j'ai pu explorer les données d'identification entre peptides et développer de nouvelles méthodes afin d'évaluer la qualité des PSM.
Comme décrit dans le Chapitre 2 (Section 2.3, page 36), un spectre de masse théorique est un ensemble de pics dont les masses sont celles des ions b et y issus de la fragmentation théorique d'un peptide. Les intensités sont habituellement mises à une valeur arbitraire égale pour tous les pics. L'intensité est donc ignorée ; ainsi les pics des spectres théoriques sont représentés uniquement par un ensemble de masses. Les spectres simulés sont ici produits de la même manière, à partir d'un peptide ; ainsi, l'expression "spectres théoriques"

désignera dans le reste du manuscrit à la fois les spectres simulés et théoriques.
Les spectres théoriques ont été identifiés par `SpecOMS` ; les caractéristiques du logiciel, les raisons pour lesquelles il a été employé ainsi que la manière dont les spectres théoriques sont identifiés sont décrites dans la Section 3.1.2.

3.1.2 Configuration du logiciel `SpecOMS`

Généralités

`SpecOMS` [David, Fertin, Rogniaux *et al.* 2017] est un logiciel d'identification de spectres en mode OMS, développé dans le cadre d'une thèse (Matthieu David, soutenue en 2019) dans les équipes ComBi (LS2N) et BIA (INRAE). Il est capable de calculer rapidement le SPC entre chaque couple de spectres au sein d'un ensemble de spectres donné, sans limite de différence de masse parente entre les spectres lorsqu'ils sont comparés. Le logiciel est composé de plusieurs modules, dont une vue d'ensemble est présentée Figure 3.1.
`SpecOMS` prend en entrée un ensemble de spectres expérimentaux à identifier, ainsi qu'une base de données de protéines, dont les peptides permettront de générer les spectres théoriques. Nous avons déjà évoqué la nécessité de prendre en compte les protéines contaminantes lors de l'analyse d'un échantillon (Chapitre 1, Section 1.1.2, page 19) ; pour identifier les spectres issus de telles protéines, une base de protéines contaminantes est aussi prise en entrée par `SpecOMS`. Avant l'exécution, l'utilisateur indique également t, qui est le seuil de SPC à partir duquel il souhaite obtenir un PSM.
À partir des spectres en entrée, la structure de données `SpecTrees` [David, Fertin et Tessier 2016] est construite. Elle est ensuite parcourue par `SpecXtract`, ce qui permet d'extraire tous les PSM avec un SPC au moins égal au seuil de t. Chaque PSM contient les informations suivantes : identifiant du spectre, séquence du peptide, SPC, et Δm, qui est la différence de masse parente entre les deux spectres du PSM.
`SpecTrees` permet de déterminer le SPC entre chaque couple de spectres avec un temps d'exécution qui dépend de t, mais qui est particulièrement intéressant. Ainsi, même si on le compare à un logiciel comme `X!Tandem`, réputé très rapide et configuré avec un nombre limité de modifications variables, `SpecOMS` est plus rapide (sur 3 000 spectres, de l'ordre de 3 minutes pour `SpecOMS` et 20 minutes pour `X!Tandem`).
À ce stade, tous les PSM avec un SPC d'au moins t sont donc obtenus. Après cette étape, chaque spectre expérimental en entrée a 0 (il est alors absent des résultats), 1

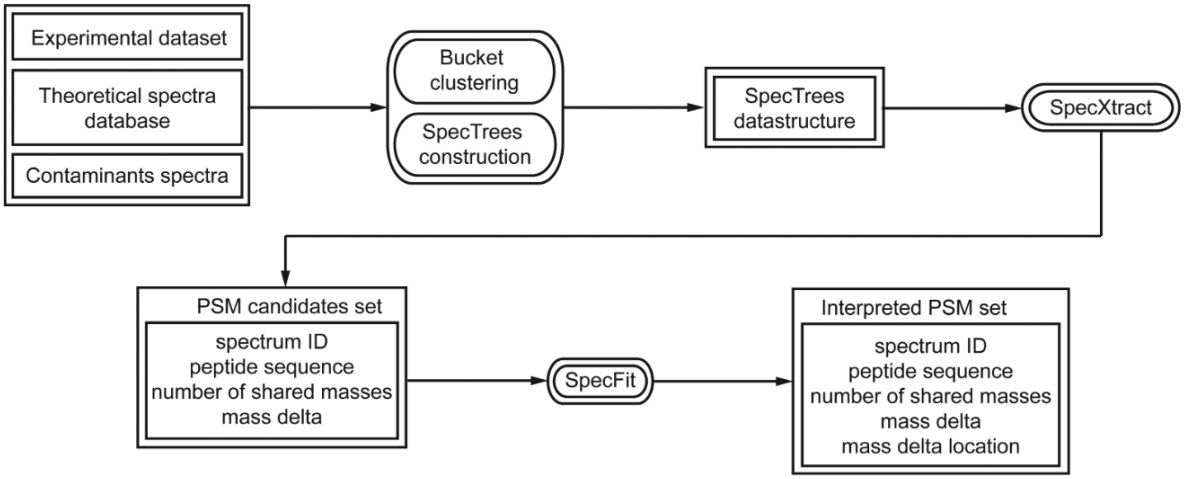


FIGURE 3.1 – **Vue d'ensemble de `SpecOMS`.** Les rectangles représentent les données, et les boîtes rondes représentent les processus. Schéma issu de [DAVID, FERTIN, ROGNIAUX *et al.* 2017].

ou n spectre(s) théorique(s) (issus des peptides candidats) qui lui sont associés, car ces spectres partagent avec lui au moins t pics. Puis, si l'utilisateur le souhaite, `SpecFit` est utilisé ; il a pour objectif de sélectionner un seul spectre théorique (et non n) pour chaque spectre expérimental. Ce spectre théorique peut être choisi suivant la valeur du SPC ; mais l'algorithme `shift` (voir ci-après) peut aussi être utilisé.

L'algorithme de shift de masse : `shift`

`SpecOMS` propose des identifications des spectres sur la base du SPC. Mais si l'utilisateur du logiciel le souhaite, `SpecOMS` est capable d'utiliser le Δm (différence de masse parente entre les deux spectres) pour générer un nouveau score de comparaison, une variante du SPC que nous appellerons le shift SPC. Dans ce cas, les PSM sont produits sur la base de ce score.

Pour déterminer le shift SPC entre deux spectres, un algorithme de réalignement, appelé le shift de masse, ou `shift`, peut être utilisé. La méthode est illustrée Figure 3.2. `shift` va tenter de réaligner (c'est-à-dire de déplacer certaines masses d'un spectre pour qu'elles soient égales à celles de l'autre, et ainsi révéler une similarité) deux spectres (le spectre expérimental et le spectre théorique issu du peptide) en partant de l'hypothèse qu'ils sont séparés par une seule modification, de masse Δm. Pour cela, `shift` simule la présence de la modification sur chaque résidu du peptide ; le Δm est donc déplacé sur chaque ré-

sidu, et pour chaque localisation, le spectre théorique correspondant est modifié et le SPC est calculé entre les deux spectres. Le SPC maximum est le shift SPC, et la localisation correspondante est également renvoyée. `shift` est réalisé pour chaque peptide candidat de chaque spectre expérimental, et renvoie un shift SPC pour chacun d'eux, sur la base duquel le meilleur peut être choisi. `shift` renvoie également une localisation qui permet d'émettre une hypothèse sur la transformation à appliquer au peptide pour retrouver la séquence du spectre. La Figure 3.2 montre que lorsque le Δm est déplacé à une certaine position sur le peptide, les spectres sont alignés, le SPC est maximisé et devient donc le shift SPC.

Une fois que tous les PSM avec un SPC supérieur ou égal à t ont été extraits, le meilleur peptide candidat pour chaque spectre peut ainsi être sélectionné soit simplement avec le SPC, soit avec le shift SPC, selon les paramètres choisis par l'utilisateur. Dans la suite de ce travail, le SPC simple sera appelé raw SPC, pour souligner sa différence avec le shift SPC.

Nous avons vu dans le Chapitre 2 (Section 2.5.1, page 41) que le SPC était un score utilisé directement ou indirectement par de nombreux outils pour capturer la similarité entre spectres. Les caractéristiques de `SpecOMS` (universalité du score et rapidité) rendent ainsi le logiciel intéressant non seulement en tant que tel pour l'identification des spectres de masse, mais aussi pour étudier et examiner les principes qui sont au cœur de la comparaison de ces spectres. `SpecOMS` est donc adapté aux études que je souhaite mettre en place pour explorer les identifications entre peptides et qui sont décrites dans ce chapitre.

Données d'entrée et paramètres de `SpecOMS`

Comme cela a été expliqué Section 3.1.1, les identifications ont été menées avec des spectres théoriques uniquement. Pour cela, `SpecOMS` a été paramétré de façon à comparer non pas des spectres expérimentaux à des spectres théoriques, mais des spectres théoriques à des spectres théoriques. L'entrée, le fonctionnement et les sorties de `SpecOMS` sont illustrés Figure 3.3.

Les spectres théoriques ont été générés à partir des peptides issus du protéome humain (Ensembl 99, [A. D. Yates *et al.* 2020], GrCh38, `http://ftp.ensembl.org/pub/release-99/fasta/homo_sapiens/pep/`). Ce protéome contient 110 210 protéines (annotation *protein-coding genes*) fournies en entrée à `SpecOMS`.

Afin de prendre en compte les contaminants, la base de données dédiée cRAP (*common Repository of Adventitious Proteins*, `http://ftp.thegpm.org/fasta/cRAP/`), de 116

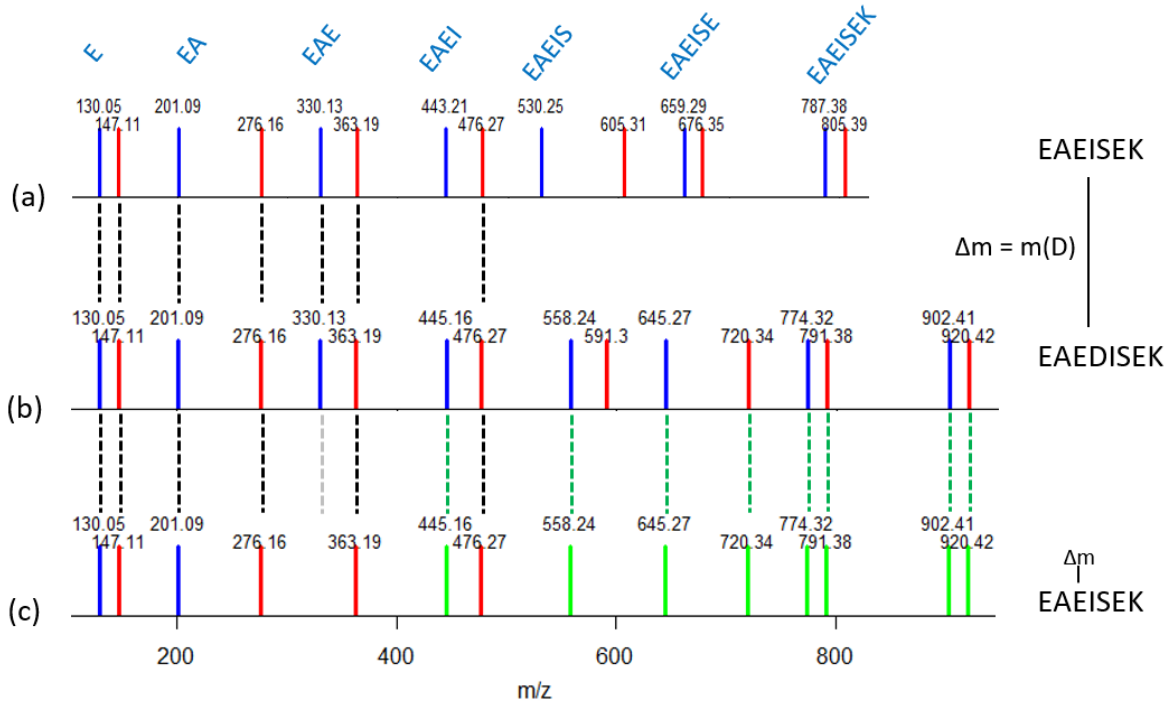


FIGURE 3.2 – **Fonctionnement de l'algorithme de shift de masse `shift` de `SpecOMS`.** `shift` est illustré avec des spectres théoriques, dont les séquences correspondantes sont connues et les intensités sont mises à une même valeur arbitraire. Supposons que l'on souhaite identifier le spectre (b), correspondant au peptide EAEDISEK, grâce au peptide EAEISEK (spectre (a), où les séquences correspondant aux ions b sont indiquées en bleu au dessus des masses) avec lequel il forme un PSM. Les ions b sont représentés en bleu, les ions y en rouge et les égalités de masse entre spectres en lignes pointillées. Le spectre (b) partage 7 masses (lignes noires pointillées) avec le spectre (a). Si le Δm est positionné au troisième résidu dans EAEISEK, le spectre (c) obtenu a 8 nouvelles masses (montrées en vert) qui s'alignent avec celles du spectre (b), et une égalité de masse disparaît (ligne grise pointillée). Le SPC passe ainsi de 7 (raw SPC) à 14 (shift SPC).

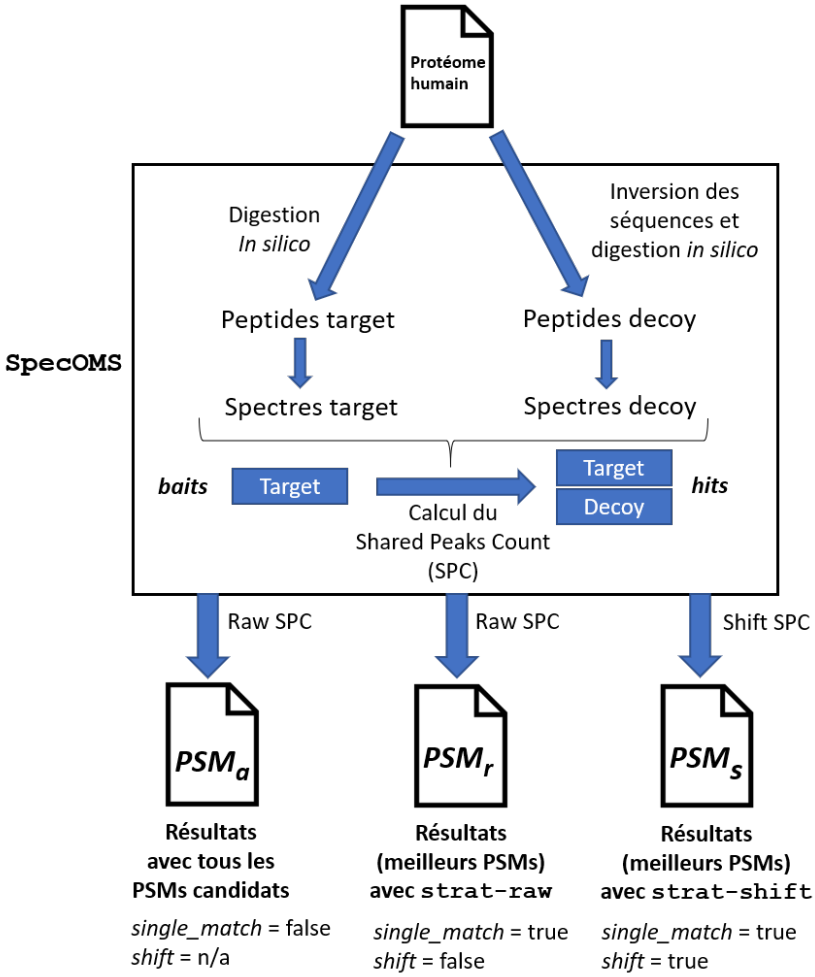


FIGURE 3.3 – **Protocole de l'obtention de tous les PSM, et des résultats selon deux stratégies : `strat-raw` et `strat-shift`.** Les protéines de la base de données (protéome humain) sont digérées en peptides qui forment la base target. Pour créer la base decoy, les protéines sont inversées avant la digestion. Des spectres théoriques sont générés à partir des peptides des deux bases de données. Les spectres issus de la base target, nommés *baits*, jouent le rôle de spectre expérimentaux. Chaque *bait* est comparé à chaque peptide (sauf lui-même) issu des bases non seulement target mais aussi decoy (*hits*). Pour `strat-raw`, le meilleur PSM est obtenu selon le raw SPC. Pour `strat-shift`, l'algorithme `shift` est appliqué et le meilleur PSM suivant shift SPC est retenu.

protéines, a été donnée en entrée à `SpecOMS`.
Le logiciel est paramétré de façon à digérer les protéines *in silico*, à la manière de la trypsine. De la digestion du protéome humain résultent 571 574 peptides distincts d'une longueur comprise entre 7 et 30 résidus, peptides sélectionnés par `SpecOMS` pour produire les spectres (paramètres $minimumPeptideLength = 7$ et $maximumPeptideLength = 30$). La base cRAP permet de produire 1 753 peptides tryptiques (issus d'une digestion après K et R) distincts d'une longueur comprise entre 7 et 30. Ces deux ensembles de peptides sont fusionnés et une fois que la redondance est supprimée, on obtient un ensemble de 572 063 peptides, transformés en spectres théoriques, que nous appellerons la *base target*. Afin de pouvoir discuter de la validation des PSM par le FDR (*False Discovery Rate*) par la stratégie target/decoy (présentée Section 2.5.3, page 47), une *base decoy* a été produite en inversant les protéines de la base target avant leur digestion (méthode "reverse"). Le paramètre $decoyBase$ de `SpecOMS` a donc été mis à "true".
Une instance de la base target joue le rôle des spectres expérimentaux que l'on souhaite identifier. Les spectres (et par extension les peptides correspondants) jouant le rôle de spectres expérimentaux sont appelés ***baits***, car ce sont eux qui seront les "appâts" afin de récupérer un ou plusieurs peptide(s) candidat(s). Chaque *bait* est comparé à la même base (target), et puisqu'on souhaite calculer le FDR, la base decoy est ajoutée à la base target avant l'étape d'identification. Dans cette base (target+decoy), les spectres jouent le rôle de spectres théoriques candidats. Les spectres - et par extension les peptides correspondants - jouant le rôle des spectres théoriques sont appelés ***hits***, nommés ainsi car un *hit* sera une "prise" d'un *bait*.
Cette étude des spectres théoriques a été mise en place afin de comparer l'efficacité de deux stratégies pour identifier des spectres MS/MS modifiés. Si nous comparons directement chaque spectre à la base de données dont il est issu, les auto-identifications, sans modification et par conséquent dépourvues d'intérêt, devraient être systématiques, i.e. un spectre aura le SPC maximum avec lui-même. Pour avoir des PSM riches en modifications et mettre à l'épreuve les stratégies d'identification étudiées, l'auto-identification est interdite lorsque les spectres théoriques sont comparés. Chaque PSM comportera donc au moins une modification de séquence (insertion/délétion/substitution de résidu(s)) qui sépare le *hit* du *bait*.
`SpecOMS` extrait de `SpecTrees` toutes les paires de spectres avec un SPC d'au moins $t = 7$ (paramètre *threshold* mis à 7). Ce seuil semble pertinent car il ne doit pas être trop bas pour ne pas produire des PSM contenant peu d'information, qui de plus seront lents à

produire et formeront un fichier de résultats volumineux. Il ne doit pas être trop haut non plus puisque le SPC peut décroître très rapidement lorsque le nombre de modifications augmente ; or, nous souhaitons autoriser la présence de PSM modifiés. De plus, il n'est pas souhaitable d'écarter les PSM qui pourraient voir leur SPC amélioré par `shift`.
Après cette étape, chaque *bait* possède un certain nombre de *hits* avec lequel il a un raw SPC d'au moins t, et donc un certain nombre de PSM candidats. Les *baits* qui n'ont aucun *hit* avec un SPC d'au moins 7 ne sont pas présents dans les résultats.
Nous obtenons à ce stade des identifications avec 1 à n $hit(s)$ pour chaque *bait*. Pour le travail décrit dans ce chapitre, je me suis intéressée aux résultats de `SpecOMS` sous trois formes (voir Figure 3.4) :

- Toutes les identifications avec, pour chaque *bait*, tous les *hits* avec lesquels il dépasse le seuil t (paramètre *single_match* mis à "false"). Les résultats correspondants sont nommés PSM_a ;
- Le meilleur *hit* pour chaque *bait* (paramètre *single_match* mis à "true") selon le raw SPC (paramètre $shift$ mis à "false"). La méthode pour les produire est nommée `strat-raw` et l'ensemble des PSM correspondants est nommé PSM_r ;
- Le meilleur *hit* pour chaque *bait* (paramètre *single_match* mis à "true") selon le shift SPC (paramètre $shift$ mis à "true"). La méthode pour les produire est nommée `strat-shift` et l'ensemble des PSM correspondants est nommé PSM_s.

Ces trois jeux de PSM produits par `SpecOMS` (environ 3 heures 30 minutes pour obtenir chaque résultat sur ordinateur de bureau sous Windows 10, Intel i7, 2,6 GHz, avec 16 Go de mémoire alloués au programme) à partir des spectres théoriques sont étudiés dans le reste du chapitre et le reste du manuscrit. PSM_a est étudié dans la Section 3.2. PSM_r et PSM_s sont comparés dans la Section 3.3. Je présente en préambule de chaque section les intérêts et les enjeux de l'étude de ces PSM.

3.2 Un réseau de peptides connectés par la MS

3.2.1 Présentation de l'étude

Afin d'avoir une première vision de l'ensemble des PSM entre spectres théoriques, j'ai d'abord exploré les identifications entre spectres théoriques issus du protéome humain en indiquant à `SpecOMS` de sélectionner pour chaque peptide (et donc chaque spectre donné) tous les peptides (et donc leurs spectres correspondants) de la base de données avec les-

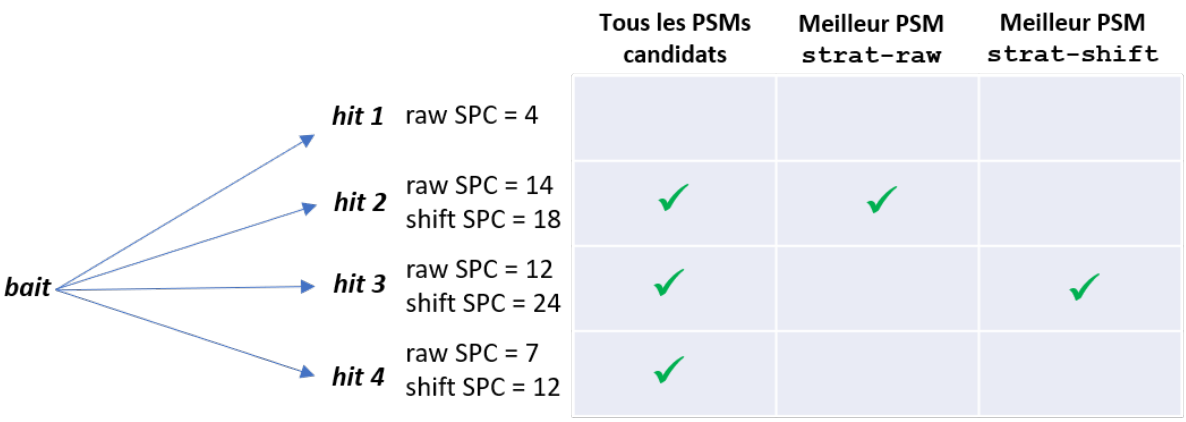


FIGURE 3.4 – **Détermination du meilleur PSM selon les trois stratégies implémentées dans `SpecOMS` et étudiées dans le chapitre.** Supposons qu'un *bait* fictif est comparé à quatre peptides (*hits*). *hit* 1 est écarté car il ne passe pas le seuil t du raw SPC minimum requis, ici 7. Les *hits* 2, 3 et 4 sont des candidats pour le *bait*. Si $\Delta m \neq 0$ pour ces *hits*, `shift` peut être appliqué, et dans ce cas une valeur de shift SPC est calculée. Pour `strat-raw`, le meilleur *hit* est *hit*2, étant donné qu'on se base sur le raw SPC pour choisir le meilleur PSM. Pour `strat-shift`, le meilleur *hit* est *hit*3, puisqu'il a le shift SPC le plus élevé.

quels il partage au moins t pics (résultats PSM_a). Dans chaque PSM, les peptides sont séparés par au moins une modification de séquence (insertion, délétion, substitution de résidus) puisque nous avons exclu les auto-identifications. Une modification de séquence a le même impact sur la similarité entre deux spectres qu'une PTM ; le SPC entre deux spectres issus du même peptide est divisé par deux si une modification sépare les deux séquences (voir un exemple Figure 3.2). Ainsi, le travail présenté dans ce chapitre peut s'appliquer à toutes les sortes de modifications, PTM ou modifications de séquences.
Comme de nombreuses données biologiques, les identifications entre spectres théoriques peuvent assez intuitivement être représentées sous la forme d'un **réseau**. En effet, chaque spectre de la base de données peut être vu comme un nœud, et si deux spectres appartiennent au même PSM, une arête est ajoutée entre ces deux nœuds. On obtient ainsi un réseau non orienté. Dans le contexte de peptides modifiés, un PSM est créé afin d'avoir une base sur laquelle s'appuyer pour déterminer la séquence du peptide qui a produit le spectre, avec des méthodes telles que `shift`. Un PSM exploité de la bonne façon peut ainsi être vu comme un chemin qui permet de passer du *hit* au *bait*, et les PSM peuvent être vus comme des chemins permettant de passer d'un peptide à l'autre dans le réseau. Cette idée a notamment été appliquée par [BANDEIRA *et al.* 2007] pour identifier les

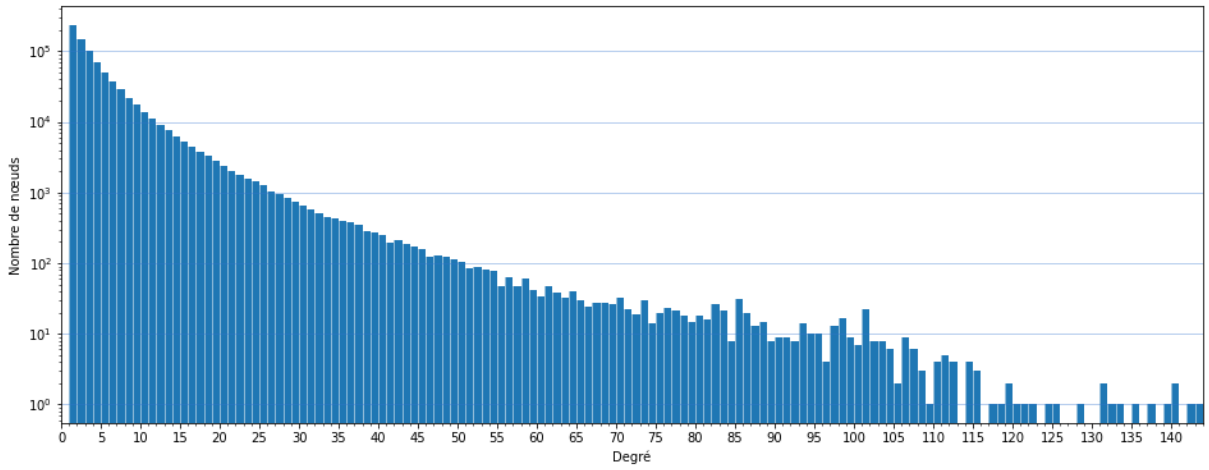


FIGURE 3.5 – **Distribution du degré des nœuds du réseau de peptides** *Rp***.** Pour chaque degré, de 1 à 145, le nombre de nœuds possédant ce degré est indiqué (échelle logarithmique en ordonnée).

spectres expérimentaux issus de peptides modifiés, grâce à un réseau de spectres identifiés de proche en proche. Afin d'avoir une première idée de la possibilité d'identifier les peptides avec un tel réseau, celui produit avec les résultats de `SpecOMS` a été exploré, et les résultats sont présentés dans cette section.

3.2.2 Étude de la similarité des spectres à l'aide d'un réseau des peptides

En exécutant `SpecOMS` sur l'ensemble des peptides du protéome humain, on a obtenu tous les PSM qui dépassent le seuil de raw SPC de 7, avec un ou plusieurs $hit(s)$ par *bait* (PSM_a). À partir des résultats obtenus, j'ai généré le réseau *Rp*, dans lequel tous les spectres théoriques de la base de données - et donc peptides correspondants - sont des nœuds reliés par une arête à partir du moment où ils partagent au moins $t = 7$ pics. *Rp* contient 798 206 nœuds et 1 830 987 arêtes.

L'une des caractéristiques les plus représentatives d'un réseau est la connectivité des nœuds ; autrement dit à quel point les peptides sont-ils reliés entre eux ? Chaque nœud dans un réseau peut être caractérisé par son degré, c'est-à-dire le nombre d'arêtes qui le relient à d'autres nœuds. Dans *Rp*, les peptides ont un degré de 1 à 143. La distribution du degré des nœuds de *Rp* est fournie Figure 3.5.

Les travaux de Barabási et Albert ([BARABÁSI et ALBERT 1999]) ont consisté à modéliser

les réseaux complexes retrouvés dans la nature (réseaux d'interaction protéine-protéine, métabolisme cellulaire, réseaux sociaux, internet, etc.). Dans ces types de réseaux, certains nœuds sont extrêmement connectés (et peuvent donc être qualifiés de *hubs*), alors que la plupart des nœuds sont peu connectés ; un tel réseau est qualifié de "sans échelle" (voir aussi [Barabási et Bonabeau 2003] et [Barabási 2009] pour des articles sur la nécessité d'étudier ce type de réseaux). L'observation de notre histogramme permet de se rendre compte que de très nombreux nœuds sont très faiblement connectés, et très peu de nœuds sont très fortement connectés. *Rp* semble ainsi posséder des propriétés d'un réseau sans échelle.

Afin d'aller plus loin dans l'exploration du réseau, il est intéressant d'en visualiser certaines parties. Cependant *Rp* est évidemment trop volumineux pour être visualisé dans son intégralité. Pour ce faire, j'ai sélectionné les PSM selon deux conditions. D'abord, seuls les nœuds avec un SPC supérieur ou égal à 10 sont conservés ; ensuite, parmi ces nœuds, seuls ceux avec un degré d'au moins 5 sont gardés. Il reste ensuite des nœuds isolés (non connectés), qui sont supprimés de *Rp*. Après ces étapes, nous obtenons un plus petit réseau de 3 661 nœuds et 14 294 arêtes, représenté Figure 3.6.

Le sous-réseau (nommé Rp') permet de mettre en évidence une hétérogénéité de la connectivité des zones. La Figure 3.7 montre des exemples de deux zones de Rp', une zone faiblement connectée (a) et une zone fortement connectée (b). Il est possible d'établir une corrélation entre certaines structures de Rp' et les propriétés des nœuds. En effet, si nous prenons en compte l'origine des nœuds (target en bleu, decoy en rouge), nous pouvons observer un lien entre les zones fortement connectées et l'origine du peptide. Nous pouvons voir dans la Figure 3.7 mais aussi dans Rp' dans sa totalité (Figure 3.6) que les peptides de la base target sont plus fortement connectés entre eux, alors que les peptides de la base decoy le sont moins.

Nous pouvons ajouter les séquences des peptides sur les nœuds qui les représentent ; une zone de Rp' avec cette information est représentée Figure 3.8. Dans certaines zones faiblement connectées, les peptides ont une séquence riche en longues successions du même résidu et une faible complexité en termes de résidus, par exemple GGGGGGGGGGSGR et GGSGGGGGGGGGGYNR. Rappelons que la base decoy est supposée représenter le hasard. Puisqu'on a pu voir un lien entre les peptides decoy et ces zones faiblement connectées, il est possible que les peptides target avec une faible complexité en termes de résidus aient davantage tendance, comme les peptides decoy, à être associés à de mauvaises identifications. Ainsi, la connectivité du réseau ainsi que la faible complexité des

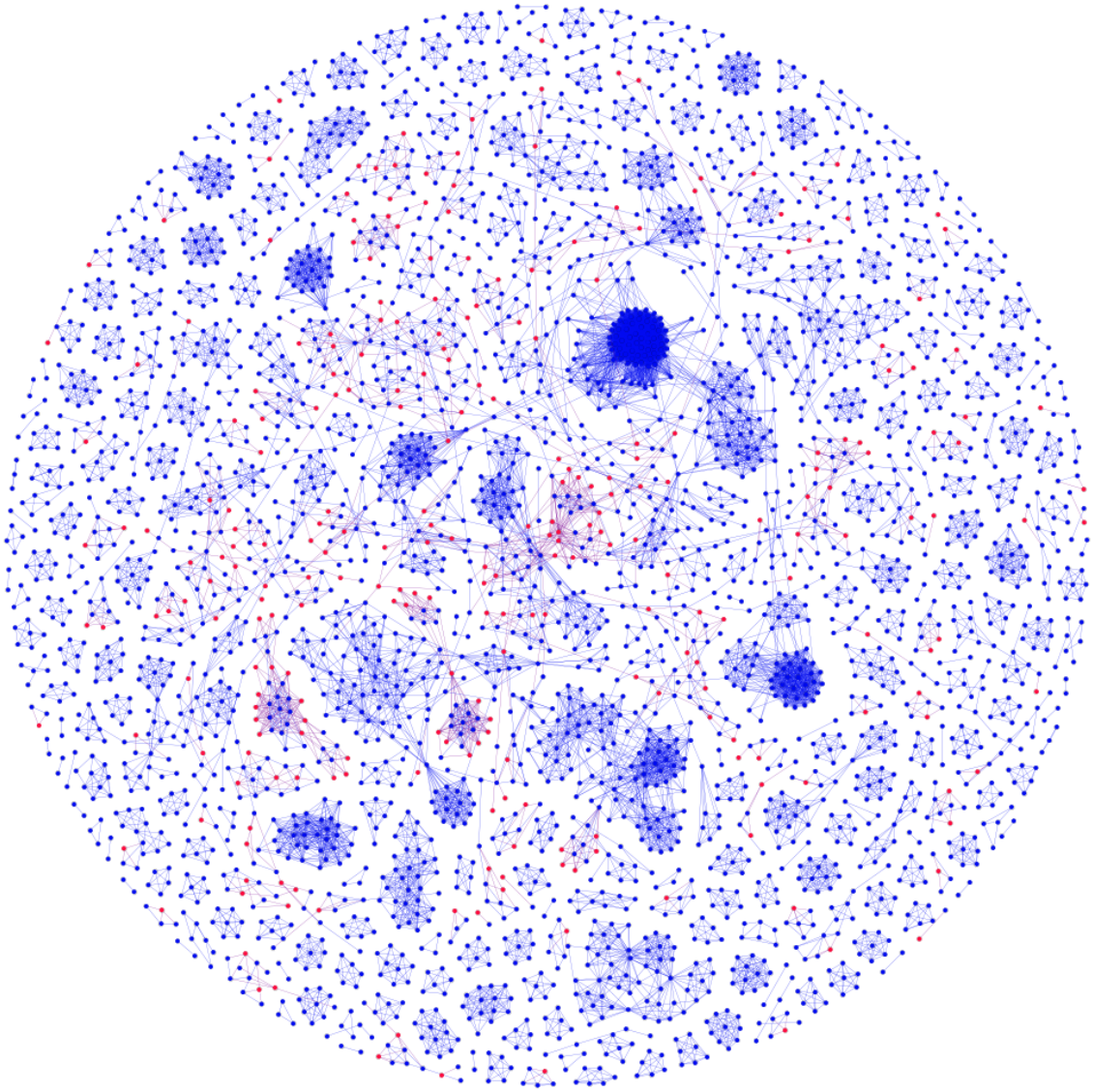


FIGURE 3.6 – **Le réseau de peptides** Rp'**.** Ce réseau contient 3 661 nœuds et 14 294 arêtes. Les nœuds sont colorés selon leur origine : target en bleu et decoy en rouge.

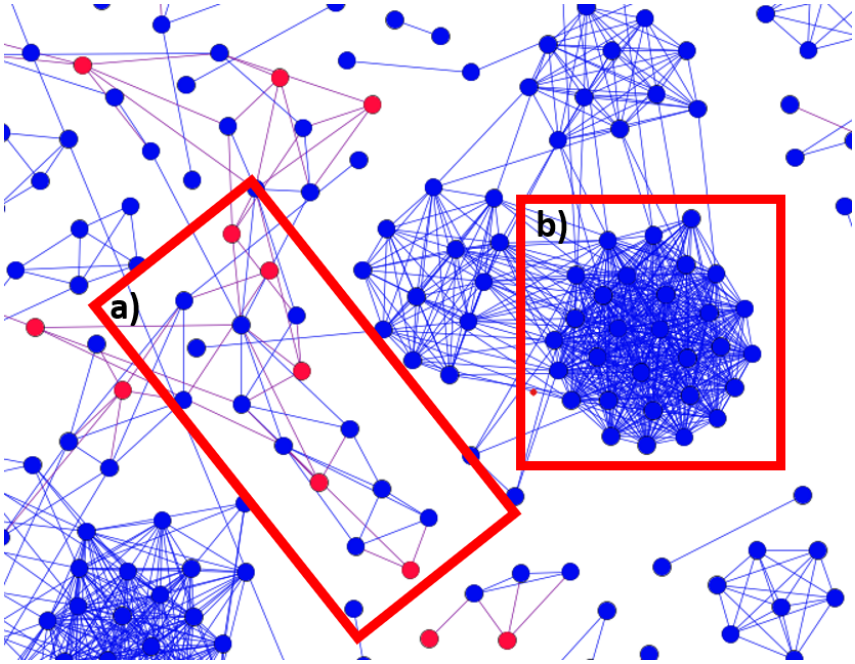


FIGURE 3.7 – **Exemples de zones de faible et forte densité dans** Rp'**.** a) Zone de faible densité b) zone de forte densité. Les nœuds sont colorés selon leur origine, target en bleu et decoy en rouge.

peptides semblent être des éléments qui pourraient être pris en compte pour évaluer la qualité des identifications dans un contexte expérimental, où les spectres expérimentaux seraient utilisés pour construire ce type de réseau.

Le travail présenté dans cette section a montré qu'il existait de nombreux nœuds fortement connectés (de nombreux *baits* ont de nombreux *hits*). Il est probable que, parmi toutes ces connexions, beaucoup ne permettent pas de retrouver la séquence du *bait* et ne sont par conséquent pas pertinentes. De plus, rappelons que les méthodes de traitement des PSM pour l'identification des séquences et des modifications peuvent être chronophages et ne peuvent parfois pas se permettre de prendre en entrée plusieurs candidats par spectre. Par conséquent, ce réseau d'identification doit être élagué afin de pouvoir se focaliser sur les identifications qui permettront de retrouver la séquence du *bait*.

Le seuil de SPC peut être un moyen de réduire le réseau d'identifications ; cependant avec un seuil trop haut, peu d'identifications sont conservées, et si le seuil est trop bas, le problème est le même que pour le réseau entier, et les identifications sont trop nombreuses.

Un moyen pertinent de filtrer les identifications est de se concentrer sur la signification d'un PSM donné (que signifie le lien ?) pour que chaque PSM soit utile pour élucider la sé-

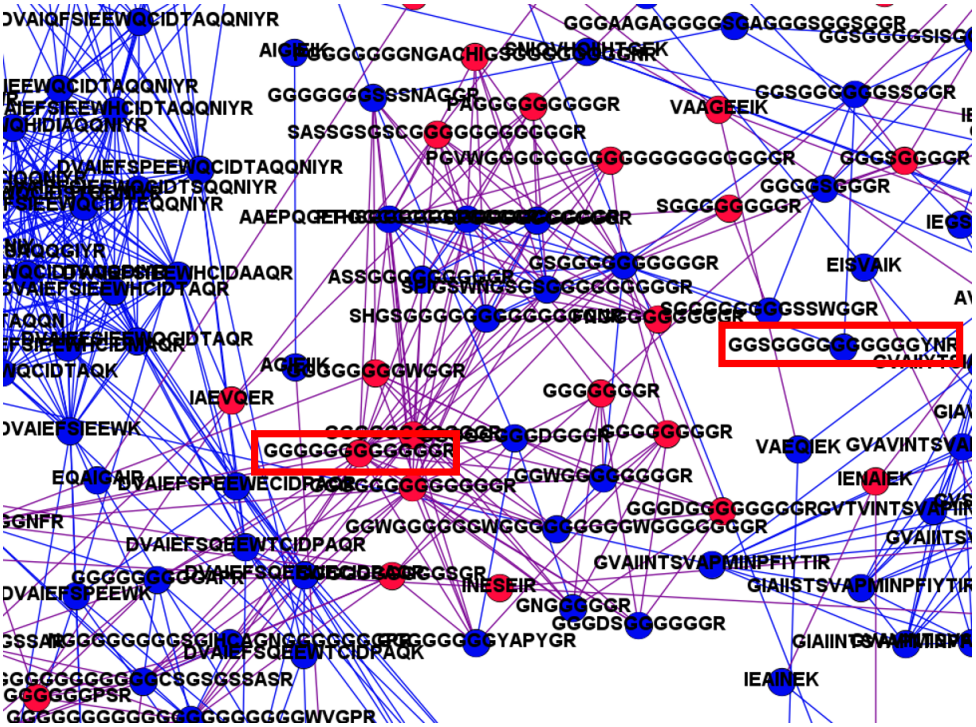


FIGURE 3.8 – **Zoom sur une partie du réseau des peptides** Rp'**.** La séquence des peptides est indiquée sur chaque nœud. Cette zone de faible densité contient de nombreux peptides avec de longues successions du même résidu (ici G, la glycine). Deux exemples de tels peptides (GGGGGGGGGGGGGR et GGSGGGGGGGGGGGYNR) sont encadrés en rouge. Les nœuds target sont colorés en bleu et les nœuds decoy en rouge.

quence correspondant au spectre. Mais comment sélectionner les liens les plus pertinents ? Une possibilité est de ne conserver qu'un seul candidat par spectre, le meilleur selon un score donné. En effet, comme expliqué dans le Chapitre 2, la plupart des outils d'identification sélectionnent un seul candidat par spectre. C'est pourquoi j'ai décidé, pour la suite de cette étude, de me concentrer sur une seule identification par spectre théorique (et non plus toutes celles qui passent le seuil de t). Dans les sections suivantes, les résultats que j'ai étudiés sont donc ceux de `SpecOMS` qui a été paramétré afin de ne conserver qu'un seul *hit* par *bait* (PSM_r et PSM_s).

3.3 Comparaison de deux stratégies de recherche OMS

3.3.1 Présentation de l'étude

Comme discuté dans le chapitre précédent (Section 2.6.3), les méthodes OMS utilisent des stratégies diverses afin de sélectionner le meilleur peptide pour un spectre parmi les différents candidats de la base de données, et ce même si des modifications séparent le peptide qui correspond au spectre expérimental du peptide candidat. Les outils présentés dans le Chapitre 2 peuvent notamment être divisés en deux catégories selon le choix qu'ils font d'utiliser ou non le Δm dans ce processus de sélection.

Le Δm représente la somme des masses des modifications qui séparent le peptide correspondant au spectre de son candidat. Il constitue une information délicate à manipuler, car il peut être ambigu, c'est-à-dire qu'il peut correspondre à une, mais aussi plusieurs, voire à de nombreuses possibilités de modifications qui séparent deux peptides associés dans un PSM. Il comporte donc une information certes parfois partielle (s'il représente plusieurs modifications, il reste une ambiguïté sur leur nature) mais cruciale pour identifier les modifications.

Une première stratégie consiste à ne pas utiliser le Δm lors de la sélection des candidats, comme dans la version initiale de MSFragger [Kong *et al.* 2017] ou bien MetaMorpheus [Solntsev *et al.* 2018], qui peuvent cependant l'utiliser pour l'interprétation des PSM *a posteriori*.

Certaines méthodes emploient une autre stratégie, et intègrent le Δm dans leur calcul de similarité, lors d'une comparaison entre un spectre et les candidats de la base de données afin de choisir le meilleur. Par exemple ANN-SoLo [Bittremieux, Meysman *et al.* 2018 ; Bittremieux, Laukens et Noble 2019] intègre dans son score les pics modifiés avec le "shifted dot product" (produit scalaire en prenant en compte toutes les masses déplacées de Δm simultanément). La recherche hybride [Burke *et al.* 2017], intégrée dans le navigateur de librairie spectrale du NIST (`https://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:nistmssearch`), fait de même avec la similarité cosinus. Pour intégrer le Δm dans le score, les pics du spectre à identifier sont donc comparés aux pics du spectre théorique issu du peptide (raw SPC) mais aussi aux pics du spectre théorique décalés de Δm pour choisir un candidat plus pertinent au regard de la valeur de Δm.

Impliquer le Δm dans le calcul du score peut ainsi être pertinent dans le sens où un spectre et un candidat qui lui est égal à une modification (ou plusieurs modifications très proches) près peuvent être réalignés ; cependant, des erreurs - pics réalignés de façon er-

ronée - peuvent survenir à cause de l'augmentation de l'espace de recherche en termes de nombre de pics créés. Étudier les différences entre les stratégies OMS qui utilisent ou non le Δm pour produire les PSM semble ainsi être adapté pour explorer la meilleure manière de sélectionner et interpréter le meilleur candidat pour un spectre modifié.
`SpecOMS` peut utiliser les deux stratégies présentées dans cette section (voir Section 3.1.2). `strat-raw` consiste à choisir le meilleur candidat sans intégrer le Δm au score de similarité (le raw SPC), alors que `strat-shift` utilise ce dernier pour sélectionner le meilleur candidat à partir du shift SPC. Puisque ces deux stratégies sont implémentées dans `SpecOMS`, nous pouvons les comparer.

3.3.2 Vue d'ensemble des PSM

Pour chaque stratégie, après la comparaison des 572 063 spectres de la base target avec les 1 148 608 spectres des bases target et decoy avec un seuil de $t = 7$, `SpecOMS` renvoie 455 404 PSM.
Le nombre de PSM renvoyés par une des deux stratégies dépend uniquement du nombre de *baits* qui ont au moins un *hit* dans la base de données. En effet, dès qu'un *bait* a au moins un *hit* avec un raw SPC d'au moins 7, alors il aura un PSM pour `strat-raw` et pour `strat-shift` (voir Figures 3.3 et 3.4). Le nombre de PSM est donc identique pour les deux stratégies.
Chaque PSM contient les informations suivantes : *bait*, *hit*, SPC, Δm, $shiftlocation$. Le *hit* est le meilleur candidat pour *bait* selon la stratégie, le SPC est le score utilisé selon la stratégie donnée (raw SPC pour `strat-raw`, shift SPC pour `strat-shift`) et Δm, exprimé en Dalton, est la différence de masse entre le *bait* et le *hit* (masse de *bait* moins masse de *hit*). $shiftlocation$ représente le numéro du résidu du *hit* pour lequel l'insertion du Δm a maximisé le SPC, qui est devenu le shift SPC. Les résultats relatifs aux deux stratégies sont présentés dans les Tables 3.1 et 3.2. Nous pouvons voir que, sur les 572 063 peptides de la base target, environ 80% partagent au moins 7 masses avec un autre peptide de la base (target ou decoy), et 23% en partagent au moins 10.
Même si elles renvoient le même nombre de PSM, les deux stratégies produisent des PSM différents. Environ 37% des PSM (167 291 *baits*) diffèrent entre PSM_r et PSM_s, c'est-à-dire que le *bait* possède un *hit* différent.
Les PSM ont été divisés en trois groupes selon la valeur de leur Δm. Le groupe G_1 représente les PSM avec $\Delta m = 0$ (le *bait* et le *hit* ont la même masse). Les groupes G_2 et G_3 représentent respectivement les PSM avec $\Delta m > 0$ (la masse du *bait* est supérieure à

celle du *hit*) et $\Delta m < 0$ (la masse du *bait* est inférieure à celle du *hit*).
Lorsque le Δm est nul, aucun shift de masse ne peut être réalisé ; le score reste le même pour les deux stratégies. Ainsi, un certain nombre de PSM avec un $\Delta m = 0$ pour `strat-raw` peuvent avoir un shift SPC élevé lorsque leur *bait* est comparé avec d'autres *hits* avec lesquels Δm est non nul, et donc ont un *hit* différent et passeront de G_1 à G_2 ou G_3. 127 949 PSM sont dans ce cas.

TABLE 3.1 – **Nombre de PSM obtenus par `strat-raw` en fonction de raw SPC**

	$\Delta m = 0$		$\Delta m < 0$		$\Delta m > 0$		Total		
Min *raw SPC*	**#target**	**#decoy**	**#target**	**#decoy**	**#target**	**#decoy**	**#target**	**#decoy**	**FDR(%)**
7	71 852	87 705	88 895	39 432	107 842	59 678	268 589	186 815	41,02
8	71 852	87 705	55 874	12 352	61 225	19 550	188 951	119 607	38,76
9	32 972	32 740	41 378	3 874	40 085	4 822	114 435	41 436	26,58
10	32 970	32 739	31 973	868	30 245	1 026	95 188	34 633	26,68
11	10 918	11 209	26 095	310	24 547	305	61 560	11 824	16,11
12	10 918	11 208	21 553	112	20 313	94	52 784	11 414	17,78
13	2 571	1 180	17 893	46	16 939	34	37 403	1 260	3,26
14	2 571	1 180	14 955	22	14 243	17	31 769	1 219	3,7
15	1 137	380	12 352	9	11 818	9	25 307	398	1,55
16	1 137	380	10 239	6	9 801	5	21 177	391	1,81
17	**672**	**61**	**8 403**	**5**	**8 085**	**4**	**17 160**	**70**	**0,41**
18	672	61	6 907	4	6 662	3	14 241	68	0,48
19	495	49	5 590	2	5 413	1	11 498	52	0,45
20	495	49	4 557	1	4 415	1	9 467	51	0,54

Chaque ligne représente un seuil de score (raw SPC). Les meilleurs PSM selon `SpecOMS` avec un raw SPC au moins égal à ce seuil sont séparés en trois groupes selon leur Δm, et présentés selon l'origine du *hit* (base target ou decoy). Un FDR $< 1\%$ est atteint lorsque le raw SPC est supérieur ou égal à 17 (ligne en gras).

La base decoy a été mêlée à la target avant l'étape d'identification des *baits*. Chaque *bait* a ainsi pour *hit* soit un peptide de la base target, soit un peptide de la base decoy. Nous pouvons ainsi appliquer la stratégie target/decoy et calculer le FDR pour chaque stratégie et pour chaque seuil de score, en divisant le nombre de PSM decoy (PSM avec un *hit* issu de la base decoy) par le nombre total de PSM. Nous obtenons ainsi pour chaque stratégie le score minimal (raw SPC ou shift SPC) à partir duquel nous considérerons un résultat acceptable au regard du FDR pour un pourcentage donné.
Avec un FDR de 1%, `strat-raw` valide les PSM avec un raw SPC ≥ 17 (17 160 PSM), alors que `strat-shift` valide les PSM avec un shift SPC ≥ 21 (57 784 PSM). Ainsi, selon la stratégie target/decoy, `strat-shift` est plus intéressante en termes de nombre de PSM validés, qui sont plus de trois fois plus nombreux comparés à `strat-raw`. 3 à

TABLE 3.2 – **Nombre de PSM obtenus par `strat-shift` selon le shift SPC**

	$\Delta m = 0$		$\Delta m < 0$		$\Delta m > 0$		**Total**		
Min shift SPC	**#target**	**#decoy**	**#target**	**#decoy**	**#target**	**#decoy**	**#target**	**#decoy**	**FDR(%)**
7	14 207	17 401	145 801	86 218	122 479	69 298	282 487	172 917	37,97
9	5 132	4 787	145 756	86 183	122 465	69 287	273 353	160 257	36,96
11	1 944	1 793	144 795	85 247	121 833	68 671	268 572	155 711	36,7
13	721	343	139 895	80 327	119 006	65 901	259 622	146 571	36,08
15	496	187	116 674	61 154	110 457	58 923	227 627	120 264	34,57
17	341	42	58 584	13 053	55 297	13 052	114 222	26 147	18,63
19	250	36	37 840	16 81	35 439	1 647	73 529	3 364	4,37
21	**201**	**12**	**29 675**	**271**	**27 908**	**251**	**57 784**	**534**	**0,92**
23	160	10	24 262	87	22 881	82	47 303	179	0,38
25	131	5	20 036	35	18 910	36	39 077	76	0,19
27	108	4	16 686	18	15 792	14	32 586	36	0,11
29	92	2	13 835	12	13 120	9	27 047	23	0,08
31	67	2	11 345	4	10 801	5	22 213	11	0,05
33	54	0	9 314	2	8 888	3	18 256	5	0,03
35	44	0	7 608	2	7 296	2	14 948	4	0,03
37	39	0	6 231	2	5 989	2	12 259	4	0,03
39	17	0	5 008	1	4 832	1	9 857	2	0,02

Chaque ligne représente un seuil de score (shift SPC). Les meilleurs PSM selon `SpecOMS` avec un shift SPC au moins égal à ce seuil sont séparés en trois groupes selon leur Δm, et présentés selon l'origine du *hit* (base target ou decoy). Un FDR $< 1\%$ est atteint lorsque le shift SPC est supérieur ou égal à 21 (ligne en gras). Seuls les résultats avec un shift SPC minimum d'une valeur impaire sont montrés ici ; en effet, le shift SPC est toujours pair, étant donné qu'un réalignement d'ions b implique aussi le réalignement des ions y correspondants. Les résultats obtenus sont ainsi les mêmes pour les scores pairs et impairs consécutifs (mis à part de rares exceptions dues à des masses redondantes).

10% (`strat-raw` et `strat-shift`) des 571 574 spectres théoriques ont donc au moins un peptide de la base target avec lequel ils partagent suffisamment de masses pour que le PSM ne soit pas considéré comme étant le résultat du hasard au regard du FDR.

Lors d'une identification par MS, lorsqu'un PSM a un $\Delta m = 0$, le logiciel utilisé suggère habituellement que le peptide identifié et celui qui a produit le spectre sont identiques. Or, nous savons que ce n'est pas le cas dans notre étude, étant donné que l'auto-identification est interdite. Très peu de PSM du groupe G_1 ont été validés par le FDR (672 PSM avec `strat-raw`, 201 PSM avec `strat-shift`), ce qui est cohérent avec la composition de ce groupe. Les groupes G_2 et G_3 représentent l'immense majorité des PSM validés à un FDR inférieur à 1%.

Cependant, une fois les PSM sélectionnés ou validés, une question au moins aussi importante est la suivante : à quel point l'information contenue dans un PSM permet-elle

d'avoir une bonne interprétation du spectre ? La réponse à cette question est une autre façon de juger de la qualité d'un PSM.
La connaissance des séquences peptidiques de chaque *bait* m'a permis de mettre au point de nouveaux indicateurs pour répondre à cette question. Dans la Section 3.3.3, je détaille ces nouveaux indicateurs, et dans la Section 3.3.4, je les applique aux résultats des deux stratégies. Les observations faites sur le réseau de peptides Section 3.2 liées à la complexité des peptides en termes de résidus permettront également de comparer les deux jeux de résultats.

3.3.3 Nouveaux critères pour évaluer les stratégies OMS

Classification des PSM en couleurs

Dans un PSM renvoyé par une méthode OMS, si le Δm est non nul, le spectre correspond à une molécule différente du spectre identifié (soit les deux peptides ont des séquences identiques, et le Δm correspond à une PTM, soit les séquences sont différentes). Ainsi, une étape de reconstruction est nécessaire si l'on veut connaître l'identité du peptide associé à un spectre expérimental. Dans le protocole mis en place pour cette étude, au moins une modification de séquence sépare le *bait* de son *hit* puisque les auto-identifications sont exclues. Afin de juger de la qualité d'une identification, il faut donc se poser la question suivante : à quel point peut-on reconstruire la séquence du *bait* avec les informations contenues dans le PSM ?
Pour répondre à cette question, j'ai mis au point une classification des PSM en trois couleurs afin de caractériser le niveau de difficulté avec lequel il est possible de reconstruire une séquence peptidique à partir du *hit*, du Δm et de la localisation déterminée par `shift` ($shiftlocation$).
Les PSM avec un $\Delta m \neq 0$ ne peuvent être expliqués que par des différences en termes de résidus, c'est-à-dire une ou plusieurs insertion(s), délétion(s) ou substitution(s) qui doivent être réalisées dans le *hit* afin de retrouver le *bait*. Ainsi, si ces opérations sont identifiées, il est possible de reconstruire la séquence du *bait*. Dans nos jeux de résultats, lorsque le *bait* et le *hit* sont séparés par une modification de séquence connue grâce à `shift`, le PSM est interprétable sans ambiguïté. C'est le cas lorsque la modification de séquence correspond à l'insertion d'un résidu unique (et identifié par Δm) dans le *hit* pour retrouver le *bait* ; c'est aussi le cas lorsqu'il faut effectuer une substitution d'un résidu en un autre, ou bien supprimer un ou plusieurs résidus dans le *hit* pour obtenir le

bait. Ce type de PSM, pour lesquels il est possible de retrouver la séquence complète du *bait* grâce à `shift`, est classé comme Vert. Cependant, il faut parfois insérer plusieurs résidus consécutifs dans le *hit* pour reconstruire le *bait*. Cette situation soulève une ambiguïté, étant donné que le Δm peut potentiellement correspondre à plusieurs ensembles distincts de résidus. Par ailleurs, pour un ensemble de résidus à insérer, leur ordre dans la séquence reste indéfini. Ces PSM sont alors classés comme Orange. Les autres PSM sont classés comme Rouges ; cette couleur est appliquée aux PSM avec un Δm nul, ou bien lorsque `shift` n'a pas permis le réalignement d'un nombre de masses suffisant pour le classer dans une autre catégorie. Des exemples des PSM des trois couleurs sont présentés Figure 3.9. L'algorithme permettant d'établir cette classification est décrit Algorithme 1. Cette classification se base principalement sur deux éléments :

— La comparaison des deux séquences du *hit* et du *bait* dans un PSM pour déterminer si elles sont séparées par l'insertion, la délétion ou la substitution d'un résidu (PSM Vert), la suppression de plusieurs résidus dans le *hit* (PSM Vert), ou bien l'insertion d'une séquence de plus d'un résidu dans le *hit* (PSM Orange) à l'endroit indiqué par $shiftlocation$;

— Le shift SPC et le Δm. Si le Δm est négatif (le *bait* est plus léger que le *hit*) et que le shift SPC est égal au double de la longueur du *bait*, toutes les masses du *bait* ont été réalignées sur celles du *hit*. On peut donc substituer plusieurs résidus du *hit* par un seul pour obtenir le *bait* (PSM Vert). Si le Δm est positif et que le shift SPC est égal au double de la longueur du *hit*, toutes les masses du *hit* ont pu être réalignées sur celles du *bait*. Il faut alors substituer un résidu du *hit* en une séquence d'au moins deux résidus pour obtenir le *bait* (PSM Orange).

À noter que les couleurs peuvent être déterminées sur PSM_r ; en effet, même si un PSM n'a pas été obtenu en se basant sur le shift SPC, `shift` peut quand même être appliqué *a posteriori* afin d'obtenir un shift SPC et $shiftlocation$ pour localiser la modification et interpréter le PSM.

LIPR

L'interprétation des spectres de masse issus de peptides repose sur une ressemblance de masses afin d'inférer une ressemblance de séquences. Mesurer la ressemblance de séquences de cette façon est pertinente dans le sens où deux ions de la même série (b ou y) avec des séquences identiques (et sans PTM) auront nécessairement la même masse. En revanche, l'inverse n'est pas forcément vrai. En effet, deux séquences différentes peuvent avoir la

Algorithme 1 : Classification Vert/Orange/Rouge d'un PSM

Entrée : Un PSM = ($bait$, hit), shift SPC, Δm, $shiftlocation$
Sortie : Une couleur (Vert, Orange ou Rouge) assignée au PSM d'entrée

si Δm *correspond à l'insertion, la délétion ou la substitution d'un seul résidu* **alors**
  couleur ← Vert #*bait* peut être retrouvé à partir du *hit* : insertion, délétion ou substitution du résidu à $shiftlocation$
sinon si *le bait est présent dans le hit en une ou deux parties* **alors**
  couleur ← Vert #Le *bait* peut être retrouvé à partir du *hit* : Δm correspond à une séquence devant être supprimée du *hit* à $shiftlocation$
sinon si *le hit est présent dans le bait en une ou deux parties* **alors**
  couleur ← Orange #une séquence de masse Δm doit être insérée dans le *hit* à $shiftlocation$ pour retrouver le *bait* ; cependant sa longueur, composition et/ou ordre en résidus ne sont pas uniques
sinon
  si $\Delta m < 0$ **alors**
    si *shift SPC* $= 2 \cdot longueur(bait)$ **alors**
      couleur ← Vert #shift SPC est maximisé, ainsi les résidus à $shiftlocation$ dans *hit* peuvent être changés en un seul résidu pour retrouver le *bait*
    sinon
      couleur ← Rouge #shift SPC n'est pas maximisé, donc shift n'a pas pu réaligner les masses, et ainsi aucune information concluante ne peut être obtenue
    fin
  sinon si $\Delta m > 0$ **alors**
    si *shift SPC* $= 2 \cdot longueur(hit)$ **alors**
      couleur ← Orange #bien que le shift SPC soit maximisé, il n'est pas possible de retrouver le *bait* à partir du *hit* sans ambiguïté
    sinon
      couleur ← Rouge #shift SPC n'est pas maximisé, ainsi aucune information concluante ne peut être obtenue
    fin
  sinon
    couleur ← Rouge #le *bait* et le *hit* sont différents par définition, $\Delta m = 0$ signifie que shift n'est pas appliqué, ainsi aucune information concluante ne peut être obtenue
fin
retourner couleur

FIGURE 3.9 – **Exemples de PSM classés dans les catégories Vert/Orange/Rouge.**

Bait	*Hit*	*Δm* (Da)	*Shift location*	Interprétation : du *hit* au *bait*
TSSDSISR	TSDSISR	87,0320	3	TSDSISR + S = TSSDSISR
YQEFQNR	EPPNPEYQEFQNR	-663,2864	6	EPPNPEYQEFQNR – EPPNPE = YQEFQNR
ETVHIPGAR	ETIPGAR	236,1273	2	ETIPGAR + Δm = ET-Δm-IPGAR
VISPEDGK	VIESPDGK	0	/	/
VCASIAQK	VTIQCQK	0	/	/
WFSIYDQR	FWSIQDYFR	-147,0684	/	/

Les deux premières lignes montrent des PSM avec un *bait* déductible sans ambiguïté de l'information donné par un PSM. De tels PSM sont classés comme Verts. Dans le premier exemple, Δm correspond à la masse de la sérine (S), qui peut être ajoutée dans le *hit* à l'endroit indiqué par `shift` ($shiftlocation$) pour reconstruire le *bait*. Dans le deuxième exemple, la valeur absolue du Δm correspond à la masse de la séquence EPPNPE, qui peut être ainsi supprimée dans le *hit* pour retrouver le *bait*. Dans la troisième ligne, le Δm peut correspondre à deux séquences peptidiques (VH ou HV). Un tel PSM est classé comme Orange. Pour les PSM des trois dernières lignes, transformer le *hit* en *bait* est plus difficile car soit le Δm est nul, soit l'ambiguïté est trop importante, bien que les séquences soient proches pour la quatrième ligne (VISPEDGK, VIESPDGK). Dans tous les cas, ces PSM sont classés comme Rouges.

même masse. C'est le cas de séquences contenant les mêmes résidus, mais dans un ordre différent (ex : AEAE et EEAA ont la même masse), mais aussi dans des situations plus complexes (ex : KE et GVT ont la même masse). Pour cette raison, des spectres de masse peuvent avoir un SPC important alors que leurs séquences sont éloignées, ce qui donne lieu à des PSM compliqués à interpréter. Il s'agit de l'une des limites de la comparaison des spectres par leur SPC.

En partant de cette constatation, j'ai développé une nouvelle mesure pour caractériser un PSM, le LIPR (*Low Information Peaks Rate*). Il a pour objectif de capturer la pertinence de l'information de séquence contenue dans un PSM et se présente donc comme un critère quantitatif sur la possibilité de l'exploiter pour retrouver la séquence du *bait*. LIPR(*bait*, *hit*) correspond au pourcentage de masses identiques entre *bait* et *hit* qui correspondent à

des séquences différentes (voir Figure 3.10 pour une illustration). Avec la comparaison de spectres théoriques, il est en effet possible d'effectuer précisément cette mesure. Si parmi toutes les masses en commun entre deux spectres, toutes correspondent à des séquences identiques, le LIPR est de 0%. Un LIPR de 100% signifie que toutes les masses partagées entre deux spectres correspondent à des séquences différentes, et donc que le PSM n'est pas exploitable pour déterminer la séquence du *bait*. Ainsi, plus cet indicateur est bas, plus le PSM sera considéré utile à la détermination de la séquence du *bait*, car il sera dépourvu du "bruit" inhérent à la comparaison des spectres par leur SPC.

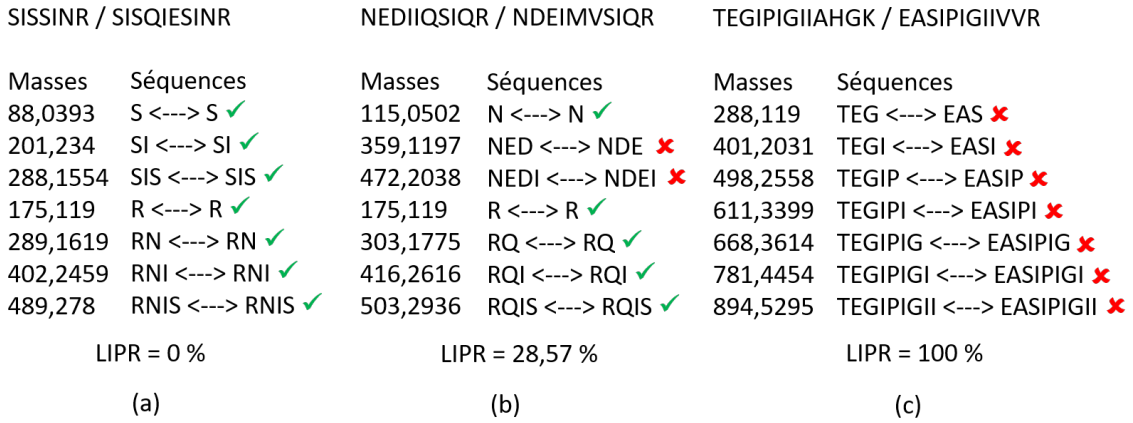


FIGURE 3.10 – **Le LIPR (*Low Information Peaks Rate*) pour 3 PSM différents.** Les exemples sont issus de l'étude sur le protéome humain (PSM_s). Les masses sont en Dalton, et la colonne "Séquences" pour chaque exemple représente les séquences de tous les ions ayant la même masse entre les deux peptides du PSM considéré. Le symbole coche vert indique des séquences identiques et une croix rouge indique des séquences différentes. (a) Aucune masse commune ne correspond à des séquences peptidiques distinctes, donc le LIPR de ce PSM est de 0% ; (b) 2 masses communes sur un total de 7 correspondent à des séquences distinctes, ainsi le LIPR de ce PSM est de $\frac{2}{7}$=28,57% ; (c) toutes les masses communes correspondent à des séquences distinctes, ainsi le LIPR de ce PSM est de 100%.

3.3.4 Application des nouveaux critères et de la complexité des peptides

Couleurs et LIPR

La classification en couleurs et le LIPR sur PSM_r et PSM_s sont présentés Figures 3.11 et 3.12. Les deux stratégies présentent des résultats sensiblement différents. Ceux-ci

TABLE 3.3 – **LIPR moyen et distribution des PSM dans les résultats spécifiques à PSM_r (SS_1) et spécifiques à PSM_s (SS_2) dans les trois catégories de couleurs.**

Ensemble de PSM	#*PSM* Verts	#*PSM* Oranges	#*PSM* Rouges	Total	LIPR (moy %)
SS_1	3 858	6 507	156 926	167 291	61,7
SS_2	62 575	6 793	97 923	167 291	19,44

peuvent être interprétés de deux manières.

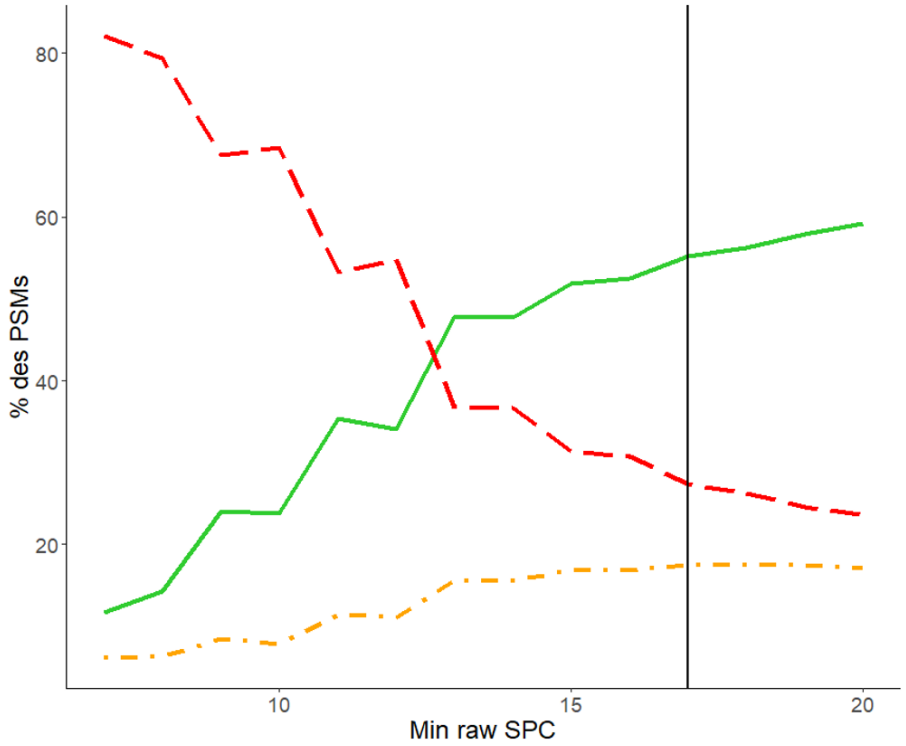
En gardant à l'esprit que le FDR est une mesure qui reste discutable et débattue pour filtrer les résultats des méthodes OMS, nous pouvons d'abord considérer les jeux de résultats sans filtre FDR. Il faut alors comparer l'ensemble des PSM, sans modifier le seuil de score (lié au FDR). Nous ne pouvons en effet pas comparer les résultats pour un certain seuil de raw SPC et de shift SPC, qui sont par définition des scores distincts, mais nous pouvons comparer les jeux de résultats totaux.

Le pourcentage de PSM Verts est plus important pour PSM_s que pour PSM_r (24% vs 11%). Selon ce critère, `strat-shift` renvoie un pourcentage plus important de PSM qui permettent de retrouver la séquence du *bait*, sans ambiguïté. Le LIPR a une valeur moyenne plus haute pour PSM_r (38,5% pour PSM_r et 22,97% pour PSM_s). Les PSM produits par `strat-shift` comportent donc une information de séquence moins bruitée, qui semble plus facilement exploitable.

Afin de mettre en lumière les différences des PSM produits par les deux stratégies, j'ai extrait les PSM qui étaient spécifiques à chacune. Les PSM spécifiques à `strat-raw` (i.e. les PSM de PSM_r que l'on ne retrouve pas dans PSM_s) sont regroupés dans l'ensemble SS_1, et les PSM spécifiques à `strat-shift` sont appelés SS_2.

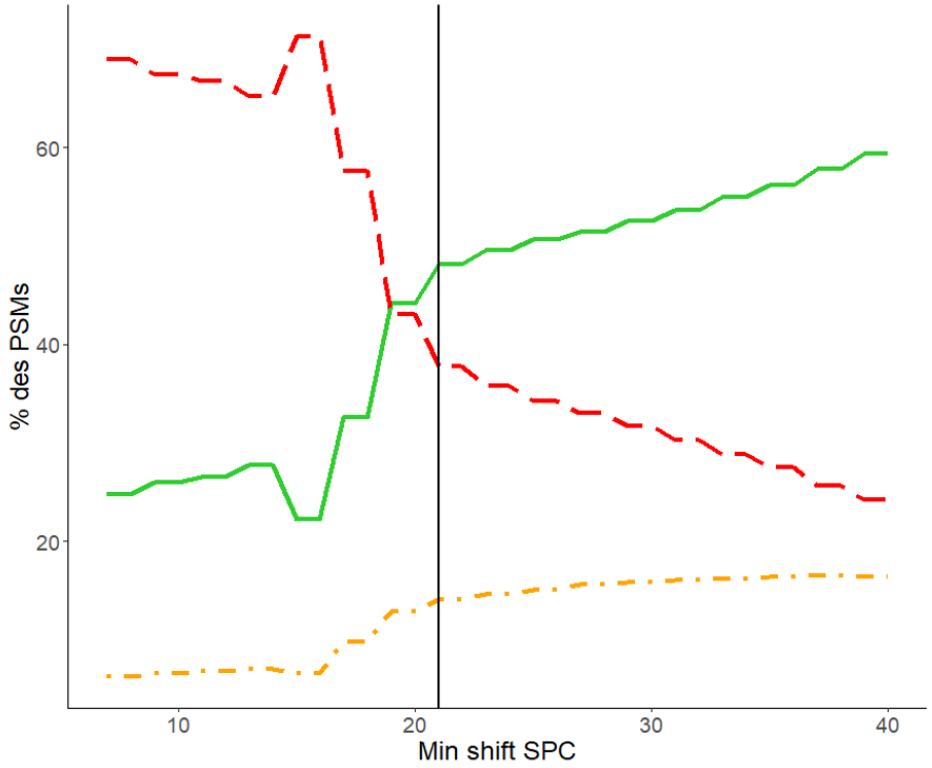
La classification en couleurs et le LIPR ont été appliqués sur les PSM SS_1 et SS_2 et les résultats sont présentés Table 3.3. Ces résultats montrent clairement que `strat-shift` est plus performante. En effet `strat-shift` renvoie bien plus de PSM Verts comparé à `strat-raw` (plus de 16 fois plus). Le LIPR est aussi bien plus élevé dans SS_1 (61,7%) que dans SS_2 (19,44%).

Les deux stratégies semblent se comporter de façon similaire dans l'évolution des indicateurs selon le seuil de SPC. Dans les jeux PSM_r et PSM_s complets, la plupart des PSM sont Rouges, mais en augmentant le seuil de score, la proportion de PSM Verts augmente et le LIPR diminue (voir les courbes des Figures 3.11 et 3.12).



Min raw SPC	#PSMs Verts	#PSMs Oranges	#PSMs Rouges	Total	LIPR (moy %)
7	51 562	28 438	375 404	455 404	38,5
10	29 697	10 160	89 964	129 821	35,79
15	12 852	4 438	8 415	25 705	5,05
17	**9 153**	**3 071**	**5 006**	**17 230**	**3,93**
20	5 460	1 676	2 382	9 518	4,73

FIGURE 3.11 – **Classification en couleurs et LIPR pour `strat-raw`.** Pourcentage de PSM dans les trois catégories de couleurs (Vert : ligne continue, Orange : ligne mixte, Rouge : ligne pointillée) selon le raw SPC (en haut). Nombre de PSM dans les trois catégories de couleurs et LIPR moyen selon le raw SPC minimum (en bas). La ligne en gras correspond au FDR $< 1\%$ (raw SPC $= 17$).



Min shift SPC	*#PSMs* Verts	*#PSMs* Oranges	*#PSMs* Rouges	Total	LIPR (moy %)
7	110 279	28 724	316 401	455 404	22,97
10	110 279	28 724	294 605	433 608	20,72
15	75 330	22 763	249 798	347 891	18,29
20	32 866	9 999	34 028	76 893	4,74
21	**27 211**	**8 334**	**22 773**	**58 318**	**2,53**
25	19 176	6 022	13 955	39 153	1,79
30	13 734	4 375	8 960	27 069	1,62
35	81 08	2 496	4 348	14 952	1,36
40	5 684	1 670	2 505	9 859	1,28

FIGURE 3.12 – **Classification en couleurs et LIPR pour `strat-shift`.** Pourcentage de PSM dans les trois catégories de couleurs (Vert : ligne continue, Orange : ligne mixte, Rouge : ligne pointillée) selon le shift SPC (en haut). Nombre de PSM dans les trois catégories de couleurs et LIPR moyen selon le minimum shift SPC (en bas). La ligne en gras correspond au FDR $< 1\%$ (shift SPC $= 21$).

Cependant une autre façon de considérer les résultats des deux stratégies est de les comparer en prenant le FDR en compte, en ne considérant que les résultats qui passent un certain seuil de FDR.
À un seuil de FDR inférieur à 1%, `strat-shift` produit presque trois fois plus de PSM Verts que `strat-raw` (27 211 vs 9 153). Ainsi ces résultats semblent montrer que `strat-shift` produit davantage de PSM sans détériorer la qualité des résultats.
À ce même seuil, les résultats en termes de LIPR sont comparables pour les deux stratégies avec 3,93% pour PSM_r et 2,53% pour PSM_s.
Par ailleurs, pour chaque stratégie, nous pouvons observer une forte différence de LIPR dans les catégories target et decoy. Dans PSM_r, les PSM target ont un LIPR moyen de 31% alors qu'il est de 49% chez les decoy. Cette différence devient plus nette lorsque le seuil de SPC est plus stringent. Par exemple, dans PSM_r, les PSM decoy avec un raw SPC supérieur ou égal à 15 ont un LIPR moyen de 64%, alors que celui de la base target n'est que de 4% (voir Figure 3.13). Le LIPR peut ainsi être vu comme un critère capturant le caractère aléatoire des PSM decoy qui contiennent davantage de masses identiques qui correspondent à des séquences différentes, en particulier pour les PSM avec de hautes valeurs de SPC.
Nous avons vu qu'en prenant en compte le FDR, `strat-shift` surpasse toujours `strat-raw` en termes de nombre de PSM validés et en termes de nombre de PSM Verts. Cependant, `strat-raw` obtient de meilleurs résultats en termes de pourcentage avec 70,9% des PSM Verts ou Oranges contre 60,9% pour `strat-shift`. Mais le plus haut pourcentage de PSM Rouges doit être remis en perspective avec le LIPR, qui a une valeur plus faible pour ces PSM, ce qui indique une proximité importante entre le *bait* et le *hit*. Le LIPR moyen est en effet de 46,23% pour les PSM Rouges produits par `strat-raw`, et de 31,71% pour les PSM Rouges produits par `strat-shift`. Ainsi, un certain nombre de ces PSM ont probablement un faible nombre d'opérations d'éditions à effectuer pour passer du *hit* au *bait*, comme des permutations de séquences. Une analyse plus en profondeur de la catégorie Rouge devrait donc facilement permettre de faire passer de nombreux PSM Rouges en PSM Oranges, en particulier pour PSM_s.
Nous pouvons enfin noter qu'une proportion importante des PSM du protéome humain sont classés comme Verts lorsque les scores dépassent un seuil de 7 (jeux de résultats totaux PSM_r et PSM_s). Si tous les *hits* associés à ces PSM Verts venaient de la base target, cela serait une bonne nouvelle pour la pertinence de la stratégie target/decoy, car ces PSM impliquent des peptides qui diffèrent par des opérations d'éditions non ambiguës

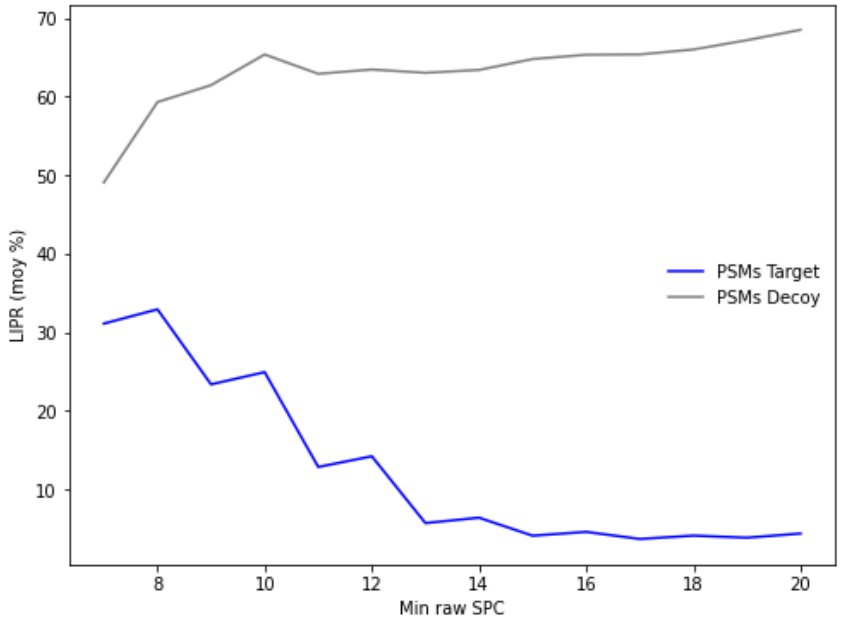


FIGURE 3.13 – **Évolution du LIPR selon l'augmentation du seuil de SPC pour PSM_r.** La courbe en bleu montre que le LIPR des PSM target est bien moins élevé que celui des PSM decoy (en gris) ; d'autre part, cette différence augmente avec le seuil de SPC.

de leur candidat le plus proche de la base de données (de 11% dans `strat-raw` à 24% dans `strat-shift`). Étant donné que ce n'est pas le cas, cela signifie que la base decoy n'est pas seulement composée de séquences "incorrectes" comme elle devrait l'être. Ainsi il est clair que pour les deux stratégies, la présence de nombreux PSM Verts dans la base decoy entrave l'identification des *baits*.
D'un autre côté, dans un contexte où nous pouvons identifier et localiser certaines modifications (comme ici avec `shift`), nous voyons que l'ajout de peptides à la base de données avant comparaison, même issus de la base decoy, peut augmenter le nombre de spectres identifiés en général.

Application des observations du réseau des peptides aux résultats

Le réseau observé dans la Section 3.2 suggérait l'existence d'un lien entre les PSM decoy et la complexité en résidus des peptides. Afin de voir si un tel lien existe dans les résultats des deux stratégies, cette complexité a été formalisée afin de pouvoir la calculer pour un PSM donné, et donc sur tous les PSM à grande échelle.
Les peptides observés dans certaines zones du réseau (voir Figure 3.8, Section 3.2.2) sont

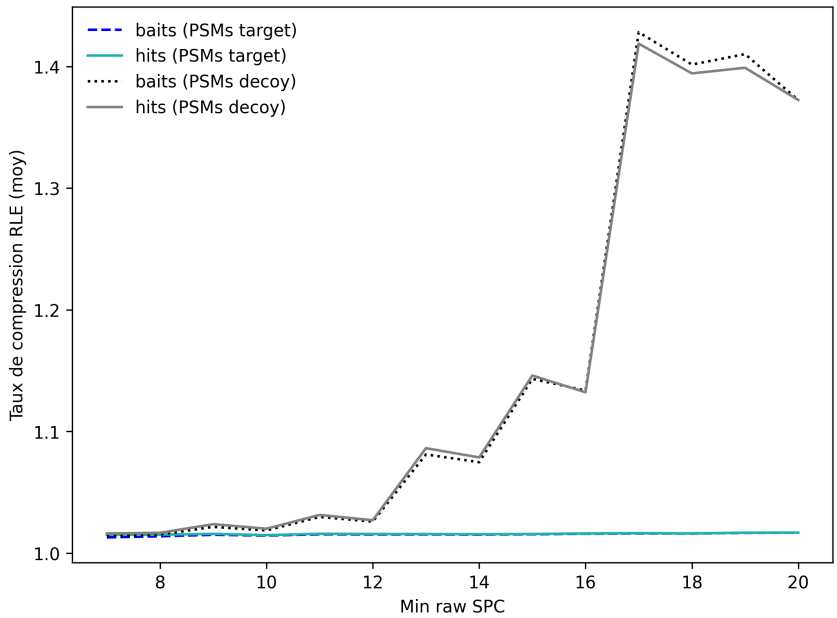


FIGURE 3.14 – **Taux de compression RLE des *baits* et *hits* pour les PSM target ou decoy (PSM_r).** Pour chaque seuil de raw SPC, le taux de compression RLE est calculé pour les *baits* et *hits* impliqués dans les PSM target ou decoy.

composés de successions de résidus identiques. Afin de capturer cette particularité, nous avons procédé à une compression de type RLE (*Run-Length Encoding*) sur les peptides. Pour cette compression, une succession d'un même résidu sera représentée par un chiffre indiquant la longueur de cette succession, immédiatement suivie de ce résidu (le nombre n'est pas indiqué si la succession est de taille 1). Par exemple, le peptide "GGGGGWGGAAR" devient "5GW2G2AR". Il est ensuite possible de calculer le taux de compression RLE. Nous divisons pour cela la taille du peptide par sa taille compressée. Dans notre exemple, le peptide est de taille 11 et sa version compressée est de taille 8. Le taux de compression RLE sera de $11/8 = 1,375$. Ainsi, plus un peptide est composé de longues successions de résidus identiques, plus ce taux sera élevé.

Afin de mettre en évidence cette caractéristique pour les peptides target et decoy, le taux de compression RLE a été calculé pour les *baits* et *hits* issus des PSM target et decoy en faisant évoluer le seuil de raw SPC pour PSM_r (Figure 3.14).

D'abord, si nous considérons séparément les PSM target d'un côté et decoy de l'autre, le taux de compression RLE des *baits* et des *hits* restent très proches avec l'augmentation du seuil de score. Cependant, nous observons une forte différence entre les catégories PSM target et decoy ; plus le seuil de score est élevé, plus la différence de taux de compression

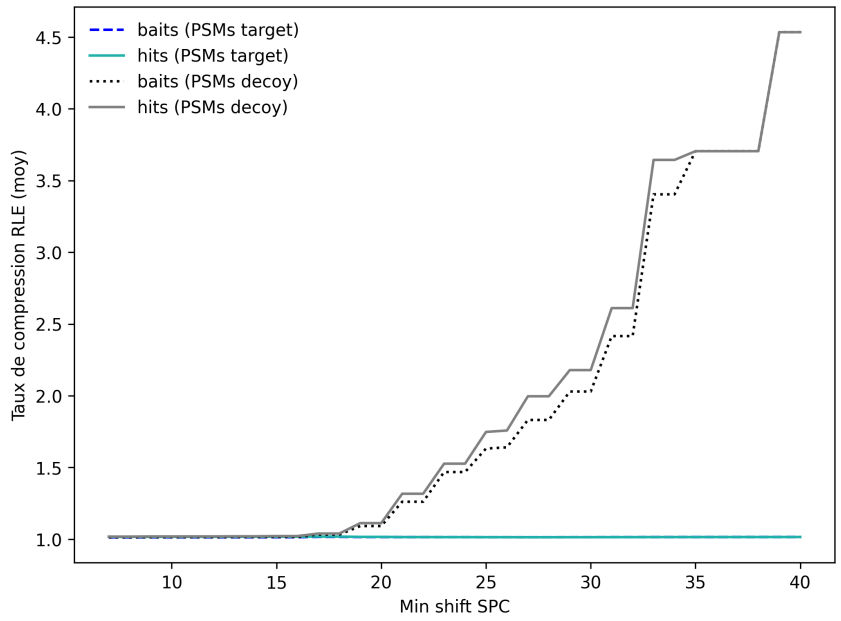


FIGURE 3.15 – **Taux de compression RLE des *baits* et *hits* pour les PSM target ou decoy (PSM_s).** Pour chaque seuil de shift SPC, le taux de compression RLE est calculé pour les *baits* et *hits* impliqués dans les PSM target ou decoy.

RLE augmente entre les peptides. Ce taux reste très proche de 1 pour les peptides impliqués dans les PSM target. La même tendance peut être observée dans les identifications PSM_s (Figure 3.15).

Une autre manière de concevoir la complexité d'un peptide en termes de séquence est de calculer sa diversité en résidus. En effet un peptide peut être incompressible, mais ne comporter que deux résidus, comme le peptide (fictif) AEAEAEAEAEAE. Afin de capturer cette particularité, pour chaque peptide, sa diversité a été calculée en divisant le nombre de résidus différents par sa longueur. Dans notre exemple, le peptide est de longueur douze mais ne comprend que deux résidus ; par conséquent, sa diversité sera de $2/12 = 0,16$. Une diversité de 1 correspondra donc à un peptide très diversifié, où chaque résidu est de nature différente.

Dans le même esprit que pour la compression RLE, ce calcul a été fait sur les *baits* et les *hits* des PSM target et decoy selon le seuil (PSM_r). Les résultats sont présentés Figure 3.16.

Les diversités en résidus des *baits* et *hits* lorsque les PSM sont séparés selon l'origine du *hit* restent très proches selon le seuil de raw SPC.

La diversité des peptides diminue pour les deux catégories, mais celle des peptides des

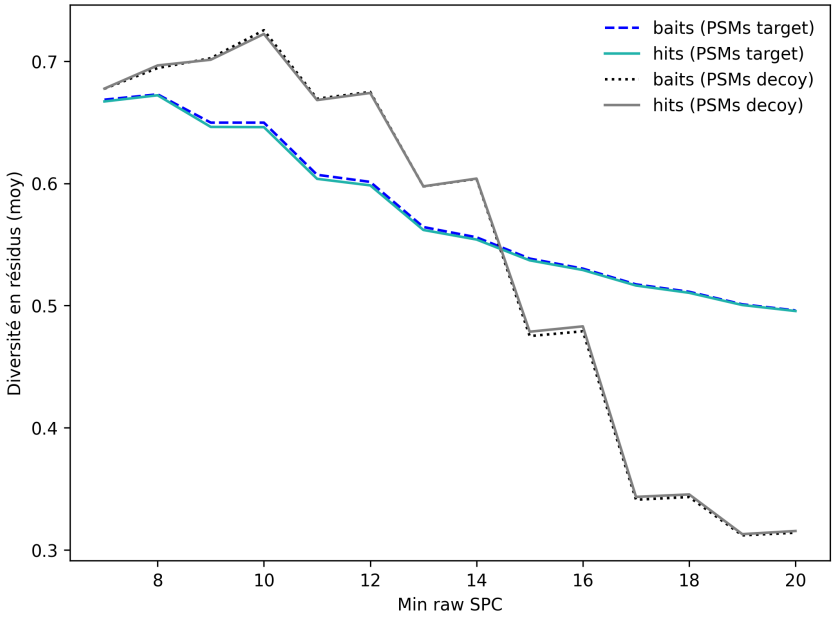


FIGURE 3.16 – **Diversité en résidus des *baits* et *hits* pour les PSM target ou decoy (PSM_r).** Pour chaque seuil de raw SPC, la diversité en résidus est calculée pour les *baits* et *hits* impliqués dans les PSM target ou decoy.

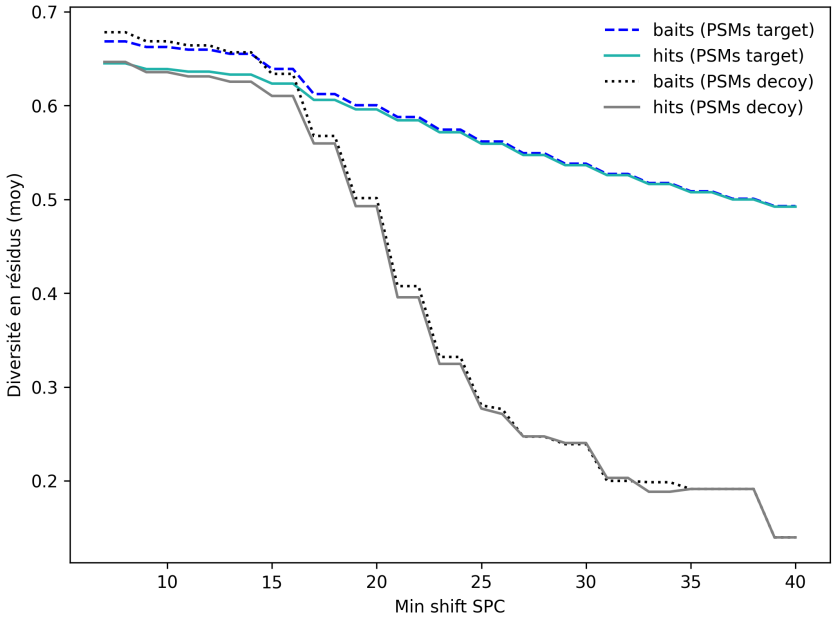


FIGURE 3.17 – **Diversité en résidus des *baits* et *hits* pour les PSM target ou decoy (PSM_s).** Pour chaque seuil de shift SPC, la diversité en résidus est calculée pour les *baits* et *hits* impliqués dans les PSM target ou decoy.

PSM decoy chute plus rapidement que celle des PSM target. La même observation peut être faite sur le jeu PSM_s (Figure 3.17).
Ces deux observations menées par le taux de compression RLE et la diversité en résidus laissent supposer qu'un lien existe entre l'origine du peptide (target ou decoy) et la complexité de la séquence d'un peptide en termes de résidus. Or les couleurs et le LIPR appliqués sur les PSM ont montré que les PSM target sont plus riches d'information que les PSM decoy. Les peptides de faible complexité semblent donc moins intéressants en termes d'information apportée dans un PSM, ce qui est cohérent avec les zones de faible densité observées dans le réseau des peptides, dans lesquelles nous supposions la présence de PSM de moindre qualité à cause d'une proportion plus importante de peptides decoy, et dans lesquelles nous avions observé des peptides peu complexes. De plus, de par leur distribution déséquilibrée dans les catégories target et decoy, ces peptides pourraient altérer le calcul du FDR. Cependant, une nuance peut être apportée, car même si le PSM identifie un peptide decoy peu complexe, ce qui donne naissance à un PSM Rouge avec un LIPR non nul, le PSM peut être informatif dans le sens où il est possible d'inférer que le peptide qui se cache derrière le spectre est composé majoritairement du même résidu que le peptide identifié.

3.4 Conclusion

Le réseau des peptides

Les spectres théoriques, et les peptides correspondants, peuvent être associés dans un réseau où les peptides sont des nœuds réunis par des arêtes si les deux peptides partagent un SPC supérieur à un seuil donné ; on obtient alors un réseau de peptides réunis par la MS.
Puisque les séquences correspondantes aux spectres sont connues, nous pouvons observer si un lien existe entre ces séquences, l'origine des peptides et la connectivité du réseau. Nous avons pu supposer l'existence d'une corrélation entre la densité du réseau, l'origine des peptides et leur complexité en termes de résidus. En effet, des zones moins denses du graphe semblent corrélées avec des peptides originaires de la base decoy avec une complexité en résidus plus faible. Ces observations pourraient aider à améliorer la qualité des identifications si on construit un tel réseau avec des spectres expérimentaux.

La meilleure stratégie pour sélectionner les PSM

Associer le Δm - indicateur des modifications entre deux spectres - et le réalignement de pics à la sélection des PSM pourrait paraître forcément intéressant pour la sélection de PSM modifiés par rapport à une méthode qui ne réalise pas cette association. Cependant, intégrer le Δm, par exemple en décalant tous les pics de Δm, élargit l'espace de recherche, ce qui risque de provoquer la sélection de PSM erronés. Ainsi, les résultats exposés dans ce chapitre n'étaient pas facilement prévisibles.
Les critères que le contexte théorique a permis de créer montrent que `strat-shift` est plus performante que `strat-raw`, en produisant des PSM qui permettent plus souvent de retrouver la séquence du *bait*, et en plus grand nombre pour un seuil de FDR équivalent. Cette stratégie devrait donc être privilégiée dans les méthodes OMS, à partir du moment où un outil de réalignement efficace est disponible afin de garder un temps de calcul raisonnable.

Les perspectives de l'environnement de spectres théoriques

Cette étude a été conduite dans un environnement "strict", avec des spectres théoriques comparés sur la base du SPC, qui représente seulement les masses communes entre deux spectres. Le SPC étant un élément important de scores de comparaison plus sophistiqués, les résultats de cette étude devraient pouvoir être applicables aux méthodes OMS qui l'utilisent. De plus, les identifications entre spectres théoriques offrent une cartographie de la proximité des peptides d'un point de vue MS en revenant aux fondamentaux des méthodes d'identification, qui se basent sur le SPC pour identifier les spectres.
Le protocole que nous avons mis en œuvre pour comparer ces deux stratégies OMS avec des peptides théoriques et de nouveaux indicateurs a permis de créer un environnement qui a fourni des éléments de réponse à la question de départ. Puisque les identifications sont complètement contrôlées, ce contexte théorique pourrait permettre de tester d'autres outils OMS afin de déterminer leurs forces et leurs faiblesses, dans le but de faire une preuve de concept, ou encore de les configurer ou les calibrer. Ces évaluations pourraient être faites avec les indicateurs présentés dans ce chapitre, mais offrent la possibilité d'en développer et en utiliser d'autres. Plusieurs exemples de tels indicateurs sont présentés dans le chapitre suivant ; nous pourrons notamment voir que la classification en couleurs peut être affinée.

Les PSM Rouges, des identifications d'intérêt

Nous avons vu que `strat-shift` recrute de nombreux PSM Rouges. Ainsi, même si cette stratégie est meilleure que `strat-raw`, de nombreux PSM ne permettent pas la détermination de la séquence du *bait*. Cependant, le LIPR plus faible pour les PSM Rouges de `strat-shift` comparé à `strat-raw` laisse penser qu'ils contiennent de l'information pertinente permettant de les transformer en PSM Oranges, voire Verts, avec davantage d'efforts de calcul. Parmi les PSM Rouges, il y a par exemple des PSM qui sont classés comme tels parce que deux modifications ou plus séparent le *hit* du *bait*. Le Δm correspond alors à la somme de plusieurs modifications, et doit donc être fragmenté pour pouvoir interpréter le PSM.

Cette problématique des modifications multiples, connue dans un contexte expérimental, semble aussi être au cœur de l'interprétation d'un certain nombre des PSM produits dans le contexte des spectres théoriques. Ainsi, ces derniers ont permis de travailler sur le développement et l'évaluation d'un algorithme ayant pour objectif l'interprétation des PSM comportant *plusieurs* modifications. Ce travail est décrit dans le chapitre suivant.

Chapitre 4

L'identification de modifications multiples

Sommaire

Préambule : L'étude présentée au Chapitre 3 a montré que seuls 21% de PSM sont interprétés lorsque le Δm est considéré de façon indivisible pour réaligner les spectres. Cela montre qu'un certain nombre de PSM comportent plusieurs modifications qui séparent les peptides du *bait* et du *hit*. Dans ce chapitre, je parle de `SpecGlob`, un algorithme que nous avons développé. Il repose sur l'alignement des spectres de masse d'un PSM donné (renvoyé par n'importe quelle méthode OMS) afin de localiser les modifications qui séparent les deux peptides, sans *a priori* sur leur nombre ou leur nature. `SpecGlob` a été utilisé sur des spectres théoriques afin d'évaluer ses performances. En effet, une fois l'alignement de `SpecGlob` réalisé, une étape supplémentaire peut être réalisée afin d'évaluer cet alignement. Enfin je discute des perspectives en lien avec `SpecGlob` et pour lesquelles de premiers résultats sont disponibles. Une partie du travail exposé dans ce chapitre a été valorisée par un poster présenté à la convention annuelle de la société américaine de spectrométrie de masse (ASMS) en 2021, et a fait l'objet d'un article déposé sur BioRXiv [Lysiak *et al.* 2022].

4.1 Motivations et objectifs

L'étude des PSM produits par `SpecOMS` à partir de spectres théoriques (voir Chapitre 3, Section 3.3.4, page 86) a permis de mettre en évidence que seuls 21% des PSM étaient interprétables sans ambiguïté par `shift`, et donc en considérant le Δm comme une seule modification. Les PSM Oranges contiennent eux aussi une seule modification, mais soulèvent un problème en termes de combinatoire de séquences possibles pour interpréter le PSM. Les PSM Rouges représentent donc une catégorie qui pourrait être étudiée davantage afin d'améliorer le taux d'interprétation des PSM en général. Il est raisonnable de supposer que parmi eux, de nombreux PSM contiennent plusieurs modifications, c'est-à-dire que le Δm doit être séparé en plusieurs parties afin d'être correctement interprété.
Des outils pour l'identification de modifications multiples ont déjà été développés. `PTMShepherd` [Geiszler *et al.* 2021] réalise un histogramme des Δm des PSM, puis annote les pics les plus intenses de cet histogramme grâce aux PTMs connues dans Unimod [Creasy et Cottrell 2004] et à celles fournies par l'utilisateur. `PTMiner` [An *et al.* 2019] regroupe

les PSM selon le Δm et tente également de les identifier par Unimod.
`Modificomb` [Savitski, Nielsen et Zubarev 2006] se base sur un histogramme de Δm pour repérer les PTMs et combinaisons de PTMs connues. `Metamorpheus` [Solntsev *et al.* 2018] recherche d'abord une base de peptides en mode OMS ; ensuite, les Δm répertoriés sur Unimod sont placés sur les protéines, et enfin une recherche avec tolérance de masse restreinte peut être faite contre cette nouvelle base de protéines. `MODPlus` [Na, Kim et Paek 2019] se base aussi sur les modifications d'Unimod pour aligner deux spectres.
Les outils existants identifient donc des modifications fréquentes et/ou connues. Or, étant donné leur importance biologique potentielle, il est nécessaire d'être capable d'interpréter les modifications rares, et de ce fait non répertoriées. `SpecOMS`, puisqu'il est capable de comparer tous les spectres les uns aux autres, met en évidence des PSM originaux. Dans le même esprit, pour être capable de mettre en évidence des modifications à la fois sans *a priori* et multiples, nous avons développé `SpecGlob`, un algorithme qui repose sur l'alignement d'un peptide sur le spectre à identifier, le peptide et le spectre appartenant à un PSM (lequel peut être fourni par n'importe quelle méthode OMS) pour mettre en évidence les modifications qui les séparent ; si ces modifications sont correctement identifiées dans le peptide, nous pouvons retrouver la séquence du peptide qui a généré le spectre.
L'objectif de `SpecGlob` est de retrouver la séquence du peptide qui a produit le spectre ; pour obtenir ce résultat, il faut savoir comment transformer le peptide candidat - dont la séquence est connue - afin de retrouver la séquence peptidique du spectre expérimental. Le résultat recherché est donc la séquence du peptide enrichie des modifications placées correctement sur ses résidus. Dans ce résultat, la somme des masses des modifications devra correspondre au Δm, séparée en plusieurs masses si plusieurs modifications séparent le peptide identifié et le peptide qui a produit le spectre à identifier.
Prenons pour exemple un PSM qui contient un spectre à identifier de séquence DYSIR (le *bait* dans l'étude entre spectres théoriques) et un peptide associé (le *hit*) de séquence DWYIR. Deux modifications de séquence séparent le *bait* et le *hit*. Le résultat souhaité est de retrouver ces deux modifications grâce à un alignement des deux spectres, c'est-à-dire l'indication qu'il faut, partant du *hit* DWYIR, d'une part supprimer W et d'autre part ajouter S avant I, ce qui permet de retrouver la séquence DYSIR associée au *bait*.

4.2 Description de `SpecGlob`

Pour présenter `SpecGlob`, je décris d'abord le principe de l'algorithme. Ensuite, je détaille son fonctionnement sur un exemple. Je fournis également le pseudocode de `SpecGlob` et une description formelle de l'algorithme. Enfin cette section est conclue par quelques exemples de résultats de `SpecGlob`. Les *baits* et *hits* permettent en effet de détailler des exemples dont les séquences sont connues.

4.2.1 Principe de l'algorithme

Alignement des résidus du *hit* sur le *bait*

Étant donné que le PSM est issu du résultat d'une recherche OMS, les peptides du *bait* et du *hit* partagent probablement une certaine similarité en termes de résidus en commun. L'idée est donc d'essayer d'aligner les résidus du *hit* (représentés par une différence de masse entre deux masses successives du spectre) sur ceux du *bait* (représentés par une différence au sein d'un couple de masses donné dans le spectre) en autorisant des décalages de masse, et donc de découvrir quels décalages de masse éventuels doivent être faits pour un alignement optimal. Ces décalages de masse correspondront aux modifications éventuelles à réaliser dans le *hit* pour obtenir la séquence du *bait*.
`SpecGlob` prend en entrée un PSM, plus précisément les masses des ions du *hit* $\{h_0,...,h_i,h_{i+1},...,h_{N-1}\}$ et les masses des ions du *bait* $\{\beta_0,...,\beta_j,\beta_{j+1},...,\beta_{M-1}\}$. h_0 et β_0 représentent l'ion H^+. `SpecGlob` considère itérativement chaque différence $h_i - h_{i-1}$ pour tout $1 \leq i \leq N-1$, qui représente la masse des résidus du *hit*. Pour chacune, il essaie de trouver la même différence de masses (le même résidu) dans le *bait*, en autorisant si nécessaire un décalage de masse pour aligner la paire de masses (aligner les résidus). Un résidu est considéré retrouvé s'il existe $0 \leq k < j \leq M-1$ tel que $\beta_j - \beta_k = h_i - h_{i-1}$. Ainsi, à la fin de l'alignement, le Δm est découpé si nécessaire en plusieurs décalages de masses non définis à l'avance, que ce soit concernant leur nombre ou leur valeur. `SpecGlob` fournit ensuite le meilleur alignement sous la forme d'une chaîne de caractères appelée ***hitModified***, qui donne les indications nécessaires pour aligner le *hit* sur le *bait*, et donc transformer le *hit* pour déterminer la séquence du *bait*. Le résultat est nommé ainsi car il correspond à la séquence du *hit* avec des indications spécifiques qui marquent les modifications à effectuer dans le *hit* pour obtenir le *bait*.
Pour chaque PSM, le $hitModified$ spécifie l'alignement de chaque résidu du *hit*. Il existe trois possibilités. (1) Deux masses consécutives du *hit* (qui correspondent à la masse d'un

résidu) sont alignées avec deux masses du *bait* sans avoir besoin d'insérer un décalage de masse ; dans ce cas ce résidu est considéré comme retrouvé dans le *bait* et est indiqué tel quel dans le $hitModified$; (2) la différence entre deux masses consécutives du *hit* est trouvée entre deux masses du *bait*, mais l'alignement de ces masses requiert l'insertion d'un décalage de masse ; dans ce cas, le résidu du *hit* est écrit dans le $hitModified$, précédé de la valeur (en Da) du décalage de masse entre crochets ; (3) enfin, si la différence de masse entre deux masses consécutives du *hit* n'est pas utilisée dans l'alignement, cela signifie que le résidu courant est considéré comme absent, et celui-ci est alors écrit dans le $hitModified$ entre crochets. La somme des décalages de masse du $hitModified$ est égale à Δm. Un exemple de $hitModified$ est donné Section 4.2.2.

Utilisation de la programmation dynamique

Puisqu'on ne souhaite pas de contrainte sur les modifications, le problème semble difficile ; il revient en effet à tester toutes les possibilités de décalages entre tous les couples de masses (représentant les résidus) afin de trouver le meilleur alignement du *hit* sur le *bait*. Néanmoins, à cause du nombre de PSM qu'une analyse MS est susceptible de fournir, l'algorithme doit être très rapide.

Le test de tous les décalages entre toutes les masses est un problème qu'il est intéressant de résoudre par programmation dynamique. En effet, en combinant les sous-solutions optimales du problème, il est possible de trouver une solution optimale au problème d'alignement des masses en autorisant des décalages. Des outils déjà cités dans le Chapitre 2 reposent sur la programmation dynamique pour produire les PSM, mais ici la programmation dynamique ne sera pas utilisée pour produire les PSM, mais pour *interpréter un PSM donné*.

Les paramètres d'entrée de l'algorithme sont les suivants :

- **deux listes de masses** qui correspondent aux masses des deux spectres d'un PSM. Seuls les N ions b du *hit* sont donnés en entrée à `SpecGlob`, puisqu'ils sont suffisants pour représenter la séquence en résidus du peptide que l'on souhaite aligner sur le *bait*. Cependant, puisqu'on souhaite que *bait* se rapproche le plus possible d'un spectre expérimental (dont le type des ions est inconnu), celui-ci est représenté par les M masses des ions b mais aussi y ;
- **une matrice de scores** D qui évalue les différentes solutions d'alignement du *hit* sur le *bait*. D est de taille $N \times M$. Chaque case $D[i][j]$ de la matrice représente le score de l'alignement de la $i^{ème}$ masse du *hit* sur la $j^{ème}$ masse du *bait*. Selon le

principe de la programmation dynamique, un choix est fait à chaque étape (chaque case de D) selon les choix qui ont été faits aux étapes précédentes (les cases remplies précédemment) afin de résoudre le problème dans sa globalité. Ainsi, la matrice D est remplie de gauche à droite et de haut en bas, et pour chaque case $D[i][j]$, les différentes façons d'aligner le résidu i du *hit* sur la masse j du *bait* sont évaluées et la plus favorable est choisie selon le système de scores (discuté ci-après) et les cases de D déjà remplies jusqu'à i et jusqu'à j. Le meilleur score de la dernière ligne de la matrice D représente donc l'alignement optimal des masses du *hit* sur les masses du *bait* ;

— **une matrice** $Origin$. La case $Origin[i][j]$ permet de connaître la case d'origine de $D[i][j]$ (la case dont la valeur a servi pour remplir la case courante) ainsi que le type d'alignement réalisé, ce qui sera nécessaire pour obtenir le résultat final par une étape de **traceback**, qui consiste à remonter dans la matrice pour produire le résultat final ;

— pour trouver le meilleur alignement, `SpecGlob` cherche à maximiser un score qui prend en compte le nombre de résidus du *hit* alignés sur le *bait*, le nombre de décalages de masses insérés, et le nombre de résidus qui n'ont pas été retrouvés au cours de l'alignement. Il prend donc aussi en entrée un **système de scores**, c'est-à-dire un ensemble de trois valeurs (s_A, s_R, s_N) affectées aux différentes possibilités : 1) résidus alignés sans décalage de masse (s_A pour $Align$) ; 2) résidus alignés avec un décalage de masse (s_R pour $Realign$) ; 3) résidu du *hit* non présent dans le *bait* (s_N pour $No-align$). Ces trois valeurs reflètent les priorités que l'on souhaite avoir pour ces trois situations lors de l'alignement. Ainsi, pour remplir chaque case $D[i][j]$, l'algorithme détermine d'abord à laquelle des trois situations elle correspond par rapport à chaque case précédemment remplie, comme détaillé dans le reste de cette section. Selon la situation, le score correspondant est ajouté à la valeur de la case déjà remplie. Enfin, le maximum de ces trois possibilités est sélectionné et devient la valeur de la case à remplir. Par ce moyen, la programmation dynamique permet de faire, à chaque étape, un compromis entre une situation que l'on considère comme favorable ou non (reflété par le système de scores) et l'alignement déjà effectué (représenté par l'état de la matrice D de i à j).

4.2.2 Exemple détaillé d'un alignement réalisé par SpecGlob

Le fonctionnement de SpecGlob est illustré ici avec l'exemple d'un PSM qui a pour *bait* DYSIR et pour *hit* DWYIR, avec Δm = -99,04 Da.
Pour l'exemple et les travaux présentés dans ce chapitre, le système de scores est le suivant : $s_A = 5$, $s_R = 2$ et $s_N = -4$. La priorité est donc de retrouver un résidu d'abord sans décalage de masse, puis avec un décalage de masse, et enfin le fait de ne pas retrouver le résidu est pénalisé. L'algorithme a été développé de façon à exploiter la similarité en termes de résidus des peptides correspondant aux spectres alignés ; ainsi régler le système de scores de cette façon permet de favoriser le nombre de résidus alignés, et donc un alignement riche en informations. La stabilité du système de scores est discutée à la fin de la Section 4.4.2.
La première étape de l'alignement est d'initialiser la matrice D, comme illustré Figure 4.1. La première ligne de D est mise à 0 car h_0 est toujours la masse de l'ion H^+. Pour chaque $0 \leq i \leq N-1$ de la première colonne, $D[i][0]$ est mis à $i \times s_N$, ce qui permet de marquer les résidus comme non trouvés lorsque l'alignement démarre.
Ensuite, D est remplie, de gauche à droite et de haut en bas à partir de la case $D[1][1]$. Pour chaque case $D[i][j]$, si le $i^{ème}$ résidu du *hit* (dont la masse est égale à $h_i - h_{i-1}$) n'est pas retrouvé dans le *bait* (aucune différence entre β_j et une masse plus petite ne correspond à la masse du résidu), SpecGlob choisit pour origine la case $D[i-1][j]$, et la valeur de cette case, à laquelle est ajouté s_N, devient la valeur de la case $D[i][j]$ à remplir. Si le résidu est retrouvé dans le *bait* avec β_j, on met en compétition deux situations : 1) le résidu est retrouvé sans décalage de masse ; 2) le résidu est retrouvé avec un décalage de masse. On peut en effet calculer le décalage de masse entre la case à remplir et une case d'origine possible $D[i-1][k]$ en comparant les masses entre les deux cases en i (au niveau du *hit*, $h_i - h_{i-1}$) et en j (au niveau du *bait*, $\beta_j - \beta_k$). Si ces deux masses sont identiques, cela signifie que l'on n'a pas de décalage de masse. Sinon, il y a un décalage de masse. Une fois la matrice D remplie, le meilleur alignement est déterminé à partir du score le plus élevé de la dernière ligne ainsi que l'étape de traceback à l'aide de la matrice $Origin$.
Les Figures 4.2 à 4.5 montrent et expliquent les étapes représentatives du remplissage de la matrice D sur notre exemple. Les cases impliquées dans le meilleur alignement (à partir duquel le $hitModified$ est produit) sont mises en valeur, avec un exemple de résidu retrouvé sans décalage de masse (Figure 4.2, première ligne), un exemple de résidu non retrouvé (Figure 4.3, troisième ligne), un exemple de résidu retrouvé avec un décalage de

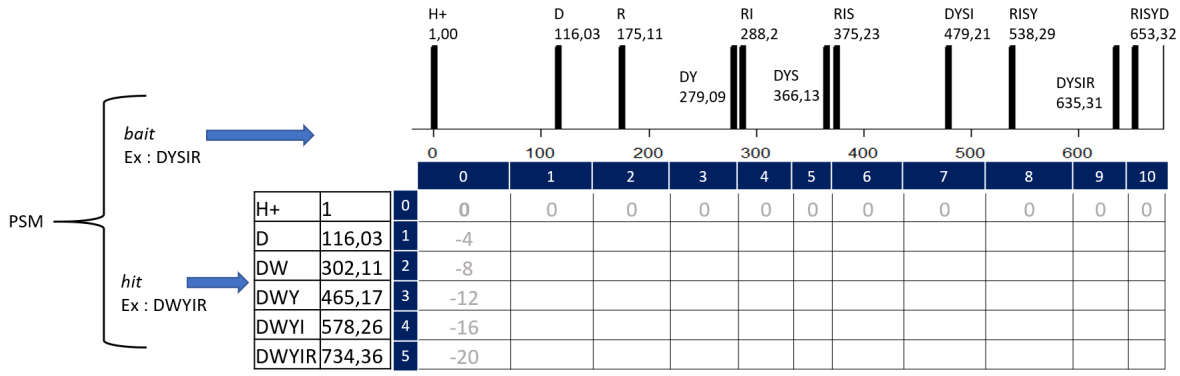


FIGURE 4.1 – **Initialisation de la matrice D.** La première ligne est initialisée à 0, et la première colonne à $s_N \times i$, i étant l'indice de la ligne et s_N le score correspondant à la situation dans laquelle on ne retrouve pas le résidu (ici, $s_N = 4$). Les masses indiquées sont en Da.

masse (Figure 4.4, quatrième ligne), et enfin la matrice complètement remplie (Figure 4.5). Ces figures illustrent la manière dont chaque case est remplie selon les différents cas.
Pour remplir chaque case $D[1][j]$ de la première ligne (Figure 4.2), `SpecGlob` détermine si la masse du résidu du *hit* qui correspond à la ligne (ici le résidu D, de masse 115,03 Da) peut être retrouvée en soustrayant à β_j une masse plus petite. Ici, c'est le cas pour deux cases. La masse de D est retrouvée en soustrayant β_1 (116,03 Da) à β_0 (1 Da). Elle est également retrouvée en soustrayant β_{10} (653,32) à β_8 (538,29). Pour la case $D[1][1]$, la seule possibilité est de venir de la case $D[0][0]$; nous n'avons pas de décalage entre les deux cases (115,03 est ajouté dans le *hit* comme dans le *bait*) ainsi `SpecGlob` ajoute s_A (de valeur 5) à $D[0][0]$ (de valeur 0). $D[1][1]$ contient donc un 5 (en bleu). Pour la case $D[1][10]$, l'algorithme considère deux possibilités d'alignements. La première est de venir de la case $D[0][8]$, et donc de retrouver D sans décalage de masse. On ajoute donc s_A (5) à la valeur de cette case (0). La valeur correspondant à cette première possibilité, 5, est conservée. La seconde possibilité est d'insérer un décalage de masse. Pour cela, `SpecGlob` considère toutes les cases de la ligne précédente d'un indice inférieur à 8 (indice de la masse qui, soustraite à la β_{10}, a permis de retrouver la masse de D). À chaque case de la ligne précédente correspond un décalage de masse possible, et une valeur (ici, toutes égales à 0) à laquelle est ajouté s_R (de valeur 2). La première possibilité de valeur 5 est donc considérée comme la plus favorable par rapport à toutes les autres cases qui impliquent l'insertion d'un décalage de masse. La case $D[1][10]$ prend donc la valeur de 5. Toutes les autres cases de la première ligne ne permettent pas de retrouver le résidu D ; ainsi,

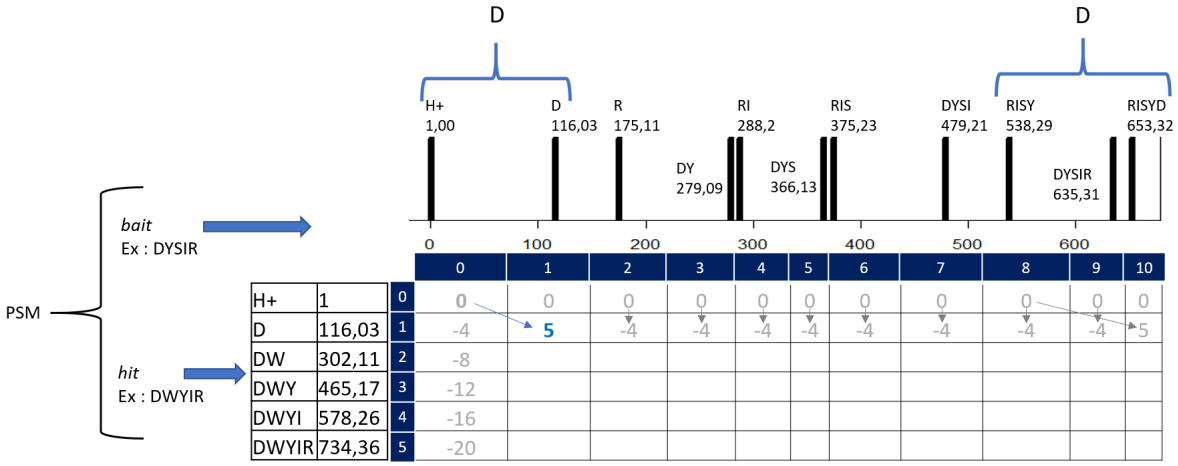


FIGURE 4.2 – **Remplissage de la deuxième ligne de la matrice** D**.** On retrouve le résidu D à deux endroits sans décalage de masse. Les masses indiquées sont en Da.

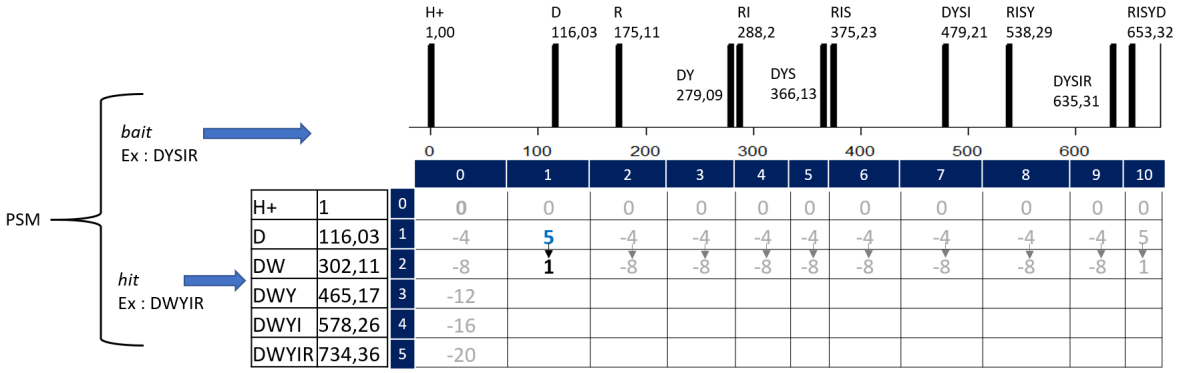


FIGURE 4.3 – **Remplissage de la troisième ligne de la matrice** D**.** Le résidu W recherché n'est pas retrouvé. Les masses indiquées sont en Da.

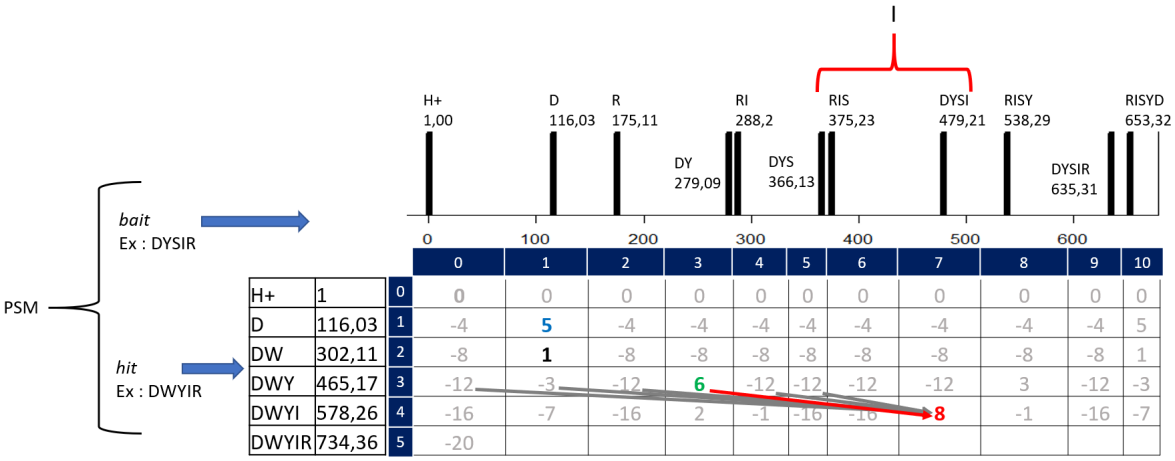


FIGURE 4.4 – **Remplissage de la cinquième ligne de la matrice D.** Le résidu I est recherché et retrouvé ; toutes les possibilités de retrouver I, sans décalage de masse (l'origine est la case $D[3][5]$) ou avec un décalage de masse (l'origine est une autre case avec $j < 5$), sont indiquées par des flèches. La flèche rouge indique I avec le décalage de masse le plus avantageux. Les masses indiquées sont en Da.

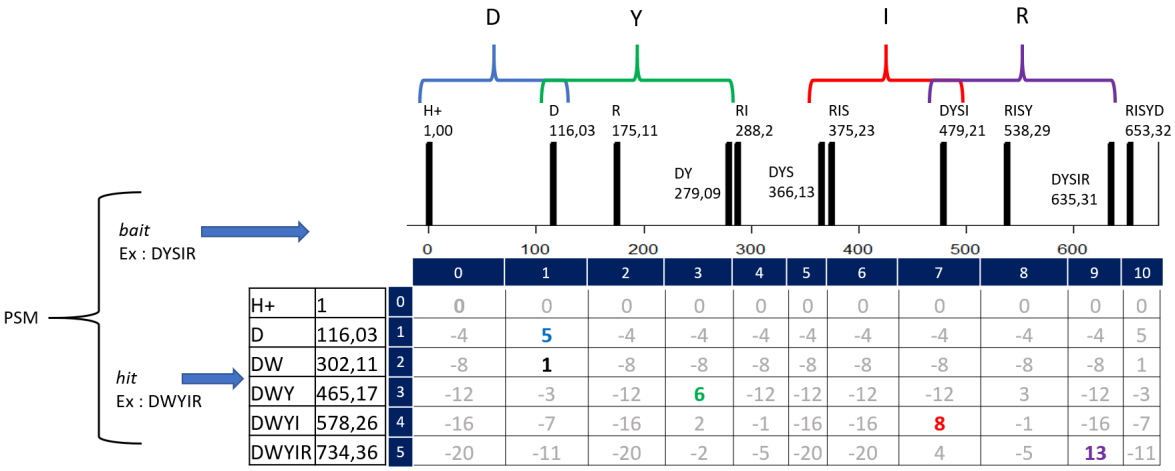
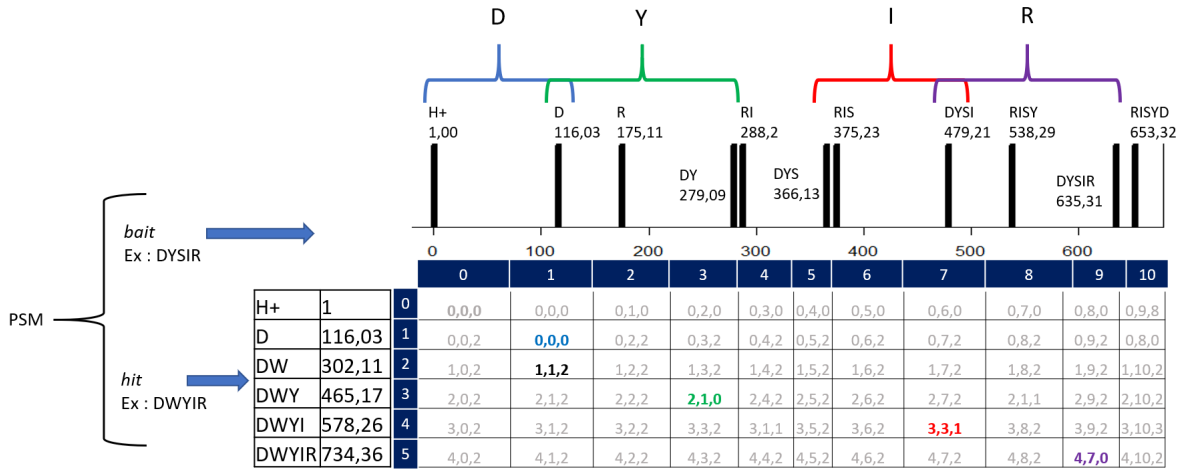


FIGURE 4.5 – **Matrice D entièrement remplie.** Le meilleur alignement a un score de 13 (en violet). Les masses indiquées sont en Da.

l'origine est la case du dessus, de valeur 0, à laquelle est ajouté s_N, de valeur -4.
Pour chaque case de la seconde ligne (Figure 4.3), W est recherché dans le *bait*. Cependant, aucune différence de masse ne permet de retrouver ce résidu, car il est absent du *bait*. Ainsi, pour chaque case, la case d'origine est la case du dessus à laquelle on ajoute s_N (de valeur -4). Le meilleur alignement contiendra la case $D[2][1]$, de valeur 1, en noir.
Pour chaque case de la cinquième ligne (Figure 4.4), I est recherché dans le *bait*. Le résidu est retrouvé entre β_2 et β_4, ainsi qu'entre β_5 et β_7. Nous montrons comment est opéré le choix pour remplir la case $D[4][7]$, que nous savons contenue dans le meilleur alignement. Pour cette case, `SpecGlob` considère la possibilité de retrouver I sans décalage de masse, auquel cas il vient de la case $D[3][5]$, de valeur -12, à laquelle on ajoute s_A, de valeur 5. La première possibilité a donc une valeur de -7. La seconde possibilité est d'insérer un décalage de masse. Toutes les cases de la ligne précédente (i - 1 = 3) d'un indice j inférieur à 5 sont alors considérées. Leur valeur est prise en compte, et on ajoute s_R (de valeur 2) au maximum de la ligne, ici 6. La valeur de 8 correspond à la seconde possibilité, qui consiste à retrouver I avec un décalage de masse à partir de $D[3][3]$. Entre -7 et 8, la seconde possibilité est sélectionnée, et $D[4][7]$ est remplie avec la valeur 8 (en rouge). Le remplissage de cette case illustre donc comment un choix (insertion d'un décalage de masse ou non) dépend à la fois du système de scores et des cases déjà remplies.
Une fois les règles appliquées pour chaque case de la matrice D, celle-ci est entièrement remplie (Figure 4.5). Dans l'exemple, l'alignement a un score total de 13 (en violet), qui est le maximum de la dernière ligne. Dans cette matrice sont mises en valeur les cases appartenant au meilleur alignement, déterminé à partir de la matrice *Origin* via une étape de traceback.
À chaque case de D remplie, les coordonnées de la case d'origine ainsi que le type d'alignement sont en effet stockés dans la matrice *Origin*. La matrice *Origin* remplie sur l'exemple est montrée Figure 4.6.
Pour chaque case remplie dans D (Figure 4.5), *Origin* stocke trois valeurs. La première indique la coordonnée i de la case d'origine, la seconde indique la coordonnée j de la case d'origine, et le troisième indique le type d'alignement entre la case courante et la case d'origine : 0 pour *Align*, 1 pour *Realign* et 2 pour $No - align$. Par exemple, la case $D[1][1]$ en bleu a pour origine la case $D[0][0]$ à laquelle elle s'aligne sans décalage (*Align*), $Origin[1][1]$ contient donc (0, 0, 0). La case $D[4][7]$ en rouge a pour origine la case $D[3][3]$ à laquelle elle s'aligne avec un décalage de masse (*Realign*), $Origin[4][7]$ contient donc (3, 3, 1).



			0	1	2	3	4	5	6	7	8	9	10
H+	1	0	0,0,0	0,0,0	0,1,0	0,2,0	0,3,0	0,4,0	0,5,0	0,6,0	0,7,0	0,8,0	0,9,8
D	116,03	1	0,0,2	**0,0,0**	0,2,2	0,3,2	0,4,2	0,5,2	0,6,2	0,7,2	0,8,2	0,9,2	0,8,0
DW	302,11	2	1,0,2	**1,1,2**	1,2,2	1,3,2	1,4,2	1,5,2	1,6,2	1,7,2	1,8,2	1,9,2	1,10,2
DWY	465,17	3	2,0,2	2,1,2	2,2,2	**2,1,0**	2,4,2	2,5,2	2,6,2	2,7,2	2,1,1	2,9,2	2,10,2
DWYI	578,26	4	3,0,2	3,1,2	3,2,2	3,3,2	3,1,1	3,5,2	3,6,2	**3,3,1**	3,8,2	3,9,2	3,10,3
DWYIR	734,36	5	4,0,2	4,1,2	4,2,2	4,3,2	4,4,2	4,5,2	4,6,2	4,7,2	4,8,2	**4,7,0**	4,10,2

FIGURE 4.6 – **Exemple de matrice** *Origin* **produite par `SpecGlob`.** Les masses indiquées sont en Da.

L'étape de traceback (Figure 4.7) consiste ensuite à remonter dans D à partir du meilleur score de la dernière ligne et grâce à $Origin$. Les informations disponibles après l'étape de traceback sont résumées dans le tableau sous la matrice. À chaque résidu du *hit* ($peptide[i]$) sont associées des informations données par le meilleur alignement : un type d'alignement ($Align$, $Realign$ ou $No-align$) et un décalage associé. Chaque résidu peut donc être écrit tel quel s'il y a un alignement de type $Align$, avec son décalage associé entre crochets si l'alignement est de type $Realign$, ou enfin directement entre crochets si l'alignement est de type $No-align$.
Le traceback produit ainsi une liste qui contient les informations nécessaires à la production du $hitModified$, c'est-à-dire pour chaque ligne - et donc chaque résidu du *hit* - le type d'alignement et les décalages de masse associés au meilleur alignement. Dans notre exemple, le résidu D est retrouvé sans décalage de masse. Il est donc écrit tel quel dans le $hitModified$. Le résidu W n'est pas retrouvé, et est donc indiqué entre crochets. Le résidu Y est retrouvé avec un alignement de type $Align$, cependant il a avec le dernier résidu retrouvé (D) un décalage de masse de valeur -186,08 Da, qui est l'inverse de la masse de W. Il est donc écrit suivi de cette masse entre crochets. Le résidu I est retrouvé avec un décalage de 87,03 Da, valeur écrite après lui entre crochets. Le résidu R est retrouvé sans décalage de masse et donc écrit tel quel. On obtient donc le $hitModified$ suivant :

D[W]Y[-186,08]I[87,03]R

Ce $hitModified$ est riche en information. En effet, il indique que W n'est pas présent dans

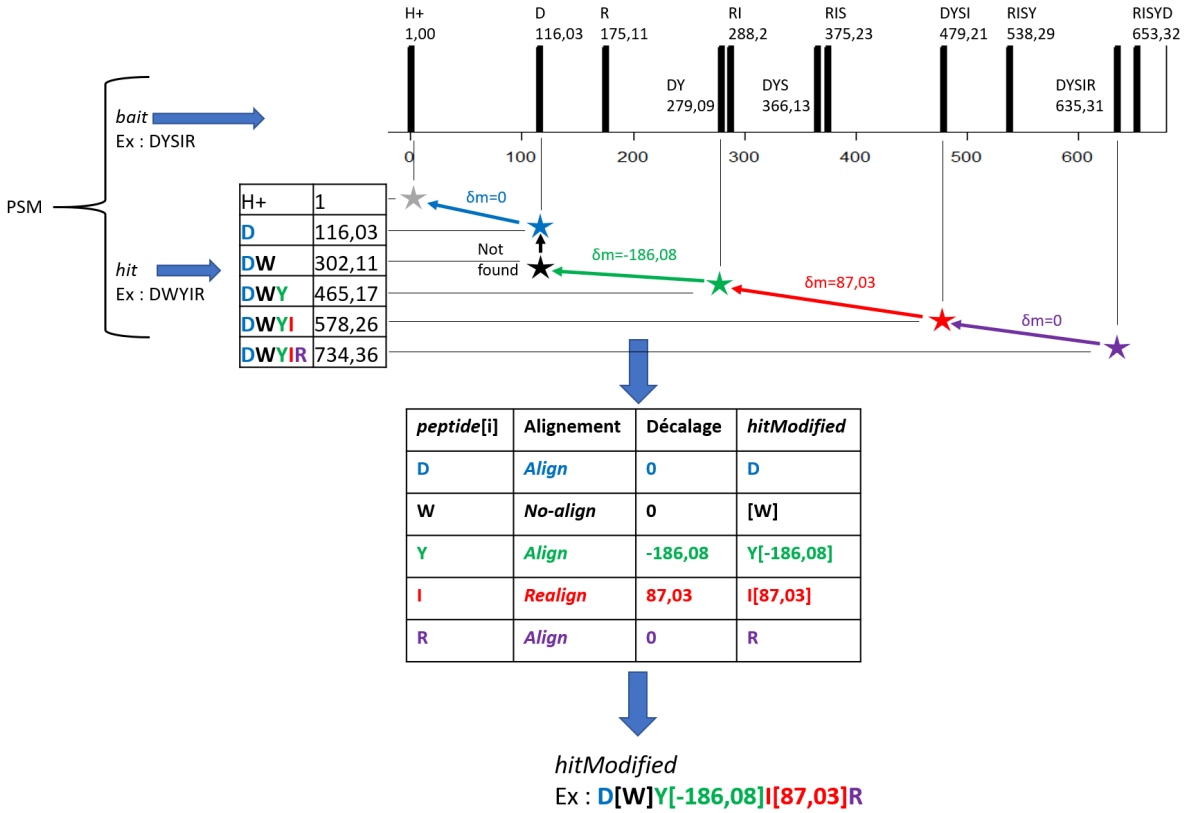


FIGURE 4.7 – **Étape de traceback et production du** *hitModified* **réalisées par** `SpecGlob`**.** Lors du traceback, un décalage de masse δm peut être associé ou non à chaque résidu du *hit*. Les masses indiquées sont en Da.

le *bait*. Y, immédiatement après, est retrouvé avec un décalage de -186,08 Da ; or cette valeur correspond à l'opposé de la masse de W (résidu de masse 186,08 Da). W n'est pas retrouvé dans le *bait* et Y doit être décalé négativement de sa valeur pour se rapprocher du *bait* ; ces deux informations indiquent donc qu'il est nécessaire de supprimer W dans le *hit* pour que la séquence corresponde au *bait*. Ensuite, I est retrouvé avec un décalage de 87,03 Da, qui doivent être insérés avant lui pour s'aligner sur le *bait*. Or 87,03 correspond à la masse d'un résidu connu, qui est S. Ainsi l'alignement indique qu'il faut insérer S avant I. En faisant ces deux opérations (délétion de W, insertion de S avant I), nous obtenons à partir de la séquence DWYIR la séquence DYSIR ; cette séquence correspond à celle du *bait*. Nous avons donc pu, grâce à notre alignement, interpréter un PSM avec deux modifications qui séparent les deux séquences correspondantes aux deux spectres, et ce sans *a priori* sur leur nombre ou leur nature.

4.2.3 Formalisation de `SpecGlob` et pseudocode

`SpecGlob` cherche donc le meilleur alignement des N masses du *hit* (stockées dans une liste $hitMasses[]$) sur les M masses du *bait* (stockées dans une liste $baitMasses[]$).
La matrice D de taille $N \times M$ est donc remplie selon des règles (voir Équation (4.1)) et le système de scores qui permettent de déterminer pour tout $1 \leq i \leq N-1$ et tout $1 \leq j \leq M-1$ la valeur de $D[i][j]$.

$$D[i][j] = \begin{cases} \max\left(D[i-1][k] + s_A ; \max\limits_{0 \leq m < k} D[i-1][m] + s_R\right) & \text{Si } aafound\text{=}\texttt{vrai} \\ D[i-1][j] + s_N & \text{Si } aafound\text{=}\texttt{faux} \end{cases} \tag{4.1}$$

La valeur de $D[i][j]$ dépend d'un booléen, $aafound$, qui est mis à $vrai$ seulement si la masse du $i^{ème}$ résidu, noté aa, du *hit* est trouvée dans le *bait*. Plus précisément, $aafound$=$vrai$ s'il existe un $k < j$ tel que :

$$baitMasses[j] - baitMasses[k] = hitMasses[i] - hitMasses[i-1]$$

Les trois valeurs du système de scores sont utilisées pour déterminer $D[i][j]$ comme suit (voir aussi l'Équation (4.1)) :

- Si $aafound$ est `vrai`, il y a deux possibilités, selon que le résidu correspondant aa est trouvé avec un décalage de masse
  - si aa est trouvé sans décalage de masse, l'alignement vient de la case $D[i-1][k]$, et le score correspondant est $D[i-1][k] + s_A$;
  - si aa est trouvé mais il faut introduire un décalage de masse, le score correspondant est $D[i-1][m] + s_R$, où m est l'indice (compris entre 0 et $k-1$) qui maximise le score. Si plusieurs valeurs de m existent, le plus grand est choisi, c'est-à-dire qu'à score égal `SpecGlob` choisit la masse la plus proche de j.

  La valeur maximum entre les deux possibilités du dessus est sélectionnée.
- Si $aafound$ est `faux`, $D[i][j]$ prend la valeur $D[i-1][j] + s_N$.

La première ligne de D est mise à 0 (donc $D[0][j] = 0$ pour chaque $0 \leq j \leq M-1$), et pour chaque $0 \leq i \leq N-1$, $D[i][0]$ est mis à $i \times s_N$. Les valeurs de toutes les cases $D[i][j]$ sont ensuite calculées pour i allant de 1 à $N-1$ et pour j allant de 1 à $M-1$, c'est-à-dire de gauche à droite et de haut en bas, selon l'Équation (4.1). Pendant le processus, les coordonnées de la case d'origine, ainsi que le type d'alignement correspondant ($Align$,

Realign ou $No-align$), sont stockés dans la matrice $Origin$.
Une fois que D est remplie, le traceback démarre de la valeur $D[N-1][p]$, où p est l'indice qui maximise la valeur de la dernière ligne de D. Si plusieurs valeurs maximum existent, `SpecGlob` choisit la case qui ajuste Δm ; on dit qu'une case de $D[N-1][j]$ ajuste Δm lorsque $baitMasses[j] - hitMasses[M-1] = \Delta m$, c'est-à-dire que la somme de tous les décalages de masse est égale à Δm ; si une telle case n'existe pas, la plus petite valeur de p est choisie. Durant l'étape de traceback, l'origine de chaque case est retrouvée en utilisant $Origin$. Selon les trois cas décrits dans l'Équation (4.1) et le décalage de masse, le traceback produit une sortie sous la forme d'une chaîne de caractères, le $hitModified$. Elle contiendra le $i^{ème}$ résidu du *hit* tel quel, avec un décalage de masse ou bien marqué comme non trouvé.
La complexité temporelle du remplissage des matrices est de $\mathcal{O}(N \times M^2)$, dû à la partie "boucle principale" de l'Algorithme 3. La complexité spatiale du remplissage est relative à la taille des matrices D et $Origin$, et est donc de $\mathcal{O}(N \times M)$. Le traceback est réalisé en temps linéaire en la taille du *hit*.
Les variables nécessaires au fonctionnement de `SpecGlob` sont présentées Algorithme 2. Ensuite, le pseudocode général de `SpecGlob`, du remplissage des matrices D et $Origin$ et enfin le traceback sont décrits respectivement Algorithmes 3, 4, et 5.

Algorithme 2 : Variables

accuracy #Valeur par défaut 0.02
baitMasses [] #*baitMasses* [] est une table contenant les masses des ions b et y du *bait*
hitMasses [] #*hitMasses* [] est une table contenant les masses des ions b du *hit*
N #Nombre d'ions b dans le *hit*
M #Nombre d'ions b et y dans le *bait*
s_A, s_R, s_N #Valeurs du système de scores (entiers)
Δm #Différence de masse parente entre le *bait* et le *hit*
peptide #Séquence en résidus du *hit*
D [][] #La table D est une matrice d'entiers à 2 dimensions, de taille N × M
Origin [][][]
#La table *Origin* est une matrice à 3 dimensions, de taille $N \times M \times 3$
#- sa première dimension représente l'indice de la case d'origine dans le *hit*
#- sa seconde dimension représente l'indice de la case d'origine dans le *bait*
#- sa troisième dimension représente le type d'alignement : 0 si le résidu est retrouvé sans décalage de masse ; 1 si trouvé avec un décalage ; 2 si non trouvé

Algorithme 3 : SpecGlob($baitMasses$ [], $hitMasses$ [])

#Fonction principale prenant en entrée le *bait* et le *hit*, initialisant les tables D et $Origin$, appelant la fonction qui remplit les tables et réalisant l'étape de traceback afin de produire la chaîne de caractères $hitModified$

Sortie : $hitModified$

#Chaîne de caractères décrivant le meilleur alignement entre le *bait* et le *hit* trouvé par notre méthode ; un résidu est écrit tel quel s'il est retrouvé sans décalage de masse ; s'il est retrouvé avec un décalage de masse, sa valeur est écrite entre crochets après le résidu ; le résidu est écrit entre crochets s'il n'est pas retrouvé

Algorithm :

#Initialisation des tables D et $Origin$

$D[0][0] \leftarrow 0$
$Origin[0][0][0] \leftarrow 0$
$Origin[0][0][1] \leftarrow 0$
$Origin[0][0][2] \leftarrow 0$
pour $i \leftarrow 1$ **à** $N-1$ **faire**
  $D[i][0] \leftarrow i \times s_N$ #Permet de marquer les résidus comme non trouvés lorsque l'alignement démarre
  $Origin[i][0][0] \leftarrow i-1$
  $Origin[i][0][1] \leftarrow 0$
  $Origin[i][0][2] \leftarrow 2$
fin
pour $j \leftarrow 1$ **à** $M-1$ **faire**
  $D[0][j] \leftarrow 0$ #Car $hitMasses[0]$ est toujours la masse de l'ion H^+
  $Origin[0][j][0] \leftarrow 0$
  $Origin[0][j][1] \leftarrow j-1$
  $Origin[0][j][2] \leftarrow 0$
fin

#Remplissage des tables D et $Origin$ (boucle principale)

pour $i \leftarrow 1$ **à** $N-1$ **faire**
  pour $j \leftarrow 1$ **à** $M-1$ **faire**
    $fillMatrices(i,j)$
  fin
fin

#Parcours de D pour trouver sa "meilleure" case $D[N-1,p]$ d'où faire le traceback

$score \leftarrow -1000$ #Le score est fixé à une petite valeur arbitraire
$p \leftarrow 0$
pour $j \leftarrow 0$ **à** $M-1$ **faire**
  $solveDm \leftarrow abs(baitMasses[j] - hitMasses[N-1] - \Delta m)$
  #$solveDm$ est utilisé de façon à favoriser une case de traceback expliquant Δm
  si $D[N-1][j] > score$ *ou* $(D[N-1][j] = score$ *et* $solveDm < accuracy)$ **alors**
    $score \leftarrow D[N-1][j]$
    $p \leftarrow j$
  fin
fin

#traceback depuis $D[N-1,p]$

$hitModified \leftarrow traceback(N-1,p)$

#Renvoi du résultat

retourner $hitModified$

Algorithme 4 : fillMatrices(i, j)

#Cette fonction prend en entrée les coordonnées (i, j) d'une case pour remplir les tables D et $Origin$, et remplit cette case selon le système de scores
Sortie : Tables de programmation dynamique D et $Origin$ remplies à (i, j)
Algorithm :
$Origin1$ [] #Ensemble de 3 valeurs utilisées pour compléter $Origin$
$Origin2$ [] #Ensemble de 2 valeurs utilisées pour compléter $Origin$
$val1, val2 \leftarrow -1000$
$aafound \leftarrow faux$ # Booléen
$k \leftarrow j - 1$
#Recherche si une différence de masse correspondant au $i^{ème}$ résidu du hit existe dans le $bait$
tant que *(non aafound) et* $(k \geq 0)$ **faire**
  $\delta \leftarrow hitMasses[i] - baitMasses[j] - hitMasses[i-1] + baitMasses[k]$
  si $abs(\delta) < accuracy$ **alors**
    #Résidu retrouvé
    $aafound \leftarrow vrai$
  fin
  sinon
    $k \leftarrow k - 1$
  fin
fin
si $aafound$ **alors**
  $Origin1[0] \leftarrow i - 1$
  $Origin1[1] \leftarrow k$
  $Origin1[2] \leftarrow 0$
  $val1 \leftarrow D[i-1][k] + s_A$
  $m \leftarrow -1$
  $maxValue \leftarrow -1000$
  pour $a \leftarrow k$ **à** 0 **faire**
    si $D[i-1][a] > maxValue$ **alors**
      $m \leftarrow a$
      $maxValue \leftarrow D[i-1][a]$
      $val2 \leftarrow maxValue + s_R$
    fin
  fin
  si $val2 > val1$ **alors**
    #Choix entre s_A et s_R de façon à maximiser le score total
    $val1 \leftarrow val2$
    $Origin1[1] \leftarrow m$
    $Origin1[2] \leftarrow 1$
  fin
fin
sinon
  #Calcul du score si le résidu n'est pas retrouvé
  si *not* $aafound$ **alors**
    $scoreMax \leftarrow D[i-1][j]$
    $val2 \leftarrow scoreMax + s_N$
    $Origin2[0] \leftarrow i - 1$
    $Origin2[1] \leftarrow j$
  fin
fin
#Choix final
si $val1 \geq val2$ **alors**
  $D[i][j] \leftarrow val1$
  $Origin[i][j][0] \leftarrow Origin1[0]$
  $Origin[i][j][1] \leftarrow Origin1[1]$
  $Origin[i][j][2] \leftarrow Origin1[2]$
fin
sinon
  $D[i][j] = val2$
  $Origin[i][j][0] \leftarrow Origin2[0]$
  $Origin[i][j][1] \leftarrow Origin2[1]$
  $Origin[i][j][2] \leftarrow 2$
fin

Algorithme 5 : traceback(i, j)

```
#Cette fonction prend en entrée les coordonnées (i, j) de la case d'où commencer
 le traceback, stocke les coordonnées des cases d'Origin impliquées dans le
 meilleur alignement, et créée le hitModified en se basant sur ces coordonnées
Sortie : Chaîne de caractères hitModified
Algorithm :
#Rassemblement de toutes les coordonnées d'Origin qui seront utilisées pour
 créer le hitModified
path #Liste de tous les points de l'alignement, initialement vide
point ← Point(i, j) #Objet avec deux coordonnées
path ← path + point
x ← i
y ← j
pour m ← i − 1 à 1 faire
    coordX ← x
    coordY ← y
    x ← Origin[coordX][coordY][0]
    y ← Origin[coordX][coordY][1]
    point ← Point(x, y)
    path ← path + point
fin
Δ ← 0
ΔAlign ← 0
hitModified ← "" #Commence avec une chaîne vide
#Production du hitModified
pour n ← 0 à longueur(path) − 1 faire
    x ← path[n].getX
    y ← path[n].getY
    Δ ← hitMasses[x]−baitMasses[y]
    typeAlignement ← Origin[x][y][2]
    si typeAlignement = 2 alors
        #Résidu non retrouvé
        hitModified ← hitModified + "[" + peptide[n]+"]"
    fin
    sinon
        si abs(ΔAlign − Δ) > accuracy alors
            #Résidu retrouvé avec un décalage de masse
            hitModified ← hitModified + peptide[n]
            hitModified ← hitModified + "[" + ΔAlign − Δ + "]"
            ΔAlign ← Δ
        fin
        sinon
            #Résidu retrouvé sans décalage de masse
            hitModified ← hitModified + peptide[n]
        fin
    fin
fin
Δ ← ΔAlign − hitMasses[N − 1]−baitMasses[M − 2]
si abs(Δ) > accuracy alors
    #Gestion du reste de décalage de masse éventuel
    #Un underscore "_" est utilisé pour diférencier un décalage de masse lié au
     au dernier résidu ou localisé après celui-ci
    hitModified ← hitModified + "_" + "[" + Δ + "]"
fin
retourner hitModified
```

4.2.4 Autres exemples de résultats

Quatre autres exemples de PSM traités par SpecGlob sont présentés Table 4.1. Les PSM traités peuvent être présentés dans un fichier résultat, dans lequel le $hitModified$ est écrit sous la forme d'une chaîne de caractères. Ensuite, il peut être analysé pour retrouver la séquence du *bait*. Le $hitModified$ indique des décalages de masse et des résidus non retrouvés qui forment les opérations (délétions, insertions, substitutions) à effectuer ; les compter revient donc à déterminer le nombre de modifications de séquences qui séparent le *bait* et le *hit*. Certaines de ces modifications permettent de retrouver une partie de la séquence du *bait*. Pour chaque PSM, ces opérations et leur nombre sont indiqués dans la Table 4.1 ; la dernière colonne indique si elles permettent de déterminer la séquence du *bait* (dont la séquence peut être utilisée après l'exécution de SpecGlob pour valider le résultat). Par exemple, pour le premier PSM, SpecGlob fournit le $hitModified$ G[I]T[-14,02]ACCITK. Cela signifie que T doit être décalé sur I, qui n'est pas retrouvé dans le *bait*, et ce décalage doit être de 14,02 Da. Or la masse de I (113,08 Da) moins 14,02 Da vaut 99,06 Da, ce qui correspond à la masse de V. Nous pouvons donc substituer I en V et obtenir la séquence GVTACCITK. Dans le contexte théorique, nous pouvons vérifier que cette séquence est bien celle du *bait*. C'est le cas, ainsi nous avons réussi notre interprétation, ce qui est indiqué dans la dernière colonne. En revanche, pour le dernier exemple le $hitModified$ vaut QVSVIA[1957,82]K. Nous devons donc insérer une séquence qui correspond à une masse de 1957,82 Da avant A dans le *hit* pour obtenir le *bait*. Cependant cette masse peut correspondre à de nombreuses séquences possibles, ce qui soulève une ambiguïté ; ainsi, il n'est pas possible de déterminer précisément la séquence du *bait*.

Nous avons donc à ce stade un algorithme qui est capable de prendre en entrée un ensemble de PSM et d'en renvoyer les meilleurs alignements selon un système de scores donné. Il est donc possible de le tester sur un grand nombre de PSM.

Afin de positionner SpecGlob par rapport à un outil existant, nous l'avons comparé à MODPlus [Na, Kim et Paek 2019] ; les résultats sont présentés dans la Section 4.3.

4.3 Comparaison de SpecGlob et MODPlus

Une façon de juger des performances de SpecGlob est de le comparer à d'autres outils d'interprétation de PSM. Comme expliqué au début de ce chapitre, les outils d'interprétation existants sont limités dans le nombre de modifications considérées ainsi que dans leur

TABLE 4.1 – **Exemples de *hitModifieds* fournis par `SpecGlob` ainsi que leurs interprétations**

bait	*hit*	*hitModified*	#Modification(s)	Du *hit* au *bait*	Interprétable avec `SpecGlob` ?
GVTACCITK	GITACCITK	G[I]T[-14,02]ACCITK	1	masse(I) – 14,02 Da = masse(V) ⟶ Substitution(I,V)	Oui
EGASDEWIR	EASDEWIR	EA[57,02]SDEWIR	1	57,02 Da = masse(G) ⟶ Insertion(G)	Oui
VCASIYQK	VSFVIFVVI PIHASIYGAK	[V][S][F][V][I][F][V]V[-791,46] [I][P][I][H]A[-300,25]SIY[G][A]K	3	791,46 Da = masse(VSFVIFV) ⟶ Délétion(VSFVIFV) masse(IPIH) – 300,25 Da = masse(C) ⟶ Substitution(IPIH, C) masse(GA) = masse(Q) ⟶ Substitution(GA, Q)	Oui
QVSVIQWSSIVH GEQCCSVWNAK	QVSVIAK	QVSVIA[1957,82]K	1	1957,82 Da = masse(?)	Non

Chaque ligne correspond à un PSM (*bait, hit*) fourni par `SpecOMS`, ainsi que la sortie de `SpecGlob`, $hitModified$, et son interprétation. Pour une meilleure lisibilité, les spectres du *bait* et du *hit* sont représentés par leurs peptides.
Ensuite, il est possible de connaître le nombre de modifications dans un PSM en comptant le nombre d'opérations (délétions, insertions, substitutions) indiquées dans le $hitModified$, puis de passer à une étape d'interprétation (colonne "du *hit* au *bait*") où l'information contenue dans le $hitModified$ est transformée par les modifications lorsque celles-ci sont possibles. La colonne la plus à droite indique si la séquence du *bait* peut être retrouvée après les opérations sur le $hitModified$.

nature. Cependant, nous pouvons comparer `SpecGlob` à un outil proche, `MODPlus` [NA, KIM et PAEK 2019], qui est dédié à la sélection et l'interprétation des PSM. `MODPlus` sélectionne dans une base de données de peptides les candidats qui partagent des *tags* avec le spectre, et ensuite aligne les spectres avec une approche de programmation dynamique pour localiser les modifications. Pour cette étape, `MODPlus` prend en entrée une liste de modifications limitée, soit des modifications issues d'Unimod [CREASY et COTTRELL 2004], soit renseignées par l'utilisateur. Ainsi, même si `SpecGlob` et `MODPlus` reposent tous les deux sur la programmation dynamique, les deux outils ont des principes de fonctionnement différents.
`SpecGlob` peut prendre en entrée les PSM renvoyés par n'importe quelle méthode OMS. Nous avons choisi de l'utiliser sur les PSM produits par `SpecOMS`, pour les raisons précisées dans le Chapitre 3. L'idée ici est de comparer `SpecOMS` et `SpecGlob` d'un côté, et `MODPlus` de l'autre, selon leur capacité à sélectionner les PSM et interpréter leurs modifi-

cations. Pour pouvoir effectuer cette comparaison, des peptides du protéome humain ont été modifiés systématiquement sur certains résidus, puis les deux outils ont été utilisés pour sélectionner les PSM et les interpréter. Pour cette étude, 50 000 peptides tryptiques avec une longueur comprise entre 12 et 25 résidus ont été choisis au hasard. Ces peptides ont été modifiés pour créer deux jeux de données. Dans le jeu que nous appellerons *ND*, les asparagines sont déamidées (N+0,984016 Da) et des adduits de sodium sont affectés à chaque acide aspartique (D+21,981943 Da). Pour le jeu que nous appellerons *SCT*, les sérines sont substituées par des alanines (S-15,99 Da), les cystéines sont carbamidométhylées (C+57,0214 Da) et les thréonines sont supprimées (T-101,0477 Da).
`SpecOMS` est utilisé pour trouver le meilleur candidat pour chaque peptide modifié (selon le shift SPC, voir Chapitre 3, page 66) et les PSM résultants sont traités par `SpecGlob`. `SpecGlob`, implémenté en Java, a été exécuté sur un ordinateur de bureau sous Windows 10 (Intel i7, 2,6 GHz) avec 16 Go de mémoire alloués au programme. `MODPlus` est utilisé à la fois pour générer les PSM, et localiser les modifications par un alignement.
Les résultats obtenus par les deux outils sont présentés Table 4.2.
La première remarque que l'on peut faire concernant la Table 4.2 concerne le temps d'exécution : `SpecOMS-SpecGlob` est particulièrement rapide par rapport à `MODPlus`. Lorsque, pour ce dernier, l'espace de recherche est limité aux modifications les plus courantes, `SpecOMS-SpecGlob` est 40 fois plus rapide, un ratio qui monte à 90 lorsque le paramétrage de `MODPlus` autorise toutes les modifications d'Unimod. Ensuite, le nombre de PSM où le peptide a été correctement identifié a été calculé. Cela signifie que pour un *bait* donné, le *hit* est le peptide à l'origine du *bait*, c'est-à-dire le peptide que l'on a modifié pour obtenir le *bait*. Nous pouvons étudier les alignements afin de voir quel pourcentage des modifications sont bien identifiées et positionnées, c'est-à-dire avec la bonne masse, et au bon endroit. Pour le jeu *ND*, `MODPlus` est très efficace, alors que `SpecOMS-SpecGlob` est meilleur pour découvrir les modifications dans le jeu *SCT*. Cela n'est pas surprenant dans le sens où *ND* est un jeu de données idéal pour `MODPlus`. En effet, dans *ND*, seules deux modifications abondantes parfaitement décrites dans Unimod ont été simulées dans les peptides. Le jeu *SCT*, quant à lui, contient des modifications parmi lesquelles la délétion de T, une modification qui n'est pas répertoriée dans Unimod. L'absence de délétions de résidus dans Unimod explique 96% des PSM incorrects fournis par `MODPlus`, parmi les peptides qui contiennent au moins un T. Les autres erreurs d'interprétation sont principalement dues à la présence de plusieurs modifications consécutives proches les unes des autres que `MODPlus` à tendance à interpréter en une seule. Ainsi, même si `MODPlus`

TABLE 4.2 – **Comparaison entre `SpecOMS-SpecGlob` et `MODPlus` sur deux jeux de spectres théoriques.**

	ND		*SCT*	
	`SpecOMS-SpecGlob`	`MODPlus`	`SpecOMS-SpecGlob`	`MODPlus`
PSM avec identification correcte (%)	84 (42 111/50 000)	98,3 (49 162/50 000)	61,7 (30 846/50 000)	44,8 (22,393/50 000)
Idem mais avec entre 1 et 3 modifications (%)	89,32 (28,984/32 449)	97,8 (35 099/35 870)	76,7 (22,924/29 868)	50,5 (15 451/30 613)
PSM avec toutes les modifications correctement localisées parmi ceux avec une identification correcte (%)	89,7 (37 782/42 111)	96,1 (47 247/49 162)	90 (27 770/30 846)	80,3 (17 987/22 393)
Modifications correctement identifiées parmi toutes (%)	75,54 (41 825/55 364)	95,1 (68 358/71 852)	67,1 (48 962/72 966)	30,9 (31 385/101 415)
Modifications correctement identifiées parmi les PSM avec identification correcte (%)	85,8 (41 519/48 353)	95,8 (68 237/70 812)	91,8 (47 735/51 986)	79,8 (29 559/37 027)
Temps d'exécution (en minutes)	5	183	5	452

`SpecOMS` suivi par `SpecGlob` d'un côté, et `MODPlus` de l'autre, ont traité deux jeux de spectres portant des modifications. Dans *ND*, les spectres théoriques correspondent à des peptides avec les résidus N et D modifiés comme suit : N+0,98 Da, D+21,98 Da, ce qui produit 0 à 13 modifications par spectre. Dans *SCT*, les spectres théoriques correspondent au peptides avec les résidus S, C et T modifiés comme suit : S-15,99 Da, C+57,02 Da, T-101,05 Da, produisant des spectres portant 0 à 15 modifications, dépendant de leur composition en résidus. Pour chaque critère, les chiffres sont donnés en pourcentage, puis en nombre entre parenthèses.

est plus efficace que `SpecOMS-SpecGlob` pour analyser les modifications connues et abondantes (comme dans le jeu *ND*), `SpecOMS-SpecGlob` est meilleur pour mettre rapidement en évidence une large variété de modifications qu'un échantillon peut contenir, dont des modifications inconnues, comme dans le jeu *SCT*. De plus, `SpecOMS-SpecGlob` montre une certaine stabilité de comportement entre les deux jeux de PSM en termes d'efficacité et de temps d'exécution. Ces résultats montrent les avantages que présente une approche sans *a priori* pour interpréter les PSM modifiés.

Nous avions vu dans le Chapitre 3 que de nombreux PSM n'avaient pas pu être interprétés ; `SpecGlob` a donc ensuite été testé sur les PSM renvoyés par `strat-shift` (présentés dans le Chapitre 3). Les résultats relatifs à ce travail sont présentés dans la section suivante (Section 4.4).

4.4 Interprétation des résultats de SpecOMS par SpecGlob

strat-shift (présentée au Chapitre 3) produit avec SpecOMS 455 404 PSM. Pour cela, tous les spectres théoriques issus de tous les peptides du protéome humain sont comparés entre eux, et ceux qui ont un SPC d'au moins 7 sont conservés. Ensuite, le meilleur *hit* par *bait* est choisi selon le shift SPC. Ces PSM ont été traités par SpecGlob afin de voir à quel point le taux d'interprétation des PSM pouvait être amélioré par notre nouvel outil qui permet de traiter les PSM comportant plusieurs modifications. Dans cette section, je détaille d'abord les observations générales que nous avons pu faire sur ces PSM traités par SpecGlob. Ensuite, j'explique comment les $hitModifieds$ ont pu être classés selon la possibilité de les reconstruire en une séquence de résidus, et comment cette reconstruction a pu être faite afin de vérifier l'efficacité de notre outil. Enfin, ces résultats sont discutés en détails.

4.4.1 Observations générales

SpecGlob finit son exécution en approximativement trois minutes pour traiter l'ensemble des 455 404 PSM.
On connaît pour chaque PSM le *bait* et le *hit* fournis par SpecOMS, ainsi que le $hitModified$ renvoyé par SpecGlob.
Dans un premier temps, le nombre de modifications par PSM a été calculé. Une modification est définie comme une opération de séquence (insertion, délétion, substitution de résidu(s)) indiquée dans le $hitModified$. Pour déterminer ces opérations, les indications du $hitModified$ (succession de résidu(s) non retrouvés et décalages de masse) doivent être traitées ensemble ou individuellement.
Dans un $hitModified$, lorsqu'un résidu n'est pas retrouvé (et donc présenté entre crochets), il y a deux possibilités. La première est que l'on retrouve un décalage après le résidu qui le suit, comme dans le $hitModified$ G[I]T[-14,02]ACCITK (revoir Table 4.1, page 118). La seconde possibilité est que l'on n'en retrouve pas, comme dans le $hitModified$ fictif EA[K][E]WIR. Dans le premier cas, "I" et "-14,02" sont à traiter ensemble, et il faudra ajouter une séquence de valeur masse(I) - 14,02 Da avant T, et donc substituer I par une autre séquence. Dans le second cas, il faudra insérer une séquence de valeur masse(KE) avant W, et donc substituer KE par une autre séquence. Dans les deux cas, il n'y a qu'une seule modification. Un décalage peut également être retrouvé seul, sans devoir être appliqué à un résidu en particulier, comme dans le $hitModified$ EA[57,02]SDEWIR. Une séquence

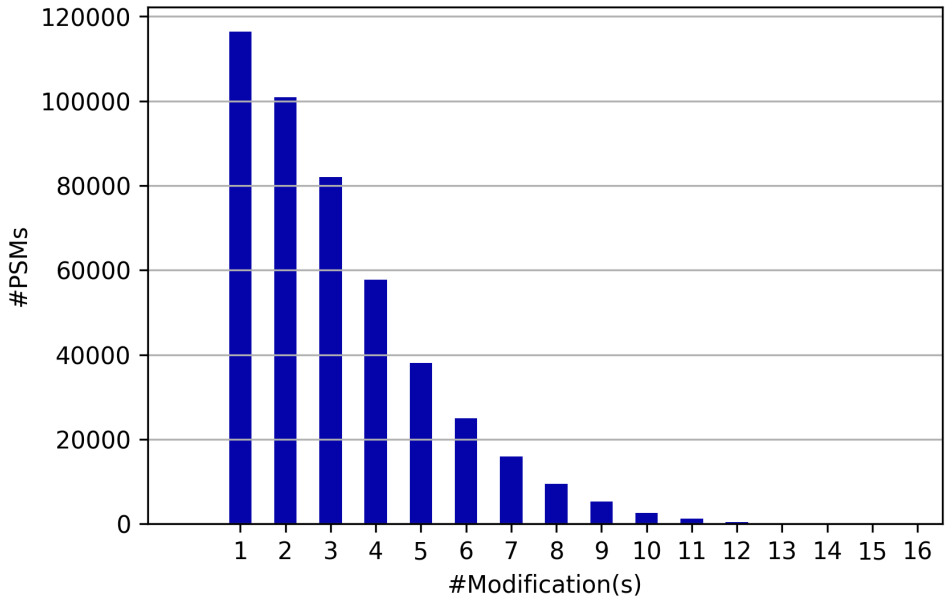


FIGURE 4.8 – **Distribution du nombre de modifications proposées par `SpecGlob` après alignement des PSM.** 455 404 PSM générés par `SpecOMS` ont été analysés par `SpecGlob`. Dans ce jeu de données, le nombre de modifications prédites par `SpecGlob` varie de 1 à 16 dans chaque PSM.

de masse 57,02 Da doit être insérée avant A. Il n'y a là aussi qu'une seule modification. Il est donc possible de parcourir tous les $hitModifieds$ de tous les PSM, et compter les modifications de cette façon. Les PSM comportent au total 1 417 713 modifications, il y a donc en moyenne un peu plus de 3 modifications par PSM. La distribution du nombre de modifications dans les PSM est présentée Figure 4.8. Nous pouvons voir qu'environ un quart des PSM ont un alignement optimal qui ne comporte qu'une seule modification, et donc une large majorité des PSM ont un alignement optimal qui a strictement plus d'une modification. Si ces alignements s'avèrent utiles pour retrouver la séquence, ce résultat confirme la nécessité de pouvoir interpréter les PSM pour lesquels *plusieurs* modifications séparent le *bait* du *hit*. Ce qui introduit cette question cruciale : à quel point, dans ce jeu de données, les $hitModifieds$ des PSM permettent-ils de retrouver la séquence du *bait* ?

4.4.2 Évaluation et reconstruction automatique d'un *baitModel*

Dans cette section, je décris comment les $hitModifieds$ peuvent être évalués en 1) déterminant à quel point leurs modifications peuvent être interprétées ; 2) étudiant les caractéristiques des ***baitModels***, qui sont supposés représenter le plus précisément possible la

TABLE 4.3 – **Exemples de *baitModels* déterminés à partir des *hitModifieds*, ainsi que la couleur de leurs modifications**.

bait / *hit* / *couleur*	*hitModified*	*baitModel*
VCASIYQK VSFVIFVVIPIHASIYGAK Vert	[V][S][F][V][I][F][V]V[-791,46][I][P][I][H]A[-300,25]SIY[G][A]K Délétion de VSFVIFV — Substitution IPIH→C — GA→Q	VCASIYQK
GATPPAPPR GAAPAPPR Orange	G[A]A[-71,04]P[198,10]APPR Délétion de A — PT/TP	GA[198,10]PAPPR
FQMPDQGMTSADDFFQGTK GMTTVDDFFQGTK Rouge	G[746,31]M[T]T[-101,05][V]A[59,00]DFFQGTK ? — Délétion de T — GT/TG/AS/SA	[746,31]GMT[158,07]DDFFQGTK

Les modifications et les PSM sont colorés selon leur degré d'ambiguïté, et les *baitModels* sont construits en appliquant les modifications décrites sous les *hitModifieds* - seules les modifications Vertes peuvent être utilisées pour construire le *baitModel*, étant donné qu'elles ne sont pas ambiguës. Le troisième exemple contient un *hitModified* avec trois modifications : la première est Rouge car le décalage de masse ne correspond à aucune masse connue du répertoire (combinaisons de masses jusqu'à trois résidus) ; la deuxième est Verte, car le premier T n'est pas aligné, mais le décalage de masse requis pour aligner le second T correspond à sa délétion ; la troisième est Orange, car la masse de V ajoutée au décalage de masse de 59 Da peut être expliqué par plusieurs combinaisons de séquences de résidus (GT, TG, AS ou SA). À cause de la modification Rouge, le PSM dans sa globalité est classé comme Rouge. Au contraire, le premier exemple contient seulement des modifications Vertes, et donc le PSM est classé comme Vert. Le second exemple est un PSM Orange, car il contient une modification Verte et une modification Orange.

séquence du *bait*, dans lequel les opérations non ambiguës indiquées dans le *hitModified* sont transformées en séquence (voir Section 4.4.2). Le lien entre ces critères et la stratégie target/decoy est également discuté.

Classification d'un *hitModified* et de ses modifications

On rappelle que dans les PSM auxquels nous nous intéressons, seules des modifications de séquences (insertions, délétions et substitutions de résidus) séparent le *hit* du *bait*. Par conséquent, certaines valeurs des décalages de masse et résidus non retrouvés présents dans le *hitModified* peuvent être interprétées grâce à la connaissance des masses des résidus. Pour déterminer à quel point un *hitModified* peut être interprété, la classification en couleurs exposée dans le Chapitre 3 a été adaptée aux résultats de `SpecGlob`. Des exemples de classification sont présentés Table 4.3.
Tout d'abord, une modification est classée comme **Verte** si elle peut être transformée en

une séquence sans ambiguïté. Une modification dont la masse correspond à l'insertion d'un seul résidu, ou à la substitution d'un résidu par un autre illustre cette situation. La délétion d'un ou plusieurs résidus consécutifs est aussi considérée comme une modification Verte. À titre d'exemple, dans la Table 4.3, le premier PSM ne contient que des modifications Vertes.
Une modification est **Orange** si elle peut être expliquée par une opération dans le *bait*, mais avec une ambiguïté en termes de séquence. Cela arrive par exemple dans le cas où la masse de la modification correspond à l'insertion de plusieurs résidus dont l'identité est connue, mais pour lesquels l'ordre dans la séquence à insérer est incertain. Nous avons choisi de limiter la classe Orange à la recherche d'une combinatoire de masses correspondant à au plus trois résidus.
Dans la Table 4.3, la modification de 198,1 Da est une modification Orange, car elle peut correspondre à PT ou TP. Enfin, une modification qui n'est ni Verte ni Orange est classée comme **Rouge**, comme la première modification du troisième PSM de la Table 4.3. Pour réaliser cette classification, les indications du $hitModified$ pourront être traitées seules ou associées ensemble, selon le type de modification, et annotées automatiquement :

- une **délétion** est caractérisée par un ou plusieurs résidus non retrouvés (entre crochets) et le résidu qui les suit est retrouvé avec un décalage dont la valeur est l'opposée de la masse de l'ensemble des résidus non retrouvés, comme pour la première modification du premier PSM de la Table 4.3. Une délétion peut être réalisée sans ambiguïté et est donc toujours classée comme une modification Verte ;
- une **substitution** peut :
  - avoir la même structure qu'une délétion, mais la somme des masses des résidus et du décalage qui les suit, notée x, est non nulle. Si x correspond à la masse d'un résidu mais ne correspond pas à la masse d'une séquence de 2 à n résidus (n étant un paramètre mis à 3 dans notre cas), cette insertion n'est pas ambiguë et est donc classée comme Verte. Si x correspond à la masse d'une séquence de 2 à n résidus, elle est classée comme Orange, car plusieurs permutations de résidus peuvent avoir une masse de x ; la détermination de la séquence correspondante est donc ambiguë. Sinon, elle sera classée comme Rouge ;
  - correspondre à un groupe de résidus non retrouvés sans décalage qui les suit. Cela signifie que l'on doit faire une substitution sans changement de masse, et donc remplacer la séquence de résidus non retrouvés par une séquence de même masse x. Selon le même principe que la substitution, la couleur dépend

TABLE 4.4 – **Distribution des modifications et des PSM dans les trois catégories de couleurs.**

	Vert	Orange (3 résidus)	Rouge	Total
#Modifications	740 458 232 925 dans les PSMs Verts 152 332 dans les PSMs Oranges 355 201 dans les PSMs Rouges	286 616 141 132 dans les PSMs Oranges 145 484 dans les PSMs Rouges	390 639 dans les PSMs rouges	1 417 713
#PSMs	132 137	114 454	208 813	455 404

La première ligne, "#Modifications", montre le nombre total de modifications Vertes, Oranges et Rouges dans les 455 404 PSM, ainsi que leur distribution dans chaque couleur de PSM (par exemple parmi les 740 458 modifications Vertes, 355 201 sont dans des PSM Rouges). La seconde ligne "#PSM" montre le nombre de PSM dans chaque catégorie de couleur.

de l'ambiguïté de x. La dernière modification du premier PSM de la Table 4.3 correspond à une substitution sans changement de masse, puisque GA et Q ont la même masse ;

— une **insertion** sera caractérisée par la présence d'un décalage de masse isolé (sans résidu non retrouvé auquel l'ajouter, comme pour une substitution) de masse x. La couleur dépend de l'ambiguïté liée à x.

Sur les PSM renvoyés par SpecOMS, 52,23% des modifications contenues dans les $hitModifieds$ sont Vertes, 20,22% sont Oranges et 27,5% sont Rouges (voir Table 4.4). Cette classification des modifications peut être étendue à l'échelle d'un PSM. En effet, lorsque toutes les modifications d'un $hitModified$ donné sont Vertes, cela signifie que l'on peut convertir sans ambiguïté le $hitModified$ en une séquence peptidique complète. Dans ce cas, le PSM correspondant est classé comme Vert. Un PSM est Orange si le $hitModified$ correspondant contient seulement des modifications Vertes et Oranges, avec au moins une modification Orange. Un PSM est Rouge sinon.
Dans les PSM, 29% sont Verts, alors que 25% sont Oranges et 46% sont Rouges (Table 4.4).

Reconstruction d'un modèle du *bait* : le *baitModel*

Transformer les modifications Vertes en résidus permet de reconstruire une séquence, complète ou partielle, qui représente un modèle du *bait*, et qui sera appelée $baitModel$. Des exemples de $baitModels$ sont présentés Table 4.3.
Pour automatiser le processus de reconstruction et pouvoir évaluer l'efficacité de SpecGlob

sur un grand nombre de PSM, nous avons développé une méthode de reconstruction. Pour créer le $baitModel$, le $hitModified$ est parcouru et, si une modification est Verte, le résidu correspondant à la masse de la modification est ajouté, ou bien les résidus sont supprimés s'il s'agit d'une délétion. Si la modification est Orange ou Rouge, la masse correspondante est positionnée à l'endroit où elle doit être insérée.
Même si le $baitModel$ correspond à une séquence incomplète (dans le cas des PSM Rouges ou Oranges), elle peut contenir une ou plusieurs séquences de résidus consécutifs, qui représentent déjà une information utile. De plus, comme montré dans la Table 4.4, plus de 50% des modifications des PSM sont Vertes. De façon à évaluer précisément le gain apporté par les modifications Vertes dans les PSM Oranges et Rouges, la plus longue séquence de résidus a été calculée, pour chaque $baitModel$ correspondant à un $hitModified$. La distribution correspondante peut être visualisée Figure 4.9. L'interprétation des modifications via le $baitModel$ permet de déplacer la distribution vers des valeurs plus importantes, et donc d'augmenter la longueur de la plus longue séquence peptidique connue en passant du $hitModified$ au $baitModel$. En moyenne, les plus longues séquences dans les $hitModifieds$ sont de 5,68 résidus, alors que la longueur va jusqu'à 7,46 dans les $baitModels$; l'interprétation des modifications Vertes permet donc d'augmenter les plus longues séquences de presque deux résidus.
Si l'on se base sur les résultats fournis par `SpecGlob` et la classification en couleurs, et parce que cette étude a été réalisée avec des spectres théoriques, nous pouvons évaluer *a posteriori* si un PSM Vert est correct. En effet, lorsque le $hitModified$ est classé comme Vert, le $baitModel$ correspondant est une séquence peptidique complète qui est supposée représenter la séquence du *bait*. Parmi les 132 137 PSM Verts, une proportion très importante de 97,3% des PSM (128 568) ont un $baitModel$ strictement égal au *bait*. Parmi les quelques cas de désaccord entre le $baitModel$ et le *bait*, nous observons le phénomène suivant : un résidu du *hit* a la même masse que la combinaison de deux résidus. Par exemple, N a la même masse que GG, ou bien Q a la même masse que GA. Dans ce cas, N (ou Q) est considéré comme présent dans le $baitModel$ étant donné que sa masse est retrouvée entre deux masses (non consécutives) du *bait*, même si le résidu n'est pas présent dans le *bait*.
Un autre phénomène, bien que moins fréquent, se produit lorsque le *bait* et le $baitModel$ sont en désaccord : les ions y interfèrent dans l'alignement, et sont par erreur considérés comme des ions b par `SpecGlob` (on rappelle que seuls les ions b du *hit* sont alignés, voir Figure 4.1, page 106). Cela amène `SpecGlob` à signaler des modifications qui sont plus

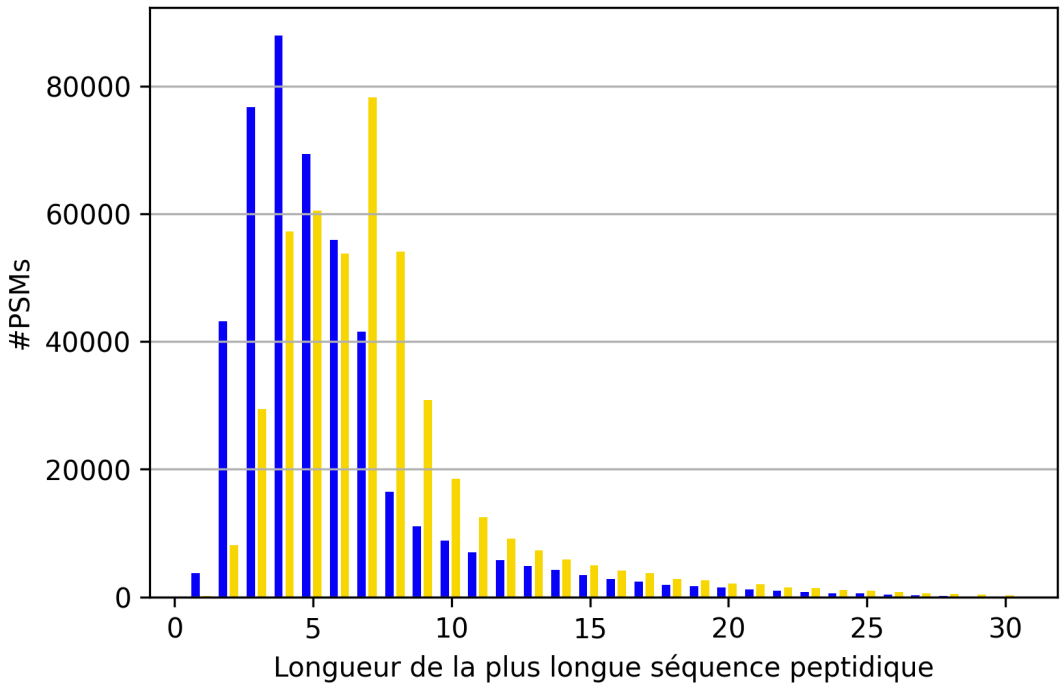


FIGURE 4.9 – **Distribution des plus longues séquences peptidiques dans les résultats de `SpecGlob` avant et après production d'un** *baitModel*. Pour chaque PSM, la longueur de la plus longue séquence de résidu(s) a été calculée dans les $hitModifieds$ (en bleu) et dans les $baitModels$ (en jaune).

tard considérées comme Vertes, mais ne le sont pas en réalité.

J'ai identifié deux raisons principales pour lesquelles un PSM peut être classé comme Rouge. La première est que la génération du $baitModel$ aurait requis une plus importante exploration de la combinatoire des masses que celle que nous avons autorisée (pour rappel, nous avons limité les séquences possibles à une taille d'au plus trois résidus). Ainsi, avec davantage de calculs, ces PSM pourraient être classés comme Orange. L'autre raison est que, à un certain point, l'alignement fourni par `SpecGlob` est simplement faux, à cause de la présence d'ions y dans le *bait* qui ont été alignés par erreur sur les ions b du *hit*. De façon à évaluer l'impact de ces derniers cas, nous avons utilisé `SpecGlob` sur le protéome humain, mais pour ce test particulier, chaque spectre généré à partir du *bait* est représenté uniquement par ses ions b. Les résultats sont présentés Table 4.5 : dans ce cas, même si le nombre de PSM Verts augmente de 7,3%, le pourcentage de PSM Rouges reste toujours important (23,85%). Nous pouvons donc conclure que la catégorie Rouge contient principalement des PSM qui requièrent des opérations complexes pour convertir le *hit* en *bait*,

TABLE 4.5 – **Classification en couleurs des PSM selon les représentations différentes des spectres des** *bait*

Méthode	#PSM Verts	#PSM Oranges	#PSM Rouges	Total
Classique	132 137	114 454	208 813	
Sans ions y	141 786	204 989	108 629	455 404
Sans les masses des ions complémentaires	135 362	148 308	171 734	

La première ligne, appelée *classique*, rappelle les résultats obtenus par la version originale de `SpecGlob` appliquée sur les *baits* non modifiés. Pour la seconde ligne, les *baits* ont été représentés seulement par les masses correspondant à leurs ions b. La troisième ligne montre les résultats obtenus avec la version de `SpecGlob` en deux étapes. Les spectres des *baits* ont été transformés après la première étape de `SpecGlob` : les ions complémentaires à ceux qui ont été alignés au cours de la première étape (supposés être des ions b) sont ainsi supposés être des ions y. De façon à éviter les mauvais alignements entre les ions b et les ions y, leurs masses complémentaires ont été retirées avant la seconde étape de `SpecGlob`. Lorsque la sortie de `SpecGlob` diffère entre la première et la seconde étape, la couleur la plus favorable (Verte, Orange ou Rouge) est choisie.

plutôt que des alignements erronés dûs à la présence de masses qui représentent du bruit. Cette mesure est importante pour la suite de nos travaux sur les spectres expérimentaux qui pourraient contenir une proportion importante de pics de bruits. Compte tenu de la précision des spectromètres de masse, nous émettons l'hypothèse que les alignements par `SpecGlob` ne sont pas trop perturbés par la présence de pics supplémentaires, ce qui devra être vérifié sur des spectres expérimentaux. Cependant, afin de limiter le nombre de masses mal alignées (qui pourrait devenir un problème dans un contexte expérimental, par nature bruité), `SpecGlob` a été modifié de façon à fonctionner en deux étapes. Dans cette seconde implémentation, `SpecGlob` est d'abord utilisé en mode standard. Ensuite, avant la seconde étape, les *baits* sont transformés comme suit : puisque les masses alignées qui sont identifiées sont supposées être principalement des ions b, leurs ions complémentaires (qui sont supposés être des ions y) sont supprimés. Cette nouvelle représentation des *baits* est ensuite utilisée pour faire une seconde itération de `SpecGlob`, et renvoyer le $hitModified$. En utilisant cette version en deux étapes, quelques milliers de PSM supplémentaires sont correctement interprétés, comme montré dans la dernière ligne de Table 4.5.

TABLE 4.6 – **Distribution des PSM dans les trois catégories de couleurs selon le signe de leur Δm.**

	$\Delta m < 0$	$\Delta m = 0$	$\Delta m > 0$
#PSM Verts	105 912	4 246	21 979
#PSM Oranges	51 452	12 952	50 050
#PSM Rouges	74 606	14 512	119 695

Chaque PSM a un Δm qui est la différence de masse entre le *bait* et le *hit*. Les PSM sont classés selon le signe de leur Δm, et ensuite selon la couleur du $hitModified$ de `SpecGlob`, donnée par la méthode de classification.

Il est raisonnable de supposer que le nombre de PSM dans chaque catégorie de couleur dépend fortement de la similarité des peptides impliqués dans un PSM. `SpecGlob` est en effet plus efficace lorsque le *hit* contient une proportion importante des résidus du *bait* et dans un ordre similaire (même si `SpecGlob` est capable de résoudre des permutations simples). De plus, les PSM avec un *hit* avec davantage de résidus qu'un *bait* (et donc un Δm positif) ont plus de chance de pouvoir s'aligner avec les résidus du *bait*, et étant donné qu'une délétion est une modification Verte, ils peuvent être plus facilement interprétés. Afin de confirmer cette hypothèse, les PSM ont été classés par Δm au sein des catégories de couleurs. Les résultats sont présentés Table 4.6.
De nombreux PSM Verts ont un Δm négatif, alors qu'au contraire de nombreux PSM Oranges et Rouges ont un Δm positif. Ainsi, l'une des raisons pour lesquels `SpecGlob` ne peut pas interpréter certains *baits* est l'absence d'une trop grande quantité de résidus dans le *hit*. Le problème peut donc être lié à la base de données de protéines, qui n'est pas représentative des spectres identifiés dans certains cas.
Afin d'évaluer la stabilité de `SpecGlob` par rapport au système de scores, la classification en couleurs des PSM a été calculée avec six autres systèmes de scores. Les résultats sont présentés Table 4.7. Nous pouvons voir que l'évolution du système de scores a peu d'impact sur la qualité des résultats fournis par `SpecGlob`.
Les systèmes de scores choisis favorisent l'insertion de décalages de masse afin d'autoriser

TABLE 4.7 – **Distribution des PSM dans les trois catégories de couleurs selon le système de scores.**

	5/2/-4	5/0/-4	3/2/-4	5/2/-6	5/4/-4	7/2/-4	5/2/-2
#PSM Verts	132 137	132 038	132 133	132 133	132 133	132 133	132 133
#PSM Oranges	114 454	117 475	114 291	114 304	114 291	114 548	114 548
#PSM Rouges	208 813	205 891	208 980	208 967	208 980	208 723	208 723

Chaque colonne correspond à un système de scores s_A / s_R / s_N différent. Le nombre de PSM dans chaque catégorie de couleur est donné pour chacun des systèmes de scores présenté.

les modifications. Cependant, nous avons pu voir qu'effectuer un décalage de masse présente un risque d'aligner des ions y sur des ions b. Dans les spectres théoriques, cela a peu d'impact sur les résultats, mais avec des spectres expérimentaux, qui présentent du bruit, le système de scores devra probablement être choisi de façon à limiter, dans une certaine mesure, l'introduction de décalages de masse.

Validation par la stratégie target/decoy

On rappelle que dans les PSM renvoyés par `SpecOMS`, un certain pourcentage (37,97%) sont des PSM decoy (c'est-à-dire avec un *hit* originaire de la base decoy), et sont supposés représenter des identifications erronées. Le travail présenté dans le Chapitre 3 nous a montré que certains de ces PSM decoy étaient complètement interprétables même en se limitant à une seule modification. Ici, grâce à la capacité de `SpecGlob` à détecter plusieurs modifications au sein d'un PSM, nous pouvons voir à quel point un match considéré comme dû au hasard, c'est-à-dire un PSM decoy, est interprétable ou non avec plusieurs modifications.
Afin de répondre à cette question, nous avons calculé le nombre de PSM target et decoy pour chaque couleur (voir Figure 4.10).
Parmi les 132 137 PSM Verts, il est intéressant de noter que 38 350 *hits* (29%) sont issu de la base decoy. Il est ainsi possible de conclure qu'une proportion importante des PSM est

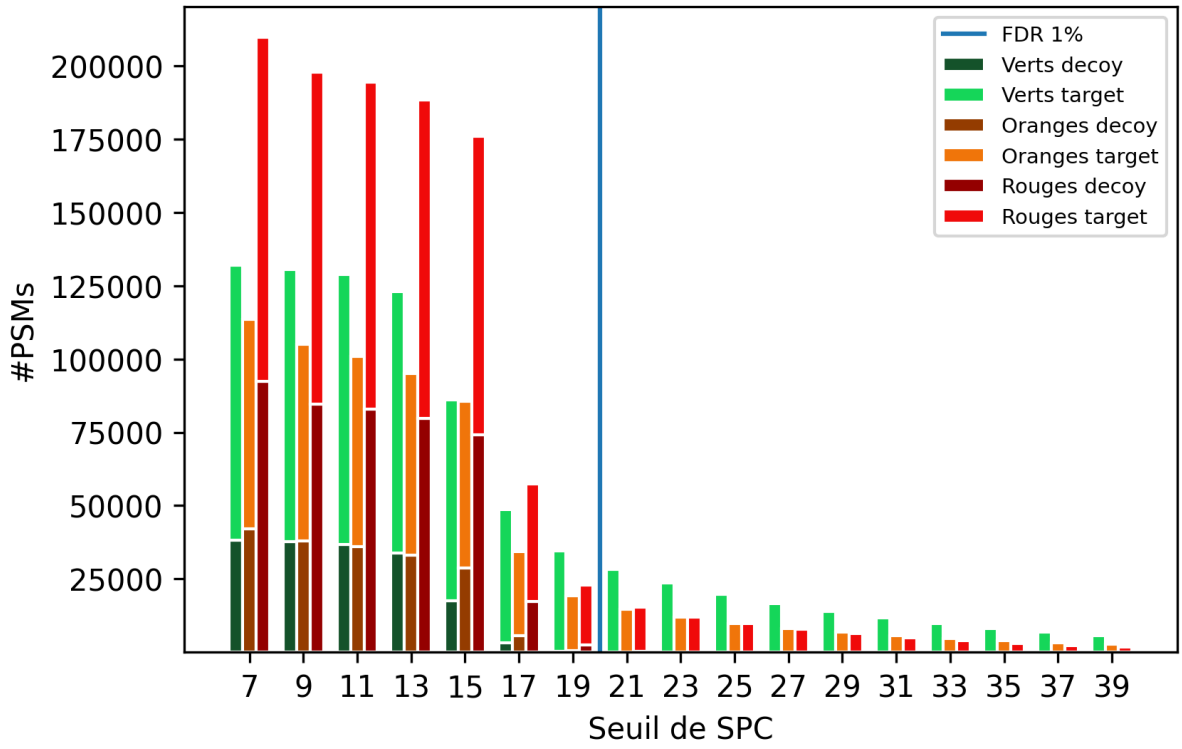


FIGURE 4.10 – **Distribution des PSM target et decoy dans les catégories de couleurs selon le seuil de SPC.** Pour un SPC allant de 7 à 39, tous les PSM avec au moins ce score sont classés dans les trois catégories de couleurs ; pour chaque catégorie de couleur, les PSM sont classés comme target ou decoy selon l'origine du *hit*. Par exemple, pour un seuil de 7, il y a 132 137 PSM Verts (93 787 targets et 38 350 decoys), 114 454 PSM Oranges (71 952 targets et 42 502 decoys) et 208 813 PSM Verts (116 748 targets et 92 065 decoys). La ligne bleue verticale montre le SPC pour lequel on atteint un FDR $< 1\%$. (dans notre cas, 21).

interprétée grâce à des *hits* decoys. Cela peut être considéré comme un problème, étant donné que la base decoy est supposée représenter le hasard. De plus, nous avons montré qu'à un FDR de 1% (SPC de 21), il reste seulement 57 784 PSM qui sont validés (12,6% des 455 404 PSM obtenus avec un seuil de SPC de 7). Cependant, seuls 28 252 (48,9%) de ces PSM validés sont Verts. Cela signifie que, en plus de perdre un grand nombre de PSM, environ la moitié des PSM restants validés par le FDR ne peuvent pas être interprétés sans davantage d'efforts, même si `SpecGlob` est capable d'aligner les spectres de peptides séparés par plusieurs modifications.

4.4.3 Discussion et améliorations possibles

`SpecGlob` aligne des paires de spectres de façon à obtenir des alignements de qualité qui peuvent permettre d'identifier la séquence du peptide du spectre que l'on analyse. Grâce aux spectres théoriques, nous sommes capables de déterminer quels alignements sont de qualité, en ce sens qu'ils permettent de reconstruire la séquence du *bait* à partir de celle du *hit*. `SpecGlob` a montré son efficacité sur les spectres théoriques ; en effet de nombreux PSM renvoyés par `SpecOMS` peuvent être interprétés par `SpecGlob`, même si plusieurs opérations d'édition sont nécessaires pour passer du *hit* au *bait*. De plus, même si le $hitModified$ ne permet pas de connaître la totalité de la séquence du *bait*, il permet souvent d'en restaurer une proportion importante. Nous obtenons alors des sous-séquences, utiles pour l'identification des peptides et des protéines, avec une efficacité qui dépend de leur longueur et de leur composition.

Augmenter l'espace de recherche en termes de candidats est nécessaire pour identifier un spectre modifié. Grâce à la programmation dynamique, `SpecGlob` fait un pas de plus dans cette direction et compare, pour chaque PSM, tous les alignements possibles entre les deux spectres, pour en retenir le meilleur, quel que soit le nombre de modifications et leur nature. Cela permet de réduire les erreurs d'interprétation dues à un *a priori*. La plupart des modifications du jeu de PSM peuvent être sans ambiguïté transformées en séquences et lorsqu'un PSM est classé comme Vert, il peut être interprété dans la grande majorité des cas. Un certain nombre de PSM sont Oranges et Rouges ; cependant, grâce au spectres théoriques, nous avons pu comprendre pourquoi.

Lorsqu'un PSM est classé comme Orange, il pourrait être interprété en testant toutes les combinaisons de séquences, en produisant le spectre correspondant et en choisissant celui qui permet d'obtenir le plus grand nombre de masses en commun entre le *bait* et le $baitModel$. Cette méthode n'a pas été implémentée puisque l'objectif n'était pas d'améliorer l'alignement fourni par `SpecGlob` entre les spectres théoriques, mais d'implémenter `SpecGlob` et de l'évaluer dans l'optique de l'adapter aux spectres expérimentaux. Avec de tels spectres, tester toutes les combinaisons qui correspondent à un décalage de masse donné, sans connaître *a priori* la nature des modifications portées par les peptides, entraînerait un temps d'exécution déraisonnable. Interpréter les PSM Rouges est encore plus compliqué, car soit la combinatoire de masse est encore plus difficile à explorer, soit l'alignement est erroné. Mais nos résultats suggèrent que ce dernier cas est minoritaire.

Filtrer les PSM afin de n'avoir qu'un seul *hit* par *bait* avant l'alignement est une étape délicate, qui peut influencer la classification en couleurs. Nous rappelons que `SpecGlob`

peut prendre en entrée des PSM renvoyés par n'importe quelle méthode OMS, mais qu'on a traité ici des PSM renvoyés par SpecOMS. Or, certains PSM prometteurs pour SpecGlob sont mis de côté par SpecOMS. Par exemple le PSM fictif (DEFGHTQR, **W**DEGHTQ**A**R) pourrait être parfaitement aligné par SpecGlob et serait donc classé comme Vert, car seules deux modifications (une côté N-ter et une avant le dernier résidu) sont nécessaires pour transformer l'un des peptides en l'autre. Malheureusement, les deux spectres correspondants n'ont qu'une seule masse en commun, car les masses de tous les ions b sont décalées, et seule la masse d'un ion y est partagée entre les deux spectres. Ainsi ce PSM ne serait pas renvoyé par SpecOMS qui, comme la plupart des méthodes, se base sur le SPC. À la place de WDEGHTQAR, un autre *hit* aurait été fourni par SpecOMS, qui aurait pu ne pas produire de PSM Vert. SpecOMS est réglé de façon à ne renvoyer qu'un seul candidat par *bait*. Utiliser SpecGlob sur plusieurs candidats (ou *hits*) avant la sélection, sur le score de l'alignement ou bien la couleur par exemple, pourrait certainement augmenter le pourcentage de PSM Verts dans les résultats finaux. Cette possibilité est explorée dans la section suivante.

4.5 Amélioration des interprétations de SpecGlob

4.5.1 Principe

Comme discuté auparavant, l'une des raisons pour lesquelles SpecGlob peut être limité dans son interprétation est la sélection d'un *hit* qui ne contient pas une information suffisante pour interpréter le *bait*. Mais SpecGlob est un algorithme rapide ; nous pouvons donc nous permettre de réaliser davantage d'alignements, entre un *bait* et plusieurs *hits* fournis par une méthode OMS (comme SpecOMS), afin de limiter ce problème. L'idée est la suivante : SpecOMS peut être réglé de façon à fournir non pas un seul *hit* par *bait*, mais tous ceux qui passent le seuil de t (voir Chapitre 3). Nous pouvons ainsi obtenir pour un *bait* plusieurs *hits*, et donc plusieurs $hitModifieds$ et plusieurs $baitModels$. Il est possible que dans certains cas, pour un *bait* donné, aucun $baitModel$ ne corresponde à une séquence complète, mais chaque $baitModel$ permet d'interpréter une partie du *bait*, et donc en combinant leurs informations, une proportion plus importante de la séquence du *bait* est élucidée par rapport à l'information apportée par un seul *hit*.
Un exemple de PSM mieux interprété si plusieurs *hits* sont considérés est présenté Table 4.8. Pour le *bait* IVHNIVEEDR, SpecOMS sélectionne quatre candidats, quatre *hits*, qui dé-

TABLE 4.8 – **Exemple d'un PSM mieux interprété si plusieurs *hits* sont considérés.**

bait	*hit*	*hitModified*	*baitModel*
IVHNIVEEDR	VIQQIVEEDR	V[113,08][I][Q][Q]I[-118,10]VEEDR	IV[251,10]IVEEDR
	VIEVHENIDR	[V]I[-99,07][E]V[-129,04]H[E]N[-129,04]ID[357,15]R	IVHNI[357,15]DR
	IVHAAIAEEIGGPVHAIAIQAR	IVH[A][A]I[-28,03][A]E[28,03]E[I][G][G][P]V[-247,19][H][A][I][A]I[-392,22][Q][A][R]_[-373,21]	IVH[114,04]IVEE[76,99]VI
	IVNEEDR	IVN[137,06]E[212,15]EDR	IVHN[212,15]EEDR

Le *bait* IVHNIVEEDR a quatre *hits* sélectionnés par SpecOMS, avec un seuil de SPC de 7. À partir de chaque PSM, SpecGlob peut produire un $hitModified$ puis un $baitModel$. Aucun $baitModel$ seul ne correspond à une séquence complète, mais si l'on regroupe les informations contenues dans le premier IV[251,10]IVEEDR et le dernier IVHN[212,15]EEDR, nous obtenons une séquence complète IVHNIVEEDR, qui est la séquence du *bait*.

passent le seuil de SPC de 7. Cela forme quatre PSM qui peuvent être fournis à SpecGlob. Nous obtenons donc quatre $hitModifieds$, à partir desquels quatre $baitModels$ sont construits. Parmi les quatre $baitModels$, aucun ne permet de former une séquence complète. Ils contiennent tous une ou plusieurs modifications Oranges ou Rouges. Cependant, si l'on regroupe l'information contenue dans le premier $baitModel$ IV[251,10]IVEEDR et l'information du quatrième IVHN[212,15]EEDR, il est possible de créer une séquence complète. En effet, la masse de 251,10 Da peut correspondre à plusieurs séquences. En alignant cette masse sur IVHN[212,15]EEDR, nous pouvons raisonnablement supposer qu'étant donné que HN a une masse de 251,10 Da, cette séquence est réellement présente dans le *bait*. Dans l'autre sens, la masse de 212,15 Da contenue dans le quatrième PSM s'aligne avec IV du premier PSM. Nous pouvons donc créer la séquence complète IVHNIVEEDR, qui s'avère être la séquence du *bait* après vérification.
Une alternative à cette méthode serait de tester toutes les séquences possibles pour une masse donnée du $baitModel$ avant de réaligner avec le *bait*. Cependant, le nombre de possibilités pour une masse peut être très élevé ; la fusion des $baitModels$ pourrait donc permettre de se focaliser sur les séquences les plus probables grâce à plusieurs candidats. L'exemple présenté Table 4.8 pose la question de l'automatisation du processus. En effet, nous voyons qu'il faut utiliser les *hits* 1 et 4, et non les *hits* 2 et 3, pour réaliser la fusion. La question de la sélection des *hits* à prendre en compte reste pour le moment en suspens. J'ai, pour commencer, résolu un problème plus simple, et développé une méthode pour réunir l'information portée par deux $baitModels$ uniquement. La section suivante explique comment cette méthode a été formalisée.

4.5.2 Test à grande échelle

Afin de voir dans quelle mesure regrouper l'information de plusieurs candidats peut améliorer l'interprétation des spectres, il a fallu mettre au point une méthode de fusion automatique de *baitModels*. Celle-ci prend en entrée deux *baitModels* et, si possible, renvoie un *baitModel* qui résume l'information portée par les deux *baitModels* en entrée, que j'appellerai *consensus*. Le principe de la méthode qui fusionne deux *baitModels* est résumé dans l'Algorithme 6.

Algorithme 6 : Principe de la fusion de deux *baitModels*

```
Entrée : Deux baitModels
Sortie : Un baitModel consensus ou "échec"
pour chaque élément des baitModels lu de gauche à droite faire
    si Les éléments sont les mêmes alors
        Éléments ajoutés au consensus
    sinon
        si Les éléments sont cohérents alors
            #Un ou plusieurs résidu(s) et une masse correspondent
            Résidus ajoutés au consensus
        sinon
            #Fusion impossible
            retourner "échec"
        fin
    fin
fin
retourner consensus
```

Le premier *baitModel* en entrée, *baitModel*1, est pris en référence. Pour chaque élément du *baitModel*1 (résidu ou décalage de masse) de gauche à droite, chaque élément du *baitModel*2 est considéré. S'ils sont égaux, ils sont ajoutés au *consensus*. Sinon, la méthode tente de les aligner. Si tous les éléments peuvent être alignés, un *consensus* est produit. Sinon, l'alignement ne fonctionne pas, et un échec est renvoyé. La fusion du premier et du dernier *baitModels* de la Table 4.8, qui permet d'obtenir un *baitModel consensus* est détaillée Table 4.9. Cependant, il arrive qu'une fusion ne fonctionne pas, comme illustré Table 4.10. Dans ce cas, la méthode de fusion renverra "échec".
Cette méthode peut être utilisée sur un grand nombre de PSM. Nous pouvons grâce à elle mesurer à quel point la classification en couleurs des PSM peut être améliorée, et si la fusion a donné un *baitModel* correct en le comparant au *bait*. Cela est fait selon la méthode décrite dans l'Algorithme 7.
Chaque *bait* a un ou plusieurs *hits* renvoyés par `SpecOMS` avec un $hitModified$ produit par `SpecGlob`. Si le *bait* a plusieurs *hits*, l'information de séquence qu'ils contiennent

TABLE 4.9 – **Exemple d'une fusion qui renvoie un** *baitModel consensus***.**

Elément *baitModel1*	Elément *baitModel2*	*Consensus*
I	I	I
V	V	V
251,10	H	H
	N	N
I	212,15	I
V		V
E	E	E
E	E	E
D	D	D
R	R	R

Le *bait* attendu est IVHNIVEEDR. À chaque étape, la méthode tente d'aligner les deux éléments des deux *baitModels*. Les deux premiers résidus s'alignent, ils sont donc ajoutés au *consensus*. Ensuite, HN s'alignent avec 251,1 Da, HN est donc ajouté au *consensus*. Même chose avec IV qui s'aligne avec 212,15 Da. IV est donc ajouté au *consensus*. Enfin, les quatre derniers résidus sont ajoutés au *consensus*.

TABLE 4.10 – **Exemple d'une fusion qui renvoie un échec.**

Elément *baitModel1*	Elément *baitModel2*	*Consensus*
I	I	I
V	V	V
H	H	H
N	114,04	N
212,15	I	I
	V	V
E	E	E
E	E	E
D	76,99	Échec
	V	

Le *bait* attendu est IVHNIVEEDR. La fusion fonctionne bien jusqu'au résidu D du *baitModel*1, qui ne peut pas s'aligner avec le décalage de 76,99 Da couplé à V. En effet, lorsqu'on retranche à D une masse de 76,99 Da, on obtient une masse de 38,03 Da. Mais si on soustrait à cette masse la masse de V, de 99,08 Da, on obtient une masse négative, et donc l'alignement a échoué. La fusion renverra donc "échec".

Algorithme 7 : Évaluation de la fusion de deux *baitModels*

Entrée : *baits*, *hits*, *hitModifieds*, *baitModels*
echecs, coherence, info ← 0
couleursAvant, couleursApres, meilleuresCouleurs, plusLonguesSequences ← []
#Listes
pour *chaque bait* **faire**
  Classification des *hits* selon le score de SpecGlob
  si *on a plus d'un hit* **alors**
    couleur1 ← couleur(*hitModified*1)
    couleur2 ← couleur(*hitModified*2)
    couleurAvant ← meilleure_couleur(couleur1, couleur2) #Vert>Orange>Rouge
    consensus ← fusion(*hitModified*1, *hitModified*2) # renvoie un consensus (succès) ou un échec
    si *succès* **alors**
      couleurApres ← couleur(*consensus*)
      si *cohérence si on compare le consensus au bait* **alors**
        #Le consensus s'aligne sur le *bait*
        coherence ← coherence + 1
        si *apporte de l'info* **alors**
          #Le consensus est différent des deux *baitModels* en entrée
          info ← info + 1
          #plusLongueSequence() calcule la plus longue séquence de résidus dans le *baitModel*
          seq1 ← plusLongueSequence(*baitModel*1)#premier *baitModel*
          seq2 ← plusLongueSequence(*baitModel*2) #second *baitModel*
          seq ← max(seq1, seq2)
          plusLonguesSequencesAvant ← plusLonguesSequencesAvant + seq
          plusLonguesSequencesApres←plusLonguesSequencesApres+plusLongueSequence(*consensus*)
    sinon
      echecs ← echecs + 1
      couleurApres ← Rouge
  sinon
    couleurAvant ← couleur(*hitModified*)
    couleurApres ← couleurAvant
  couleursAvant ← couleursAvant + couleurAvant
  couleursApres ← couleursApres + couleurApres
  meilleurCouleur ← meilleure_couleur(couleurAvant, couleurApres)
  meilleuresCouleurs ← meilleuresCouleurs + meilleurCouleur
fin
retourner echecs, coherence, info, couleursAvant, couleursApres, meilleuresCouleurs, plusLonguesSequencesAvant, plusLonguesSequencesApres

peut être fusionnée. Le choix des deux $baitModels$ à fusionner est fait selon le score de `SpecGlob`, ou dans l'ordre des PSM renvoyés par `SpecOMS` en cas d'égalité. Nous espérons ainsi sélectionner des *hits* qui permettront d'obtenir une proportion importante de la séquence du *bait*. Si ce *bait* n'a qu'un seul *hit*, sa couleur est prise en compte. Si le *bait* a plus d'un *hit*, la meilleure des couleurs des deux meilleurs *hits* est comptée (Vert est meilleure que Orange qui est meilleure que Rouge). Ces couleurs sont indiquées en haut de la Figure 4.11. Ensuite, si le *bait* a au moins deux *hits*, la fusion est appliquée selon le principe décrit dans l'Algorithme 6. Si la fusion est un échec, l'échec est compté. Sinon, la méthode détermine la couleur du *consensus*. Cela permet de faire une comparaison entre la couleur avant et après fusion ; la meilleure couleur est déterminée et comptée. Si la fusion a fonctionné, il est possible de vérifier si le *consensus* est cohérent avec le *bait*, c'est-à-dire s'il s'y aligne bien ; cette vérification peut être effectuée par la méthode fusion elle-même. Si c'est le cas, nous pouvons vérifier si la fusion a apporté de l'information, c'est-à-dire si le *consensus* n'est pas égal à l'un des deux $baitModels$ qui ont été fusionnés. Pour un compromis en termes de temps de calcul et de précision des résultats, la méthode a été testée sur les 100 000 premiers *baits* renvoyés par `SpecOMS` sur le protéome humain (voir Chapitre 3, Section 3.1.2, page 67). Les résultats sont présentés Figure 4.11.
Sur les 100 000 *baits*, 1 759 ne peuvent pas subir de fusion car ils n'ont qu'un seul *hit*. 61 986 ont une fusion qui ne fonctionne pas (l'algorithme de fusion renvoie "échec"). Si nous considérons la meilleure couleur entre avant et après fusion, celle-ci permet, sur les 100 000 *baits*, de passer de 27 067 à 32 567 PSM Verts. 24 857 PSM ont un *consensus* cohérent avec le *bait*, dont 8 699 apportent davantage d'information par rapport aux deux $hitModifieds$ en entrée. De plus, la fusion permet d'augmenter la taille moyenne de la plus longue séquence de résidus retrouvée, qui passe de 6,94 à 8,49 résidus.
Fusionner les $baitModels$, même sur seulement deux candidats par *bait*, donne donc des résultats prometteurs pour élucider une proportion plus importante de la séquence des *baits*. La prochaine étape consisterait à adapter cette méthode à plusieurs candidats afin d'augmenter encore le nombre de séquences identifiées.

4.6 Conclusion

L'étude des PSM renvoyés par `SpecOMS` à partir de spectres théoriques (Chapitre 3) suggérait que dans de nombreux PSM, plusieurs modifications séparaient le *bait* et le *hit*. Ce résultat motive le développement de méthodes capables d'indiquer, au sein d'un PSM

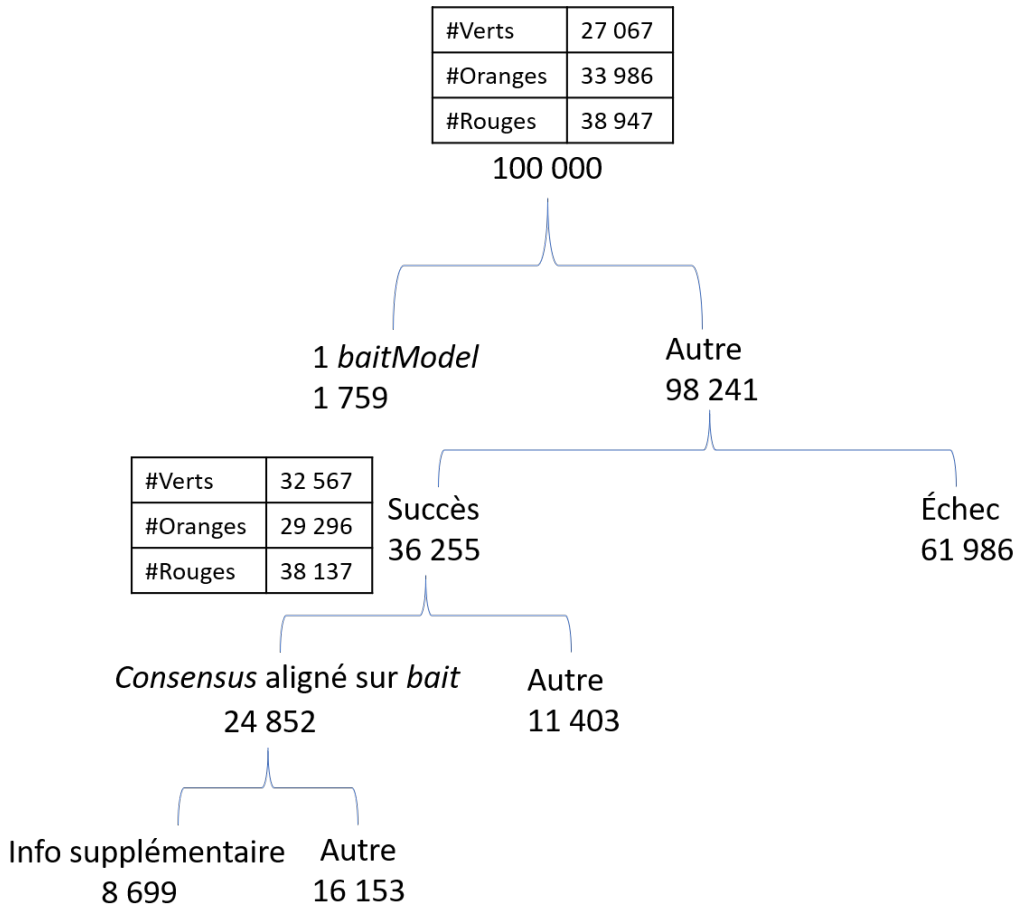


FIGURE 4.11 – **Résultats sur la fusion de 100 000** *baits***.** Si un *bait* a au moins deux *hits*, les deux *baitModels* correspondants peuvent subir la méthode de fusion décrite Algorithme 6. La fusion renvoie soit un *consensus*, soit un échec. Si nous obtenons un succès, nous pouvons vérifier s'il n y a pas d'erreur, c'est-à-dire si le *consensus* s'aligne bien sur le *bait*. Enfin, nous pouvons voir si le *consensus* permet d'obtenir une information de séquence plus importante que les deux *baitModels* d'entrée pris séparément, c'est-à-dire que le *consensus* n'est pas égal à l'un d'entre eux. La classification en couleurs est également appliquée à deux étapes : avant la fusion (nous prenons en compte la couleur du $hitModified$ si le *bait* n'a qu'un *hit* et la meilleure des deux premiers selon le score de `SpecGlob` sinon) et après si la fusion renvoie un *consensus*. Dans ce dernier cas, la meilleure couleur entre avant et après la fusion est indiquée.

donné, quelles modifications sont présentes, autrement dit comment modifier le *hit* afin de retrouver la séquence du *bait*.
Les outils existants limitent la prise en compte des modifications, dans leur nombre et leur nature. À cause des nombreuses modifications inconnues qui peuvent être présentes dans les échantillons, il est nécessaire de développer un outil capable de s'affranchir de ces limites. C'est pour cela que `SpecGlob` a été développé. Il repose sur la programmation dynamique pour aligner les spectres correspondant au *bait* et au *hit*. En alignant les masses du *hit* sur celles du *bait* et en autorisant des décalages, les modifications qui séparent les deux peptides à l'origine des spectres peuvent être représentées sous la forme d'un $hitModified$. Ce résultat est une chaîne de caractères, donc facilement manipulable pour des analyses qui, grâce aux spectres théoriques, indiquent à quel point `SpecGlob` est efficace.
D'abord, `SpecGlob` a prouvé son efficacité par rapport à un outil existant, `MODPlus`, avec des spectres simulés "parfaits". En modifiant systématiquement les peptides théoriques sur certains résidus et en les identifiant sur leur base de données d'origine par ces outils, `SpecGlob` permet d'identifier davantage de modifications que `MODPlus` lorsque celles-ci ne sont pas représentées dans Unimod.
`SpecGlob` a ensuite pu être testé sur les résultats de `strat-shift`. Une méthode a été mise au point pour 1) classer les modifications en trois couleurs selon leur niveau d'interprétabilité 2) transformer ces modifications en séquences de résidus correspondantes, lorsque c'est possible, sous la forme d'un $baitModel$.
Sur le jeu de spectres théoriques produits par `SpecOMS`, `SpecGlob` permet de reconstruire complètement la séquence de nombreux *baits*. Et même si la séquence n'est pas complètement reconstruite, la majorité des modifications sont complètement interprétables. Ainsi, une certaine proportion de la séquence peut être révélée, et permettre ainsi une identification du peptide.
Une fois les séquences reconstruites, la connaissance de la séquence réelle du *bait* permet de vérifier l'alignement. Dans l'immense majorité des cas, la séquence reconstruite correspond bien au *bait*.
L'efficacité de `SpecGlob` peut être améliorée principalement à deux niveaux. D'abord, nous avons pu voir que les ions y du *bait* peuvent interférer avec l'alignement. Nous avons également pu voir qu'il était possible de résoudre une partie de ce problème en tentant de détecter et éliminer les ions y.
Une seconde limitation est liée aux PSM eux-mêmes. Pour un *bait* donné, si un seul *hit* est

sélectionné, il peut ne pas contenir une information suffisante pour interpréter son *bait*. Dans cette situation, si c'est possible, il est intéressant d'utiliser l'information présente dans plusieurs *hits* et d'accumuler l'information présente dans plusieurs $baitModels$ pour avoir une interprétation plus précise d'un *bait*. J'ai pu implémenter cette méthode pour la fusion de deux $baitModels$, et montrer qu'une proportion plus importante de PSM Verts pouvait être atteinte de cette façon.

Le filtre par FDR des PSM avant un alignement est également discutable, dans cette étude théorique où les PSM contiennent plusieurs modifications. Des PSM complètement interprétables par `SpecGlob`, comme par exemple le PSM (GVPTEVK, GDPITVK), qui existe dans les résultats de `SpecOMS`, sont rejetés à cause d'un SPC bas (8 dans l'exemple) qui ne passe pas le seuil de FDR de 1%. Comme montré Figure 4.10, un grand nombre de PSM Verts viennent de la base decoy. Cela signifie qu'un *bait* de la base target peut avoir un *hit* qui vient de la base decoy, et quand même résulter en une interprétation complète par `SpecGlob`, en dépit de plusieurs modifications et d'un SPC relativement bas. Ainsi, même si augmenter le seuil réduit le FDR, de nombreux PSM intéressants sont rejetés pendant cette étape de filtrage. Ces résultats laissent donc penser que le filtre FDR basé sur un score n'est pas approprié avant d'appliquer `SpecGlob`. Une estimation statistique de la fiabilité de l'alignement final serait plus appropriée. Des estimations comme celles des outils `PTMProphet` [D. D. Shteynberg *et al.* 2019] ou `PTMiner` [An *et al.* 2019] pourraient être très utiles dans ce cas. Cependant, cette question est en dehors du cadre de cette étude, dont l'objectif était d'évaluer la qualité des résultats.

La manière dont le $hitModified$ est exprimé est utile en elle-même pour décrire de façon compacte les alignements incomplets, car l'information qu'elle contient va au-delà de la localisation des modifications, en mettant en avant les séquences de résidus détectés dans le spectre. Une alternative est de visualiser graphiquement les spectres avec les masses annotées, ce qui donne accès à plus de détails, mais est très chronophage pour l'utilisateur. De plus, les $hitModifieds$, qui sont des chaînes de caractères, sont faciles à manipuler par des scripts variés, ce qui permet de rassembler des informations pour répondre à certaines questions, comme je l'ai fait en les classant en trois couleurs ou bien en générant les $baitModels$.

Les résultats présentés ici sont très encourageants pour l'adaptation de `SpecGlob` aux spectres expérimentaux. En effet, les erreurs d'alignement induites par les ions y ont été prises en compte dans la conception des *baits*, erreurs dont l'importance peut être réduite en détectant et supprimant les ions y. De plus, il est clair que certains ions b peuvent

manquer dans un spectre expérimental. Cependant, il est possible de prendre en compte cette limite et de générer les ions complémentaires des ions du spectre. Ainsi, si un ion b est manquant mais que son ion y complémentaire est présent, l'ion b sera généré et pourra intervenir dans l'alignement.
`SpecGlob` a déjà été adapté aux spectres expérimentaux à l'INRAE, et les premiers tests sont prometteurs. L'ensemble de ce travail est donc encourageant pour poursuivre le développement de `SpecGlob` et son utilisation sur de véritables spectres.

Conclusion et perspectives

Pour identifier précisément les protéines par spectrométrie de masse (MS), il est nécessaire de pouvoir identifier les spectres MS2 issus de peptides modifiés, c'est-à-dire qui portent des PTM ou bien des modifications de séquence par rapport à leur candidat le plus proche d'une base de données. Pour cette identification, il est possible d'utiliser les méthodes OMS (*Open Mass Search*) qui autorisent une différence de masse parente (Δm) importante lors de la comparaison entre un spectre expérimental à identifier et un spectre théorique généré à partir d'un peptide candidat d'une base de données. Ces méthodes doivent mettre en œuvre des stratégies variées pour 1) sélectionner un peptide candidat proche de celui qui a généré le spectre à identifier, c'est-à-dire générer un PSM ; 2) identifier et localiser les modifications éventuelles qui séparent les deux peptides impliqués dans un PSM, et donc interpréter le PSM.
Pour ce faire, les méthodes OMS sont confrontées à de nombreux défis, notamment lorsque les spectres de masses correspondent à des peptides qui portent plusieurs modifications inconnues. Afin d'améliorer l'identification des peptides modifiés, les différentes stratégies OMS méritent d'être étudiées en profondeur pour mieux les comprendre et donc mieux les utiliser, et les améliorer.
Pour mon travail de thèse, j'ai choisi de me placer dans un contexte où le rôle des spectres expérimentaux est joué par des spectres théoriques. Cela permet de mettre à l'épreuve les stratégies d'identification étudiées en favorisant la présence de PSM modifiés, par exemple en empêchant l'auto-identification, ou encore en ajoutant des modifications sur les peptides avant de générer les spectres correspondants. Puisque les séquences des deux peptides impliqués dans un PSM sont connues, les identifications peuvent être évaluées précisément ; il est donc possible de déterminer à quel point les stratégies ont été efficaces pour sélectionner les candidats et interpréter les modifications dans les PSM. Nous pouvons ainsi mettre en évidence des phénomènes difficilement détectables avec des spectres expérimentaux.
`SpecOMS` est capable de calculer rapidement le SPC entre tous les spectres d'une base de données, sans limite de Δm. De plus, il peut utiliser le Δm pour déterminer un nouveau score et localiser une potentielle modification. Ce logiciel d'identification semble donc être

adapté pour avoir un premier aperçu des identifications entre spectres théoriques.

Vers un réseau de spectres expérimentaux

Nous avons obtenu avec `SpecOMS` un ensemble de PSM (*bait*, *hit*) produits à partir des peptides issus du protéome humain (Chapitre 3). Ces résultats peuvent être modélisés sous la forme d'un réseau dans lequel les peptides sont séparés par au moins une modification de séquence, ce qui offre donc une cartographie de peptides reliés par la MS. Le réseau des peptides semble présenter, à l'instar de nombreux réseaux retrouvés dans la nature, des propriétés sans échelle, avec peu de peptides très connectés et de nombreux peptides peu connectés. De plus, la densité des connexions paraît reliée à l'origine (base target ou decoy) des peptides. Étudier de plus près les propriétés du réseau pourrait permettre de mieux déterminer la qualité des identifications lorsque les nœuds représenteront des spectres expérimentaux ; nous pourrions en effet obtenir certaines informations selon l'environnement (par exemple la densité) du spectre dans le réseau.
Ce travail sur le réseau contient donc de nombreuses informations qui, au vu de ces études préliminaires, devront être explorées davantage. Cependant, afin de recentrer mon travail, je me suis concentrée sur l'exploration des données plus réduites.

De l'interprétation d'une à plusieurs modifications

Les méthodes OMS peuvent être caractérisées notamment par leur utilisation ou non du Δm pour sélectionner le meilleur candidat pour un spectre. Ces deux stratégies ont été implémentées dans `SpecOMS`. Pour comparer les stratégies, j'ai pu développer des critères utilisant les informations disponibles dans les spectres théoriques. L'analyse des PSM renvoyés par ces deux stratégies grâce à ces critères a d'abord permis de conclure qu'utiliser le Δm permet d'obtenir davantage de PSM dont l'information permet plus facilement de connaître la séquence du *bait*, et semble donc être la stratégie à privilégier. À ce stade, l'utilisation de spectres théoriques paraît donc déjà très prometteuse pour évaluer les résultats de méthodes OMS.
Une autre information donnée par ce travail est que nous obtenons au final un grand nombre de PSM dont la séquence ne peut pas être retrouvée. Une hypothèse est que ces PSM contiennent plusieurs modifications, ce qui rend leur interprétation inaccessible par `SpecOMS`.
Afin d'être capable d'interpréter les PSM contenant plusieurs modifications sans *a priori*

sur leur nombre et leur nature, nous avons développé SpecGlob, qui repose sur la programmation dynamique pour aligner les résidus du *hit* sur son *bait* dans un PSM (Chapitre 4). Les spectres théoriques ont permis de tester l'efficacité de SpecGlob. Tout d'abord, nous avons comparé SpecGlob à MODPlus. Pour cela, des modifications ont été intégrées aux spectres théoriques avant l'étape d'identification par SpecOMS et SpecGlob d'un côté, et MODPlus de l'autre. Ces modifications ont permis de mettre en évidence les différences entre SpecGlob et MODPlus. D'abord, dans le contexte des spectres théoriques, MODPlus est meilleur pour interpréter les PSM contenant des modifications déjà répertoriées, alors que SpecGlob est plus efficace pour les modifications non répertoriées. Ensuite, SpecOMS/SpecGlob a un temps d'exécution bien plus faible que celui de MODPlus. Afin d'aller plus loin dans les capacités d'interprétation de SpecGlob, celui-ci a été appliqué sur les résultats de SpecOMS lorsque tous les spectres théoriques issus du protéome humain sont comparés les uns aux autres. Une classification a pu être développée pour caractériser les résultats de SpecGlob. Il en ressort que SpecGlob permet d'interpréter complètement de nombreux PSM. De plus, même si SpecGlob ne peut pas interpréter complètement certains PSM, il est capable de déterminer une certaine proportion de la séquence du *bait* ; les séquences obtenues, même partielles, peuvent être utiles à l'identification des peptides et des protéines. SpecGlob a donc pu montrer son efficacité en termes de temps de calcul et d'efficacité dans le contexte théorique.

À ce stade, SpecGlob n'était utilisé qu'avec un seul *hit* par *bait* ; ainsi, si le *hit* n'est pas suffisant pour couvrir une partie importante de la séquence du *bait*, celle-ci ne pourra pas être déterminée. Pour limiter ce phénomène, j'ai travaillé sur la fusion de plusieurs $baitModels$ afin d'améliorer l'interprétation d'un *bait* donné. Il est en effet probable que si plusieurs *hits* sont disponibles pour un *bait* donné (par exemple grâce à SpecOMS), même si aucun *hit* (et donc aucun $baitModel$) ne permet d'interpréter complètement le *bait*, il existe plusieurs $baitModels$ qui permettent de déterminer une partie différente de la séquence du *bait*. Dans cette situation, ces informations méritent d'être accumulées. Les premiers résultats, avec le développement d'une méthode qui fusionne deux $baitModels$, montrent qu'il serait intéressant de fusionner un plus grand nombre de $baitModels$ pour interpréter un *bait* donné, et à essayer de résoudre les questions que cela soulève (quels $baitModels$ fusionner ? dans quel ordre ? etc.).

Vers une adaptation de `SpecGlob` aux spectres expérimentaux

Dans le travail décrit ici, afin d'avoir une vision précise de l'efficacité d'une stratégie donnée, les spectres théoriques sont des simplifications des spectres expérimentaux ; ils ne comportent ni bruit, ni pics manquants. Cependant, tout au long de ce travail, il était nécessaire de garder à l'esprit que les méthodes d'analyse et d'identification développées devaient être applicables aux caractéristiques des spectres expérimentaux. Ainsi certaines précautions ont été prises au cours du développement des outils ; par exemple, dans `SpecGlob`, les ions b et y sont pris en compte dans le *bait* afin qu'il se rapproche d'un spectre expérimental. Puisque les résultats de `SpecGlob` poussent à l'utiliser sur des spectres expérimentaux, `SpecGlob` continue à être développé à l'INRAE de Nantes afin de l'améliorer, de permettre une meilleure application de l'outil à des spectres expérimentaux, et de le rendre plus convivial à travers le développement d'une interface graphique. Le logiciel a pu été testé, entre autres, sur des collections de spectres issus d'échantillons d'ovalbumine, plus ou moins chauffés avec ou sans l'addition de glucose. Les résultats obtenus ont permis de produire l'hypothèse que des sites de modifications sur cette protéine sont en lien avec son allergenicité ; cette étude a été réalisée par Mehdi Cherkaoui (post-doctorant) et a fait l'objet d'une présentation lors de la convention de l'ASMS 2022. Ainsi, tout le travail d'élaboration et d'amélioration de l'algorithme réalisé sur les spectres théoriques est pertinent et la simplification réalisée dans la modélisation des spectres n'est pas incompatible avec l'adaptabilité des résultats aux spectres expérimentaux.

Vers un réseau d'identifications par `SpecGlob`

En permettant de découvrir tout ou une partie de la séquence d'un spectre à l'aide d'un peptide, `SpecGlob` offre des perspectives prometteuses pour identifier un grand nombre de spectres. On pourrait en effet imaginer l'appliquer sur un ou plusieurs peptides candidats pour identifier un spectre donné. Ces PSM pourraient ensuite être vus comme un réseau dans lequel un spectre expérimental, une fois identifié, pourrait lui-même jouer le rôle de peptide candidat ; on pourrait ainsi, grâce à `SpecGlob`, propager les identifications de spectre en spectre en tirant partie du nombre de spectres dans le réseau.

Des critères complémentaires au FDR

L'utilisation de spectres théoriques a permis d'établir de nouveaux critères pour évaluer la qualité des PSM. Nous avons pu mesurer à quel point ces critères étaient corrélés ou

non avec le FDR, afin de juger de sa pertinence dans notre contexte. De nombreux PSM complètement interprétables ont des identifications dans la base decoy. De plus, augmenter le seuil de score jusqu'à atteindre un FDR de 1% écarte de nombreux PSM interprétables. Ces résultats semblent montrer que, dans un cadre OMS, des méthodes différentes du FDR sont nécessaires pour juger de la qualité des résultats.

Une mise en valeur des méthodes OMS par des analyses théoriques contrôlées

D'une façon générale, nous avons montré que l'utilisation de spectres théoriques avec un contrôle total sur les identifications offre une vision précise des résultats renvoyés par une méthode OMS. En effet, en connaissant les séquences des peptides impliqués dans un PSM, il est possible de les examiner, de comprendre la nature du lien entre ces deux peptides et de mettre au point des critères originaux pour les caractériser. Certains de ces critères pourraient être adaptés à l'analyse de PSM expérimentaux. Le travail réalisé pendant cette thèse met en valeur des méthodes OMS pour sélectionner et interpréter les PSM, mais nous avons aussi vu qu'un environnement théorique pouvait se montrer utile pour développer un outil complet (`SpecGlob`) et le tester avant de le partager à la communauté.

BIBLIOGRAPHIE

AEBERSOLD, R., J. N. AGAR *et al.* (2018), « How many human proteoforms are there ? », *in* : *Nature Chemical Biology* 14.*3*, p. 206-214, DOI : 10.1038/nchembio.2576 (cité page 19).

AEBERSOLD, R. et M. MANN (2016), « Mass-spectrometric exploration of proteome structure and function », *in* : *Nature* 537.*7620*, p. 347-355, DOI : 10.1038/nature19949 (cité page 2).

ALHAJJ, M. et A. FARHANA (2022), « Enzyme Linked Immunosorbent Assay », *in* : *StatPearls*, Treasure Island (FL) : StatPearls Publishing (cité page 23).

AN, Z. *et al.* (2019), « PTMiner : Localization and Quality Control of Protein Modifications Detected in an Open Search and Its Application to Comprehensive Post-translational Modification Characterization in Human Proteome », *in* : *Molecular & Cellular Proteomics : MCP* 18.*2*, p. 391-405, DOI : 10.1074/mcp.RA118.000812 (cité pages 100, 141).

ASLAM, B. *et al.* (2017), « Proteomics : Technologies and Their Applications », *in* : *Journal of Chromatographic Science* 55.*2*, p. 182-196, DOI : 10.1093/chromsci/bmw167 (cité pages 2, 21).

BANDEIRA, N. *et al.* (2007), « Protein identification by spectral networks analysis », *in* : *Proceedings of the National Academy of Sciences of the United States of America* 104.*15*, p. 6140-6145, DOI : 10.1073/pnas.0701130104 (cité page 72).

BARABÁSI, A.-L. (2009), « Scale-Free Networks : A Decade and Beyond », *in* : *Science* 325.*5939*, p. 412-413, DOI : 10.1126/science.1173299 (cité page 74).

BARABÁSI, A.-L. et R. ALBERT (1999), « Emergence of scaling in random networks », *in* : *Science (New York, N.Y.)* 286.*5439*, p. 509-512, DOI : 10.1126/science.286.5439.509 (cité page 73).

BARABÁSI, A.-L. et E. BONABEAU (2003), « Scale-free networks », *in* : *Scientific American* 288.*5*, p. 60-69, DOI : 10.1038/scientificamerican0503-60 (cité page 74).

BEDFORD, M. T. et S. G. CLARKE (2009), « Protein Arginine Methylation in Mammals : Who, What, and Why », *in* : *Molecular Cell* 33.*1*, p. 1-13, DOI : 10.1016/j.molcel.2008.12.013 (cité page 16).

Bellman, R. (1952), « On the Theory of Dynamic Programming », *in* : *Proceedings of the National Academy of Sciences* 38.*8*, Publisher : National Academy of Sciences Section : Mathematics, p. 716-719, doi : 10.1073/pnas.38.8.716 (cité page 58).

Berman, H. M. *et al.* (2000), « The Protein Data Bank », *in* : *Nucleic Acids Research* 28.*1*, p. 235-242, doi : 10.1093/nar/28.1.235 (cité page 26).

Bittremieux, W., K. Laukens et W. S. Noble (2019), « Extremely Fast and Accurate Open Modification Spectral Library Searching of High-Resolution Mass Spectra Using Feature Hashing and Graphics Processing Units », *in* : *Journal of Proteome Research* 18.*10*, p. 3792-3799, doi : 10.1021/acs.jproteome.9b00291 (cité pages 57, 78).

Bittremieux, W., P. Meysman *et al.* (2018), « Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing », *in* : *Journal of Proteome Research* 17.*10*, p. 3463-3474, doi : 10.1021/acs.jproteome.8b00359 (cité pages 57, 78).

Bogdanow, B., H. Zauber et M. Selbach (2016), « Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides », *in* : *Molecular & Cellular Proteomics : MCP* 15.*8*, p. 2791-2801, doi : 10.1074/mcp.M115.055103 (cité page 51).

Braconi, D. *et al.* (2018), « Foodomics for human health : current status and perspectives », *in* : *Expert Review of Proteomics* 15.*2*, p. 153-164, doi : 10.1080/14789450.2018.1421072 (cité page 2).

Brodbelt, J. S. et D. H. Russell (2015), « Focus on the 20-Year Anniversary of SEQUEST », *in* : *Journal of the American Society for Mass Spectrometry* 26.*11*, p. 1797-1798, doi : 10.1007/s13361-015-1264-1 (cité page 42).

Burgess, R. R. (2018), « A brief practical review of size exclusion chromatography : Rules of thumb, limitations, and troubleshooting », *in* : *Protein Expression and Purification* 150, p. 81-85, doi : 10.1016/j.pep.2018.05.007 (cité page 22).

Burke, M. C. *et al.* (2017), « The Hybrid Search : A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics », *in* : *Journal of Proteome Research* 16.*5*, p. 1924-1935, doi : 10.1021/acs.jproteome.6b00988 (cité pages 58, 78).

Chen, C., H. Huang et C. H. Wu (2017), « Protein Bioinformatics Databases and Resources », *in* : *Protein Bioinformatics*, sous la dir. de C. H. Wu, C. N. Arighi et K. E. Ross, t. 1558, Series Title : Methods in Molecular Biology, New York, NY : Springer New York, p. 3-39, doi : 10.1007/978-1-4939-6783-4_1 (cité page 25).

Cherepanova, N., S. Shrimal et R. Gilmore (2016), « N-linked glycosylation and homeostasis of the endoplasmic reticulum », *in* : *Current opinion in cell biology* 41, p. 57-65, doi : 10.1016/j.ceb.2016.03.021 (cité page 17).

Chick, J. *et al.* (2015), « An ultra-tolerant database search reveals that a myriad of modified peptides contributes to unassigned spectra in shotgun proteomics », *in* : *Nature biotechnology* 33.*7*, p. 743-749, doi : 10.1038/nbt.3267 (cité page 56).

Chick, J. M. *et al.* (2015), « A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides », *in* : *Nature Biotechnology* 33.*7*, p. 743-749, doi : 10.1038/nbt.3267 (cité page 51).

Clauser, K. R., P. Baker et A. L. Burlingame (1999), « Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching », *in* : *Analytical Chemistry* 71.*14*, p. 2871-2882, doi : 10.1021/ac9810516 (cité page 33).

Cliquet, F. *et al.* (2009), « Comparison of Spectra in Unsequenced Species », *in* : *4th Brazilian Symposium on Bioinformatics (BSB 2009)*, Lecture Notes in BioInformatics (LNBI), Issue : 5576, Porto Alegre, Brazil, p. 24-35, doi : 10.1007/978-3-642-03223-3_3 (cité page 54).

Cox, J. et M. Mann (2007), « Is Proteomics the New Genomics ? », *in* : *Cell* 130.*3*, p. 395-398, doi : 10.1016/j.cell.2007.07.032 (cité page 19).

Craig, R., J. C. Cortens *et al.* (2006), « Using Annotated Peptide Mass Spectrum Libraries for Protein Identification », *in* : *Journal of Proteome Research*, p. 7, doi : 10.1021/pr0602085 (cité page 47).

Craig, R. et R. C. Beavis (2003), « A method for reducing the time required to match protein sequences with tandem mass spectra », *in* : *Rapid Communications in Mass Spectrometry* 17.*20*, p. 2310-2316, doi : 10.1002/rcm.1198 (cité page 53).

— (2004), « TANDEM : matching proteins with tandem mass spectra », *in* : *Bioinformatics (Oxford, England)* 20.*9*, p. 1466-1467, doi : 10.1093/bioinformatics/bth092 (cité page 53).

Creasy, D. M. et J. S. Cottrell (2004), « Unimod : Protein modifications for mass spectrometry », *in* : *PROTEOMICS* 4.*6*, p. 1534-1536, doi : 10.1002/pmic.200300744 (cité pages 26, 55, 100, 118).

Cunningham, F. *et al.* (2021), « Ensembl 2022 », *in* : *Nucleic Acids Research* 50.*D1*, p. D988-D995, doi : 10.1093/nar/gkab1049 (cité page 26).

Curto, P. *et al.* (2019), « A Pathogen and a Non-pathogen Spotted Fever Group Rickettsia Trigger Differential Proteome Signatures in Macrophages », *in* : *Frontiers in Cellular and Infection Microbiology* 9, p. 43, doi : 10.3389/fcimb.2019.00043 (cité page 33).

Dancík, V. *et al.* (1999), « De novo peptide sequencing via tandem mass spectrometry », *in* : *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 6.*3-4*, p. 327-342, doi : 10.1089/106652799318300 (cité page 41).

David, M., G. Fertin, H. Rogniaux *et al.* (2017), « SpecOMS : A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes », *in* : *Journal of Proteome Research* 16.*8*, Publisher : American Chemical Society, p. 3030-3038, doi : 10.1021/acs.jproteome.7b00308 (cité pages 4, 57, 65, 66).

David, M., G. Fertin et D. Tessier (2016), « SpecTrees : An Efficient Without a Priori Data Structure for MS/MS Spectra Identification », *in* : *Algorithms in Bioinformatics*, Lecture Notes in Computer Science, p. 65-76, doi : 10.1007/978-3-319-43681-4_6 (cité pages 57, 65).

Desiere, F. *et al.* (2004), « Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry », *in* : *Genome Biology*, p. 12, doi : 10.1186/gb-2004-6-1-r9 (cité page 46).

Deutsch, E. W. *et al.* (2015), « Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics », *in* : *PROTEOMICS - Clinical Applications* 9.*7-8*, p. 745-754, doi : 10.1002/prca.201400164 (cité page 49).

Devabhaktuni, A. *et al.* (2019), « TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets », *in* : *Nature Biotechnology* 37.*4*, p. 469-479, doi : 10.1038/s41587-019-0067-5 (cité pages 56, 59).

Edman, P. (1949), « A method for the determination of amino acid sequence in peptides », *in* : *Archives of Biochemistry* 22.*3*, p. 475 (cité page 24).

Elias, J. E. et S. P. Gygi (2007), « Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry », *in* : *Nature Methods* 4.*3*, p. 207-214, doi : 10.1038/nmeth1019 (cité page 48).

— (2010), « Target-decoy search strategy for mass spectrometry-based proteomics », *in* : *Methods in Molecular Biology (Clifton, N.J.)* 604, p. 55-71, doi : 10.1007/978-1-60761-444-9_5 (cité page 48).

Eng, J. K., B. Fischer *et al.* (2008), « A Fast SEQUEST Cross Correlation Algorithm », *in* : *Journal of Proteome Research* 7.*10*, p. 4598-4602, doi : 10.1021/pr800420s (cité page 44).

Eng, J. K., T. A. Jahan et M. R. Hoopmann (2013), « Comet : An open-source MS/MS sequence database search tool », *in* : *Proteomics* 13.*1*, p. 22-24, doi : 10.1002/pmic.201200439 (cité page 44).

Eng, J. K., A. L. McCormack et J. R. Yates (1994), « An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database », *in* : *Journal of the American Society for Mass Spectrometry* 5.*11*, p. 976-989, doi : 10.1016/1044-0305(94)80016-2 (cité page 42).

Fenn, J. B. *et al.* (1989), « Electrospray Ionization for Mass Spectrometry of Large Biomolecules », *in* : *Science* 246.*4926*, p. 64-71, doi : 10.1126/science.2675315 (cité page 32).

Frewen, B. E. *et al.* (2006), « Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries », *in* : *Analytical Chemistry* 78.*16*, p. 5678-5684, doi : 10.1021/ac060279n (cité page 47).

Garavelli, J. S. (2004), « The RESID Database of Protein Modifications as a resource and annotation tool », *in* : *Proteomics* 4.*6*, p. 1527-1533, doi : 10.1002/pmic.200300777 (cité page 27).

Geiszler, D. J. *et al.* (2021), « PTM-Shepherd : Analysis and Summarization of Post-Translational and Chemical Modifications From Open Search Results », *in* : *Molecular & Cellular Proteomics* 20, p. 100018, doi : 10.1074/mcp.TIR120.002216 (cité page 100).

Griffiths, J. (2008), « A Brief History of Mass Spectrometry », *in* : *Analytical Chemistry* 80.*15*, p. 5678-5683, doi : 10.1021/ac8013065 (cité page 31).

Griss, J. *et al.* (2016), « Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets », *in* : *Nature Methods* 13.*8*, p. 651-656, doi : 10.1038/nmeth.3902 (cité page 51).

Henzel, W. J. *et al.* (1993), « Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. », *in* : *Proceedings of the National Academy of Sciences* 90.*11*, p. 5011-5015, doi : 10.1073/pnas.90.11.5011 (cité page 36).

Horlacher, O., F. Lisacek et M. Müller (2016), « Mining Large Scale Tandem Mass Spectrometry Data for Protein Modifications Using Spectral Libraries », *in* : *Journal*

of Proteome Research 15.*3*, p. 721-731, DOI : 10.1021/acs.jproteome.5b00877 (cité pages 57, 58).

HUANG, K.-Y. *et al.* (2019), « dbPTM in 2019 : exploring disease association and crosstalk of post-translational modifications », *in* : *Nucleic Acids Research* 47.*D1*, p. D298-D308, DOI : 10.1093/nar/gky1074 (cité page 27).

HUNT, D. F. *et al.* (1992), « Characterization of Peptides Bound to the Class I MHC Molecule HLA-A2.1 by Mass Spectrometry », *in* : *Science, New Series* 255.*5049*, p. 1261-1263, DOI : 10.1126/science.1546328 (cité page 41).

JIMÉNEZ-MUNGUÍA, I. *et al.* (2018), « Proteomic and bioinformatic pipeline to screen the ligands of S. pneumoniae interacting with human brain microvascular endothelial cells », *in* : *Scientific Reports* 8.*1*, p. 5231, DOI : 10.1038/s41598-018-23485-1 (cité page 33).

JISNA, V. A. et P. B. JAYARAJ (2021), « Protein Structure Prediction : Conventional and Deep Learning Perspectives », *in* : *The Protein Journal* 40.*4*, p. 522-544, DOI : 10.1007/s10930-021-10003-y (cité page 26).

JOHNSON, P. E. *et al.* (2011), « Current perspectives and recommendations for the development of mass spectrometry methods for the determination of allergens in foods », *in* : *Journal of AOAC International* 94.*4*, p. 1026-1033 (cité page 35).

KAMATH, K. S., M. S. VASAVADA et S. SRIVASTAVA (2011), « Proteomic databases and tools to decipher post-translational modifications », *in* : *Journal of Proteomics* 75.*1*, p. 127-144, DOI : 10.1016/j.jprot.2011.09.014 (cité pages 1, 19, 27).

KARAS, M. et F. HILLENKAMP (1988), « Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons », *in* : *Analytical Chemistry* 60.*20*, p. 2299-2301, DOI : 10.1021/ac00171a028 (cité page 32).

KHOURY, G. A., R. C. BALIBAN et C. A. FLOUDAS (2011), « Proteome-wide post-translational modification statistics : frequency analysis and curation of the swiss-prot database », *in* : *Scientific Reports* 1, p. 90, DOI : 10.1038/srep00090 (cité pages 17, 19).

KONG, A. T. *et al.* (2017), « MSFragger : ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics », *in* : *Nature Methods* 14.*5*, p. 513-520, DOI : 10.1038/nmeth.4256 (cité pages 56, 78).

LAARSE, S. A. M., A. C. LENEY et A. J. R. HECK (2018), « Crosstalk between phosphorylation and O-GlcNAcylation : friend or foe », *in* : *The FEBS Journal* 285.*17*, p. 3152-3167, DOI : 10.1111/febs.14491 (cité page 17).

Lam, H. (2011), « Building and Searching Tandem Mass Spectral Libraries for Peptide Identification », *in* : *Molecular & Cellular Proteomics : MCP* 10.*12*, R111.008565, DOI : 10.1074/mcp.R111.008565 (cité page 46).

Lam, H. *et al.* (2007), « Development and validation of a spectral library searching method for peptide identification from MS/MS », *in* : *Proteomics* 7.*5*, p. 655-667, DOI : 10.1002/pmic.200600625 (cité page 46).

Lysiak, A. *et al.* (2021), « Evaluation of open search methods based on theoretical mass spectra comparison », *in* : *BMC Bioinformatics* 22.*2*, p. 65, DOI : 10.1186/s12859-021-03963-6 (cité page 63).

— (2022), « SpecGlob : rapid and accurate alignment of mass spectra differing from their peptide models by several unknown modifications », *in* : *bioRXiv*, DOI : 10.1101/2022.05.31.494131 (cité page 100).

Mann, M. et M. Wilm (1994), « Error-tolerant identification of peptides in sequence databases by peptide sequence tags », *in* : *Analytical Chemistry* 66.*24*, p. 4390-4399, DOI : 10.1021/ac00096a002 (cité page 45).

Meftahi, G. H. *et al.* (2021), « Applications of western blot technique : From bench to bedside », *in* : *Biochemistry and Molecular Biology Education* 49.*4*, p. 509-517, DOI : 10.1002/bmb.21516 (cité page 23).

Montecchi-Palazzi, L. *et al.* (2008), « The PSI-MOD community standard for representation of protein modification data », *in* : *Nature Biotechnology* 26.*8*, p. 864-866, DOI : 10.1038/nbt0808-864 (cité page 27).

Na, S., N. Bandeira et E. Paek (2012), « Fast Multi-blind Modification Search through Tandem Mass Spectrometry », *in* : *Molecular & Cellular Proteomics : MCP* 11.*4*, p. M111.010199, DOI : 10.1074/mcp.M111.010199 (cité pages 56, 59).

Na, S., J. Kim et E. Paek (2019), « MODplus : Robust and Unrestrictive Identification of Post-Translational Modifications Using Mass Spectrometry », *in* : *Analytical Chemistry* 91.*17*, p. 11324-11333, DOI : 10.1021/acs.analchem.9b02445 (cité pages 56, 59, 101, 117, 118).

Nesvizhskii, A. I. (2010), « A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics », *in* : *Journal of Proteomics* 73.*11*, p. 2092-2123, DOI : 10.1016/j.jprot.2010.08.009 (cité page 49).

Noor, Z. *et al.* (2021), « Mass spectrometry–based protein identification in proteomics—a review », *in* : *Briefings in Bioinformatics* 22.*2*, p. 1620-1638, DOI : 10.1093/bib/bbz163 (cité page 19).

Omenn, G. S. (2012), « The HUPO Human Proteome Project (HPP), a Global Health Research Collaboration », *in* : *Central Asian Journal of Global Health* 1.*1*, p. 37, DOI : 10.5195/cajgh.2012.37 (cité page 1).

Pazos, M. et K. Peters (2019), « Peptidoglycan », *in* : *Bacterial Cell Walls and Membranes*, sous la dir. d'A. Kuhn, t. 92, Series Title : Subcellular Biochemistry, Cham : Springer International Publishing, p. 127-168, DOI : 10.1007/978-3-030-18768-2_5 (cité page 14).

Perkins, D. N. *et al.* (1999), « Probability-based protein identification by searching sequence databases using mass spectrometry data », *in* : *Electrophoresis* 20.*18*, p. 3551-3567, DOI : 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2 (cité page 45).

Pevzner, P. A., Z. Mulyukov *et al.* (2001), « Efficiency of database search for identification of mutated and modified proteins via mass spectrometry », *in* : *Genome Research* 11.*2*, p. 290-299, DOI : 10.1101/gr.154101 (cité page 56).

Pevzner, P. A., V. Dančík et C. L. Tang (2000), « Mutation-Tolerant Protein Identification by Mass Spectrometry », *in* : *Journal of Computational Biology* 7.*6*, Publisher : Mary Ann Liebert, Inc., publishers, p. 777-787, DOI : 10.1089/10665270050514927 (cité pages 56, 58).

Riffle, M. *et al.* (2022), « Discovery and Visualization of Uncharacterized Drug–Protein Adducts Using Mass Spectrometry », *in* : *Analytical Chemistry* 94.*8*, Publisher : American Chemical Society, p. 3501-3509, DOI : 10.1021/acs.analchem.1c04101 (cité page 64).

Rodriguez, E. L. *et al.* (2020), « Affinity Chromatography : A Review of Trends and Developments over the Past 50 Years », *in* : *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences* 1157, p. 122332, DOI : 10.1016/j.jchromb.2020.122332 (cité page 22).

Sakurai, T. *et al.* (1984), « PAAS 3 : A computer program to determine probable sequence of peptides from mass spectrometric data », *in* : *Biological Mass Spectrometry* 11.*8*, p. 396-399, DOI : 10.1002/bms.1200110806 (cité page 40).

Savitski, M. M., M. L. Nielsen et R. A. Zubarev (2006), « ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel

types of modifications, and fingerprinting complex protein mixtures », *in* : *Molecular & cellular proteomics : MCP* 5.*5*, p. 935-948, DOI : 10.1074/mcp.T500034-MCP200 (cité pages 55, 101).

SAWASDIKOSOL, S. (2010), « Detecting Tyrosine-Phosphorylated Proteins by Western Blot Analysis », *in* : *Current Protocols in Immunology* 89.*1*, DOI : 10.1002/0471142735.im1103s89 (cité page 23).

SEARLE, B. C. *et al.* (2005), « Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm », *in* : *Journal of Proteome Research* 4.*2*, p. 546-554, DOI : 10.1021/pr049781j (cité page 55).

SHTEYNBERG, D. *et al.* (2011), « iProphet : Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates », *in* : *Molecular & Cellular Proteomics* 10.*12*, p. M111.007690, DOI : 10.1074/mcp.M111.007690 (cité page 49).

SHTEYNBERG, D. D. *et al.* (2019), « PTMProphet : Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline », *in* : *Journal of Proteome Research* 18.*12*, p. 4262-4272, DOI : 10.1021/acs.jproteome.9b00205 (cité page 141).

SMITH, C. M. *et al.* (1997), « The protein kinase resource », *in* : *Trends in Biochemical Sciences* 22.*11*, p. 444-446, DOI : 10.1016/s0968-0004(97)01131-6 (cité page 27).

SOLNTSEV, S. K. *et al.* (2018), « Enhanced Global Post-translational Modification Discovery with MetaMorpheus », *in* : *Journal of Proteome Research* 17.*5*, p. 1844-1851, DOI : 10.1021/acs.jproteome.7b00873 (cité pages 55, 78, 101).

THIEDE, B. *et al.* (2005), « Peptide mass fingerprinting », *in* : *Methods* 35.*3*, p. 237-247, DOI : 10.1016/j.ymeth.2004.08.015 (cité page 37).

TYANOVA, S., T. TEMU et J. COX (2016), « The MaxQuant computational platform for mass spectrometry-based shotgun proteomics », *in* : *Nature Protocols* 11.*12*, p. 2301-2319, DOI : 10.1038/nprot.2016.136 (cité page 50).

UNIPROT CONSORTIUM (2021), « UniProt : the universal protein knowledgebase in 2021 », *in* : *Nucleic Acids Research* 49.*D1*, p. D480-D489, DOI : 10.1093/nar/gkaa1100 (cité page 25).

VEIGA LEPREVOST, F. da *et al.* (2020), « Philosopher : a versatile toolkit for shotgun proteomics data analysis », *in* : *Nature Methods* 17.*9*, p. 869-870, DOI : 10.1038/s41592-020-0912-y (cité page 50).

Venne, A. S., L. Kollipara et R. P. Zahedi (2014), « The next level of complexity : Crosstalk of posttranslational modifications », *in* : *Proteomics* 14.*4-5*, p. 513-524, doi : 10.1002/pmic.201300344 (cité page 17).

Vizcaíno, J. A. *et al.* (2016), « 2016 update of the PRIDE database and its related tools », *in* : *Nucleic Acids Research* 44.*D1*, p. D447-D456, doi : 10.1093/nar/gkv1145 (cité page 46).

Walsh, C. T., S. Garneau-Tsodikova et G. J. Gatto (2005), « Protein Posttranslational Modifications : The Chemistry of Proteome Diversifications », *in* : *Angewandte Chemie International Edition* 44.*45*, p. 7342-7372, doi : 10.1002/anie.200501023 (cité pages 1, 16, 17).

Wang, D. *et al.* (2019), « A deep proteome and transcriptome abundance atlas of 29 healthy human tissues », *in* : *Molecular Systems Biology* 15.*2*, doi : 10.15252/msb.20188503 (cité pages 16, 33).

Yates, A. D. *et al.* (2020), « Ensembl 2020 », *in* : *Nucleic Acids Research* 48.*D1*, Publisher : Oxford Academic, p. D682-D688, doi : 10.1093/nar/gkz966 (cité page 67).

Yates, J. R. *et al.* (1998), « Method To Compare Collision-Induced Dissociation Spectra of Peptides : Potential for Library Searching and Subtractive Analysis », *in* : *Analytical Chemistry* 70.*17*, p. 3557-3565, doi : 10.1021/ac980122y (cité page 47).

Yu, F. *et al.* (2020), « Identification of modified peptides using localization-aware open search », *in* : *Nature Communications* 11.*1*, p. 4065, doi : 10.1038/s41467-020-17921-y (cité page 58).

Zhang, G. *et al.* (2014), « Overview of Peptide and Protein Analysis by Mass Spectrometry », *in* : *Current Protocols in Molecular Biology* 108.*1*, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb1021s108, p. 10.21.1-10.21.30, doi : 10.1002/0471142727.mb1021s108 (cité page 24).

Zhang, K. *et al.* (2019), « High-Throughput Experimentation : Where Does Mass Spectrometry Fit ? », *in* : *Spectroscopy Supplements*, Special Issues-07-01-2019 17.*3*, Publisher : MJH Life Sciences, p. 8-15 (cité page 33).

Zhu, Z., J. J. Lu et S. Liu (2012), « Protein separation by capillary gel electrophoresis : A review », *in* : *Analytica Chimica Acta* 709, p. 21-31, doi : 10.1016/j.aca.2011.10.022 (cité page 22).

GLOSSAIRE

Δm Différence de masse parente entre un spectre expérimental et un spectre théorique, positive si le spectre expérimental a une masse parente plus importante que celle du spectre théorique. 50

bait Peptide, ou par extension spectre qui en résulte, qui joue le rôle du spectre expérimental dans le travail d'identification entre spectres théoriques. 70

baitModel Interprétation du spectre expérimental à partir de l'alignement $hitModified$. Les décalages de masse introduits dans l'alignement sont interprétés en séquences de résidus quand cela est possible. 122

hit Peptide, ou par extension spectre qui en résulte, qui joue le rôle du spectre théorique dans le travail d'identification entre spectres théoriques. 70

hitModified Résultat de l'alignement du *hit* sur le *bait*, fourni par `SpecGlob`. Cette chaîne de caractères indique pour chaque résidu du *hit* s'il est aligné dans le *bait*, et si oui, si un décalage de masse doit être inséré pour aligner le résidu sur le *bait*. 102

SPC *Shared Peaks Count*. Nombre de masses en commun partagées par deux spectres MS2. 44

shift SPC Score de comparaison dérivé du SPC. Pour le calculer, le Δm entre deux spectres est considéré comme une seule modification qui est déplacée sur chaque résidu du peptide correspondant au spectre théorique, et le SPC est calculé pour chacun ; le maximum parmi ces SPC est le shift SPC. 59

acide aminé Molécule qui possède un groupe carboxylique (COOH), un groupe amine (NH_2) ainsi qu'un radical, ou chaîne latérale, qui est spécifique à chacun. 10

alignement spectral Alignement de deux spectres MS2 (certaines masses sont décalées afin d'obtenir davantage de masses identiques), par exemple par programmation dynamique, afin d'évaluer leur similarité. 58

bibliothèque spectrale Base de données de spectres MS2 formée à partir de spectres expérimentaux identifiés, à laquelle des spectres à identifier peuvent être comparés.

46

C-ter Extrémité C-terminale (côté du groupement carboxylique COOH) d'un acide aminé, et donc d'un peptide ou d'une protéine. 10

Dalton ou Da. Unité de masse couramment utilisée en physique et chimie, masse du douzième d'un atome de carbone 12. 10

decoy La base decoy est une base de spectres théoriques "leurres" qui est créée pour filtrer des interprétations de spectres. Les identifications provenant de cette banque sont censées être incorrectes. 48

enzyme Protéine dédiée à la catalyse de certaines réactions chimiques ; le nom d'une enzyme est identifiable grâce au suffixe -ase. Par exemple une protéase est une enzyme spécialisée dans la dégradation d'une autre protéine. 14

FDR *False Discovery Rate*. Valeur statistique associée au taux de faux-positifs, calculée afin de contrôler la qualité des identifications. Le FDR peut être calculé selon la stratégie target/decoy. Pour cela, à partir de la base de protéines étudiées (target), une base supposée représenter des mauvaises identifications est créée (base decoy). Les PSMs sont produits avec pour candidats proposés les bases de peptides target et decoy. Le FDR peut alors être calculé en divisant le nombre de PSMs avec une identification dans la base decoy divisé par le nombre de PSMs au total, et le seuil de score peut être augmenté jusqu'à diminution du FDR à un seuil acceptable, habituellement 1%. 48

masse parente Masse du peptide fragmenté, associée au spectre MS2 correspondant. 39

MS *Mass Spectrometry* : spectrométrie de masse. 30

N-ter Extrémité N-terminale (côté du groupement amine NH_2) d'un acide aminé, et donc d'un peptide ou d'une protéine. 10

peptide Molécule de quelques résidus à quelques dizaines de résidus, naturelle ou bien issue de la fragmentation d'une protéine. 13

ppm Partie par million. Le ppm est une fraction valant un millionième. Elle est utilisée en MS pour décrire la résolution, la précision ou la tolérance non pas de façon absolue (Dalton) mais relative à la masse d'un peptide. 32

précision Différence entre une valeur mesurée par un instrument et sa valeur réelle. 32

protéase Une protéase est une enzyme qui clive la chaîne de résidus (c'est-à-dire brise des liaisons peptidiques) de façon spécifique ou non. Par exemple, la trypsine est une protéase qui coupe après les lysines (K) et arginines (R) dans un peptide. 34

protéine Enchaînement de résidus produit selon l'information génétique. Les protéines ont des séquences et des structures très variées, ainsi que des fonctions qui le sont tout autant au sein du vivant. 10

protéomique Identification et quantification des ensembles de protéines - ou protéomes - qui sont exprimés dans une dimension spatio-temporelle donnée, par une cellule, un tissu ou un organisme, dans des conditions données. 19

PSM *Peptide Spectrum Match*. Association d'un spectre et d'un peptide. 42

PTM *Post-Translational Modification*. Modification chimique appliquée sur un ou plusieurs résidu(s) d'une protéine après sa traduction. 16

réseau Graphe dont les nœuds et les arêtes ont des attributs. 72

résidu Acide aminé impliqué dans une liaison peptidique ; qualifie donc un acide aminé impliqué dans un peptide ou une protéine. 13

résolution Plus petite différence mesurable par un instrument. 32

spectre de masse Donnée produite par un spectromètre de masse à partir d'un ion, sur laquelle une intensité est représentée selon un ratio m/z. Ces spectres peuvent être des spectres MS1 produits à partir des peptides, ou des spectres MS2 produits à partir des ions issus de la fragmentation d'un peptide. 36

target La base target est une base de spectres théoriques produite à partir d'une base de protéines natives et des peptides qui en résultent. 48

tolérance Différence de masse parente (Δm) acceptée pour sélectionner les spectres candidats pour un spectre donné. 43

traceback Terme appliqué à la programmation dynamique. Une fois que la matrice de programmation dynamique est remplie, l'étape de traceback consiste à remonter dans la matrice de scores pour retrouver les étapes qui ont conduit au meilleur score. L'étape de traceback de `SpecGlob` permet ainsi de reconstruire l'alignement des spectres. 104

DOCTORAT BRETAGNE LOIRE / MATHSTIC

Nantes Université

Titre : Développement de méthodes informatiques pour l'évaluation et l'amélioration de l'identification par spectrométrie de masse des peptides modifiés

Mot clés : Bioinformatique, protéomique, spectrométrie de masse, peptide, OMS

Résumé : La spectrométrie de masse (MS) est l'une des méthodes privilégiées pour identifier les protéines. Elles sont habituellement identifiées à partir de leurs peptides. Pour cela, les spectres obtenus à partir des peptides sont comparés à une base de données de spectres théoriques à l'aide d'un score de similarité, et les PSM (*Peptide-Spectrum Matches*) produits aident à l'identification des spectres. Cependant, la plupart des spectres générés ne peuvent pas être correctement identifiés. L'une des raisons est que les protéines, et donc les peptides résultants, portent des modifications, ce qui complexifie l'identification des spectres issus de ces peptides. Pour résoudre ce problème, les approches OMS (*Open Mass Search*) offrent des éléments prometteurs. Cependant ces méthodes sont encore confrontées à certains obstacles, surtout lorsque le peptide comporte des modifications inconnues. De plus, l'évaluation de leurs résultats est complexe. Dans le cadre de cette thèse, j'ai d'abord évalué des stratégies OMS à l'aide de spectres théoriques jouant le rôle de spectres expérimentaux, ce qui a permis de développer de nouveaux critères pour évaluer les PSM. À la suite de ce travail, j'ai développé `SpecGlob`, un algorithme qui repose sur la programmation dynamique pour aligner les masses contenues dans un PSM, afin d'identifier plusieurs modifications inconnues qui séparent le peptide qui a généré le spectre à identifier du peptide candidat.

Title: Development of computational methods to evaluate and improve the identification by mass spectrometry of modified peptides

Keywords: Bioinformatics, proteomics, mass spectrometry, peptide, OMS

Abstract: Mass spectrometry (MS) is one of the main methods used to identify proteins. They are usually identified from their peptides. To this end, spectra obtained from their peptides are compared to a database of theoretical spectra using a similarity score, and resulting PSMs (*Peptide-Spectrum Matches*) are used to identify spectra. Nevertheless, most of the spectra that are generated cannot be properly identified. One of the reasons is that proteins, and thus resulting peptides, carry modifications that complexify the identification of corresponding spectra. To solve this problem, OMS (*Open Mass Search*) methods offer promising elements. But these methods still face obstacles, especially when the peptide carries unknown modifications. Moreover, the evaluation of their results is complex. During my PhD, I first evaluated OMS strategies with theoretical spectra playing the role of experimental spectra, which enabled the development of new criteria to evaluate PSMs. Following this work, I developed `SpecGlob`, an algorithm that relies on dynamic programming to align masses of spectra of a PSM, aiming at identifying several unknown modifications seperating the peptide that generated the spectrum from the candidate peptide.