Nouvelles approches de quantification des variations du protéome au niveau des protéines intactes : analyses expérimentales et computationnelles

Nicolas Sénécaut

Ecole doctorale 474 Frontières de l'Innovation en Recherche et Education (FIRE)

Thèse de doctorat en Biologie et Informatique

Mention « Gène, Omiques, Bioinformatique et Biologie des systèmes »

Nouvelles approches de quantification des variations du protéome au niveau des protéines intactes : analyses expérimentales et computationnelles

Par Nicolas SÉNÉCAUT

Institut Jacques Monod
UMR7592 - Université Paris Cité
Centre National de la Recherche Scientifique

Soutenue publiquement le 22 juin 2022

Devant un jury composé de :

Emmanuelle LEIZE-WAGNER	Directrice de recherche CNRS, Université de Strasbourg, Rapportrice
Willy BIENVENUT	Chargé de recherche CNRS, Université Paris-Saclay, Rapporteur
Fernando RODRIGUES-LIMA	Professeur, Université Paris Cité, Président du jury
Chiara GUERRERA	Ingénieure de Recherche, Université Paris Cité, Examinatrice
Denis MESTIVIER	Professeur, Université Paris-Est Créteil (UPEC), Examinateur
Jean-Michel CAMADRO	Directeur de recherche CNRS, Institut Jacques Monod (Directeur de thèse)
Gaëlle LELANDAIS	Professeure, Université Paris Sud (co-Directrice de thèse)

REMERCIEMENTS

A mon Directeur de Thèse le Docteur Jean-Michel Camadro

Capitaine Jean-Michel, merci d'avoir cru en moi et de m'avoir pris sous ton aile et de m'avoir appris tant de nouvelles choses. La biochimie est un monde fantastique et vaste dont tu as été un guide. Toutes les idées ou hypothèses étaient bonnes à prendre et à développer, mon esprit de curiosité était au bon endroit. Encore merci, en espérant que nos caps maritimes se croiseront (peut-être au large de Belle-Île ?).

Au Professeure Gaëlle Lelandais

Gaëlle tu m'as fait confiance dès notre première réunion. Merci pour tout, ton soutien, ton temps, ton aide et tes conseils précieux. Je n'ai plus peur des statistiques grâce à toi, chaque problème possède une solution simple. Je te remercie en particulier pour la relecture de ce manuscrit et tes commentaires pertinents.

Aux membres de mon jury

Je remercie le Professeur Fernando RODRIGUES-LIMA d'avoir accepté de présider mon jury de thèse. Je remercie également la Docteure Emmanuelle LEIZE-WAGNER et le Docteur Willy BIENVENUT d'avoir accepté d'être mes rapporteurs et d'avoir pris le temps de relire ce manuscrit. Merci enfin aux examinateurs, la Docteure Chiara GUERRERA et le Professeur Denis MESTIVIER pour avoir pris le temps de me suivre durant ce doctorat et pour avoir accepté d'examiner de ce travail de thèse.

Au laboratoire Camadro de l'IJM

Je remercie les membres du laboratoire de l'équipe Camadro présents ou passés : Élodie Rousseau (la reine des Western-Blot), Thomas Denecker (mon cousin de thèse), Valérie Serre, Françoise Auchère, Pierre Poulain. A tous merci pour votre aide votre soutien, vos disponibilités et vos conseils précieux. Une pensée pour Emmanuel Lesuisse qui nous a récemment quittés.

A la plateforme de protéomique de l'Institut Jacques Monod

Je remercie les membres passés ou présents de la plateforme ProteoSeine@IJM : Bastien Morlet (pour ton accueil sur la plateforme quand j'étais en stage de master), Camille Garcia, Thibaut Léger, Samuel Terrier (qui m'a fait découvrir la subtilité de la protéomique).

Mention spéciale à Véronique Legros, pour ton aide, ton temps et tes discussions enrichissantes. Merci à Guillaume Chevreux pour tes conseils et ton travail qui nous ont propulsés dans la protéomique du futur. Enfin, merci à Laurent Lignières, qui arrivé au début de mon aventure, a toujours su m'aider, me conseiller, et avec qui j'ai partagé le meilleur et le pire en particulier durant le retraitement des données de spectrométrie de masses.

Service support

Je remercie également tous les membres de l'Institut Jacques Monod pour les nombreux services rendus. En particulier les administratifs qui ont un rôle crucial dans le soutien à la recherche.

Au CRI (Centre de Recherche Interdisciplinaire) maintenant LPI

Après presque 9ans d'études (Licence FDV, Master AIV, Doctorat FIRE) il est l'heure de se quitter et de vous remercier. Merci à tous les enseignants, chercheurs et les coordinateurs d'études. En particulier Typhaine et Virginie, Jeannette, Chiara, Élodie et Camille.

Et enfin un grand merci à mes proches (famille et amis), ceux grâce à qui je suis là aujourd'hui.

<u>Titre :</u> Nouvelles approches de quantification des variations du protéome au niveau des protéines intactes : analyses expérimentales et computationnelles

Résumé :

La spectrométrie de masse à haute sensibilité et à haute résolution est un outil de mesure adapté permettant de décrire qualitativement et quantitativement la dynamique et les variations de la composition du protéome. La complexité des spectres de masse produits s'explique par la présence d'isotopes dans les molécules biologiques analysées. Le « Simple Light Isotope Metabolic labeling » (SLIM), est une méthode de protéomique quantitative précédemment développée au laboratoire, au cours de laquelle la composition isotopique des protéines est modulée *in vivo*, par apport d'une source de carbone uniformément marquée au ^{12}C. Ceci permet la synthèse des acides aminés et donc des protéines ne contenant que du ^{12}C, les clusters isotopiques expérimentaux ainsi obtenus permettent la quantification après multiplexage des échantillons marqués et non marqués issus de conditions biologiques différentes en une seule analyse.

Durant cette thèse nous avons optimisé la version « Bottom-up » de la méthode SLIM (bSLIM) par le développement d'un workflow automatique de traitement des données brutes (logiciel KNIME), utilisant des méthodes algébriques permettant la modélisation des clusters isotopiques théoriques. Une description quantitative fine de chaque peptide est obtenue en mesurant le rapport entre les intensités isotopiques expérimentales et théoriques. Les valeurs quantitatives obtenues ont été validées expérimentalement par la comparaison de deux protéomes issus de souches de levure proches. L'observation des variations d'abondance des protéines marqueurs d'auxotrophies, démontre la sensibilité de la méthode. En complément de l'usage d'outils OpenSource et performants, nous avons produit un deuxième workflow similaire dont l'utilisation de ressources informatiques propriétaires (logiciel Proteome Discoverer) pour optimiser l'extraction des intensités expérimentales des isotopologues dans les clusters isotopiques. Ce sont des données expérimentales critiques pour notre méthode de quantification. Nous avons par ailleurs développé des méthodes statistiques robustes dont l'objectif est d'évaluer et d'estimer le taux de fausses découvertes dans les résultats obtenus. Nous avons également adapté les méthodes de résolutions algébriques permettant la quantification dans l'éventualité de marquage incomplet des protéines, que nous avons expérimenté sur des organismes présentant des auxotrophies t.q. des lignées cellulaires humaines. De manière analogue, une preuve de faisabilité de marquage indirect a été développée avec le nématode *C. elegans*, ouvrant ainsi les perspectives pour des applications futures de la méthodes SLIM aux sujets impliquant des organismes eucaryotes supérieurs.

Le développement au préalable du bSLIM nous a permis d'aborder les difficultés rencontrées lors de l'observation par spectrométrie de masse au niveau des protéines intactes (approches Top-down) et le déploiement de la méthode tSLIM. En effet, la masse élevée de ces objets biologiques pose plusieurs contraintes méthodologiques et technologiques, la principale étant qu'à cette échelle de masse, le nombre d'isotopes lourds dans les molécules est très grand. Les clusters isotopiques observés dans les spectres expérimentaux présentent donc une faible intensité de l'ion monoisotopique, rendant d'autant plus difficile sa mesure. C'est pourquoi nous avons développé une solution utilisant la distribution de Poisson de l'intensité des isotopologues dans chaque massif isotopique comme modèle théorique, pour décrire les clusters isotopiques en tSLIM, ce qui permet la quantification des protéines intactes. En parallèle, nous avons développé un protocole de préparation des protéines intactes adapté au l'analyse Top-Down. Le développement d'outils dédiés de traitement des données nous a permis de développer une méthode originale d'identification en top-down permettant, en plus de la quantification, la caractérisation des protéoformes.

<u>Mots-clés :</u> Spectrométrie de masse, Simple Light Isotope Metabolic labeling (SLIM), Biologie computationnelle, Protéomique Quantitative

<u>Title:</u> New quantification approaches of proteome variations at the intact proteins level: experimental and computational analyses

Abstract :

High sensitivity and high resolution mass spectrometry is a suitable tool to describe qualitatively and quantitatively the dynamics and variations of the proteome composition. The complexity of the mass spectra produced is explained by the presence of isotopes in the biological molecules analyzed. Simple Light Isotope Metabolic labeling (SLIM), is a quantitative proteomics method previously developed in the laboratory, in which the isotopic composition of proteins is modulated *in vivo*, by the addition of a carbon source uniformly labeled with ^{12}C. This allows the synthesis of amino acids and therefore proteins containing only ^{12}C, the experimental isotopic clusters thus obtained allowing the quantification between different biological conditions in a single analysis.

During this thesis we optimized the bottom-up version of the SLIM method (bSLIM) by developing an automatic workflow for processing raw data (KNIME software), using algebraic methods to model theoretical isotopic clusters. A fine quantitative description of each peptide is obtained by measuring the ratio between experimental and theoretical isotopic intensities. The quantitative values obtained were validated experimentally by comparing two proteomes from closely related yeast strains. The observation of abundance variations of auxotrophy marker proteins demonstrated the sensitivity of the method. In addition to the use of OpenSource and powerful tools, we have produced a second similar workflow using proprietary computer resources (Proteome Discoverer software) to optimize the extraction of experimental intensities of isotopologues in isotopic clusters. These are critical experimental data for our quantification method. We have also developed robust statistical methods to evaluate and estimate the false discovery rate in the obtained results. We have also adapted algebraic resolution methods for quantification in the event of incomplete protein labeling, which we have tested on organisms with auxotrophies (human cell lines). Similarly, a proof of concept for indirect labeling has been developed with the nematode, thus opening the perspectives for future applications of the SLIM method to subjects involving higher eukaryotic organisms.

The prior development of bSLIM allowed us to address the difficulties encountered when analyzing intact proteins by mass spectrometry (top-down approaches) and the deployment of the tSLIM method. Indeed, the high mass of these biological objects poses several methodological and technological constraints, the main one being that at this mass scale, the number of heavy isotopes in the molecules is very large. The isotopic clusters observed in the experimental spectra thus present a weak intensity of the monoisotopic ion, making its measurement even more difficult. Therefore, we have developed a solution using the Poisson distribution of the isotopic intensity in each isotopic cluster as a theoretical model to describe the isotopic clusters in tSLIM, which allows the quantification of intact proteins. In parallel, we have developed a protocol for the preparation of intact proteins adapted to Top-Down analysis. The development of dedicated data processing tools allowed us to develop an original top-down identification method allowing, in addition to the quantification, the characterization of proteoforms.

<u>Keywords:</u> Mass spectrometry, Simple Light Isotope Metabolic labeling (SLIM), Computational biology, Quantitative proteomics

TABLE DES MATIERES

ABRÉVIATIONS

SLIM : Simple Light Isotope Metabolic

12C : Condition Marquée SLIM 12C

NC : Condition Naturelle (non-marqué)

MIDAs : Molecular Isotopic Distribution Analysis

PTMs : Post-Translational Modifications

FFId : Feature Finder Identifications

SILAC : Stable Isotope Labeling by Amino acids in Cell culture

TMT : Tandem Mass Tag

iTRAQ : Isobaric Tag for Relative and Absolute Quantification

LFQ Label Free Quantification

RAId : Robust Accurate Identification

HUPO : Human Proteome Organization

LISTE DES FIGURES

LISTE DES TABLEAUX

AVANT-PROPOS : CONTEXTE GENERAL ET ORGANISATION DU MANUSCRIT

J'ai réalisé ma thèse à l'Institut Jacques Monod, sous la direction du Dr Jean-Michel Camadro et de la Pr Gaëlle Lelandais. J'ai travaillé dans un contexte de recherche interdisciplinaire alliant biologie expérimentale et bioinformatique. Pour chacun de mes projets, j'ai ainsi alterné des périodes de travail pendant lesquelles j'ai appliqué ou mis au point des protocoles de préparation d'échantillons biologiques, avec des périodes pendant lesquelles j'ai réalisé des simulations informatiques et écrit des programmes pour automatiser et rendre reproductibles mes analyses. Je me suis intéressé à la problématique de l'étude des protéines par spectrométrie de masse. A l'instar des autres "omiques" (génomique et transcriptomique), la protéomique est une discipline de recherche qui connait des évolutions rapides et spectaculaires, à la fois sur le plan technologique (performance des appareils utilisés) et sur le plan des applications en recherche fondamentale et appliquée. Celles-ci résultent de capacités nouvelles à identifier et quantifier les protéines de façon systématique et précise.

Mon manuscrit de thèse est organisé en quatre grandes parties. La première (nommée « Introduction générale ») a pour objectif de présenter les enjeux scientifiques et le contexte général dans lequel mes travaux de thèse s'inscrivent. Nous verrons que l'étude de la dynamique du protéome des organismes et la quantification des variations de composition associées permettent d'étudier leurs réponses à des conditions biologiques spécifiques. La quantification de chaque protéine, nécessite de bien comprendre la structure fine des protéines, d'utiliser des appareils de mesure spécifiques que sont les spectromètres de masse, et enfin d'appliquer des stratégies originales qui allient variabilités et contraintes biologiques avec les méthodes d'analyse

computationnelle des observations par ces outils. J'expliquerai comment la spectrométrie de masse permet l'étude des protéines et je discuterai les avantages et les limites des analyses classiquement réalisées dans les laboratoires de recherche en biologie. Enfin, je présenterai la méthode SLIM (Simple Light Isotope Metabolic Labeling), développée par Jean-Michel Camadro avant mon arrivée au laboratoire. La méthode SLIM est à l'origine de mes travaux de thèse, qui ont consisté en la mise au point des variantes bSLIM et tSLIM, dont les détails seront présentés dans les sections suivantes (nommées respectivement « Développement de la méthode bSLIM » et « Développement de la méthode tSLIM »). La dernière partie de mon manuscrit, nommée « Conclusion et perspectives de la thèse » sera consacré à la présentation des ouvertures possibles de mes travaux de thèses.

Partie 1 : Introduction générale

Chapitre 1 : Protéomique : enjeux scientifiques et contexte générale

I. Étude des fonctions des protéines

I.1 Relation structure – fonction

Les protéines sont des macromolécules biologiques constituées d'une chaîne d'acides aminés, des molécules plus petites aux propriétés physico-chimiques diverses. Les protéines sont structurées en trois dimensions (des structures tri-dimensionnelles). Elles sont impliquées dans de nombreuses fonctions biologiques (Sleator, 2012)(Liebermeister et al., 2014) telles que :

- La structure et l'organisation des cellules : pour la formation ou le maintien de macrostructures cellulaires comme les filaments d'actines qui composent le cytosquelette.
- La signalisation : pour la communication entre les cellules.
- Le transport des éléments : pour la réalisation de mouvements et de transports de molécules. Par exemple les dynéines, complexes protéiques, responsables du transport de vésicules vers le centre de la cellule le long des fibres de microtubules (mouvement rétrograde), à la différence des Kinésines qui effectuent le transport dans le sens inverse (antérograde).
- Les mécanismes de régulation de l'expression des gènes : nécessaires pour la transcription des gènes par exemple (facteur de transcription).
- La production d'énergie : nécessaire pour de multiples réactions cellulaires (ATP synthase, pompes à protons).
- Les membranes biologiques : 27% des protéines d'origine humaine, interviennent dans la membrane des cellules (Almén et al., 2009). Cela est expliqué par le fait que les cellules humaines forment un organisme complexe et donc la communication est une tâche très importante.
- Les réactions enzymatiques : la majorité des protéines sont des enzymes, catalysant des réactions biochimiques spécifiques. Par exemple dans le

peroxysome d'une levure, la protéine CAT1 (catalase) protège par oxydoréduction les cellules des effets destructeurs du peroxyde d'hydrogène (Figure 1).

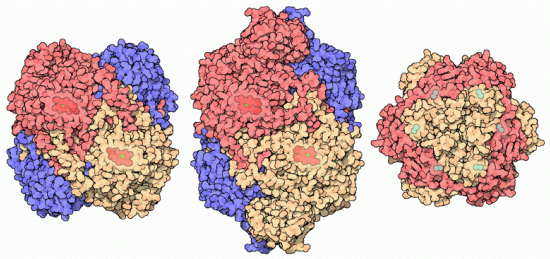


Figure 1: Représentation tri-dimensionnelles de trois protéines, une d'origine humaine à gauche et deux d'origine bactérienne au milieu et à droite. Ces protéines sont des HPII et catalases de type manganèse, elles permettent d'effectuer une même fonction : protéger la cellule de l'oxydation (Source David Godsell [1]).

Ainsi, les protéines sont organisées en structures parfaitement définies et spécifiques, c'est à dire avec des conformations structurales, des sites actifs et des domaines d'hydrophobicité dans le but d'accomplir ces fonctions essentielles à la biologie cellulaire (Uhlén et al., 2015). La Figure 2 illustre la grande diversité des formes et des structures naturelles des protéines. Elle a été dessinée par l'artiste et biologiste structuraliste David S. Goodsell. Toutes les protéines sont représentées avec une échelle normalisée, ces dessins sont dérivés principalement d'observation de structures cristallographiques.

[1] https://pdb101.rcsb.org/motm/57

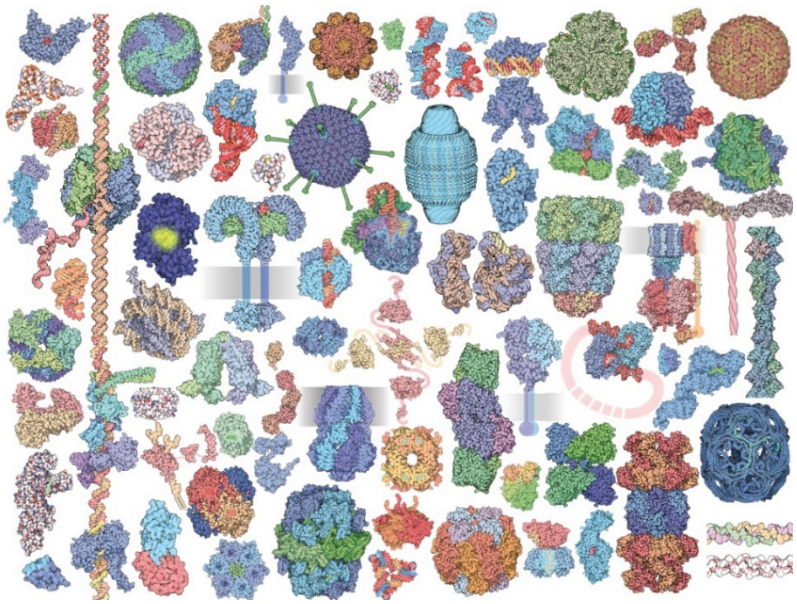


Figure 2: Représentation de la diversité des formes et des structures naturelles des protéines. Illustration réalisée par l'artiste et biologiste structuraliste David S. Goodsell
[1]https://pdb101.rcsb.org/

I.1.1 Quelques exemples emblématiques

- **Exemple de l'ATP synthase**

L'Adénosine Triphosphate (ATP) est une molécule permettant un stockage d'énergie dans la cellule. L'énergie est en effet libérée lors du clivage de certaines liaisons chimiques de la molécule, libérant alors un groupement phosphate. Cette énergie est essentielle à de nombreuses réactions cellulaires. Dans ce contexte, l'ATP-Synthase est une « machine moléculaire » associée à la membrane des mitochondries et des chloroplastes (Rak et al., 2009). Elle permet la formation d'ATP. Cette macrostructure biologique s'illustre par un mouvement mécanique étonnant, initié par un gradient d'ions hydrogènes (protons). C'est l'un des exemples les plus notables de l'intérêt de la structure des protéines sur leurs fonctions.

Ce complexe protéique unique est composé de deux moteurs biologiques, l'un moléculaire, l'autre chimique (Figure 3). Plus précisément, la base de la structure est constituée par un moteur moléculaire, le complexe F_0. Ce dernier est composé d'une première sous-unité de domaine fortement hydrophobe, ancrée à la membrane et jouant le rôle de "stator". Une deuxième sous-unité, jouant le rôle de "rotor", permet

[1] https://pdb101.rcsb.org/

sous l'effet de la force protomotrice (un flux continu de protons généré par un gradient transmembranaire) la rotation de la structure supérieure.

Le deuxième complexe F_1 est un moteur chimique, dont la rotation de la partie centrale permet la modification de la structure des sites actifs des trois unités β, ces dernières permettant la formation d'ATP à partir d'ADP et de phosphate. Les unités α participent quant à elles au maintien de la structure en périphérie. L'ensemble des unités statiques sont maintenues grâce à un ensemble vertical de protéines ancrées à la membrane.

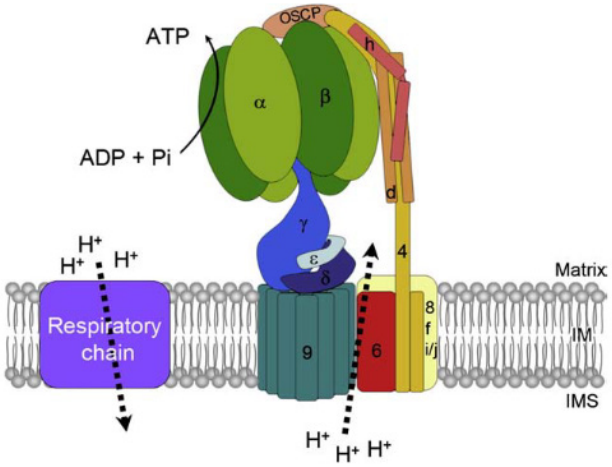


Figure 3 : Organisation spatiale de l'ATP synthase de la levure Saccaromyces cerevisiae. Cette représentation est issue de (Rak et al., 2009)

- **Exemple des Histones**

Nous pouvons citer un autre exemple de la relation forte entre structure et fonction des protéines, l'effet des modifications post-traductionnelles (PTMs). Ces changements consistent en l'ajout ou la modification de groupements chimiques constituant les protéines, avec pour effet la modification de leurs structures et ainsi leurs fonctions. En particulier, l'acétylation des lysines (ajout ou suppression d'un groupement acétyle sur une lysine), trouve son application la plus connue sur les histones. Ces protéines, par leurs fortes affinités avec l'ADN, permettent la formation des nucléosomes, supports du maintien de la chromatine. La modulation de l'état de compaction de la chromatine a un effet sur la transcription, la réplication ou la réparation de certains gènes. L'interaction entre l'ADN chargé négativement et les histones s'explique par leurs compositions riches en acides aminés aux charges positives, tels que la lysine, l'arginine et l'histidine. L'ajout de groupements acétyles

(issus de la molécule d'acétyl-coenzyme A) sur les lysines des histones est régulé par les familles d'enzymes nommées lysines acétyltransférases (KATs) et lysines désacétylases (KDAs). L'addition d'un tel groupement de charges négatives sur les histones apporte donc une neutralité dans l'équilibre électrostatique, les rendant donc moins fortement liées avec l'ADN (Zentner & Henikoff, 2013). L'état de modifications des histones régule ainsi l'expression des gènes (on parle également de code histone (Zhong et al., 2013)) et permet la transmission d'informations non codantes, définitions même de l'épigénique (Tableau 1).

Table 1. Acetylation sites on histones

Histone	Acetylation site	Functional consequence
H3	K9 (PCAF,Gcn5),	Transcriptional activation
	K14 (CBP, p300, PCAF, Gcn5)	Transcriptional activation
	K18 (CBP, p300, PCAF, Gcn5)	Transcriptional activation, H3 removal, cell cycle
	K56 (Rtt109), (Hst3/4 HDAC)	Histone deposition and chromatin assembly; DNA repair
H4	K5 (CBP, p300, HBO1, HAT1)	Transcriptional activation
	K8 (TIP60, HBO1, Esa1, CBP, p300)	DNA repair
	K12 (TIP60, HBO1, HAT1, Esa1,ATAC)	Transcriptional activation
	K16 (TIP60, Esa1, Sas2, Mof/MYST1), (Sirtuins)	rDNA silencing, development; chromatin architecture
H2A	K5 (CBP, p300)	Transcriptional activation
	K14 in H2Az (Esa1, Gcn5, NuA4)	Transcriptional activation
H2B	K12 (CBP,p300)	Chromatin architecture
	K15 (CBP, p300)	Chromatin architecture
	K6,K11(Hos3 HDAC), K16, K17,K21, K22	Transcriptional activation; cell survival

Tableau 1 : Principaux sites d'acétylation des histones et leurs effets fonctionnels (Selvi & Kundu, 2009)

I.1.2 Considérations évolutives

Lors de l'étude des protéines dans les différents organismes vivants, nous pouvons observer des homologies en termes de structure (des protéines ayant la même configuration tri-dimensionnelle), mais également des homologies en termes de fonctions, c'est à dire des protéines ayant des structures totalement différentes mais réalisant des fonctions similaires. Généralement, cette homologie s'explique par un mécanisme de coévolution (protéines paralogues) ou d'évolution séparée (protéines homologues). Différents organismes possèdent des protéines dont les structures permettent d'effectuer des fonctions similaires ou spécifiques à chacune (Todd et al., 2001). Ainsi, l'étude de la similitude et de la divergence des structures des protéines permet de comprendre l'histoire évolutive de ces organismes. Notamment, il est

attendu que deux protéines provenant de deux espèces distinctes soient issues d'un même précurseur si :

- Leurs séquences ADN sont similaires,
- Leurs séquences protéiques sont similaires,
- Leurs structures tri-dimensionnelles sont similaires : SCOP (Andreeva et al., 2020) est une base de données qui les référence et les classe par domaines structuraux et espèces (Structural Classification of Proteins),
- Les enzymes possèdent des substrats similaires,
- Leur mécanisme catalytique est très semblable.

L'un des exemples classiques de cette évolution est la famille des métalloprotéines, comme l'hémoglobine et l'hémocyanine. En effet, ces variants possèdent des affinités métalliques diverses (fer ou cuivre) mais répondent toutes à une même fonction : le maintien et le transport de l'oxygène, dans les érythrocytes ou hémolymphe respectivement.

I.2 Relation séquence – structure

I.2.1 Propriétés des acides aminés

Tous les acides aminés standards incorporés dans les protéines, appelés "protéinogènes", sont des molécules composées d'un agencement chimique identique. Toutefois, les acides aminés diffèrent par la nature de leurs chaînes latérales, offrant une diversité de groupements chimiques et apportant des caractéristiques uniques à chaque acide aminé. Les vingt-et-un acides aminés composant les protéines des Eucaryotes sont constitués d'un atome de carbone central auquel sont liés la chaîne latérale mais également deux groupements, permettant la liaison peptidique. Le premier groupement est un « acide carboxylique », un acide faible (groupement qui libère facilement des protons H^+) et qui comportera alors une charge négative au pH physiologique (pH = 7.4 dans le sang). Le deuxième groupement, « amine », comportant un atome d'azote donc « aminé », et dont la charge sera positive à pH basique et physiologique. Ainsi, le pH intracellulaire (pH 7.2 dans le cytosol des cellules), favorise l'état "zwitterionique" durant lequel les groupements portent individuellement des charges mais les molécules possèdent un équilibre électronique

global neutre. Ce squelette commun à tous les acides aminés donne ainsi son nom à la famille des molécules que forment les acides aminés.

- **La diversité chimique des acides aminés**

Nous pouvons classer les acides aminés en cinq groupes, en fonction de la similitude de leurs propriétés et structures chimiques (Figure 4).

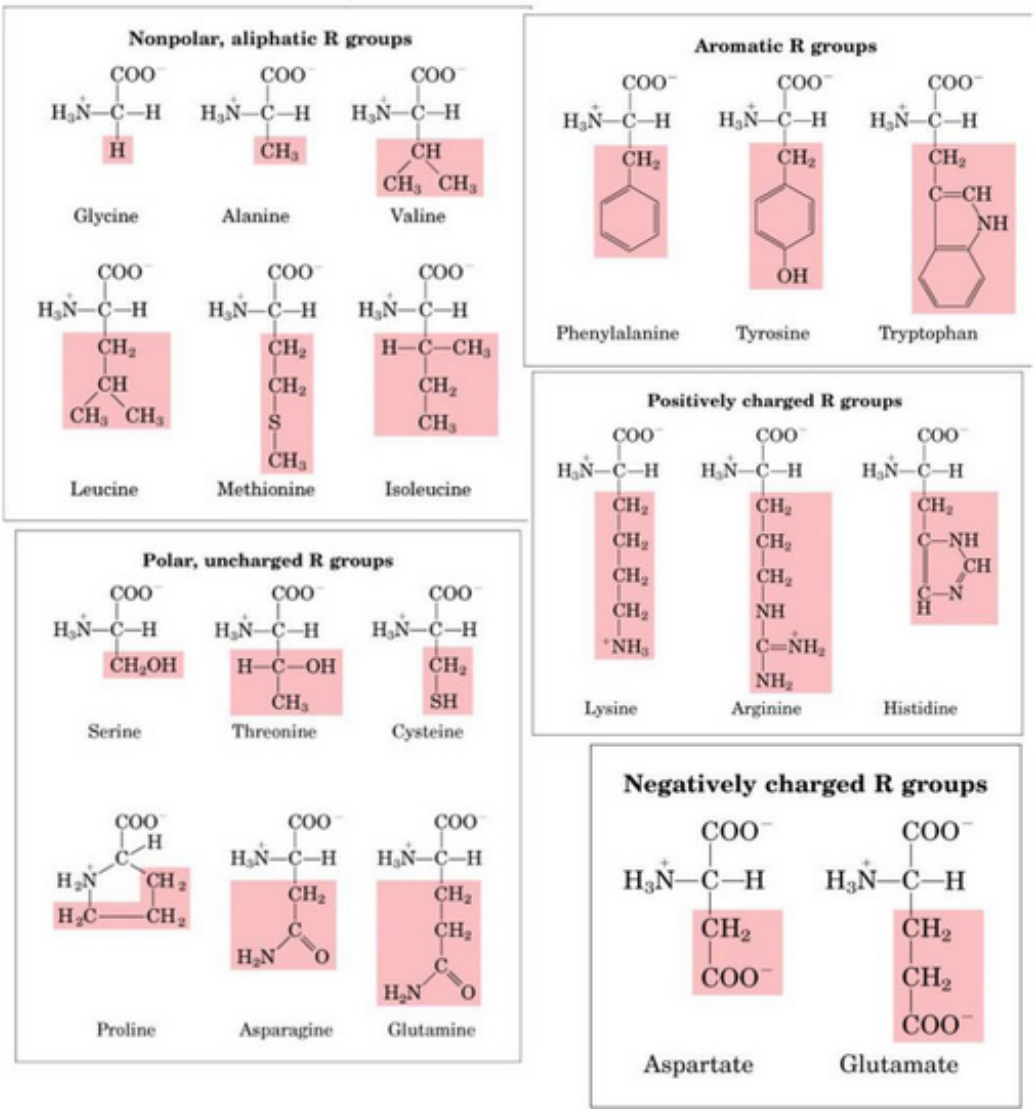


Figure 4 : représentation des groupements chimiques des acides aminés, classés en cinq groupes. Cette image est extraite de la présentation utilisée par Lennart Martens[1]

[1] Support de la présentation « MS-based proteomics data analysis » par Lennart Martens (Sep 27, 2021)

Le premier groupe représenté en Figure 4 rassemble les acides aminés dit aliphatiques, c'est-à-dire les acides aminés dont la chaîne latérale ne porte pas de charge, et dont la composition (une chaîne carbonée), les rend non-polaires, donc peu solubles avec les milieux aqueux (molécules hydrophobes). Il comporte, la Glycine (Gly G), l'Alanine (Ala A), la Valine (Val V), la Leucine (Leu L), l'Isoleucine (Iso I) et la Méthionine (Met M). Il est à noter que la Méthionine est un cas particulier puisque cette molécule possède un atome de soufre. Cet acide aminé, correspondant au codon (AUG), initie généralement les séquences protéiques lors de leur synthèse, puis est excisée lors des étapes de maturation de la protéine (modification post-traductionnelle) par une aminopeptidase.

Les acides aminés polaires et non-chargés ne comportent pas de groupements chimiques chargés, mais ces espèces possèdent des atomes très électronégatifs (comme l'oxygène), ayant pour effet de les rendre polaires, et donc hydrophiles. Deux d'entre eux, la Sérine (Ser S) et la Thréonine (Thr T) possèdent une fonction hydroxyle (-OH) permettant la formation de liaisons hydrogènes particulièrement importantes pour la formation de structures tri-dimensionnelles. Ce groupe contient également la Cystéine (Cys C), le deuxième acide aminé contenant un atome de soufre. Par la présence d'un groupement thiol (-SH), cet acide aminé est fondamental pour le maintien et la formation de ponts disulfures (liaison de deux cystéines par les -SH formant une molécule de cystine), induisant une contrainte structurale importante pour la conformation des protéines. L'Asparagine (Asn N) et la Glutamine (Gln Q) sont dérivées de l'acide aspartique (Asp D) et de l'acide glutamique (Glu E) respectivement, par l'ajout d'un groupement NH_2 sur la chaîne latérale. Le dernier acide aminé de ce groupe est la Proline (Pro P). Sa chaîne latérale est reliée à l'amine primaire, formant un cycle qui induit ainsi un angle de liaisons peptidiques différent, induisant un changement de la configuration de la structure tri-dimensionnelle de la séquence protéique.

Le groupe des acides aminés aromatiques comporte les trois acides aminés les plus volumineux. La chaîne latérale est un résidu cycle aromatique, les rendant apolaires. Le cycle phénol, dont le premier composé de ce groupe est la Phénylalanine (Phe F), le deuxième la Tyrosine (Y Tyr) possède à son extrémité un groupement alcool permettant la formation de liaisons hydrogènes avec les autres molécules. Le Tryptophane (Trp W) est également particulièrement hydrophobe et volumineux, en

faisant un acide aminé relativement peu abondant dans les protéines. Il a également tendance à être localisé au centre des structures tridimensionnelles, nécessitant alors un espace conséquent.

L'Acide aspartique (Asp D) et l'Acide glutamique (Glu E), forment le groupe des acides aminés chargés négativement. En effet, ces deux molécules sont munies d'un deuxième groupement carboxyle ($-COO^-$) conférant ainsi leurs charges négatives.

Enfin, le dernier groupe contient les trois derniers acides aminés Lysine (Lys K), Arginine (Arg R) et Histidine (His H). Ces molécules ont un groupement amine leur conférant une charge positive. L'Histidine est utilisée comme accepteur et donneur de protons (H^+), dans de nombreuses réactions enzymatiques au sein de processus biologiques.

La connaissance de la nature de chacun des acides aminés composant les protéines permet de mieux comprendre la réactivité et la fonction de celles-ci. Elle permet également de pouvoir mieux les observer et dans des conditions optimales. Notamment, les protocoles de préparation d'échantillons en protéomique fondés sur la spectrométrie de masse utilisent la trypsine, une enzyme dont la fonction est de cliver des séquences protéiques, après un résidu Lysine ou Arginine à condition qu'il ne soit pas précédé par une Proline. Ceci, en plus d'autres raisons évoquées par la suite de cette thèse, permet le maintien d'une charge électrostatique positive primaire à tous peptides nouvellement formés.

- **Une masse qui fait débat**

The amino acid masses*

1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C_3H_5ON	71.03711	71.0788
R	Arg	$C_6H_{12}ON_4$	156.10111	156.1875
N	Asn	$C_4H_6O_2N_2$	114.04293	114.1038
D	Asp	$C_4H_5O_3N$	115.02694	115.0886
C	Cys	C_3H_5ONS	103.00919	103.1388
E	Glu	$C_5H_7O_3N$	129.04259	129.1155
Q	Gln	$C_5H_8O_2N_2$	128.05858	128.1307
G	Gly	C_2H_3ON	57.02146	57.0519
H	His	$C_6H_7ON_3$	137.05891	137.1411
I	Ile	$C_6H_{11}ON$	113.08406	113.1594
L	Leu	$C_6H_{11}ON$	113.08406	113.1594
K	Lys	$C_6H_{12}ON_2$	128.09496	128.1741
M	Met	C_5H_9ONS	131.04049	131.1926
F	Phe	C_9H_9ON	147.06841	147.1766
P	Pro	C_5H_7ON	97.05276	97.1167
S	Ser	$C_3H_5O_2N$	87.03203	87.0782
T	Thr	$C_4H_7O_2N$	101.04768	101.1051
W	Trp	$C_{11}H_{10}ON_2$	186.07931	186.2132
Y	Tyr	$C_9H_9O_2N$	163.06333	163.1760
V	Val	C_5H_9ON	99.06841	99.1326

Figure 5 : Masse des acides aminés engagés dans une liaison peptidique. Cette image est extraite de la page Web, d'un projet étudiant (Isotopident) sur le site d'Expasy[1].

Les acides aminés ont des masses spécifiques, définies à partir de l'ensemble des atomes les composant (Figure 5). Les masses présentées ci-dessus sont calculées à partir de leurs formules chimiques dans le cadre d'une liaison peptidique, c'est-à-dire sans résidu H et OH de masse 18.01 daltons. La glycine est la molécule chimique la plus légère (masse moléculaire 57.02 Da) tandis que la molécule tryptophane, de composition chimique plus complexe, possède la masse la plus importante (masse moléculaire 186.07 Da). Dans la mesure où dans une expérience de spectrométrie de masse, la masse permet de discriminer les ions afin de déterminer leur composition en acides aminés, celle-ci doit être connue de façon aussi précise que possible. La mesure de masse laisse toutefois des incertitudes dans la discrimination entre les molécules qui possèdent des masses identiques comme la Leucine et l'Isoleucine. Ce qui est complexe à différencier de manière évidente par spectrométrie de masse doit être

[1] http://education.expasy.org/student_projects/isotopident/htdocs/aa-list.html

déterminé, au besoin, par un couplage avec d'autres approches d'études biochimiques ou par la mesure de l'encombrement stérique des molécules (comme la mobilité ionique). Cependant, les masses des acides aminés sont définies de manière uniforme mais ne sont pas toutes considérées de la même façon au sein de la communauté scientifique. En effet, les valeurs massiques déterminées sont une source de variabilité selon leurs définitions et le contexte. Par exemple, la valeur de masse d'une molécule est différente selon qu'on considère la masse moyenne, la masse monoisotopique ou enfin la masse des molécules en condition physiologique (pH 7.4). Ceci explique pourquoi il faut être vigilant lors de l'utilisation de différents logiciels. L'effet de cette imprécision de mesure sera détaillé dans les applications où cela a un impact important, notamment pour l'études des protéines intactes.

- **Diversité de production des acides aminés**

La production des acides aminés diffère selon les organismes qui possèdent des voies de biosynthèse spécifiques. Toutefois la plupart sont issus de métabolites produits par la glycolyse et le cycle de Krebs. En particulier, la levure *S. cerevisiae* a pour particularité d'être autotrophe. Elle est capable de produire tous les acides aminés nécessaires à sa croissance à partir du catabolisme d'une source carbonée, comme le glucose, le glycérol ou l'acétate (Figure 6). En revanche, chez les cellules eucaryotes complexes comme l'Humain ou le nématode, une source extérieure de certains acides aminés dit "essentiels" est nécessaire, ceux-ci sont dit auxotrophes. Chez l'Homme, ils sont au nombre de 10, il s'agit de la Méthionine, la Leucine, la Valine, la Lysine, l'Isoleucine, la Phénylalanine, le Tryptophane, l'Histidine, le Thréonine et enfin l'Arginine. La supplémentation est rendue possible par l'assimilation lors de la digestion des composés indispensables présents dans l'alimentation. Artificiellement, il est possible d'interrompre par délétion génétique des voies métaboliques essentielles à la production de certains acides aminés chez la levure, rendant les cellules dépendantes à un apport extérieur de ces molécules et permettant la sélection de souches particulières.

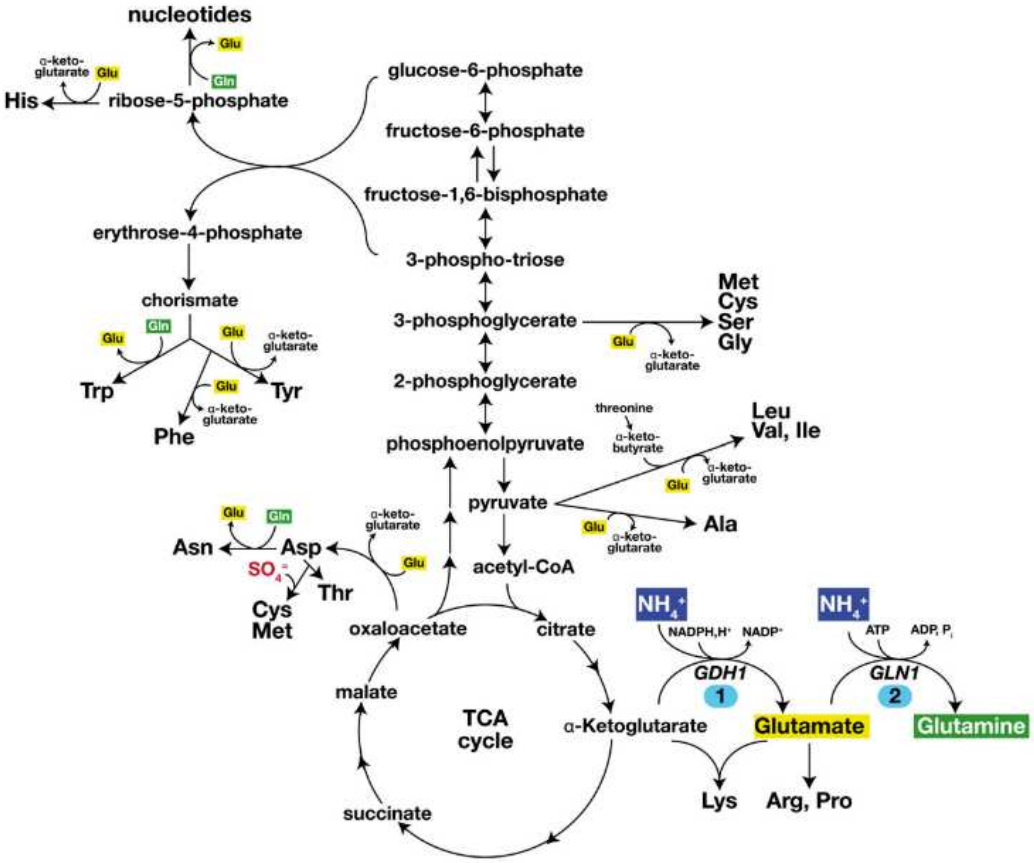


Figure 6 : Biosynthèse des acides aminés à partir du catabolisme du glucose chez la levure S. cerevisiae. Cette image est extraite de (Ljungdahl & Daignan-Fornier, 2012)

I.2.2 Rôle du pH

Les fonctions des protéines sont impactées par le pH (Figure 7). En effet, la valeur du pH conditionne l'état chimique des acides aminés (voir ci-dessus). Dans les cellules eucaryotes, le pH intracellulaire est un paramètre qui varie fortement : de 7.2 pour le noyau et le cytosol, à 8 pour la matrice mitochondriale (Casey et al., 2010). C'est ainsi qu'un acide aminé comme l'Histidine, dont la valeur du *pI* (point isoélectrique) est de 7.59, portera ou non une charge électrique, en fonction de sa localisation cellulaire. Par ailleurs, comme nous l'avons vu précédemment, la matrice mitochondriale possède, en plus d'une concentration forte en Acétyl-CoA, un pH élevé qui favorise très fortement de manière non-enzymatique l'acétylation des protéines (Wagner & Payne, 2013).

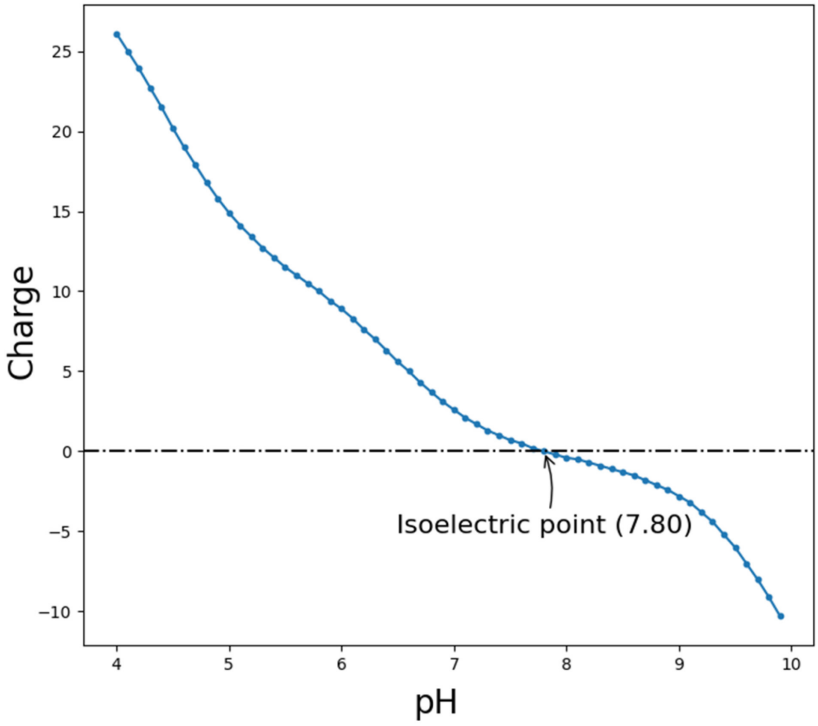


Figure 7 : Courbe de titration du suivi de la charge de la protéine Myoglobine de cheval estimée en fonction du pH données issues du site de modélisation Protcalc[1].

L'étude des protéines nécessite des étapes de préparation des échantillons, pour lesquelles une attention particulière doit être portée aux tampons utilisés (étapes d'extraction et/ou de purification), si l'objectif est de ne pas altérer l'état naturel des protéines. Ainsi, lors d'analyses par spectrométrie de masse, la préparation des échantillons est une étape très importante, qui conditionne la qualité finale des résultats. Il faut respecter les conditions physico-chimiques des objets biologiques, pour être en mesure de les observer de façon optimale. Généralement, la dénaturation des protéines, en prévision de leur digestion, s'effectue de manière douce par simple élévation progressive de la température. De plus, afin de séparer les ponts disulfures, une étape de réduction des protéines est réalisée par l'application d'agent réducteur comme le dithiothréitol (DTT), une molécule possédant deux groupements thiols (il en existe d'autre comme le TCEP, le beta-mercaptoéthanol, le DTE, etc..). Cette étape est suivie par une étape d'alkylation, durant laquelle l'échantillon biologique est mis en contact d'agents alkylants comme l'iodoacétamide (IAA), afin d'empêcher la

[1] http://protcalc.sourceforge.net/

reformation d'éventuels ponts disulfures entre les cystéines. Cela génère une carboxyamidométhylation, associée à une modification artificielle de masse (57.02 Da).

I.2.3 Bases de données de séquences et de structures des protéines

- **Les protéines à l'échelle du Protéome**

Le *protéome*, mot introduit par le chercheur australien Marc Wilkins en 1994 (contraction de PROTÉine et GénOME) est l'ensemble des protéines présentes dans une cellule à un moment donné et dans des conditions données. Une fois séquencé, un génome est décrit de façon relativement statique, car ne changeant que très peu. En revanche, la dynamique de l'expression des gènes, face à des contraintes environnementales et en réponse à des stimuli spécifiques, conduit à de multiples variations dans la composition et l'abondance des protéines. Le protéome est ainsi très dynamique. Son étude, la plus exhaustive possible, a pour objectif d'observer et de caractériser au niveau moléculaire le comportement d'un organisme en réponse à certaines conditions.

Les modifications post-traductionnelle (PTMs) modifient la structure des protéines, ce qui peut induire une modification de leurs fonctions (Figure 8). C'est d'autant plus critique que certaines fonctions sont des éléments essentiels de l'expression de nombreux gènes induisant ainsi certains phénotypes. La méthylation et la glycosylation sont des PTMs courantes d'ajout de groupement à des acides aminés. D'autre modifications existent, elles consistent en l'ajout de peptides ou de modifications des acides aminés composant la protéine initiale, comme l'ubiquitination. L'étude des mécanismes moléculaires ayant lieu lors d'adaptation ou en réponse à des conditions biologiques particulières permettrait de mieux caractériser les systèmes biologiques mis en jeux dans les organismes.

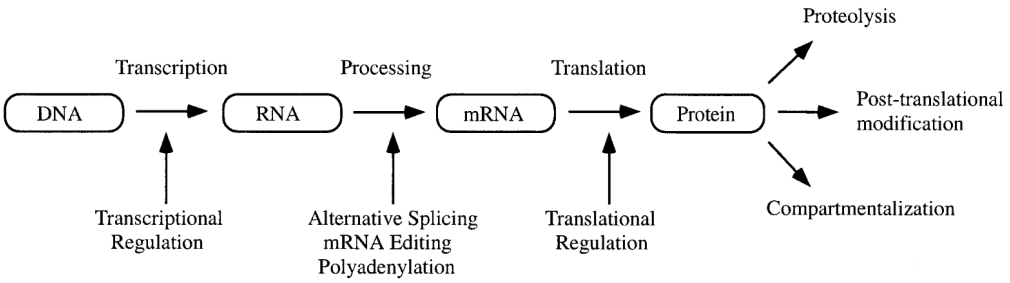


Figure 8 : Un gène exprimé peut donner de multiples formes de protéines, aux multiples fonctions. Cette image est extraite de (Graves & Haystead, 2002)

L'étude de la structure des protéines et de leur abondance dans les cellules permet d'analyser et d'interpréter l'état du métabolisme cellulaire, ainsi que d'observer les éventuelles perturbations ou adaptations, en fonction de conditions biologiques caractéristiques. Dans ce contexte, la communauté scientifique a créé plusieurs banques de données, plus ou moins exhaustives en fonction des organismes. Ces banques ont toutes des spécialités et permettent d'étudier les gènes et les protéines à plusieurs échelles.

- **Séquence protéique**

UniprotKB[1] est une banque de données qui rassemble, pour un grand nombre d'organismes, les séquences ainsi que des annotations fonctionnelles des protéines composant le protéome. Cette base de données regroupe d'une part des annotations protéiques expérimentales publiées (initiative SwissPot) et d'autre part des plus indirectes, telles que des annotations probables (initiative TrEMBL). La banque Uniprot regroupe également des informations provenant d'autres banques comme des informations génomiques, des données structurales, des informations d'observation issues d'expériences publiées. C'est notamment sur le site Web de cette base de données que l'on peut extraire l'ensemble des séquences protéiques prédites ou annotées du protéome d'un organisme d'intérêt.

- **Spécificités par organisme**

Recenser les séquences et les annotations fonctionnelles disponibles pour les protéines est une tâche très chronophage. Elle nécessite également une expertise importante. C'est la raison pour laquelle les communautés de chercheurs travaillant sur une même espèce mutualisent leurs efforts au sein de banques de données spécifiques. C'est le cas de la levure *S. cerevisiae* (base de données SGD[2]), de la levure pathogène *Candida albicans* (base de données CGD[3]) ou du nématode *Caenorhabditis elegans* (base de données WormBase[4]). Ces trois exemples correspondent à des espèces avec lesquelles j'ai travaillé au cours de ma thèse.

[1] UniprotKB https://www.uniprot.org/

[2] SGD: Saccharomyces Genome Database https://www.yeastgenome.org/

[3] CGD: Candida Genome Database http://www.candidagenome.org/

[4] https://wormbase.org/

Un autre point important est le recueil des structures tri-dimensionnelles des protéines. Protein Data Bank est une banque de données qui recueille les structures tridimensionnelles des protéines, obtenues de manière expérimentale, par cristallographie ou RMN. Ces informations sont utiles pour, par exemple, étudier la diversité des domaines structuraux des protéines et de mieux comprendre les relations entre conformation et possibles fonctions.

- **Informations sur l'abondance des protéines dans les organismes**

La banque de données PaxDb (Wang et al., 2012) recense et compile les abondances relatives des protéines dans les protéomes. Les données d'abondance sont exprimées en ppm (partie par million) et correspondent à une normalisation du nombre de protéines par cellule de l'organisme décrit. Ainsi, la protéine la plus abondante possède la plus forte valeur et la protéine la moins abondante (ou non-exprimée) une valeur nulle, la somme totale des données est égale à un million. Cette standardisation des valeurs permet la comparaison entre échantillons et entre organismes sans biais de la taille du protéome ou de l'abondance réelle. Les données sont majoritairement issues de données de protéomique quantitative par spectrométrie de masse. Les valeurs d'intensité (ou nombre d'ions, une explication de cette notion sera présentée par la suite) sont pondérées par la taille de la séquence protéique et par le nombre de peptides « détectables » par spectrométrie de masse (Wang et al., 2015).

- **Composition moyenne des protéomes**

Le Tableau 2 résume la taille du protéome pour 5 organismes modèles. Plus l'organisme modèle est complexe, plus le nombre de protéines est important. Par exemple, les organismes unicellulaires *S. cerevisiae*, C. *albicans* et *E. coli* ont un protéome de taille plus faible par rapport aux organismes multicellulaires *H. sapiens* et *C. elegans*.

Organismes	Souches	Taille du protéome	Protéines Révisées	Sources
E. coli	K12	4 448	4 400	https://www.uniprot.org/proteomes/UP000000625
S. cerevisiae	S288C	6 062	6050	https://www.uniprot.org/proteomes/UP000002311
C. albicans	SC5314	6 035	1 027	https://www.uniprot.org/proteomes/UP000000559
C. elegans	Bristol N2	26 548	4 352	https://www.uniprot.org/proteomes/UP000001940
H. sapiens	-	79052	20361	https://www.uniprot.org/proteomes/UP000005640

Tableau 2 : Composition du protéome de différents organismes.

Disposer de multiples banques de données dans lesquelles les compositions des protéomes de nombreux organismes sont stockées permet de mener des études globales de leur composition. Pour exemple, l'usage des différents acides aminés peut être comparé entre les organismes (Figure 9 ci-dessous). Ainsi, il est observé que l'acide aminé Sérine est moins présent dans le protéome de l'organisme procaryote *E. coli,* par rapport aux eucaryotes. Le protéome exprimé de l'humain est enrichi en acides aminés soufrés (Cystéine et Méthionine).

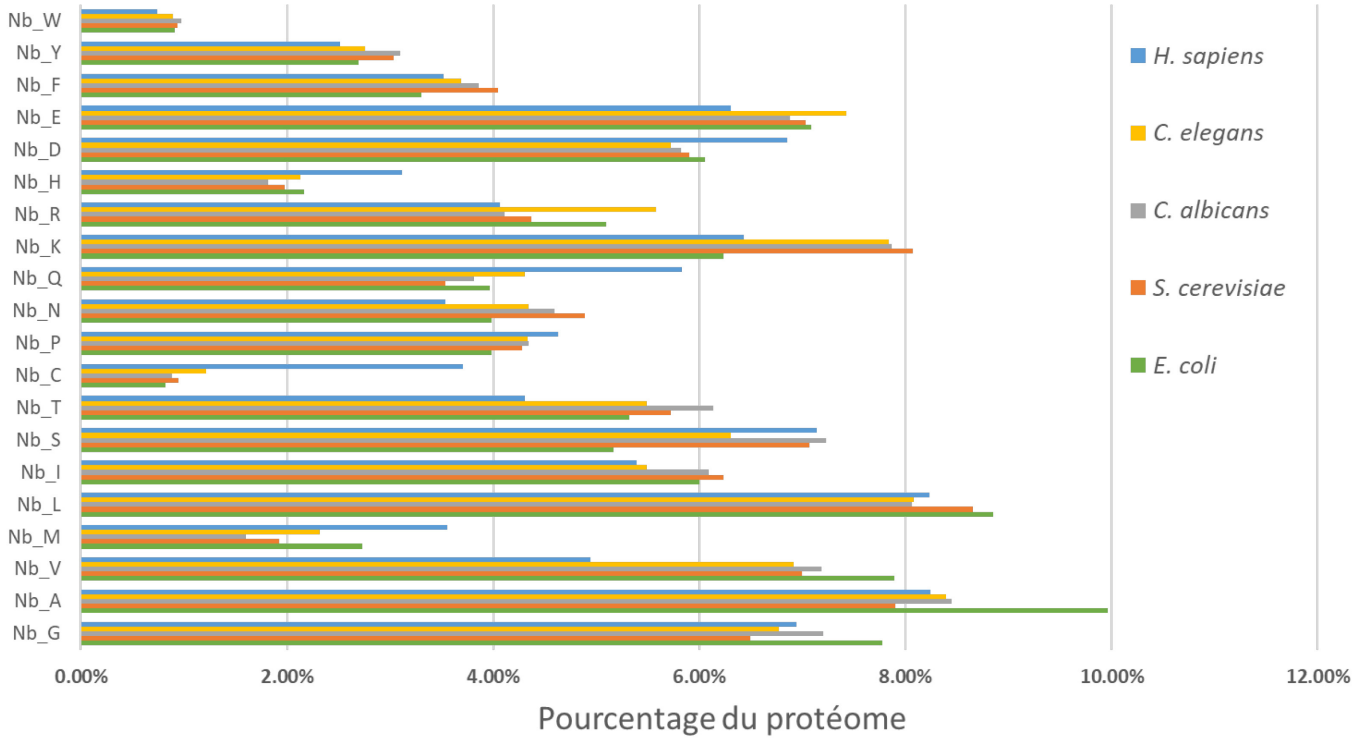


Figure 9 : Enrichissement des acides aminés dans les protéomes de différents organismes (représenté en pourcentage). Graphique créé à partir de données issues des banques Uniprot et PaxDb.

I.3 Prédiction de la structure des protéines à partir de la séquence en acide aminés

I.3.1 Les intérêts pour la médecine et les biotechnologies

Dans un contexte de santé publique où la résistance des souches bactériennes aux antibiotiques est fréquente et le développement de traitements anti-cancéreux est un enjeu important, la recherche de nouveaux remèdes thérapeutiques est cruciale. En particulier, la possibilité de prédire la structure des protéines à partir de la séquence en acides aminés ouvre des perspectives intéressantes, telles que la conception de nouveaux médicaments ayant une grande affinité (*drug design,* efficacité ligand-cible plus forte) et une haute sélectivité (spécificité ligand-récepteur), limitant ainsi les effets secondaires. En effet, une étude précise de la structure de la protéine cible permet de réaliser une recherche exhaustive de molécules candidates, pouvant reconnaître spécifiquement la protéine cible et impacter sa fonction. Pour cela, un *screening computationnel* (McInnes, 2007) permet de trouver les ligands les plus à même d'être efficaces contre une cible. Les paramètres du modèle de sélection et de validation

peuvent être fondés d'une part sur la structure des molécules, et d'autre part sur le type de liaisons qu'elles mettraient en place. D'ailleurs, une approche hybride permet d'allier les deux méthodes afin d'obtenir des résultats plus probants (Lin et al., 2020).

La prédiction de la structure des protéines peut permettre des avancées en biotechnologie. Si les protéines enzymatiques représentent une grande diversité en termes de fonctions, les types de réactions catalysées, elles, sont spécifiques. Ainsi, l'évolution par un mécanisme de sélection naturelle a permis de retenir des enzymes performantes et permettant d'effectuer des réactions chimiques dans des conditions extrêmes (température, pression, milieu) et de manière peu coûteuse en énergie pour la cellule. De nombreuses entreprises de biotechnologie utilisent les avantages de cette ressource biologique naturelle dans des secteurs d'applications très variés. Dans le but d'améliorer les performances des enzymes utilisées, une évolution dirigée peut être effectuée en laboratoire. L'une des stratégies possibles est un criblage, qui consiste à produire de manière aléatoire des séquences protéiques issues de mutations génétiques, lors d'amplification PCR à fort taux d'erreur (epPCR). Une autre stratégie utilise la capacité des phages à infecter et transmettre des plasmides chez les bactéries, *phage assisted continuous evolution* (PACE) (Esvelt et al., 2011). Durant ce procédé, la survie des bactéries porteuses de la séquence codant pour la synthèse de la protéine cible, est modulée par l'efficacité des mutations avantageuses accumulées par cette même enzyme. Ainsi, par une pression de sélection induite artificiellement et par une évolution dirigée, les bactéries survivantes sont celles qui produisent le plus efficacement la protéine recherchée. Les voies d'applications de ces méthodes sont variées, allant du secteur énergétique comme les bio-carburants ou la bioluminescence, au secteur médical pour la fabrication de médicaments à grande échelle ou de protéines recombinantes comme l'insuline.

I.3.2 Objectif majeur de la bio-informatique

En 1972, le Dr. Anfinsen obtient le prix Nobel en chimie pour son travail sur l'étude de la séquence et la structure de la protéine ribonucléase A. Une de ses théories, nommée le « dogme Anfinsen », propose que la structure des protéines est guidée exclusivement par leur séquence, c'est à dire leurs séquences en acides aminés (Anfinsen, 1972). Si ce principe est relativement solide pour des protéines de structure simple, il est plus fragile pour beaucoup d'autres, même si la thermodynamique permet que le repliement et la structure soit formée biologiquement *in-vivo* en quelques

minutes seulement. Mais la modélisation de repliement *in-silico* est une tâche complexe. En effet, une résolution purement thermodynamique (fruit de l'enthalpie libre) produit un nombre colossal de solutions possibles, dont la validation requiert un temps infini. C'est le paradoxe décrit par le biologiste Cyrus Levinthal (Levinthal's paradox). Mieux comprendre les mécanismes de repliements des protéines est un objectif de recherche qui se décompose en deux parties. La première vise à modéliser le plus fidèlement possible la structure tridimensionnelle finale de la protéine, tandis que la deuxième vise à trouver la séquence temporelle des événements qui permettent le repliement des protéines (Scheraga et al., 2006). Ces deux concepts de recherche sont complémentaires, par le fait que la résolution du premier aide à conceptualiser plus efficacement le deuxième (Dill et al., 2008).

Ces 20 dernières années, la prédiction de la structure tridimensionnelle des protéines, à partir de l'information de leurs séquences, s'est grandement améliorée grâce à l'essor de la bio-informatique. La résolution de calculs considérés comme impossibles en un temps raisonnable peut désormais être réalisée. Les méthodes de bio-informatique pour la prédiction de la structure des protéines se décomposent en deux familles : la modélisation par homologie et la modélisation *de novo*.

- **Modélisation par homologie**

Si la prise en compte de l'homologie entre séquences a été l'objet de nombreux développements, l'utilisation de la structure en parallèle permet d'obtenir des résultats plus pertinents (Carpentier et al., 2019). L'une des observations est que, évolutivement, la structure des protéines est beaucoup plus conservée que la séquence (Illergård et al., 2009). Cela peut s'expliquer par le fait que la mutation de la séquence d'acides aminés a moins d'effet sur la fonction de la protéine, par rapport à une modification de sa structure. C'est pourquoi la première approche de résolution est fondée sur l'utilisation de structures déjà connues et homologues à la protéine étudiée. Cette approche euristique utilise l'existant pour conceptualiser la structure inexistante par homologie. Elle part du postulat que si deux protéines sont très semblables en séquences d'acides aminés, alors, elles pourraient être similaires en termes de structures. Actuellement la banque de structure des protéines (PBD) contient plus de 188 000 structures déterminées expérimentalement par Cristallographie, Microscopie Electronique ou par RMN (Résonance Magnétique Nucléaire). La recherche d'homologie de domaines et de motifs structuraux conservés est faite par alignement de type PSA (Pairwire

Sequence Alignement). Mais également, et d'une manière générale, le MSA (Multiple Sequence Alignement) permet d'effectuer la recherche de domaines entre plusieurs séquences protéiques d'espèces différentes. La recherche d'homologie entre les protéines est actuellement une tâche computationnellement aisée, puisque de nombreux outils existent et ont été développés pour les études génomiques. De plus, les protéines possèdent des résidus très conservés puisque très fortement impliqués dans la structure.

- **Modélisation *de novo***

La stratégie *de-novo* (ou *ab initio*), a pour but d'effectuer une modélisation de la structure d'une protéine exclusivement à partir de l'étude des interactions entre les résidus (les acides aminés) qui composent la séquence protéique. La conformation tridimensionnelle de chaque domaine structural est ainsi déterminée en appliquant des règles de la physicochimie de la séquence protéique et en considérant les interactions entre les différents atomes composant les acides aminés. Tout cela est indépendant d'éventuelles autres structures préalablement déjà déterminées. Cette méthode stochastique est intéressante car elle se libère de tout *a priori* sur les données et permet de modéliser des protéines, qui n'ont jamais pu être observées expérimentalement, les protéines membranaires en particulier.

- **Modélisation soutenue par la science participative**

Le biochimiste David Backer et son équipe s'intéressent tout particulièrement à cette problématique et encouragent un usage des deux familles de méthodes de modélisations. Cependant, si la modélisation de la structure finale de la protéine est obtenue, la modélisation de son processus de repliement reste souvent à définir. Afin de progresser sur ce questionnement, ce groupe de recherche a développé un projet participatif de partage de ressources computationnelles. Le projet est nommé Rosetta@Home, et il permet d'effectuer des simulations numériques sur des ordinateurs dispersés partout dans le monde. Dans une même optique, le jeu foldit a été développé, proposant une méthode ludique et originale d'étude du repliement des protéines. Ce jeu utilise la capacité d'apprentissage et de cognition de l'être humain pour répondre aux défis du repliement des protéines. L'initiative a pour objectif d'étudier également comment le cerveau humain fait face à ce type de problème, travaillant souvent de manière originale et sans biais artificiel, contrairement à un algorithme. C'est un exemple emblématique de science participative. A noter qu'un

module complémentaire a été ajouté et permet de créer des automatisations personnelles de séquence de repliement appelées « recettes ». Ces recettes peuvent se partager entre utilisateurs, et correspondent finalement à de véritables algorithmes de repliements inédits. En particulier, l'un d'entre eux, noté comme très performant a même été publié (Khatib et al., 2011).

I.3.3 AlphaFold et son intelligence artificielle

- **Compétition CASP**

La compétition CASP (Critical Assessment of protein Structure Prediction) initiée par John Moult, a pour objectif de comparer les performances des modèles de prédiction des structures protéiques (Mosimann et al., 1995). Le principe est simple. Dans une première étape, les différents concurrents utilisent leurs algorithmes de prédiction et soumettent une structure pour un ensemble de protéines, dont seules les séquences en acides aminés sont publiques. Dans un second temps, les organisateurs du concours rendent publique les structures expérimentalement déterminées par cristallographie des mêmes protéines, et calculent des scores d'homologie (distance globale) avec les modèles prédits par les compétiteurs. Plusieurs critères sont utilisés afin d'évaluer au mieux la prédiction, notamment à partir des différents domaines structuraux qui composent la protéine cible. Le gagnant est l'équipe qui a proposé une structure qui ressemble le plus à la structure expérimentale.

La société DeepMind, appartenant à Google, est célèbre pour ses algorithmes utilisant une approche d'apprentissage supervisée pour résoudre des problèmes computationnels ouverts (par exemple AlphaGo pour le jeu de Go (Silver et al., 2016)). En 2018, lors de la compétition CASP 13 (Kryshtafovych et al., 2019), une équipe de cette entreprise présente AlphaFold, un algorithme utilisant du *deep-learning* pour prédire la structure tri-dimensionnelle des protéines (Senior et al., 2020). La communauté des chercheurs en biologie structurale a ainsi pu observer de manière flagrante les performances d'une stratégie de résolution fondée sur le *deep-learning* (Pearce & Zhang, 2021b, 2021a).

L'algorithme d'AlphaFold procède tout d'abord au traitement d'une carte de contact correspondant à un alignement des différents acides aminés composant la séquence. Puis, l'apprentissage préalablement effectué permet de produire une carte

de distance, présentant le schéma prédictif de la distance entre domaines structuraux. La deuxième partie de l'algorithme utilise une approche de type *de novo* afin de prédire la position des angles de chaque liaison peptidique. Un modèle neuronal a été préalablement entrainé en réunissant les nombreuses structures expérimentales disponibles sur le site de PDB.

Lors de la compétition suivante, CASP 14 en mai 2020, la deuxième version d'AlphaFold (AlphaFold2 (Jumper et al., 2021)) termine première du classement (Figure 10) (Kryshtafovych et al., 2021), avec un score parfois supérieur à 90% d'homologie avec la structure expérimentale. A cette valeur, la variation de la distance de chaque atome entre le modèle structural théorique et la structure expérimentale est inférieure à 2 angströms (inférieure à la distance naturelle entre les atomes d'un composé).

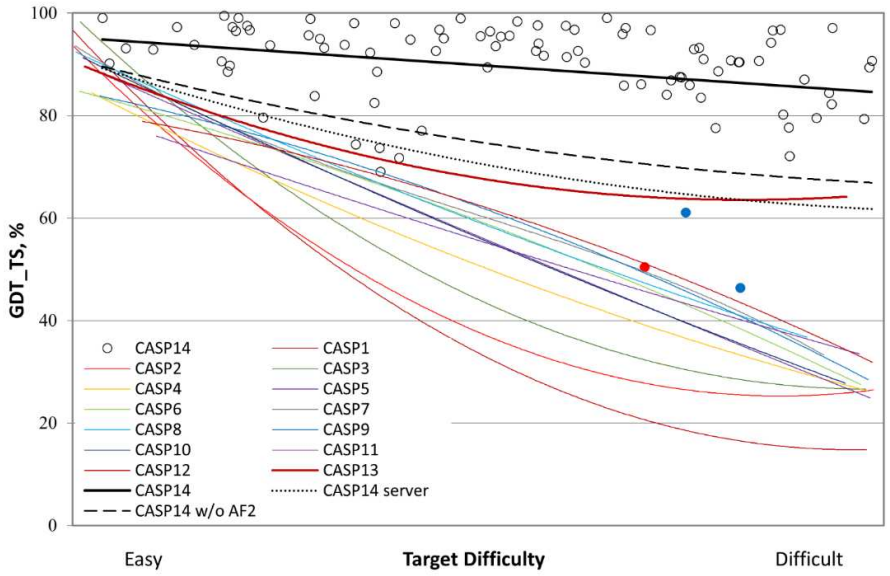


Figure 10 : Graphique représentant l'évolution des scores globaux des modèles prédictifs lors des compétitions CASP (Kryshtafovych et al., 2021).

AlphaFold2 est très différent de la première version publiée deux ans plus tôt. La modification la plus importante réside dans le fait que l'algorithme a été beaucoup plus entrainé. Ainsi, si dans AlphaFold la carte distance était fondée sur l'exécution d'une classification de type *Convulationnal Neural Network,* dans AlphaFold 2, toutes les séquences étudiées par la MSA, sont considérées comme pertinentes et utilisées

comme des données d'entrée pour le module suivant. De plus, une étape de correction a été ajoutée afin de corriger les sous-domaines modèles par une structure calculée de manière *de novo* à l'aide d'un réseau neuronal permettant le calcul de la position de chaque acide aminé qui avait tendance à être surestimée dans la version précédente. Pour chaque structure prédite, AlphaFold présente tout au long de la séquence un niveau de confiance correspondant au score obtenu lors du calcul des angles entre chaque résidu (Figure 11).

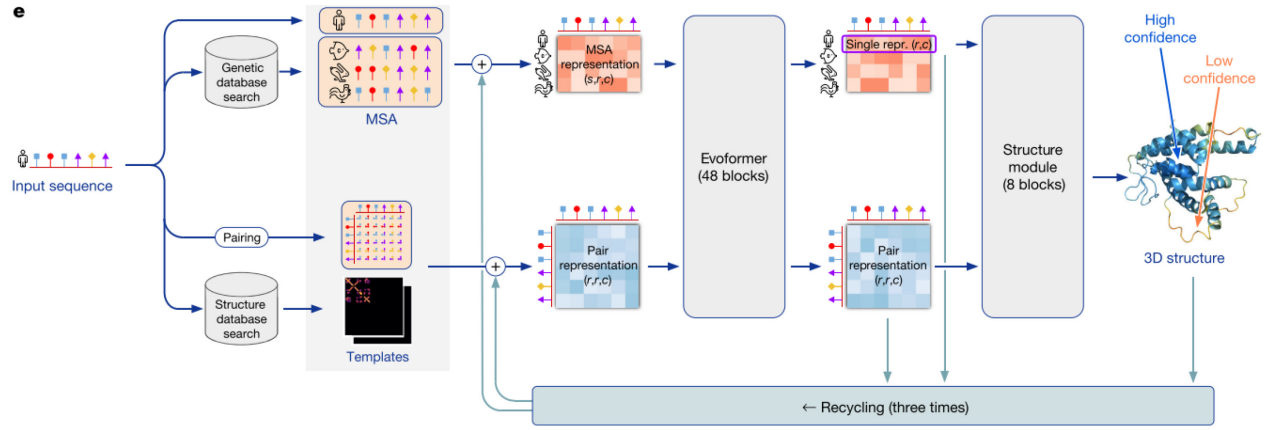


Figure 11 : Représentation schématique du fonctionnement d'AlphaFold2 (Issu de l'article original de l'équipe (Jumper et al., 2021)).

- **Nouvelles informations disponibles**

Cet algorithme est considéré comme révolutionnaire dans le domaine de la biologie structurale et de nombreuses applications utilisant ce modèle sont en cours de développement, telles que des applications de type *drug design* (Diabètes de type 2) ou des études macroscopiques des protéomes. En particulier, (A. Y. Chen et al., 2022) est une étude qui détaille la détermination du *pKa* des protéines à l'aide des données de modélisation récemment publiées par AlphaFold. Une autre étude (Spitz et al., 2022), est appliquée cette fois à l'étude de la structure d'une protéine (toxine hémolysine) lors de sa sécrétion par la bactérie *E. coli*. Les auteurs participent à une approche hybride par une description structurale d'une part grâce au prédiction d'AlphaFold mais également à l'aide d'études fonctionnelles grâce à des expérience *in silico*.

Récemment, l'équipe de DeepMind a également publié des améliorations de l'exhaustivité de leur base de données, notamment concernant la modélisation de la

quasi-totalité (98%) du protéome humain (Tunyasuvunakool et al., 2021). Par ailleurs, les conclusions des publications et des conférences récentes de l'entreprise font référence à des travaux actuels et futurs. Outre l'amélioration du modèle existant, l'algorithme est en cours d'optimisation pour répondre à des problématiques adjacentes. En particulier, la modélisation de complexes protéiques « multichain » qui reste une thématique non résolue (Evans et al., 2021)

- **Limitation du modèle prédictif**

Cependant, les prédictions d'AlphaFold sont à considérer avec prudence, en particulier lorsque le score de certains domaines est bas (inférieure à 70). Par exemple (Azzaz & Fantini, 2022) utilisent l'interface Robetta (Baek et al., 2021) un autre algorithme de prédiction issue du projet *Rosetta* soutenu et développé par l'équipe de David Backer (Ovchinnikov et al., 2018) leur permettant de mettre en lumière cette limitation (Figure 12).

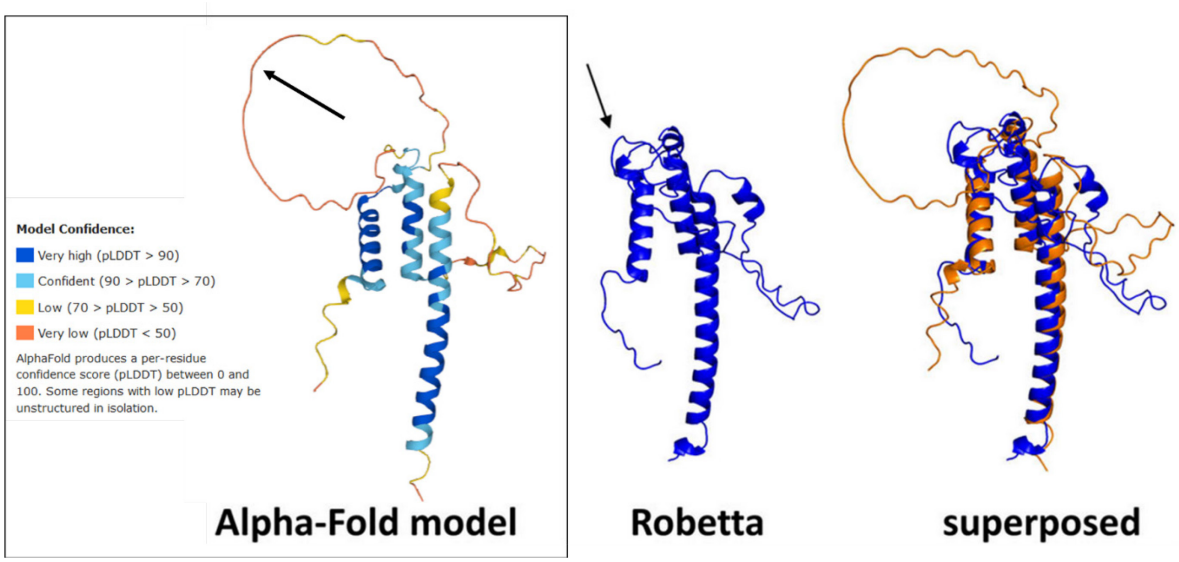


Figure 12 : Mid1- interacting protein 1 (Uniprot #Q9NPA3) Figure Adaptée de (Azzaz & Fantini, 2022).

II. Utilisation de la spectrométrie de masse pour étudier les protéines

II.1 Principal général

II.1.1 Protéomique et spectrométrie de masse

L'étude des protéomes a pour objectifs 1) d'identifier les différentes protéines présentes dans un échantillon à un moment donné et dans des conditions données et 2) de quantifier leurs abondances respectives. Ainsi, la dynamique de l'expression des gènes est observée. La structure des protéines ayant un rôle déterminant sur leurs fonctions (voir section précédente), cela représente également une donnée essentielle à prendre en compte. Pour étudier les protéomes, il existe plusieurs stratégies expérimentales. Elles sont décrites dans ce chapitre.

- **Méthodes d'analyses protéomiques**

La première stratégie expérimentale est fondée sur les méthodes d'analyses biochimiques, c'est à dire les approches qui exploitent les caractéristiques physico-chimiques des éléments à discriminer (ici des protéines) afin de les identifier et de les étudier. Ce type d'expérience permet la mesure des protéomes à petite échelle, c'est à dire par petits groupes de protéines ou à l'échelle de la molécule unique. Dans ce contexte, la méthode classiquement utilisée est une électrophorèse de protéines de type SDS-PAGE (Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis) en deux dimensions. Cette méthode permet une première séparation des composés selon leur valeur du point isoélectrique, puis dans un deuxième temps, une séparation selon leur valeur de masse moléculaire. Une séparation simple selon la valeur de la masse permet également d'identifier une protéine (à condition qu'elle soit très abondante par rapport aux autres composés de l'échantillon). Si la masse permet de localiser de façon macroscopique les bandes d'intérêt, seule une technique de révélation par sélection spécifique permet son identification/quantification. C'est le cas des "dot-blot" de protéines (Western Blot) durant lesquels un transfert sur une membrane des protéines après migration est réalisé, puis un anticorps primaire spécifique de la protéine d'intérêt est utilisé. À la suite de cela, un anticorps secondaire spécifique permet la révélation grâce à système de visualisation comme par exemple le rayonnement ultra-violet.

Une autre méthode utilisée est l'ELISA (Enzyme-Linked Immunosorbent Assay), fondée sur une méthode immunologique, ce qui rend la détection sensible et très spécifique. Initialement prévue pour une quantification d'antigènes en solution (fluide biologique), cette méthode est applicable aux protéines par la réalisation d'une gamme étalon très utilisée en biochimie clinique.

Un élément important à prendre en compte est que ces méthodes sont sensibles aux protocoles de préparation et restent coûteuses, d'une part en temps (migration) et d'autre part financièrement (les anticorps coûtent chers). De plus, ces méthodes ne permettent l'identification et le dosage que de petits ensembles de protéines. Les résultats restent des mesures qualitatives entre conditions et non quantitatives comme on pourrait le souhaiter. Enfin, les méthodes biochimiques ne permettent pas de réaliser des études structurales, ni de discriminer les différentes formes de protéines, en fonction de la présence d'éventuelles modifications post-traductionnelles (PTMs) et de leurs positions dans la séquence en acides aminés.

Une mesure des niveaux d'expression de certaines protéines, au niveau macroscopique peut-être réalisée par des appareils comme les puces à protéines (*proteins array)*(Hall et al., 2007). Permettant un criblage à haut débit, cet outil est expliqué par la spécificité des anticorps utilisés.

Enfin, les méthodes fondées sur des techniques de chromatographie consistent en une phase ayant des propriétés de séparation par rétention des différents composants de l'échantillon étudié. Le suivi de la trace d'élution à la sortie de la colonne est réalisé par une mesure UV (280 nm), dépendante des cycles aromatiques de la tyrosine et du tryptophane. En effet, la phénylalanine, autre acide aminé possédant un cycle aromatique de coefficient d'extinction molaire n'absorbant significativement qu'à 270nm max.

Une autre méthode d'analyse protéomique se fonde sur l'observation *in vivo* de la localisation et du niveau d'expression des protéines à l'échelle de la cellule, par

l'utilisation de fluorophore en imagerie microscopique à fluorescence. La protéine d'intérêt est ainsi marquée avec un tag fluorescent tel que la GFP (Green Fluorescent Protein), une protéine de 26 886 Da issue de la méduse *Aequorea victoria*. Cette protéine possède un assemblage de trois acides aminés (Sérine, Tyrosine et Glycine) générant un chromophore de structure chimique capable d'émettre des photons lumineux (509nm) sous l'effet d'excitation à une longueur d'onde spécifique (395nm ou 470nm) (Cody et al., 1993). Les microscopes à fluorescence permettent de les observer. Cependant, cette protéine est d'une taille massive, ce qui implique au niveau biologique un comportement proche mais tout de même écarté de la condition naturelle. Une mesure de la quantité par la valeur de l'intensité de la fluorescence peut tout de même être mise en place.

Toutes les méthodes présentées ci-dessus permettent d'étudier de manière précise et analogique des protéines *dans des échantillons simples*. Afin d'étudier un ensemble de protéines en mélange complexe, c'est-à-dire à plus grande échelle, la protéomique fondée sur de la spectrométrie de masse est utile, car élargissant l'analyse à l'échelle du protéome entier, avec une grande profondeur. La "protéomique MS" est aujourd'hui une technique d'analyse essentielle en biologie permettant la caractérisation de la structure et de la fonction des protéines.

- **Méthode de séparation des composants de l'échantillon**

Les spectromètres de masse sont composés d'une source d'ionisation, d'un analyseur de masse et d'un détecteur d'ions (Figure 13). L'ionisation est une étape essentielle, d'autant plus critique qu'elle permet de faire la transition d'échantillons en phase liquide à une phase gazeuse et par la même occasion d'apporter des charges électriques, rendant possible la manipulation de leurs trajectoires à l'aide de charges inverses. L'analyseur de masse permet la séparation des molécules selon le rapport masse sur charge (m/z). Le détecteur d'ions mesure une quantité ionique équivalente à une intensité pour chaque masse. La combinaison de ces trois éléments forme un spectromètre de masse, un outil puissant permettant la détection de molécules très peu abondantes et de manière la plus exhaustive possible dans le cadre de la limite de détections. Les molécules chimiques composant l'air ambiant serait un obstacle à l'analyse, c'est la raison pour laquelle un environnement sous vide est nécessaire.

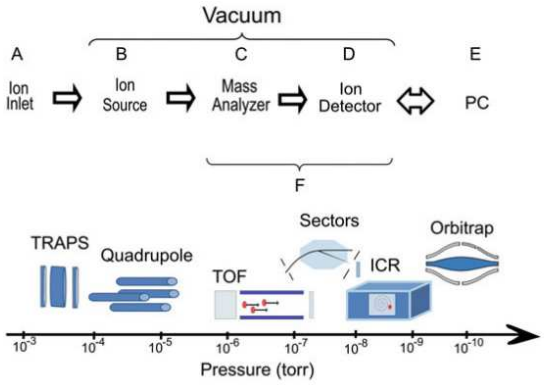


Fig. 1 Simple mass spectrometer overview. (**a**) Ion inlet: In shotgun proteomics peptides are separated by liquid chromatography either off line or on line. (**b**) The ion source converts molecules from solid or liquid phase into ionized species in the gas phase. (**c**) The mass analyzer separates the ions according to their mass-to-charge ratio. (**d**) The detector records a signal that is electronically amplified and stored as *m/z* versus intensity. (**e**) A computer typically controls liquid flow, vacuum, electrostatic fields, records data, and processes data during a sample run. (**f**) The mass analysis of ions takes place at low pressures typically 10^{-5}–10^{-10} mbar depending on the mass analyzer used

Figure 13: Représentation du fonctionnement d'un spectromètre de masse classique. Cette figure est extraite de (Matthiesen, 2020)

Sur la période pendant laquelle mon travail de thèse a été réalisé, j'ai pu travailler avec plusieurs spectromètres de masse, disponibles au sein de la plateforme protéomique de l'IJM. Ils avaient le point commun d'être tous fondés sur la technologie Orbitrap (Fusion, Qexactive Plus), développée par l'entreprise Thermo Fisher Scientific. Le principe général du fonctionnement de ce type d'analyseur est présenté en détail, dans la partie suivante. Certains résultats, obtenus au cours de ma thèse, sont issus d'expériences réalisées sur d'autres machines, *tims-TofPro 2* (Bruker Daltonics) notamment. Cet appareil a été installé plus récemment sur la plateforme (Octobre 2021), et exploite une technologie supplémentaire de séparation des ions (mobilité ionique) et une mesure de masse différente.

L'ionisation est la première étape lors d'une analyse par spectrométrie de masse. C'est la plus critique car tous les systèmes de mesure de masse en analyse protéomique procèdent à une caractérisation de la masse sur la charge des ions. La plupart des systèmes d'acquisition effectuent leurs mesures sur des ions chargés positivement. En effet, les peptides ou protéines possèdent généralement des charges positives en N-terminale et par la présence de Lysine (K) et d'Arginine (R). C'est la raison pour

laquelle les échantillons sont d'une part, analysés en condition acide, et d'autre part traités biochimiquement afin de favoriser l'apport de charge positive.

Plusieurs types de sources d'ionisation existent, l'Electrospray Source Ionisation (ESI) (Yamashita & Fenn, 1984) est la méthode classiquement utilisée en protéomique. Cette technologie est fondée sur le principe physique de nébulisation (transition liquide en nuage électronique (cône de Taylor)). Ce type de système encourage la réception de charge électriques (protons) car les peptides sont soumis à cette dépolarisation forte. Le débit en amont de l'ESI ainsi que la concentration en électrolyte composant le flux, influent sur la détection (sensibilité) et sur la séparation en masse dans le spectromètre. Le système présente de nombreux avantages : maintenance peu contraignante ainsi qu'une plus-value dans l'analyse de molécules polaires. Un autre aspect de la source est d'évaporer le solvant (l'acétonitrile classiquement utilisé est très volatile) afin de concentrer les molécules d'intérêt. Les paramètres de la source sont déterminants pour l'analyse en cours et pour le retraitement des données *a posteriori* (perte de spray ou fragmentation dans la source). En particulier, l'équilibre entre tension électrique de désolvatation (voltage), température des éléments, densité des gaz pour « ioniser » correctement sans fragmenter (démontré en Top-down). De la même manière, il faut veiller aux protocoles d'analyse durant l'acquisition. Les limitations résident dans le cas où les molécules sont apolaires, la technique perd en sensibilité. Les molécules s'ionisent de manière hétérogène, induisant un biais lors de l'analyse quantitative. La condition chimique influe grandement sur le processus d'ionisation. La source est à l'origine de plusieurs artefacts, comme des adduits de masse sur les molécules, provenant du solvant utilisé. Durant une analyse de peptides/protéines, les adduits de molécules de sodium et ammonium sont très fréquents. L'une des difficultés lors de l'étape d'ionisation est le phénomène de *suppression ionique*. Dans ce cas, les ions sont mélangés à d'autres ions possiblement mono-chargés, comme des polluants plastiques (lors de la préparation de l'échantillon) ou des résidus de solvant (lors de la séparation de l'échantillon). La suppression ionique est à réduire au maximum, car cela impacte très fortement les mesures de l'intensité des ions, la précision des mesures en masse mais surtout l'exhaustivité de l'identification lors du traitement informatique des données.

Lors de la calibration du spectromètre de masse, on procède à une injection continue d'une solution « calibrante », c'est-à-dire une solution composée d'une mixture de molécules de masses et d'abondances connues. L'analyse d'un échantillon « simple » (composé de quelques protéines seulement) est également possible lors d'une injection directe. Ainsi lors du développement de nouvelles méthodes de mesure de masses l'injection directe est privilégiée puisque cela permet de s'abstenir de la complexité de la dimension temporelle. Cependant, une fois les méthodes d'acquisition du spectromètre développées, il est essentiel de réaliser les analyses en ligne, généralement à la suite d'une chromatographie séparative pour les échantillons complexes.

Les objets biologiques étudiés sont mélangés dans un échantillon complexe constitué de peptides ou de protéines. Cet échantillon ne peut pas être injecté dans le spectromètre de masse en une seule fois. En effet, le nombre d'ions présentés simultanément serait trop important et induirait une saturation des signaux à mesurer par le détecteur et donc une impossibilité d'individualiser les différents composés observés par le spectromètre de masse. Ainsi, en amont de l'analyse en masse, les différents objets biologiques qui composent un échantillon sont généralement séparés selon leurs propriétés physico-chimiques et élués en fonction du temps, par une technique séparative telle que la chromatographie. La chromatographie en phase liquide (LC) est la méthode séparative classiquement utilisée en analyse protéomique, en particulier dans un contexte de couplage avec un spectromètre de masse (montage LC-MS/MS). Cette méthode de séparation utilise une colonne en phase inverse pour fractionner les différents constituants d'un mélange, par exemple chaque peptide ou chaque protéine. Les composants sont élués en fonction du temps par rétention hydrophobique et l'utilisation d'un gradient de solvant organique polaire (généralement de l'acétonitrile). Plusieurs paramètres influent sur cette séparation, en particulier la température, le débit et la nature des solvants utilisés. Par ailleurs, la longueur de la colonne est importante pour améliorer la résolution de séparation. Plus la colonne est longue plus la séparation sera précise, mais au détriment du temps de rétention qui sera plus long. De même, la taille des molécules composant la phase (granulométrie de l'ordre de grandeur de l'angström) dans les colonnes influent sur la résolution, plus la granulométrie est fine plus la colonne sera résolutive.

Lors du traitement des données, les différents paramètres de séparation des peptides et protéines sont à prendre en compte afin de modéliser le plus fidèlement possible les différents profils d'élution des objets. En particulier, les outils bio-informatiques de retraitement des données (d'identification et de quantification) font appel à des algorithmes comportant de nombreux paramètres. Ces valeurs correspondent à la description du profil d'élution (temps de la trace d'élution, largeur à mi-hauteur, hauteur, profil gaussien, etc…).

II.1.2 Augmentation de la résolution des spectromètres de masse, implications pour les usages en biologie

Comme décrit précédemment, la spectrométrie de masse est une option expérimentale pour étudier le protéome dans sa globalité. Cela est d'autant plus pertinent que la mesure de la masse par les spectromètres de masse est précise. Dans ce contexte, des avancées importantes ont été réalisées ces dernières années. La première avancée réside dans le fait que les spectromètres sont de plus en plus sensibles et résolutifs. Ainsi, les mesures gagnent d'une part, en précision au niveau de la mesure des masses, et d'autre part en l'intensité (spectres de haute résolution). Ces mesures sont également obtenues avec une meilleure reproductibilité. De plus, des progrès importants ont été réalisés, vis-à-vis des méthodes d'ionisation, de fragmentation en MS/MS, ainsi que les méthodes séparatives telles que la chromatographie en phase liquide. La deuxième avancée est l'amélioration considérable des outils bio-informatiques disponibles pour analyser les données brutes. En effet, les ressources informatiques sont de plus en plus performantes, et les algorithmes d'identification et de quantification des protéines associent leurs résultats à des paramètres statistiques robustes et de plus en plus perfectionnées, tout cela étant obtenu avec des temps de calculs faibles. L'un des exemples les plus notables est l'utilisation d'algorithmes de *Machine Learning* dans le calcul des scores d'identification dans l'algorithme Percolator (Käll et al., 2007)).

Les appareils actuels sont donc hautement sensibles et possèdent des meilleures résolutions que les appareils précédents. Cela aide le développement de nouvelles méthodes d'étude et de quantification des ions. En particulier, l'utilisation de

l'abondance des isotopologues dans les spectres de masse est un exemple caractéristique, permis par ces progrès technologiques.

II.1.3 Les applications biologiques de la protéomique

Les utilisations possibles des techniques d'analyse protéomique par spectrométrie de masse sont nombreuses et trouvent leurs applications dans des secteurs divers. Dans le milieu hospitalier, en bactériologie notamment, la protéomique permet l'identification de pathogènes de manière rapide (Tsakou et al., 2020). C'est un élément important compte tenu du fait que la temporalité est une donnée critique pour une bonne prise en charge des patients par exemple dans le cas d'infections graves (septicémie en particulier). Puisqu'il n'est pas nécessaire d'isoler au préalable le pathogène (procédure coûteuse en temps) comme, par exemple, la culture bactérienne sur gélose (Kondori et al., 2021). Une autre application dans le secteur médical est l'identification de biomarqueurs afin de réaliser des diagnostics à partir de prélèvements moins invasifs sur les patients. Cela permet de dépister, de façon plus précoce, des maladies mais également, lors de chirurgies ablatives, de pouvoir réaliser une démarcation plus nette entre des zones saines ou cancéreuses de tissus, préservant ainsi des tissus qui habituellement (et par mesure de précaution) sont retirés (ce qui ralentit le temps de rétablissement) (Le Faouder et al., 2014).

- **Cluster Isotopiques**

Les acides aminés, entités de base qui composent les protéines (voir ci-dessus), ont des formules chimiques brutes composées des cinq éléments principaux : le carbone, l'hydrogène, l'azote, l'oxygène et pour deux cas particuliers (acides aminés cystéine et méthionine), le soufre. L'Averagine est un acide aminé de composition moyenne purement théorique de formule brute $C_{4.9384}$ $H_{7.7583}$ $N_{1.3577}$ $O_{1.4773}$ $S_{0.0417}$ (Senko et al., 1995). La masse moléculaire associée est de 111.1254 Dalton. Cette notion « d'acides aminés moyens » a été obtenue par l'étude systématique des compositions en acides aminés des protéines connues dans plusieurs organismes vivants. Cette notion est intéressante car elle quantifie les abondances des différents types d'atomes dans les acides aminés. Le carbone apparaît ainsi comme étant le deuxième élément chimique le plus abondant dans les acides aminés. Nous verrons

l'importance de cette observation pour la méthode d'analyse SLIM, utilisée et améliorée au cours de ma thèse.

Ainsi, tous les éléments chimiques possèdent naturellement des isotopes stables. L'abondance isotopique de chaque atome est généralement considérée comme constante (Berglund & Wieser, 2011). Concrètement ces abondances sont représentées par des probabilités d'observation de chaque isotope stable, au sein d'une molécule quelconque (Tableau 3). Cependant dans la réalité ces valeurs fluctuent selon les endroits et certains organismes modulent l'abondance de ces différents isotopes. Au cours de cette thèse nous admettrons ces valeurs moyenne comme fixée et ne variant pas de manière naturelle. En ce qui concerne les molécules du vivant, nous pouvons ainsi observer que l'élément chimique carbone, étant donné son nombre dans les acides aminés (en moyenne 4.92) et son abondance forte de présence d'isotopes (supérieure à 1%), est le principal contributeur à l'apport d'isotopes dans les molécules. Les autres éléments sont soit moins présents, soit possèdent une abondance isotopique plus faible les rendant moins contributeurs.

^{12}C = 98.93 %	^{13}C = 1.07 %		
^{1}H = 99.9885 %	^{2}D = 0.0115 %		
^{16}O = 99.757 %	^{17}O = 0.0373 %	^{18}O = 0.2057 %	
^{14}N = 99.632 %	^{15}N = 0.368 %		
^{32}S = 94.93 %	^{33}S = 0.76 %	^{34}S = 4.29 %	^{36}S = 0.02 %
^{31}P = 100 %			

Tableau 3 : Abondance isotopique de chaque élément chimique présent dans la matière biologique. Ces valeurs sont extraites de (Audi & Wapstra, 1993).

Ainsi, une même molécule peut avoir des masses différentes, expliquées par la présence d'isotopes dans sa composition, induisant un incrément de masse correspondant à la différence de masse entre ces isotopes (adduit de neutrons). Or la

masse des neutrons est différentes selon son origine (éléments chimiques) la différence de masse est présentée dans le Tableau 4.

élément	$masse\ neutron$
C	1.00335484
H	1.006276746
N	0.9970349
O	1.00421708
S	0.99938776

Tableau 4 : Masse du neutron en fonction des éléments chimiques

Les instruments actuels couramment utilisés sur les plateformes de spectrométrie de masse ne permettent pas de déterminer l'origine de ces adduit, mais seulement de les observer (incrément de masse). C'est la raison pour laquelle il est admis que la différence de masse entre deux isotopologues (donc un neutron) correspond à la masse d'un neutron issu du carbone entre le ^{13}C et le ^{12}C).

Une expérience de spectrométrie de masse peut être vue, dans le cas de l'observation d'une molécule donnée, comme un tirage aléatoire des différentes formes de cette même molécule (Figure 14). La loi de probabilité étant la distribution de l'abondance isotopique de chaque élément chimique multipliée par son nombre, analogue à une mesure de la fréquence d'apparition des isotopologues lors d'un tirage aléatoire. Pour une lecture plus détaillée, il est utile de lire (Matthiesen, 2020).

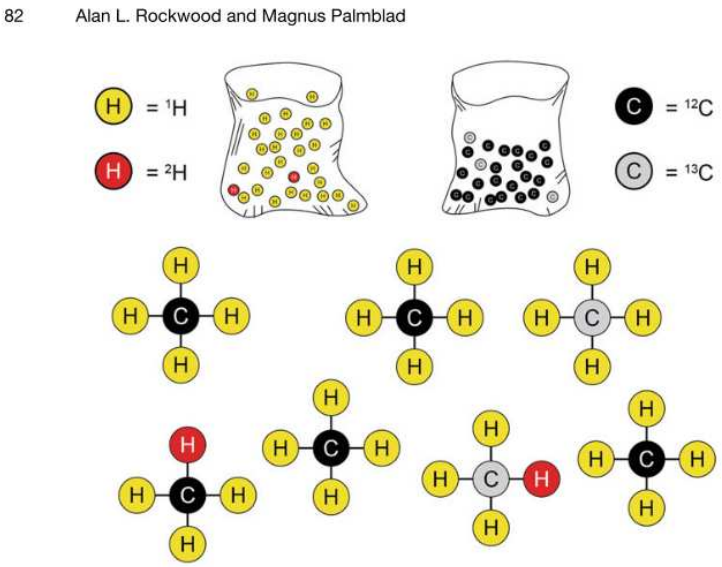


Fig. 1 Calculating isotopic distributions is a combinatorial problem and the branch of mathematics concerned with these types of problems is *combinatorics*. One can think of the task being analogous to building molecules by drawing differently colored balls representing the atoms from different bags, for instance methane (CH_4) by drawing four hydrogens and one carbon (returning each ball after it was drawn). Hydrogen and carbon have two isotopes each: $^{1}H/^{2}H$ and $^{12}C/^{13}C$. This gives many possible combinations of isotopes, or isotopologs, three of which are illustrated above. The problem of calculating or estimating isotopic distributions is generally about predicting the frequencies by which these isotopologs occur

Figure 14 : Illustration de l'analogie d'un tirage aléatoire d'atome chimique. Cette Figure est extraite de (Rockwood & Palmblad, 2020)

Comme expliqué précédemment, les spectromètres de masse actuels sont très résolutifs et sensibles. L'écart de masse possiblement observé est équivalent à la masse d'un neutron. Ceci permet d'observer pour chaque molécule, les différentes formes isotopiques de cette même molécule, se caractérisant par plusieurs pics de masses différentes, appelée "cluster isotopique" (Figure 15).

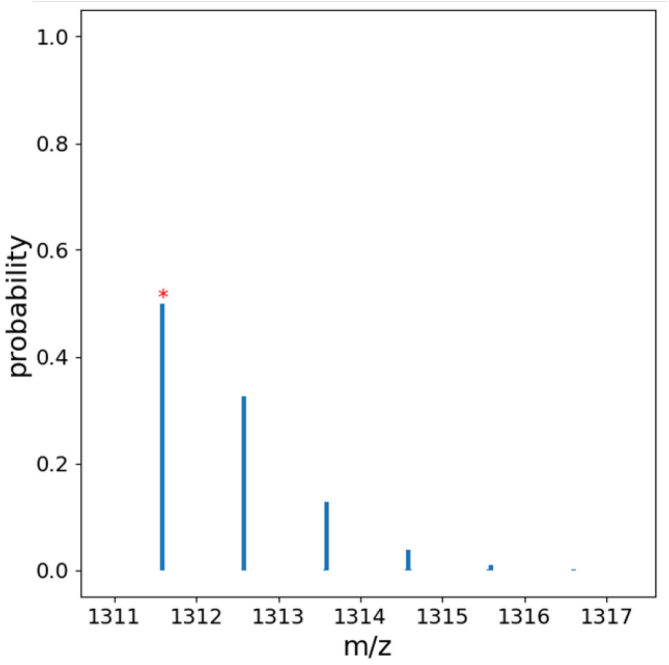


Figure 15 : Un cluster isotopique issu de la modélisation du peptide YGKPNTTDSNTN à l'aide de l'algorithme MIDAs. Cet algorithme (Alves et al., 2014) est présenté plus loin dans le texte, il a été utilisé à de multiples étapes de mes travaux de thèse.

Au sein d'un cluster isotopique, l'intensité de chaque pic, appelé isotopologue, correspond à la quantité des différentes formes « isotopiques » de la molécule. Le premier pic de masses (représenté par un astérix * sur la Figure 15) est appelé monoisotopique M_0. C'est un cas particulier puisqu'il révèle la quantité de molécules composées exclusivement des isotopes légers de chaque élément chimique, c'est à dire que tous les atomes de Carbone sont ^{12}C, tous les atomes d'Hydrogène sont ^{1}H, tous les atomes d'Azote sont ^{14}N, tous les atomes d'Oxygène sont ^{16}O, tous les atomes de Soufre sont ^{32}S et enfin tous les atomes de Phosphore sont ^{31}P. L'intensité de ce pic peut être simplement modélisée. En effet celle-ci est directement proportionnelle à la somme des probabilités d'obtenir pour chaque type d'atome (C, H, N, O et S), leur isotope le plus stable, élevée à la puissance de leurs nombres respectifs dans la molécule étudiée. Ainsi, si l'on note $P(^{12}C)$ la probabilité d'occurrence de l'isotope ^{12}C et nC le nombre d'atomes de carbone composant la molécule, on aura:

$$M_0 = P(^{12}C)^{nC} * P(^{1}H)^{nH} * P(^{14}N)^{nN} * P(^{16}O)^{nO} * P(^{32}S)^{nS} \qquad (1\text{-}1)$$

Les autres isotopologues, par définition, sont tous composés de la combinaison des différentes formes de molécules possédant un nombre variable d'isotopes. Dans les systèmes de mesure de masse courants, à une masse d'isotopologue donnée, il est impossible de savoir quels sont les isotopologues à l'origine d'un delta de masse observé. Il est uniquement possible de réaliser un calcul des lois de probabilités combinatoires et d'obtenir une probabilité de présence de certains isotopes. Ces calculs reposent sur des modèles de lois de probabilités polynomiales, et se complexifient d'autant plus que le nombre d'isotopes possibles augmente.

Ainsi, dans le cas du premier isotopologue (de rang 1) M_1, la probabilité (modélisant une intensité expérimentale) correspond à la somme des possibilités d'obtenir un isotope « lourd » de chaque atome sachant que les autres sont "légers". Le M1 possède un incrément de masse de 1 Dalton, dont l'élément chimique responsable de ce neutron supplémentaire est d'origine inconnue. L'intensité théorique normalisée est donc la combinaison de l'apport isotopique de chacun des éléments chimiques soit :

$$
\begin{aligned}
\mathrm{M1} = {} & nC * \mathrm{P}(^{12}\mathrm{C})^{nC-1} * P(^{13}C) * P(^{1}H)^{nH} * \mathrm{P}(^{14}N)^{nN} * P(^{16}O)^{nO} * P(^{32}S)^{nS} \\
& + \mathrm{P}(^{12}\mathrm{C})^{nC} * nH * P(^{1}H)^{nH-1} * \mathrm{P}(^{2}H) * \mathrm{P}(^{14}N)^{nN} * P(^{16}O)^{nO} * P(^{32}S)^{nS} \\
& + \mathrm{P}(^{12}\mathrm{C})^{nC} * P(^{1}H)^{nH} * nN * \mathrm{P}(^{14}N)^{nN-1} * \mathrm{P}(^{15}N) * P(^{16}O)^{nO} * P(^{32}S)^{nS} \\
& + \mathrm{P}(^{12}\mathrm{C})^{nC} * P(^{1}H)^{nH} * \mathrm{P}(^{14}N)^{nN} * nO * P(^{16}O)^{nO-1} * P(^{17}O) * P(^{32}S)^{nS} \\
& + P(^{12}C)^{nC} * P(^{1}H)^{nH} * P(^{14}N)^{nN} * P(^{16}O)^{nO} * nS * P(^{32}S)^{nS-1} * P(^{33}S) \qquad (1\text{-}2)
\end{aligned}
$$

L'intensité des autres isotopologues est plus difficile à obtenir de manière exacte, puisque le calcul décrit sous la forme d'une loi polynomiale prend en compte les différentes combinatoires d'occurrence des isotopes lourds.

L'une des conséquences les plus notables de cette complexité isotopique réside dans le fait que plus la masse des molécules observées augmente, plus elles sont composées d'atomes. Au vu du nombre croissant d'atomes, la probabilité d'observer des isotopologues "lourds" est vérifiée puisque, en suivant notre analogie, le nombre de « tirages aléatoires augmente. Ceci a pour effet de diminuer la probabilité d'observer des ions légers, représentés par le signal monoisotopique. Or sa masse, représentant la masse neutre de la molécule, est la valeur qui est utilisée pour effectuer une recherche informatique en banque de données, et ainsi permettre l'identification de la molécule.

Ainsi plus la masse augmente, plus l'intensité du signal monoisotopique diminue (Figure 16) ce qui a pour conséquence d'augmenter l'imprécision sur la mesure de la masse neutre, rendant plus complexe l'identification de la molécule. Ceci est une limitation qui est d'autant plus forte lors d'études de molécules complexes à haute masse telles que les protéines intactes.

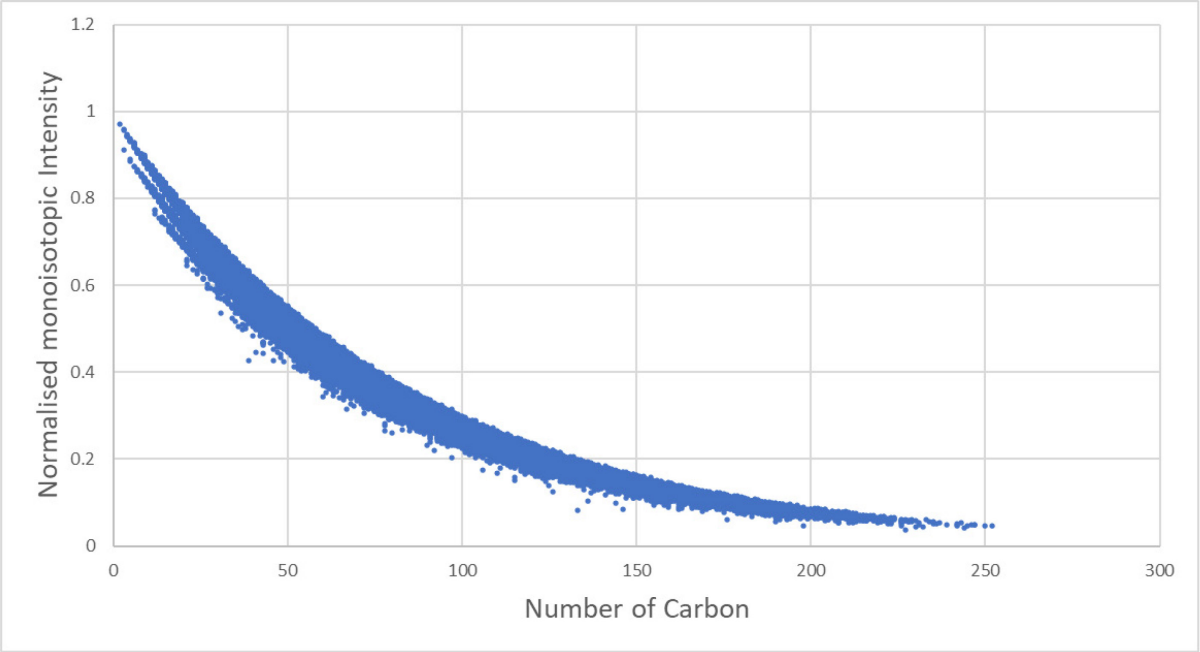


Figure 16 : Intensité normalisée de l'ion monoisotopique en fonction du nombre de carbone composant les peptides (peptidome de S. cerevisiae).

II.2 Mesure de l'intensité des isotopologues

II.2.1 Exemple d'un algorithme de modélisation des clusters isotopique théorique : MIDAs

Pour une molécule donnée la modélisation théorique d'un spectre de masse correspond à la détermination de la distribution des masses et des intensités correspondant aux isotopologues. La description des lois de probabilité d'apparition de chaque isotopologue dans un cluster isotopique (sous forme d'intensité normalisée) est une fonction mathématique combinatoire. Aussi, le défi réside dans la résolution d'un problème mathématique de complexité croissante, en fonction de leurs rangs.

Le logiciel MIDAs (*Molecular Isotopic Distribution Analysis*) (Alves et al., 2014) permet de réaliser ce type de modélisation des spectres de masse théorique. Il a été créé par des physiciens spécialisés en spectrométrie de masse, Gelios Alves et Aleksey Ogurtsov. Tous les deux travaillent au laboratoire *Quantitative Molecular Biological Physics* du NCBI (National Center Biotechnology Information), dirigé par Yi-Kuo Yu. L'algorithme MIDAs permet de résoudre des équations décrivant la distribution des isotopologues au sein d'un massif isotopique, de deux manières différentes. La première méthode procède à une résolution binomiale tandis que la seconde méthode utilise une résolution de type transformée de Fourier. Chacune de ces deux méthodes possède également deux niveaux de résolution, une proche de la réalité et nommée « coarse grain », et une deuxième hautement résolutive et nommée « fine grain ». Ainsi, MIDAs permet le calcul des intensités théoriques des isotopologues qui constituent un cluster isotopique, à partir d'une séquence d'acide aminés ou d'une formule chimique. L'algorithme est écrit en langage de programmation C++ et peut être utilisé *via* une version web[1]. Les données expérimentales utilisées par l'algorithme, telles que la masse et l'abondance isotopique de chaque atome, sont par défaut celles indiquées dans des tables de références en physique nucléaire (Audi & Wapstra, 1993, 1995). Il est toutefois possible de modifier ces valeurs, en changeant par exemple la probabilité naturelle d'occurrence des différents éléments chimiques. Cette possibilité est très utile pour analyser des données issues de la méthode SLIM (voir ci-dessous) et c'est une originalité de MIDAs par rapport aux autres algorithmes disponibles. De plus, le code source est « libre » a pu être implémenté dans d'autres algorithmes développés en C++ au cours de cette thèse.

L'obtention et l'utilisation de spectres théoriques sont des tâches centrales pour mon travail de thèse. En effet, ils permettent d'une part une observation fine du comportement des clusters isotopiques, et d'autre part les clusters isotopiques théoriques sont utilisés pour la réalisation de simulations numériques, dans un but de confronter à des données expérimentales.

[1] https://www.ncbi.nlm.nih.gov/CBBresearch/Yu/midas/index.html

II.3 Stratégies d'analyse « Bottom-up » et « Top-down »

II.3.1 Définition et introduction

- **L'étude des protéines intactes**

Nous l'avons vu précédemment, la fonction des protéines est étroitement déterminée par leur structure et par leur organisation dans l'espace (structures tridimensionnelles). L'ensemble des variations structurales d'une même protéine est appelé protéoformes (Schaffer et al., 2019). Une observation précise de la composition et des variations du protéome, en termes d'identification des protéoformes d'une part, mais aussi en termes de quantification d'autre part, permet de caractériser la dynamique du protéome. L'analyse d'un échantillon biologique par spectrométrie de masse est idéale pour décrire ces variations et pour répondre à ces objectifs. En particulier, l'étude des protéines intactes par spectrométrie de masse, l'approche Top-down, est la méthode d'analyse qui permet de conserver et d'observer les modifications post-traductionnelles des protéines. Toutefois, les protéines sont des objets massifs de haut poids moléculaire généralement entre 5-6 kDa et 50-150 kDa mais parfois jusqu'à 300kDa (Catherman et al., 2014). L'observation de tels objets par spectrométrie de masse a été initiée il y a déjà plusieurs décennies mais est restée limitée en raison de plusieurs contraintes.

La première contrainte a été technologique, puisque les spectromètres de masse de l'époque ne permettaient pas de réaliser des analyses avec une résolution suffisamment haute. Or, dans le cas de l'analyse de protéines intactes, les masses étant particulièrement élevées, leur observation et leur mesure précise sont particulièrement difficiles. C'est pourtant un critère strictement requis pour mettre au point cette approche.

De même, un échantillon biologique standard contient plusieurs milliers de protéines distinctes et d'abondances variables. La réussite de l'étude exhaustive d'un tel échantillon nécessite des protocoles de séparation et de fractionnement des protéines particuliers qui n'étaient pas assez développés. En particulier, les méthodes de séparation de type "online", "offline" ou "en infusion" n'étaient également que très peu optimisées pour réaliser les mesures des quantités sans biais.

Enfin, la dernière limitation demeurait dans les méthodes de traitement des données spectrales. En effet, la place centrale de la bio-informatique pour l'analyse des

données nécessitait d'une part des développements conséquents en termes de robustesse statistique des résultats, et d'autre part en termes de méthodes algorithmiques de traitement du signal (déconvolution, peak-picking, etc.).

- **Le « Bottom-up », une réduction de l'échelle de masse**

L'étude par spectrométrie de masse de petites molécules chimiques est bien réalisée et documentée. Ainsi, face aux limitations rencontrées dans l'analyse de protéines entières, la communauté scientifique a décidé de développer des protocoles d'analyse des fragments de protéines de plus petite taille, les peptides. Avec pour objectif de réduire la contrainte de masse, les protéines sont digérées en fragments peptidiques, pour lesquels le nombre d'acides aminés est bien plus petit. Cette stratégie d'analyse d'un échantillon est nommée "Bottom-up". Elle consiste donc en la mesure de la masse de peptides composés généralement de 6 à 22 acides aminés seulement. La procédure d'analyse est la suivante. Tout d'abord, le mélange complexe de protéines en solution est digéré à l'aide d'une protéase, par exemple la Trypsine. C'est une enzyme digestive qui clive la liaison peptidique après un résidu Lysine ou un résidu Arginine sauf s'ils sont précédés d'un résidu Proline. La présence d'une charge positive, nécessaire à l'analyse des ions est garantie, d'une part par les acides aminés chargés composant donc le peptide (R, K, etc.), et d'autre part par les protocoles mis en place lors de l'utilisation de sources d'ionisations (comme l'Electro Spray Ionisation par exemple). L'avantage principal de cette approche réside dans le fait que la masse à observer par spectrométrie de masse est grandement réduite (inférieur à 3 ou 4 kDa).

II.3.2 Identification de séquence

- **Analyse nano-LC-MS/MS**

En analyse de masse par la stratégie Bottom-up, la séparation des peptides par chromatographie est réalisée par des colonnes qui permettent des débits nanométriques (*nl/min*) sous des pressions extrêmement importantes (1200 bars). L'objectif principal de ces valeurs est de réduire le temps de rétention (et donc le temps nécessaire pour une analyse), sans perdre en résolution (la façon dont les particules sont séparées). Le protocole le plus utilisé en analyse LC-MS/MS Bottom-up classique est un gradient d'acétonitrile en condition acide, permettant l'élution des peptides

hydrophiles en premier, puis l'élution progressive des peptides hydrophobes, au fur et à mesure du temps de rétention. Ainsi, les choix de la colonne et de la forme du gradient sont des étapes critiques auxquelles il faut réfléchir avec soin, pour permettre des séparations peptidiques de qualité optimale.

Le chromatogramme dure en générale 120 minutes, pour une colonne de 75 cm de silice immobilisée en C18. Cela permet un temps d'élution d'une demi-minute pour un peptide standard. Cependant, il est à noter que certains peptides seront élués plus rapidement ou plus lentement, en fonction de leur abondance et leurs propriétés physicochimiques. En sortie de colonne, les gouttelettes du gradient sont soumises à une tension électrique élevée, dans la source Electrospray, induisant la formation de microgouttelettes (formant un cône de Taylor) permettant la désolvatation et leur transfert en phase gaz (principes de nébulisation). Ce processus permet aux peptides d'acquérir des protons supplémentaires à leurs surfaces et ainsi de recevoir une charge électrique. En Bottom-up, les ions acquièrent ainsi entre deux et huit charges par peptides. Les particules monochargées sont souvent des contaminants et peuvent être observées dans les spectres de masse comme des *patterns*, produit généralement par l'analyse de polymères.

- **Spectrométrie de masse en tandem MS/MS**

Les rapports masse sur charge des peptides observés dans le système sont mesurés une première fois (MS1). La fragmentation des peptides/ions et la deuxième mesure du rapport m/z de ceux-ci (MS2) permet d'obtenir des spectres de masse additionnels et caractéristiques. En effet, Ils présentent un ensemble composé des masses des différents fragments, ce qui permet une identification du peptide précurseur de manière non-ambigüe. Lors de la MS1, l'algorithme constructeur permet la sélection des ions les plus intenses, ils seront les précurseurs sélectionnés pour la fragmentation et la mesure MS2. Les avancées des constructeurs de spectromètres de masse consistent en l'introduction de plusieurs modes de fragmentation dans les systèmes. Par exemple, les ions candidats pour la fragmentation « ions précurseurs » qui étaient maintenus dans le vide jusqu'alors, vont rencontrer de l'azote inerte, dans le cas d'une fragmentation de type « HCD » (Higher Energy Collision Dissociation). Dans le cas du CID (Collision-Induced Dissociation), il s'agira d'hélium. D'autres méthodes de fragmentation impliquent la combinaison des deux méthodes (EtHcD) et l'utilisation de laser dans le cas de l'UVPD. Une fragmentation génère une « carte de

fragmentation » qui a la caractéristique d'être prédictible (Biemann, 1990). Ainsi, on sait que la rupture de la liaison peptides par HCD ou CID favorisera la formation d'ions *b* et *y* (Figure 17). La production théorique de ces cartes de fragmentation et leur confrontation avec les données réelles permettent l'identification des particules (peptides) de manière très précise (Steen & Mann, 2004).

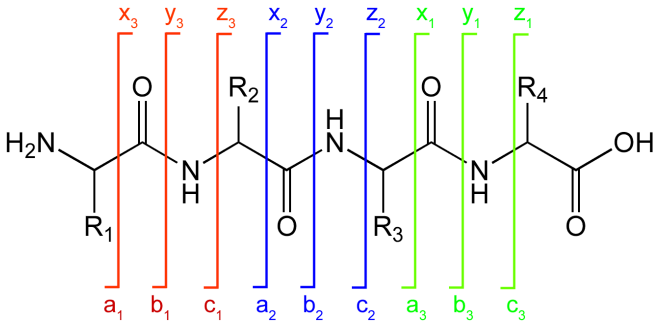


Figure 17 : Carte de fragmentation d'un polypeptide. Cette illustration est issue de [1]

A cet effet, l'interrogation des spectres débute par la digestion in-silico d'une banque de données de séquences protéiques (protéome) afin d'obtenir un peptidome. Pour chaque spectre MS2, le peptidome théorique est tout d'abord filtré par rapport à la masse du précurseur (du peptides), puis les cartes de fragmentations théoriques sont confrontées aux spectres de fragmentations expérimentales. Un séquençage a donc lieu afin d'annoter chacun des « pics » du spectre. Au final, plus l'annotation d'un spectre est complète, plus le score de confiance est élevé pour le peptide candidat.

- **Score d'identification**

Les identifications des protéines/peptides par séquençage des spectres de fragmentation est rendue possible par l'utilisation d'un score de confiance. Ainsi, dans le cas où plusieurs séquences peptidiques candidates pourraient expliquer la masse du précurseur observé dans un spectre expérimental, un score de fragmentation permet de les discriminer. Pour cela, l'identifiant ayant le score représentant la plus grande probabilité de présence est mis en valeur (Hogan et al., 2005). Pour illustrer cette notion de score d'identification, nous allons décrire l'équation définissant le score d'identification d'un logiciel d'interrogation en banque de données nommé RAId

[1] https://commons.wikimedia.org/wiki/File:Peptide_fragmentation.gif

(Robust Acurate Identification). Ce logiciel a été développé par l'équipe du QMBP (Alves & Yu, 2005). Il a l'avantage de disposer de plusieurs scores paramétrables à chaque interrogation. Il s'agit du XCorr (utilisé dans le logiciel Sequest (T. Liu et al., 2007)) ainsi que l'Hyperscore et le Kscore tous deux utilisés dans le logiciel X!tandem (Duncan et al., 2005)). Au cours de ma thèse, j'ai porté une attention particulière au score « RAId _DbS », développé par les auteurs (Alves et al., 2010) :

$$\text{RAId } S(\pi) = \frac{1}{T(\pi)} \sum_{i=1}^{T(\pi)} \ln(I_i)\, e^{-\Delta m_i}\, \theta(1-\Delta m_i)$$

Équation 1-3 : Score d'identification utilisé par RAId.

Avec :

- $T(\pi)$ qui correspond au nombre de fragments théoriques du peptide étudié (fragmentation *in-silico*).
- I_i qui correspond à l'intensité expérimentale de l'ion (issue du fragment i).
- Δm_i qui correspond à la différence entre la masse théorique (fragment *in-silico*) et la masse expérimentale de l'ion étudié.
- $\emptyset$ () qui est une fonction de Heaviside permettant de séparer les valeurs en 2 groupes. En effet, la valeur ressortie est nulle, si la valeur d'entrée est inférieure à 0, autrement la valeur est 1.

Enfin le score d'identification obtenu est pondéré par rapport à la taille de la séquence du peptide, dans le but d'effectuer une réaffectation statistique du score de confiance. L'un des grands avantages apportés par cet algorithme est la prise en compte d'un calcul des identifications réelles (vrai positif) et des identifications de peptides incorrectes (faux positif). Ainsi, la notion de *FDR* (False Discovery Rate) correspond au nombre d'identifications de peptides qui pourraient être obtenues de manière aléatoire et possédant la même masse. Le score *FDR* est mis en perspective avec des corrections statistiques, afin de produire la valeur E (E val) plus représentative que la traditionnelle p-value.

II.3.3 Identification systématique des PTMS

La recherche en banque de données est très couteuse en temps de calcul et les choix des paramètres lors des interrogations influencent grandement les

identifications réalisées. Ainsi l'ajout lors de la recherche de possibles modifications post-traductionnelles à des positions variables augmente grandement le nombre d'explorations de la base de données et pose un problème de combinatoire non trivial. Là où une recherche classique teste la possibilité de présence de deux ou trois modifications post-traductionnelles par recherche seulement (en général il s'agit de la Méthylation de l'Acétylation et de Phosphorylation), de nombreux peptides modifiés ne sont pas observés (puisque non recherchés correctement).

Au laboratoire, un travail original a été réalisé par Thomas Denecker, avec pour objectif d'améliorer l'identification de peptides singuliers, en partant de listes plus étendues de modifications pots traductionnelles possibles et les moyens de calcul de la grille de calcul de l'Institut Français de Bioinformatique (publication en cours de rédaction).

II.3.4 Méthodes de quantification

Les méthodes de quantification en protéomique fondées sur de la spectrométrie de masse reposent sur le traitement des échantillons (*in-vitro*) ou sur l'incorporation des marqueurs dans le métabolisme des organismes étudiées (*in-vivo*) (Bantscheff et al., 2012). L'approche ***in vitro*** repose sur la quantification du signal brut ou d'un marqueur ajouté en post-lyse. Une possible limitation est l'erreur de mesure des spectromètres induite, soit une problématique de reproductibilité des spectres. Dans l'approche ***in vivo,*** les cellules sont mises en contact avec différent marqueurs (Gouw et al., 2010). Ceux-ci peuvent produire de possibles effets sur leurs croissances, ou limitations par leurs natures sur les souches et/ou métabolisme compatible. Nous allons étudier les principales méthodes de quantifications existantes (Figure 18).

- **TMT / iTRAQ**

La stratégie utilisée par ce type d'approche se fonde sur un marquage chimique ***in vitro*** permettant un multiplexage de nombreux échantillon durant une seule analyse par spectrométrie de masse. Chaque « tag » est composé d'un groupement « reporter », d'une « balance » et d'un groupe réactif (groupement électrophile), à-même de réagir avec le groupement amine en partie N-terminal des peptides. Les signaux de la masse correspondant au « reporter » est utilisé pour la quantification des signaux d'intensité entre les différents pools de peptides. Le kit de quantification TMT de Thermo Fisher Scientifique n'a eu de cesse d'augmenter la capacité d'analyse

simultanée passant au départ de 6-plex à 12-plex et récemment 18-plex. Ce qui permet donc en théorie d'analyser en simultané 18 conditions. Le marché est partagé entre l'*isobaric Tag for Relative and Absolute Quantitation* (iTRAQ) de l'entreprise Sciex (Ross et al., 2004)(Choe et al., 2007) et le *Tandem Mass Tag* (TMT) de l'entreprise Thermo Fisher Scientific (McAlister et al., 2012) (Dayon et al., 2008) (Thompson et al., 2003).

- **SILAC *in vivo***

La méthode de quantification, SILAC (Stable Isotope Labeling by Amino acids in Cell culture) (Ong et al., 2002) (X. Chen et al., 2015) repose sur une incorporation de marqueurs isotopiques ***in vivo***. Pour cela deux cultures sont effectuées, l'une dans du milieu normal, l'autre dans du milieux « lourd » (comportant les acides aminés marqués). Le marquage des protéines est réalisé par l'incorporation des différents acides aminés lors de la biogenèse des protéines. Il est à rappeler que seuls quelques acides aminés seront présents et donc marqués. La quantification s'effectue après l'analyse par spectrométrie de masse. Dans un spectre deux clusters isotopiques sont observés chacun représentant les peptides d'une condition. Le ratio de l'intensité des clusters permet d'obtenir une valeur quantitative. Cette technique présente de grande limitation comme l'utilisation de souches biologique spécifiques (auxotrophes) et du fait que les acides aminés ne sont pas parfaitement incorporés. De plus au niveau du traitement du signal la mesure entre deux clusters isotopiques différents est donc soumise à la variation de mesure du spectromètre de masse.

- **Isotopes lourds *in vivo***

D'autres méthodes utilisent les isotopes lourds comme précurseur pour un marquage ***in vivo***. Il s'agit en particulier de l'oxygène ^{18}O (Ye et al., 2009)(Stewart et al., 2001)(Miyagi & Rao, 2007) et plus récemment le deutérium (hydrogène ^{2}H) dans les travaux de Rovshan Sadygov (Sadygov, 2018) et (Borzou et al., 2019). Cependant l'usage d'isotopes lourds implique la mesure des signaux bruts de deux clusters isotopiques (l'un marqué, l'autre naturel). Une double mesure implique donc potentiellement une double erreur d'extraction des signaux et une perte d'information (présence intense des deux clusters). En particulier les méthodes de calcul ont été présentées dans cet article (Sadygov, 2021). De plus comme nous le verrons par la suite ces méthodes ne sont pas applicables pour l'études des protéines intactes (Top-Down).

- **Label Free Quantification *in vitro* (LQT)**

La méthode de quantification de type Label-free est la plus utilisée car elle n'influe pas sur la biologie des systèmes et s'implémente facilement (Neilson et al., 2011). La procédure de quantification repose sur une comparaison brute entre les différentes conditions/injections. Un travail conséquent au niveau du retraitement des données est effectué afin d'aligner les runs selon le temps de rétention. En particulier le logiciel Progenesis procède de manière automatique à l'alignement des chromatogrammes toutefois l'utilisateur peut améliorer les traitements en corrigeant manuellement par un ajout de point de référence (Léger et al., 2016). Les autres logiciels (FFId, Maxquant, Peaks etc …) fonctionnent de manière totalement automatique en utilisant uniquement le XIC (courant d'ion total) comme valeur discriminatoire d'alignement. Une fois les runs alignés, une comparaison de l'intensité des clusters isotopiques est effectuée. La valeur d'intensité correspond au volume en trois dimensions du profil d'élution du cluster isotopique selon le temps de rétention et la masse. Deux principaux indicateurs sont utilisés : le dénombrement spectral (nombre de MS2) ou par extraction des intensités expérimentales. Enfin les ratios de quantification sont produits à partir d'un run sélectionné comme référence. Cette méthode de quantification présente un besoin important de reproductibilité du chromatogramme lors des séparations et lors de l'analyse par spectrométrie de masse. Les avantages reposent sur le fait qu'il n'y a pas de limite absolue dans le nombre d'échantillons à analyser tant que la reproductibilité est maintenue ce qui est de plus en plus le cas (en particulier avec l'introduction de l'Evosep (Melby et al., 2021)). Cependant, cette méthode se limite par deux points principaux. D'une part l'alignement n'est pas toujours idéal et les « features » quantifiés s'en retrouvent donc erronés ou très limités (réduction drastique des données exploitables). D'autre part, une perte d'information a lieu lors de l'étape d'association entre des objets quantifiés et les identifications (Jin et al., 2021).

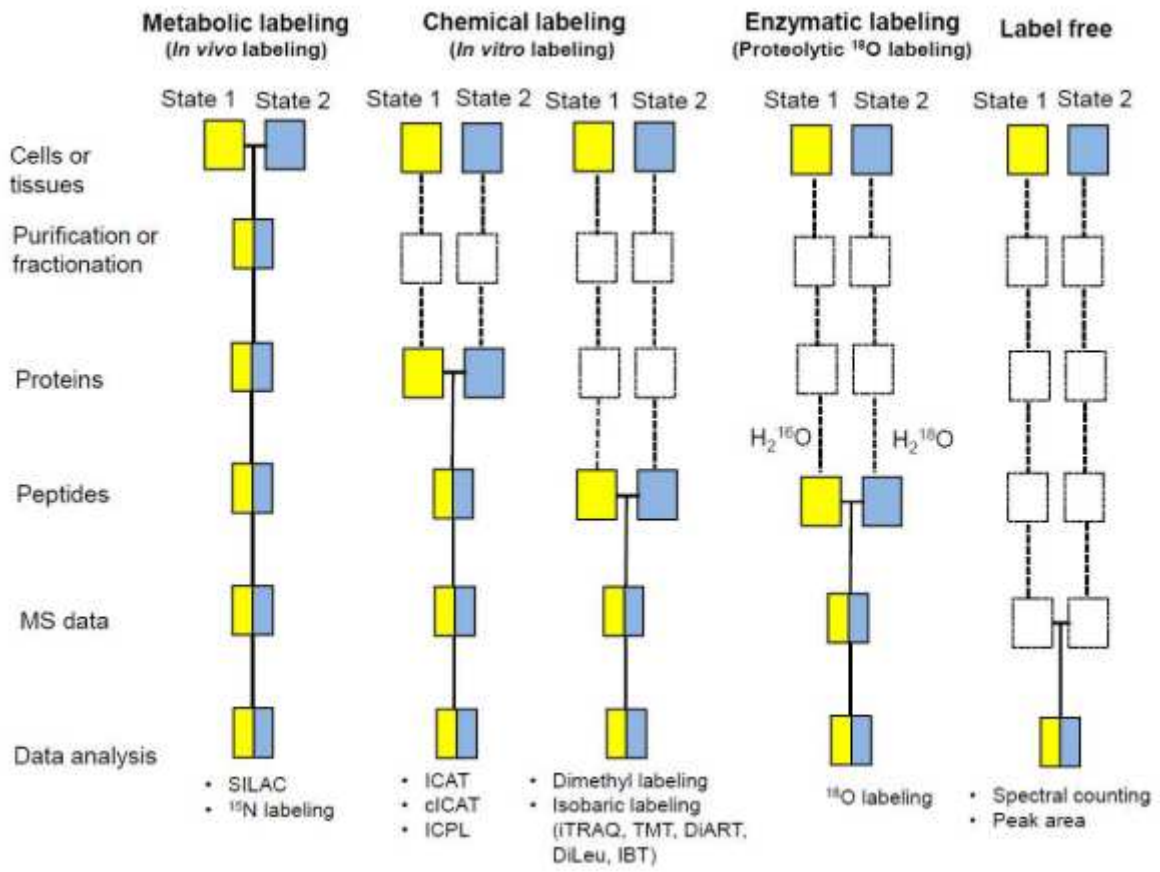


Figure 18 : Principales méthodes de quantifications existantes. Cette figure est extraite de (X. Chen et al., 2021)

Les appareils actuels sont hautement sensibles et possèdent des hautes résolutions permettant la quantification des peptides en utilisant les mesures d'intensité des ions parents (MS1). Cependant, lors de la réalisation d'une expérience de protéomique quantitative, seulement une fraction des protéines qui composent le protéome sont identifiées, et seulement une partie d'entre elles (quelques milliers) sont *in fine* quantifiées (Figure 19). Ainsi, chez un organismes tel que la levure (*C. glabrata* ou *C. albicans*), c'est seulement 30% des protéines composant le protéome qui sont quantifiées de façon reproductible (Lelandais et al., 2019). C'est la raison pour laquelle, lors d'une analyse du protéome par spectrométrie de masse, il est indispensable de faire la distinction entre l'identification d'une part, et la quantification d'autre part. Des ions analysés lors de la MS1 avec une bonne qualité de signal, ne sont pas forcément analysables avec une aussi bonne qualité, après la **fragmentation** en MS2.

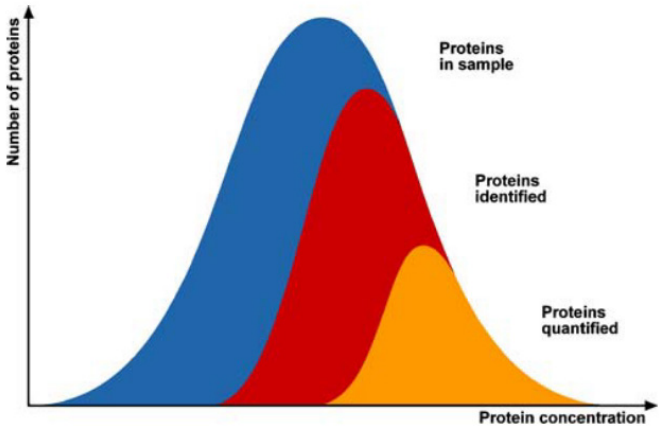


Figure 19 : Représentation schématique de la fraction d'un protéome identifiée et/ou quantifiée par spectrométrie de masse. Cette figure est extraite de (Bantscheff et al., 2007).

La valeur du taux d'abondance des protéines (quantification relative) est généralement exprimée en « fold-change ». Classiquement, une valeur supérieure à $Log_2(FC) = 2$, est considérée comme révélatrice d'un différentiel d'abondances entre les conditions relatives comparées (Old et al., 2005). A cette valeur, il y a quatre fois plus de protéines exprimées dans une condition par rapport à l'autre. Des évaluations statistiques de la "significativité" des mesures sont obtenues par la mise en application des tests statistiques de type *ANOVA* (les mesures des variables sont considérées indépendantes). L'hypothèse nulle (H0) teste un modèle dans lequel les variations observées entre les conditions étudiées seraient uniquement dues au hasard (fluctuations d'échantillonnage). La probabilité des observations, compte tenu de ce modèle H0, est représentée par la P-value. Ainsi, une représentation du type "Volcano plot" est souvent utilisée pour visualiser les résultats d'une analyse de données (Figure 20).

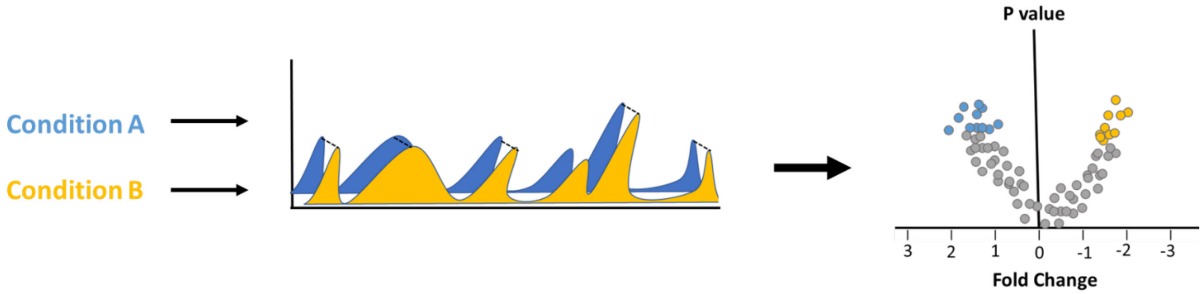


Figure 20 : Représentation schématique de quantification par la méthode label-free. Les abondances des protéines calculées dans les conditions A et B sont comparées (abondances relatives)

L'approche Bottom-up, bien que plus simple à réaliser d'un point de vue technique, n'est pas optimale. Elle induit des compromis qui limitent (parfois drastiquement) les conclusions biologiques des analyses. Ainsi, il est utile de garder à l'esprit que le protéome « réel » peut être bien différent du protéome « observable », composée uniquement des peptides les plus abondants. Une bonne couverture de séquences des protéines, en termes de peptides identifiés à des positions non-ambiguës des séquences protéiques, est nécessaire pour obtenir des identifications fiables. Cela est encore plus complexe dans le cas de recherches de modifications post-traductionnelles (voir II.3.3). Enfin, les traitements biochimiques pour la préparation des échantillons (digestion et dénaturation) peuvent induire des biais biologiques irréversibles sur les composés (adduit sodium etc..). Cela se révèle par de grandes variations, entre échantillons et entre réplicas biologiques.

II.3.5 Mise en place de la stratégie d'analyse « Top-down »

Si les approches Bottom-up sont maintenant performantes et bien maitrisées par la communauté scientifique pour étudier des problématiques biologiques de plus en plus complètes (Lermyte et al., 2019) les avancées technologiques, à la fois en termes d'appareils (spectromètres de masse) et d'outils bio-informatiques pour l'analyse des données, permettent des développements originaux de type « Top-down ».

L'étude des modifications structurales des protéines, associée à une mesure des abondances relatives des protéoformes, permet d'aborder des questions biologiques diverses. En particulier, une équipe interdisciplinaire de spécialistes en protéomique Top-down (bio-informaticiens, biologistes, physiciens, et instrumentistes) s'est constituée pour codévelopper et joindre les avancées dans le domaine : le Consortium for TopDown Proteomics (CTDP). Les réalisations de ce groupe sont illustrées par le lancement d'un projet de séquençage des protéoformes. Les protocoles développés spécifiquement pour l'analyse des protéines intactes sont également régulièrement publiés (Donnelly et al., 2019). Un exemple concret est le projet qui a pour objectif de

décrire la fonction des protéoformes composant le protéome humain « The Human Proteoform projet » (L. M. Smith et al., 2021). En effet, les cellules humaines possèdent environ 20000 gènes, et chaque protéine exprimée peut présenter une diversité de modifications post-traductionnelles (au moins 400 sont décrites). Ceci génère un nombre théorique de protéoformes possibles colossal, supérieur à 1 million. Plusieurs protéines caractéristiques de maladies seraient expliquées par une forme structurale particulière de celles-ci (Figure 21). La détermination de ces protéoformes permet d'effectuer un suivi de l'évolution des maladies et le développement de nouvelles cibles thérapeutiques.

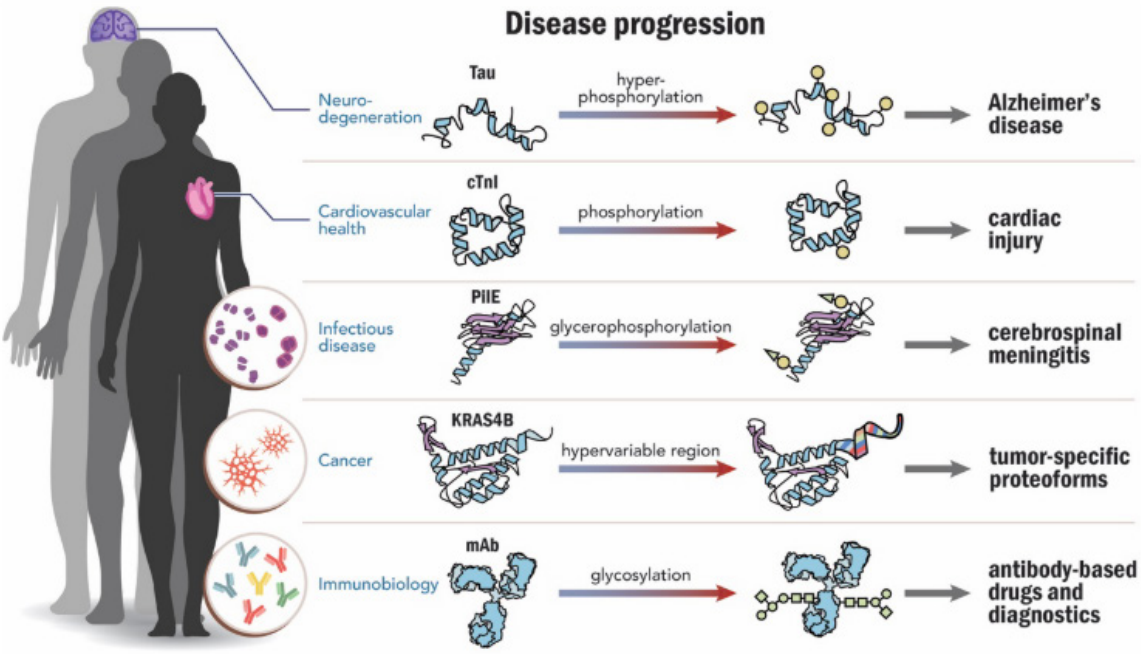


Figure 21 : Protéoformes impliquées dans des maladies humaines. Cette figure est extraite de (L. Smith et al., 2020)

II.4 Avancées technologiques récentes

II.4.1 Technologie des analyseurs

Les spectromètres de masse actuels sont fabriqués à partir de plusieurs nouvelles technologies. Celles-ci améliorent la qualité des signaux expérimentaux mesurés (ou acquis), permettant ainsi d'étendre les applications possibles de la spectrométrie de masse en biologie. Bien que les données produites par les spectromètres de masse soient toujours de même nature, à savoir des mesures de masse et d'intensité des ions, les différents appareils ont chacun des spécificités, en ce

qui concerne les méthodes d'analyses des ions mais également les technologies mises en jeux.

Pareillement, les progrès récents des systèmes de chromatographie en phase liquide (LC) permettent de séparer des échantillons biologiques de plus en plus complexes, tout en produisant des profils d'élution de grande résolution et de manière reproductible. L'intérêt de ces nouveaux modes de séparation des échantillons biologiques est révélé quand ils sont couplés à une analyse par spectrométrie de masse.

Dans ce contexte, les améliorations en spectrométrie de masse peuvent s'initier depuis différents aspects. Les paramètres les plus importants sont :

La vitesse. En effet, la fréquence d'analyse et de séparation des ions qui entrent dans le spectromètre est déterminante. Toute lenteur empêche d'une part l'analyse de multiples espèces d'ions qui seraient éluées conjointement, et d'autre part limite l'observation approfondie des différents ions présents à un instant donné dans l'appareil.

La sensibilité. Ce critère permet d'analyser des ions très peu intenses, c'est-à-dire pour lesquels le rapport signal sur bruit serait très faible.

La résolution. Elle permet de mesurer très précisément les masses, facilitant ainsi l'étape d'identification ultérieure. Par ailleurs la résolution des spectromètres est définie par un rapport entre la masse mesurée et la différence minimum pour que deux ions soient discriminés.

La gamme de masse. Les objets biologiques étudiés ont des masses différentes. Leur analyse simultanée nécessite des gammes de masse aussi larges que possible. C'est d'autant plus critique vis-à-vis des applications récentes en analyse des protéines intactes. Celles-ci nécessitent des outils capables de mesurer des ions de très grandes masses variables.

La discrimination des ions. En complément du critère de sensibilité (voir ci-dessus), ce critère permet, dans un échantillon complexe, d'améliorer la discrimination des ions et de faciliter leur identification et leur quantification, même s'ils sont en très faible quantité dans le groupe complexe étudié.

Le tableau ci-dessous (Tableau 5) présente plusieurs modèles types qui sont vendus actuellement. Ils représentent l'éventail des technologies disponibles sur le marché. Les technologies de séparations associées reposent sur la mesure de la mobilité ionique des ions analysée (voir II.4.2).

Instruments	Constructeur	Technologie de Mesure Masse	Technologie de séparations ioniques associée	Format de Fichier produit
Lumos	ThermoFisher Scientific	Quadrupôle et Orbitrap	FAIMs	.raw
TimsTOFPro 2	Bruker	TOF	Tims	.d
SELECT SERIES Cyclic IMS	Waters	Quadrupôle et TOF	Cyclique	.pkl
FT-ICR	Bruker	Ion Cyclonic Resonance	-	-

Tableau 5 : Technologies des différents appareils vendus actuellement

Les progrès récents dans le développement des technologies de spectrométrie de masse ont permis d'obtenir des spectres de meilleure qualité avec une haute résolution isotopique. Les traitements et les interprétations des données brutes produites, sont réalisés par des algorithmes développés par les constructeurs des appareils. Leur qualité est déterminante afin d'exploiter pleinement la qualité des mesures produites. Les analyseurs sont les outils centraux pour l'observation des ions, il est donc important d'étudier avec précision leurs technologies pour comprendre leurs spécificités et caractéristiques.

- **Temps de vol (TOF)**

L'analyseur TOF (*Time Of Flight*, Temps de vol) utilise une méthode d'analyse qui repose sur le principe des lois de la conservation et de la transformation de l'énergie potentielle en énergie cinétique. Concrètement, la mesure du temps nécessaire à des particules chargées pour parcourir une distance connue permet de calculer une vitesse qui est fonction du rapport masse sur charge.

Les ions sont définis dans un système soumis aux lois non relativistes, ce qui permet d'écrire :

- L'énergie cinétique $E_c = \frac{1}{2}mv^2$, où m est la masse de la particule et v sa vitesse,
- L'énergie potentielle $E_p = zU$ où z est la charge électrique de la particule et U la différence de potentiel (tension).

Si l'on pose l'égalité dans le cas d'un transfert d'énergie totale, la relation suivante est obtenue :

$$E_c = E_p \Leftrightarrow \mathrm{zU} = \frac{1}{2}\mathrm{m}v^2$$

Comme la vitesse v est égale à la distance parcourue divisée par le temps : $\mathrm{v} = \frac{\mathrm{d}}{t}$

Cela permet d'obtenir :

$$t^2 = \frac{d^2}{2U}\frac{m}{z}$$

Dans un spectromètre de masse, le rapport $\frac{\mathrm{d}^2}{2\mathrm{U}}$ est une constante. Ainsi nous avons une relation proportionnelle entre le temps de vol au carré et le rapport masse sur charge.

Soit : $\frac{m}{z} = t^2/cte$

Ce type d'analyseur est une technologie ancienne (le premier brevet de William E. Stephens date de 1952), mais se perfectionne continuellement et gagne en résolution. Par exemple, en 1973 le *reflectron* a été inventé (Mamyrin et al., 1973). Il s'agit d'une technologie qui permet de renvoyer les ions et de réajuster la vitesse de particules de même rapport masse sur charge, mais de mesures d'énergie cinétique

différentes. Cette technologie est toutefois extrêmement sensible aux éventuels écarts de température qui influent sur les mesures de temps de vol. Cela peut être résolu par l'ajout d'une cellule thermostatée autour du tube de vol (comme le cas du TimsTOF Pro 2 possédant un *chiller*, fixant la température à 25°C précisément). L'analyse par temps de vol permet d'obtenir une séparation des ions qui est en relation directe (linaire) avec leur masse, sans limite absolue. Cela en fait un outil idéal pour effectuer des expériences dans lesquelles les masses mesurées varient beaucoup, comme l'analyse des protéines intactes.

- **Quadripôles**

Le quadripôle est la plus ancienne technologie. Elle est rapide et sensible, permettant une sélection performante des ions pour l'analyse. Un quadripôle est composé de quatre électrodes linéaires soumises à une tension alternative. Le mécanisme a été mis au point par Paul Wolfgang en 1955, à partir d'une adaptation de l'étude des résonances électromagnétiques et des trajectoires des ions calculées selon les équations de Mathieu.

Dans les spectromètres actuels, cet analyseur est placé au début de la trajectoire optique et permet de « trier » les ions. En effet, l'analyseur parcourt les fréquences et permet de guider uniquement les ions possédant une certaine résonance. Les autres ions, dits « non-résonants », ont alors une trajectoire chaotique et sortent du quadripôle. Ainsi, ils ne poursuivent pas leur parcours vers la suite de l'optique instrumentale.

- **Cellules Orbitrap**

En 1923 K. Kingdon (Kingdon, 1923) met au point la "trappe ionique" puis R. D. Knight (Knight, 1981) améliore son fonctionnement. Alexander Makarov, physicien en 1990, s'intéresse à la problématique de miniaturisation de cet outil et développe au sein de l'entreprise Thermo Fisher l'analyseur Orbitrap (A. Makarov, 2000; A. Makarov et al., 2005; A. A. Makarov, 1999) (Hu et al., 2005)(Van De Waterbeemd et al., 2018). L'injection des ions de manière tangentielle à l'Orbitrap est une étape critique pour cet analyseur. Les premières machines utilisant cette technologie Orbitrap ont été commercialisées en 2005. C'est une technologie de pointe qui permet la miniaturisation de principes de physique électronique. Les principes sont les suivants. Les ions gravitent en spirale autour d'un noyau central, soumis à un dipôle électrique. Les particules émettent un champ électromagnétique qui va être détecté par les

bobines/détecteurs. Le signal produit est de la forme d'ondes de radiofréquences sinusoïdales en fonction du temps, appelées "transient" (Figure 22) (A. Makarov et al., 2019)

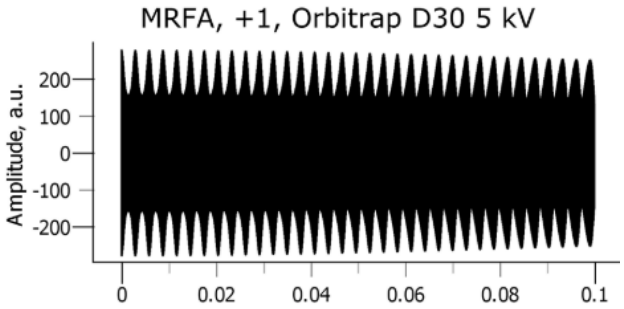


Figure 22 : Simulation du temps de transient (Molécule de MRFA, monochargée, sur une cellule Orbitrap de type D30 et de tension 5kV). Cette figure est extraite de (Nagornov et al., 2020)

Une transformée de Fourier est appliquée afin d'extraire toutes les fréquences, à l'aide d'équations simples reliant la fréquence au ratio charge/masse. Ainsi, après la mise en application de la transformée de Fourier sur un spectre de fréquence, nous avons :

$$\omega = \sqrt{VcK0q/m}$$

Avec :

- ω = fréquence d'oscillation axiale,
- Vc = Voltage (de la cellule intérieure "noyau", courant négatif),
- Ko = constante définie selon concepteur,
- q/m = le rapport charge sur masse.

Plus l'outil fonctionne avec des gammes de fréquence élevées, plus la résolution est importante. Cet analyseur possède une très forte précision et permet d'obtenir des spectres de très haute résolution. Malheureusement, cela est souvent réalisé au détriment du temps d'acquisition, réduisant alors la fréquence d'analyse. De plus, l'Orbitrap ne possède pas une résolution linaire par rapport à la masse, ce qui en fait un outil précis pour certaines gammes de masse restreintes uniquement.

II.4.2 Technologie de mobilité ionique : *Tims* (Trapped ion mobility spectrometry)

Cette technologie permet une observation de la conformation des ions analysés, en particulier l'encombrement stérique des molécules. C'est une manière différente de séparer les ions. Une nouvelle dimension s'ajoute à celle du rapport masse sur charge. Des applications originales sont permises par cette technologie, jusqu'alors impossibles en spectrométrie de masse, telle que la discrimination des molécules isomères. Enfin, cela permet de fournir, en plus des données spectrales, des informations sur la mobilité des ions, afin d'identifier au mieux les peptides. La mobilité ionique apporte de nouvelles potentialités de séparation des ions et de leur étude, mais également permet de produire des spectres de masse de meilleure qualité.

Chapitre 2 : Méthode SLIM

I. Principe de fonctionnement

I.1 Le constat

En 2000, plusieurs articles scientifiques de l'équipe d'Allan Marshall (Marshall et al., 1997; Rodgers et al., 2000; Shi et al., 1998) démontrent que la réduction de la complexité isotopique d'un échantillon biologique, notamment par un apport augmenté de ^{12}C et ^{14}N dans les protéines, facilite la détermination de la masse monoisotopique des objets étudiés. Une dizaine d'années plus tard, l'équipe de Neil Kelleher explique de manière théorique la plus-value d'un tel protocole de préparation des échantillons *in vivo*. Il abaisse de façon importante le niveau du bruit par rapport au niveau du signal, ce qui en fait une stratégie particulièrement prometteuse en analyse « Top down » des protéines intactes (Compton et al., 2011).

I.2 Le SLIM

À partir de ces constatations, notre laboratoire a développé une méthode d'analyse nommée SLIM, Simple Light Isotopes Metabolic Labeling (Léger et al., 2017). C'est une méthode de quantification des protéines, par leur "marquage" métabolique *in vivo*. Le terme "marquage" signifie ici qu'une réduction de la complexité isotopique des protéines est réalisée, permettant (comme vu précédemment) une modulation de l'intensité de l'ion monoisotope. Cela a pour effet des conséquences bénéfiques pour les analyses des molécules biologiques (peptides ou protéines) par spectrométrie de masse. La mesure de la masse monoisotopique est plus précise, ce qui facilite l'étape d'identification des peptides/protéines. Nous avons pu développer une méthode de quantification originale.

I.2.1 Incorporation du ^{12}C dans les protéines, *via* le métabolisme des organismes

La méthode SLIM tire profit de la capacité des organismes à synthétiser des acides aminés marqués U-[^{12}C] *in vivo*, à partir d'une source de carbone telle que le glucose, exclusivement U-[^{12}C] (Figure 23).

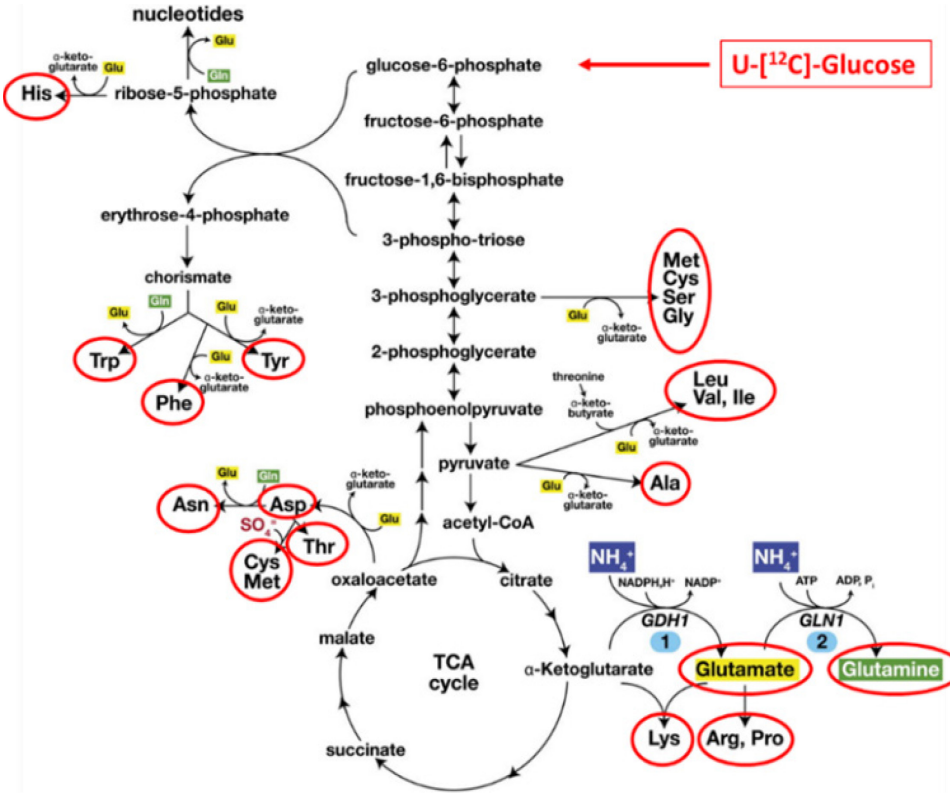


Figure 23 : Illustration de l'incorporation de ^{12}C dans les acides aminés par le métabolisme de la levure. Cette figure est extraite de (Sénécaut et al., 2022)

Pour la méthode SLIM, les organismes sont cultivés avec du glucose pour lequel tous les atomes de carbone sont de type ^{12}C. Dans cette situation, la synthèse des acides aminés est réalisée avec uniquement des atomes ^{12}C, et les protéines, *in fine*, sont également intégralement "marquées". Ce changement (artificiel) de l'abondance isotopique du carbone, a pour conséquence une augmentation de l'intensité de l'ion monoisotopique, permettant une modification du cluster isotopique (Figure 24).

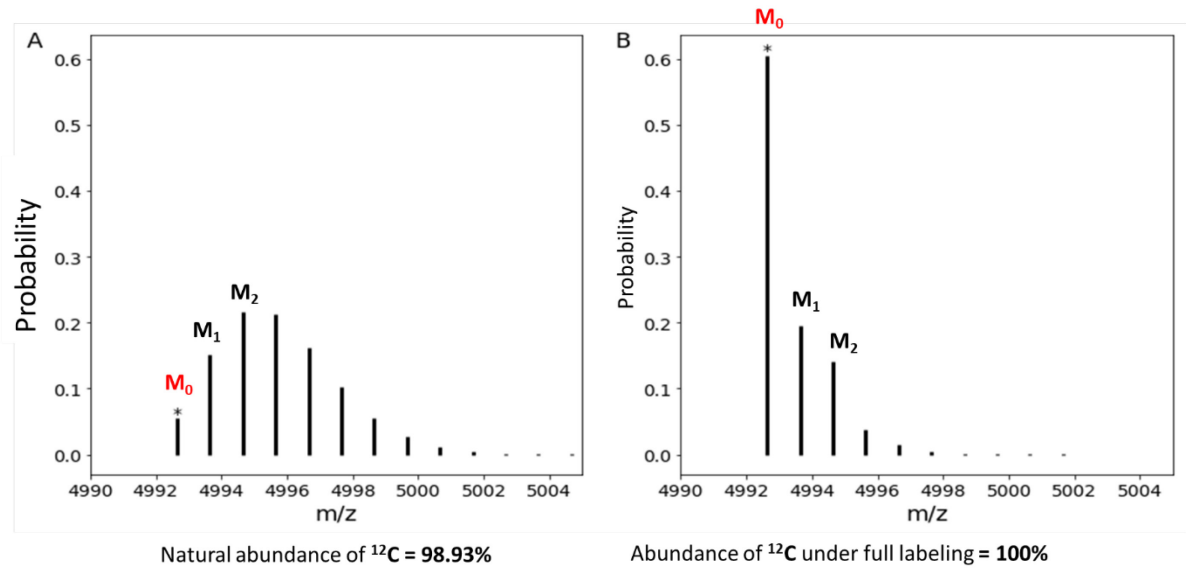


Figure 24 : Deux clusters isotopiques théoriques, le cluster naturel en A, la condition 12C SLIM Labeling en B.

I.2.2 Premier effet : une amélioration de la mesure des masses

A nouveau, l'augmentation de l'intensité de l'ion monoisotopique a pour effet de faciliter la mesure de la masse monoisotopique. C'est une valeur expérimentale pour laquelle la précision de la mesure est déterminante pour l'identification des ions, lors de la recherche en banque de données (augmentation de l'intensité du signal par rapport au bruit). Cela conduit à un score d'identification et une couverture de séquence protéique plus élevée dans les expériences de spectrométrie de masse.

I.2.3 Deuxième effet : réalisation de mesures quantitatives

Les spectres « marqués » ainsi nouvellement formés ont un profil typique et une signature se démarquant des mêmes types de spectres analysés sur des protéines possédant une abondance isotopique naturelle. Cela nous a permis de développer une méthode de quantification innovante et originale en protéomique, indépendante de la masse des ions. L'effet quantitatif par marquage *in vivo* des protéines est le point central développé dans cette thèse.

Cette stratégie développée au laboratoire fonctionne de la façon suivante. Les cellules sont cultivées dans deux conditions différentes. Dans une première condition,

dite « naturelle » (notée NC pour “Natural Condition”), les cellules sont mises en culture avec une source de carbone possédant une abondance naturelle des différents éléments, comme du glucose commercial. Dans la deuxième condition, dite “marquée” (et notée 12C), la source de carbone est du glucose exclusivement U-[^{12}C]. Le mélange 1:1 des protéines provenant des deux conditions permet, après préparation et analyse par spectrométrie de masse, de déterminer la quantité relative de chaque peptide d’une condition par rapport à l’autre (Figure 25). Un des avantages de cette méthode, est que l’incorporation du carbone U-[^{12}C] se fait de manière constante, pour tous les acides aminés que l’organisme est capable de métaboliser. Cela permet d’obtenir des résultats indépendants des fréquences des acides aminés constituant les protéines, en particulier vis-à-vis de la Lysine ou l’Arginine (acides aminés couramment utilisés en marquage SILAC, et dont la présence est alors indispensable).

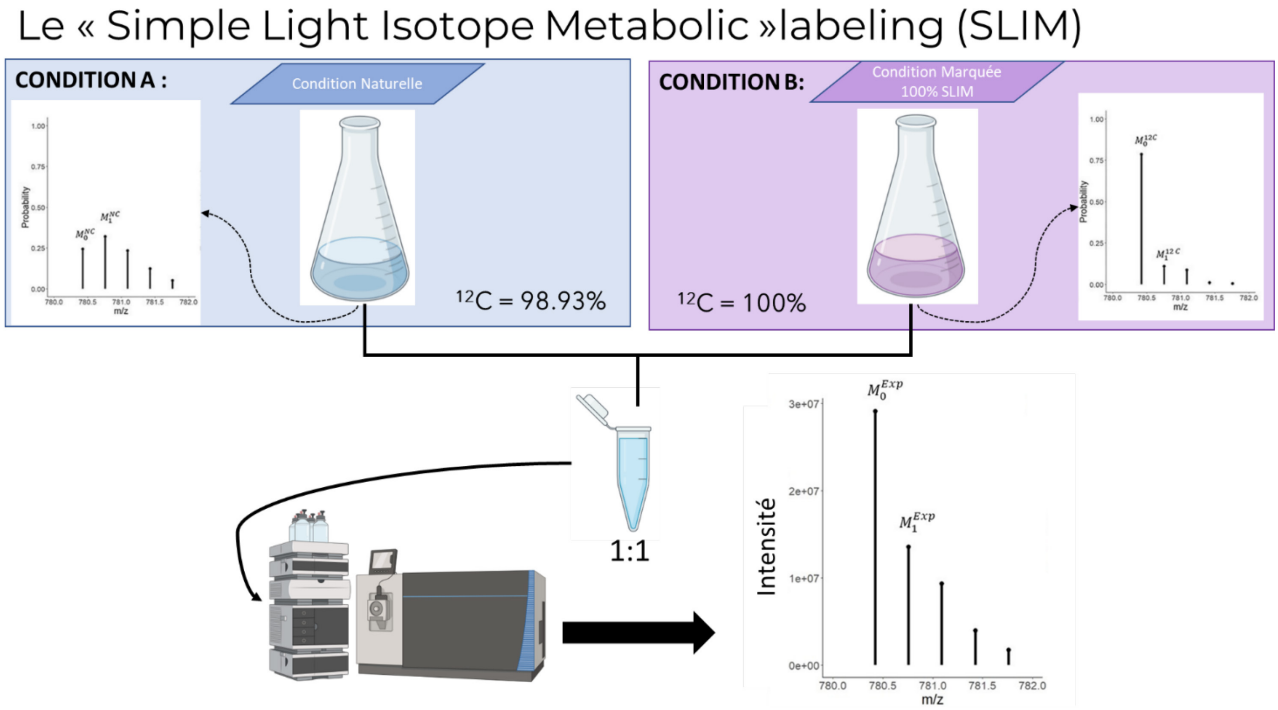


Figure 25 : Schéma de principe du SLIM labeling

Cette méthode SLIM est la première stratégie nouvelle de quantification exposée dans la littérature, depuis l'introduction du marquage métabolique (SILAC et dérivés) et du marquage isotopique des peptides/protéines (i-CAT, i-TRAQ, TMT), introduits dans les années 1999-2004. De plus, c'est la seule méthode réellement applicable à l'analyse quantitative des protéines intactes. En effet, seule l’intensité des

isotopologues varie d'une condition à une autre, les valeurs des masses demeurant strictement identiques. De plus, l'obtention des valeurs quantitatives se fait de manière moins bruitée car la mesure se fonde sur la mesure des intensités des isotopologues d'un seul et unique cluster.

II. Améliorations nécessaires

L'article originalement publié par l'équipe en 2017 (Léger et al., 2017), présente les développements de la méthode SLIM ainsi que son utilisation pour des applications biologiques. Outre la présentation du bénéfice de l'incorporation de ^{12}C dans les protéines pour l'identification et la quantification en spectrométrie de masse, une observation originale de mesure du temps de demi-vie des protéines sous l'effet de différents inhibiteurs de protéase a été également exposée. Par ailleurs, la publication de la méthode a été mise en avant en septembre 2017 dans la lettre informative de Mascot (Figure 26), un moteur de recherche en banque de données développé par Matrix Science.

Néanmoins, plusieurs limitations ont été soulignées. Par exemple, la méthode d'analyse et le retraitement des données brutes issues du spectromètre de masse ont été considérés comme trop complexes. De plus, la méthode d'extraction de données brutes était fondée sur un logiciel payant et propriétaire, « Progenesis QI for metabolomics ». Par ailleurs la méthode présentée, déjà complexe, ne présentait pas la possibilité de travailler avec des organismes possédant une ou plusieurs auxotrophies. L'étude de ces organismes nécessitent un apport d'acides aminés exogènes non-marqués dans le milieu, dont la présence a des conséquences sur le signal mesuré. Enfin, des statistiques complémentaires devaient être associées, afin de joindre un score de confiance à la quantification finale obtenue. En définitive et comme discuté dans la publication originale, seuls les bénéfices du SLIM avaient été démontrés en Bottom-up, alors que les réelles forces de cette technique trouvaient leurs applications pour l'études des protéines intactes (Top-down). Sur ce point, tout le travail restait à faire.

Featured publication using Mascot

Here we highlight a recent interesting and important publication that employs Mascot for protein identification, quantitation, or characterization. If you would like one of your papers highlighted here please **send us a PDF or a URL**.

A Simple Light Isotope Metabolic labeling (SLIM-labeling) strategy: a powerful tool to address the dynamics of proteome variations *in vivo*

Thibaut Leger, Camille Garcia, Laetitia Collomb and Jean-Michel Camadro

Molecular & Cellular Proteomics, in press, published August 18, 2017

The authors have developed a new approach to address some of the limitations imposed by current quantitative proteomics methods. With metabolic labelling or stable isotope reagents, there are challenges due to the variable mass shifts or the ion suppression from additional peptides present causing lower identification efficiencies.

The paper describes bottom-up proteomic analyses of the pathogenic yeast, *C. albicans*, grown on a synthetic medium containing either normal glucose or depleted glucose containing only ^{12}C carbon atoms as the sole carbon source. The use of ^{12}C resulted in an increase in the intensity of the monoisotopic ion, markedly improving bottom-up proteomics analyses.

^{12}C incorporation resulted in better overall scores per identified peptide with average scores up to 28% higher leading to an average increase of the protein identification scores of 36%. The number of identified peptides and proteins in *C. albicans* also increased by 14% and 11%, respectively, when applying a 1% FDR filter after ^{12}C enrichment *in vivo*.

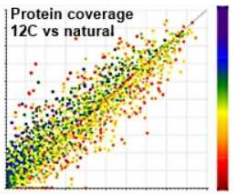


Figure 26 : newsletter matrix science (https://www.matrixscience.com/nl/201709/newsletter.html)

III. Objectifs de ma thèse

Ma thèse avait ainsi 3 grands axes de recherche fondés sur les développements complémentaires de la méthode SLIM. Le premier était le développement d'un pipeline d'analyse bio-informatique (workflow) robuste d'analyse des données bSLIM (Bottom-up SLIM). Ce workflow se devait d'être facilement utilisable, versatile et *OpenSource*, pour être distribué plus largement au sein de la communauté scientifique.

Le deuxième objectif était le développement de la méthode sur des organismes eucaryotes supérieurs. Une partie des acides aminés correspondant à l'auxotrophie, seront à ajouter dans le milieu provoquant une dilution isotopique. Ainsi l'implémentation de nouveaux calculs est nécessaire en réponse à un marquage partiel.

Enfin le dernier objectif est de développer la méthode top-SLIM (Top-down SLIM) pour l'analyse et développer la quantification des protéines intactes.

Partie 2: Développement de la méthode bSLIM

Avancées en protéomique quantitative par une stratégie « Bottom-up »

Si la méthode SLIM avait déjà été mise au point au laboratoire avant le début de ma thèse, des améliorations étaient nécessaires (voir ci-dessus). Tout au long de ma thèse, j'ai travaillé en utilisant une démarche scientifique que nous avons souhaitée robuste et rigoureuse. Ainsi, nos réflexions ont systématiquement débuté par un temps d'observation des données, de façon à formuler des hypothèses ou proposer des modélisations. Celles-ci, ont ensuite été éprouvées par la mise au point de simulations numériques. Enfin, des validations finales expérimentales ont été réalisées.

Chapitre 1 : Présentation de la méthode bSLIM

Comme expliqué précédemment avec la méthode SLIM, la quantification des peptides (puis des protéines) est rendue possible par la mise au point d'une solution algébrique, nécessitant l'extraction de valeurs expérimentales contenues dans les spectres de masse.

Dans une première partie, nous verrons le principe de la méthode de suivi du marquage, permettant de contrôler la qualité de l'expérience. Cette étape est critique puisque, dans le cas où le marquage ne serait pas correct, la mesure de quantification finale serait erronée. L'indicateur permettant le suivi du marquage consiste à calculer l'incorporation de ^{12}C dans les molécules biologiques, observées par spectrométrie de masse.

Dans la deuxième partie, l'incorporation de ^{12}C sera considérée comme maximale et l'expérience de quantification SLIM peut être réalisée. Cette méthode consiste à effectuer un mélange des échantillons obtenus dans deux conditions biologiques différentes, dont l'une seulement est marquée. Dans celle-ci, l'incorporation de ^{12}C est connue et correspond à celle de la source de carbone (glucose) utilisée. L'autre condition, elle, possédera des abondances isotopiques naturelles. La détermination de la fraction molaire des peptides non marqués permettra de calculer les valeurs quantitatives associées.

I. Calcul de la probabilité de l'isotope ^{12}C à partir des mesures expérimentales des ions M_0 et M_1 (et de la séquence des peptides – in silico)

I.1 Valeur d'incorporation théorique du ^{12}C dans les molécules

Une molécule chimique est composée d'un mélange d'isotopes de chacun des atomes la composant. Dans le vivant, la probabilité d'obtenir des isotopes lourds et

instables (^{14}C par exemple) est extrêmement faible. C'est la raison pour laquelle elle est négligée. Selon les lois de conservation de la matière, la somme des abondances de chacun des isotopes stables des éléments chimiques, exprimée sous forme de probabilités, est une constante. Cette valeur est analogue à un tirage aléatoire avec remise, d'une certaine forme d'isotope parmi l'ensemble des atomes de l'élément. Ainsi pour une molécule donnée, la probabilité de tirage de l'isotope stable de carbone ^{13}C, notée $P(^{13}C)$ parmi un ensemble total d'atomes de carbone est de 1.07%. Pour le ^{12}C, la probabilité notée $P(^{12}C)$ est de 98.93%.

Il est possible d'écrire :

$$P(^{12}C) + P(^{13}C) = 1$$

La méthode SLIM de quantification par spectrométrie de masse repose sur le marquage d'une condition biologique, en effectuant une modification de l'abondance isotopique des protéines de cette condition donnée. Ce marquage *in-vivo* est rendu possible d'une part par la capacité à utiliser des méthodes algébriques de suivi et de contrôle de la condition marquée, et d'autre part d'effectuer une quantification relative.

Lors d'une expérience SLIM, il est essentiel de suivre l'incorporation des atomes de ^{12}C dans les protéines, puisque cela permet de confirmer l'état du marquage dans les échantillons biologiques étudiés. L'incorporation d'atomes de ^{12}C dans les molécules s'observe en spectrométrie de masse par un changement caractéristique des spectres obtenus. L'objectif est donc de développer des méthodes algébriques permettant de suivre le profil du spectre afin d'en révéler l'incorporation de ^{12}C apparente. Par définition et s'agissant d'un enrichissement, cette valeur est strictement supérieure ou égale à la valeur de la condition naturelle. Néanmoins, elle restera uniquement théorique puisque lors d'une procédure de marquage type, une population hétérogène de molécules est générée. Celle-ci est composée d'une première partie de molécules entièrement « naturelles » et donc ayant incorporé les isotopes lourds des éléments chimiques. La deuxième partie est composée des molécules enrichies en ^{12}C et donc n'ayant pas incorporé d'isotopes lourds du carbone. Un spectromètre de masse observe une quantité finie de molécules de cette population mixte, et ne différencie pas les deux parties. L'appareil produit donc des spectres qui correspondent à une somme des deux sous-ensembles de la population, analogue à un mélange "quantité à

quantité". Les variations dans la composition du protéome se traduisent ainsi par des variations de quantités relatives entre les deux échantillons, et donc par la quantité d'espèces observées par le spectromètre de masse.

Au début de ce projet, nous modélisions l'incorporation du ^{12}C dans les peptides/protéines d'un échantillon en utilisant le calcul d'une corrélation linéaire par rapport à des mélanges théoriques effectués. Cependant, si cette modélisation est pertinente pour des cas extrêmes où la population des molécules est composée exclusivement de molécules provenant soit de la condition totalement marquée (100%SLIM, notée 12C), soit de la condition naturelle (notée NC), cela n'est pas le cas pour les autres séries de valeurs intermédiaires, sous forme des différents mélanges des deux conditions. Or, dans un contexte de mélange, les valeurs d'incorporation calculées sont purement théoriques et expliquées par le fait que ce sont des observations simultanées, donc à l'origine d'un état composite. La non-linéarité des valeurs de l'incorporation du ^{12}C est montrée en Figure 27 et s'explique par le fait que cette valeur représente un calcul théorique de l'incorporation du ^{12}C à partir du ratio M_1/M_0 (en rouge sur la figure). Néanmoins, les valeurs d'incorporations de ^{12}C dans les molécules dépendantes du taux de ^{12}C dans la source de carbone utilisée, seront strictement linaires dans le cas d'une source répondant au mélange attendu (en bleu).

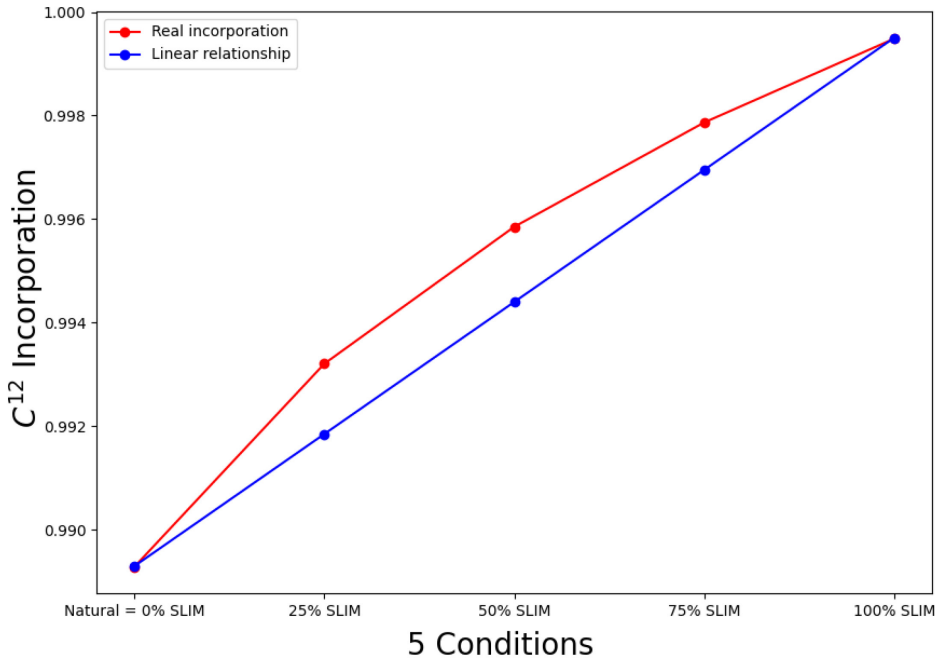


Figure 27 : Valeur d'incorporation théorique de ^{12}C dans le peptide in-silico de la levure S. cerevisiae.

I.2 Calcul de la probabilité de l'isotope ^{12}C à partir des mesures expérimentales des ions M_0 et M_1

Trouver un moyen de suivre l'incorporation des atomes de ^{12}C dans les protéines, à partir des spectres de masse obtenus, est donc un objectif essentiel pour le développement de la méthode SLIM. Dans la première partie, nous avons défini les valeurs expérimentales en suivant les règles décrites, selon un modèle théorique. Un autre aspect à étudier réside au sein même des données des spectres de masse, en particulier les intensités des isotopologues qui permettent la différenciation entre la condition marquée et la condition non marquée (Figure 28). En exprimant les équations binomiales décrivant l'intensité du monoisotopique et du premier isotopologue, une relation linéaire existe entre le quotient des intensités par rapport à l'incorporation de ^{12}C. Cela nous permet d'étudier le comportement du rapport, et en particulier d'émettre des hypothèses sur les règles algébriques répondant à cette observation.

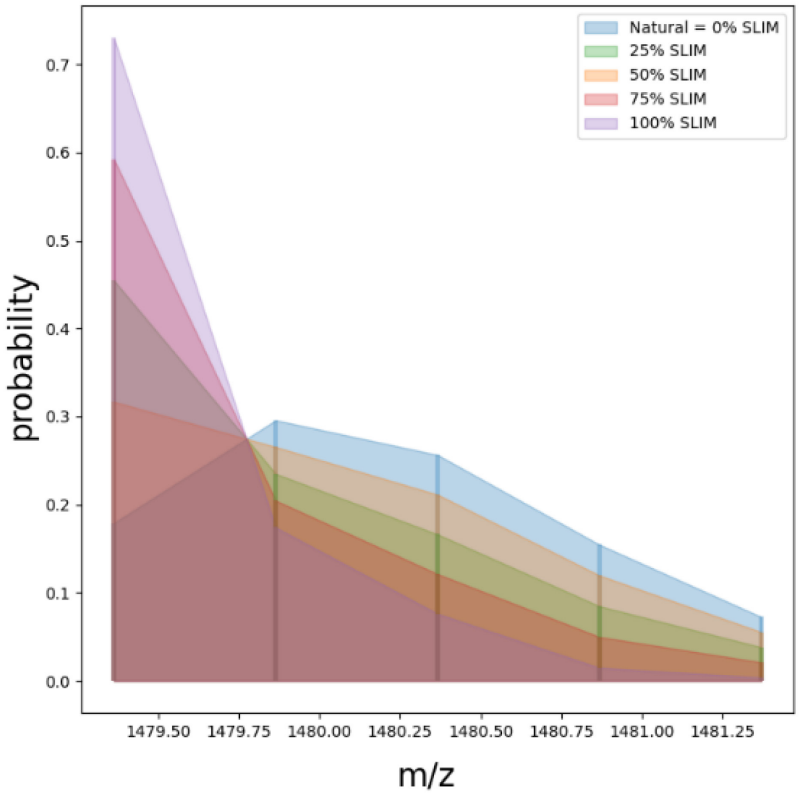


Figure 28 : évolution d'un cluster isotopique en fonction des différents ratios de mélanges des conditions marquée et non marquée.

Il est à noter que, plus l'incorporation du ^{12}C est importante (déplétion en ^{13}C), plus le rapport des intensités M_1/M_0 est petit (intensité du monoisotopique très grande en comparaison avec l'intensité du M_1). Le coefficient directeur de la relation linéaire est donc proportionnel à l'apport de ^{12}C dans les peptides. Les valeurs des coefficients directeur de chaque mélange, déterminées de manière théorique par des simulations numériques, sont comprises entre 0,01 et 3,6. Cela nous permet de calculer l'incorporation du ^{12}C à partir des données théoriques et expérimentales, en exprimant les données expérimentales en fonction des données théoriques. Cependant les données théoriques étant normalisées, il est nécessaire de normaliser les intensités expérimentales. Le développement d'une méthode de normalisation indépendante de l'intensité des autres isotopologues du massif isotopique est une nécessité. Dans le cas d'un ratio d'intensité, une simplification permet de se passer de l'intensité du *cluster*.

$$M_0 = \frac{M_0^{exp}}{\sum_0^n M_i^{exp}} \; et \; M_1 = \frac{M_1^{exp}}{\sum_0^n M_i^{exp}} \qquad (2\text{-}1)$$

Ce qui permet d'écrire la relation suivante, déterminante car à l'origine d'une simplification importante de la méthode de calcul entre la méthode SLIM et la méthode bSLIM :

$$=> \frac{M_1}{M_0} = \frac{M_1^{exp}}{M_0^{exp}} \qquad (2\text{-}2)$$

Dans le cas d'un marquage bSLIM, on note $P'(^{12}C)$ la probabilité de l'incorporation du ^{12}C dans les peptides, et $P'(^{13}C)$ la probabilité de l'incorporation résiduelle des ^{13}C.

$$\frac{M_1^{exp}}{M_0^{exp}} = nC \times \frac{P\prime(^{13}C)}{P\prime(^{12}C)} + nH \times \frac{P(^{2}H)}{P(^{1}H)} + nN \times \frac{P(^{15}N)}{P(^{14}N)} + nO \times \frac{P(^{17}O)}{P(^{16}O)} + nS \times \frac{P(^{33}S)}{P(^{32}S)} \qquad (2\text{-}3)$$

Le nombre d'atomes de carbones étant constant, cela nous permet d'écrire.

$$P'(^{13}C) = 1 - P'(^{12}C) \qquad (2\text{-}4)$$

Ainsi, le calcul (2-3) peut se simplifier comme cela :

$$\frac{M_1^{exp}}{M_0^{exp}} = nC \times \frac{[1-P\prime(^{12}C)]}{P\prime(^{12}C)} + B \qquad (2\text{-}5)$$

Où B est la somme des apports de chaque atome naturel restant (H, N, O, S)

$$B = nH \times \frac{P(^2H)}{P(^1H)} + nN \times \frac{P(^{15}N)}{P(^{14}N)} + nO \times \frac{P(^{17}O)}{P(^{16}O)} + nS \times \frac{P(^{33}S)}{P(^{32}S)}. \quad (2\text{-}6)$$

Ce qui nous permet d'exprimer $P'(^{12}C)$ l'incorporation du ^{12}C dans les peptides, en fonction de l'intensité expérimentale du M_0 et du M_1.

$$P'(^{12}C) = \frac{nC \times M_0^{exp}}{M_1^{exp} + (nC - B) \times M_0^{exp}} \quad (2\text{-}7)$$

Grâce à ce calcul, il est possible de suivre l'incorporation ^{12}C et ainsi de contrôler l'intensité du marquage.

Un résultat important de ce travail est la capacité à suivre l'incorporation des atomes de ^{12}C dans les molécules, avec un calcul algébrique qui utilise des données expérimentales (spectres de masses) et théoriques (formule chimique d'identification et constante physique). Le deuxième résultat est le fait que dans le cadre d'un mélange de peptides venant des conditions NC et 12C, les acides aminés seront tous soit intégralement marqués, soit intégralement non-marquée, et le spectre de masse observé permettra d'extraire expérimentalement les valeurs d'incorporation. De plus, dans le cas où la source de carbone ne serait pas entièrement marquée il y aurait une dilution du marquage. Par exemple, dans le cadre d'une source de carbone à demi marquée, l'abondance théorique de ^{12}C dans les acides aminés correspondrait à la moitié de la valeur du taux de ^{12}C naturelle, soit environ 99.94%. Ainsi, le spectre examiné serait donc le résultat d'observations de molécules, dont la moitié seulement des acides aminés composants les peptides seront marqués.

Cela permet également de suivre la fraction molaire des peptides non-marqués de façon indépendante aux probabilités de ^{12}C et permet d'avoir une valeur de quantification dans le cas d'une expérience classique en bSLIM.

II. Calcul de la fraction molaire pour la comparaison des abondances des peptides entre conditions

Comme vu précédemment lors d'une expérience bSLIM, un mélange à quantité égale (ou bien 1:1) des peptides venant d'une condition naturelle et d'une condition marquée est réalisé, puis analysé par spectrométrie de masse en une seule fois. Lors de la première mesure en masse MS1, un cluster isotopique expérimental bSLIM est obtenu (Figure 29). Il est la combinaison d'un cluster isotopique théorique naturel (en bleu) et d'un cluster isotopique théorique provenant d'une condition 100% 12C (en rouge). Ainsi dans cette application, le paramètre d'incorporation de ^{12}C est fixé et correspond à une valeur naturelle ou bien totalement marquée (Figure 30).

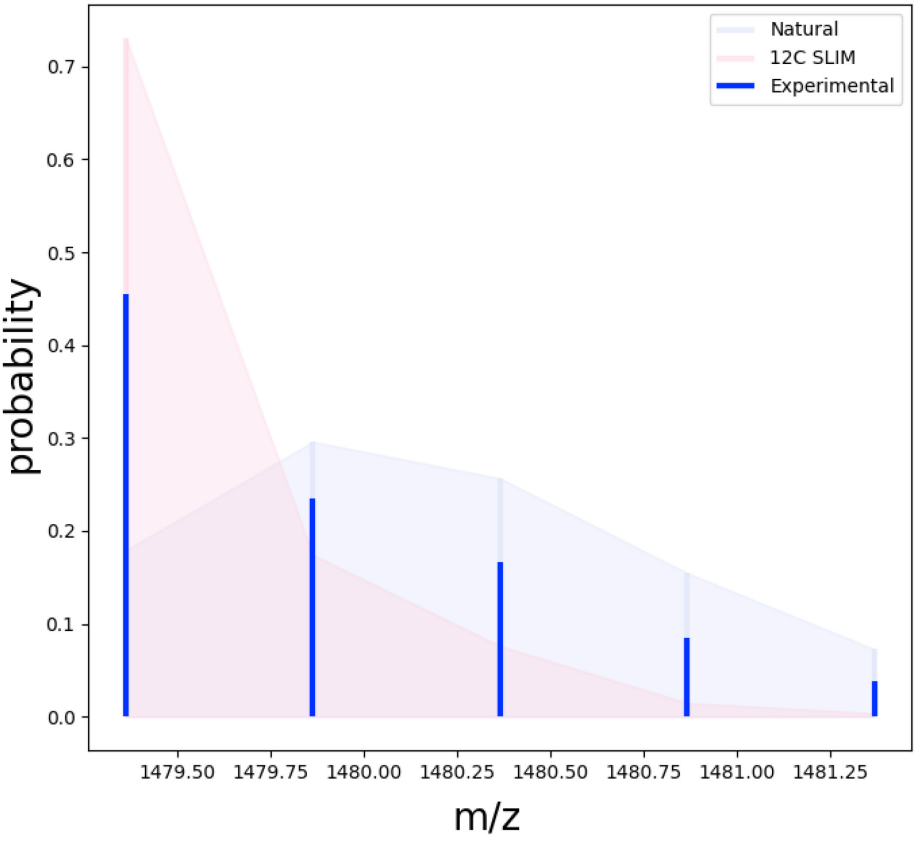


Figure 29 : Quantifier en bSLIM revient à évaluer dans un cluster expérimental l'apport respectif des clusters bleu et rouge (valeurs théoriques ici, provenant de la modélisation du peptide de séquence YIGAGISTIGLLGAGIGIAIVFAALINGVSR).

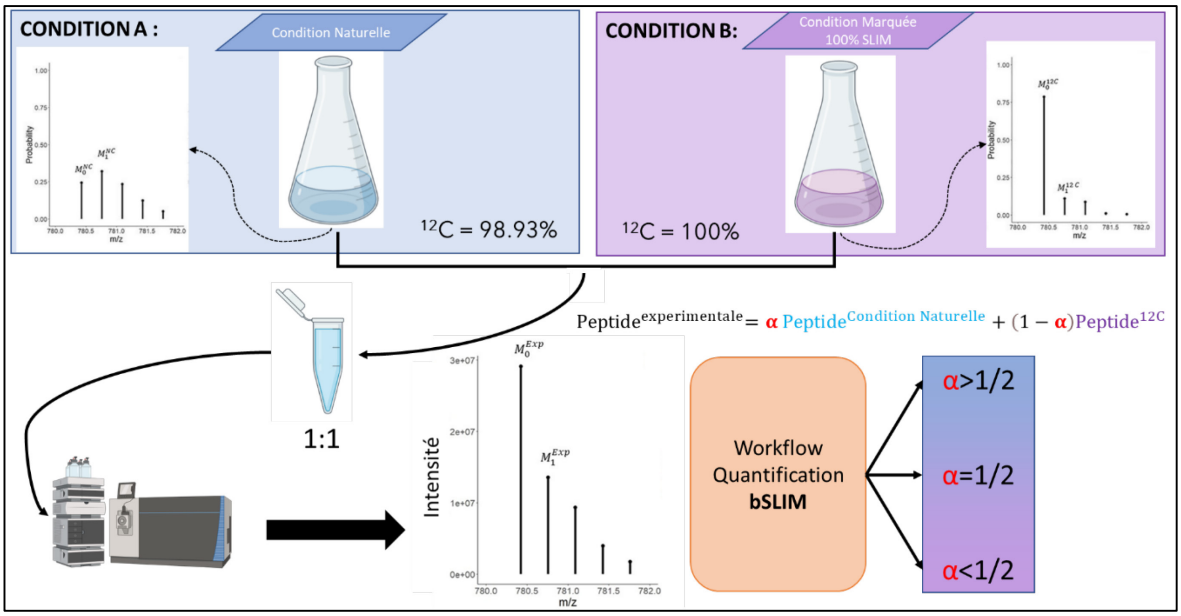


Figure 30: Schéma de principe de la quantification bSLIM (version Bottom-up du SLIM)

L'enjeu de la quantification bSLIM est donc pour tout cluster isotopique expérimental observé et identifié en masse, de mesurer l'apport respectif de chacun de ces deux clusters théoriques. Cette information quantitative correspond à la fraction molaire des peptides non-marquée et est notée alpha (**α**).

$$\text{Peptide}^{\text{experimental}} = \boldsymbol{\alpha}\ \text{Peptide}^{\text{Condition Naturelle}} + (1 - \boldsymbol{\alpha})\text{Peptide}^{12C}$$

Ainsi un cluster expérimental correspond à α, l'apport des molécules provenant de la condition naturelle (produisant un cluster Naturel) et (1 - α) l'apport des molécules provenant de la condition marquée (produisant un cluster 100% SLIM).

Si l'on note les différentes distributions isotopiques des conditions :

$$peptide^{exp} = (M_0^{exp}, M_1^{exp}, M_2^{exp}, M_3^{exp} \dots M_n^{exp})$$

$$peptide^{NaturalCondition} = (M_0^{NC}, M_1^{NC}, M_2^{NC}, M_3^{NC} \dots M_n^{NC})$$

$$peptide^{SLIM\ 12C} = (\mathrm{M}_0^{12C}, \mathrm{M}_1^{12C}, \mathrm{M}_2^{12C}, \mathrm{M}_3^{12C} \dots \mathrm{M}_n^{12C})$$

Cela nous permet d'exprimer cette égalité :

$$(M_0^{exp}, M_1^{exp}, M_2^{exp}, M_3^{exp} \dots M_n^{exp}) = \boldsymbol{\alpha} * (M_0^{NC}, M_1^{NC}, M_2^{NC}, M_3^{NC} \dots M_n^{NC}) + (1 - \boldsymbol{\alpha}) * (M_0^{12C}, M_1^{12C}, M_2^{12C}, M_3^{12C} \dots M_n^{12C})$$

Or, nous pouvons exprimer les intensités expérimentales de M_0 et de M_1 par une expression théorique :

$$\frac{M_0^{exp}}{\sum_0^n M_i^{exp}} = \alpha M_0^{NC} + (1-\alpha) M_0^{12C} \text{ et } \frac{M_1^{exp}}{\sum_0^n M_i^{exp}} = \alpha M_1^{NC} + (1-\alpha) M_1^{12C} \quad (2\text{-}8)$$

Ainsi, la valeur du rapport l'intensité M_1/ M_0 en fonction de α, et permet de poser l'équation suivante.

$$\frac{M_1^{exp}}{M_0^{exp}} = \frac{\alpha M_1^{NC} + (1-\alpha) M_1^{12C}}{\alpha M_0^{NC} + (1-\alpha) M_0^{12C}} = \frac{\alpha (M_1^{NC} - M_1^{12C}) + M_1^{12C}}{\alpha (M_0^{NC} - M_0^{12C}) + M_0^{12C}} \quad (2\text{-}9)$$

$$M_1^{exp} * (\alpha (M_0^{NC} - M_0^{12C}) + M_0^{12C}) = M_0^{exp} * (\alpha (M_1^{NC} - M_1^{12C}) + M_1^{12C}) \quad (2\text{-}10)$$

$$M_0^{exp} \alpha M_1^{NC} - M_0^{exp} \alpha M_1^{12C} - M_1^{exp} \alpha M_0^{NC} + M_1^{exp} \alpha M_0^{12C} = M_1^{exp} M_0^{12C} - M_0^{exp} M_1^{12C} \quad (2\text{-}11)$$

$$\alpha (M_0^{exp} M_1^{NC} - M_0^{exp} M_1^{12C} - M_1^{exp} M_0^{NC} + M_1^{exp} M_0^{12C}) = M_1^{exp} M_0^{12C} - M_0^{exp} M_1^{12C} \quad (2\text{-}12)$$

Une fois résolu, on obtient une expression claire et qui permet d'exprimer la fraction molaire α des peptides non-marqués.

$$\alpha = \frac{M_0^{12C} M_1^{exp} - M_0^{exp} M_1^{12C}}{M_0^{exp} M_1^{NC} - M_0^{exp} M_1^{12C} - M_1^{exp} M_0^{NC} + M_1^{exp} M_0^{12C}} \quad (2\text{-}13)$$

Le ratio des deux valeurs quantitatives, $\frac{NC}{12C} = \frac{\alpha}{(1-\alpha)}$ peut être exprimé comme étant une valeur de quantification relative, analogue à un taux d'expression qui, sous forme de $\text{Log}_2(Ratio)$, peut être interprété biologiquement.

Nous avons donc développé un modèle théorique robuste qui permet de quantifier les peptides de deux conditions, sous le terme de « fraction molaire ». La fraction molaire peut être très facilement calculée à l'aide des valeurs expérimentales et des valeurs théoriques du couple M_1 et M_0.

III. Cas des marquages incomplets

La méthode décrite ci-dessus est strictement limitée à des organismes *autotrophes* c'est-à-dire dont le métabolisme permet la synthèse de tous les acides aminés, à partir d'une source unique de carbone. Dans ce cas précis, le marquage est uniforme et idéal, puisque le ^{12}C est incorporé de manière parfaitement prévisible dans les séquences protéiques.

Cependant, la grande majorité des organismes d'intérêt dans les laboratoires, possèdent soit des délétions de gènes essentiels à la biosynthèse de certains acides aminés (souche de levure particulière), soit sont des organismes *auxotrophes*, c'est-à-dire qui ont besoin d'un apport d'acides aminés exogènes dans le milieu. Par exemple, les lignées de cellules humaines requièrent l'ajout des 10 acides aminés essentiels pour leur croissance. Toutefois, la supplémentation du milieu avec des acides aminés U-[^{12}C] n'est pas possible (à l'heure actuelle l'entreprise Eurisotope n'en produit pas) et aurait un prix très élevé. De plus, les eucaryotes supérieurs (organismes multicellulaires) ne pourraient être cultivés de cette manière puisque le complément en acides aminés se fait par le tractus digestif. C'est pourquoi il existe un besoin important de développer la méthode pour pouvoir réaliser des analyses, dans ce cas précis pour lequel le marquage est incomplet. Il s'agit de prendre en compte dans les calculs le cas particulier où il y a eu un apport de sources de carbone contenant du ^{13}C exogène rendant l'incorporation de ^{12}C incomplète et plus difficilement prévisible dans les peptides.

III.1 Calcul et modélisation de l'intensité des isotopologues dans des organismes auxotrophes

Pour toute séquence peptidique il est essentiel de différencier les acides aminés pouvant être synthétisés et qui seront ainsi marqués, de ceux ne le pouvant pas et qui demeureront à l'état naturel. Ce paramètre est calculé grâce à un regroupement des formules chimiques pour chaque acide aminé. Par exemple la souche de levure *Saccaromyces cerevisiae* « BY4742 » présente une auxotrophie pour la base nucléique Uracile et trois acides aminés : l'Histidine (H), la Leucine (L) et la Lysine (K).

Dans ce contexte, le peptide suivant TENLHQSLTGCLNDYSNAFGK peut être décomposé en deux peptides théoriques : l'un totalement « marquable » (TENQSTGCNDYSNAFG), l'autre ne pouvant être enrichi en ^{12}C (LHLLK).

Posons *x* l'ensemble des atomes de carbone composant ce peptide. Notons *nCa* le nombre d'atomes de carbones provenant des acides aminés de la première séquence et *nCb* ceux de la deuxième séquence. Soit mathématiquement $x = nCa + nCb$.

Dans le cas de la condition naturelle, l'expression littérale de l'intensité normalisée des isotopologues ne varie pas. Cependant, dans le cas de la condition 12C, nous fixons l'abondance p(^{12}C) à 100%. Or l'expression de l'intensité de l'ion monoisotopique de ce peptide résulte de la composition de deux populations de source d'atomes de carbone. La probabilité d'abondance de ^{12}C par les acides aminés marquables (carbones *a)* est décrite par la notation $P(^{12}C_a) = 100\%$. Pour le carbone *b* décrit par les acides aminés non-marquables, l'abondance isotopique des ^{12}C est $P(^{12}C_b) = 98.93$ abondance naturelle.

Ainsi, l'intensité du monoisotopique de la condition 12C est décrite par :

$$M_0^{12C.max} = P(^{12}C_a)^{nCa} * P(^{12}C_b)^{nCb} * P(^{1}H)^{nH} * P(^{14}N)^{nN} * P(^{16}O)^{nO} * P(^{32}S)^{nS}$$

De manière identique, le premier isotopologue est décrit par la combinaison de l'apport isotopique de chacun des éléments chimiques et des deux populations de carbones, soit :

$$\begin{aligned} M_1^{12C.max} = {} & nCa * P(^{12}C_a)^{nCa-1} * P(^{12}C_b)^{nCb} * P(^{13}C) * P(^{1}H)^{nH} * P(^{14}N)^{nN} * P(^{16}O)^{nO} * P(^{32}S)^{nS} \\ & + nCb * P(^{12}C_b)^{nCb-1} * P(^{12}C_a)^{nCa} * P(^{13}C) * P(^{1}H)^{nH} * P(^{14}N)^{nN} * P(^{16}O)^{nO} * P(^{32}S)^{nS} \\ & + P(^{12}C_b)^{nCb} * P(^{12}C_a)^{nCa} * nH * P(^{1}H)^{nH-1} * P(^{2}H) * P(^{14}N)^{nN} * P(^{16}O)^{nO} * P(^{32}S)^{nS} \\ & + P(^{12}C_b)^{nCb} * P(^{12}C_a)^{nCa} * P(^{1}H)^{nH} * nN * P(^{14}N)^{nN-1} * P(^{15}N) * P(^{16}O)^{nO} * P(^{32}S)^{nS} \\ & + P(^{12}C_b)^{nCb} * P(^{12}C_a)^{nCa} * P(^{1}H)^{nH} * P(^{14}N)^{nN} * nO * P(^{16}O)^{nO-1} * P(^{17}O) * P(^{32}S)^{nS} \\ & + P(^{12}C_b)^{nCb} * P(^{12}C_a)^{nCa} * P(^{1}H)^{nH} * P(^{14}N)^{nN} * P(^{16}O)^{nO} * nS * P(^{32}S)^{nS-1} * P(^{33}S) \end{aligned}$$

Dans les calculs, les atomes de carbone *a et b* sont donc considérés comme des atomes distincts et possédant des abondances isotopiques différentes.

III.2 Fraction molaire corrigée pour la comparaison des abondances des peptides entre conditions dans des organismes auxotrophes

L'objectif de la méthode est d'obtenir la valeur de la fraction molaire des peptides, provenant de la condition naturelle. Les calculs d'expression de l'intensité des isotopologues nous permet ainsi de modéliser fidèlement M_1 et M_0, en tenant compte du biais dû à l'auxotrophie. Ainsi l'expression de la fraction molaire alpha (**α**) comprenant l'expression de $M_0^{12C.max}$ et $M_1^{12C.max}$ est la suivante :

$$\alpha = \frac{M_0^{12C.max} M_1^{exp} - M_0^{exp} M_1^{12C.max}}{M_0^{exp} M_1^{NC} - M_0^{exp} M_1^{12C.max} - M_1^{exp} M_0^{NC} + M_1^{exp} M_0^{12C.max}} \qquad (2\text{-}14)$$

III.3 L'incorporation réelle de ^{12}C

Les calculs de l'abondance expérimentale de ^{12}C pouvant être incorporé dans les acides aminés marquables du peptide est notée $P'(^{12}C_a)$. Cette valeur, fixée jusqu'alors pour les calculs de quantification, doit être déterminée avec précision pour la validation de la méthode. De ce fait, comme détaillé précédemment et par analogie, nous pouvons poser l'égalité suivante.

$$\frac{M_1^{exp}}{M_0^{exp}} = nC \times \frac{[1 - P'(^{12}C_a)]}{P'(^{12}C_a)} + B' \qquad (2\text{-}15)$$

Où B' est la somme des apports de chaque atome naturel restant (H, N, O, S), ainsi que le carbone *b*.

$$B' = nCb * \frac{P(^{13}C_b)}{P(^{12}C_b)} + nH \times \frac{P(^{2}H)}{P(^{1}H)} + nN \times \frac{P(^{15}N)}{P(^{14}N)} + nO \times \frac{P(^{17}O)}{P(^{16}O)} + nS \times \frac{P(^{33}S)}{P(^{32}S)} \quad (2\text{-}16)$$

L'expression de l'abondance isotopique expérimentale de ^{12}C pouvant être incorporée dans les peptides provenant des acides aminés marquable est donc la suivante :

$$P'(^{12}C_a) = \frac{a \times M_0^{exp}}{M_1^{exp} + (a - B\prime) \times M_0^{exp}} \qquad (2\text{-}17)$$

Toutefois, la valeur de l'abondance expérimentale de ^{12}C pouvant être incorporée dans le peptide est la composition des deux probabilités soit :

$$P'(^{12}C) = \frac{a \times P\prime(^{12}C_a) + b \times P(^{12}C_b)}{a+b} \qquad (2\text{-}18)$$

IV. Mise en œuvre informatique

Développer un workflow d'analyse est un besoin important pour traiter ces données atypiques. De plus, afin d'être plus facilement utilisé par d'autres laboratoires que le nôtre, le workflow doit être simple et facile à installer. Nous avons pour cela choisi d'utiliser la plateforme d'environnement *Konstanz Information Miner* (KNIME) (Fillbrunn et al., 2017). C'est un outil populaire de virtualisation d'algorithmes indépendants, *via* l'utilisation d'une machine virtuelle Java et un fonctionnement par « nœud ». Chaque développeur peut participer au développement d'une "boite" personnelle contenant un algorithme et qui sera intégrée sur la plateforme par l'intermédiaire d'un nouveau nœud publique ou "restrain". Dans le domaine de la protéomique et de la spectrométrie de masse, la solution informatique OpenMS se regroupe également dans une même architecture, et permet d'exécuter les algorithmes puissants de traitement du signal en spectrométrie de masse. Le logiciel a pu être intégré à l'architecture de KNIME, et disposant des outils nécessaires à nos développements, nous avons donc sélectionné cette solution informatique. Plus récemment, certains algorithmes que j'ai développés en C++ ont été intégrés à un nœud, permettant une interface utilisateur simplifiée et facilement intégrable dans des workflows. Au final, le workflow bSLIM est utilisable, téléchargeable facilement, et il très aisément modifiable. Notamment, les différents filtres et calculs peuvent être adaptés afin d'optimiser les analyses de données diverses, issues d'expériences différentes.

IV.1 Identification

La première étape, une fois l'acquisition réalisée avec un spectromètre de masse, est de convertir les données spectrales en données informatiques plus facilement utilisables par les algorithmes. Ensuite les informations des signaux expérimentaux nécessitent d'être identifiées, telles que les séquences peptidiques à l'origine des signaux. Cela permet de connaitre la formule chimique, essentielle aux nombreux calculs ci-dessus développés, ainsi que d'extraire les données expérimentales associées. Concrètement les spectres de fragmentation (MS2) servent à l'étape d'identification, tandis que les données de la première mesure de masse (MS1) servent à l'étape de quantification. Ces dernières servent également à associer un score d'identification à la valeur quantitative, à la base de l'expérience en protéomique.

Ainsi, les fichiers expérimentaux bruts du spectromètre de masse, au format propriétaire « .raw » de l'entreprise *Thermo Fisher Scientific (*constructeur de l'appareil) sont convertis en fichier ouvert « .mzML » à l'aide de l'outil « MS-Convert » (Holman et al., 2014) de la suite algorithmique *Proteowizard* (Kessner et al., 2008). Durant cette conversion, une étape de *peak-picking* est également faite sur les MS1 et MS2 (Figure 31). En résumé, un algorithme de reconnaissance des pics permet de réduire les données en mode « profile » en centroïde. C'est-à-dire que les intensités et les positions en masse correspondent au barycentre et au maximum de la gaussienne décrite par le signal initial.

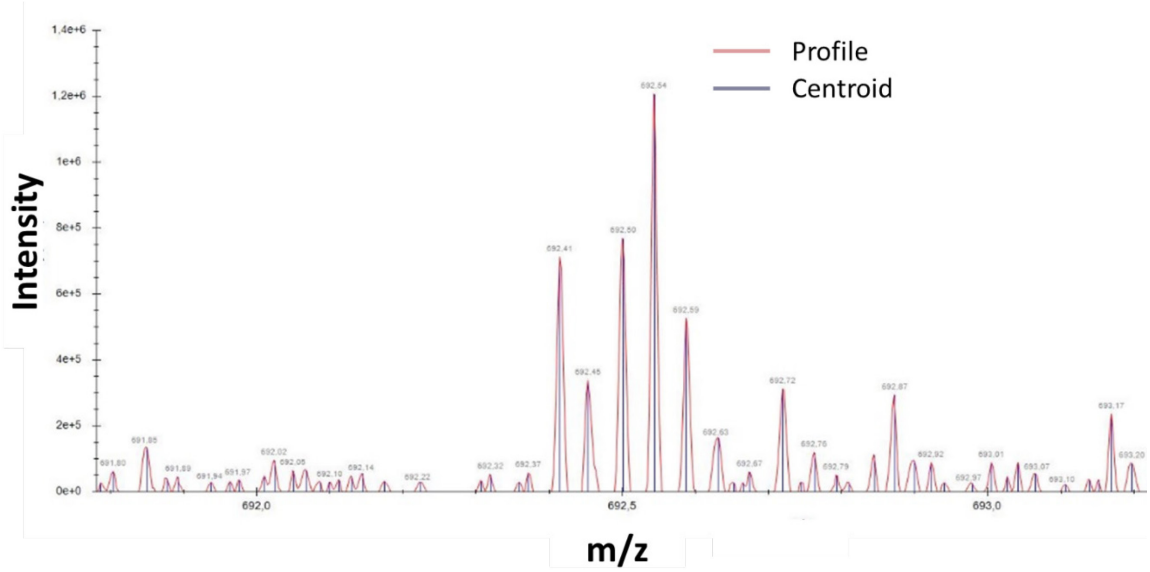


Figure 31: Exemple d'un spectre expérimental représentant la transformation des données avant et après l'étape de peak-picking.

Dans le but d'obtenir les identifications associées aux spectres, une recherche dans les banques de données est ensuite réalisée. Pour cela, on crée une banque de séquences au format « fasta », contenant les séquences protéiques du protéome de l'organisme référence. Par exemple la souche S288C de la levure *Saccaromyces cerevisiae* est disponible sur le site Uniprot, référence UP000002311. Les données d'identification issues d'analyses par spectrométrie de masse reflètent une probabilité de présence d'une séquence dans la banque. Dans le but de renforcer le score de significativité des identifications, une évaluation du taux de faux positifs est ajoutée. Ainsi, la banque de données est complétée par des séquences protéiques identiques, mais dont l'ordre des résidus est mélangé aléatoirement (technique « shuffle »). A chaque identification une séquence protéique réelle pourra être confrontée une éventuelle identification, dite « Decoy » (totalement aléatoire). Cette procédure est mise en œuvre par l'intermédiaire de l'algorithme « DecoyDatabase » d'OpenMS (Figure 32 A).

Enfin, une recherche en banque de données est effectuée à l'aide du logiciel X!Tandem (Duncan et al., 2005) implémenté dans le nœud KNIME « *XTandemAdapter* ». Divers filtres sont utilisés et les peptides ne répondant pas aux critères fixés sont écartés. Les peptides identifiés, qualifiés de *peptide spectrum matches* (PSM), sont alors triés selon leurs scores d'identification « *PeptideIndexer* » et par la suite, l'algorithme PSM*Feature extractor* détermine les statistiques associées à l'identification. Ils sont utilisés pour réaliser un filtrage à un taux de faux positivité inférieur à 5% via PercolatorAdapter. Les peptides aléatoires « Decoy » sont aussi écartés des fichiers d'identification et enfin un calcul comportant une correction Epifany est réalisée, pour l'ajout de critères de filtres statistiques supplémentaires (Pfeuffer et al., 2019).

Tous ces algorithmes de la suite logiciel OpenMS sont incorporés dans des nœuds du logiciel d'environnement KNIME (Konstanz Information Miner) formant ainsi le workflow complet d'identification (Figure 32 B).

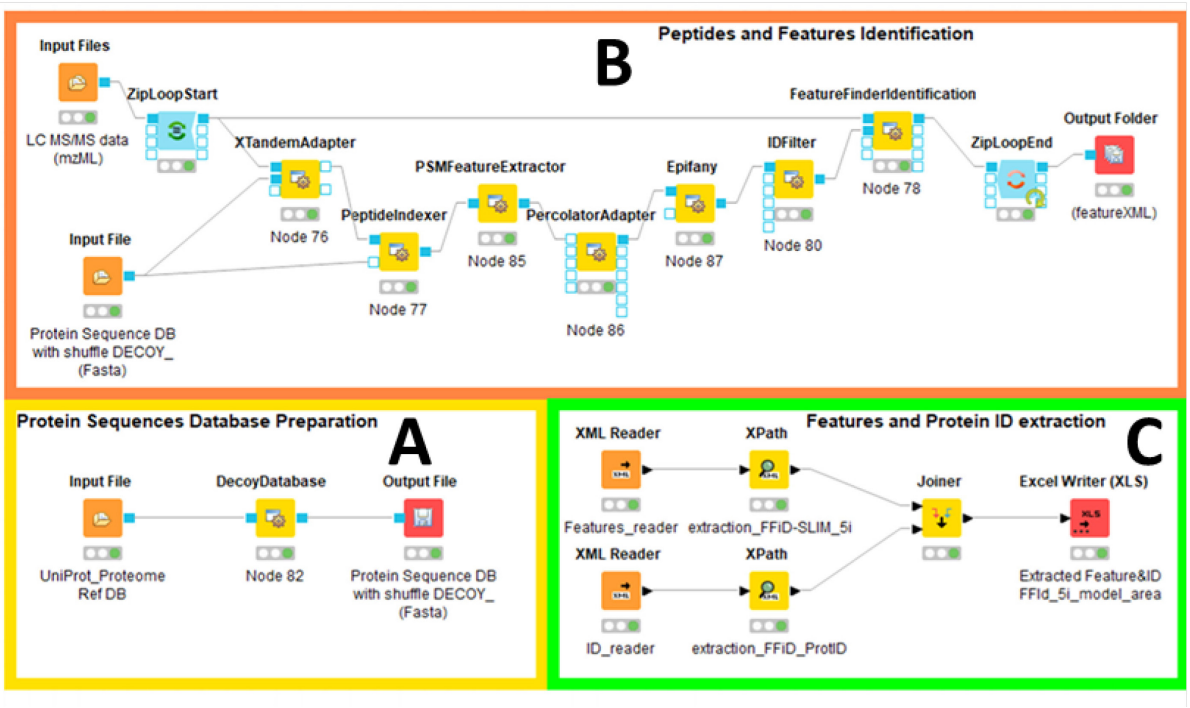


Figure 32 : Workflow d'identification KNIME créé et utilisé dans le cadre des expériences bSLIM. Cette figure est extraite de la publication (Sénécaut et al., 2021) (publication personnelle)

IV.2 L'identification, un facteur limitant de la quantification

L'un des grands enjeux en protéomique quantitative de type *label free* est la liaison entre identification et quantification. Particulièrement, une perte importante des données a lieu durant l'attribution d'une identification statistiquement robuste, à une valeur de quantification.

« Feature Finder Identification » est un outil initialement prévu pour effectuer de la quantification de type « label-free », c'est-à-dire une méthode dont les valeurs quantitatives sont issues d'extractions brutes des intensités peptidiques (Weisser & Choudhary, 2017) Figure 33. L'originalité de cet outil est que son fonctionnement repose sur une identification déjà préalablement effectuée afin de servir de point de repère pour procéder à une extraction des valeurs quantitatives. L'algorithme génère ainsi une fenêtre de masse qui est utilisée pour extraire dans le temps de rétention, les valeurs d'intensités expérimentales.

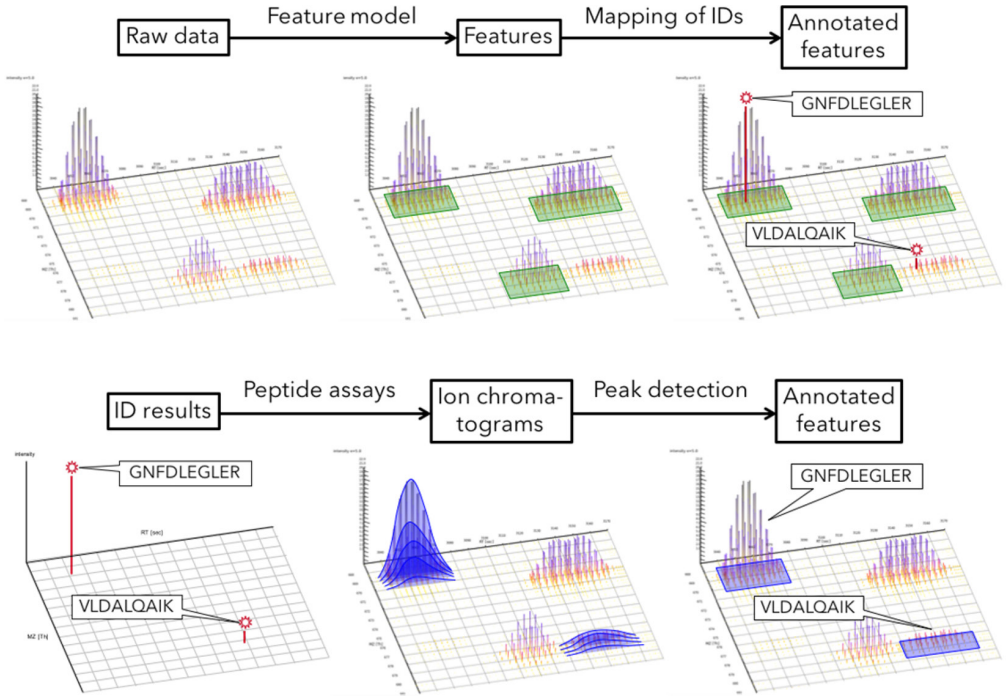


Figure 33 : Méthode de quantification de type label free fondée sur une extraction des données issues d'identification. Figure extraite de (Weisser & Choudhary, 2017)

IV.3 Fonctionnement de l'outil

L'algorithme Feature Finder Identification nécessite donc deux fichiers d'entrée au format XML : les données spectrales *.mzml* d'une part, et les données d'identification *.idxml* d'autre part. Ce dernier fichier est "balisé", c'est à dire qu'il contient les annotations de chaque peptide identifié ainsi que les informations de : la charge, la masse, le temps de rétention, le numéro du scan, le score, ainsi que la séquence peptidique responsable de son identification, et enfin le nom de la protéine associée. Ces informations sont essentielles pour la localisation précise du spectre parmi l'ensemble des données spectrales. Cela permet de délimiter une fenêtre à deux dimensions : masse (m/z) et temps de rétention (RT) appelée « feature ». Cette zone liée à l'identification (trace de masses) permet à l'algorithme de procéder à une extraction de l'enveloppe isotopique sur l'ensemble du temps chromatographique. Nous avons choisi cet algorithme car il permet une extraction des données expérimentales de manière exhaustive, mais également car il permet de réaliser une correction des valeurs expérimentales par l'application d'un modèle théorique. En effet, l'intensité de chacun des isotopologues le long du temps de rétention, est compatible avec un modèle de courbe de Gauss. C'est la raison pour laquelle, un

modèle gaussien est appliqué sur chaque trace d'élution des peptides, le long du temps de rétention. L'algorithme d'approximation du modèle théorique utilise la méthode dite des moindres carrées et utilisant l'algorithme Levenberg–Marquardt[1]. Les valeurs des paramètres initiaux sont déterminées selon des règles algébriques admises. Si le modèle est considéré comme valide, l'aire sous la courbe décrite par l'ion monoisotopique est utilisée comme mesure de l'intensité du peptide. C'est pourquoi l'algorithme de *Feature Finder Identification (FFId)* a été identifié comme pertinent vis à vis de notre problématique. Néanmoins, une modification de l'algorithme original est requise dans le but d'obtenir des informations complémentaires. La modification consiste en l'ajout d'une boucle supplémentaire afin de réaliser un ajustement gaussien et de produire les valeurs d'intensité pour chacun des isotopologues de l'enveloppe isotopique étudiée. Cette modification a été faite par le développeur de ce logiciel, Hendrik Weisser, puis incorporée dans le logiciel OpenMS (Version 2.6). Ainsi, l'algorithme dispose d'un nouveau paramètre décrivant le nombre d'isotopologues à extraire. Simultanément, la nouvelle version a été implémentée dans l'architecture KNIME et ajoutée dans notre workflow de quantification.

IV.4 Quantification à l'aide des différents nœuds de calcul

Notre objectif premier est donc de déterminer la fraction molaire des peptides à partir des valeurs des intensités des spectres. A cet effet, notre méthode de calcul utilise les valeurs des intensités expérimentales et théoriques des isotopologues M_0 et M_1. L'obtention des valeurs théoriques de l'intensité des isotopologues est déterminée par une résolution algébrique, comme décrit plus haut. A cet effet, plusieurs nœuds de calcul KNIME permettent de décomposer/décompter les acides aminés de la séquence peptidique afin de produire la formule chimique brute. Par la suite, le calcul des valeurs d'intensités théoriques consiste à résoudre les équations binomiales. Toutefois, le retraitement des valeurs expérimentales extraites par FFId nécessite un traitement particulier. En effet, le fichier produit par FFId est un .featureXML. C'est un fichier

[1] https://fr.wikipedia.org/wiki/Algorithme_de_Levenberg-Marquardt

balisé qui à chaque identification associe les valeurs expérimentales extraites, puis celles du modèle confronté. Une réduction de l'ensemble des données en sélectionnant des descripteurs et en produisant une matrice est requise, car ce fichier est très volumineux (taille de l'espace disque occupé supérieur à 3Go). Pour cela, dans notre workflow KNIME, nous procédons à un retraitement du fichier par le nœud Xpath permettant la sélection des données indispensables.

Par ailleurs, afin de faciliter les conversions de fichiers ultérieures, toutes les modifications post-traductionnelles identifiées et concaténées aux séquences peptidiques sont la forme de mots clés, sont réannotées. Par exemple, les *Oxydation* sont remplacées par un O, les *Acétylisation* par un B et les *Phosphorylation* par un Z. Lors de cette étape, réalisée par plusieurs nœuds KNIME, une procédure importante de filtrage des peptides est réalisée. Ceci dans le but d'écarter des résultats finaux les peptides dont les valeurs sont aberrantes. Ces « outliers » s'expliquent premièrement par des erreurs d'extraction des valeurs expérimentales et deuxièmement par des effets « artefacts » dus au mode d'analyse par spectrométrie de masse. Plus concrètement, il s'agit des peptides dont le modèle d'intensité des isotopologue M_0 et M_1, valeurs critiques dans ce calcul, ne serait pas valide (ou bien présentant de trop faibles mesures d'intensités, voire nulles). Il est à noter que le seuil correspondant à la valeur d'intensité "minimum" est strictement dépendant de l'appareil utilisé. Par exemple, notre workflow utilise une intensité supérieure à 1E6 pour les technologies *Orbitrap*, mais de seulement 1E2 pour la technologie Temps de vol du *tims-TOF Pro2*. Enfin, les peptides dont la masse est inférieure à 800 Da sont écartés, car ayant été observés comme induisant des valeurs aberrantes.

Ainsi, à l'issue du workflow présenté en Figure 34, la valeur finale de quantification pour chaque peptide est déterminée. Elle est exprimée comme étant le ratio de la fraction molaire sur la valeur complémentaire. Ceci permettant une comparaison relative entre les deux conditions analogues à un taux d'expression ou « fold change ». Toutefois, cette valeur est ensuite exprimée sous forme de logarithme base deux, à l'instar des pratiques habituelles dans les études effectuées en protéomique quantitative.

L'étape suivante réside dans l'association des valeurs quantitatives des peptides appartenant aux mêmes protéines. A cet effet, un script en langage R a été développé et a été inséré dans le workflow KNIME (nœud dédié du type « R Snippet »). La procédure d'association réside dans la détermination pour chaque protéine de la moyenne, de la déviation standard et de la médiane des valeurs quantitatives de l'ensemble des peptides associés. Dans une autre étape et selon les préférences de l'utilisateur, la sélection des trois peptides les plus intenses (« top 3 ») est utilisé pour décrire les protéines.

Par la suite, deux fichiers au format du tableur *Excel* sont produits, l'un contenant le descriptif des valeurs quantitatives et intermédiaires des différents nœuds de calcul au niveau des peptides, l'autre contenant uniquement le descriptif des valeurs quantitatives au niveau des protéines.

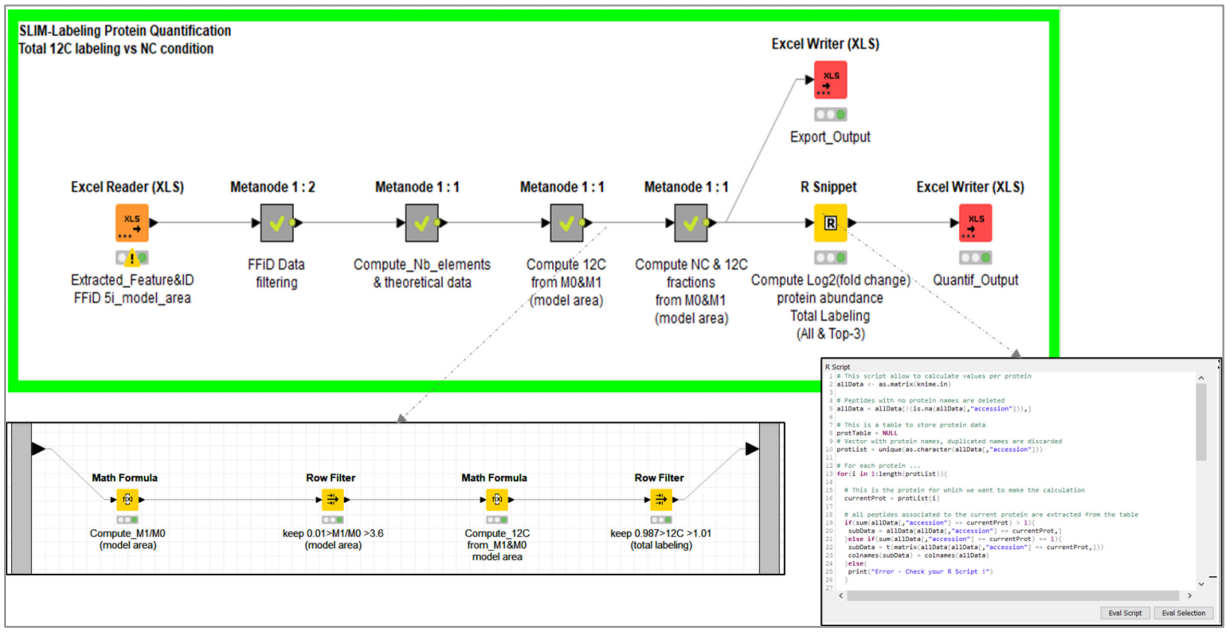


Figure 34 : Deuxième partie du Workflow KNIME de quantification en bSLIM à la suite d'une extraction par FFId.

Chapitre 2 : Simulations

Le développement de la méthode bSLIM a été rendu possible par la réalisation de modélisations et par l'utilisation de simulations numériques afin de confirmer les hypothèses émises. En particulier, cette étape a permis la validation des calculs permettant la quantification des abondances des protéines, à partir des données spectrales.

I. Suivi de l'intensité des Isotopologues en fonction du temps de rétention

Notre méthode repose donc sur une formule algébrique utilisant les valeurs des intensités expérimentales des isotopologues M_0 et M_1. Ces deux valeurs sont d'autant plus critiques qu'une variation légère (engendrée par une erreur de mesure par exemple) influe très fortement les résultats finaux quantitatifs. Or, pris indépendamment, les valeurs d'intensités des isotopologues sont « bruitées », c'est-à-dire que l'appareil effectue des corrections du signal qui sont variables le long de la masse, et donc du cluster isotopique. Une solution à cette limitation est d'obtenir plusieurs scans (MS1) pour le même ion. C'est justement l'un des avantages de procéder à une élution des peptides sur le temps de rétention par usage de la chromatographie. Concrètement, l'intensité d'une molécule extraite sur le temps est décrite par une courbe de Gauss. La forme de cette courbe est strictement dépendante du chromatogramme, c'est-à-dire qu'elle varie en fonction de la colonne et les solvants utilisés (Acétonitrile par exemple). C'est la raison pour laquelle une méthode fiable d'extraction des intensités expérimentales de chaque isotopologues le long du temps de rétention est nécessaire.

Ma première tâche a donc été de trouver des moyens logiciels et computationnels pour vérifier ces informations, et observer une amélioration lors du retraitement des données. Ensuite, j'ai cherché à obtenir une solution algorithmique facilement implémentable.

Pour rappel, une courbe de gausse est définie selon la fonction suivante :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Soit f(x) l'intensité de l'isotopologue de rang n, μ centre (ou moyenne) la courbe décrite par la fonction de Gauss, σ l'écart type de la courbe à mi-hauteur et x la position dans le temps de rétention.

Dans l'algorithme FFId, c'est la fonction utilisée afin d'être confrontée aux valeurs expérimentales extraites. Néanmoins, pour que le modèle d'élution d'un peptide soit valide, il faut que les paramètres décrits par fonction de la courbe de Gauss ne soient pas trop éloignés des valeurs limites déterminées à partir des valeurs expérimentales (Tableau 6).

Étiquette	Signification
(0 valid)	Le modèle est considéré comme valide selon les paramètres du filtre
(1 out of bound)	Le modèle n'est pas pertinent car l'aire est incorrecte
(2 out of bound)	Le modèle n'est pas pertinent car le centre du modèle μ n'est pas centré
(3 out of bound)	Le modèle n'est pas pertinent car il dépasse de trop de temps du côté gauche
(4 out of bound)	Le modèle n'est pas pertinent car il dépasse de trop de temps du côté droit

Tableau 6 : Différents contrôles qualité de FFId sur le modèle d'élution des isotopologues.

Avec un tel modèle, les isotopologues sont décrits par des critères beaucoup plus contrôlés et donc les valeurs d'intensité varient de façon moindre que celles mesurées par l'appareil. La fonction est décrite par plusieurs descripteurs, qui tous décrivent l'abondance relative des isotopologues. Différents indicateurs ont été testés, comme la hauteur de la courbe, la sommes des intensités sur le temps de rétention, et enfin la largeur à mi-hauteur. Au final, c'est l'aire de la courbe décrite qui a été observée comme la plus fiable, et le plus représentative des objets observés. Cependant, l'aire du modèle gaussien appliqué et la somme des valeurs des intensités sont mathématiquement

identiques si le pas entre deux mesures est petit. Or, la mesure de l'intensité des isotopologues est, sur un Orbitrap, extrêmement variable. Ainsi, prendre la valeur du modèle appliqué semble plus juste, surtout que l'ajustement d'un modèle gaussien permet de quantifier des ions présents de manière très peu intense, voire de rectifier des mesures erronées effectué par le détecteur d'ions (Figure 35).

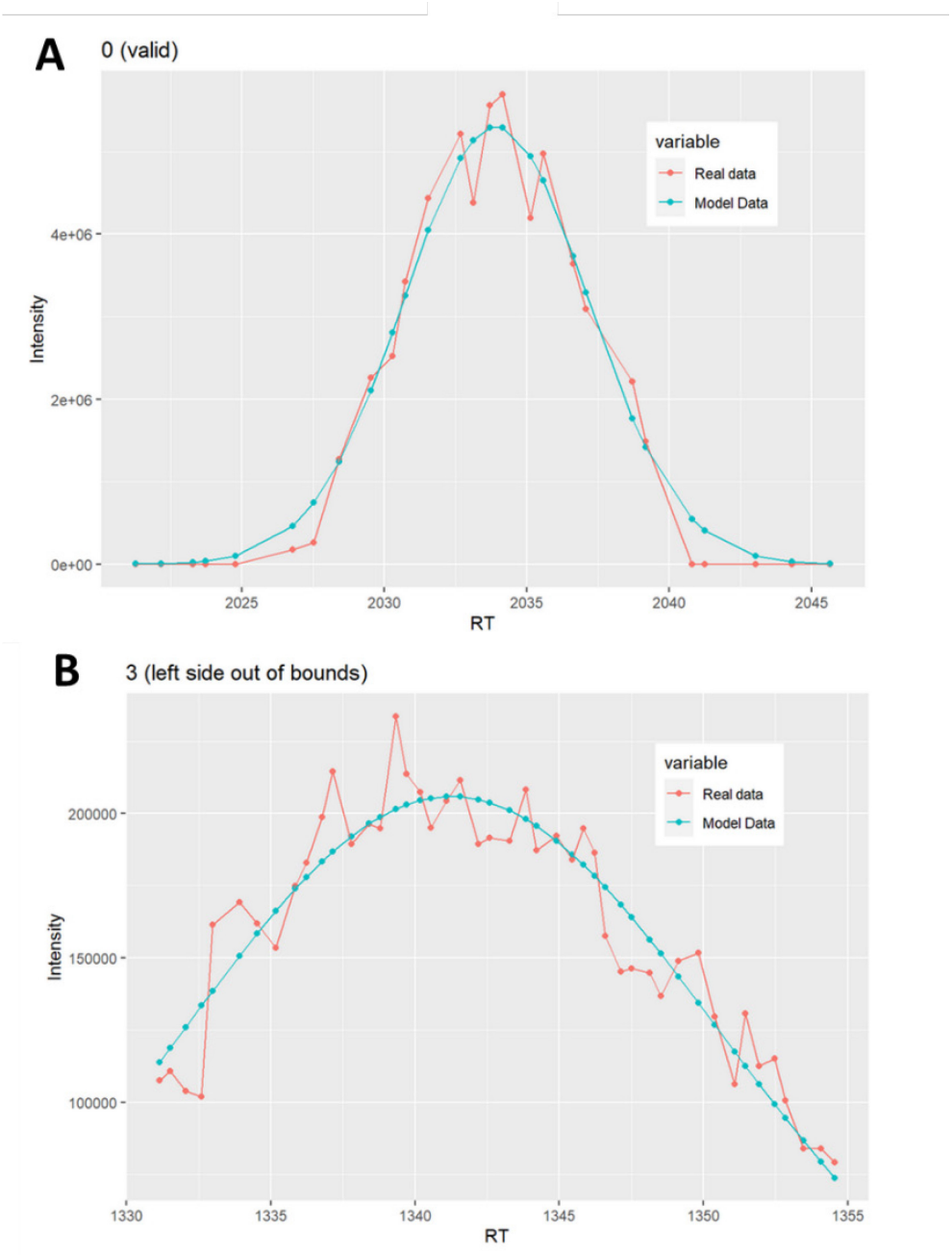


Figure 35 : Représentation de données brutes extraites (en rose) et du modèle gaussien obtenu par FFId (en bleu). Le contrôle qualité valide le modèle appliqué en A mais pas celui en B.

II. Relations entre le rapport des intensités des ions M_0 et M_1, le nombre de carbone des peptides et l'incorporation de ^{12}C

Les formules algébriques exprimant les intensités normalisées du monoisotopique M_0 et du premier isotopologue M_1, donne la possibilité de déterminer leurs valeurs théoriques. Cela permet notamment d'observer la relation entre le nombre d'atomes de carbones de la molécule, et le rapport des intensités M_1/M_0. Nous pouvons remarquer la relation stricte entre le taux de condition 12C et le coefficient directeur de la courbe. En effectuant une régression linéaire de ces observations, nous pouvons extraire le coefficient directeur de la droite obtenue. Celui-ci diminue au fur et à mesure que les mélanges sont enrichis en condition 100% SLIM 12C (Figure 36). Les mélanges sont très importants car ces conditions représentent, dans un contexte des mélanges expérimentaux de conditions marquées et non marquées, des variations d'abondances entre protéines. Ces mélanges *in-silico* sont calculés en pondérant les deux clusters extrêmes NC et 12C par leurs abondances théoriques (25%, 50%, 75%).

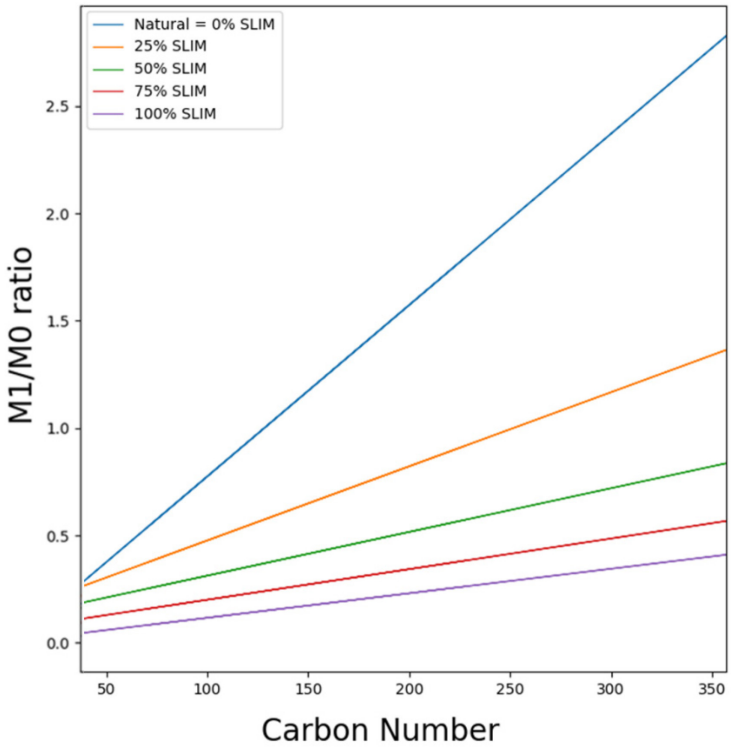


Figure 36 : Régression linéaire du rapport M_1/M_0 dans les 5 conditions en fonction du nombre de carbones.

De la même manière que nous avons pu, à partir des valeurs des intensités théoriques de M_1 et M_0 et en utilisant les équations présentées précédemment $P'(^{12}C) = \frac{nC \times M_0^{exp}}{M_1^{exp} + (nC\text{-}B) \times M_0^{exp}}$ (2-7), observer l'évolution du taux d'incorporation de ^{12}C dans les peptides (Figure 37).

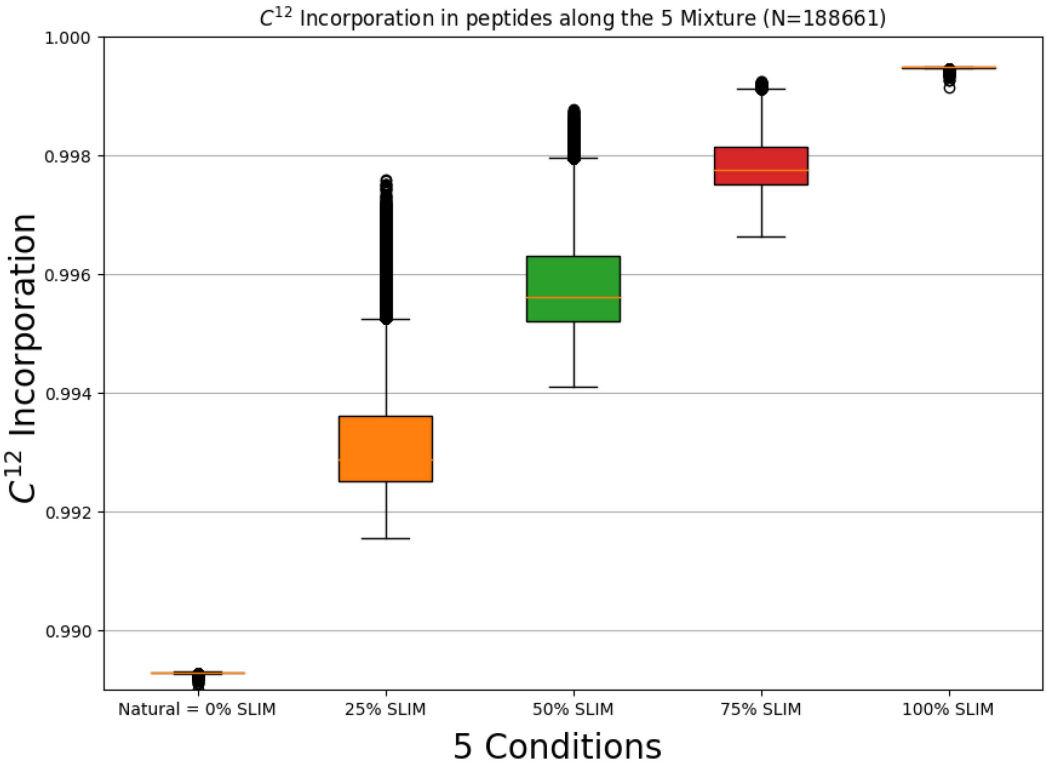


Figure 37 : Incorporation théorique du protéome tryptique de la levure S. cerevisiae.

Nous pouvons y observer en particulier la confirmation de l'hypothèse selon laquelle l'incorporation de ^{12}C n'est pas linéaire. Mais également, la confirmation que dans le cadre des mélanges, une modulation du coefficient quantitatif des clusters isotopiques théoriques prévaut à une modulation directe des valeurs d'incorporation de ^{12}C.

III. Comparaison des spectres expérimentaux et théoriques pour un peptide donné (de séquence connue)

Toujours de manière théorique, nous avons réalisé le calcul de la fraction molaire des peptides en effectuant une simulation d'une gamme de mélange peptides 100-0% bSLIM ^{12}C illustré en Figure 38. Par ailleurs les premières données expérimentales extraite avant tout filtres et amélioration ultérieur du workflow publié sont présentée dans l'histogramme de la Figure 39.

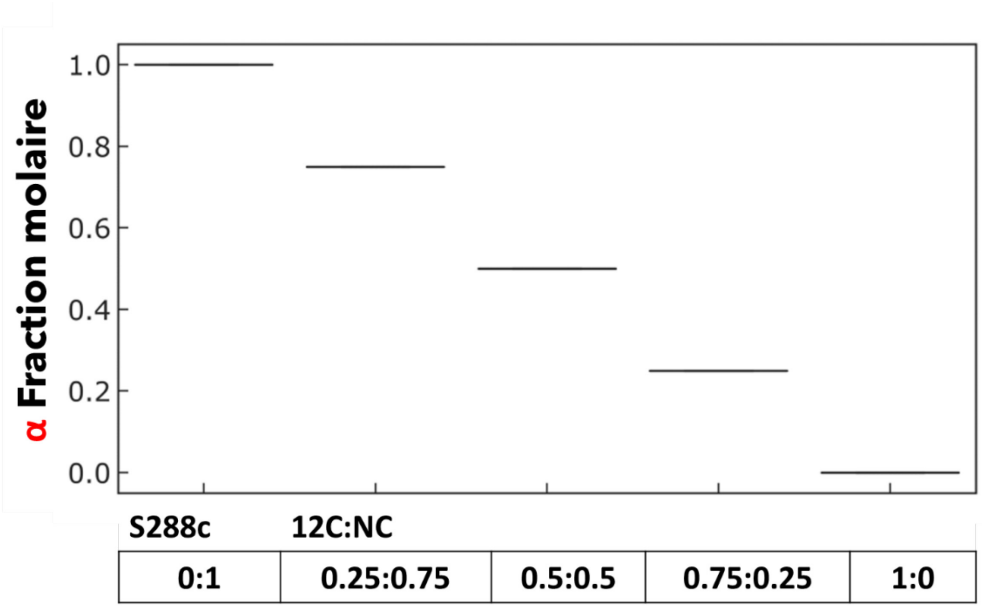


Figure 38 : Fraction molaire dans des mélanges théoriques (n=188 661)

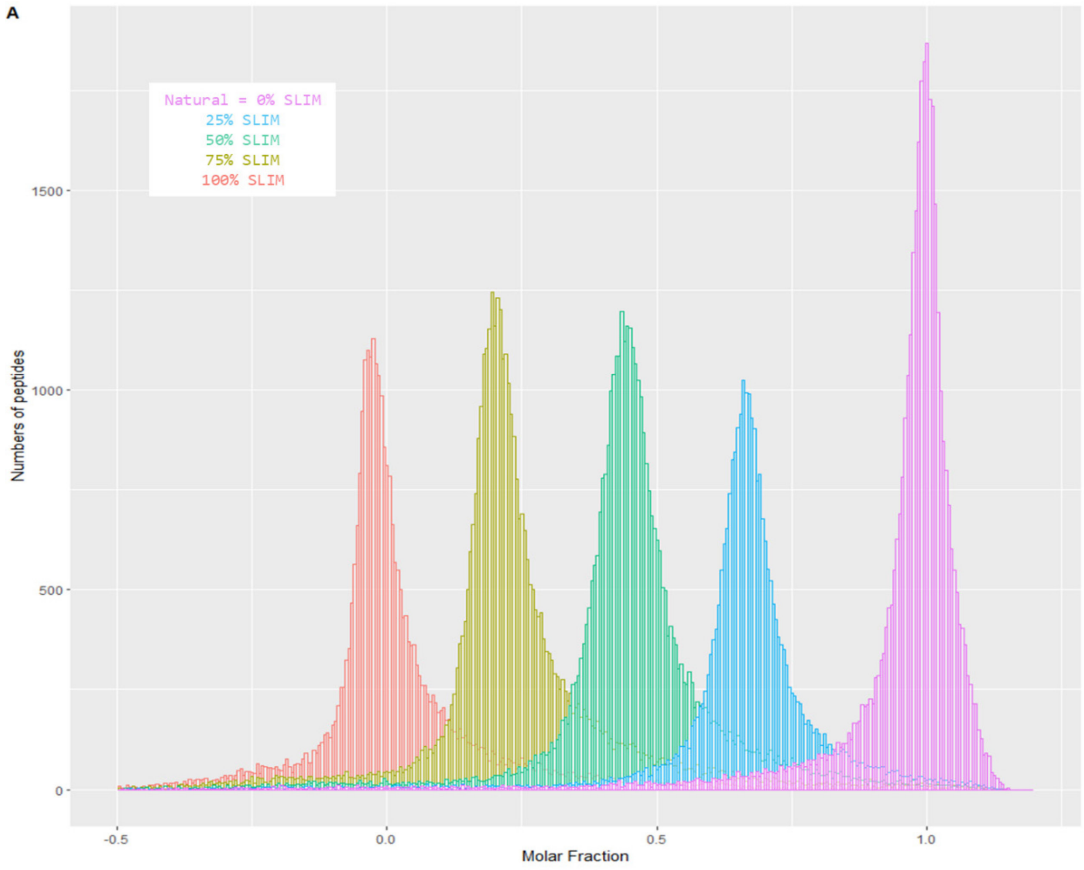


Figure 39 : Histogrammes des valeurs de la fraction molaire en fonction des mélanges.

IV. « Simulation des auxotrophies »

La résolution algébrique dans le cas d'une auxotrophie est plus complexe, et seule la simulation numérique nous a permis d'être certain de sa résolution. Pour cela, nous avons modifié l'algorithme MIDAs pour incorporer un nouvel élément chimique qui serait chimérique de l'élément carbone et possédant artificiellement une distribution isotopique identique. Nous avons choisi le Lutécium car cet élément possède une masse atomique relativement proche du carbone. Ainsi, seul son abondance isotopique (^{23}Lu -> ^{24}Lu) a été changée afin de répondre à la problématique.

Durant les simulations, les carbones entièrement marquables (C_a) sont représentés par les atomes de carbones de la formule brute, tandis que les atomes de carbone non-marquables (C_b) sont associés au Lutécium.

Ainsi un peptide aura donc pour formule chimique chimérique :

$$C_aH_wN_xO_yS_z + Lu_b$$

Ce qui permet la production d'un spectre de masse théorique proche d'un peptide expérimental dans le cas d'une auxotrophie.

V. Stratégie de calcul d'un taux de FDR

Les données d'identification obtenues par spectrométrie de masse sont par définition des probabilités de présence de certaines molécules par rapport à d'autre. Un besoin important est donc de faire la distinction entre les identifications et les quantifications réelles de celles incorrectes et aléatoires, les faux positifs. A la suite de l'identification par un logiciel de recherche en banque de données, un score de confiance est associé à chaque identification. Toutefois, évaluer la robustesse des valeurs quantitatives est nécessaire. Cela est rendu possible en utilisant une méthode discriminative par un test de fiabilité, à partir d'un calcul du taux de faux positivité (FDR). Dans la communauté scientifique il est généralement admis qu'un taux de FDR supérieur à 5% est le signe d'un résultat peu fiable. Mais ce seuil peut être discuté car il faut l'adapter à chaque expérience, qui sont chacune uniques (Eidhammer et al., 2013).

Classiquement, la méthodologie d'étude de la pertinence des résultats par un calcul de FDR, utilise un modèle dans lequel toutes les fluctuations des valeurs quantitatives seraient aléatoires. De cette manière, les protéines variant de façon plus « significative » au regard de ce modèle aléatoire, auront plus de chance d'être fiables, c'est à dire associées à une confiance plus élevée. Travailler avec des fluctuations aléatoires est donc une stratégie classique de calcul de critères de confiance statistique. En bSLIM les distributions de probabilités ne sont pas connues. C'est pourquoi nous avons développé une stratégie statistique, robuste et descriptive pour cette nouvelle méthode de quantification. L'intégralité des données de quantification se fonde sur l'extraction de signaux bruts, l'outil statistique devra donc être basé sur une étude comparable de traitement du signal.

Nous avons ainsi développé une stratégie de calcul des FDR, directement inspirée de la méthode SAM (Significance Analysis of Microarrays (Tusher, Tibshirani, & Chu, 2001)). SAM est une méthode ancienne d'évaluation graphique de la significativité des scores de quantification de l'expression des gènes, développée pour les analyses de données issues des puces à ADN. Elle se fonde sur le calcul d'un log(ratio) utilisé comme mesure de la quantification des transcrits. Cela est analogue aux mesures obtenues avec la méthode bSLIM, où un log(ratio) de la fraction molaire des peptides non-marqués (via une mesure d'intensité) est utilisé comme valeur de quantification. De plus, cette méthode est facilement implémentable, et repose par la procédure suivante : dans un premier temps, un score est calculé pour chaque donnée expérimentale, puis sont triés par ordre croissant selon leurs valeurs (Figure 40 A). Dans un deuxième temps, les données expérimentales brutes sont permutées de façon aléatoire et les « nouvelles valeurs » sont conservées dans une matrice. Un score sera calculé et enfin utilisé pour réaliser un tri par ordre croissant (Figure 40 B). La permutation aléatoire de l'ensemble des données génère un bruit de fond (empirique) qui permet une analyse des fluctuations observées au-dessus des fluctuations aléatoires. Ces scores sont comparés graphiquement aux "scores réels" obtenus à partir des données originales (Figure 40 C). Les protéines pour lesquelles les scores réels diffèrent le plus des scores aléatoires sont par définition celles qui sont les plus "statistiquement significatives".

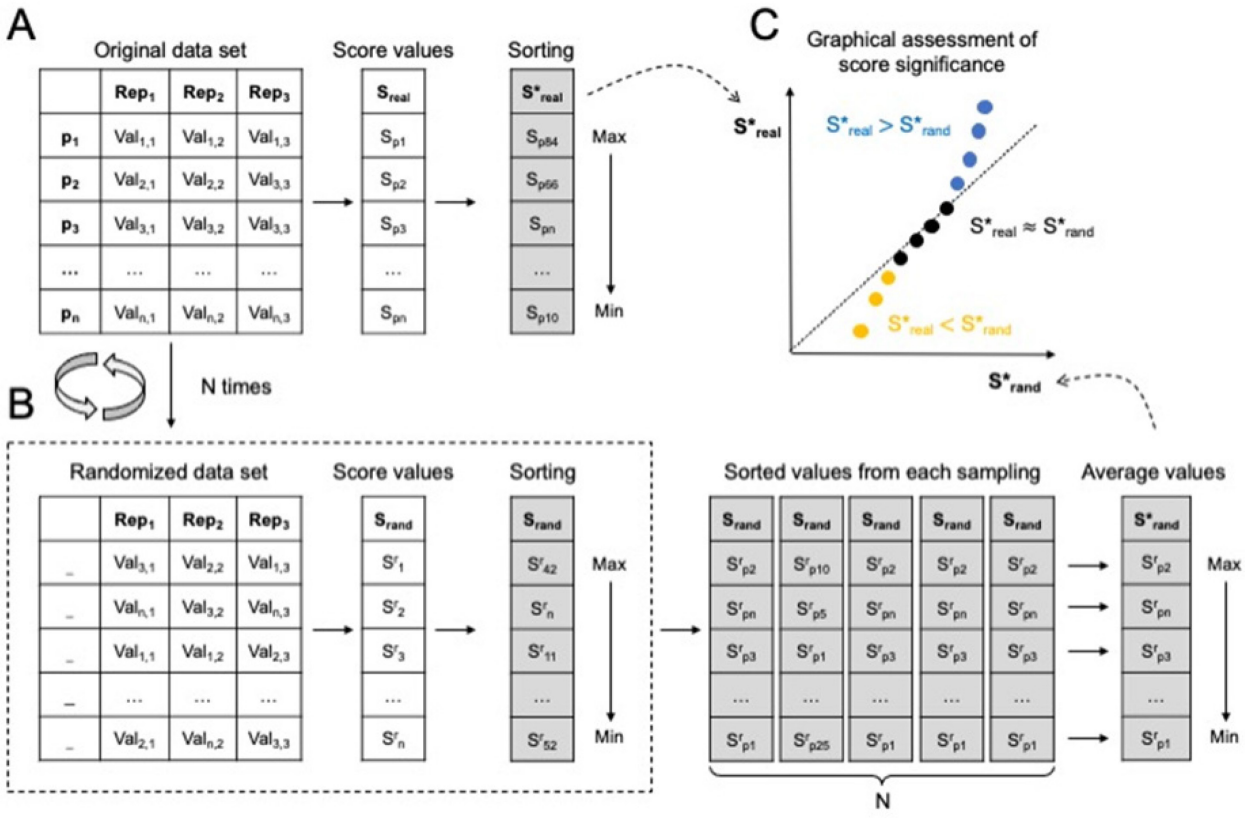


Figure 40 : Représentation du modèle statistique développé. Figure extraite de la publication (Sénécaut et al., 2022)

VI. Généralisation du workflow à d'autres logiciels (Minora)

Si de nombreux logiciels de traitement des données issues de spectrométrie de masse existent, la communauté des protéomiciens reste influencée par les solutions informatiques fournies par les constructeurs des appareils. En particulier, les laboratoires munis de spectromètres utilisant la technologie Orbitrap sont incités à traiter leurs données sur la plateforme « Proteome Discoverer » développée par *Thermo Scientific*. C'est la raison pour laquelle, dans le but d'étendre l'utilisation de la méthode de quantification bSLIM, nous avons implémenté un workflow qui permet d'utiliser le format de fichier propriétaire de Thermo (pdresult). Ces fichiers sont le résultat d'interrogations effectuées dans le cadre d'un workflow *Proteome Discover*. Ce fichier contient les identifications et l'extraction brute de l'intensité de chacun des isotopologues associés en utilisant l'algorithme Minora (Horn et al., 2016). Ici, les données de ce logiciel seront utilisées comme des valeurs d'entrée des quantifications

pour la méthode bSLIM. A cet effet nous avons mis au point une procédure impliquant des requêtes SQL conditionnées dans un script R, implémentable dans un workflow KNIME. Dans ce script et à l'issue des requêtes, diverses procédures permettent d'ordonner les données afin d'extraire les valeurs essentielles à la suite workflow de quantification comme l'aire des isotopologues, la masse des ions et leurs identifications associées.

L'ensemble des workflows présentés est disponible sur la plateforme Zenodo[1], un dépôt public développé par le CERN et permettant l'archivage et le référencement pour toute la communauté scientifique par la délivrance d'un identifiant numérique « DOI ».

Workflow	Lien DOI
bSLIM Quantification FFId KNIME 4.2.3	https://doi.org/10.5281/zenodo.4467812
bSLIM Identification FFId KNIME 4.2.3	https://doi.org/10.5281/zenodo.4467789
bSLIM Quantification Minora KNIME 4.2.3	https://doi.org/10.5281/zenodo.4467829
bSLIM Scoring SAM	https://doi.org/10.5281/zenodo.4467882

[1] https://zenodo.org/

Chapitre 3 : Données réelles

I. Calculs a posteriori de la probabilité de l'isotope ^{12}C

Bien que les études portant sur la pathogénicité de la levure *Candida albicans* soient importantes dans le domaine de la santé publique, la communauté des biologistes qui travaillent sur ce sujet est plus restreinte. La présentation et la démonstration de l'intérêt de notre méthode aurait dans ce contexte un impact limité. Notre méthode vise à être appliquée par des laboratoires de biochimie, en lien étroit avec une plateforme de spectrométrie de masse. C'est pourquoi nous avons choisi de changer de modèle d'étude (par rapport au choix qui avait été réalisé pour le SLIM) et de travailler avec la levure *Saccharomyces cerevisiae,* organisme modèle par excellence, très utilisé au laboratoire. Toutefois, des applications concrètes sont toujours en cours de développement afin de répondre à des questions biologiques en suspend pour C. *albicans*.

Ainsi, j'ai effectué des cultures cellulaires et des préparations d'extraits protéiques de *S. cerevisiae*. De ce fait, plusieurs nouvelles acquisitions en spectrométrie de masse et en particulier une gamme complète de mélange peptides ont été réalisées. Par ailleurs, une quantification bSLIM a été effectuée par une analyse de mélange de deux souches de levures différentes dans le but de répondre à plusieurs objectifs.

Tout d'abord, une étude de la sensibilité et des capacités résolutives de la méthode a été réalisée, en étudiant les variations biologiques entre les protéomes des deux souches de *S. cerevisiae* : la souche de référence S288c (MATα SUC2 gal2 mal2 mel flo1 flo8-1 hap1 ho bio1 bio6) et la souche BY4742 (MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0), une souche de laboratoire dérivée de S288c. Dans un deuxième temps, nous avons réalisé une étude de robustesse et d'ajustement de la méthode bSLIM en étudiant l'incorporation du ^{12}C au sein des protéines dans des mélanges contenant des conditions "naturelles" et "marquées" en proportion variable. Enfin, nous avons analysé l'incorporation de ^{12}C dans le protéome de la souche BY4742 qui est une souche auxotrophique pour la Leucine, la Lysine, l'Histidine et l'Uracile. Cela induit un biais

partiel, nécessitant l'optimisation des méthodes computationnelles de traitement des données (voir III).

Les protocoles de mise en culture sont disponibles dans la publication issue de cette thèse (ci-après).

Les résultats nous ont permis d'observer l'incorporation de ^{12}C dans les peptides expérimentaux produits. Comme la souche BY4742 présente une auxotrophie pour L, H, K, la théorie prédit qu'en moyenne 20% des acides aminés d'un peptide composé de L, H, K ne seront pas marqués. Par conséquent les carbones auront tendance à garder une composition isotopique naturelle, ce qui induit une dilution du signal isotopique 100% C12. Cependant, nos calculs utilisent la notion de carbones B, le nombre d'atomes de carbones non-marquables issus des acides aminés exogènes, permettant d'obtenir une valeur corrigée d'incorporation du ^{12}C (Figure 41).

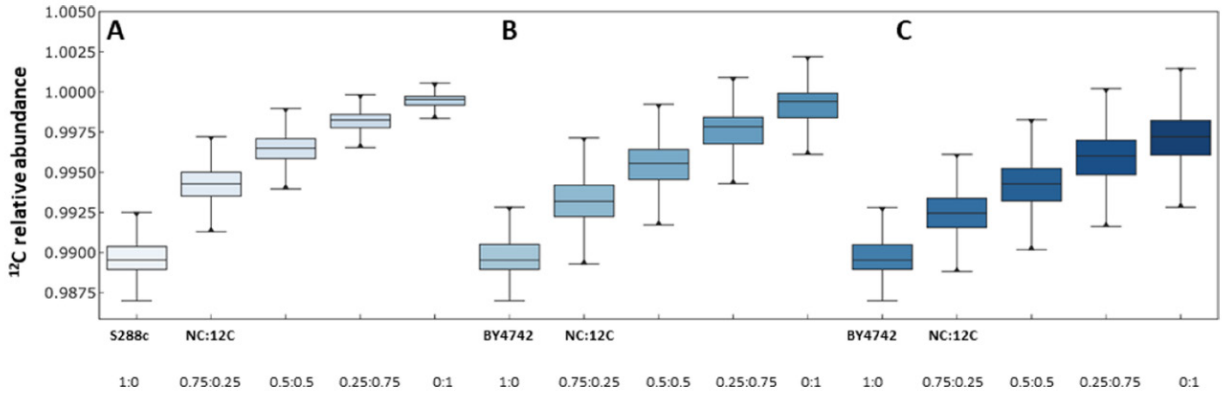


Figure 41 : Incorporation du ^{12}C dans les peptides expérimentaux, A dans la souche sauvage S288C, B dans la souche auxotrophe BY4247, et C la valeur de l'incorporation totale.

II. Utilisation de la probabilité de l'isotope ^{12}C entre deux échantillons pour la protéomique quantitative

Afin de tester la sensibilité et la finesse de notre méthode de quantification, nous avons étudié les variations des abondances protéiques entre les deux souches de *S. cerevisiae*. La première souche décrite comme sauvage S288c, et l'autre souche « mutante » BY4742 (Ura3, Leu2, Lys2, His3). Après culture et extraction des

protéines, un mélange 1:1 de même quantité de protéines des deux conditions est tout d'abord digéré (par exemple 40µg = 20µg S288c /20µg BY4742).

Nous avons produit huit échantillons issus de deux réplicas biologiques (échantillons différents mais provenant de la même condition) et de deux réplicas techniques (même échantillons mais deuxième injection dans le spectromètre de masse). Par ailleurs, nous avons procédé à un *label-swap*, c'est-à-dire que dans un cas le mutant était cultivé en condition normale et la souche sauvage en condition marquée, et dans l'autre cas les deux conditions ont été inversées. Une fois les valeurs expérimentales et quantitatives déterminées, nous avons observé les profils d'expression des protéines en détail.

II.1 Validation de la robustesse de la méthode

Notre analyse a permis d'identifier 4162 protéines dans les huit échantillons, cependant la quantification n'a été réalisée que sur les 3354 protéines identifiées par au moins huit peptides.

Les valeurs quantitatives mesurées pour les protéines liées à l'auxotrophie, ont clairement démontré la seule contribution de la souche S288c. En effet, ces gènes avaient été délétés chez le mutant. Concrètement, les protéines exprimées par les gènes KO sont quantifiées de manière surexprimées avec un $\log_2$(ratio), allant de 3 jusqu'à 6 pour la souche S288c. Cela démontre que seule la souche sauvage possède les protéines associées aux gènes délétés. Cependant, la protéine His3 peu abondante dans les cellules de type sauvage (Ho et al., 2018) n'est pas facile à détecter et à quantifier. A titre d'exemple, cette protéine présente une valeur quantitative expérimentale de 12.3 ppm dans la banque d'expression protéique PaXDB, démontrant ainsi qu'elle est peu abondante aux regards des autres protéines contenues dans la levure. Nous avons donc une validation expérimentale que notre méthode quantitative est robuste puisque les protéines cibles montrent des profils d'expression attendus.

II.2 Validation de la sensibilité de la méthode

Nous avons utilisé l'algorithme Autoclass, un outil de classification Bayésien non-supervisé (Achcar et al., 2009), afin de rassembler les protéines ayant des profils

d'expression similaires. Parmi les différentes classes, celles ayant des profils très typiques d'une surexpression dans une condition ou dans une autre ont été minutieusement analysées.

Cette méthode est également sensible puisque nous avons pu observer des variations biologiques d'une grande finesse s'expliquant par les *variations* entre les deux souches isogéniques. Notamment nous avons pu observer et quantifier au sein de mêmes classes sur-exprimées dans la souche S288c, un grand nombre de protéines codées par des gènes impliqués dans des voies métaboliques incluant la respiration. Particulièrement dans une classe typique, les marqueurs d'autotrophie sont présents, signe que la méthode de quantification est correcte.

Deux classes contiennent 13 et 22 protéines sur-exprimées (3 et 2 taux d'expression) dans la condition S288c. Les protéines ainsi exprimées sont impliquées dans les voies métaboliques de la biosynthèse du glutamate (*Gdh1, Gdh3, Gln1, Glt1*). Or l'expression du gène *Gln1* est connue pour être régulée par l'apport d'azote et les ressources disponibles en acides aminés. Comme tous les acides aminés chez la souche sauvage sont biosynthétisés par le catabolisme du glucose, ce résultat suggère que leur synthèse peut permettre la vitalité et le développement mais en quantité assez limitée pour activer des voies métaboliques spécifiques en réponse à un manque de nutriment. Par ailleurs, un nombre important de gènes exprimant des protéines impliquées dans des fonctions mitochondriales clés sont présents (e.g., *Cox5A, Cox9, Cox12, Cox13, Cox2, Tim11, Aco1, Aco2*, et d'autre sous-unités de la F0–F1 ATPases). Cela est en accord avec les données publiées par (Young & Court, 2008) ainsi que (Dimitrov et al., 2009), démontrant l'importance du fond génétique, dans la stabilité et la conservation du maintien de l'ADN mitochondrial et dans les fonctions essentielles de la respiration.

De manière surprenante, bien que la souche de référence S288c ait été sélectionnée pour son auxotrophie à la biotine (mutation *bio1* et *bio6*), d'autres gènes impliqués dans la voie métabolique de biosynthèse de la biotine sont sous-exprimés chez BY4742 (Bio4, Bio2, Bio3). Cette voie permet la biosynthèse de la biotine, à partir de *7-keto-8-aminopelargonate*, et sert de précurseur en dérivant du métabolisme de l'alanine.

De la même manière, nous avons observé une troisième classe contenant 50 gènes codant pour des protéines sous-exprimées chez S288c (par un facteur 2 dans le

taux d'expression). Un grand nombre des protéines sur-exprimées chez le mutant BY4742 sont impliquées dans le métabolisme des acides aminés exogènes ajoutés au milieu de culture pour répondre à l'auxotrophie. Notamment la mutation du gène Leu2, dont la protéine exprimée est impliquée dans la dernière étape de biosynthèse de la leucine (*3-Isopropylmalate* en *(2S)-2-Isopropyl-3-oxosuccinate*) est également impliqué dans le métabolisme de la valine et de l'isoleucine comme molécule précurseur de l'*2- oxobutanoate* .

En réponse à ces mutations, une compensation des voies métaboliques impactées a été observée dans la souche BY4742. L'analyse des protéines impliquées dans la biosynthèse de la lysine montre une sur-expression dans le s288c s'expliquant par l'absence de Lys5, une protéine cible de Lys2. Cela est notamment révélé par le fait que les groupes de gènes *Ilv2,3,5* (biosynthèse de la valine), *Leu1,4,5,9* (biosynthèse de la Leucine) et *Bat1,2* (étapes finales du métabolisme de biosynthèse de la leucine et de la valine) sont tous sur-exprimés chez BY4742. Une observation similaire est faite pour la voie métabolique de l'histidine avec une suppression de *His1, 4, 7* et *His2*. Toutefois exprimées de manière moindre que celles vues précédemment (Leucine par exemple) en accord avec la littérature. Une grande partie des protéines impliquées dans la biosynthèse de l'arginine (*Arg4, 8, 5, 6, 1, 7*) ont été quantifiées comme plus abondantes dans la condition BY4742. Cela est dû au fait que la voie métabolique de biosynthèse de l'arginine est affectée par l'altération des fonctions mitochondriales.

Enfin, comme attendu, un grand nombre de protéines se caractérisent par des profils d'expression identiques entre les deux souches. La souche de levure S288c (dite sauvage) est issue de culture et d'optimisation en boulangerie. Elle est donc issue de nombreux croisement intra souche afin d'obtenir la meilleure souche dédiée à une fermentation idéale au détriment de son bien-être métabolique. Cette souche a donc un fond instable au niveau métabolique.

Les données brutes du spectromètre de masse obtenues durant cette thèse, d'interrogation et de quantification, ont été rendues publiques pour la communauté

scientifique sur le site « Proteomics Identifications Database » (PRIDE[1]) sous le nom de projet PXD021329.

Ces travaux ont été publiée en 2021 dans Journal of Proteome Research.

[1] https://www.ebi.ac.uk/pride/

Journal of proteome research

pubs.acs.org/jpr

Article

Novel Insights into Quantitative Proteomics from an Innovative Bottom-Up Simple Light Isotope Metabolic (bSLIM) Labeling Data Processing Strategy

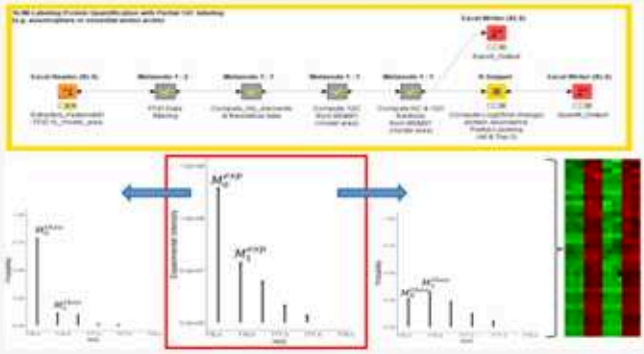
Nicolas Sénécaut, Gelio Alves, Hendrik Weisser, Laurent Lignières, Samuel Terrier, Lilian Yang-Crosson, Pierre Poulain, Gaëlle Lelandais, Yi-Kuo Yu, and Jean-Michel Camadro*

Cite This: J. Proteome Res. 2021, 20, 1476−1487

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Simple light isotope metabolic labeling (SLIM labeling) is an innovative method to quantify variations in the proteome based on an original *in vivo* labeling strategy. Heterotrophic cells grown in U-[^{12}C] as the sole source of carbon synthesize U-[^{12}C]-amino acids, which are incorporated into proteins, giving rise to U-[^{12}C]-proteins. This results in a large increase in the intensity of the monoisotope ion of peptides and proteins, thus allowing higher identification scores and protein sequence coverage in mass spectrometry experiments. This method, initially developed for signal processing and quantification of the incorporation rate of ^{12}C into peptides, was based on a multistep process that was difficult to implement for many laboratories. To overcome these limitations, we developed a new theoretical background to analyze bottom-up proteomics data using SLIM-labeling (bSLIM) and established simple procedures based on open-source software, using dedicated OpenMS modules, and embedded R scripts to process the bSLIM experimental data. These new tools allow computation of both the ^{12}C abundance in peptides to follow the kinetics of protein labeling and the molar fraction of unlabeled and ^{12}C-labeled peptides in multiplexing experiments to determine the relative abundance of proteins extracted under different biological conditions. They also make it possible to consider incomplete ^{12}C labeling, such as that observed in cells with nutritional requirements for nonlabeled amino acids. These tools were validated on an experimental dataset produced using various yeast strains of *Saccharomyces cerevisiae* and growth conditions. The workflows are built on the implementation of appropriate calculation modules in a KNIME working environment. These new integrated tools provide a convenient framework for the wider use of the SLIM-labeling strategy.



KEYWORDS: *In vivo metabolic labeling, light carbon isotope, ^{12}C, quantitative proteomics, data processing workflow, OpenMS, KNIME, yeast*

INTRODUCTION

We recently developed a new method to quantify protein half-lives *in vivo* based on an original labeling strategy, simple light isotope metabolic labeling (SLIM labeling),[1] derived from the pioneering work of the group of Marshall.[2−4] Briefly, we grew yeast cells on a synthetic medium containing glucose as the sole source of carbon (referred to as the normal carbon condition (NC): 98.93% ^{12}C, 1.07% ^{13}C) and shifted the cells to the same medium containing U-[^{12}C]-glucose as the sole source of carbon (referred to as the ^{12}C condition: 100% ^{12}C). This allowed us to follow the incorporation of the ^{12}C-amino acids into proteins rapidly synthesized by the cells over time. SLIM labeling is, with the whole-organism heavy water labeling methodology,[5,6] one of the few methods that allows an integral labeling of the cellular components, without bias in functional group requirement or amino-acid composition. The SLIM-labeling strategy provides several advantages for MS-based proteomics developments. The isotope cluster for peptides remains strictly bound to the natural monoisotopic ion, leading to a large increase in the intensity of the monoisotopic ion of peptides and proteins. This allows for better precision in the monoisotopic mass measurement, with higher identification scores and protein sequence coverage[1] in mass spectrometry-based experiments. This method, initially developed for signal processing and quantifying the incorporation rate of ^{12}C into peptides, was based on a multistep process

Received: June 29, 2020
Published: February 11, 2021

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

that was difficult to implement for many laboratories. It involved (1) extracting the intensities of isotopologues using the commercial software Progenesis QI for metabolomics, (2) aligning the "Features" files with the peptide identification files (Mascot files), (3) calculating the ratio of the monoisotope intensity to the intensity of all isotopologues (called R_{iso}), (4) establishing the equation of the theoretical R_{iso} values for each peptide sequence, calculated using the MIDAs application[7] as a function of ^{12}C enrichment, after determining the elemental composition of each peptide and therefore the exact number of carbons it contains, and (5) calculating the abundance of ^{12}C in each experimental peptide by fitting its experimental R_{iso} to the closest theoretical R_{iso}. The data were then filtered to eliminate outliers based on monoisotopic intensity, peptide size, and ^{12}C composition derived from the adjustment. Protein abundance was calculated on the basis of the abundance of their three most intense peptides (Top-3). The absence of an automated processing tool contributed to limiting the widespread diffusion of the method.

To overcome these limitations, we established the basis of a new theoretical background to process the experimental data in SLIM-labeling experiments and developed the appropriate procedures based on open-source resources, using, in particular, dedicated OpenMS modules[8] for peptide identification in conjunction with a modified version of FeatureFinderIdentification.[9] This solution provides a Gaussian fit of the intensities of every isotopologue chromatographic trace (mass trace) for peptides with a validated identification, allowing extraction of the relevant information. The only experimental data required for high-quality quantification are the abundance of M_0 and M_1, the monoisotopic ion, and the +1 isotopologue of the peptides identified with high confidence. Computation of the ^{12}C abundance of peptides and the molar fraction of peptides from the NC and ^{12}C conditions in multiplexing experiments is performed by implementing appropriate calculation modules in a KNIME working environment.[10,11] We also present the theoretical basis for establishing appropriate filters to obtain high-quality processed data from experimental datasets. These new integrated tools provide a convenient framework that enables wider use of the bSLIM-labeling strategy.

Theoretical Basis of bSLIM-Labeling Data Processing

The isotope cluster of every peptide Pept is defined as a series of isotopologues (M_0, M_1, M_2, ... M_n). Considering a peptide of elemental composition: $C_x H_y N_v O_w S_u$, the isotope cluster can be described as the monoisotopic ion associated with a distribution of neutrons among the various elements present in the molecule. The intensities of the isotopologues can be estimated by the enumeration of the isotopic contributions of each element. In proteins, the elements are (^{12}C, ^{13}C), (^{1}H, ^{2}H), (^{14}N, ^{15}N), (^{16}O, ^{17}O, ^{18}O), and (^{32}S, ^{33}S, ^{34}S, ^{36}S). Phosphorous is present only as the stable isotope ^{31}P. A classical approach combines polynomial distributions for all of the atoms in a molecule to take into account all of the possible permutations of neutrons of additional origin.[12]

Accordingly, the normalized intensity of the monoisotopic ion M_0 is given by

$$M_0 = \prod_{\substack{i=x,y,v,w,u \\ X=C,H,N,O,S}} P(^{A}X)^{i} \quad (1)$$

where $P(^{A}X)$ is the probability of occurrence, expressed as relative abundance, of the element X of mass number A. The normalized intensity of M_1 is given by the polynomial expansion to the ^{13}C, ^{2}H, ^{15}N, ^{17}O, and ^{33}S terms of the distribution (elements containing one extra neutron)

$$M_1 = \sum_{\substack{X=C,H,N,O,S \\ i=x,y,v,w,u}} \left[i \times P(^{A}X)^{i-1} \times P(^{A+1}X) \times \prod_{\substack{i=x,y,v,w,u \\ Y\neq X=C,H,N,O,S}} P(^{A}Y)^{i} \right] \quad (2)$$

Although formal expressions for higher-order isotopologues have been produced,[13] the present method uses only the M_0 and M_1 intensities. The developed expressions of M_0 and M_1 are provided in Supporting File S1.

Under standard conditions,[14] $P(^{12}C) = 0.9893$; $P(^{13}C) = 0.0107$; $P(^{1}H) = 0.999885$; $P(^{2}H) = 0.000115$; $P(^{14}N) = 0.99632$; $P(^{15}N) = 0.00368$; $P(^{16}O) = 0.99757$; $P(^{17}O) = 0.00038$; $P(^{18}O) = 0.00205$; $P(^{32}S) = 0.9493$; $P(^{33}S) = 0.0076$; $P(^{34}S) = 0.0429$; $P(^{36}S) = 0.0002$.

The quantification of bSLIM labeling allows determination of the amount of ^{12}C present in every peptide in experiments in which the carbon isotope composition is manipulated to favor ^{12}C incorporation. This amount is expressed as $P'(^{12}C)$ and the residual amount of ^{13}C as $P'(^{13}C)$.

For every peptide, it is possible to have experimental access to at least M_0^{exp} and M_1^{exp}. Equations 1 and 2 provide the theoretical values of M_0 and M_1, with the sum of the intensity (probability) of all of the isotopologues normalized to 1.

It is therefore possible to set the equivalence

$$M_0 = \frac{M_0^{exp}}{\sum_0^n M_i^{exp}} \text{ and } M_1 = \frac{M_1^{exp}}{\sum_0^n M_i^{exp}} => \frac{M_1}{M_0} = \frac{M_1^{exp}}{M_0^{exp}}$$

Equations 1 and 2 are rearranged to set

$$\frac{M_1^{exp}}{M_0^{exp}} = x \times \frac{P'(^{13}C)}{P'(^{12}C)} + y \times \frac{P(^{2}H)}{P(^{1}H)} + v \times \frac{P(^{15}N)}{P(^{14}N)} + w \times \frac{P(^{17}O)}{P(^{16}O)} + u \times \frac{P(^{33}S)}{P(^{32}S)} \quad (3)$$

Let us define B as the sum of the terms that remain identical under the NC and ^{12}C conditions

$$B = y \times \frac{P(^{2}H)}{P(^{1}H)} + v \times \frac{P(^{15}N)}{P(^{14}N)} + w \times \frac{P(^{17}O)}{P(^{16}O)} + u \times \frac{P(^{33}S)}{P(^{32}S)}$$

Therefore

$$\frac{M_1^{exp}}{M_0^{exp}} = x \times \frac{P'(^{13}C)}{P'(^{12}C)} + B$$

since

$$P'(^{13}C) = 1 - P'(^{12}C)$$

then

$$\frac{M_1^{exp}}{M_0^{exp}} = x \times \frac{[1-P'(^{12}C)]}{P'(^{12}C)} + B$$

and $P'(^{12}C)$ is obtained as

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

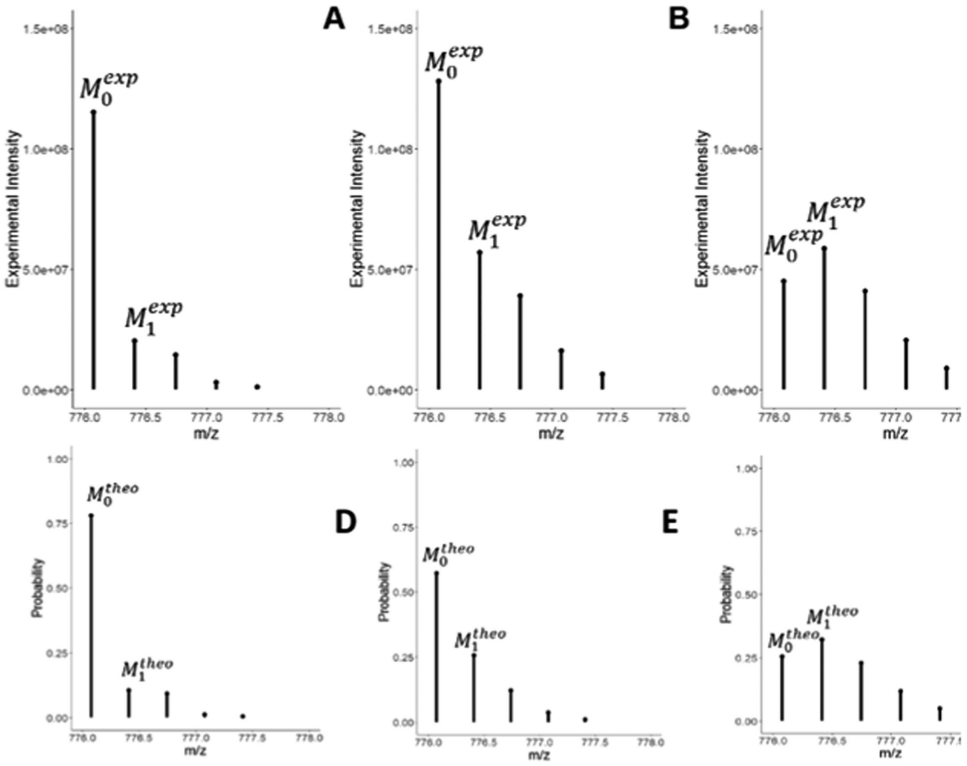


Figure 1. Experimental spectra (top) of the peptide AGITVMNEIGLDPGIDHLYAVK (3+) from the protein YNR050c and the corresponding theoretical spectra (bottom), under 12C (A and D) and NC (C and F) conditions, and from the same peptide from a 1:1 mixture of total extract from the 12C and NC conditions (yeast strain S288c) (B and E).

$$P'(^{12}\mathrm{C}) = \frac{x \times M_0^{\mathrm{exp}}}{M_1^{\mathrm{exp}} + (x - B) \times M_0^{\mathrm{exp}}} \tag{4}$$

This equation makes it possible, for example, to follow the ^{12}C incorporation rate in the kinetics of SLIM labeling *in vivo*.

Quantitative Analysis of SLIM Labeling

When mixing given amounts of tryptic digests from the NC and 12C conditions, each peptide amount is the sum of the molar fraction of the NC peptide and the 12C peptide present in the initial samples.

Let us define the molar fractions of NC peptide (α) and the 12C peptide $(1 - \alpha)$

$$\mathrm{Pept}^{\mathrm{exp}} = \alpha \mathrm{Pept}^{\mathrm{NC}} + (1 - \alpha)\mathrm{Pept}^{12\mathrm{C}} \tag{5}$$

The isotope cluster of $\mathrm{Pept}^{\mathrm{exp}}$ is defined as (M_0^{exp}, M_1^{exp}, M_2^{exp}, ... M_n^{exp}), while those of $\mathrm{Pept}^{\mathrm{NC}}$ and $\mathrm{Pept}^{^{12}\mathrm{C}}$ are defined as (M_0^{NC}, M_1^{NC}, M_2^{NC}, ... M_n^{NC}) and ($M_0^{^{12}\mathrm{C}}$, $M_1^{^{12}\mathrm{C}}$, $M_2^{^{12}\mathrm{C}}$, ... $M_n^{^{12}\mathrm{C}}$), respectively, as illustrated in Figure 1.

Considering the normalized expression of M_0 and M_1, we may write the following equalities

$$\frac{M_0^{\mathrm{exp}}}{\sum_0^n M_i^{\mathrm{exp}}} = \alpha M_0^{\mathrm{NC}} + (1 - \alpha)M_0^{12\mathrm{C}}$$

and

$$\frac{M_1^{\mathrm{exp}}}{\sum_0^n M_i^{\mathrm{exp}}} = \alpha M_1^{\mathrm{NC}} + (1 - \alpha)M_1^{12\mathrm{C}}$$

Therefore

$$\frac{M_1^{\mathrm{exp}}}{M_0^{\mathrm{exp}}} = \frac{\alpha M_1^{\mathrm{NC}} + (1 - \alpha)M_1^{12\mathrm{C}}}{\alpha M_0^{\mathrm{NC}} + (1 - \alpha)M_0^{12\mathrm{C}}} = \frac{\alpha(M_1^{\mathrm{NC}} - M_1^{12\mathrm{C}}) + M_1^{12\mathrm{C}}}{\alpha(M_0^{\mathrm{NC}} - M_0^{12\mathrm{C}}) + M_0^{^{12}\mathrm{C}}}$$

The molar fraction α of nonlabeled peptide in the mixture is therefore

$$\alpha = \frac{M_0^{12\mathrm{C}}M_1^{\mathrm{exp}} - M_0^{\mathrm{exp}}M_1^{12\mathrm{C}}}{M_0^{\mathrm{exp}}M_1^{\mathrm{NC}} - M_0^{\mathrm{exp}}M_1^{12\mathrm{C}} - M_1^{\mathrm{exp}}M_0^{\mathrm{NC}} + M_1^{\mathrm{exp}}M_0^{12\mathrm{C}}} \tag{6}$$

Knowing α allows us to compute the molar fraction of ^{12}C-labeled peptide $(1 - \alpha)$, and the ratio

$$R = \frac{\mathrm{NC}}{12\mathrm{C}} = \frac{\alpha}{(1 - \alpha)}$$

which can be classically taken as a fold change in omics studies and analyzed as $\log 2(R)$.

Cases of Incomplete SLIM Labeling

The ideal situation described above applies only for heterotrophic organisms for which all of the amino acids can be synthesized from a single U-[^{12}C]-carbon source, such as U-[^{12}C]-glucose, U-[^{12}C]-glycerol, or U-[^{12}C]-acetate, such as wild-type prokaryotic or eukaryotic microorganisms. However, many laboratory strains and cells of higher eukaryotes either carry mutations in genes required for the synthesis of selected amino acids or rely on exogenous essential amino acids for their growth. These amino acids must be added to the growth media and are currently not available as U-[^{12}C]-amino acids. They will be incorporated into proteins in the NC form only. Therefore, to

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

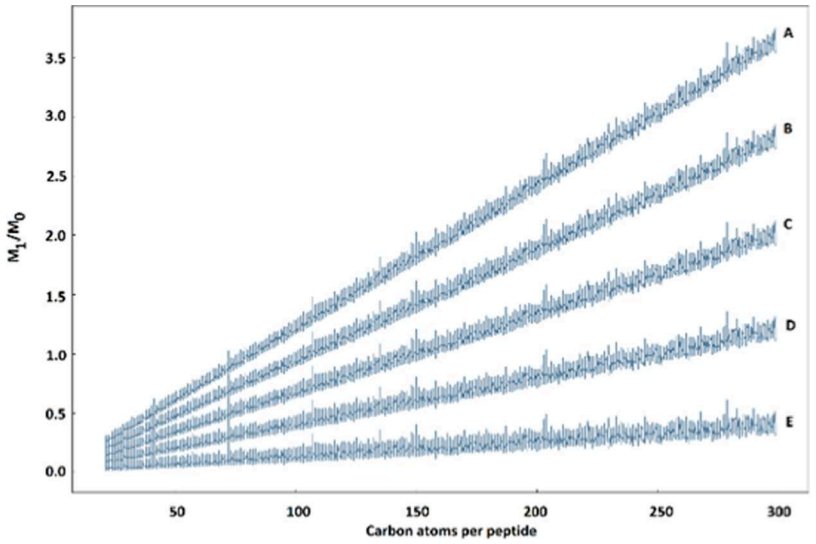


Figure 2. Relationship between the M_1/M_0 ratios and the number of carbon atoms per peptide at various ^{12}C enrichment levels. A: 98.93% ^{12}C (NC condition); B: 99.1975% ^{12}C; C: 99.465% ^{12}C; D: 99.7325% ^{12}C; and E: 100% ^{12}C (^{12}C condition).

quantify the relative abundance of the peptide originating from the NC condition and that from the 12C condition, we define a new parameter, $M_i^{^{12}C.max}$, corresponding to the maximum value of the relative intensity reached by any isotopologue under 12C-labeling conditions. To compute this parameter, the peptide sequence is split into (i) the part that may be labeled and (ii) the part that will remain nonlabeled due to the presence of NC amino acids in the growth medium. For example, in a yeast strain auxotroph for histidine, leucine, and lysine, a peptide of the sequence: TEN**LH**QS**L**TGC**L**NDYSNAFG**K** will be considered as (TENQSTGCNDYSNAFG) + (**LHLLK**), the first part being 100% ^{12}C-labeled under the 12C condition, whereas the second part remains NC, even under the 12C condition. To take into account both contributions to compute the $M_i^{^{12}C.max}$ parameter, we decompose the carbon composition x of the amino acids into two fractions: Carbon-A (C_A), composed of a atoms for the part of the peptide that will be fully ^{12}C-labeled in the 12C condition, and Carbon-B (C_B), composed of b atoms for the part of the peptide that remains nonlabeled (thus computed as in the NC condition), with $x = a + b$.

Therefore

$$M_0 = \prod_{\substack{i=a,b,y,v,w,u \\ X=C_A,C_B,H,N,O,S}} P(^AX)^i \tag{7}$$

and

$$M_1 = \sum_{\substack{X=C_A,C_B,H,N,O,S \\ i=a,b,y,v,w,u}} \left[i \times P(^AX)^{i-1} \times P(^{A+1}X) \times \prod_{\substack{i=a,b,y,v,w,u \\ Y\neq X=C_A,C_B,H,N,O,S}} P(^AY)^i \right] \tag{8}$$

$M_0^{^{12}C.max}$ and $M_1^{^{12}C.max}$ are obtained by setting $P'(^{12}C_A)$ equal to 1, and $P'(^{12}C_B)$ equal to 0.9893

The quantification of the molar fraction (α) of the peptide originating from the NC condition is therefore computed as

$$\alpha = \frac{M_0^{12C.max}M_1^{exp} - M_0^{exp}M_1^{12C.max}}{M_0^{exp}M_1^{NC} - M_0^{exp}M_1^{12C.max} - M_1^{exp}M_0^{NC} + M_1^{exp}M_0^{12C.max}} \tag{9}$$

Therefore

$$P'(^{12}C_A) = \frac{a \times M_0^{exp}}{M_1^{exp} + (a - B') \times M_0^{exp}} \tag{10}$$

With

$$B' = b \times \frac{P'(^{13}C_B)}{P'(^{12}C_B)} + B$$

However, the probability of occurrence of ^{12}C originating both from the 12C-labeled peptides and from the naturally occurring ^{12}C from nonlabeled peptides is given by

$$P'(^{12}C) = \frac{a \times P'(^{12}C_A) + b \times P'(^{12}C_B)}{a + b} \tag{11}$$

Further Extension of the bSLIM-Labeling Strategy. The bSLIM-labeling method described thus far takes advantage of the possibility to synthesize U-[^{12}C]-amino acids *in vivo* from a single U-[^{12}C]-source of carbon, such as glucose. It is possible to further reduce the complexity of the peptide/protein isotope clusters by combining the ^{12}C labeling with ^{14}N labeling, using a single U-[^{14}N]-source of nitrogen, such as ammonium sulfate or another convenient compound.

The overall processing of the experimental data remains similar to that described above, except that the $u \times \frac{P(^{15}N)}{P(^{14}N)}$ term from eq 3 is now equal to 0 when $P(^{15}N)$ is equal to 0 (100% ^{14}N condition).

It is therefore possible to set and process experimental data from the four different combinations of labeling: Normal Carbon/Normal Nitrogen, Normal Carbon/14N, 12C/Normal Nitrogen, and 12C/14N.

Definition of the Practical Computation Parameters

We assessed the theoretical ranges of the various parameters used to process the experimental datasets.

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

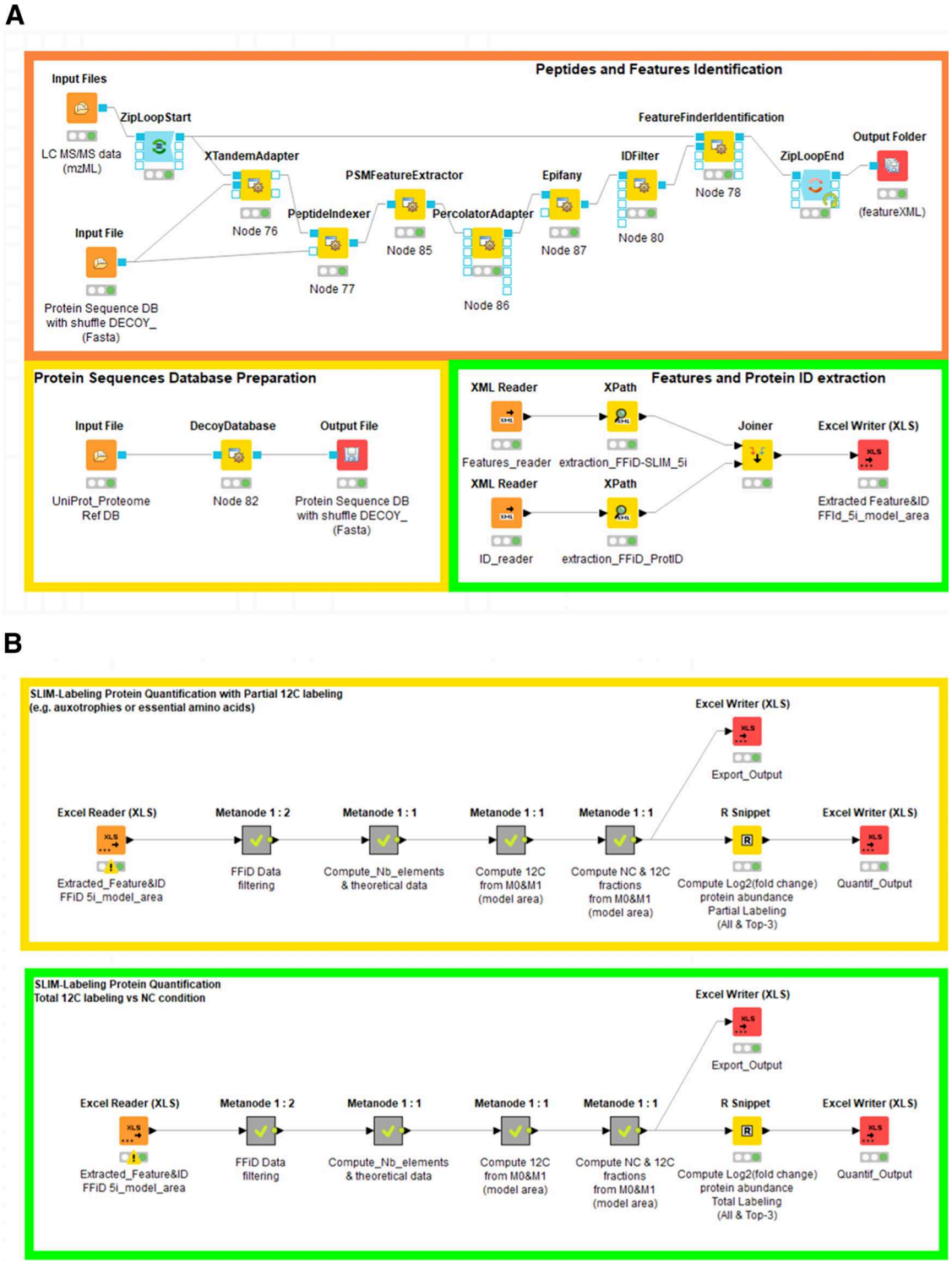


Figure 3. A. KNIME workflows for protein sequence database preparation (yellow box) for peptide identification and feature extraction (orange box) and parsing of the. featureXML files and extraction of all relevant features and Protein IDs (green box). B. KNIME workflows for ^{12}C enrichment evaluation and quantitative analysis of the differential abundance of proteins in normal carbon versus 12C-labeling conditions for total labeling (green box) and partial labeling (yellow box).

The relative abundance of ^{12}C incorporated in the SLIM-labeling experiments must be in the range $0.9893 \leq P'(^{12}C) \leq 1$

The molar fraction of unlabeled peptide must be in the range $0 \leq \alpha \leq 1$

As shown in eq 3, $\frac{M_1^{exp}}{M_0^{exp}}$ is a linear function of the mass of the peptide, reflected by the number of carbon atoms x and B and the contribution of the other elements in the mass of the peptide.

Figure 2 shows the plot of $\frac{M_1^{exp}}{M_0^{exp}}$ over the number of carbon atoms for tryptic peptides computed for various abundances of

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

total ^{12}C from the whole proteome of the yeast *Saccharomyces cerevisiae*, restricted to the mass range 600−6000 Da used in bottom-up proteomics experiments. This allows us to set the range of acceptable experimental $\frac{M_1^{exp}}{M_0^{exp}}$ values to $0.01 \leq \frac{M_1^{exp}}{M_0^{exp}} \leq 3.6$

Implementation of the Experimental Data Processing Workflow

The workflow is composed of two distinct components for (i) peptide identification and feature extraction (Figure 3A) and (ii) computation of $P'(^{12}C)$ and $\log_2(R)$ (Figure 3B).

Peptide Identification and Feature Extraction. This pipeline is based on OpenMS (v2.6.0) plug-in nodes used inside the KNIME (v4.2.2) environment.

Experimental mass spectrometry .raw files were converted to mzML files using the MS-convert tool from Proteowizard.[15,16] The *S. cerevisiae* protein sequence database consisted of the UniProt UP000002311 reference proteome augmented with shuffled entries produced using the DecoyDatabase node of OpenMS,

In the present workflow, we use X!Tandem[17] as a protein sequence database search engine, with XTandemAdapter linked to the PeptideIndexer, peptide spectrum matches (PSM) Feature extractor, PercolatorAdapter, and Epifany[18] for FDR filtering and IDFilter OpenMS nodes. The X!Tandem *E*-value threshold was set to 0.001 and FDRs were computed at 1%, both at the peptide and protein level, using the posterior error probability (PEP) scores of the peptide spectrum matches (PSMs). The resulting idXML file was processed, together with the initial mzML file, using a modified version of the FeatureFinderIdentification (FFId) node.[9] This version allows modeling the mass trace of every isotopologue by applying a Gaussian fit (or Exponential Modified Gaussian fit) to its chromatographic elution profile. This allowed extraction of the features (height of the peak, area, and detailed statistics) of five isotopologues on every isotope cluster, associated with quality scores. These data were retrieved from the FFId output files in featureXML format. We then used XPath to generate two. tsv files, one containing all of the features associated with every identification, the second containing the corresponding protein ID and accession number. To process batches of experimental files, we made use of KNIME's ZipLoopStart and ZipLoodEnd nodes.

Computation of $P'(^{12}C)$ and $\log_2(R)$. The features- and "ProtID"-containing files were joined to restore the association between the peptide and protein identifications.

We defined a first KNIME "metanode" with a series of filters to select the peptides with high-quality modeled features, flagged as "0(valid)" by FFId for M_0 and M_1. The other filters are database search engine-specific and can be easily edited depending on the preferred peptide identification strategy. In the present workflow, based on XTandemAdapter, we transformed certain post-translational modifications into a custom one-letter code: "(oxidation)" as O, "(acetyl)" as B, and "(phospho)" as Z. We included filters to remove the peptides with the "(Gln->pyro-Glu)" and "(Glu->pyro-Glu)" modifications, since FeatureFinderIdentification tended to attribute the value 0 to M_0 from the peptides identified with these modifications.

We also set two filters to remove small (<800 Da) and low-intensity peptides (M_0_model_area <1E6 in our experiments).

The second metanode computes the amino-acid composition of each peptide and therefore their elemental composition, including the elements present in the post-translational modifications. We used the Unimod[19] elemental composition of oxidation: O, acetylation: H(2)C(2)O, and phosphorylation: HO(3)P.

In cases of incomplete labeling, we introduced specific nodes to compute the number of nonlabeled residues in the peptide sequence (auxotrophies, H, L, and K in the example above) and the number of carbon atoms associated with these residues. As described above, these atoms are therefore referred to as Carbon-B (C_B) in the rest of the "partial labeling workflow". The settings can be easily modified in the configuration dialog box associated with these nodes to adjust for the actual experiments.

The elemental composition of each peptide is used to calculate the theoretical M_0 and M_1 values at the natural ^{12}C abundance (98.93%) and when ^{12}C is 100%. In partial labeling experiments, we compute $M_0^{^{12}C.max}$ and $M_1^{^{12}C.max}$, taking into account the fraction of carbon atoms (C_A) that can be labeled, and the fraction (C_B) not labeled.

The third metanode calculates the experimental $\frac{M_1^{exp}}{M_0^{exp}}$ ratio and filters the data according to the maximum theoretical ratio in the mass range considered (0−6000 Da). The next node computes the abundance of ^{12}C according to eq 4.

The last metanode determines the molar fractions of normal carbon (NC) and ^{12}C-labeled peptides in multiplexing experiments by implementing eq 6. Next, the ratio $\frac{NC}{12C}$ is calculated and log 2-transformed.

Integration of bSLIM-Labeling Quantification Data at the Protein Level. To edit variations of protein abundance, we integrated an R script into the workflow through the KNIME R-snippet node. This script allows the grouping of peptides per protein and calculation of the mean and median of the log 2 of the $\frac{NC}{^{12}C}$ ratios, both for all of the quantified peptides and using only the three peptides with the highest M0_model_area intensity (Top-3).

The final results are written into an Excel file for further biological interpretation.

The overall SLIM-labeling data processing workflow is presented in Figure 2A,B. When processing a 1.53 Gb mzML file with the *S. cerevisiae* protein sequence database, the overall process from peptide identification to protein quantification is executed in an average of 1355 s per file of experimental data (on Intel Xeon Gold 6134, 8 Cores, 3.2−3.7 GHz Turbo, 24.75 Mo Cache and 16 Go 2666 MHz DDR4 RAM). Most of this time was dedicated to the peptide identification and filtering steps including Percolator and Epifany (343 s), FFId (305 s), and features and protein ID extraction (707 s).

The complete workflows are provided as Executable files E1 and E2. The files can be directly imported into a valid KNIME instance and run as provided, with the appropriate adjustments of certain parameters (protein sequence database, ion intensity threshold, auxotrophies) specific to the user's experimental conditions. The complete list of the parameters used for each node of the pipelines is provided in Supporting Files S2 and S3.

EXPERIMENTAL PROCEDURES

Strains and Growth Conditions

S. cerevisiae strains S288c (MATα SUC2 gal2 mal2 mel flo1 flo8-1 hap1 ho bio1 bio6)[20] and its isogenic derivative BY4742

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

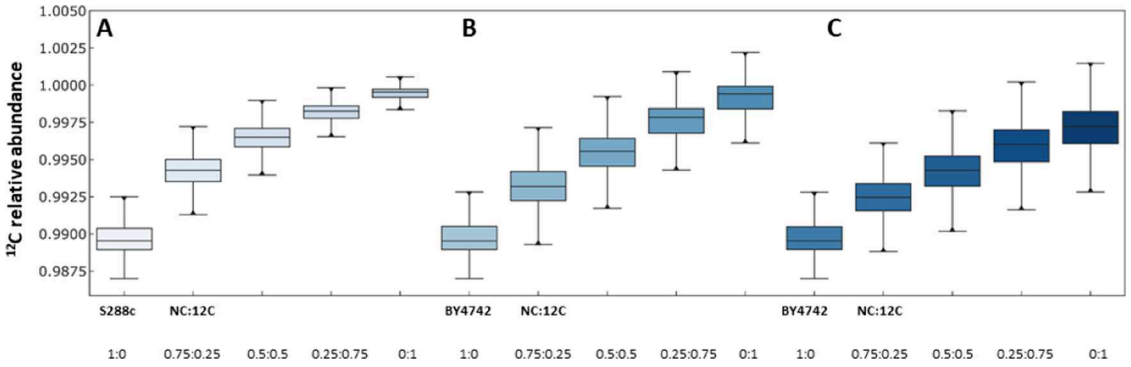


Figure 4. ^{12}C incorporation levels in the peptides from the experimental standard curves made of mixtures of known amounts of ^{12}C-labeled and nonlabeled peptides (100% NC, 75% NC-25% 12C, 50% NC-50% 12C, 25% NC-75% 12C, and 100% 12C) from (A) the prototrophic yeast strain S288c (n = 4) and (B) the auxotrophic yeast strain BY4742 (n = 4). (C) Corresponding total ^{12}C incorporation into proteins from BY4742.

(MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0)[21,22] were used throughout this work. S288c cells were grown on a defined synthetic medium (Yeast Nitrogen Base, with ammonium sulfate, without amino acids, Difco BD Life) with glucose (0.5% w/v) as the sole source of carbon. BY4742 cells were grown under the same conditions, with filter-sterilized L-Histidine, L-Leucine, L-Lysine, and Uracil having 20, 100, 50, and 20 mg/L final concentrations, respectively. We ran two types of cultures using either regular glucose (N-Glc) or glucose containing only ^{12}C carbon atoms (U-[^{12}C]-glucose, Euriso-Top, Saint-Aubin, France; filter-sterilized; ^{12}C-Glc). The carbon isotopic composition of N-Glc is 98.93% ^{12}C/1.07% ^{13}C. Cell cultures (100 mL) were inoculated at an initial OD_{600} of 0.03 with liquid precultures (10 mL of the same medium inoculated at an OD_{600} of 0.03 with colonies freshly grown on YPD agar plates) and incubated in an orbital shaker at 30 °C. The cells were collected by centrifugation for 5 min at 1200g. The cell pellet was resuspended in 10 vol/g cells of lysis buffer pH 7.5 consisting of 20 mM Tris HCl, 140 mM NaCl, 5 mM $MgCl_2$, 1 mM DTE, and one tablet of protease inhibitor cocktail 2X Sigma fast/10 mL lysis buffer. The cell extracts were prepared by vortexing the cell suspension for six cycles of 2 min in the presence of a 2 vol/vol cell suspension of acid-washed glass beads (0.45−0.5 mm Ø). The cell lysate was diluted twice in the same buffer and large cell debris removed by centrifugation for 10 min at 1200g. The supernatant, referred to hereafter as the cell homogenate, was rapidly frozen in 5 mg protein/mL aliquots and stored at −80 °C. After a BCA protein assay, specific amounts of protein extracts from ^{12}C-Glc and N-Glc cultures were mixed for liquid chromatography with tandem mass spectrometry (LC-MS/MS) processing.

LC-MS/MS Acquisition

Protein extracts (40 μg) were precipitated with acetone at −20 °C. The protein pellets were collected by centrifugation. The acetone supernatant was carefully removed and the protein pellet resuspended in 20 μL 25 mM NH_4HCO_3 containing sequencing-grade trypsin (0.2 μg/μL aliquots, Promega) and incubated overnight at 37 °C. The resulting peptides were desalted using ZipTip μ-C18 pipette tips (Pierce, Thermo Fisher Scientific) and analyzed on a Q Exactive Plus coupled to a Nano-LC Proxeon 1000 equipped with an Easy Spray ion source (all from Thermo Scientific). Peptides were separated by chromatography with the following parameters: Acclaim PepMap100 C18 pre-column (2 cm, 75 μm i.d., 3 μm, 100 Å), LC EASY-Spray C18 column (50 cm, 75 μm i.d., 2 μm bead size, 100 Å pore size) operated at 55 °C, 300 nL/min flow rate, gradient from 95% solvent A (water, 0.1% formic acid) to 35% solvent B (100% acetonitrile, 0.1% formic acid) over a period of 97 min, followed by column regeneration for 23 min, giving a total run time of 2 h. Peptides were analyzed in the Orbitrap cell, in full ion scan mode, at a resolution of 70 000 (at m/z 200), with a mass range of m/z 375−1500 and an Automatic Gain Control (AGC) target of 3×10^6. Fragments were obtained by higher-energy C-trap dissociation (HCD) activation with a collisional energy of 30%, and a quadrupole isolation window of 1.4 Da. MS/MS data were acquired in the Orbitrap cell in Top20 mode, at a resolution of 17 500 with an AGC target of 2×10^5, with a dynamic exclusion of 30 s. MS/MS spectra of the most intense precursors were acquired first. Peptides with an unassigned charge state or those that were singly charged were excluded from the MS/MS acquisition. The maximum ion accumulation times were set to 50 ms for MS acquisition and 45 ms for MS/MS acquisition.

Experimental Validation of the Bottom-Up Quantification Pipeline

We collected experimental datasets using the cell extracts from the yeast cultures grown under NC or 12C conditions to test and validate our data processing pipelines.

The first two datasets consisted of a series of samples of (i) S288c and (ii) BY4742 peptides. In both cases, peptides were analyzed as 1:0, 0.75:0.25, 0.5:0.5, 0.25:0.75, and 0:1 vol/vol mixtures of identical amounts of total protein extracts prepared from the NC and 12C conditions, respectively. For BY4742, we expected lower ^{12}C incorporation rates, as the His, Leu, and Lys residues remain unlabeled.

The third dataset consisted of an analysis of 1:1 mixtures of identical amounts of protein from S288c and BY4742 extracts prepared under the 12C or NC condition (S288C-12C vs BY4742-NC and S288c-NC vs BY4742-12C, with biological duplicates and technical replicates). This allowed quantitative analysis of proteome differences between two yeast strains, often referred to as "wild-type" strains, in a very stringent condition, as only a few proteins are expected to vary significantly. The quantitative results obtained were analyzed using a non-supervised Bayesian clustering algorithm AutoClass@IJM.[23,24] The information content of the various classes was evaluated using the *Saccharomyces* Genome Database[25] (SGD) annotation files and GO-term analysis tools (SGD and BiNGO[26]).

Figure 4 shows the inferred ^{12}C abundance in the various samples as box plots and the correlation with the theoretical values expected from an incomplete labeling of the peptides, due to the presence of nonlabeled amino acids in the growth media.

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

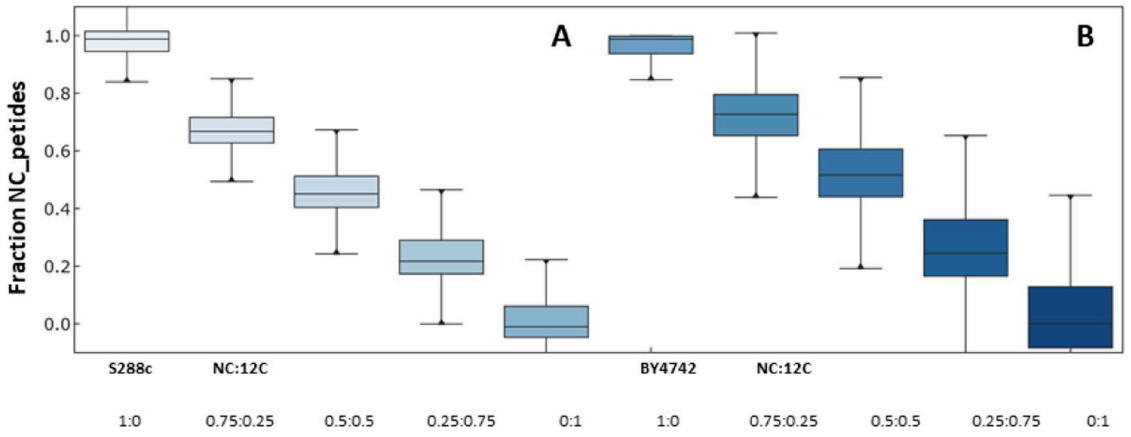


Figure 5. (A) Molar fraction of the nonlabeled peptides in proteins of the prototrophic yeast strain S288c from the experimental standard curves made of mixtures of known amounts of ^{12}C-labeled and nonlabeled peptides (100% NC, 75% NC-25% 12C, 50% NC-50% 12C, 25% NC-75% 12C, and 100% 12C) (n = 4). (B) Molar fraction of the nonlabeled peptides in proteins of the auxotrophic yeast strain BY4742 from the experimental standard curves made of mixtures of known amounts of ^{12}C-labeled and nonlabeled peptides (100% NC, 75% NC-25% 12C, 50% NC-50% 12C, 25% NC-75% 12C, and 100% 12C).

RESULTS AND DISCUSSION

Here, we describe an easy-to-use and straightforward workflow to process bSLIM-labeling quantitative proteomics data. The workflow (Figure 4A,B) is implemented in the KNIME environment and uses open-source resources from the OpenMS suite and dedicated R scripts.

The workflow integrates a series of nodes for peptide identification through a database search engine, the calculation of appropriate FDR metrics, feature detection for high-confidence identified peptides, and extraction of the intensity of every isotopologue in every isotope cluster of these peptides. Although we used X!Tandem in this study, other search engines are supported by OpenMS, such as MS-GF+, Mascot, or OMSSA. These search engines may be used with *ad hoc* parameters and produce the idXML files required for feature detection. Typically, we identified approximately 25 000 peptides with X!Tandem *E*-values better than 1E-3 at a 1% FDR.

A key node in this workflow is OpenMS FeatureFinder-Identification (FFId). This node was originally developed for label-free quantitative proteomics, with an efficient algorithm to extract the raw data (m/z and intensity) of peptides at the MS1 level, find chromatographic peaks (features), and model peptide elution by a Gaussian fit. A control algorithm provides a quality score for each feature. Here, we used a modified version of FFId that allows extraction of the data (raw and modeled) from every isotopologue trace of each identified peptide isotope cluster. All of the analyses were performed using the "Model Area" values, a metric commonly used in modern label-free quantification software.[27] Since most of the developments of the SLIM-labeling quantification method presented here rely on accurate M_0 and M_1 intensity measurements, we selected only the data that passed the quality control step (flagged as "0 = valid") for further processing. Applying this stringent filter reduces the total number of entries to an average of 14−16 000.

All of the datasets were acquired using an Orbitrap Q-Exactive Plus mass spectrometer. Although this instrument is remarkable in terms of mass accuracy, there may be some concerns about the precision of ion intensity measurements in isotope clusters[28−30] using the Orbitrap as a mass detector, with errors above 20% in the M1 intensity measurement for low-abundance ions. We attempted to overcome this limitation by running samples at different AGC target values (from 1E5 to 5E6) and different resolutions (from 7.5 to 120 K) in the MS1 acquisition step. However, this did not significantly improve the output of the analysis workflow. The datasets obtained in our standard conditions were nonetheless considered to be of sufficient quality to validate the methods and workflows.

These experimental uncertainties have important consequences when computing the incorporation of ^{12}C in SLIM-labeling experiments and the fraction of nonlabeled peptides in quantitative experiments. Indeed, we use the experimental M_0 and M_1 data with reference to the corresponding theoretical values computed both under the NC condition and under 100% enrichment in the 12C condition, together with the elemental composition of the peptides. Under ideal conditions, the $P'(^{12}C)$ value should be in the range of 0.9893−1, and the nonlabeled peptide fraction α between 0 and 1. However, when running our workflow to compute these factors, we found a large fraction of peptides (3−4 000) with α values (as an example) between 1.005 and 1.007 instead of 1, or between −0.008 and −0.009 instead of 0. We chose to correct the value of these outliers to 1 or 0, respectively. After these corrections, the average number of high-quality data points to process was 11−12 000 per dataset.

Figure 4A, computed using eq 4, shows the ^{12}C incorporation levels in the peptides from the experimental standard curves, composed of mixtures of known amounts of ^{12}C-labeled- and nonlabeled peptides (100% NC, 75% NC-25% 12C, 50% NC-50% 12C, 25% NC-75% 12C, and 100% 12C) from the yeast strain S288c. Equation 4 states that the variation of ^{12}C levels follows an equilateral hyperbola, as observed in the experimental data.

The SLIM-labeling and protein quantification strategy is well adapted to prototroph organisms, such as wild-type microorganisms that can grow using glucose as the sole source of carbon to sustain all amino-acid biosynthesis. However, incomplete labeling is likely to occur in most biological studies involving laboratory strains with growth dependence for specific amino acids, used as selection markers, or in cell cultures requiring essential amino acids, such as human cell lines. We adapted the SLIM-labeling data processing to these situations by enumerating the amino-acid residues that cannot be ^{12}C labeled in every peptide sequence and calculating the number of associated carbon atoms. We therefore defined a new element

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

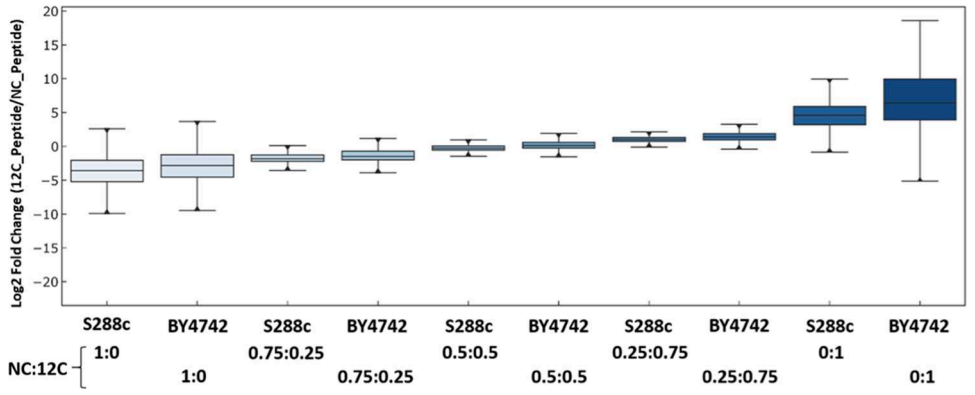


Figure 6. Log 2 transform of the NC/12C ratios at the peptide level from the experimental standard curves made of mixtures of known amounts of ^{12}C-labeled and nonlabeled peptides (100% NC, 75% NC-25% 12C, 50% NC-50% 12C, 25% NC-75% 12C, and 100% 12C) from the S288c and BY4742 strains.

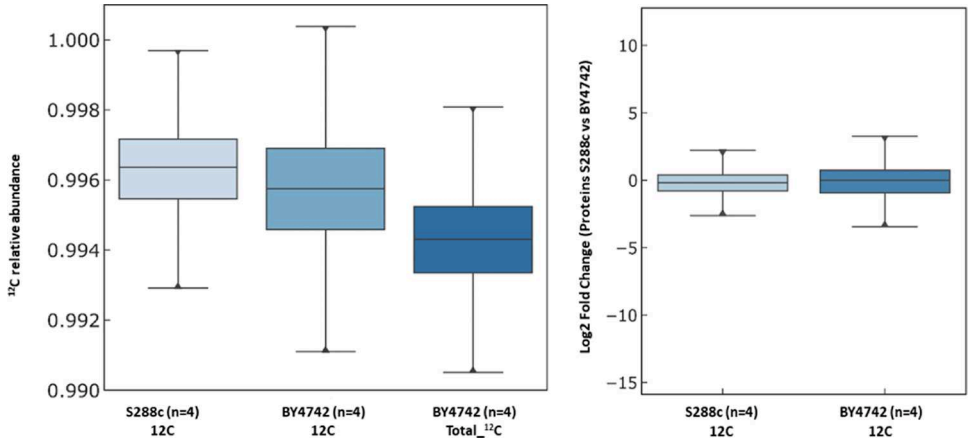


Figure 7. Left: Distribution of ^{12}C incorporation in protein samples from 1:1 mixtures of the S288c-12C and BY4742-NC conditions (left), from the BY4742 -12C and S288c-NC conditions (center), and total ^{12}C calculated for the later sample (right). Right: Distribution of the log 2 transform of the NC/12C ratios at the protein level from the 1:1 sample mixtures from the S288c-12C and BY4742-NC conditions (left) and the BY4742-12C and S288c-NC conditions (right).

(referred to as "carbon-B") in the elemental composition, which represents the isotope distribution of natural carbon. This element is incorporated into eqs 4 and 7 with the same status as hydrogen, nitrogen, oxygen, and sulfur, which have an isotope distribution that remains invariant under full ^{12}C-labeling conditions. We numerically tested this approach using MIDAs (on the command line), in which we added this new element with the features of regular carbon. The numerical simulations showed that eq 9 indeed allowed us to quantify only the exclusive contribution of the NC peptide from the NC condition but not that related to the NC part of the peptides (auxotrophies) from the 12C condition.

Analogous to Figure 4A,B, computed using eq 10, shows the ^{12}C incorporation levels in the peptides from the experimental standard curves for the yeast strain BY4742, which requires exogenous histidine, leucine, and lysine for growth. These residues cannot therefore be labeled in the 12C condition. The experimental data show that the incorporation in ^{12}C is similar to that measured in the extracts from the S288c strain, meaning that the SLIM-labeling strategy allows accurate analysis of ^{12}C incorporation rates, even when there is some heterogeneity in the distribution of nonlabeled amino acids in the identified peptides. In a typical experiment with 10 950 peptides (average length: 16 residues), the content of H, L, and K represented from 0 to 62% of the peptide composition, with a mean value of 16.5% (average size of the peptides: 16 residues). However, since a part of the peptides could not incorporate extra ^{12}C due to auxotrophy, the total amount of ^{12}C calculated using eq 11 remained lower (Figure 4C), with a broader distribution of ^{12}C at each point.

We obtained a linear relationship between the NC fraction and the expected amount of the NC:12C peptides for the points of the standard curve, both under conditions of full labeling (S288c, Figure 5A) and conditions of partial labeling (BY4742, Figure 5B).

Although the standard curves were established by mixing controlled amounts of peptides from the same strain, grown in the same medium but with different glucose source (NC and 12C conditions), it is still possible to calculate the fold change (FC) for each peptide, defined as the ratio of the ion count from the 12C sample to that from the original NC sample calculated using the bSLIM-labeling quantification. The log 2(FC)s were indeed centered on 0 in the 1:1 mix (50% of each sample) and varied according to the expected ratio, both with the S288c (total labeling) and BY4742 (partial labeling) samples (Figure 6).

The variability under the extreme conditions (1:0 and 0:1, NC:12C) reflects that one of the terms of the ratio tends to 0, and thus the ratio tends to infinity.

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

Overall, these results underscore the soundness of the theoretical grounds for the data processing in SLIM-labeling quantification.

We tested the specificity and sensitivity of the SLIM-labeling quantitative strategy by comparing the differences in protein abundance at the proteome level between the two yeast strains S288c and its isogenic derivative BY4742, the latter strain carrying null alleles for the URA3, HIS3, LEU2, and LYS2 genes. These strains are considered to be prototypical "wild-type" laboratory strains. We performed a label swap to compare the quantitative data of ^{12}C-labeled S288c versus NC BY4742 extracts and BY4742 12C labeling (giving rise to partial labeling) versus NC S288c extracts. The distributions of ^{12}C incorporation (Figure 7A) and of the log 2(FC) (Figure 7B) were similar to those expected from samples of closely related proteome composition and abundance.

The overall analysis allowed us to identify 4162 proteins. However, the quantification was performed on only the 3354 proteins identified by at least eight peptides in the eight samples resulting from two biological and two technical replicates.

The proteins encoded by the genes used as auxotrophic markers (Ura3, Leu2, Lys2, His3) exhibited an average log 2(FC), ranging from 3 to 6 for strain S288c. His3 is a protein not easy to detect and quantify, since it is a low-abundance protein in wild-type cells.[31] Here, the quantification and the measured FC for the auxotrophy-related proteins clearly showed the sole contribution of the S288c proteins when observed in the S288c-NC vs BY4742-12C condition, indicating the power of the bSLIM method.

Several unexpected differences were also uncovered by this analysis. Two classes of 13 and 22 proteins were overexpressed more than 3- and 2-fold, respectively, in the S288c genetic background (Supporting Figure 1). A third class, consisting of 50 proteins, were underexpressed more than 2-fold in S288c. Among the overexpressed proteins in S288c, the samples were representatives of the three pathways for glutamate synthesis (Gdh1, Gdh3, Gln1, Glt1). Gln1 is known to be regulated by nitrogen source and amino-acid limitation. As all of the amino acids in S288c are synthesized endogenously from glucose, this result suggests that their synthesis may be sufficient to sustain growth but sufficiently limited to trigger specific regulation pathways. A number of proteins involved in mitochondrial functions were also more abundant in S288c than BY4742 (*e.g.*, Cox5A, Cox9, Cox12, Cox13, Cox2, Tim11, Aco1, Aco2, and many subunits of the F_0−F_1 ATPases). This is consistent with the data of Young and Court[32] and Dimitrov et al.,[33] showing the importance of the genetic background on mtDNA stability and respiratory competence in yeast. Although the S288c strain was selected based on its auxotrophy for biotin due to *bio1* and *bio6* mutations, other proteins of the biotin biosynthesis pathway (Bio4, Bio2, Bio3) were specifically found to be more highly produced by S288c than BY4742. These proteins are involved in the *de novo* biotin synthesis from 7-keto-8-aminopelargonate, a precursor derived from alanine metabolism (for a recent review, see Perli et al.[34]).

Among the proteins overexpressed in BY4742, many appear to be related to the main pathways involved in the synthesis of histidine, leucine, and lysine, the amino acids added to the BY4742 growth media to supplement the genetic auxotrophy present in this strain. The *leu2* mutation affects not only the penultimate step of leucine biosynthesis (3-Isopropylmalate to (2S)-2-Isopropyl-3-oxosuccinate reaction) but also an early step of the valine and isoleucine biosynthesis pathways at the 2-oxobutanoate production reaction. Our data show that certain compensatory mechanisms were activated in BY4742, as the Ilv2,3,5 proteins were overexpressed, as well as Leu1,4,5,9 and Bat1,2, which are directly involved in terminal reactions of leucine and valine biosynthesis. We made a similar observation concerning the histidine regulon, in which His1, 4, 7, and 2 were more abundant in the BY4742 genetic context. However, the extent of the differences in protein synthesis was less than that observed for the leucine regulon. Analysis of the proteins of the lysine biosynthesis pathway showed a general increase in the abundance of proteins involved in lysine biosynthesis in the S288c context in our dataset, presumably caused by the absence of the Lys5 target protein Lys2. Surprisingly, most of the proteins involved in arginine biosynthesis (Arg4, 8, 5, 6, 1, 7) were also more abundant in the BY4742 context. This may be related to the observed decreased in mitochondrial function, which affects the arginine biosynthesis pathway.[35] The global analysis of these differences in protein abundance is presented in Supporting Data (AutoClass classification), in which the proteins with similar expression profiles are found in well-defined clusters. Not surprisingly, most proteins exhibited similar abundance in S288c and BY4742. The detail of certain representative clusters is presented in Supporting Figure 1, in which protein abundance varying by a factor >2 are shown. Most of the proteins discussed above are found in these clusters.

CONCLUSIONS

Here, we present an original and highly efficient workflow to process bSLIM-labeling data based on open-source resources. The overall data analysis process is rapid, *i.e.*, less than 20 min from raw data to protein quantification, and the input and output use standard file formats (mzML, .tsv, Excel files). The workflows are robust and well adapted to processing data in experiments in which proteins are fully labeled with ^{12}C, as well as in experiments in which proteins are only partially labeled, due to the requirement of essential amino acids not yet available as U-[^{12}C]-AA. The workflows are easy to edit for adaptation for specific experiments. It is possible to extend them with new functionalities and to add additional tools and features to the analysis. For example, we used X!Tandem as a database search engine for peptide identification in the present workflow, but it is possible to add other search engines and combine the results files in a single idXML file for feature extraction. The sensitivity of the quantification method allowed us to compare the proteome of two isogenic yeast strains that theoretically differ solely by the absence of four proteins that are missing due to their corresponding gene deletion. Indeed, we were able to demonstrate their absence, but we also revealed a number of other differences, showing that the overall physiology of these strains is different, especially in terms of their mitochondrial function and the amino-acid metabolism. This has important consequences for all studies based on comparative genomics in the yeast *S. cerevisiae*. These factors should all help us to promote dissemination of the bSLIM-labeling quantitative proteomics approach as a cost-effective method, with great potential for applications both for bottom-up and top-down analyses.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00478.

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

Supplemental File S1: Development of the relative intensity calculation for the isotopologues M_0 and M_1 from a peptide sequence. Supplemental File S2: Excel folder with the complete description of the parameters used for each node of the Executable files E1 and E2. Supplemental File S3: Text file describing the regular expressions associated with each node of the Executable files E1 and E2. Supplemental File S4 (.zip format):. cdt file of the AutoClass clustering of the S288c vs BY4742 SLIM-labeling-based quantitative analysis. Supplemental Figure F1: Clusters of proteins with similar expression profiles with characteristic fold changes in the quantitative analysis of a 1:1 mixture of proteins from the prototrophic yeast strain S288c and the auxotrophic yeast strain BY4742 in label-swap experiments (data from technical duplicates of biological replicates; $n = 4$). The partial labeling condition on the left columns corresponds to BY4742 grown on ^{12}C-glucose, while the right columns correspond to the ^{12}C-glucose labeling of S288c Executable file E1 (.zip format). KNIME workflows (.knwf file) for protein sequence database preparation, peptide identification, and feature extraction and parsing of the. featureXML files, and extraction of all relevant features and Protein IDs. File name: bSLIM_5i_w_model_area_Identification_FFId. Executable file E2 (.zip format). KNIME workflows (.knwf file) for ^{12}C enrichment evaluation and quantitative analysis of the differential abundance of proteins in normal carbon versus ^{12}C-labeling conditions for total labeling or partial labeling. File name: bSLIM_5i_w_model_area_R-snippet_Quantif_Prot_total_partial_labeling (PDF)

Supplemental File S2_(Parameter_Knime4.2.2Workflow_Nov2020) (XLSX)

Supplemental File S4_(AutoClass_S288cvsBY4742_12C_NC&Glc_swap.cdt) (ZIP)

E1_K4.2.2_Oms2.6.0_Identification_FFId_(model 5i)_Nov2020 (ZIP)

E2_K4.2.2_Oms2.6.0_FFiD_5i_w_model_area_R-snippets_Quantif_Prot_total&partial_labeling_-Nov2020 (ZIP)

AUTHOR INFORMATION

Corresponding Author

Jean-Michel Camadro − *Mitochondria, Metals, and Oxidative Stress Group, Université de Paris CNRS, Institut Jacques Monod, 75013 Paris, France; ProteoSeine@IJM, Université de Paris CNRS, Institut Jacques Monod, 75013 Paris, France;* orcid.org/0000-0002-8549-2707; Email: jean-michel.camadro@ijm.fr

Authors

Nicolas Sénécaut − *Mitochondria, Metals, and Oxidative Stress Group, Université de Paris CNRS, Institut Jacques Monod, 75013 Paris, France;* orcid.org/0000-0001-7948-6776

Gelio Alves − *National Center for Biotechnology Information, NLM, NIH, Bethesda, Maryland 20894, United States*

Hendrik Weisser − *STORM Therapeutics Limited, Cambridge CB22 3AT, U.K*

Laurent Lignières − *ProteoSeine@IJM, Université de Paris CNRS, Institut Jacques Monod, 75013 Paris, France*

Samuel Terrier − *ProteoSeine@IJM, Université de Paris CNRS, Institut Jacques Monod, 75013 Paris, France*

Lilian Yang-Crosson − *Mitochondria, Metals, and Oxidative Stress Group, Université de Paris CNRS, Institut Jacques Monod, 75013 Paris, France*

Pierre Poulain − *Mitochondria, Metals, and Oxidative Stress Group, Université de Paris CNRS, Institut Jacques Monod, 75013 Paris, France*

Gaëlle Lelandais − *Institut de Biologie Intégrative de la Cellule, 91190 Orsay, France*

Yi-Kuo Yu − *National Center for Biotechnology Information, NLM, NIH, Bethesda, Maryland 20894, United States;* orcid.org/0000-0002-6213-7665

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.0c00478

Author Contributions

N.S., G.A., P.P., Y.-K.Y., and J.-M.C. conceptualized the method. N.S., H.W., L.Y.-C., P.P., G.L., and J.-M.C. contributed to the development of the post-processing algorithms. N.S. and J.-M.C. designed and implemented the post-processing algorithms. N.S., L.L., and S.T. performed the experiments. N.S., G.L., and J.-M.C. analyzed the data. N.S. and J.-M.C. wrote the manuscript with input from all of the authors.

Notes

The authors declare no competing financial interest.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[36] partner repository with the dataset identifier PXD021329. A training set with. mzml files from partial labeling, and total labeling data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD021502.

ACKNOWLEDGMENTS

This work was supported in part by Agence Nationale de la Recherche grant ANR-18-CE44-0014. NS received a thesis grant-in-aid from the Center for Interdisciplinary Research (CRI-Paris). The English text was edited by Alex Edelman & Associates.

REFERENCES

(1) Léger, T.; Garcia, C.; Collomb, L.; Camadro, J. M. A Simple Light Isotope Metabolic Labeling (SLIM-labeling) Strategy: A Powerful Tool to Address the Dynamics of Proteome Variations In Vivo. *Mol. Cell. Proteomics* **2017**, *16*, 2017−2031.

(2) Marshall, A. G.; Senko, M. W.; Li, W. Q.; Li, M.; Dillon, S.; Guan, S. H.; Logan, T. M. Protein molecular mass to 1 Da by C-13, N-15 double-depletion and FT-ICR mass spectrometry. *J. Am. Chem. Soc.* **1997**, *119*, 433−434.

(3) Shi, S. D.; Hendrickson, C. L.; Marshall, A. G. Counting individual sulfur atoms in a protein by ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry: experimental resolution of isotopic fine structure in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11532−11537.

(4) Rodgers, R. P.; Blumer, E. N.; Hendrickson, C. L.; Marshall, A. G. Stable isotope incorporation triples the upper mass limit for determination of elemental composition by accurate mass measurement. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 835−840.

(5) Wang, D.; Liem, D. A.; Lau, E.; Ng, D. C.; Bleakley, B. J.; Cadeiras, M.; Deng, M. C.; Lam, M. P.; Ping, P. Characterization of human plasma proteome dynamics using deuterium oxide. *Proteomics - Clin. Appl.* **2014**, *8*, 610−619.

(6) Borzou, A.; Sadygov, V. R.; Zhang, W.; Sadygov, R. G. Proteome Dynamics from Heavy Water Metabolic Labeling and Peptide Tandem Mass Spectrometry. *Int. J. Mass Spectrom.* **2019**, *445*, No. 116194.
(7) Alves, G.; Ogurtsov, A. Y.; Yu, Y. K. Molecular Isotopic Distribution Analysis (MIDAs) with adjustable mass accuracy. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 57−70.
(8) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H. C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmstrom, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **2016**, *13*, 741−748.
(9) Weisser, H.; Choudhary, J. S. Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *J. Proteome Res.* **2017**, *16*, 2964−2974.
(10) Warr, W. A. Scientific workflow systems: Pipeline Pilot and KNIME. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 801−804.
(11) Fillbrunn, A.; Dietz, C.; Pfeuffer, J.; Rahn, R.; Landrum, G. A.; Berthold, M. R. KNIME for reproducible cross-domain analysis of life science data. *J. Biotechnol.* **2017**, *261*, 149−156.
(12) Yergey, J. A. A general-approach to calculating isotoic distributions for mass spectrometry. *J. Mass Spectrom.* **2020**, *52*, 337−349.
(13) Wang, B.; Sun, G.; Anderson, D. R.; Jia, M.; Previs, S.; Anderson, V. E. Isotopologue distributions of peptide product ions by tandem mass spectrometry: quantitation of low levels of deuterium incorporation. *Anal. Biochem.* **2007**, *367*, 40−48.
(14) Audi, G.; Wapstra, A. H. The 1993 Atomic Mass Evaluation. 1. Atomic Mass Table. *Nucl. Phys. A* **1993**, *565*, 1−65.
(15) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr. Protoc. Bioinf.* **2014**, *46*, 13−24.
(16) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534−2536.
(17) Duncan, D. T.; Craig, R.; Link, A. J. Parallel tandem: a program for parallel processing of tandem mass spectra using PVM or MPI and X!Tandem. *J. Proteome Res.* **2005**, *4*, 1842−1847.
(18) Pfeuffer, J.; Sachsenberg, T.; Dijkstra, T. M. H.; Serang, O.; Reinert, K.; Kohlbacher, O. EPIFANY: A Method for Efficient High-Confidence Protein Inference. *J. Proteome Res.* **2020**, *19*, 1060−1072.
(19) Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4*, 1534−1536.
(20) Mortimer, R. K.; Johnston, J. R. Genealogy of principal strains of the yeast genetic stock center. *Genetics* **1986**, *113*, 35−43.
(21) Brachmann, C. B.; Davies, A.; Cost, G. J.; Caputo, E.; Li, J.; Hieter, P.; Boeke, J. D. Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **1998**, *14*, 115−132.
(22) Winston, F.; Dollard, C.; Ricupero-Hovasse, S. L. Construction of a set of convenient Saccharomyces cerevisiae strains that are isogenic to S288C. *Yeast* **1995**, *11*, 53−55.
(23) Achcar, F.; Camadro, J. M.; Mestivier, D. AutoClass@IJM: a powerful tool for Bayesian classification of heterogeneous data in biology. *Nucleic Acids Res.* **2009**, *37*, W63−W67.
(24) Camadro, J.-M.; Poulain, P. AutoClassWrapper: a Python wrapper for AutoClass C classification. *J. Open Science Software* **2019**, *4*, 1390−1392.
(25) Cherry, J. M.; Hong, E. L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; Chan, E. T.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S. R.; Fisk, D. G.; Hirschman, J. E.; Hitz, B. C.; Karra, K.; Krieger, C. J.; Miyasato, S. R.; Nash, R. S.; Park, J.; Skrzypek, M. S.; Simison, M.; Weng, S.; Wong, E. D. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **2012**, *40*, D700−D705.
(26) Maere, S.; Heymans, K.; Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **2005**, *21*, 3448−3449.
(27) Al Shweiki, M. R.; Monchgesang, S.; Majovsky, P.; Thieme, D.; Trutschel, D.; Hoehenwarter, W. Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J. Proteome Res.* **2017**, *16*, 1410−1424.
(28) Erve, J. C.; Gu, M.; Wang, Y.; DeMaio, W.; Talaat, R. E. Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2058−2069.
(29) Xu, Y.; Heilier, J. F.; Madalinski, G.; Genin, E.; Ezan, E.; Tabet, J. C.; Junot, C. Evaluation of accurate mass and relative isotopic abundance measurements in the LTQ-orbitrap mass spectrometer for further metabolomics database building. *Anal. Chem.* **2010**, *82*, 5490−5501.
(30) Weber, R. J.; Southam, A. D.; Sommer, U.; Viant, M. R. Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification. *Anal. Chem.* **2011**, *83*, 3737−3743.
(31) Ho, B.; Baryshnikova, A.; Brown, G. W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.* **2018**, *6*, 192−205 e3.
(32) Young, M. J.; Court, D. A. Effects of the S288c genetic background and common auxotrophic markers on mitochondrial DNA function in *Saccharomyces cerevisiae*. *Yeast* **2008**, *25*, 903−912.
(33) Dimitrov, L. N.; Brem, R. B.; Kruglyak, L.; Gottschling, D. E. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. *Genetics* **2009**, *183*, 365−383.
(34) Perli, T.; Wronska, A. K.; Ortiz-Merino, R. A.; Pronk, J. T.; Daran, J. M. Vitamin requirements and biosynthesis in *Saccharomyces cerevisiae*. *Yeast* **2020**, *37*, 283−304.
(35) Braun, R. J.; Sommer, C.; Leibiger, C.; Gentier, R. J.; Dumit, V. I.; Paduch, K.; Eisenberg, T.; Habernig, L.; Trausinger, G.; Magnes, C.; Pieber, T.; Sinner, F.; Dengjel, J.; van Leeuwen, F. W.; Kroemer, G.; Madeo, F. Accumulation of Basic Amino Acids at Mitochondria Dictates the Cytotoxicity of Aberrant Ubiquitin. *Cell Rep.* **2015**, *10*, 1557−1571.
(36) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**, *47*, D442−D450.

https://dx.doi.org/10.1021/acs.jproteome.0c00478
J. Proteome Res. 2021, 20, 1476−1487

III. Validation des statistiques descriptives

Nous avons développé un algorithme en langage R, facilement implémentable à la suite du workflow de quantification bSLIM sous KNIME (voir ci-dessus).

Afin d'observer la robustesse de notre méthode d'évaluation statistique, nous avons utilisé les résultats de l'analyse bSLIM des deux souches isogéniques de *Saccharomyces cerevisiae*. Cela est illustré par une étude de la représentation graphique des scores de significativité des valeurs quantitatives des protéines (Figure 42). Dans cette figure, les protéines qui sont sous-exprimées dans la souche BY4742 par rapport à la souche de référence S288c sont colorées en jaune. Celles sur-exprimées sont colorées en bleu. Les valeurs entre les bornes le long de la droite identité sont proches de celles du "bruit de fond" généré par les permutations des valeurs expérimentales (modèle aléatoire). Toutes les protéines dont les gènes codants sont supprimés dans BY4742 apparaissent comme les plus significativement diminuées (voire absentes). De plus, les taux de FDR pour chacune des valeurs quantitatives de ces protéines d'intérêt sont très bas, exprimant ainsi une confiance forte dans les résultats observés. Cela est une démonstration de la sensibilité et de la spécificité de la méthode de quantitative.

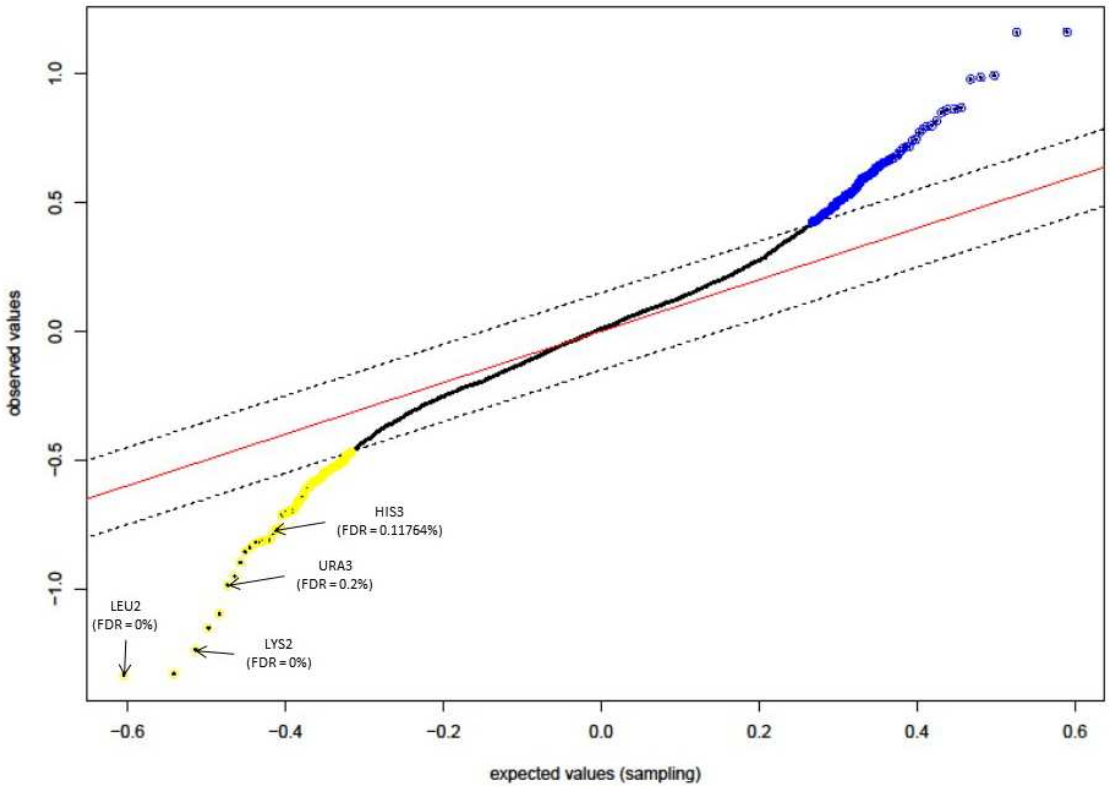


Figure 42 : Résultats expérimentaux de la comparaison entre les données expérimentales et théoriques (issues de mélanges aléatoires) démontrant la significativité des valeurs quantitatives.

Tout récemment un chapitre du livre Yeast Functional Genomics de la série « Methods in Molecular Biology » a été publiée et dont je suis le premier auteur.

Chapter 16

Quantitative Proteomics in Yeast: From bSLIM and Proteome Discoverer Outputs to Graphical Assessment of the Significance of Protein Quantification Scores

Nicolas Sénécaut, Pierre Poulain, Laurent Lignières, Samuel Terrier, Véronique Legros, Guillaume Chevreux, Gaëlle Lelandais, and Jean-Michel Camadro

Abstract

Simple light isotope metabolic labeling (bSLIM) is an innovative method to accurately quantify differences in protein abundance at the proteome level in standard bottom-up experiments. The quantification process requires computation of the ratio of intensity of several isotopologs in the isotopic cluster of every identified peptide. Thus, appropriate bioinformatic workflows are required to extract the signals from the instrument files and calculate the required ratio to infer peptide/protein abundance. In a previous study (Sénécaut et al., J Proteome Res 20:1476–1487, 2021), we developed original open-source workflows based on OpenMS nodes implemented in a KNIME working environment. Here, we extend the use of the bSLIM labeling strategy in quantitative proteomics by presenting an alternative procedure to extract isotopolog intensities and process them by taking advantage of new functionalities integrated into the Minora node of Proteome Discoverer 2.4 software. We also present a graphical strategy to evaluate the statistical robustness of protein quantification scores and calculate the associated false discovery rates (FDR). We validated these approaches in a case study in which we compared the differences between the proteomes of two closely related yeast strains.

Key words Mass spectrometry, Proteomics, Quantification, Isotopic labeling, Yeast, Metabolism, bSLIM, Proteome Discoverer, Minora, KNIME

1 Introduction

The yeast *Saccharomyces cerevisiae* is a biological model that is widely used for the development and validation of global analytical methods in functional genomics and genetics. Yeast has been extensively studied for many years, resulting in a solid understanding of its physiology and metabolism. Yeast is the first eukaryotic

Gaëlle Lelandais and Jean-Michel Camadro are co-senior authors.

Frédéric Devaux (ed.), *Yeast Functional Genomics: Methods and Protocols*,
Methods in Molecular Biology, vol. 2477, https://doi.org/10.1007/978-1-0716-2257-5_16, © The Author(s) 2022

organism for which the genome was fully sequenced [1]. This has opened up new avenues for the exploration of living organisms, notably through the analysis of gene and protein expression using the amazing recent technical developments in transcriptomics and proteomics. In-depth knowledge of the yeast genome has enabled the construction of complete collections of haploid or diploid strains carrying modified alleles, for example, disruptions, deletions, and ORF or promoter fusions with a large number of reporter genes for use as probes to assess gene function and associated regulatory networks, laying the foundation for systems biology.

One specific aspect of yeast is its ability to grow in the presence (aerobiosis) or absence (anaerobiosis) of molecular oxygen. This is made possible by a metabolic switch that allows passage from respiratory to fermentative metabolism, provided that the carbon source available to the yeast can be metabolized by fermentation. This requires processes in which mitochondrial functions are essential, making yeast a critical organism for deciphering the genetics, biochemistry, and physiology of this energy producing organelle.

Yeast cells are capable of growing on synthetic media in which the vitamins and essential trace elements are provided or on complex media containing yeast hydrolysates or peptones. In wild-type yeast grown on synthetic media, cell metabolism is based on the assimilation of organic nitrogen (usually ammonium sulphate or chloride) and the catabolism of a single carbon source, such as glucose, glycerol, or acetate. The genetics of yeast has been brought to light through the selection and use of mutants affected in the synthesis of certain nucleotides (e.g., Ura3) or amino acids (e.g., Lys, His, Leu, Met, Arg, Trp). Such auxotrophic mutants require the addition of the defective bases and/or amino acids in the synthetic growth media.

Beyond the identification of proteins in complex extracts, mass spectrometry-based proteomic analysis allows the quantification of differences in the proteome between several biological states. Several bottom-up quantitative proteomics approaches have been reported [2], providing critical information in yeast biology. They are based either on in vitro labeling of peptides by isobaric chemical probes, releasing fragments in MS/MS, of which the measured intensity reflects the abundance of the protein in the initial extract (e.g., ICAT or TMT labeling), or on differential metabolic labeling in vivo, in which the cells are cultured in the presence of "light" (unlabeled) or "heavy" (labeled) amino acids that will be incorporated into the proteins, allowing their quantification after tryptic digestion and the measurement of a heavy–light ratio for each peptide/protein (e.g., SILAC and derived methods) (for a review, *see* [3]). Despite the fact that TMT- and SILAC-based quantitative proteomics allow multiplexing multiple samples in a single run, one of the most widely used proteomics approaches is

Table 1
Relative abundance of the stable isotopes of the elements found in proteins

$^{12}C = 98.93\%$	$^{13}C = 1.07\%$		
$^{1}H = 99.9885\%$	$^{2}D = 0.0115\%$		
$^{16}O = 99.757\%$	$^{17}O = 0.0373\%$	$^{18}O = 0.2057\%$	
$^{14}N = 99.632\%$	$^{15}N = 0.368\%$		
$^{32}S = 94.93\%$	$^{33}S = 0.76\%$	$^{34}S = 4.29\%$	$^{36}S = 0.02\%$
$^{31}P = 100\%$			

still the label-free approach, in which individual LC-MS/MS runs are compared and the intensity of the peptide ions in MS1 are measured [4] to determine differences in protein abundance.

Recently we presented an innovative quantification method, called simple light-isotope metabolic (SLIM) labeling [5]. The SLIM-labeling strategy uses the fundamental property of the living matter in which all the biomolecules are basically composed of carbon, nitrogen, hydrogen, oxygen, and sulfur, with several additional elements, such as phosphorus, selenium, or iodine. Most of these elements, except phosphorus, are naturally present in the form of several stable isotopes for which the abundance is fixed (Table 1). It is thus possible to infer their isotopic abundance in biomolecules, such as amino acids by solely taking into account their average elemental composition: $C_{4.9384}$ $H_{7.7583}$ $N_{1.3577}$ $O_{1.4773}$ $S_{0.0417}$ [6].

This has important consequences in terms of high-resolution MS-based peptide/protein analysis. Every peptide is measured as a series of ions (m/z) in an isotope cluster of similar charge (z) but with the mass ranging from the monoisotopic mass m_0, containing only the lightest isotopes of each element, to higher masses resulting from the statistical distribution of additional neutrons present in the stable isotopes (isotopologs). The intensities of the various isotopologs within an isotope cluster therefore depend on the elemental composition of the peptide and follows a Poisson distribution that can be accurately modeled using dedicated software, such as MIDAS [7]. The basic principle of SLIM labeling is to manipulate the elemental composition of proteins in vivo from the natural abundance of the isotopes of the atoms present in proteins (C, H, N, O, S, P), defining the "NC" (natural carbon) condition as the condition named "12C" in which the proteins are enriched in the light isotope of carbon (^{12}C) and, eventually, nitrogen (^{14}N) (referred as to "12C14N" condition). Considering the main routes for amino-acid biosynthesis in yeast (Fig. 1) [8], we hypothesized that providing yeast cells with U-[^{12}C]-glucose as the sole carbon source would result in the rapid synthesis of U-[^{12}C]-

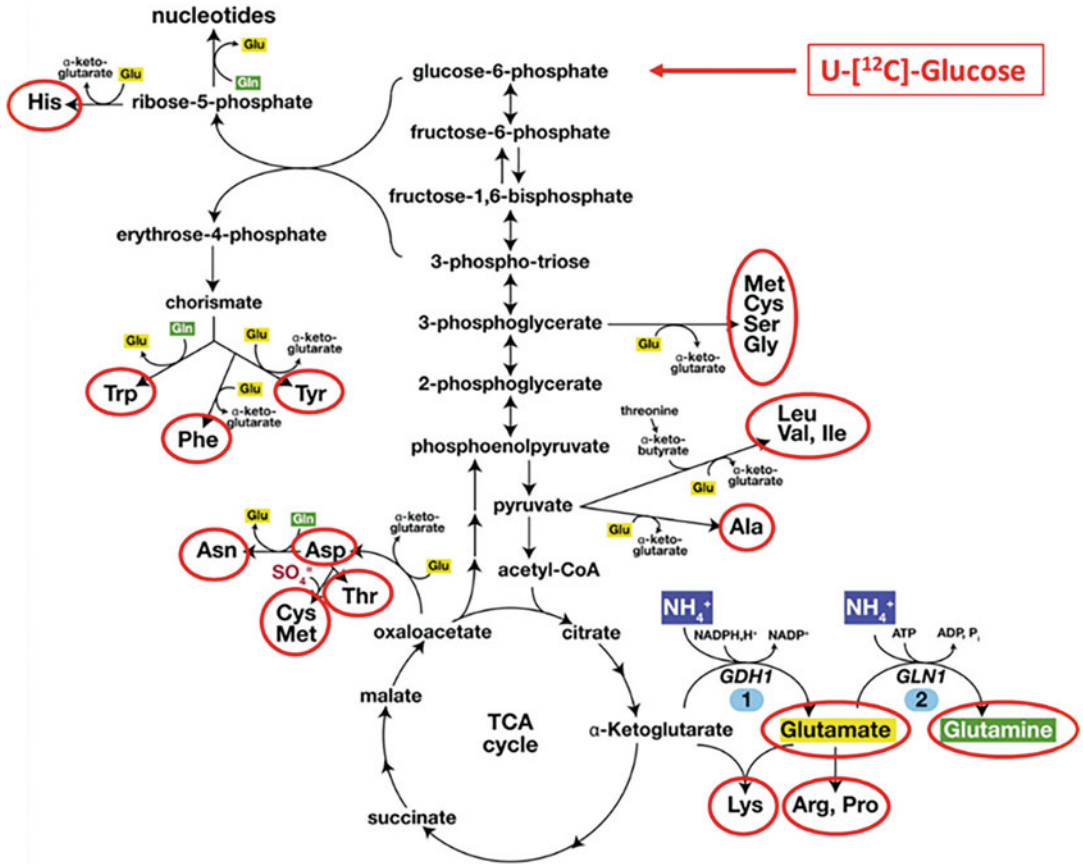


Fig. 1 Yeast Metabolism (adapted from Ljungdahl PO et al. YeastBook 2012 [5]

amino acids and their incorporation into newly synthesized proteins. Applying this labeling method allowed us to experimentally evaluate the half-life of the proteome in *Candida albicans*, and measure the effect of the proteasome inhibitor MG132 and a broad-specificity serine-protease inhibitor, PMSF, on the dynamics of the proteome in this organism [5].

Increasing the amino-acid content in ^{12}C (and to a lesser extent in ^{14}N) results in a different and simpler isotopic cluster that always remains within the boundaries of that observed with a natural isotopic composition, but with the intensity of the monoisotopic ion greatly enhanced. This has significant impact on downstream analyses, that is, allowing better signal-to-noise discrimination, more precise mass determination, and better MS/MS fragmentation spectra. As a result, higher scores for peptide identification and protein sequence coverage are obtained (*see* characteristic mass spectra in Fig. 2a–c). We took advantage of these characteristics to develop a new quantitative proteomics method in which peptides originating from the NC condition are mixed in equimolar amounts with peptides from the 12C condition. The intensity of

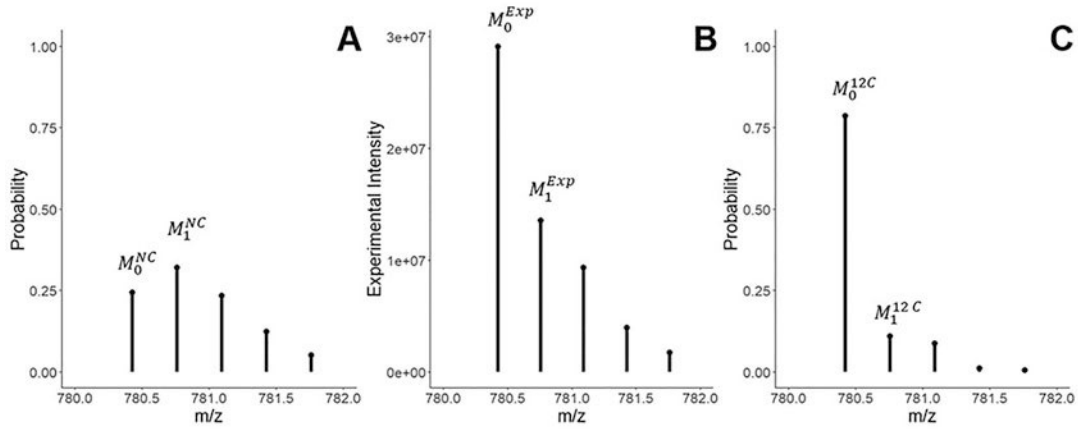


Fig. 2 Theoretical spectra of the peptide DQPILFWGGATAVGQMLIQLAK from the protein YNL134c under NC (**a**) and 12C (**c**) conditions and the corresponding experimental spectrum (**b**) of the same peptide from a 1:1 mixture of total extract from the 12C and NC conditions

every isotopolog in any isotope cluster is thus the sum of the intensity of the isotopologs from each condition (Fig. 2b). Therefore, measuring the ratio between the experimental values of the monoisotopic ion (M_0) and the next ion containing one more neutron (M_1), *modulo* the values of their theoretical intensity, expressed as the probability of occurrence, in each condition allows calculation of the molar fraction of the peptide originating from the NC- and 12C conditions. We recently described the full formalism for the quantification of ^{12}C incorporation into proteins/peptides and its use in quantitative proteomics, and we developed the data processing tools required to smoothly run SLIM labeling experiments [9].

One critical step in the SLIM-labeling quantification procedure is the accurate extraction of the intensities for all isotopologs in each isotope cluster from the experimental spectra. In our initial study, we used commercial software, Progenesis QI for metabolomics, but it does not provide the possibility to automatically link the quantification files with the identification files [5]. This prompted us to develop another workflow, referred to as bSLIM [9], in which only the intensities of the identified peptides are used to extract the data using an OpenMS node, FeatureFinderIdentification [9], which was modified to fit every mass trace in the isotope cluster. This approach only required us to install and run the KNIME (Konstanz Information Miner) environment for computation [9], together with the latest versions of OpenMS [10–12], and hence is fully independent of any commercial software.

Here, we present an alternative integrated procedure that takes advantage of the tools available in the proprietary software suite Thermo Scientific Proteome Discoverer. Proteome Discoverer (PD v2.4) is a popular program for the analysis of peptide-centric

proteomics data, with a high level of integration with Thermo Fisher Scientific high-resolution, high-sensitivity Orbitrap instruments. This analytical platform includes many algorithms developed by Thermo or third parties, such as the IMP Protein Chemistry facility (https://pd-nodes.org/) [13, 14] and others. Proteome Discoverer is therefore widely used on a routine basis in many MS-based proteomics laboratories. It associates and integrates both raw spectra processing and filtering, peptide identification through database searches in data-dependent analyses, diverse quantification routines, and convenient spectra viewers. The output of the Proteome Discoverer analyses is written in ".msf" or ".pdresult" files. A key feature of ".pdresult" files is that they are SQLite relational databases (https://www.sqlite.org) that can be queried using SQL. The possibility to visualize individual annotated spectra for peptides up to the level of their isotope clusters with their associated intensity prompted us to develop the appropriate tools to extract this data and use it as input for our bSLIM labeling quantitative proteomics strategy. We accessed the individual mass trace intensities by taking advantage of the capabilities of the newly developed node, Minora, initially designed for label-free quantification.

We also present a solution to assess the robustness of the protein quantification scores calculated using bSLIM which was missing from our previous data analyses workflow. Derived from the SAM (Significance Analysis of Microarrays [15]) method, the general idea is to randomize the original bSLIM output data sets multiple times and calculate the associated "random scores." These scores are graphically compared to the "real scores" obtained from the original bSLIM data. Proteins for which the real scores vary the most from the random scores are thus easily detected and worth considering for further analyses. In this chapter, we present a case study to illustrate the different outputs from the various workflows that we developed. We compared differences between the proteomes of two "wild-type" *Saccharomyces cerevisiae* strains with the same genetic background, but with one strain (BY4742) harboring the deletion of four genes (Ura3, His3, Leu2, and Lys2) relative to the reference strain S288c (*see* **Note 1** for data availability). The proteomes of the two strains are expected to be very similar and therefore represent a challenging test to assess the sensitivity and specificity of our quantification methods.

Overall, we expect that these alternative solutions implemented in the bSLIM data analysis workflow will be useful for proteomics laboratories running Orbitrap-based mass spectrometers, which are very familiar with Proteome Discoverer. This is an original way to combine the completeness and reproducibility of routine proprietary software with the power of open-source tools.

2 Materials

1. Reagents for yeast synthetic growth media and appropriate supplements.
2. SLIM-labeling specific reagent: U-[^{12}C]-glucose (e.g., Cambridge Isotope Laboratories).
3. Lysis buffer: 40 mM HEPES–KOH, pH 7.5, 350 mM NaCl, 10% glycerol, 0.1% Tween-20.
4. Acid-washed silica beads (0.45–0.5 mm Ø).
5. 200 μg/mL trypsin solution, prepared by dissolving 20 μg trypsin (Proteomic grade) in 100 μL of 1 mM HCl.
6. Cold Acetone.
7. 50 mM ammonium carbonate (NH_4HCO_3).
8. 0.1% formic acid (MS grade).
9. Dry incubator at 37 °C.
10. Vacuum dryer (Speed Vac).
11. Low-binding microcentrifuge tubes.
12. 4–12% polyacrylamide gradient gels.
13. Coomassie blue (MS friendly, such as SimplyBlue SafeStain, Invitrogen).
14. Bradford protein assay reagent.
15. An instrument setup for LC-MS/MSMS data acquisition (*see* **Note 2**).
16. Appropriate software suites for quantification and identification of the peptide/protein content of the samples analyzed (Fig. 3).

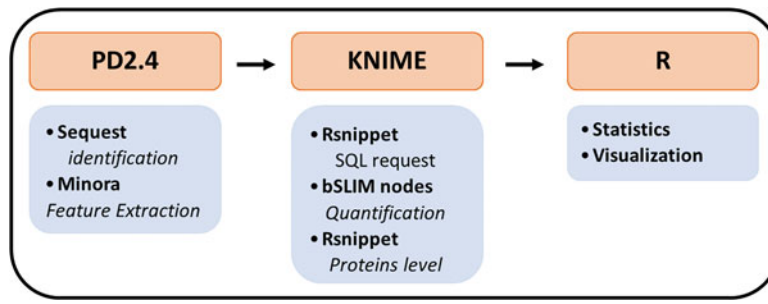


Fig. 3 General organization of the data processing workflows

3 Methods

3.1 Cell Growth and Preparation of Protein Extracts

1. Grow the cells to be compared in a synthetic medium with either regular glucose (NC-condition), or U-[^{12}C]-glucose (12C-condition) as the sole carbon source (*see* **Note 3**).
2. At the appropriate cell density (mid-exponential phase of growth), collect the cells by centrifugation for 10 min at 4000 × *g* at 4 °C.
3. Wash the cell pellet with cold water, resuspend the cells in lysis buffer at a cell density of 0.6 g/mL, and lyse the cells by adding 0.32 mL acid-washed, heat-sterilized silica glass beads (0.45–0.5 mm∅) to 0.6 mL cell suspension and vortexing the resulting suspension three times for 5 min, leaving the tubes on ice for 5 min between each vortexing.
4. Centrifuge the lysed cells for 5 min at 3000 × *g* and collect the supernatant, referred to as the cell homogenate.
5. Carefully measure the protein concentration using the Bradford Protein microassay and validate the protein measurement by running an aliquot on an SDS-PAGE gel and staining with Coomassie Blue.
6. Precipitate a 50-μg protein aliquot using 6 vol. cold acetone for 2.5 h at −20 °C.
7. Resuspend the dry protein pellet in 50 mM ammonium carbonate buffer by heating for 15 min at 95 °C.
8. Add 5 μL of 200 μg/mL trypsin stock solution and incubate for 12 h at 37 °C in a dry incubator.
9. Remove all solvents by vacuum drying.
10. Resuspend the peptides in 0.1% formic acid.
11. Carefully mix an equal amount of peptides from the NC- and 12C-conditions.
12. Inject the samples, typically 5 μg in ≤5 μL, into the LC-MSMS instrumental setup (*see* **Note 4**).
13. Ensure that your instrumental setup allows the isotopic resolution of all the peptides analyzed (*see* **Note 5**).
14. Save the ".raw" files for data processing and signal extraction.
15. Create a folder to gather all the ".raw" files from one project together.

3.2 Data Processing Workflows

1. Install the appropriate computational resources.
   (a) Proteome Discoverer 2.4 or higher with a valid activation key.

(b) KNIME (v4.2.3) with all available extensions (https://www.knime.com/downloads): the OpenMS nodes (v2.6.0) are part of the "community nodes."

(c) R (v4.0.2), including the dplyr, dbplyr, RSQLite, sqldf, readr, raster, RMySQL packages and libraries (https://cran.r-project.org/bin/windows/).

3.2.1 Proteome Discoverer Analysis

1. Open Proteome Discoverer 2.4.
2. Create a new study and add your Thermo Fisher Scientific mass spectrometry *.raw* files.
3. Select the appropriate "processing." The basic processing workflow is composed of the following nodes: spectrum files, spectrum selector, sequestHT (1.1.0.189), Percolator (3.02.1), and IMP-ptmRS. To this Processing workflow, add the "Minora Feature Detector" node linked to the "Spectrum Files" node and set the correct advanced parameters) (*see* **Note 6**).
4. Select the appropriate "consensus" workflow composed of the following nodes: MSF Files, PSM Grouper, Peptide Validator, Peptide and Protein Filter with a link to Protein annotation, Protein Scorer with a link to Protein FDR Validator, and Protein Grouping (*see* **Note 7**).
5. Enable the postprocessing node "Display Settings."
6. Run the analysis, **one file at the time**, by giving nonambiguous names to the output files. The produced results files from PD2.4 have the extension .pdresult and are used as input in our KNIME workflow.

3.2.2 Isotopolog Intensity Extraction and Peptide/Protein Quantification Using a Dedicated bSLIM KNIME Workflow (Fig. 4)

1. Open KNIME 4.2.3.
2. Import from https://zenodo.org (DOI 10.5281/zenodo.4467829), the bSLIM quantification workflow "*File > Import KNIME workflow*" (file extension is ".knwf"). The workflow is a modification of the original workflow presented in our previous study [9]. The adaption is very simple (disconnection between metanode 1.2 "FFiDData filtering" and connection with the Row filter "erase ModifiedPeptides").

   This workflow contains three main parts.

   (a) The ".pdresult" file, which extracts all data concerning every peptide, including their identification and the intensity of the isotopologues in the isotope clusters. This procedure uses an R script written specifically for this study. It is embedded in a dedicated RSnippet as a part of the KNIME quantification workflow (*see* **Note 8** on SQL request formalism).

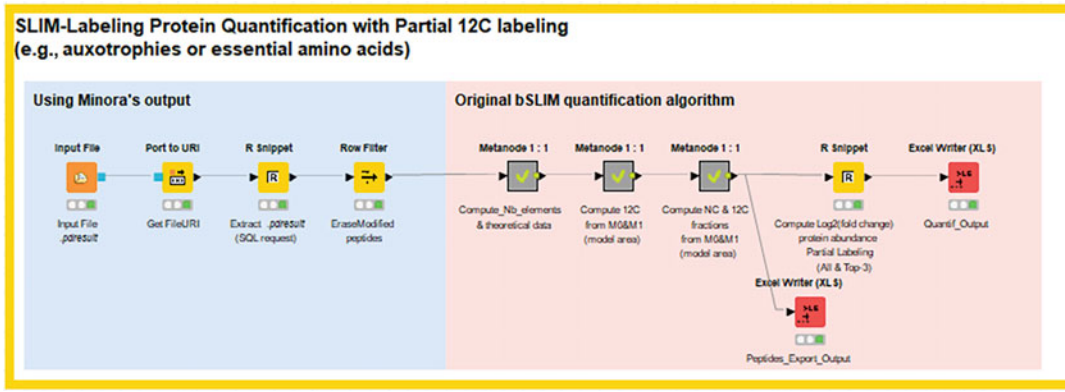


Fig. 4 KNIME workflow of the integrated ".pdresult" processing node connected to the quantification nodes

(b) The peptide/protein quantification workflow based on our previous study [9]. As previously described, two biological conditions are considered: complete labeling of the proteins (true wild-type strain) or partial labeling of the proteins (strains that are auxotrophic for specific amino acids).

(c) The procedure to compute the statistics on the identified peptides/proteins. These two procedures use R scripts embedded in specific RSnippets.

3. Check that your installation of Rserver allows all RSnippets to run smoothly.

4. There are two possible cases:

(a) The samples come from an autotrophic yeast strain and the SLIM labeling is complete: follow "*cases of total labeling experiment.*"

(b) The samples come from auxotrophic cells for which certain amino acids are not labeled. The SLIM labeling is thus incomplete: follow "*case of incomplete labeling experiment.*" In the latter case, the essential amino acids are defined as containing "Carbon B," corresponding to carbon atoms of natural isotopic abundance. It is therefore necessary to open the meta-node "Compute_Nb_elements& theoretical data" and then the meta-node "Compute Elemental Composition" and, finally, the carbon B calculus node (Compute Nb_Carbon-B (Ex: HKL)) to add the correct total number of carbons to each exogenous amino acid by typing the regular expression in the form "(Nb_H*6)+ (Nb_K*6)+ (Nb_L*6)", as exemplified for the BY4742 strain auxotrophic for histidine (H), lysine (K), and leucine (L). The number of carbon atoms for each amino acid is shown in Table 2.

Table 2
Number of carbon atoms per amino acid

Amino acid	A	C	D	E	F	G	H	I	K	L
Nb carbon atoms	3	3	4	5	9	2	6	6	6	6
Amino acid	M	N	P	Q	R	S	T	V	W	Y
Nb carbon atoms	5	4	5	5	6	3	4	5	11	9

5. Set the Excel exporter nodes with an appropriate name for file output (scores for quantifications at proteins or peptide levels) (*see* **Note 9**).

3.2.3 Statistics and Graphical Assessment of Score Significance

For statistical analysis of differential expression, we reproduced the SAM methodology, adapting it to the specifics of the quantitative measurements of protein abundance that are obtained at the end of the bSLIM workflow (Fig. 5) (*see* **Note 10**). Workflows available at https://zenodo.org (DOI 10.5281/zenodo.4467882).

1. Aggregate the protein quantification results from individual experiments (replicates) into a single ".tsv" table: Accession/ "name of column-2"/ "name of column-3"/ . . . / "name of column-n." Typically, column-2 to –n represents the log2(Fold change) of protein abundance per experimental condition.
2. Load the R scripts developed in this study to compute the scoring functions, and save the analysis.
3. Use the graphical package ggplot2, within the script, to produce the figures showing differentially expressed proteins.
4. Retrieve the table produced by the script containing the proteins for which the over- or underexpression is statistically significant between the different experimental conditions.

3.3 Case Study: BY4742 Vs S288c Proteome Comparison

To test and illustrate the different outputs from the various workflows, we compared the differences between the proteomes of two "wild-type" *Saccharomyces cerevisiae* strains with the same genetic background but with one strain (BY4742) harboring the deletion of four genes (Ura3, His3, Leu2, and Lys2) versus the reference strain S288c. The proteomes of the two strains are expected to be very similar (*see* **Note 3** for experimental details).

As shown in Fig. 6, the graphical representation of the distribution of quantification quality scores shows the efficiency of the workflows to identify proteins that are underexpressed (yellow) or overexpressed (blue) in the laboratory wild-type strain BY4742 relative to the reference strain S288c. All the proteins encoded by the genes that are deleted in BY4742 appear as the most significantly diminished (indeed absent) in BY4742, showing the sensitivity and specificity of the proposed quantification methods.

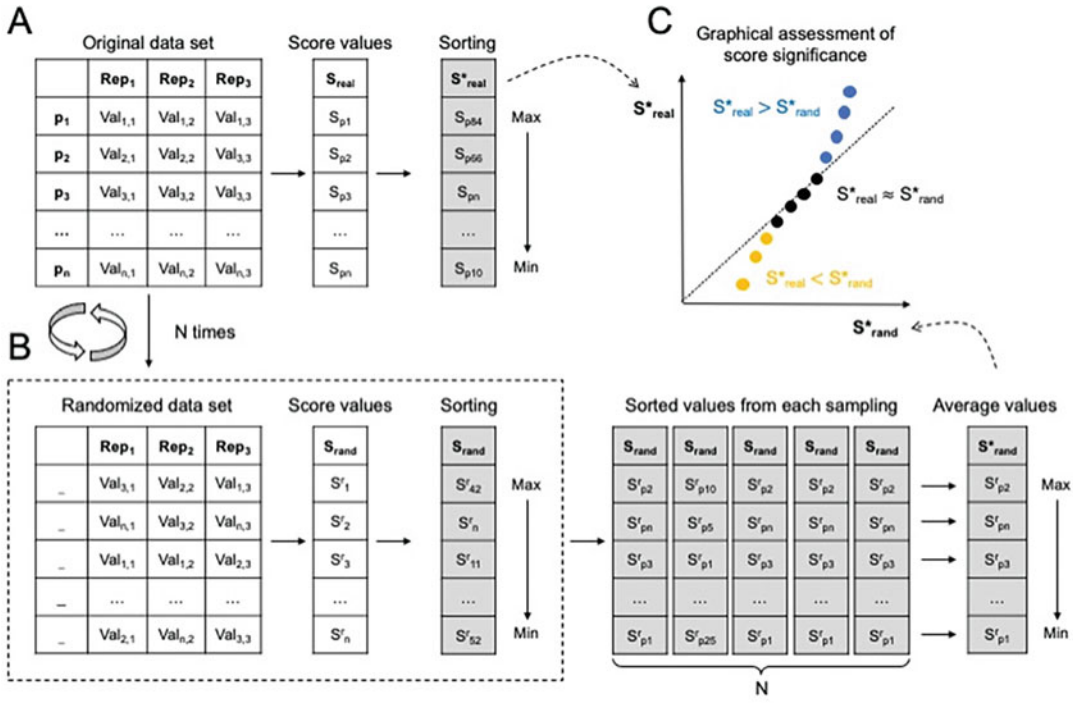


Fig. 5 Schematic representation of the methodology used to assess the relevance of the bSLIM results exported for each protein. (**a**) The original bSLIM output data set is organized as a table in which the proteins are presented in rows and the repetitions of the experiments in columns. For all proteins, score values are calculated and are next sorted from the highest to the lowest. (**b**) Random data set is created from the original dataset, randomly sampling the values in each column. New scores are next calculated and sorted from the highest to the lowest. This process is repeated N times, resulting in a final table with N columns, comprised of the sorted values from each sampling. The average values of all sorted scores are finally calculated. Note that the maximal average value is derived from calculation of the mean between the maximal score values obtained in each sampling. (**c**) The significance of the scores obtained from the original data set are graphically assessed by plotting the S^*real values (**a**) against the S^*rand values (**b**). Significant scores are those that are higher (colored in blue) or lower (colored in yellow) than the random scores. False discovery rates are finally calculated, comparing for each protein score, the average number of other scores in the table of sorted values (from each sampling, *see* (**b**)) that are higher (respectively lower) than the number of scores that are higher (respectively lower) in the original dataset. This is the same method as detailed in [15]

4 Notes

1. The original data sets are publicly available in the ProteomeXchange platform under the Pride submission number PXD021329.
2. The instrumental setup in our laboratory consists of Orbitrap Fusion Tribrid ETD and Orbitrap Q-Exactive Plus mass spectrometers, equipped with Easy-Spray nanoelectrospray ion sources. The LC setup consists of Easy nano-LC Proxeon 1000 or 1200 systems equipped with an Acclaim PepMap100

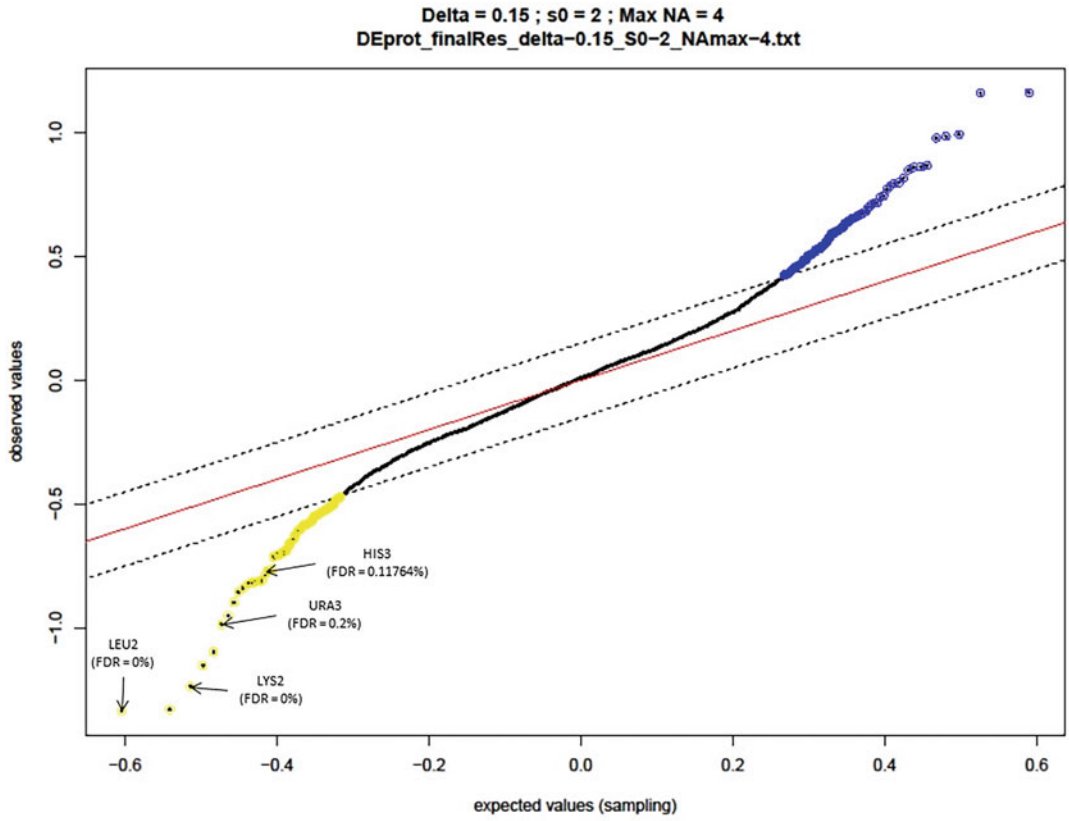


Fig. 6 Experimental distribution of the statistical distribution of the protein quantification quality scores in the characterization of the differences between the BY4742 and S288c proteomes

C18 precolumn and a Pepmap-RSLC Proxeon C18 column. These devices are all from Thermo Fisher Scientific (Bremen, Germany and San Jose, CA, USA).

3. In the case study presented here, the *S. cerevisiae* strains S288c (MATα SUC2 gal2 mal2 mel flo1 flo8-1 hap1 ho bio1 bio6) [16] and its isogenic derivative BY4742 (MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0) are grown on a synthetic medium made of 6.7 g/L Yeast Nitrogen Base (YNB) with ammonium sulfate, without amino acids, with 0.5% glucose as the sole carbon source. The auxotrophy of the BY4742 strain is complemented with uracil (20 mg/L), histidine (20 mg/L), and leucine and lysine (both 30 mg/L). The carbon source was either regular D(+)-glucose anhydrous, defining the normal condition (NC condition), or U-[^{12}C]-glucose (Cambridge Isotope Laboratories), defining the 12C condition. The 10% glucose stock solutions were filter-sterilized.
4. Liquid chromatography coupled to mass spectrometry data acquisition:

In the case study presented here, the chromatographic separation of peptides was performed using the following parameters: Acclaim PepMap100 C18 precolumn (2 cm, 75 µm i.d., 3 µm, 100 Å), Pepmap-RSLC Proxeon C18 column (75 cm, 75 µm i.d., 2 µm, 100 Å), and 300 nL/min flow. The chromatographic separation of peptides was obtained with a gradient consisting of 95% solvent A (water, 0.1% formic acid) to 35% solvent B (99.9% acetonitrile, 0.1% formic acid) in 90 min, followed by column regeneration for 15 min, giving a total run time of 1 h and 45 min.

5. Peptides masses were analyzed in the Orbitrap cell in full ion scan mode at a resolution of 70,000 with a mass range of m/z 375–1500 and an AGC target of 3.10^6. MS/MS were performed in a Top 20 DDA mode. Peptides were selected for fragmentation by Higher-energy C-trap Dissociation (HCD) with a Normalized Collisional Energy of 27%, and a dynamic exclusion of 30 s. Fragment masses were measured in the Orbitrap cell at a resolution of 17,500, with an AGC target of 2.10^5. Monocharged peptides and unassigned charge states were excluded from the MS/MS acquisition. The maximum ion accumulation times were set to 50 ms for MS and 45 ms for MS/MS acquisitions respectively.
6. All MS/MS data are processed using the SequestHT (v1.1.0.189) node. The mass tolerance is set to 6 ppm for precursor ions and 0.02 Da for fragments when using an Orbitrap Q-Exactive Plus mass spectrometer. The following alterations are used for various modifications: carbamidomethylation (C), if the sample is reduced and alkylated, and oxidation (M). Phosphorylation (STY) and acetylation (K, N-term) are generally added for additional analyses of trypsin digests. The maximum number of missed cleavages by trypsin is limited to two. MS/MS data are searched against the Uniprot *Saccharomyces cerevisiae* reference proteome UP000002311 (https://www.uniprot.org/proteomes/UP000002311, 6049 protein counts).
7. The Consensus workflow is very basic, because using the Minora node, as presented here, strictly requires that only one results file is processed per run (Fig. 7: Proteome Discoverer 2.4 consensus workflow for presentation).
   (a) The R snippet uses SQL query to link the table of identification with the isotopic intensities. The data are then incorporated in the KNIME workflow.
   (b) The computation is rapid and can be performed as a side analysis during the bSLIM experiment. In cases of auxotrophy, only the amino acids synthesized by the yeast are labeled, whereas the exogenous amino acids that need to

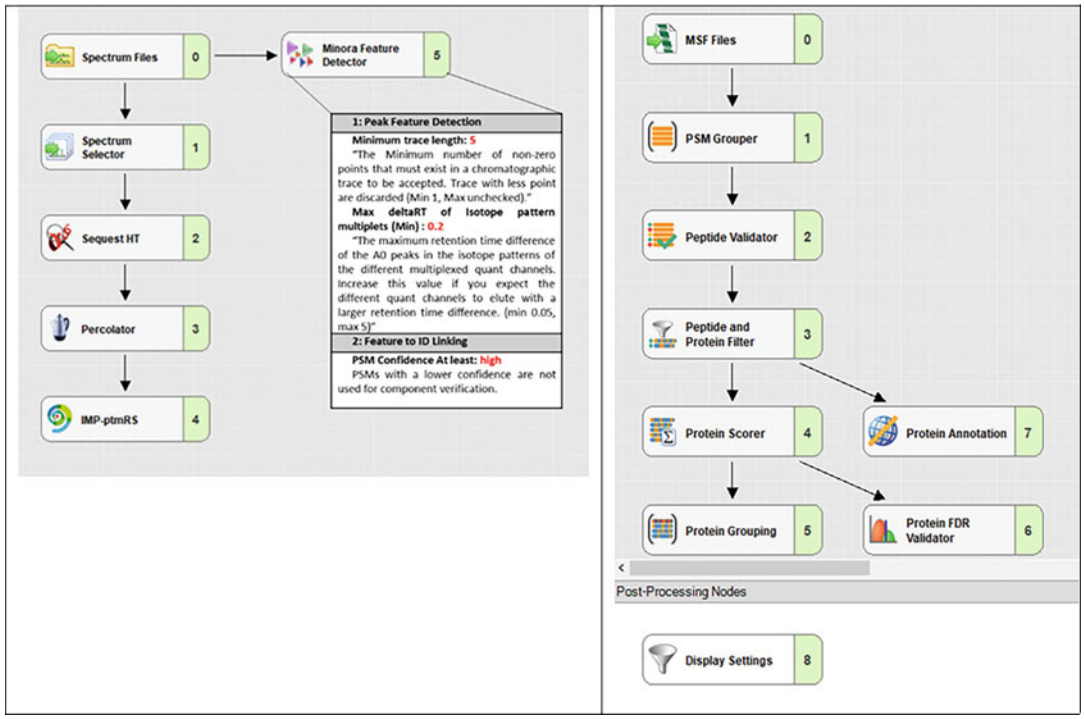


Fig. 7 Details of Proteome Discoverer 2.4 processing (left panel) and consensus (right panel), including the specific Minora node parameters to extract the isotopolog intensities for every isotope cluster

be added to the media are not, resulting in mixed labeling. Quantification is possible with the introduction of a new calculation to accommodate this type of analysis. Experimental data were analyzed using BY4742 auxotrophic yeast.

8. For each identified peptide, we extracted the following items, which are combined in a single output table: FeatureId / MassOverCharge / ParentProteinAccessions / ParentProteinDescriptions / MasterScanNumbers / RetentionTime / Charge / Sequence / Modifications / MonoisotopicMassOverCharge / Area / Intensity / NumberOfIsotopes / MassOverChargeIsotope / PeakHeight (M_0) / PeakArea (M_0) / PeakHeight (M_1) / PeakArea (M_1) / … / PeakHeight (M_4) / PeakArea (M_4).
   (a) The ".pdresult" file is an SQLite relational database and the information can be accessed using SQL queries. The database contains all PD search parameters, variables, and results. In our quantification workflow, we only need to access three tables that contain the relevant information:
      - *TargetPsms*, containing the peptide IDs.
      - *LcmsFeatures*, with the description of the MS1 cluster used for the identification.

- *LcmsPeaks*, which is the most important in this study, because it contains the abundance of each individual isotopolog contained in each independent identified isotopic cluster.

(b) To extract and produce the final output table from the database (".pdresult"), we created an original workflow to:

- Import the ".pdresults" file.
- Create explicit path names to access the requested data, using the Uniform Resource Identifier (URI) node of KNIME.
- Embed the SQL commands into an RSnippet to retrieve the expected data and create two tables (Feature_Data and Peak_Data in the R code). These tables are further joined using a connection link between them as an intermediate table used as a "dictionary" of ID equivalents.
- Define the ordered rank of the isotopologs extracted by sequential numbering of the lines related to each PSM (Protein Spectrum Match).

(c) Within the R script, the complete SQL request for generating Table Feature_Data is as follows:

```
 select TargetPsms.MassOverCharge, TargetPsms.ParentProteinAccessions, TargetPsms.ParentProteinDescriptions, TargetPsms.MasterScanNumbers, TargetPsms.RetentionTime, TargetPsms.Charge, TargetPsms.Sequence, TargetPsms.Modifications, LcmsFeatures.MonoisotopicMassOverCharge, LcmsFeatures.Area, LcmsFeatures.Intensity, LcmsFeatures.NumberOfIsotopes, LcmsFeatures.Id as FeatureId
 from TargetPsms, TargetPsmsLcmsFeatures, LcmsFeatures
 where TargetPsms.PeptideID = TargetPsmsLcmsFeatures.TargetPsmsPeptideID
 and TargetPsmsLcmsFeatures.LcmsFeaturesId = LcmsFeatures.Id
```

(d) Within the R script, the complete SQL request for generating Table Peak_Data is as follows:

```
 select LcmsFeatures.Id as FeatureId, LcmsPeaks.MassOverCharge, LcmsPeaks.PeakHeight, LcmsPeaks.PeakArea
 from LcmsFeatures, LcmsFeaturesLcmsPeaks, LcmsPeaks
 where LcmsFeaturesLcmsPeaks.LcmsFeaturesId = LcmsFeatures.Id
 and LcmsFeaturesLcmsPeaks.LcmsPeaksId = LcmsPeaks.Id
```

(e) The intensity of the isotopolog ions is defined by the peak area.

(f) We restrict the number of isotopologs during extraction in the final dataset to five, as we only require M_0 and M_1 for further quantification calculations.

9. The produced table is used in the bSLIM workflow for complete or incomplete labeling with the correct exogenous (non-labeled) amino acids given in parameters. The code proceeds to the ratio of M_1 over M_0 to quantify the molar fraction, the key variable for the quantification. The ratio of the molar fraction/(1-molar fraction) is calculated. For protein levels, all top N peptides log2(Ratio) are grouped together to obtain classical fold changes for biological interpretations.
10. A key question in proteomics data analysis is the distinction between noteworthy (or significant) results from other observations, which are false positives, that is, acquired by random chance. Indeed, the large amount of data arising from proteomics technologies is associated with an increase in the possibility to observe atypical "by chance" values in the dataset. In this context, statistical methodologies generally assume that all variations in the data are due to random fluctuations and, accordingly, derive a probability to observe variations that are greater than those present in the data. Random fluctuations can be modelled in two different ways. In the first, a mathematical function is chosen (often normal or student laws) and a statistical hypothesis is used to discriminate "significant" from "nonsignificant" observations, based on a predefined error rate (generally 5%). In the second, random permutations of the original dataset are performed to define empirical distributions, which will be used to assess potential random fluctuations. It is a remarkably interesting approach, especially when the theoretical probability distributions of the studied parameters are not demonstrated, as is the case with the bSLIM output dataset.

Acknowledgments

We would like to thank Dr. Bernard Delanghe (Thermo Fisher Scientific) for providing us with the table structure of the ".pdresult" file that allowed us to develop the present workflow. NS received a thesis grant from CRI-Paris. This work was supported by the ARN, grant ANR-18-CE44-0014. The English text was edited by Alex Edelman & Associates (http://www.alexedelman.com/).

References

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al (1996) Life with 6000 genes. Science 274:546. 563–567
2. Wilm M (2009) Quantitative proteomics in biological research. Proteomics 9:4590–4605
3. Chahrour O, Cobice D, Malone J (2015) Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. J Pharm Biomed Anal 113:2–20
4. Leger T, Garcia C, Videlier M, Camadro JM (2016) Label-free quantitative proteomics in yeast. Methods Mol Biol 1361:289–307
5. Leger T, Garcia C, Collomb L, Camadro JM (2017) A simple light isotope metabolic labeling (SLIM-labeling) strategy: a powerful tool to address the dynamics of proteome variations in vivo. Mol Cell Proteomics 16:2017–2031
6. Senko MW, Beu SC, McLafferty FW (1995) Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. J Am Soc Mass Spectrom 6:52–56
7. Alves G, Ogurtsov AY, Yu YK (2014) Molecular isotopic distribution analysis (MIDAs) with adjustable mass accuracy. J Am Soc Mass Spectrom 25:57–70
8. Ljungdahl PO, Daignan-Fornier B (2012) Regulation of amino acid, nucleotide, and phosphate metabolism in Saccharomyces cerevisiae. Genetics 190:885–929
9. Sénécaut N, Alves G, Weisser H, Lignieres L, Terrier S, Yang-Crosson L, Poulain P, Lelandais G, Yu YK, Camadro JM (2021) Novel insights into quantitative proteomics from an innovative bottom-up simple light isotope metabolic (bSLIM) Labeling data processing strategy. J Proteome Res 20:1476–1487
10. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K et al (2008) OpenMS - an open-source software framework for mass spectrometry. BMC Bioinform 9:163
11. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich HC, Gutenbrunner P, Kenar E et al (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods 13:741–748
12. Pfeuffer J, Sachsenberg T, Alka O, Walzer M, Fillbrunn A, Nilse L, Schilling O, Reinert K, Kohlbacher O (2017) OpenMS—a platform for reproducible analysis of mass spectrometry data. J Biotechnol 261:142–148
13. Doblmann J, Dusberger F, Imre R, Hudecz O, Stanek F, Mechtler K, Durnberger G (2019) apQuant: accurate label-free quantification by quality filtering. J Proteome Res 18:535–541
14. Griss J, Stanek F, Hudecz O, Durnberger G, Perez-Riverol Y, Vizcaino JA, Mechtler K (2019) Spectral clustering improves label-free quantification of low-abundant proteins. J Proteome Res 18:1477–1485
15. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98:5116–5121
16. Mortimer RK, Johnston JR (1986) Genealogy of principal strains of the yeast genetic stock center. Genetics 113:35–43

IV. Suivi de l'incorporation du ^{12}C dans des cellules eucaryotes supérieures et un organisme multicellulaire simple

Une des nouvelles applications de cette méthode de quantification est la culture puis le suivi de l'incorporation du ^{12}C dans des cellules eucaryotes complexes et un organisme multicellulaire simple. Dans ces contextes, la difficulté d'implémentation de la méthode réside essentiellement dans l'adaptation des protocoles classiques de composition des milieux de culture. Ainsi il faut limiter au maximum tout apport de sources de carbone ayant une composition naturelle exogène et donc métabolisable par les organismes étudiés. En effet, cela diluerait les carbones ^{12}C réduisant ainsi son enrichissement dans la condition marquée.

IV.1 Lignées cellulaires

Nous avons choisi les cellules de lignées cellulaires *HEK* d'une tumeur du rein. Une telle culture dans du milieu RPMI modifié implique l'ajout d'un mélange de 11 acides aminés. Ces acides aminés possèdent donc tout le long de l'expérience une abondance isotopique naturelle. Les deux conditions sont maintenues en cultures pendant 6 "passages", c'est à dire que les cellules, une fois arrivées à confluence, sont en partie seulement réimplantées dans du milieu frais (voir Figure 43). Durant cette étape, des cellules restantes sont collectées pour analyser l'incorporation de ^{12}C dans les protéines comme présenté en (Figure 44). Ce projet a été réalisé au laboratoire en collaboration avec l'équipe Ladoux/Mége de l'Institut Jacques Monod.

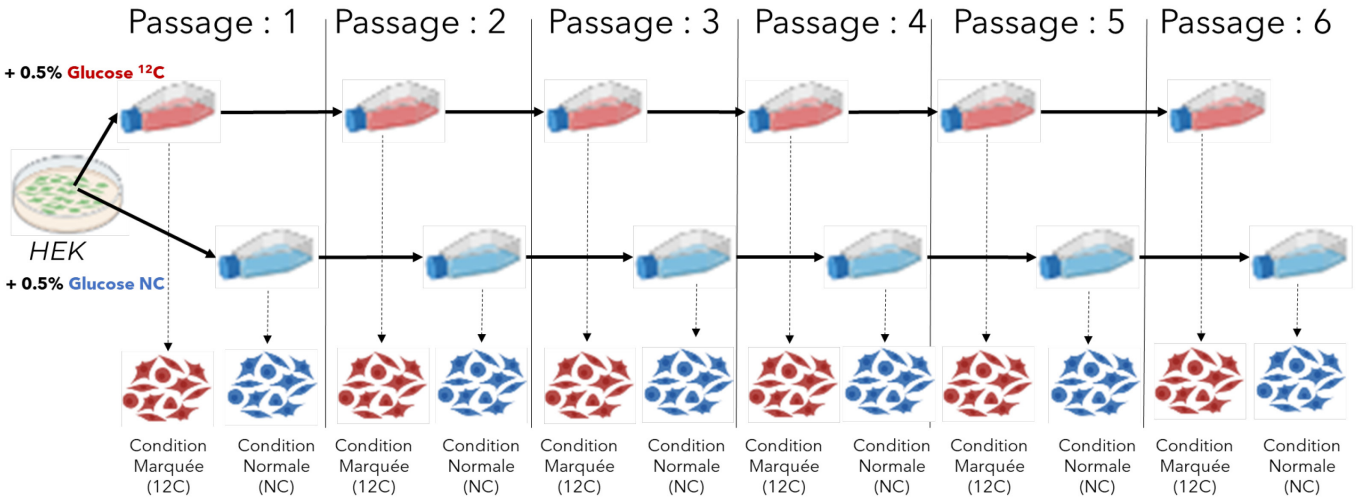


Figure 43 : Plan d'expérience du marquage bSLIM sur des lignées cellulaires humaines (HEK)

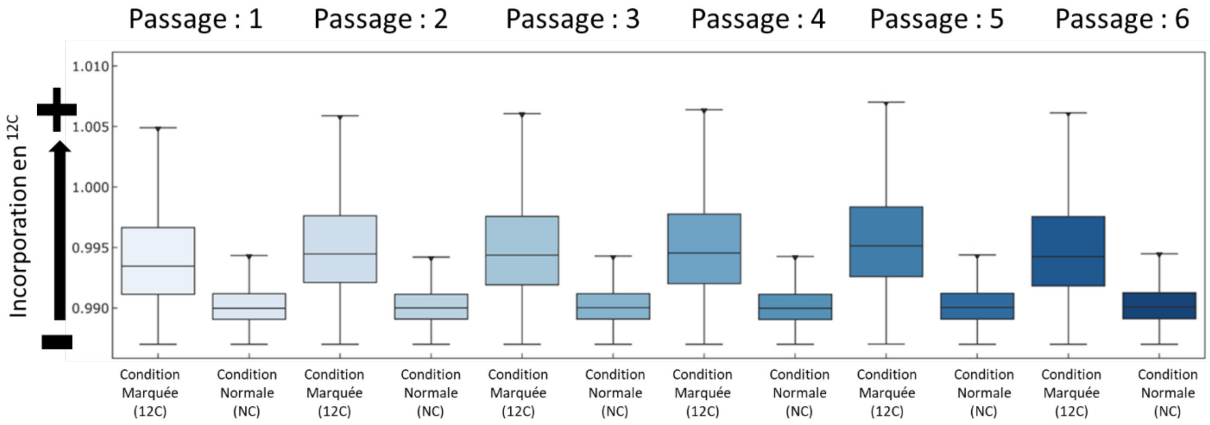


Figure 44 : Suivi de l'incorporation du ^{12}C dans les différentes conditions

IV.2 Organismes multicellulaires

Le choix d'un organisme pluricellulaire est soumis à la façon dont l'incorporation du ^{12}C pourra être maximale. Nous avons choisi pour le développement et la démonstration du concept bSLIM, le nématode *Caenorhabditis elegans*. Afin de réaliser un marquage des cellules du vers, nous utilisons la source de nourriture comme point d'entrée de la source de carbone ^{12}C, cela a été réalisé au laboratoire en collaboration avec l'équipe Dumont de l'Institut Jacques Monod. Dans un premier temps une culture bactérienne de deux conditions bSLIM est réalisée, les bactéries *Escherichia coli* de souche *OP50* incorporent ainsi du ^{12}C. Puis dans un deuxième

temps, les bactéries ainsi marquées sont récupérées et déposées dans une boite de pétri afin de servir de source de nutriment aux vers. Ainsi, par “effet secondaire”, les vers incorporent le ^{12}C de protéines issues de la digestion des bactéries (Figure 45).

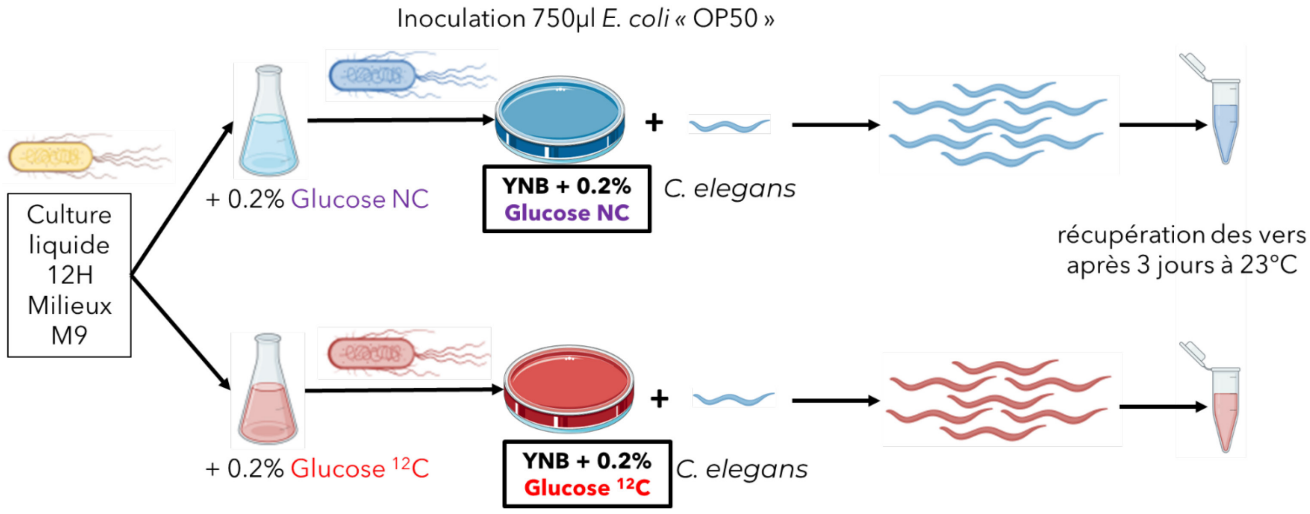


Figure 45 : Plan d’expérience du marquage bSLIM sur l’organisme multicellulaire C. elegans

L’incorporation du ^{12}C dans les protéines de *E. coli* et de *C. elegans* ont été mesurée et sont présenté en Figure 46 et Figure 47.

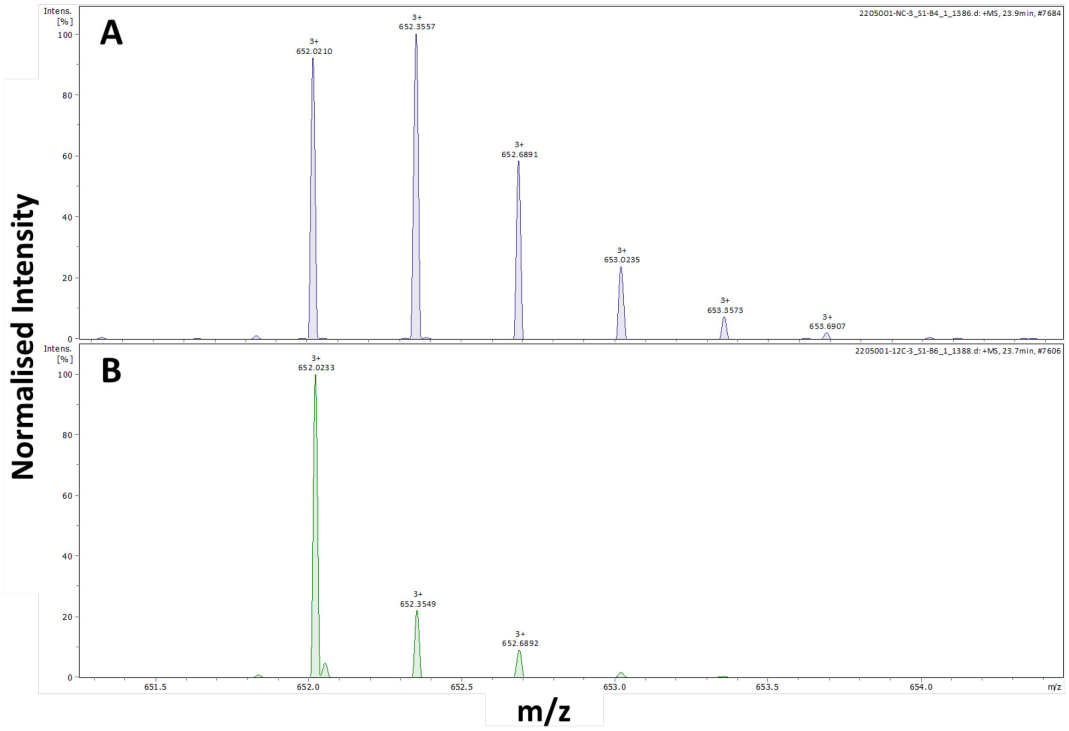


Figure 46 : Spectre de masse d’une famille d’ions provenant d’un digestat de lysat cellulaire de C. elegans. En condition NC en A et 12C en B. Cette analyse a été effectuée sur le nouvel appareil « tims-TOF Pro 2 ».

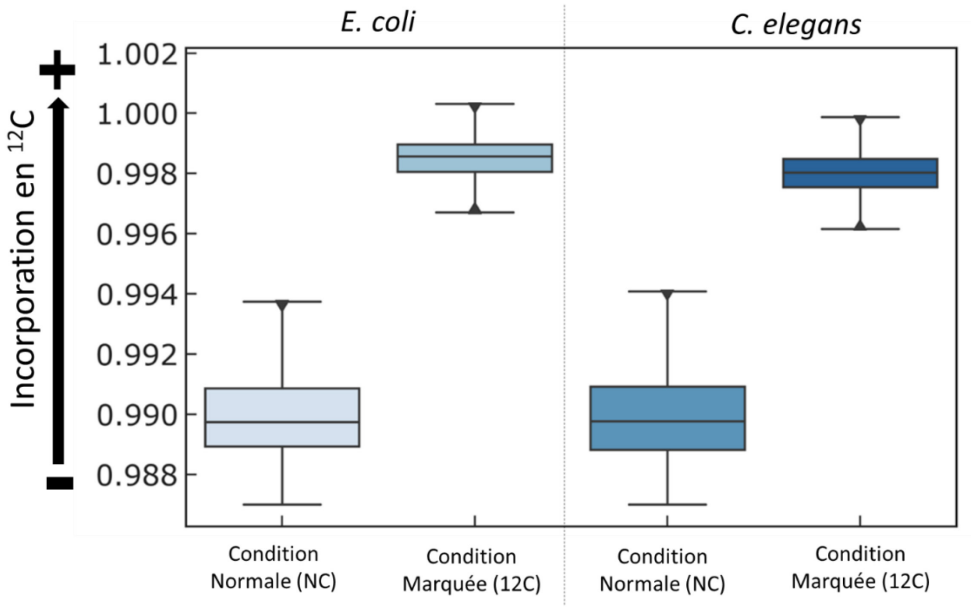


Figure 47 : Suivi de l'incorporation du ^{12}C dans les différentes conditions, dans les bactéries E. coli (souche OP50) et le nématode C. elegans

Ces deux expériences constituent une application concrète du bSLIM et ouvre la porte à des applications biologiques plus complexe.

Une publication présentant ces avancées novatrices du bSLIM est en cours d'écriture, dont je suis le co-premier auteur avec Laurent Lignières ingénieur sur la plateforme de spectrométrie de masse de l'Institut Jacques Monod.

Extending the range of SLIM-labeling application: from human cell line in culture to *Caenorhabitis elegans* whole organism labeling

Authors

Laurent LIGNIERES[1*], Nicolas SENECAUT[2*], Tien NGUYEN[3], Laura BELLUTTI[4], Samuel TERRIER[1], Véronique LEGROS[1], Guillaume CHEVREUX[1], Gaëlle LELANDAIS[5], René-Marc MÈGE[3], Julien DUMONT[4], Jean-Michel CAMADRO[1,2,$]

Affiliations

1-ProteoSeine@IJM, Paris, France
2-"Mitochondria, Metals and Oxidative stress" group, IJM Paris, France
3-"Cell Adhesion and Mechanics" group, IJM, Paris, France
4-"Cell Division and Reproduction" group, IJM, Paris, France
6-I2BC, Gif-sur-Yvette, France

* Contributed equally to this work

$ Corresponding author: jean-michel.camadro@ijm.fr

Abstract

The simple light isotope metabolic labeling technique relies on the *in vivo* biosynthesis of amino acids from U-[^{12}C]-labeled molecule provided as the sole carbon source. The incorporation of the resulting U-[^{12}C]-amino acids into proteins presents several key advantages for mass-spectrometry based proteomics analysis since it induces more intense monoisotopic ions, with better signal to noise ratio and higher identification scores and protein sequence coverage in bottom-up analysis. The initial studies using the SLIM-labeling strategy were performed on prototrophic eukaryotic microorganisms, the yeasts *Candida albicans* and *Saccharomyces cerevisiae*. Although in this later case, we used strains with genetic markers leading to some amino acids auxotrophies, we hypothesized that the SLIM-labeling strategy may apply to cells relying on numerous essential amino acids (amino acids the cell cannot synthetize) for their growth, provided they are able to synthesize at least some amino acids from exogenous source of carbon.

In order to extend the range of SLIM-Labeling applications, we evaluated the feasibility of the direct labeling incorporation into proteins of human cells in culture, that require a large number of essential amino-acids to support their growth, and an indirect labeling strategy in the worm *C. elegans*.

Partie 3: Développement de la méthode tSLIM

Avancées pour l'étude de protéines intactes par une stratégie « Top-down »

Chapitre 1 : Problématiques de l'approche Top-down au regard du Bottom-up

La stratégie Top-down consiste en l'étude de protéines intactes par spectrométrie de masse. Elle permet d'obtenir des informations pertinentes sur les structures des protéines et en particulier une observation fine des modifications post-traductionnelles. Cette approche s'oppose au Bottom-up où les protéines sont étudiées après digestions en molécules plus petites, les peptides. Pour permettre une étude complète des protéines, une haute couverture de séquence est requise, l'analyse exige donc que de nombreux peptides soient observés. Or, dans une expérience classique en Bottom-up, certaines zones de la séquence des protéines produisent des peptides atypiques : ils sont soit trop hydrophobes donc difficilement séparables par la chromatographie, soit sont trop petits ou trop grands pour être observés dans les spectres de masse. Cette zone "sombre" de couverture des séquences protéiques constitue donc une grande limitation à l'approche Bottom-up. De plus, les protocoles de préparation de l'échantillon (digestion, réduction et alkylation) réduisent fortement le maintien des protéoformes contenant des modifications post-traductionnelles. C'est la raison pour laquelle, dans l'objectif d'étudier de manière rigoureuse les protéines, la stratégie Top-down est la plus adaptée. Cependant la difficulté de cette stratégie d'analyse par spectrométrie de masse réside dans le fait d'étudier des protéines entières et donc de très haute masse (supérieur à 15kDa). Cela est d'autant plus critique que les méthodes séparatives de l'échantillons sont complexes mais également que l'optimisation de l'analyse en masse, en particulier de la fragmentation des protéines, n'est pas aisée.

Dans le cas de l'analyse des protéines intactes en Top-down, les protocoles de séparation chromatographique des protéines sont différents. En effet, les protéines intactes sont des objets massifs et ayant des propriétés physico-chimiques particulières. Le couplage LC–MS est une approches méthodologique importante car

cruciale dans l'analyse des protéines intactes. C'est d'autant plus critique que les colonnes ainsi que les protocoles de séparation chromatographique doivent être adaptés avec soin afin de réaliser des séparations chromatographiques reproductibles. Un autre aspect est que plus la masse des objets est grande plus la charge doit être importante. En effet les spectromètres mesurant un rapport masse sur charge, un objet massif doit être fortement chargé pour être observé dans la fenêtre d'acquisition des spectres.

A l'heure actuelle, les méthodes de quantification en Bottom-up ne sont pas transposables à l'analyse Top-down. En effet il n'existe aucune méthode quantitative développée en Top-down reposant sur un marquage *in-vivo*. Par ailleurs bien que le "Stable Isotope Labelling by Amino acids in Cell culture (SILAC) ait été développé de manière purement théorique (Waanders et al., 2007) la mise en pratique d'un tel marquage n'est pas optimum car les données spectrales sont naturellement complexes sans modification de la chimie des molécules (condition naturelle). Les nouvelles données ajouteraient donc une dimension de complexité supplémentaire. D'autant plus que le suivi de l'incorporation des acides aminés « lourds » dans les protéines est non prévisible, la difficulté étant d'autant plus grande à cette échelle de masse.

Pour des raisons identiques, le "Tandem Mass Tag" (TMT) n'a pas été développé car cette méthode se fonde sur une fragmentation exhaustive des objets biologiques. Elle n'aurait pas de sens en approche Top-down.

Actuellement, une des rares stratégies de quantification des protéines disponibles en approche Top-down est le label-free (LFQ). Cette méthode repose sur l'utilisation des informations associées au courant d'ions total (TIC) de chaque cluster isotopique comme valeurs quantitatives des protéoformes identifiées. Dans un premier temps, les affectations des ions issus de l'analyse d'un spectre de fragmentation (MSMS) sont utilisées pour identifier par un séquençage la protéine intacte précurseur, c'est à dire à l'origine de ce spectre. Cette identification est rendue possible par l'utilisation d'une banque de données protéiques associant les protéines à leurs séquences. Les masses calculées à partir des séquences en acides aminés sont comparées à des masses dites « moyennes » correspondant au centre de masse du cluster isotopique. Ainsi, un calcul de la masse monoisotopique n'est pas nécessaire à

l'identification des protéines mais est rendu possible par l'utilisation de l'*Averagine*, mais également en utilisant l'abondance naturelle des isotopes de chaque élément chimique. Une fois l'identification des protéoformes effectuée pour chaque échantillon analysé, un alignement des pics chromatographiques est réalisé afin de pouvoir comparer les intensités brutes des clusters isotopiques des mêmes ions entre les échantillons. Cela implique donc une haute reproductibilité entre les conditions et exige une procédure peu usitée de séparations chromatographiques superposables.

Il existe d'autres limitations comme l'exhaustivité de la banque de données et la diversité des modifications post-traductionnelles non systématiquement prise en compte (effet combinatoire critique). En effet selon la composition en acides aminés cible d'une séquence protéique donnée, il y a une multitude de positions possibles pour les PTMs.

Les données d'analyse par spectrométrie de masse des protéines intactes sont donc complexes à obtenir et à traiter. Actuellement plusieurs méthodes et logiciels permettent de réaliser un retraitement des données. Cependant, ces méthodes de quantification sont encore à l'heure actuelle très peu développées.

Chapitre 2 : Analyse par spectrométrie de masse des protéines intactes

I. États de charges à l'intérieur d'un scan

Une protéine, étant donné sa taille importante, possède dans sa séquence un grand nombre d'acides aminés ionisables (acide aspartique, acide glutamique, lysine, arginine, histidine par exemple). Ces adduits multiples de protons possibles sur des acides aminés différents, ont pour effet de créer plusieurs états possibles de charge pour une même protéine (z = 1, z = 2, etc.). Ces états de charges ont des positions dans un spectre de masse définies en fonction de la masse divisée par la charge. Si la charge est différente, un *pattern* d'états de charge (de forme logarithmique) est observé pour chaque protéoforme étudiée (Figure 48).

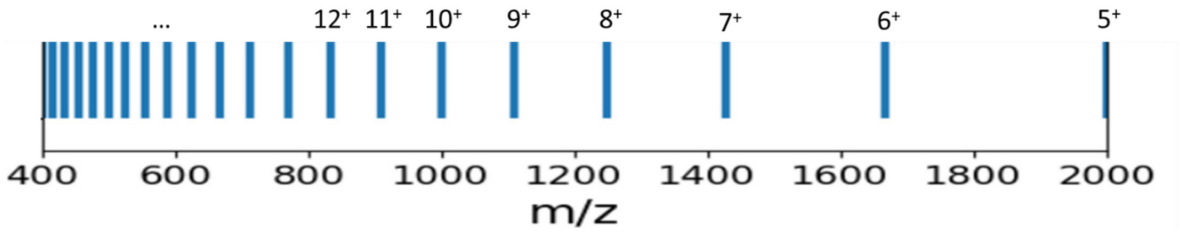


Figure 48 : Présentation de la position théorique des états de charges successifs pour une protéine de masse 10 000 Da.

En plus des différents états de charges, une même protéine peut également être observée plusieurs fois au cours d'une expérience de spectrométrie de masse sur des temps de rétention successifs (Reid & McLuckey, 2002; Rožman & Gaskell, 2012).

II. Détermination de la masse

En analyse Top-down, l'un des effets les plus notables de l'incrément en masse est que cela influe sur la composition isotopique des molécules. En effet, à des masses très élevées, la probabilité d'observer des ions « légers » est de plus en plus petite (illustré en Figure 49). Le corolaire étant que la probabilité d'observer des ions isotopologues, molécules possédant des neutrons supplémentaires est grande.

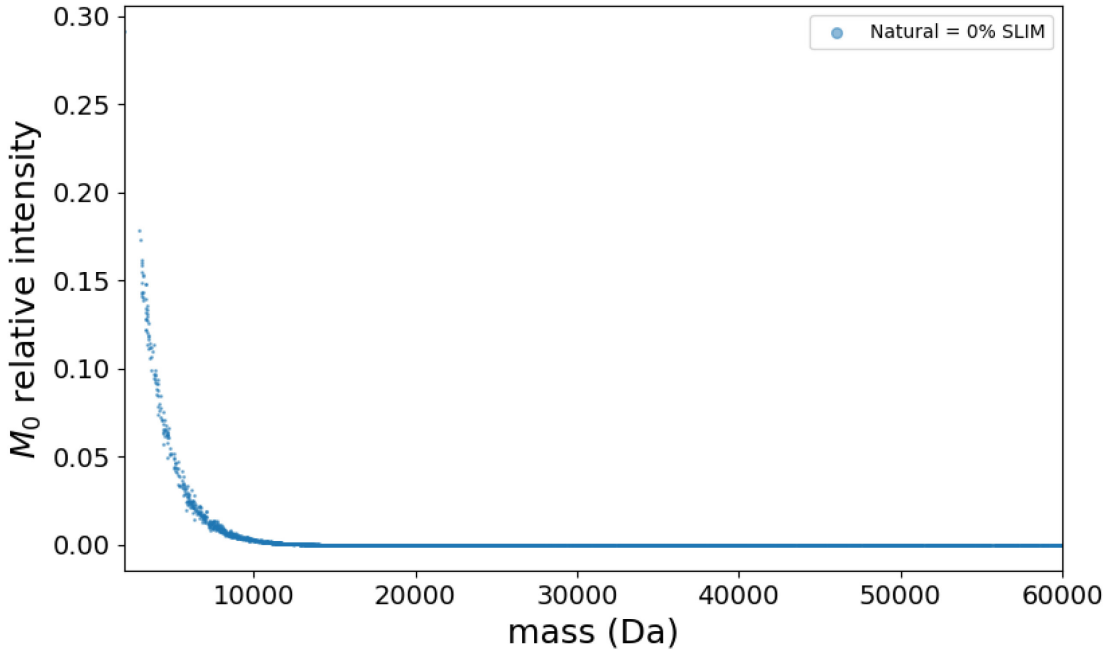


Figure 49 : Probabilité de l'intensité de l'ion monoisotopique pour chaque protéine de S. cerevisiae (n=6721) en fonction de la masse en dalton des protéines.

Pour permettre une recherche en banques de données, la mesure de la masse doit être la plus précise possible afin d'obtenir des identifications pertinentes. En approche Bottom-up, la masse du monoisotopique peut être utilisée, car elle s'observe facilement et de manière intense et son calcul algébrique est aisé. Dans les clusters isotopiques des protéines intactes, l'accès à la masse du monoisotopique est beaucoup plus complexe car souvent très peu intense (filtrée lors de la suppression dans le bruit de fond). Les algorithmes travaillent donc à partir des données de masse moyenne (Figure 50). Cette solution est beaucoup moins précise, car l'annotation du rang isotopique est ambiguë (Figure 51). La mesure de masse obtenue permet tout de même des identifications en se fondant sur le modèle de composition élémentaire moyenne des acides aminés (*Averagine*) à une abondance naturelle des isotopes.

Averagine: $C_{4.9384}\ H_{7.7583}\ N_{1.3577}\ O_{1.4773}\ S_{0.0417}$ (Senko et al., 1995)

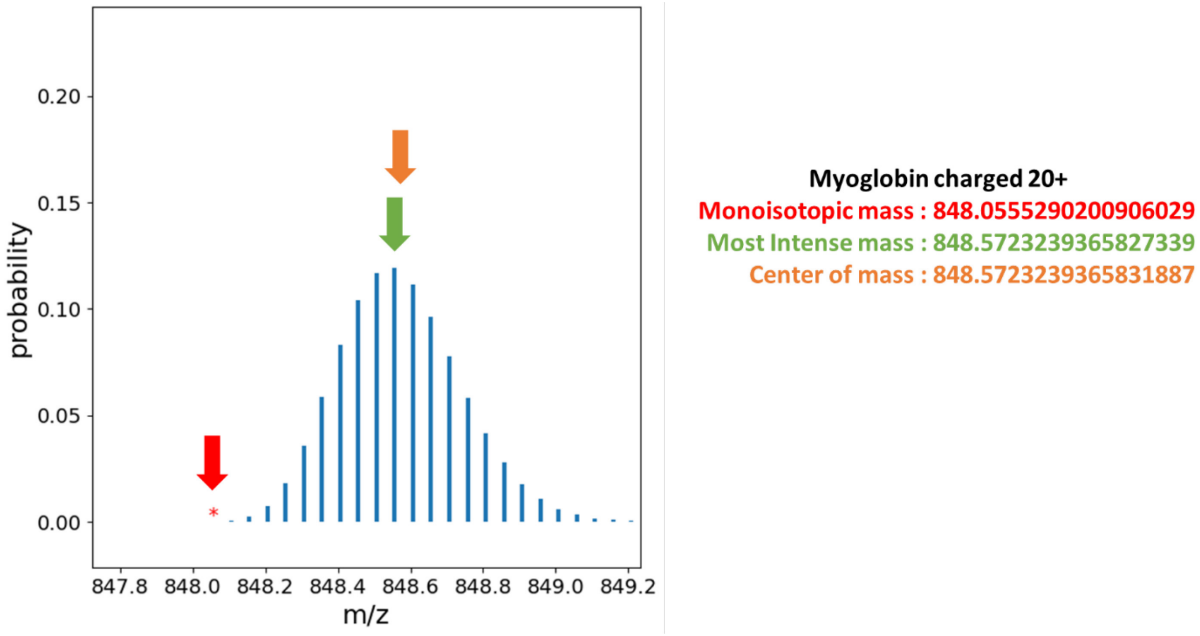


Figure 50 : Spectre de masse théorique de la Myoglobine de Cheval P68082 chargée 20⁺ (protéines de masse monoisotopique 16941.9723Da).

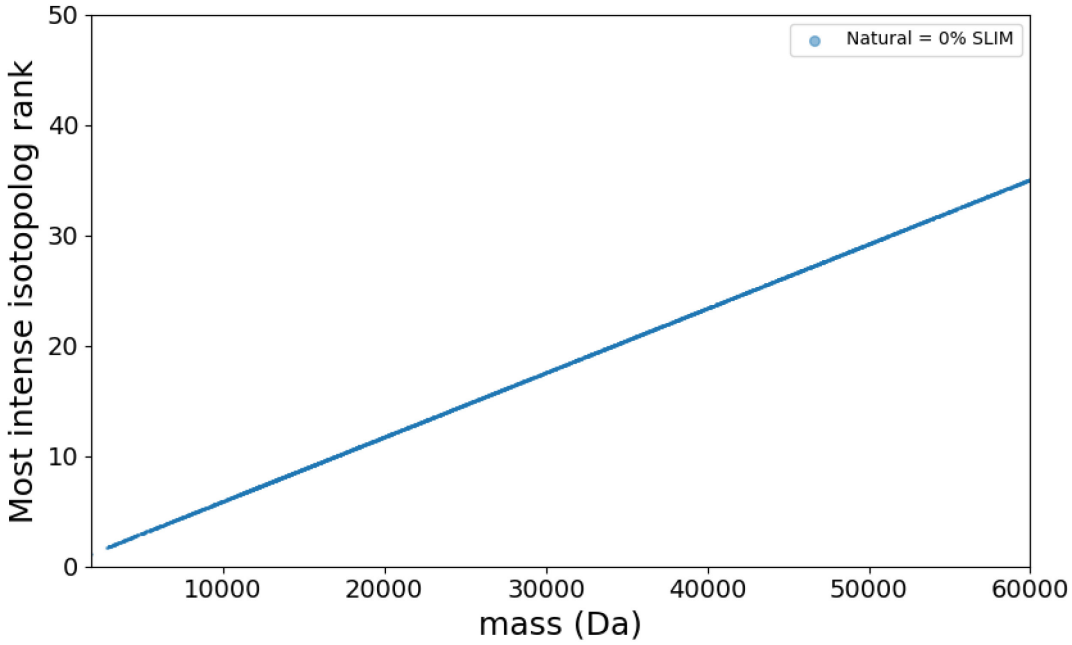


Figure 51 : Position de l'isotopologue le plus intense en fonction de la masse de chaque protéine de S. cerevisiae (n=6721)

En résumé, deux grandes familles de logiciels de traitement des données de masse existent. La première regroupe les logiciels qui fonctionnent avec la masse neutre moyenne, tandis que la deuxième regroupe les logiciels qui font appel aux masses neutres monoisotopiques.

III. Détermination de la charge

III.1 Déconvolution

La première étape de retraitement des données brutes de masse est le calcul de la masse neutre de la protéine qui est à l'origine des états de charges observés (Figure 52). Pour cela, une approche dite de « déconvolution des spectres » est mise en application.

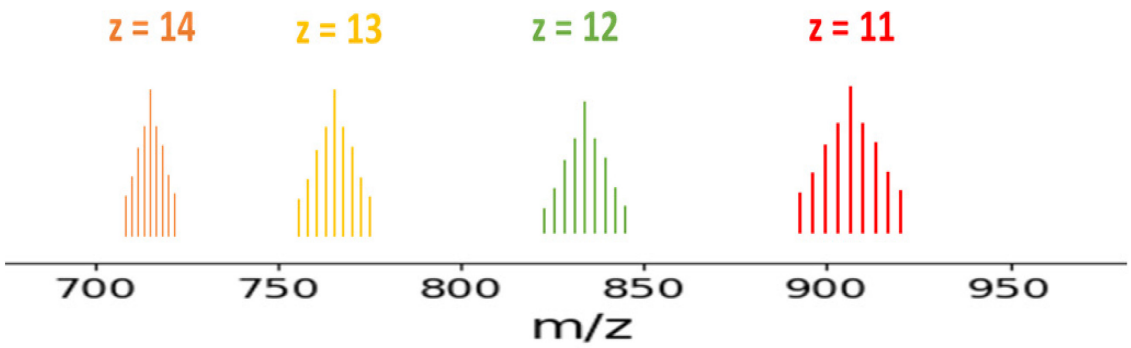


Figure 52 : Exemple d'attribution des états de charges au sein d'un spectre MS pour une protéine de masse 10 kDa.

La procédure de déconvolution est relativement simple puisqu'il s'agit d'attribuer artificiellement, selon un ordre précis, les charges des massifs observés. Les paramètres sont les valeurs limites des charges définies par l'utilisateur (par exemple z=1 -> z=60). Une fonction de convergence par un calcul de déviation standard est appliquée. Lorsque les massifs ont été correctement assignés en masse, la masse neutre (déconvoluée) de chaque état de charge est similaire. L'écart de masse sert de mesure d'imprécision car la valeur finale de masse neutre dépend strictement des valeurs de masse/z extraites du spectre comme données d'entrée.

Le logiciel en ligne ESIprot online[1] (Winkler, 2010) permet de calculer les valeurs de charges de chaque massif isotopique à l'aide de la position en masse/z observée. Bien que très performant, ce logiciel développé en python ne fonctionne qu'avec des données ponctuelles (pas d'extraction de données brutes) et n'est pas simple à implémenter dans un workflow d'analyse complet. C'est pourquoi je l'ai réécrit en C++ afin d'être inclus dans l'algorithme complet d'analyse développé dans le cadre de cette thèse. Il est disponible publiquement sur un hébergeur de code source : GitHub[2]. Il est à noter que dans ce logiciel, la masse déconvoluée correspond à une masse monoisotopique chargée 1 fois (possédant donc un proton additionnel) MH^+.

D'autres logiciels de déconvolution utilisent des données brutes et permettent la réalisation d'une extraction automatique des massifs existant : **Top-FD** (X. Liu et al., 2010), **FlashDeconv** (Jeong et al., 2021). Pour chaque scan MS, ces logiciels génèrent une liste des valeurs en masse des massifs déconvolués et leurs attributions de charge au sein du spectre. Une intensité correspondant à la somme des valeurs d'intensité du massif isotopique est également associée.

Au sein de l'architecture de la suite de logiciels OpenMS a été développé FlashDeconv qui se présente comme une stratégie originale de traitement des données d'analyse par spectrométrie de masse des protéines intactes. L'algorithme de déconvolution de FlashDeconv utilise la simplification des valeurs par une approche logarithmique pour limiter l'attribution des états de charges à une simple recherche de *pattern*.

En effet, le cœur de cet algorithme réside dans l'équivalence suivante :

$$log\ (\frac{m}{z}) = log\ (m) - log\ (z) \qquad (3\text{-}1)$$

Les données de masses de chacun des spectres sont tout d'abord converties en log(m/z), puis connaissant la valeur *m/z*, il suffit d'attribuer *z* pour retrouver *m*. C'est un algorithme rapide qui peut être utilisé en temps réel lors de l'acquisition des spectres afin d'obtenir des analyses du protéome avec une plus grande exhaustivité.

[1] https://www.bioprocess.org/esiprot/esiprot_form.php

[2] https://github.com/Nohic56

III.2 Écart de masse entre les isotopologues

Comme vu précédemment, les biomolécules sont composées d'éléments chimiques qui possèdent naturellement plusieurs isotopes stables, dont les abondances relatives sont connues. Dans un spectre de masse, chaque ion est ainsi observé dans un état de charge donné (voir ci-dessus Figure 52). Cet état de charge est lui-même représenté par un « cluster isotopique », c'est-à-dire un ensemble d'isotopologues qui sont séparés par un écart de masse (Figure 53) correspondant à la masse d'un neutron issue du carbone (et des autres éléments chimiques comme l'Hydrogène, l'Azote, l'Oxygène, le Soufre) divisé par la charge (soit 1.00335/z, où z est la charge de l'ion associé au cluster isotopique).

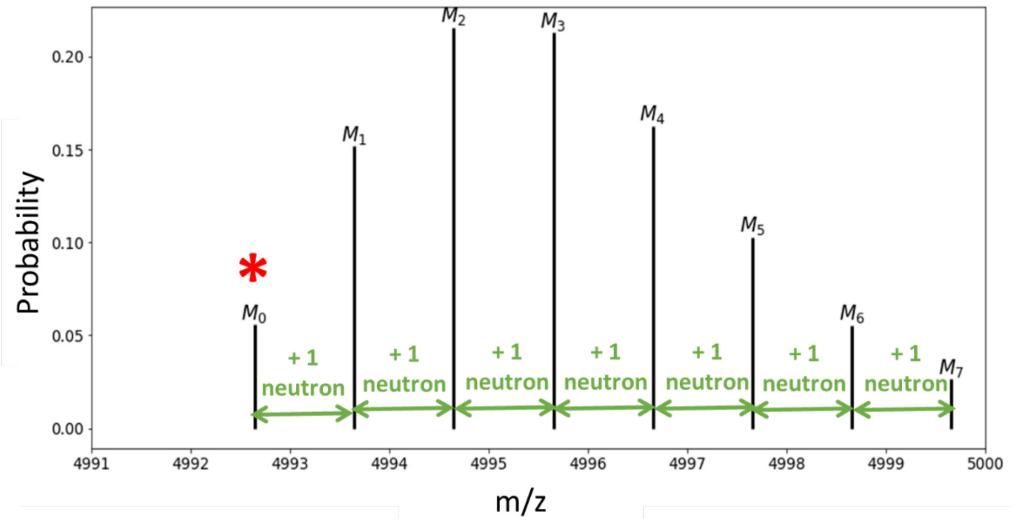


Figure 53 : Cluster isotopique théorique présentant l'écart entre les isotopologues d'une molécule de 4000 Da monochargée.

Ainsi au sein d'un cluster isotopique quelconque, l'état de charge de l'ion associé peut être retrouvé par la mesure de l'écart en masse observée entre les isotopologues qui composent le massif isotopique. Il est à noter que la masse du neutron dépend de l'élément chimique (Audi & Wapstra, 1995). En moyenne, pour les calculs réalisés en spectrométrie de masse de bonne résolution, on utilise une masse approximative du neutron en prenant celle de l'élément carbone. Cette masse s'obtient par le calcul suivant : masse ^{13}C- masse ^{12}C = 1.003355. Cependant, si l'on effectue une moyenne des masses neutroniques des différents éléments chimiques, on obtient la valeur 1,008665. Pour des molécules de faible masse, l'écart de masse est relativement faible et l'approximation justifiée. Néanmoins, c'est une information à garder à l'esprit

puisque l'écart de masse augmente avec le rang isotopique, ce qui sera à l'origine d'incertitude. C'est la raison pour laquelle, les isotopologues sont décrits eux-mêmes par un ensemble de différents signaux des espèces isotopiques, selon la nature de l'élément chimique. Il est à noter qu'en spectrométrie de masse à très haute résolution (par exemple les FT-ICR), ces sous-isotopologues s'observent et sont donc mesurés. C'est la raison pour laquelle le logiciel de calculs théoriques MIDAs (Alves et al., 2014) possède deux modèles de résolutions des équations polynômiales : le C*oarse grain,* une solution peu résolutive et le F*ine grain,* une solution plus complète dans le cas où les sous-isotopologues s'observent (Figure 54).

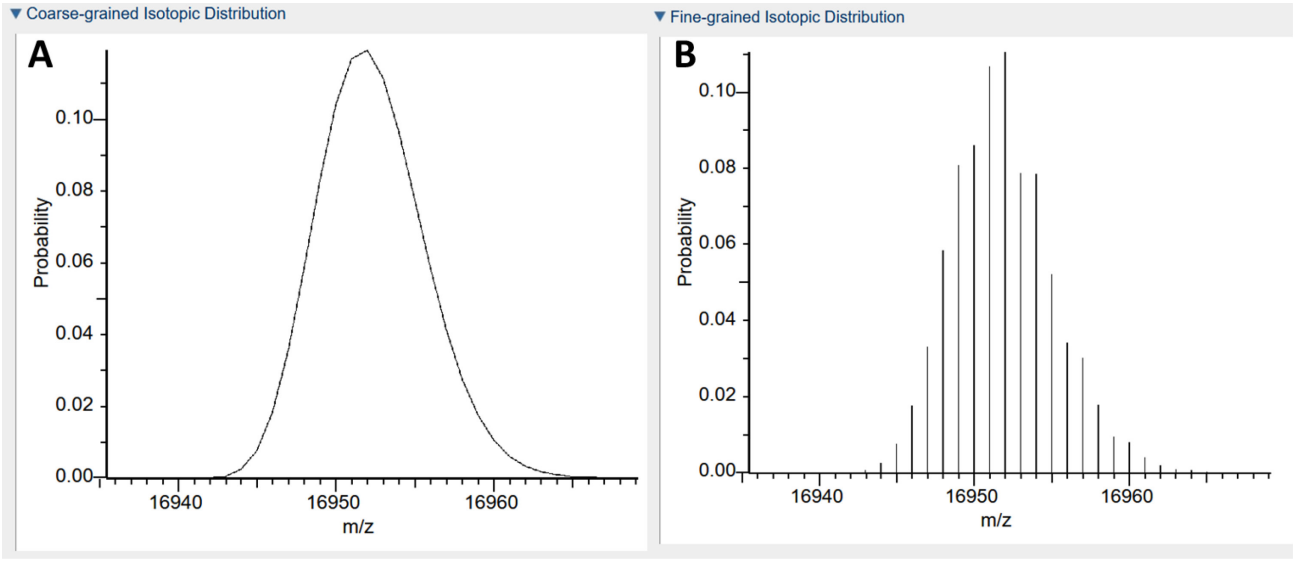


Figure 54 : Différence entre le modèle Coarse-grained (A) et Fine-grained (B) de MIDAs. La séquence utilisée est celle de la Myoglobine de Cheval (P68082) de masse monoisotopique 16941.9723Da)

Chapitre 3 : Quantification par la méthode tSLIM

Notre méthode de quantification SLIM est fondée sur un calcul impliquant des valeurs expérimentales et des valeurs théoriques modélisées. La stratégie de quantification SLIM repose sur le principe d'effectuer un mélange de façons équimolaires de deux échantillons provenant d'une condition non-marquée (NC) et d'une condition marquée (12C) au carbone ^{12}C. La variation d'abondance entre les deux conditions s'exprime par une modulation du facteur de molarité déterminée expérimentalement. Le calcul consiste en la mesure de l'intensité des isotopologues continue dans le cluster isotopique expérimental.

Dans son application en approche Bottom-up (le bSLIM), la modélisation des clusters isotopiques est rendue possible par les données d'identification qui associent pour chaque cluster une séquence d'acide aminés. Celle-ci est aisément convertie en formule chimique utilisable pour les calculs de modélisations des clusters isotopiques théoriques (Ex : MIDAs). Le bSLIM repose sur la mesure du rapport M_1/M_0 fonction de l'incorporation de ^{12}C dans les peptides. Cependant cette méthode exige d'avoir accès aux valeurs expérimentales d'intensité de ces deux isotopologues. En tSLIM, ces isotopologues (M_0, M_1 et éventuellement les isotopologues de rang plus élevés) ne sont plus visibles, c'est à dire que leurs intensités sont trop faibles pour être mesurées par le spectromètre de masse. Nous avons donc développé une autre méthode afin de quantifier les protéines. Concrètement la mesure des clusters isotopiques expérimentaux sert de support à la modulation d'un cluster théorique provenant des deux conditions (NC et 12C) permettant la détermination du facteur d'abondance. Ainsi, nous avons eu besoin de développer une méthode afin de modéliser théoriquement les clusters isotopiques des deux conditions indépendamment de la connaissance de la séquence en acides aminés.

I. Modélisation des clusters isotopiques dissociée de la formule chimique

Nous avons repris la définition absolue de cluster isotopique et étudié les différentes approches pour réaliser une modélisation de manière purement théorique. Notamment, (Breen et al., 2000) décrivent le fait qu'un massif isotopique peut être modélisé sous la forme d'une distribution de Poisson des intensités des isotopologues qui le constituent. Ainsi, quelle que soit la composition chimique et l'abondance isotopique de la molécule biologique étudiée, l'intensité et la position des isotopologues au sein d'un cluster isotopique peut être déterminée mathématiquement.

Pour tout x entier ≥ 0, représentant le numéro de l'isotopologue,

$$x \in [0, +\infty[$$

l'intensité normalisée des isotopologues au sein d'un massif est analogue à une valeur de probabilité définie par l'expression suivante :

$$f(x) = \frac{e^{-A}}{x!} \times A^x \qquad (3\text{-}2)$$

où A, entier strictement positif, est l'espérance de la loi, défini comme la moyenne des masses du cluster modélisé. Ce paramètre requiert d'être estimé avec la plus grande précision car il contraint très grandement la loi.

Pour cela, (Valkenborg et al., 2007) ont démontré que le paramètre A, centre de masse de la distribution de Poisson suit une relation linaire avec la masse monoisotopique des clusters isotopiques modélisés (Figure 55).

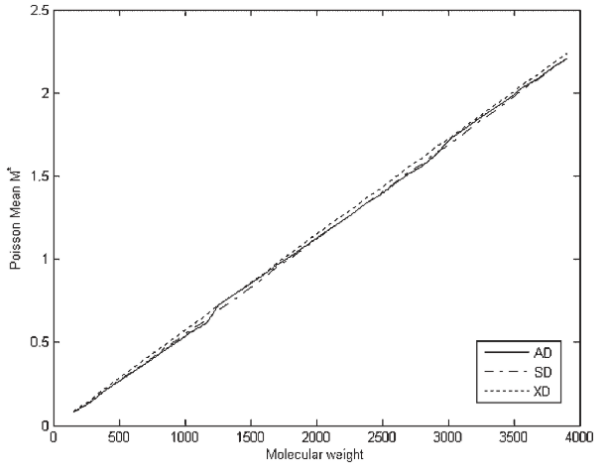


Figure 55 : Centre de la distribution de Poisson en fonction de la masse monoisotopique des molécules. Cette figure est extraite de (Valkenborg et al., 2007).

Plus récemment, ces paramètres ont été confirmés par (Sadygov, 2018), dans une étude plus exhaustive de l'influence des éléments chimiques. En particulier, ce modèle permet une meilleure prise en compte de l'élément chimique Soufre.

Cela nous permet d'écrire :

$$A = a * M_0 + b \quad (3\text{-}3)$$

Où a est le coefficient directeur de la relation linéaire et b le facteur correctif de la fonction affine formée par la relation entre A et M_0.

La détermination précise de la masse monoisotopique (M_0) est donc un élément clé dans l'étude des protéines intactes.

L'application du marquage SLIM à la problématique Top-down (le tSLIM) est donc intéressante car il permet d'allier les avantages de l'analyse des protéines intactes à la puissance de notre méthode de quantification d'un protéome entier. Pour cela, les données de masses sont à traiter différemment de celles d'une analyse classique Top-down. En effet, en raison de l'enrichissement isotopique en ^{12}C des protéines issues de la condition 12C, les *clusters* isotopiques de chaque protéoforme sont modifiés par rapport à la condition Naturelle NC. Ces modifications sont d'autant plus fortes que la masse de la molécule étudiée est grande (Figure 56).

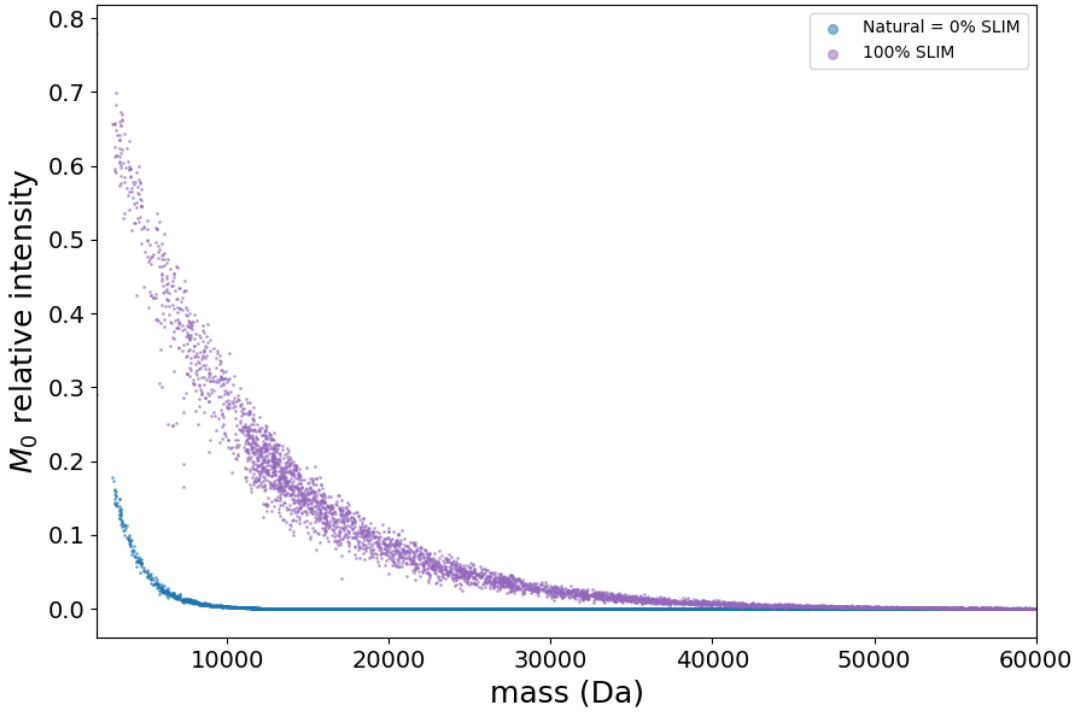


Figure 56 : Intensité normalisée de l'ion monoisotopique dans les conditions NC et 12C. Cette figure a été réalisé par la simulation numérique de cluster isotopique à partir des séquences protéiques du protéome de S. cerevisiae (n=6721).

Les objectifs de l'application du marquage SLIM en Top-down sont donc de permettre la modélisation correcte des spectres de masses puis la quantification tant en condition NC qu'en 12C.

Au sein d'un spectre de masse, bien que les positions des masses des isotopologues restent strictement identiques entre les deux conditions NC et 12C, les intensités des pics varient fortement. Cela est observable lors du suivi de la mesure de la position de l'isotopologue le plus intense en fonction de la masse (Figure 57).

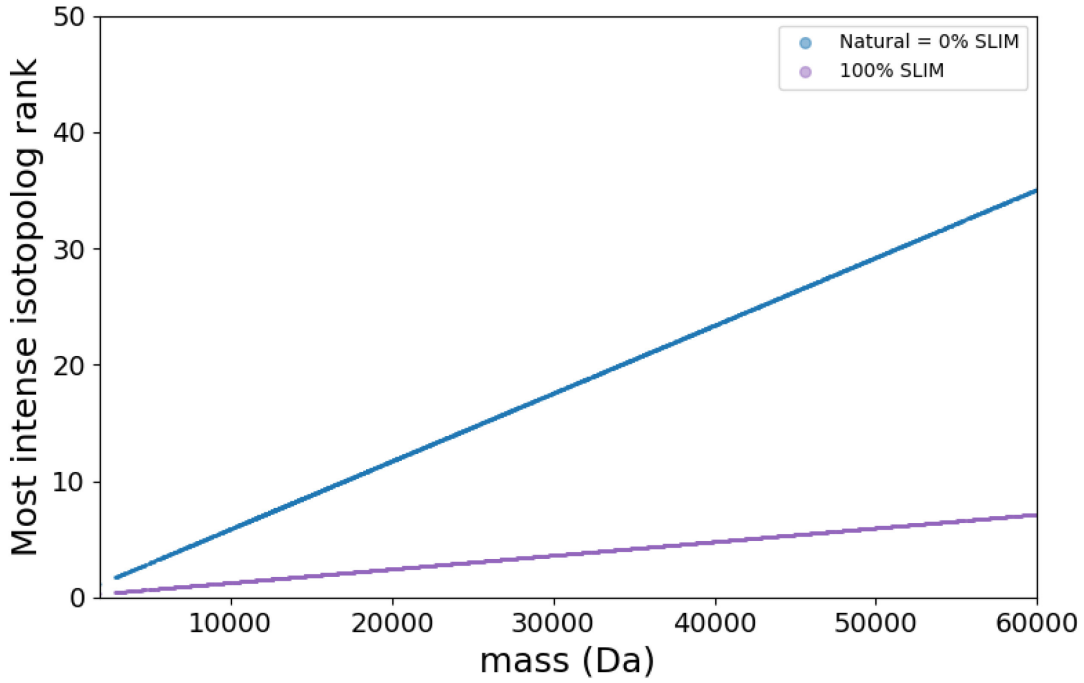


Figure 57 : Différence en dalton entre la masse monoisotopique et la masse de l'isotope le plus intense en conditions NC naturelle et 12C. Données de modélisation à partir du protéome de S. cerevisiae (n=6721).

II. Utilisation de la distribution de Poisson pour le calcul de la fraction molaire pour la comparaison des abondances des protéines intactes

Dans une expérience type de tSLIM, des protéines provenant de la condition NC sont mélangées en quantité équimolaire avec des protéines provenant de la condition 12C. Au sein des spectres de masse expérimentaux, chaque cluster isotopique de protéines intactes sera donc composé d'un spectre entièrement « marqué » (12C) et d'un autre spectre entièrement « non-marqué » (NC). Le facteur de molarité *Alpha*

sera la contribution du spectre naturel par rapport au spectre marqué, équivalent à la fraction molaire des protéines issues de la condition non-marquée contenue dans l'échantillon étudié. Cette valeur quantitative est utilisée pour mesurer l'abondance relative des protéines entre les deux conditions comparées (12C et NC) lors de l'expérience effectuée.

Ainsi nous pouvons écrire que l'observation d'un spectre expérimental en tSLIM est définie part :

$$SpectreProteine^{\,experimentale} = \boldsymbol{\alpha} * SpectreProteine^{Naturel} + (1 - \boldsymbol{\alpha}) * SpectreProteine^{SLIM\ 12C}$$

En termes d'intensité, chacune des valeurs des isotopologues est la résultante d'une somme entre les intensités des isotopologues marqués (12C) et non marqués (NC), pondérée par un facteur d'abondance.

Comme expliqué précédemment, à la différence du Bottom-up SLIM les clusters isotopiques en Top-down sont très complexes et difficilement modélisable, parce que nécessitant un accès au isotopologues de rangs élevés ce qui difficilement accessible par des développements polynomiaux. Aussi, nous avons développé une nouvelle méthode de modélisation fondée sur une distribution de Poisson de l'intensité des isotopologues que l'on utilise pour décrire chaque cluster isotopique. Dans la quantification tSLIM, Il est nécessaire de proposer les valeurs des paramètres définissant les deux distributions de Poisson, l'une issue de la condition naturelle (NC) notée A, l'autre issue de la condition 12C marquée et notée B.

Soit A le coefficient Naturel et B le coefficient de la condition marquée 12C respectivement :

$$A = a * M_0 + b \qquad (3\text{-}4)$$

$$B = c * M_0 + d \qquad (3\text{-}5)$$

Nous parvenons donc à cette équation où nous posons $f(x)$ comme l'expression de la probabilité d'abondance de l'isotopologue de rang x. Celle-ci est analogue à une valeur d'intensité normalisée.

$$f(x) = \alpha \times \left[\frac{e^{-A}}{x!} \times A^x\right] + (1-\alpha) \times \left[\frac{e^{-B}}{x!} \times B^x\right] \qquad (3\text{-}6)$$

Cette équation nous permet de modéliser les clusters isotopiques expérimentaux issus d'un mélange entre les conditions NC et 12C comme présenté de manière théorique en Figure 58.

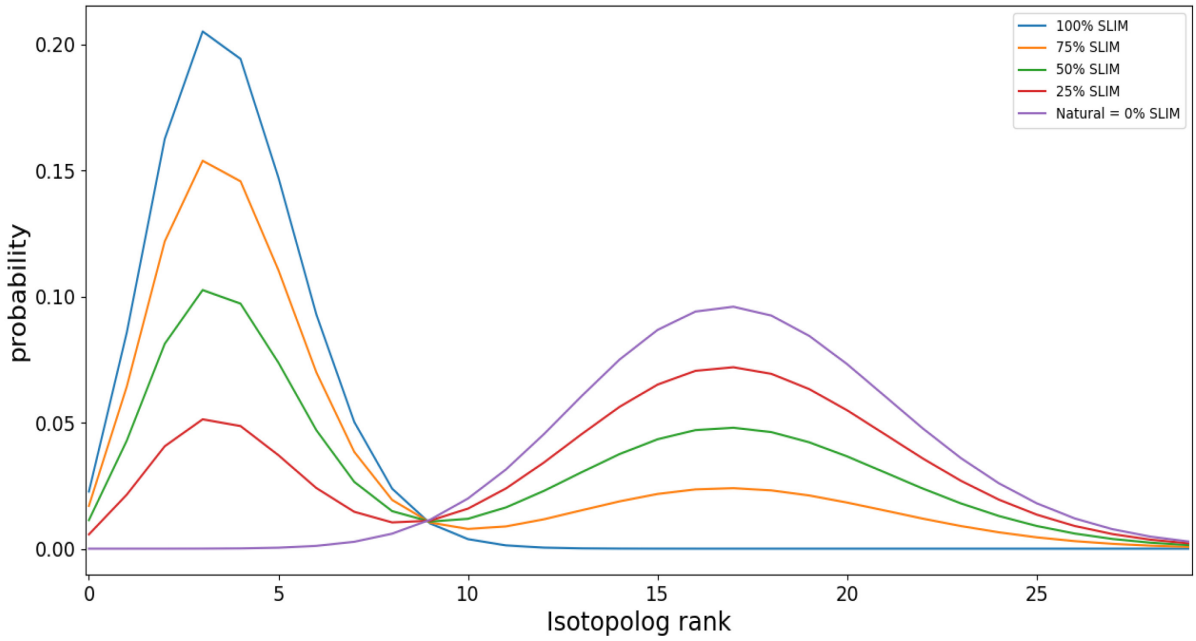


Figure 58 : Exemple de distribution théorique de différents mélanges pour une masse de 30 000 dalton

II.1 Détermination de *alpha*, la fraction molaire des protéines non-marquées

Sachant que les paramètres A et B sont déterminés à partir de la masse de la protéine non chargée, seule la valeur du facteur *alpha* reste à déterminer.

Etudions les valeurs remarquables des calculs de modélisation des distributions isotopiques à partir des équations de Poisson.

- Pour l'ion monoisotopique, le calcul de son intensité normalisée $f(0)$ est :

$$f(0) = \alpha \times e^{-A} + (1-\alpha) \times e^{-B} \qquad (3\text{-}7)$$

$$f(0) = \alpha \times (e^{-A} - e^{-B}) + e^{-B} \qquad (3\text{-}8)$$

Ainsi l'expression de *alpha* peut être déterminée par :

$$\alpha = \frac{f(0)-e^{-B}}{e^{-A}-e^{-B}} \qquad (3\text{-}9)$$

- Pour le premier isotopologue de rang 1, on résout $f(1)$

$$f(1) = \alpha \times [e^{-A} \times A] + (1-\alpha) \times [e^{-B} \times B] \qquad (3\text{-}10)$$

$$f(1) = \alpha \times (A \times e^{-A} - B \times e^{-B}) + (e^{-B} \times B) \qquad (3\text{-}11)$$

L'expression de *alpha* associée est ainsi :

$$\alpha = \frac{f(1)-(B\times e^{-B})}{(A\times e^{-A}-B\times e^{-B})} \qquad (3\text{-}12)$$

Ce qui permet d'écrire :

$$\alpha = \frac{f(0)-e^{-B}}{e^{-A}-e^{-B}} = \frac{f(1)-(B\times e^{-B})}{(A\times e^{-A}-B\times e^{-B})} \qquad (3\text{-}13)$$

$$\alpha = (f(0) - e^{-B}) \times \ (A \times e^{-A} - B \times e^{-B}) = \ [f(1) - (B \times e^{-B})] \times (e^{-A} - e^{-B}) \qquad (3\text{-}14)$$

Les valeurs d'intensité exprimées ci-dessus sont analogues à une intensité relative de chaque isotopologue. Les intensités expérimentales ont donc besoin d'être normalisées.

Ainsi si l'on exprime les valeurs d'intensité expérimentales normalisées pour le monoisotopique et le 1er isotopologue respectivement :

$$f(0) = \frac{f(0)_{exp}}{\sum_0^i f(x_i)_{exp}} \text{ et } f(1) = \frac{f(1)_{exp}}{\sum_0^i f(x_i)_{exp}}$$

Le rapport entre les valeurs expérimentales permet de simplifier l'équation à cette égalité :

$$\frac{f(0)}{f(1)} = \frac{f(0)_{exp}}{f(1)_{exp}}$$

Ce qui permet d'écrire plus simplement :

$$\frac{f(0)_{exp}}{f(1)_{exp}} = \frac{\alpha\times(e^{-A}-e^{-B})+e^{-B}}{\alpha\times(A\times e^{-A}-B\times e^{-B})+(B\times e^{-B})} \qquad (3\text{-}15)$$

En exprimant *alpha*, on obtient :

$$\alpha = \frac{e^{-B} \times (B \times f(0)_{exp} - f(1)_{exp})}{e^{-A} \times (f(1)_{exp} - A \times f(0)_{exp}) + e^{-B} \times (B \times f(0)_{exp} - f(1)_{exp})} \quad (3\text{-}16)$$

Ce rapport permet donc d'obtenir la fraction molaire des protéines non-marquées (provenant de la condition NC). Tout comme le bSLIM, le rapport $\frac{\alpha}{1-\alpha}$ est analogue à une valeur de *fold change*, la valeur de référence de quantification couramment utilisée en biologie.

Dans le cadre d'une expérience tSLIM, dans la condition 12C les molécules présentent une abondance isotopique particulière et les spectres de masse sont alors extrêmement modifiés. Comme présenté lors de la modélisation des clusters isotopiques en Top-down, le marquage SLIM provoque une modification drastique de la distribution de l'intensité des isotopologues.

En particulier, dans cette condition l'ion monoisotopique est observable malgré la grande masse des molécules. Cela permet de mesurer le plus fidèlement la masse de l'ion monoisotopique utilisé dans le cadre de la modélisation des clusters isotopiques théoriques. En premier lieu, la quantification s'effectue à partir des données expérimentales des MS1. Une extraction de la valeur des isotopologues est l'unique descripteur expérimentale permettant la quantification. De plus, une redondance de la mesure est rendue possible par l'études des différents clusters isotopiques formant les états de charge de la molécule d'intérêt.

Ainsi, l'obtention de valeurs quantitatives de manière relative entre deux conditions permet de dresser des réponses et des conclusions à des états biologiques.

Si les calculs algébriques permettant une quantification ont été développés, la difficulté réside dans une approche fondée sur la mesure de quelques valeurs. De plus, utiliser une méthode nécessitant une intensité forte pour les premiers isotopologues, telle que dans la condition 12C, induit un biais. C'est-à-dire que les protéines quantifiées de manière robuste auront une tendance à être surexprimées dans la condition 12C. C'est la raison pour laquelle nous avons développé une autre méthode

afin de quantifier chaque cluster isotopique par un ajustement dynamique des données expérimentales sur un modèle de distribution de l'intensité des isotopologues.

Pour rappel, un cluster isotopique expérimental est la combinaison des deux clusters, possédant l'un une distribution isotopique naturelle et l'autre une composition isotopique 12C. Les deux clusters théoriques sont modulés par le coefficient de quantification, alpha.

$$(M_0^{exp}, M_1^{exp}, M_2^{exp}, M_3^{exp} \dots M_n^{exp}) = \boldsymbol{\alpha} * (M_0^{NC}, M_1^{NC}, M_2^{NC}, M_3^{NC} \dots M_n^{NC}) + (1 - \boldsymbol{\alpha}) * (M_0^{12C}, M_1^{12C}, M_2^{12C}, M_3^{12C} \dots M_n^{12C})$$

Notre objectif est donc, pour tout cluster expérimental observé et extrait, de déterminer de manière itérative, la valeur de ce coefficient.

Pour cela, à chaque cluster isotopique expérimentalement déterminé est généré un cluster isotopique naturel et un autre 12C. En effet, la valeur précise de la masse monoisotopique obtenue grâce au marquage ^{12}C permet de résoudre les équations de la distribution de poisson et de décrire fidèlement les clusters isotopiques dans les deux conditions (comme illustré en Figure 59).

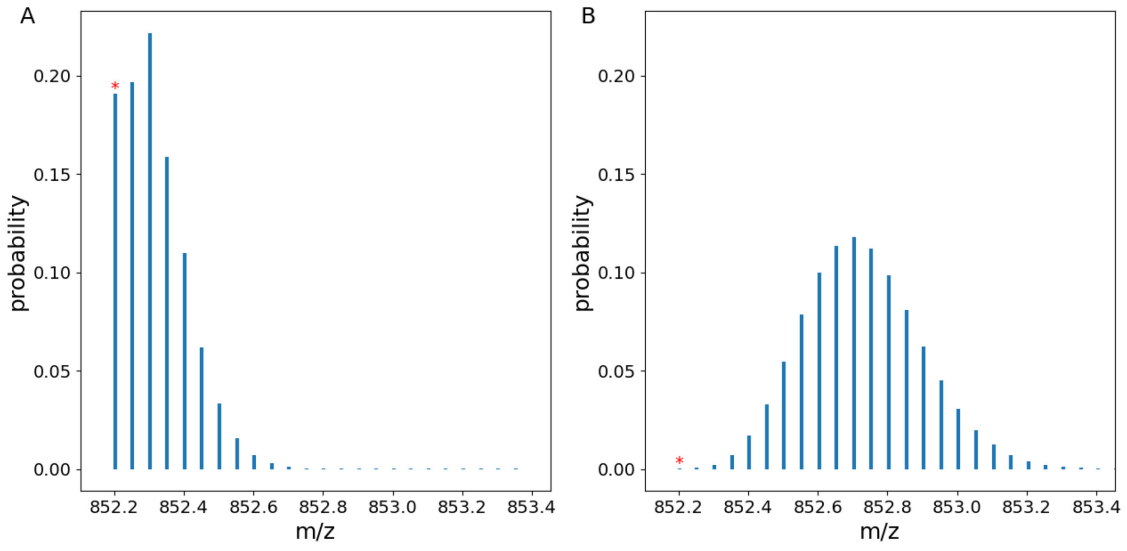


Figure 59 : Deux distributions isotopiques théoriques pour la Myoglobine de Cheval, A condition 100% 12C et B condition naturelle NC. Données théoriques obtenues par MIDAs.

De manière itérative, le facteur de quantification alpha est incrémenté pour une valeur de 0 (100% SLIM) à 1 (0%SLIM = NC) par pas successif de 0.01 correspondant à une valeur quantitative de 1%. Pour chaque cluster isotopique théorique ainsi produit, un score de quantification est réalisé. Pour cela l'écart quadratique moyen entre les valeurs théoriques et expérimentales est déterminé, s'exprimant sous la forme de Root Mean Squared Deviation (RMSD).

$$RMSD = \sqrt{\frac{1}{N} * \sum_{i=0}^{N} (Experimental - Theoretical)^2} \qquad (3\text{-}17)$$

Ainsi, la valeur quantitative « alpha » pour laquelle le RMSD, l'écart entre les deux clusters expérimentale et théorique, est minimum détermine la valeur quantitative du cluster isotopique expérimental étudié. Cette méthode de quantification s'illustre donc par le fait que seuls les signaux expérimentaux mesurés lors d'une acquisition en masse permettent de quantifier fidèlement les molécules biologiques indépendamment de leurs identifications.

Chapitre 4 : Simulations numériques pour explorer les propriétés de la modélisation de Poisson

I. Estimation graphique des paramètres de la loi de Poisson

Nous avons procédé à des simulations numériques à partir de la banque de séquences du protéome de *S. cerevisiae,* afin de déterminer les valeurs des paramètres décrivant la distribution de Poisson, de façon à obtenir des valeurs cohérentes avec les données expérimentales.

Ainsi, j'ai développé sous KNIME un workflow de simulation numérique qui permet premièrement de générer un peptidome complet (digestion tryptique *in silico*) à partir du protéome de *S. cerevisiae*. En effet, nous avons choisi de travailler sur le peptidome pour pouvoir comparer nos résultats directement avec ceux présentés par (Breen et al., 2000). Puis deuxièmement j'ai développé un algorithme en C^{++} utilisant MIDAs afin de générer les clusters isotopiques théoriques correspondant à chaque peptide du protéome donné. Enfin, en utilisant l'une des propriétés mathématiques de la distribution de Poisson selon laquelle la somme des produits de chaque probabilité multipliée par son rang, est égale à l'espérance de cette loi :

$$A = \sum_{n=0}^{\infty} n \times I_n \qquad (3\text{-}18)$$

J'ai pu tracer la courbe de la position de l'ion monoisotopique par rapport au centre de la distribution, A. Ce qui nous a permis, après ajustement sur une fonction affine, de déterminer les valeurs expérimentales pour un organisme donné. Cela a été fait dans les deux conditions NC et en 12C. Cependant, le graphique représente des courbes parallèles correspondant à des ensembles de peptides possédant un nombre variable d'atomes de soufre visibles sur la Figure 60. Dans ce cas, pour chaque atome

de soufre additionnel, l'ordonnée à l'origine (variable b et d) est doublée (cf. Tableau 7).

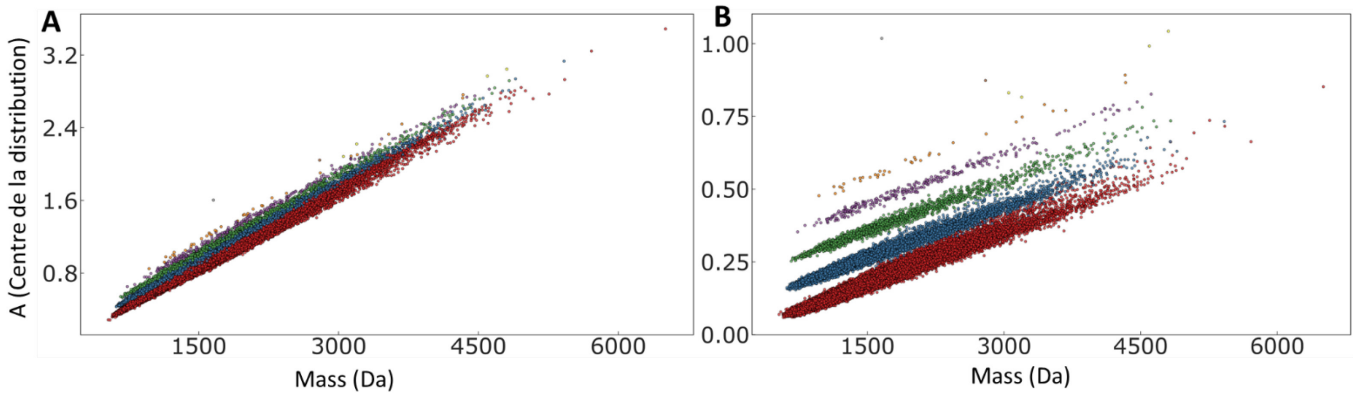


Figure 60 : Représentation du centre de la distribution en fonction de la masse des peptides

		NC		12C	
Nb de soufre	Nb de peptides	a	b	c	d
0	48923	0.00057258	0.07665648	0.00011747	0.08854432
1	10552	0.00057412	0.14726698	0.0001161	0.18218916
2	1827	0.00057056	0.22806662	0.0001172	0.27138535
3	296	0.00056693	0.30563559	0.00011799	0.36224859
4	37	0.00054178	0.46097817	0.0001229	0.4401034
5	4	-	-	-	-
6	1	-	-	-	-

Tableau 7 : Paramètres estimés par régression linaire (package R lm).

II. Estimation algébrique des paramètres de la loi de Poisson

Comme vu précédemment, le calcul algébrique de l'intensité normalisée de l'ion monoisotopique dans un spectre de masse se définit comme étant l'apport de l'intensité de chacun des éléments chimiques légers composant la particule.

Ainsi :

$$I_0 = P(^{12}C)^{nbC} * P(^{1}H)^{nbH} * P(^{14}N)^{nbN} * P(^{16}O)^{nbO} * P(^{32}S)^{nbS} \quad (3\text{-}19)$$

La modélisation de l'intensité du monoisotopique correspond à une résolution de la loi de Poisson au rang 0 :

$$f(0) = \frac{e^{-A}}{0!} \times A^0 = e^{-A} \qquad (3\text{-}20)$$

Selon la relation mathématique ci-dessus décrite, partons de l'égalité suivante :

$$I_0 = f(0) \qquad (3\text{-}21)$$

$$ln(I_0) = ln\big(f(0)\big) \qquad (3\text{-}22)$$

D'une part :

$$ln(e^{-A}) = -A \qquad (3\text{-}23)$$

Et d'autre part :

$$ln(I_0) = nbC \times ln\big(Pr^{12}C\big) + nbH \times ln\big(Pr\ ^1H\big) + nbN \times ln\big(Pr^{14}N\big) + nbO \times ln\big(Pr^{16}O\big) + nbS \times ln\big(Pr^{32}S\big) \qquad (3\text{-}24)$$

Sachant que A = a*M_0+b :

$$ln\big(f(0)\big) = -(a \times [nbC \times 12 + nbH \times 1.007825 + nbN \times 14.003074 + nbO \times 15.994915 + nbS \times 31.972071] + b) \qquad (3\text{-}25)$$

En utilisant la notion d'*Averagine,* la composition moyenne en élément chimique dans les acides aminés $C_{4.9384}\ H_{7.7583}\ N_{1.3577}\ O_{1.4773}\ S_{0.0417}$ (Senko et al., 1995), pour toute protéine il existe un facteur N, dans le produit qui explique la masse M_0 :

$M_0 = N \times M_{av}$.

Il est ainsi possible de résoudre l'équation en utilisant cette propriété pour chacun des éléments chimiques.

D'une part :

$$Ln(I_0) = 4.9384 \times ln\big(Pr^{12}C\big) + 7.7583 \times ln\big(Pr\ ^1H\big) + 1.3577 \times ln\big(Pr^{14}N\big) + 1.4773 \times ln\big(Pr^{16}O\big) + 0.0417 \times ln\big(Pr^{32}S\big) \qquad (3\text{-}26)$$

$$Ln(I_0) = -0.06493725 \qquad (3\text{-}27)$$

Et d'autre part :

$$Ln\big(f(0)\big) = -(a \times [4.9384 \times 12 + 7.7583 \times 1.007825 + 1.3577 \times 14.003074 + 1.4773 \times 15.994915 + 0.0417 \times 31.972071] + b) \qquad 3\text{-}28$$

$$Ln\big(f(0)\big) = -(a * 111.0543 + b) \qquad (3\text{-}29)$$

Or :

$$Ln(I_0) = ln\big(f(0)\big) = -A = -(a * M + b) \qquad (3\text{-}30)$$

Ainsi :

$$a = \frac{Ln(I_0)+b}{-M} = \frac{-0.0647873}{-111.0543} = 0.00058338 \qquad (3\text{-}31)$$

Ainsi, a est le coefficient directeur de la droite linéaire et b le facteur correctif (l'ordonnée à l'origine) de la fonction affine formée par la relation entre A et M_0. Notons que le facteur de correction b permet la prise en compte des protéines enrichies en soufre par rapport à la probabilité de présence moyenne.

Ces valeurs expérimentales sont cohérentes avec (Breen et al., 2000).

II.1 Application dans le cas d'une incorporation de ^{12}C

Les mêmes calculs peuvent être résolus afin de modéliser et de déterminer le coefficient directeur, seules les valeurs d'abondance de l'atome ^{12}C varient.

Il est ainsi possible, de résoudre l'équation en utilisant cette propriété pour chacun des éléments chimiques.

D'une part :

$$Ln(I_0) = 4.9384 \times ln\big(Pr\,^{12}C^{SLIM}\big) + 7.7583 \times ln\big(Pr\,^{1}H\big) + 1.3577 \times ln\big(Pr^{14}N\big) + 1.4773 \times ln\big(Pr^{16}O\big) + 0.0417 \times ln\big(Pr^{32}S\big) \qquad (3\text{-}32)$$

$$Ln(I_0) = \text{- } 0.01166168$$

Et d'autre part :

$$Ln\big(f(0)\big) = -(a \times [4.9384 \times 12 + 7.7583 \times 1.007825 + 1.3577 \times 14.003074 + 1.4773 \times 15.994915 + 0.0417 \times 31.972071] + b) \qquad (3\text{-}33)$$

$$Ln\big(f(0)\big) = -(a * 111.0543 + b) \qquad (3\text{-}34)$$

Or :

$$Ln(I_0) = ln\big(f(0)\big) = -A = a * M + b \qquad (3\text{-}35)$$

Ainsi :

$$a = \frac{Ln(I_0)+b}{-A} = \frac{-0.01166168}{-111.0543} = 0.00010501 \qquad (3\text{-}36)$$

Ce qui nous permet d'obtenir la valeur suivante **0.00010501** permettant de modéliser le plus justement possible les clusters isotopiques composés de 100% d'atomes de carbone ^{12}C.

II.2 Exemple de résolution algébrique pour une protéine donnée.

Nous allons maintenant mettre en application ces calculs afin de démontrer la sensibilité de résolution algébrique. Tout d'abord nous étudierons une protéine standard c'est-à-dire, de formule chimique simple, puis nous étudierons une protéine comprenant deux atomes de soufre.

- **Exemple de la protéine EXP1**

La protéine EXP1 de la levure (*Saccaromyces cerevisiae*) a pour particularité de ne pas avoir d'acide aminé soufré (cystéine ou méthionine hormis la méthionine initiatrice). Elle ne comporte ainsi pas d'atome de soufre dans sa formule chimique brute :

Q07541 = EXP1_YEAST (ER export of PMA1 protein 1)

De séquence protéique :

NLYGYFLLLIIVIAFIALLPLFSGIGTFKLTKPKSSATAQSATGKLGKREYLKKKL DHTNVLKFDLKDTEESLGHDSASASSASRKFEIDSKTGLKRRVIGQYNKDPNDFDFD IDDLINEDELDERREEEKKLKKYNGKKNEAYEGFV

De formule brute: $C_{758}H_{1203}N_{199}O_{232} + H^+$

De masse monoisotopique : 16806.8527574200925301 Da

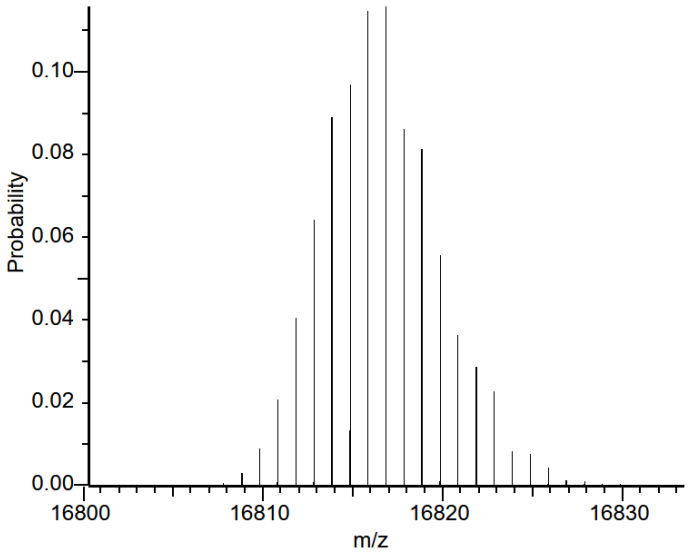


Figure 61 : Spectre de masse théorique modélisé par MIDAs de la protéine EXP1 ôté de la méthionine initiatrice.

Posons les équations :

$$Ln(I_0) = 758 \times ln\left(Pr^{12}C\right) + +1203 \times ln\left(Pr\ ^1H\right) + 199 \times ln\left(Pr^{14}N\right) + 232 \times ln\left(Pr^{16}O\right) + 0 \times ln\left(Pr^{32}S\right) \quad (3\text{-}37)$$

$$Ln(I_0) = -9.63214997 \quad (3\text{-}38)$$

Résolution algébrique de $f(0)$

$$ln\left(f(0)\right) = -A = -(a * M + b) = -(16941.9723284200917988 * 0.00058473 + b) = -\ (9.9064794795 + b) \quad (3\text{-}39)$$

L'intensité théorique du monoisotopique calculé par la modélisation d'une distribution de poisson est donc de 6.55859E-05.

- **Exemple de la Myoglobine :**

Posons les équations pour la protéine Myoglobine de cheval (*Equus caballus*) chargée 1, ôté de la méthionine initiatrice mais possédant tout de même deux méthionines dans la séquence, apportant deux atomes de soufre.

De séquence :

P68082 = MYG_HORSE (*Myoglobin*)

GLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKT
EAEMKASEDLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFIS
DAIIHVLHSKHPGDFGADAQGAMTKALELFRNDIAAKYKELGFQG

De formule brute : $C_{769}\ H_{1212}\ N_{210}\ O_{218}\ S_2 + H^+$

De Masse monoisotopique : 16941.9723284200917988 Da

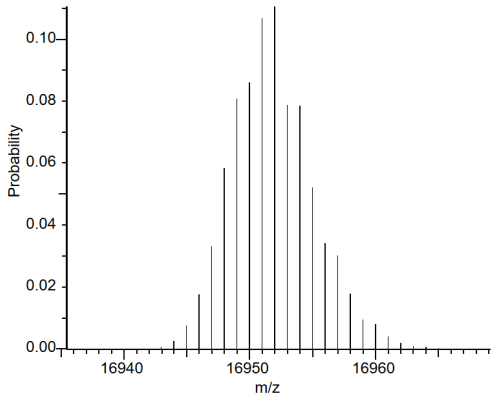


Figure 62 : Spectre de masse théorique modélisé par MIDAs de la protéine MYG ôté de la méthionine initiatrice

Posons les équations :

$$Ln(I_0) = 769 \times Ln(Pr12C) + 1212 \times Ln(Pr1H) + 210 \times Ln(Pr14N) + 218 \times Ln(Pr16O) + 2 \times Ln(Pr32S) \quad (3\text{-}40)$$

$$Ln(I_0) = -9.83072757 \quad (3\text{-}41)$$

Résolution algébrique de $f(0)$

$$ln\big(f(0)\big) = -A = -(a * M + b) = -(16941.9723284200917988 * 0.00058473 + b) = -(9.90655153 + b) \quad (3\text{-}42)$$

L'intensité théorique du monoisotopique calculé par la modélisation d'une distribution de poisson est donc de 5.37736E-05.

III. Point isobestique dans les distributions de Poisson

L'une des observations faites en modélisant des clusters isotopiques à différents mélanges de conditions NC et 12C est la présence d'un point isobestique (Figure 63). Ce point isobestique est bien plus flagrant pour les masses élevées d'ions observées en Top-down. Le rang isotopique de sa position (distance au monoisotopique) est non entier et donc observable sauf en traçant une courbe d'intensité. À cette valeur, quelques soit la valeur d'incorporation de mélange SLIM, l'intensité de ce point isobestique sera identique.

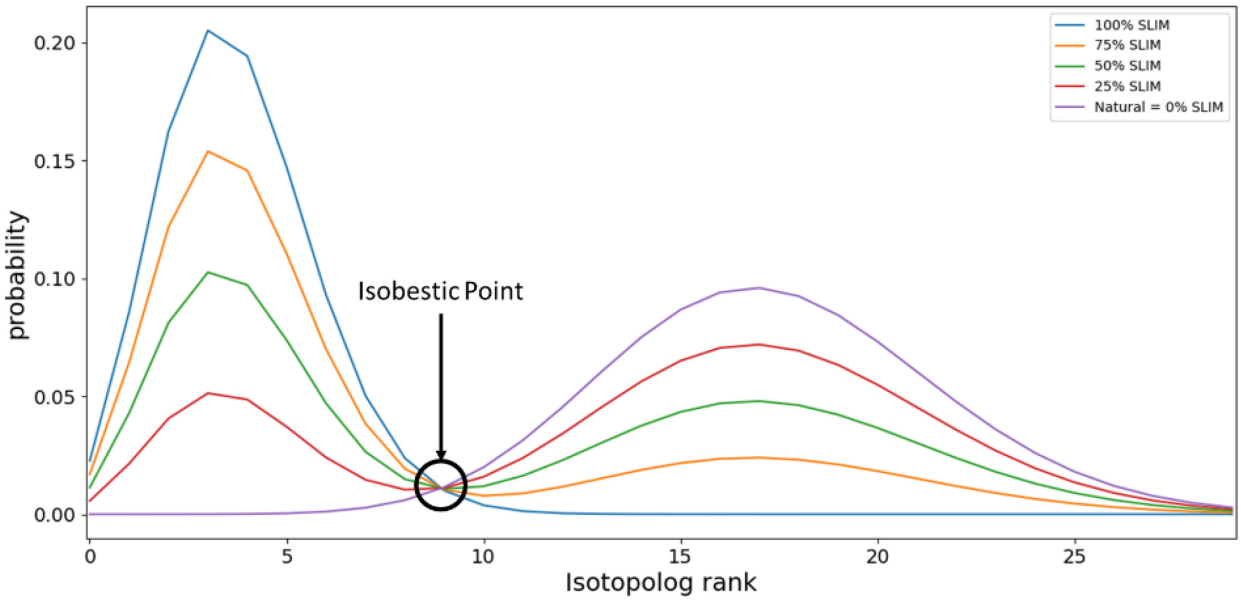


Figure 63 : Lors de la representation d'une distribution théorique de différents mélanges, un point isobestique s'observe

On peut donc poser les applications numériques du calcul de sa position. A cette valeur toutes les lois de poisson sont à égalité.

Ainsi à ce point nous pouvons écrire de manière algébrique :

$$\forall \alpha \in [0, \infty[, f(x) = \alpha \times \left[\frac{e^{-A}}{x!} \times A^x\right] + (1-\alpha) \times \left[\frac{e^{-B}}{x!} \times B^x\right] = cste \quad (3\text{-}43)$$

Pour chaque mélange de la gamme :

$$0.25 \times \left[\frac{e^{-A}}{x!} \times A^x\right] + 0.75 \times \left[\frac{e^{-B}}{x!} \times B^x\right] = 0.5 \times \left[\frac{e^{-A}}{x!} \times A^x\right] + 0.5 \times \left[\frac{e^{-B}}{x!} \times B^x\right] \quad (3\text{-}44)$$

Ce qui donne en simplifiant les termes :

$$e^{-A} \times A^{x} = e^{-B} \times B^{x} \qquad (3\text{-}45)$$

Si l'on utilise le logarithme népérien :

$$-a \times M_0 + x \times Ln(a \times M_0) = -b \times M_0 + x \times Ln(b \times M_0) \qquad (3\text{-}46)$$

$$=> \; -a \times M_0 + x \times [Ln(a) + Ln(M_0)] = -b \times M_0 + x \times [Ln(b) + Ln(M_0)] \qquad (3\text{-}47)$$

Soit en simplifiant l'équation :

$$x = M_0 \times \frac{(b-a)}{Ln(b)-Ln(a)} \qquad (3\text{-}48)$$

Validation par une résolution numérique :

$x = M_0 \times 0.000278968$

Si M_0 = 30,000 alors $x = 8.36904$ ce qui correspond au point isobestique obtenue en Figure 63.

Chapitre 5 : Au-delà de la quantification … l'identification des protéines dans les approches Top-down

La quantification est intéressante puisqu'elle permet d'observer des variations d'abondance des protéines. Toutefois sans identification associée, cette quantification reste d'un intérêt limité pour la compréhension des mécanismes biologiques mis en jeux. L'indentification de protéines dans l'approches Top-down est donc un défi majeur. Au cours de ma thèse, j'ai abordé différents aspects de la problématique de l'identification des protéines par spectrométrie de masse.

I. Stratégies reposant sur les données spectrales de la condition naturelle (NC)

La difficulté dans l'analyse des protéines intactes par spectrométrie de masse réside dans le retraitement des données et dans le fait d'associer une identification aux masses observées expérimentalement. Le couplage MS/MS permet d'obtenir des informations complémentaires aux spectres initiaux, puisque les spectres de fragmentation des protéines intactes donnent des fragments de tailles diverses et dont la mesure en masse, plus faible, permet de déterminer la séquence des objets ainsi analysés. La mesure des masses doit être la plus précise possible afin d'être la plus proche des masses théoriques et donc d'identifier le plus justement possible les séquences d'acide aminés associées. Une identification de manière plus exhaustive aborde également la recherche des protéoformes. Il s'agit d'une même protéine mais avec des variations de masses, correspondant à d'éventuelles modifications post-traductionnelles (PTMs). Ces protéoformes sont d'autant plus nombreux à la mesure que la protéine étudiée comporte dans sa séquence des acides aminés cible de PTM. Ceci pose une limitation de combinatoires lors de la recherche en banque de données, et a fortiori si les PTMs sont diverses.

Lors du retraitement des données brutes issues du spectromètre de masse en approches Top-down, les premiers spectres à être étudiés sont les MS1. Ces spectres contiennent les informations massiques des protéines, c'est-à-dire les « protéines intactes », avec les différents états de charge. Étant observés sous forme de cluster isotopique, certains très intenses seront sélectionnés pour une fragmentation puis une deuxième mesure de masse (MS2). Cela dresse donc la possibilité d'étude des protéines à plusieurs échelles par une mesure de la masse « intacte » variant selon les isoformes des molécules et des fragments de celle-ci.

Toutefois, entre la sortie de la colonne de séparation chromatographique et le détecteur de masse se déroule au sein de la source électrospray des réactions biochimiques pouvant altérer l'état naturel des protéines étudiées. Ceci a pour effet, notamment durant la recherche en banque de données, de devoir prendre en compte la spécificité biochimique des protéines analysées. Par exemple, beaucoup de protéines ne seront plus munies de la méthionine initiatrice (Ben-Bassat et al., 1987). Certaines protéines vont, sous l'effet du potentiel électrique et de la température de la source, subir une dégradation et produire des fragments.

Dans le cadre de cette thèse, j'ai utilisé trois logiciels d'identification de données Top-down. Chacun disposent de spécificité permettant un retraitement des données spectrales de manière différente. Il s'agit de *Intact (nom commercial* Intact Mass™), un logiciel utilisant les données des MS1, et de *TopPIC* et *ProSightPC,* tous deux utilisant les données de fragmentation (MS2).

I.1 Utilisation de l'application « Intact »

Intact (Protein Metrics Inc, 2021) est le premier logiciel que nous utilisons. C'est initialement un moteur de recherche en banque de données prévu pour la recherche exhaustive d'anticorps intacts Ce n'est pas une application conçue pour réaliser des identifications absolues de protéines dans un mélange. Elle permet de trouver des associations entre des séquences protéiques au sein d'une banque de séquences et pour des masses expérimentales observées. Ce logiciel utilise une petite banque de données

(composée de moins de 200 séquences) et calcule la masse exacte moyenne de chaque protéine pour comparer les masses théoriques à celles observées expérimentalement. Cette application ne permet pas de faire de la fragmentation *in silico* et n'utilise exclusivement que les masses déconvoluées contenues dans les spectres MS1. Toutefois, il est possible de procéder à une recherche de PTMs par incrémentation successives de masses de ces PTMs, plus exhaustive mais limitée par la combinatoire.

Dans un premier temps l'algorithme découpe, par reconnaissance de l'évolution de l'intensité (Total Ion Curent), le chromatogramme en plusieurs traces d'élution chromatographique correspondant aux mêmes objets. Par la suite, une somme des différents spectres de masses (MS1) est effectuée le long de la tranche du temps de rétention précédemment définie. Puis un algorithme de réduction du bruit de fond est utilisé, MaxEnt (Le, 2004). Le logiciel procède ensuite à une déconvolution du spectre ainsi produit. Enfin, dans le but d'augmenter artificiellement la résolution, un algorithme utilisant la méthode des dérivés en série est appliqué (Figure 64). Au final, le logiciel conserve, pour chaque tranche de temps de rétention, 5 masses déconvoluées intenses.

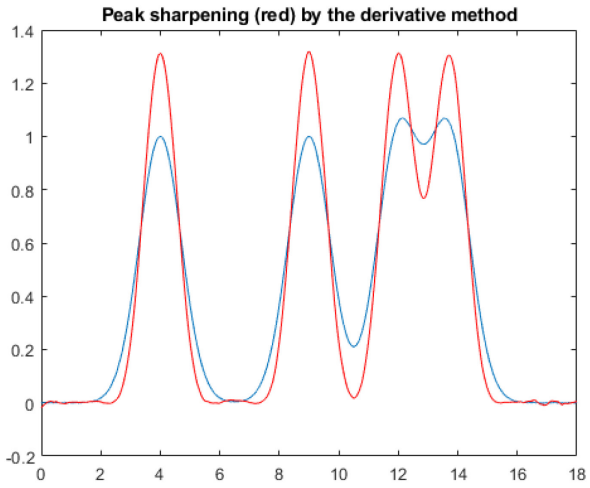


Figure 64 : Illustration de l'augmentation de la résolution via la méthode des dérivées. Les données originales sont en bleu et les données issues du traitement en rouge. Cette figure est extraite du Livre du Professeur émérite Tom O'Haver (Haver, 2022)

Dans un deuxième temps, chaque masse expérimentale sélectionnée est confrontée aux masses théoriques obtenues par la lecture de la banque de séquences protéiques.

Le fichier d'export du logiciel n'est pas utilisé car il ne contient pas les données essentielles à notre workflow, aussi seul un export particulier du tableau de données est réalisé pour obtenir la liste des Séquences, Masse et temps chromatographique des ions identifiés dans l'analyse.

I.2 Utilisation du logiciel « TopPIC »

Le deuxième logiciel d'identification TopPIC (Top-down mass spectrometry-based Proteoform Identification and Characterization) (Kou et al., 2016) que nous utilisons, procède à une recherche exhaustive des *proteoform spectrum-matches (PrSMs)* à l'aide des données spectrales de fragmentations (Figure 65).

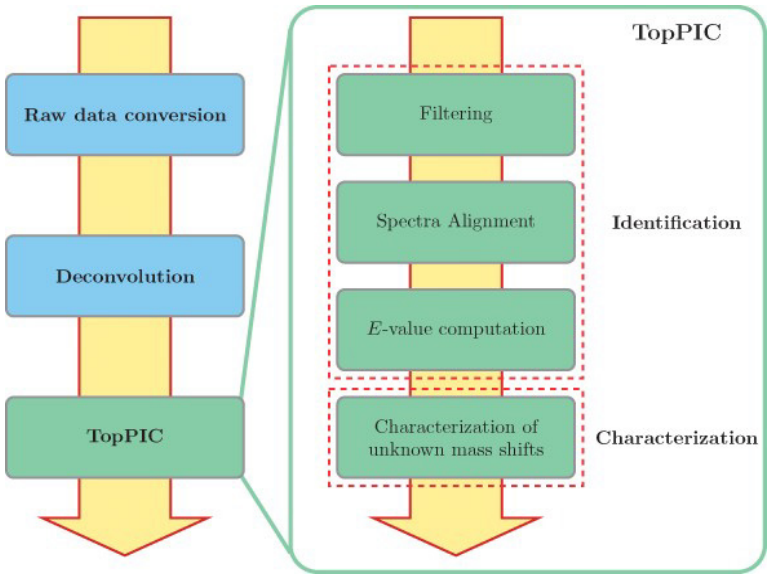


Figure 65 : Procédure du fonctionnement de TopPIC s'inscrivant dans un workflow de retraitement des données spectrales en Top-down.

La procédure d'utilisation de ce logiciel est décomposée en 3 grandes parties.

- La première est la conversion du fichier machine, format propriétaire de Thermo Fisher « .raw », en format ouvert « .mzml ». De plus, lors de cette étape, est réalisée un « peak-piking » dans le but de convertir les données de masse acquisses en mode profil en centroïdes. Cette étape est réalisée à l'aide du logiciel MSConvert.
- La deuxième partie est la déconvolution des spectres MS1 et MS2 à l'aide du logiciel topFD.
- Enfin dans la troisième partie, le logiciel TopPIC est utilisé avec en données d'entrée une banque de séquences protéiques au format fasta et le fichier de sortie de topFD, contenant la liste des spectres MS2 déconvolués (ms2.msalign).

Une recherche en banque de données est alors faite, en utilisant les masses contenues dans les spectres de fragmentation. Pour chaque masse recherchée parmi la séquence de la banque protéique, une flexibilité est ajoutée (+/- 500 dalton) dans le but de ne pas manquer l'identification de certaines formes particulières de protéines. Notamment pour chaque identification contenant un apport de masse, une identification de sa nature est déterminée puis annotée, et le cas échéant la valeur de la masse est simplement écrite dans le fichier de sortie. Toutefois, la flexibilité de masse ajoutée présente des limitations pour l'identification des modifications persistentes. Par exemple une ubiquitination est une modification qui consiste en l'incrémentation de la masse d'environ +8 500 Da des protéines.

Le logiciel, dans un premier temps, travaille avec des données de masses monoisotopiques, obtenues par ajustement dynamique sur une loi de distribution isotopique définie par une abondance naturelle et une composition chimique basée sur le modèle d'Averagine. Les données de séquençage des masses de fragmentation sont confrontées aux masses des précurseurs (MS1), c'est-à-dire des masses des protéines intactes.

L'utilisation de ce logiciel induit des biais, notamment sur le fait que les protéoformes fragmentées en MS2 (valeurs de fragmentation) sont soumises à la sélection par l'algorithme du spectromètre de masse. Pour obtenir des résultats

d'identifications satisfaisants il faut des MS2 de grande qualité et des protéines analysées de manière intense (intensité suffisamment forte pour être sélectionnées comme précurseur) et correspondant aux critères non-contrôlables de la machine (ajustement de la taille de sélection des précurseurs seulement), afin d'être sélectionnés comme précurseur.

	TopPIC	Intact Mass
Données de masse issue de	MS2	MS1
Masse Utilisée	Monoisotopique recalculé	Moyenne
Déconvolution	Spectre par Spectre (topFD)	Spectres moyennés par intervalle de temps

Tableau 8 : Comparaison entre les deux stratégies utilisées par les logiciels d'identification en Top-Down.

I.3 Utilisation du moteur de recherche « ProsightPC»

Nous avons également utilisé le moteur de recherche en banque de données Top-down, *ProSightPC* disponible au sein de l'architecture *Proteome Discover* de l'entreprise Thermo Fisher. Ce logiciel commercial payant[1] est recommandé en particulier pour des analyses Top-down effectué sur des spectromètres de masse de marque Thermo Fisher. Le développement du logiciel est toujours en cours d'optimisation et supporté en particulier par l'équipe de Neil L Kelleher (Greer et al., 2022). Ce logiciel possède un principe de fonctionnement analogue à TopPIC, c'est-à-dire que les spectres MS2 sont tout d'abord déconvolués puis interrogés à l'aide d'une banque de données. La banque est préalablement conçue à partir d'un fichier contenant les séquences protéiques voulues puis en paramétrant les modifications post-traductionnelles désirées. Cet outil est performant pour identifier des protéoformes très abondantes. Toutefois, les ions identifiés nécessitent une fragmentation de haute qualité ce qui crée un biais pour étudier de manière profonde les protéomes.

[1] https://www.proteinaceous.net/prosightpd

I.4 Biais intrinsèque des logiciels en condition 12C

Les moteurs de recherche en banque de données prévus pour le Top-down ont été développés sur des spectres naturels. En particulier, ils ont été optimisés pour extraire et interpréter des clusters isotopiques naturels dont la masse retenue pour la recherche correspond à la masse monoisotopique calculé à partir de l'isotopologue le plus intense. En effet, à cette échelle de masse, l'intensité du monoisotopique est très peu intense et donc la détermination de sa position est en pratique souvent impossible expérimentalement. Pour cela, l'abondance naturelle des isotopes ainsi que la composition chimique définie en utilisant le modèle de l'Averagine sont utilisées afin d'obtenir la masse monoisotopique attendue.

Dans le cadre du marquage SLIM, l'abondance des isotopes est changée, les spectres de la condition 12C sont alors extrêmement modifiés. Comme présenté lors de la modélisation des clusters isotopiques, le marquage SLIM provoque une modification drastique de la distribution de l'intensité des isotopologues. Ces différents logiciels d'identifications présentent donc un biais puisque l'hypothèse selon laquelle le cluster isotopique présente une abondance isotopique naturel n'est plus vérifiée. Cela induit une erreur sur le calcul de la masse monoisotopique utilisée pour la recherche en banque de données. Les identifications sont donc soit erronées soit inexistantes. Cela a été déterminée expérimentalement (cf. Tableau 10Tableau 8).

II. Une piste pour la résolution du problème : l'exploitation de la reproductibilité des séparations chromatographiques pour l'extraction des données d'identifications

Nous avons travaillé sur l'optimisation de la reproductibilité des séparations chromatographiques des protéines dans différentes conditions de marquage tSLIM, comme montré ci-après (Partie 3Chapitre 6 :III.5). Nous avons décidé d'utiliser les données d'identifications obtenues dans la **condition naturelle**, afin d'attribuer l'identification des protéines de même temps de rétention issues des conditions 12C.

Cette identification est essentielle. Premièrement afin de garantir la séquence protéique et la formule chimique utilisée dans les calculs. Deuxièmement car cela nous permet d'obtenir une description précise de la position des isotopologues et donc de pouvoir extraire le plus justement l'intensité expérimentales des isotopologues dans le but de procéder à une quantification tSLIM par l'intermédiaire des calculs algébriques présentés ci-dessus.

Pour cela deux logiciels d'identifications *Intact* Mass et *TopPIC* ont été utilisés, maximisant ainsi le nombres d'entrées utilisées dans le but d'extraire des valeurs quantitatives. Ainsi, les deux fichiers de sortie des logiciels d'identification et les spectres bruts servent de données d'entrées à un algorithme que j'ai développé (Figure 66). L'algorithme est écrit en C^{++} pour allier rapidité, flexibilité et compatibilité avec les autres outils bio-informatiques de la communauté. De plus, son implémentation dans un nœud KNIME (écrit en java) est facilité afin d'ouvrir son utilisation à un public non expert. L'algorithme de quantification de protéoformes identifiées se décompose en plusieurs parties.

Dans un premier temps, les fichiers d'entrée contenant les identifications sont lus puis stockés en mémoire vive. Concrètement, l'algorithme permet une lecture des différents fichiers, séparés par des virgules, spécifique à chaque logiciel d'identification. En particulier le fichier produit par le logiciel *Intact Mass* est un tableau contenant une liste des temps de rétention minimum et maximum de chaque tranche chromatographique. Pour chacune d'entre elles, cinq masses obtenues par la déconvolution sont présentées munies des séquences et protéoformes dans le cas d'une identification. De manière analogue, le fichier d'identification produit par le logiciel TopPIC présente un tableau contenant uniquement les valeurs descriptives des spectres de fragmentation identifiés. Dans ce cas, il faut associer la liste des identifications TopPIC fondées sur les masses des fragments produits en MS2 avec les spectres MS1 contenant les valeurs de masse et les intensités des protéines. Ces données sont déterminantes pour extraire les valeurs expérimentales requises par la quantification tSLIM des protéines.

La combinaison de ces deux approches de descriptions des données spectrales permet de caractériser au mieux les spectres expérimentaux en termes

d'identifications. En effet, l'enjeu principal pour répondre à l'objectif de la quantification SLIM est d'obtenir pour chaque identification de protéoforme la masse utilisée par notre algorithme dans le but d'extraire les valeurs d'intensité expérimentales des isotopologues.

La deuxième partie de mon algorithme consiste donc à la résolution algébrique des valeurs quantitatives précédemment décrites. En particulier, les données expérimentales extraites durant la première partie sont confrontées aux données théoriques.

```
Read the fasta file and store in memory
IF the file is "intact"
    allocate space in memory FOR later
    Read the file and store in memory
    variable MaxRT is EndTime*60
    variable MinRT is StartTime*60
    open the rawFile and store in memory
    From the rawfile extract the correct scan number from the retetnion time provided
    From the fasta extract the correct sequence according to acession
    FOR each sequence
        compute the chemical composition
    ENDFOR
    FOR each Charge from 5 to 50
        IF calculated mass over charge is less than 500
            break
        ELSEIF calculated mass over charge is more than 4500
            next
        ELSE
            execute the function get_ms_clusters_dist_Mono
            with parameter experimental_distribution_isotopic,experimental_distribution_isotopic_RT,Time_RT,SNb,File_Offset, MASS, Charge, ITmin, MinRT,MaxRT,Ep)
            FOR each MS1 Extracted
                IF the memory is not empty
                    FOR each Isotopologs extracted
                        Isotopologs intensity is local maximum
                    ENDFOR
                ENDIF
            ENDFOR
            FOR each Isotopologs extracted
                Do gaussian feat
                keep in memory the parameters
                IF the gaussian area is more then 10000
                    print the Isotpologs descritpion data in a file
                ENDIF
            ENDFOR
            FOR each Isotopologs extracted and contained in the ProteoFORm database
                experimental values are gaussian area;
                Temp_Proteins.Exp_Dis = experimental_distribution_isotopic; // MZ couple IT
                Grand_A is set to 0.00058338 * monoisotopic mass
                Grand_B is set to 0.00010501 * monoisotopic mass
                IF cluster intensity summed is more 1000
                    Keep in memory the proteoFORm
                ENDIF
            ENDFOR
        ENDELSE
    ENDFOR
NEXT proteoFORm
```

Dans la dernière partie, les valeurs quantitatives ainsi déterminées sont écrites dans un fichier texte. Ces valeurs critiques sont utilisées par la suite pour procéder à des statistiques descriptives.

```
OPEN the file "Quantification.tsv"
    FOR proteoform in kept in memory
        PRINT in a file the various variable
            MHplus,Charge,Sequence,Description,Sequence.size,Feature,AA_BeFORe_Pos,AA_After_Pos,Chem,Target_Mass,Mass,Residus,
            Carbone,Hydrogen,Nitrogen,Oxygen,Sulfur,Alpha,Grand_A,Grand_B,FinalFM_Poisson,FinalFM_Midas,Best_Cal_Score,SumIT
        FOR each Isotopologs extracted and contained in the proteoform database
            PRINT Gauss_area
        ENDFOR
    ENDFOR
CLOSE the file
```

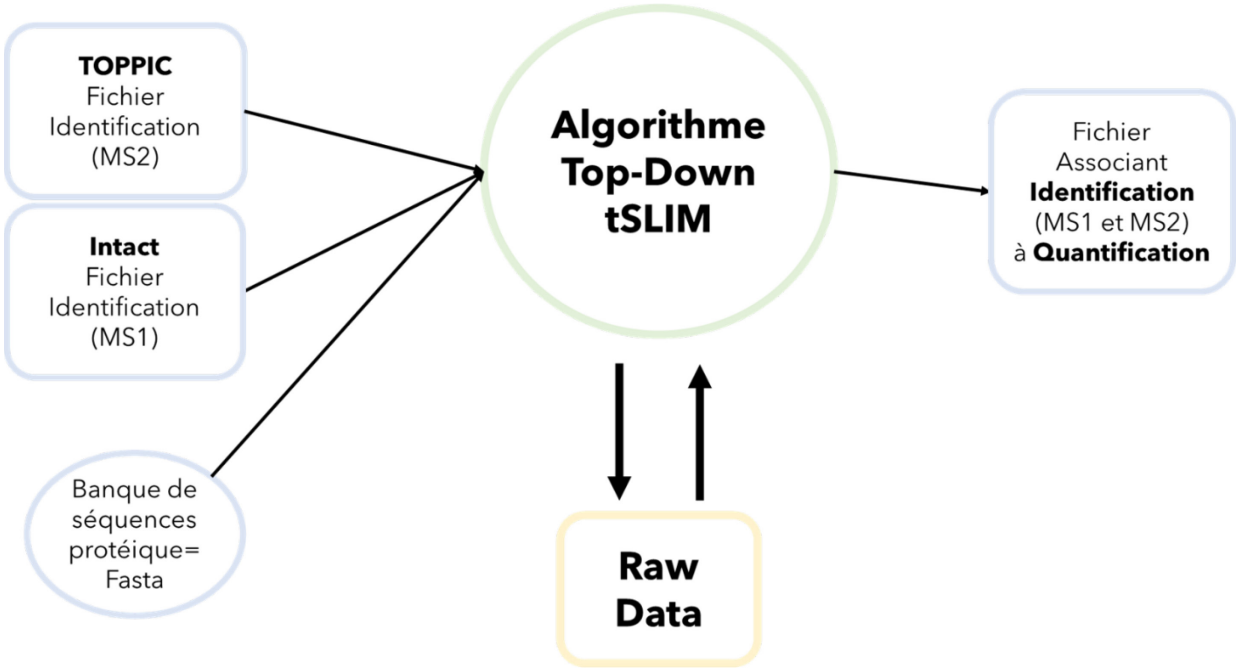


Figure 66 : Représentation schématique de l'algorithme de quantification développé durant cette thèse.

III. Essais d'Identification à partir des protocoles de quantification tSLIM

Notre workflow de quantification en tSLIM est totalement indépendant des logiciels existants qui ajoutent des étapes supplémentaires et des biais dans le retraitement des données expérimentales. En particulier, nous avons inversé la procédure classiquement utilisée en Top-down. La quantification est effectuée dans un premier temps puis nous procédons à une identification par une méthode originale que

nous avons développée. À cet effet, durant l'étape de quantification, nous avons déterminé le coefficient alpha, correspondant à la fraction molaire des protéines non-marquée. Ce facteur a été déterminé par un ajustement d'une distribution de Poisson décrite par la masse neutre de la protéine (Partie 3Chapitre 3 :II.1).

A l'aide de cette masse neutre, nous sommes en mesure de proposer l'identité de cette protéine. En effet, nous avons développé un algorithme d'identification de type *de novo*, en générant pour une masse déterminée toute les séquences protéiques possibles à partir d'un protéome de référence. Pour cela, nous avons développé un algorithme semblable à celui du logiciel Biomarker@IJM (Mestivier et Camadro non publié) qui permet de produire une banque de donnée exhaustive dans le but d'effectuer une interrogation *de novo* avec l'algorithme (OMMSA). Concrètement, pour un protéome donné, une fenêtre glissante permet l'extraction de toutes les protéines ou morceaux de protéines dont la séquence en acides aminés est compatible avec la masse recherchée. Une marge d'erreur est ajoutée pour plus de souplesse, mais il est à noter qu'elle influe grandement sur le nombre d'entrées finales. Ce séquençage brut des données de masse permet d'obtenir un éventail de séquences protéiques et ce de façon dépendante de l'organisme étudié. Leurs compositions en acides aminés varient de manière extrême et les compositions chimique également (Figure 67).

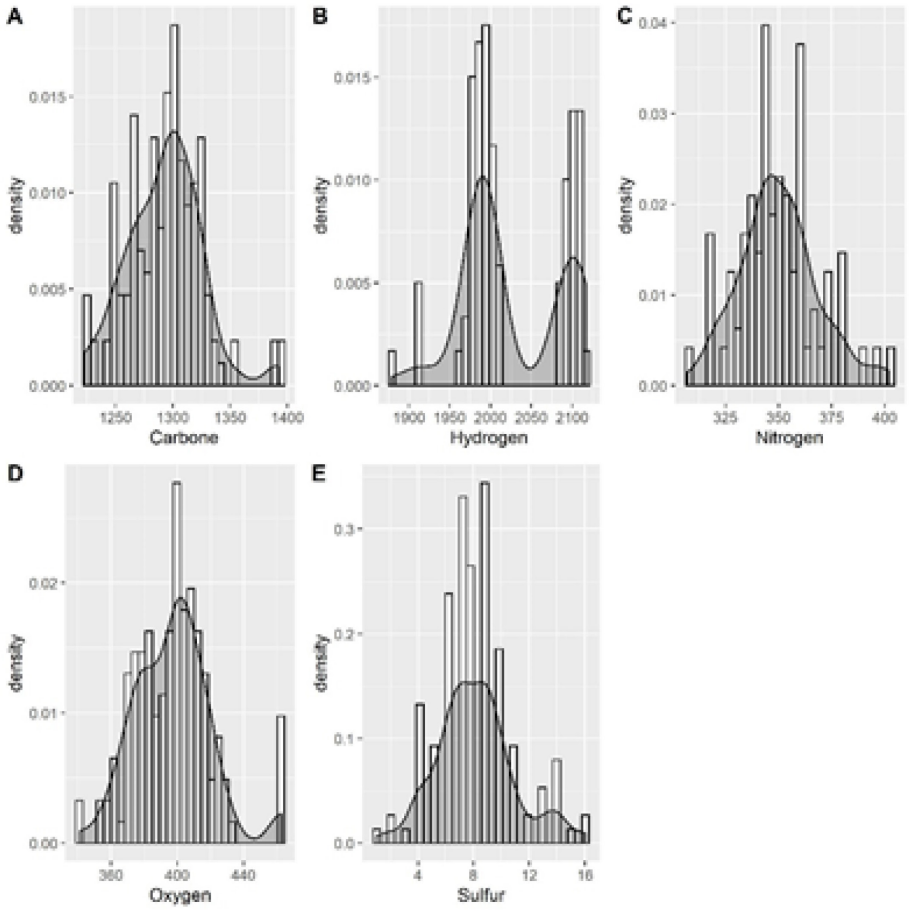


Figure 67 : Distribution du nombre de chaque élément chimique pour les séquences obtenues ayant une masse de 29066.37 Da +- 0.5 Da parmi la banque de séquence du protéome de S. cerevisiae.

Puis pour chaque séquence, nous calculons la distribution isotopique théorique du cluster isotopique avec une haute résolution fournie par l'algorithme MIDAs (la valeur du coefficient alpha, l'abondance des deux conditions étant fixée). Enfin, pour chaque cluster isotopique théorique produit, l'écart quadratique moyen avec le cluster isotopique expérimental est déterminé. Parmi les séquences candidates, celles possédant le RMSD le plus faible sont celles qui expliquent de manière la plus probable les clusters isotopiques observés. Dans le but d'augmenter la détermination de l'identité et dans le cas où une fragmentation MS2 aurait eu lieu une fragmentation *in silico* est réalisée.

Si l'intensité générale du cluster varie entre les états de charges, les valeurs des ratios des intensités des isotopologues ne varient que très peu. Nous admettons que l'abondance isotopique des molécules reste constante pour les différents fragments de

la molécule quelles que soient leurs masses. Ceci implique que les spectres de fragmentations peuvent également être utilisés afin de réaliser une quantification. De plus cela présente l'avantage d'obtenir des clusters isotopiques dont la masse est réduite par rapport à la masse de la protéine précurseur facilitant ainsi les calculs qui seraient alors semblables à ceux utilisés en bSLIM.

Chapitre 6 : Développements expérimentaux en réponse aux défis du Top-down

L'étude des protéoformes implique un changement de dimension dans l'analyse par le couplage chromatographie en phase liquide et spectrométrie de masse. Les protocoles d'acquisition sont à redéfinir et à mettre au point. Les problématiques dans le développement d'une approche différente de celle des protocoles habituels sont nombreuses et leurs implémentations nécessitent également une partie de retraitement de données nouvelles.

I. Optimisation des protocoles de séparation chromatographique des protéines intactes

Nous tirons avantage de deux grandes avancées : l'utilisation de colonnes inédites de séparation des protéines et l'utilisation d'un agent chimique augmentant la sensibilité lors de l'analyse par le spectromètre de masse. Pour comprendre tout l'enjeu et la réflexion à développer pour pouvoir réaliser une séparation de protéine en solutions, il faut comprendre les caractéristiques des colonnes. Dans cette partie, je vais présenter les innovations de la plateforme réalisé par Manel Khelil berbar, Laurent Lignières, Véronique Legros et Guillaume Chevreux (en cours de publication). Celles-ci m'ont permis d'obtenir des résultats originaux de séparation d'un mélange complexe de protéines ribosomiques de levure (Partie 3Chapitre 6 :III.5).

Les colonnes de chromatographie sont déterminées par les caractéristiques physiques (taille, diamètre, température) mais également par les caractéristiques chimiques de la phase stationnaire employée. Dans l'objectif de maximiser la résolution de la séparation chromatographique, les paramètres de la phase stationnaire sont à étudier. L'efficacité d'une colonne chromatographique est décrite par le nombre de plateaux théoriques qu'elle propose. Un plateau se définit comme un découpage théorique de la colonne où à un instant *t* la phase mobile admet un équilibre

avec la phase stationnaire. Le long de la colonne se trouve ainsi un nombre de plateaux noté N. Dans le but de comparer les colonnes entre elles, on utilise une valeur dite de hauteur équivalente en plateaux théoriques (noté H) :

$H = \frac{L}{N}$ (L longueur de la colonne)

Dans le but d'obtenir une condition de séparation idéale le paramètre choisi vise à obtenir une hauteur équivalente minimale (Figure 68).

$$H = A + \frac{B}{v} + (C_M + C_S) * v \qquad (3\text{-}49)$$

Avec :

A, le paramètre de la diffusion de Eddy

B, le coefficient de diffusion longitudinale

C_M, e transfert de masse de la phase mobile

C_S, le transfert de masse de la phase stationnaire

V, la vitesse de la phase mobile

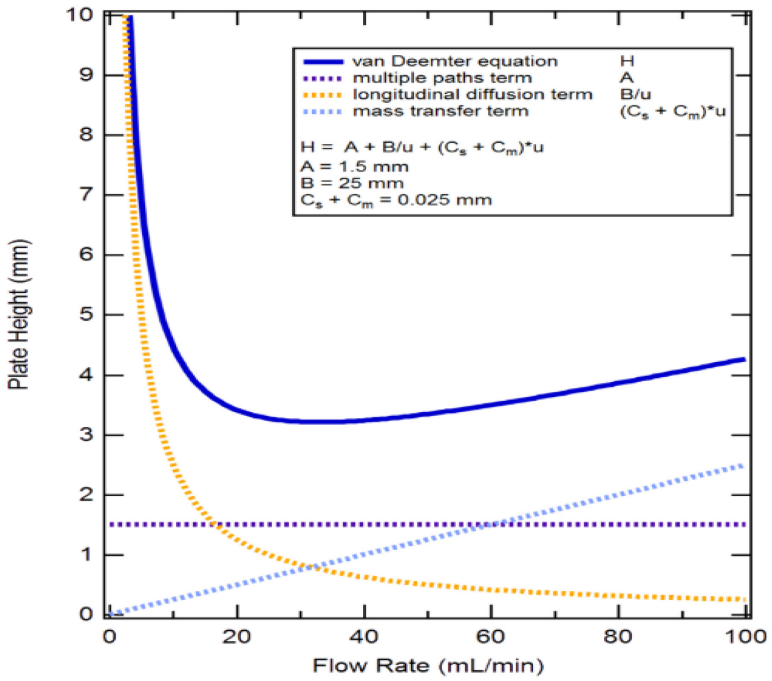


Figure 68: Illustration des équations de Van Deemter. Cette illustration est extraite de la page Wikipédia[1].

[1] https://en.wikipedia.org/wiki/Van_Deemter_equation

Cette équation (3-49) nous permet de comprendre que réduire les paramètres d'effet de masse (C_M et C_S) sont l'un des moyens d'augmenter la résolution, la sélectivité et l'efficacité des colonnes.

Le développement des systèmes chromatographiques à haute pression UPLC (Ultra Performance Liquid Chromatography), a permis d'augmenter significativement la résolution et la sensibilité des séparations. Pour cela, la stratégie de réduire le transfert de masse est rendue possible par la diminution de la granulométrie de la phase. La réduction de la granulométrie permet de gagner en homogénéité lors de la séparation et donc de minimiser la perte de résolution. De plus, cela autorise une augmentation du débit de la phase mobile sans impacter la sensibilité et permet un gain de temps. Le débit étant toujours selon les équations de Van Deemter fonction du diamètre de la colonne.

Un autre moyen de réduire le paramètre de transfert de masse est l'augmentation de la taille des pores des billes de la phase stationnaire. A titre de comparaison, dans les colonnes utilisées pour la séparation des peptides (C18) le diamètre des pores est de 120 à 180Å. Le diamètre des pores de la phase stationnaire composant les colonnes impliquées dans la séparation des protéines (C4) est de l'ordre de 300 Å.

La colonne Bioresolve™ mAb RP Polyphenyl est une colonne alliant deux technologies plus adaptées à l'analyses de molécules biologiques massives. En effet, cette colonne dispose de particules dont le diamètre des pores est de 450 Å (donc le transfert de masse est réduit) d'une part, et les particules greffées d'une surface chimique phényle disposent de la technologie « core shell » d'autre part. Cette colonne est initialement développée pour des applications du secteur pharmaceutique, comme la séparation d'anticorps. Le diamètre de la colonne est de 2mm ce qui implique un débit de la phase mobile de l'ordre de 250µl/min. Toutefois, dans le but de réaliser des séparations de protéines en vue d'analyse par spectrométrie de masse Top-down, la plateforme de l'Institut Jacques Monod a développé une collaboration avec la société Waters afin d'obtenir des colonnes identiques mais de diamètre moindre. Ainsi le

diamètre de ces nouvelles colonnes de 300µm implique un débit de 5µl/min, et un fonctionnement de l'appareil chromatographique UHPLC est en mode capillaire.

Le principe de fonctionnement de la technologie « core shell » (Figure 69) repose sur la composition multicouche des billes de silice, avec un noyau solide au centre entouré d'une couche poreuse extérieure (Guillarme & Fekete, 2013). Cette technologie permet d'augmenter l'homogénéité des composées durant la séparation, garantissant ainsi une haute résolution.

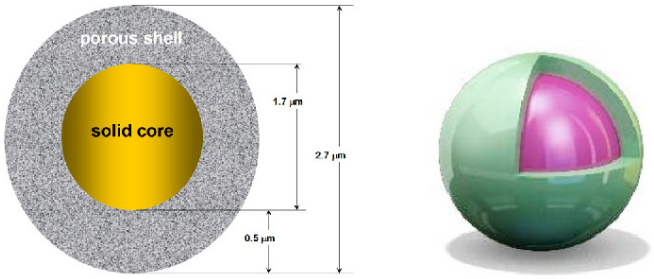


Figure 69: représentation de la technologie « core shell » composant les colonnes chromatographiques de séparation de protéines utilisée durant cette thèse. Cette figure est extraite de (Guillarme & Fekete, 2013)

La phase stationnaire de cette colonne possède une originalité puisqu'elle est enrobée d'une surface poly-phényle à la différence des phases classiquement utilisées constituées de silice liée à une chaîne de butyle (C4). L'article de (Bobály et al., 2018) dresse une comparaison entre plusieurs colonnes et permet de conclure sur les effets positifs d'utiliser une telle colonne. Premièrement, la capacité des pics chromatographique est augmentée par rapport à une colonne C4 cela entraine que les pics d'élution seront plus résolutifs. Deuxièmement, la colonne est plus séparative car deux composés dont le point d'élution est proche seront d'autant mieux discriminés. Troisièmement, cela permet de réduire la concentration de l'agent d'appariement des ions (donc la sensibilité en spectrométrie de masse augmente). En résumé, les caractéristiques techniques de cette colonne, sont : nanoEase© M/Z bioresolve™ mAb RP Polyphenyl Column, le diamètre des pores est de 450 Å, la taille des billes est de 2.7µm et la colonne possède un diamètre de 300µm ainsi qu'une longueur de 150mm.

Il est à noter toutefois que le dispositif de chromatographie en phase liquide est suivi d'une analyse en tandem par spectrométrie de masse, le couple LC-MS. Toute modification de la chimie de l'échantillon dans le but d'obtenir des séparations de meilleure qualité aura des effets potentiellement délétères sur la sensibilité et la capacité de détection par le spectromètre de masse. Un compromis est donc à trouver. En particulier, les agents d'appariement des ions tels que l'acide formique ou le TFA (acide trifluoroacétique) ont pour but de rendre homogène la surface ionique des protéines. Cependant, si cela permet d'obtenir des séparations chromatographiques de grande qualité (réduction des interactions *silanol*). Cela s'effectue au détriment de l'ionisation dans la source d'ionisation du spectromètre de masse. En effet, l'agent d'appariement des ions aura tendance à s'évaporer en captant des proton H^+ réduisant ainsi l'ionisation des composés. Cela se traduit par une sensibilité de l'analyse par le spectromètre de masse divisée par 10 dans le cas d'utilisation de TFA. Cependant, cette réduction de sensibilité peut être diminuée en utilisant d'autres acides. En effet, le protocole utilisé sur la plateforme de l'Institut Jacques Monod utilise de l'acide difluoroacétique (DFA), réduisant par trois seulement la sensibilité de l'analyse en masse et permettant donc d'obtenir des séparations de qualité (Figure 70) et des spectres de meilleure résolution (Lardeux et al., 2021)(Nguyen et al., 2019).

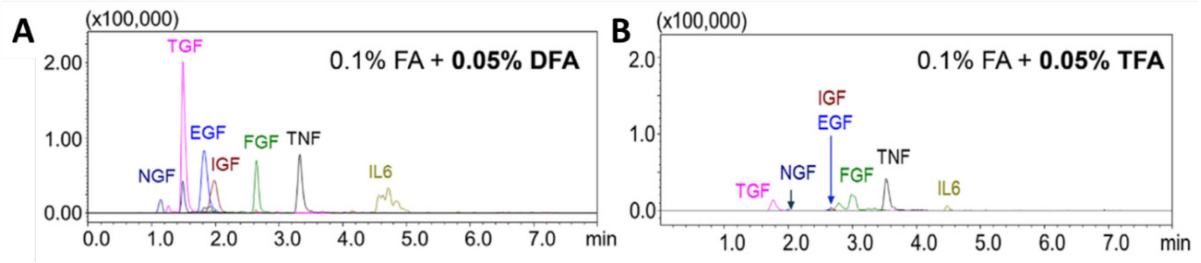


Figure 70 : séparation LC-spectrométrie de masse de sept facteurs de croissance. L'effet bénéfique du DFA (A) par rapport au TFA (B) dans la composition de la phase mobile est démontré. Cette figure est extraite de (Maráková et al., 2020)

Toutefois, l'un des points à soulever dans l'utilisation de ce genre de colonnes est que cela limite la quantité de matière initiale. Dans le cadre du couplage LC-MS, le spectromètre de masse nécessite d'analyser une quantité définie et constante le long du temps de rétention. Chaque pic chromatographique se doit d'être suffisamment intense, aussi la relation entre la quantité de matière déposée en amont de la séparation

et en sortie de colonne est donc optimisé au maximum. Cependant, notre système actuel de séparation chromatographique très sensible ne permet pas le chargement d'une quantité de matière supérieure à 50µg, si l'on veut obtenir une séparation. Bien que présentant l'avantage d'une grande sensibilité et d'une grande résolution la surface d'échange de ce type de colonne est plus limité que celle des colonnes classiques dont la quantité de matière éventuelle à injecter est plus grande.

II. Optimisation des acquisitions en masse :

Dans le but de développer les méthodes d'analyses et les protocoles de séparation en chromatographie en phase liquide, plusieurs échantillons qualifiés de « standards » et de complexité graduelle ont été sélectionnés.

Le Pierce™ Intact Protein Standard Mix est un échantillon standard constitué d'un mélange de 6 protéines recombinantes purifiées en solution (Tableau 9). Le ProtMix contient des protéines de différentes tailles permettant une gamme de masse étendu allant de 9 kDa à 68 kDa.

Cet échantillon reproductible a permis le développement des conditions optimales de séparation des protéines et d'utilisation de la nouvelle colonne. Pour cela, un programme d'étude a été conduit dans lequel la même quantité de protéines est injecté dans les conditions chromatographiques variables. Différents paramètres ont été essayés, la température de la colonne, le débit et la teneur en acétonitrile dans la phase mobile. Durant ce cas d'étude, trois grands indicateurs ont été utilisé afin de décrire les séparations (Figure 71) :

- Le facteur d'élution, correspondant à la moyenne de temps de rétention des composées.

- Le facteur de séparation, correspondant à l'écart entre les différents pics d'élution des composés. Plus l'écart est grand, plus la colonne est séparative.

- Le facteur de résolution, correspondant à une valeur quantitative décrivant les pics d'élutions chromatographique.

TABLE 1. Thermo Scientific Pierce Intact Protein Standard Mix. The theoretical masses include known sequence variants and disulfide bonds. Uniprot accession numbers correspond to the original protein sequences.

Protein Name	Protein Accession	Theo. Average Mass (Da)	Theo. Mono Mass (Da)
Human IGF-I LR3*	P05019 (40-118)	9111.47	9105.34872
Human Thioredoxin	Q99757(60-166)	11865.52	11858.04393
Streptococcus dysgalactiae Protein G	P06654(223-413)	21442.61	21429.75915
Bovine Carbonic Anhydrase II*	P00921	28981.29	28963.6881
Streptococcus Protein AG (*chimeric)*	P02976, P19909	50459.74	50429.84641
Escherichia coli Exo Klenow	P00582(324-928)	68001.15	67959.42515

*Proteins may undergo partial deamidation in acidic conditions.

Tableau 9 : Composition et masse des six protéines contenues dans l'échantillon « Thermo Scientific Pierce Intact Protein Standard Mix ».[1]

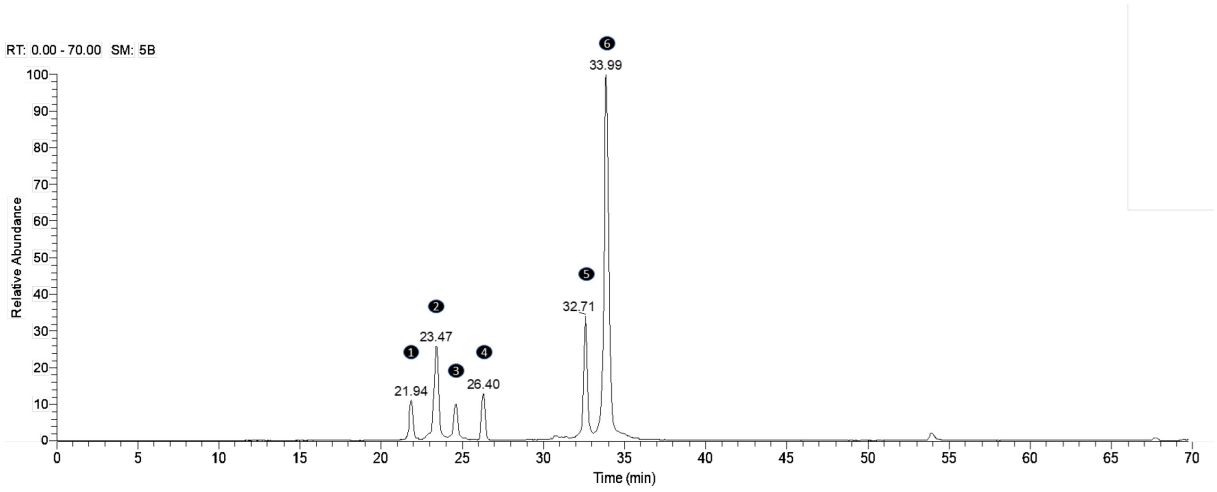


Figure 71 : Séparation du ProtMix, pour le protocole optimum : un débit de 6.00µl/min, une phase mobile composée de 65% d'ACN et une température de colonnes de 60°C.

II.1 NIST anticorps

Une fois les paramètres optimums de séparation déterminés, une première application biologique a lieu dans le but d'étudier l'identification de PTMs. Les anticorps sont des macroprotéines de plusieurs milliers de dalton et comportant un nombre important de PTMs, ce qui en fait un sujet d'étude idéal pour l'observation de modification post-traductionnelle en approches Top-down. C'est la raison pour laquelle l'étude de l'anticorps monoclonal (mAbs) standard a été développé par le

[1] https://www.thermofisher.com/order/catalog/product/A33526

National Institute of Standards and Technology (NIST). Cet anticorps recombinant de IgG1k humaine est très utilisé en routine comme sujet de démonstration et pour le développement de nouvelle méthode d'étude pharmaceutique. Dans ce contexte, l'anticorps massif est digéré à certaines zones cibles, clés dans la discrimination des PTMs (Chevreux et al., 2011). Ce projet a permis la démonstration du fort pouvoir résolutif de cette colonne dans les conditions de séparation développées auparavant.

III. Top-down Validation expérimentale : Jeux de données Ribosome

III.1 Ribosomes chez la levure

La levure est un organisme modèle au laboratoire, car son métabolisme est très semblable aux cellules humaines. Les conclusions biologiques, tirées d'expériences sur cet organisme, peuvent être transposées chez l'Homme. L'étude des ribosomes chez la levure permet donc d'étudier au mieux la physiopathologie de certaines mutations impliquant la voie de biosynthèse des ribosomes dans les lignées cellulaires humaines. L'ensemble des pathologies humaines qui touchent la structure et la fonction des protéines ribosomiques et donc du ribosome sont critiques et sont appelés ribosomopathies. Par exemple, la DBA (Anémie de Blackfan-Diamond) et la SDS (Syndrome de Shwachman) sont deux pathologies humaines dont la nature et la localisation des mutations parmi les protéines ribosomiques ainsi que leurs effets biologiques ont été décrites (Ellis et al., 2010). Une étude plus détaillée de la structure de ces protéines permettrait de mieux comprendre les mécanismes mis en jeux, lors de telles pathologies (Benjamin et al., 1998; Videler et al., 2005). Notamment, la protéomique fondée sur de la spectrométrie de masse en protéines intactes, ouvre une dimension supplémentaire pour l'étude de ces protéines clés dans la vie de la cellule (Van De Waterbeemd et al., 2018).

III.2 Biosynthèse des ribosomes cytosoliques chez la levure

Le ribosome cytosolique de la levure est une entité constituée de 2 sous-unités totalisant 79 protéines ribosomales (Nakao et al., 2004) et 4 brins d'ARN formant un complexe (Maguire & Zimmermann, 2001; Ramakrishnan & Moore, 2001; Sanyal & Liljas, 2000). En 1998, lors du séquençage du génome de *S. cerevisiae,* il a été démontré que 137 gènes codent pour 46 protéines pour la grande sous-unité d'une part, et 33 protéines pour la petite sous-unité d'autre part (Planta & Mager, 1998) (Figure 72). La grande sous-unité 60S est appelée ainsi à cause de la valeur de son coefficient de sédimentation (unité Svedberg) et de la même manière, la petite sous-unité est nommée 40S.

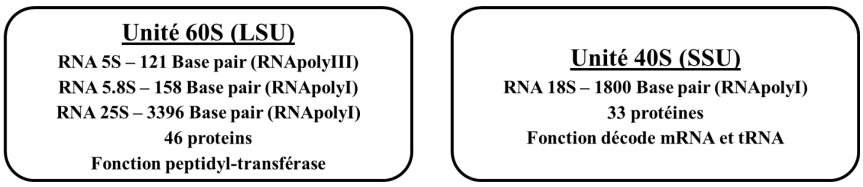


Figure 72: Les 2 sous-unités constituant le ribosome de la levure S. cerevisiae. Les données sont issues de (Woolford & Baserga, 2013)

La biosynthèse du complexe ribosomique requiert une activité de 79 ARN et plus de 200 protéines aidant à l'assemblage de l'entité. La production des ribosomes est étroitement liée au cycle cellulaire. Durant la phase exponentielle, une seule cellule de levure synthétise jusqu'à 2000 ribosomes par minute (Warner, 1999). Une dérégulation de l'assemblage et la production des ribosomes au sein de la cellule implique des phénotypes particuliers. De plus, la mutation de certaines protéines ribosomales est léthale. Notamment, 64 protéines ribosomales ont été décrites comme extrêmement sensibles pour la formation du complexe (Steffen et al., 2012).

Pour développer la méthode d'analyse des protéines intactes en spectrométrie de masse, nous avons choisi le ribosome de levure comme modèle d'étude. En effet, les 79 protéines composant le ribosome de *S. cerevisiae* possèdent des masses dans une gamme intéressante (Figure 73) allant de 3 kDa pour L47 à 45 kDa pour L3 (Masse monoisotopique = 45 948.72 Da). De plus, la préparation d'extraits cellulaires et la purification des protéines ribosomiques est aisée ainsi que reproductible en concentration et en qualité.

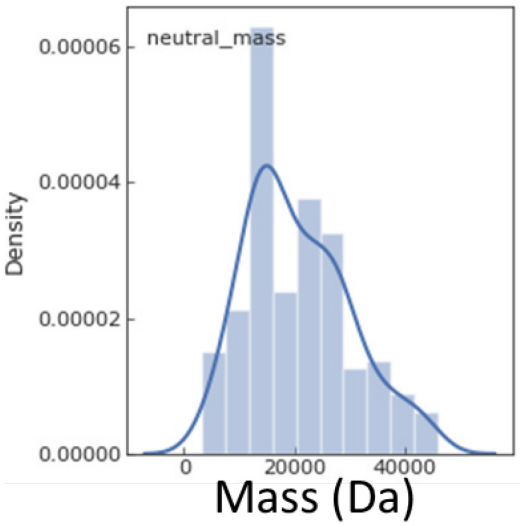


Figure 73 : Représentation de la gamme de masse des protéines ribosomales de la Levure S. cerevisiae.

III.3 Préparation de l'échantillon

Les protocoles d'analyse des protéines intactes par spectrométrie de masse requièrent d'injecter des échantillons sans contaminants, notamment par des résidus de polymère plastique de type PEG (polyéthylène glycol), ni de milieux trop salins induisant de la suppression ionique. Les échantillons sont ainsi préalablement nettoyés par précipitation et/ou dessalage par membrane semi-perméable. Tout traitement biochimique doit être fait avec soin afin de ne pas perturber la chimie de l'échantillon, autre que la dénaturation. En effet, l'étape finale et critique dans cette analyse est la dénaturation, effectuée ici au dernier instant avant l'injection sous l'action d'acide DFA (acide difluoroacétique). Bien que ce type d'analyse soit extrêmement sensible, les protocoles de mesure nécessitent toutefois d'avoir des protéines en concentration suffisante. En effet, avec un regard probabiliste, comparé au Bottom-up, la probabilité d'observer des protéines intactes est plus faible que celle d'observer un des nombreux morceaux qu'une protéine peut générer après digestion.

- **Gradient de sucrose**

Albert Heck, décrit un protocole pour purifier les protéines ribosomiques afin de permettre leurs compatibilités pour l'analyse en spectrométrie de masse (Van De Waterbeemd et al., 2018). Ce protocole, utilisant une étape d'ultracentrifugation sur un gradient de sucrose, a été adapté à notre expérience à l'aide d'autres protocoles du

laboratoire (Panasenko, 2012). Ainsi, l'échantillon cellulaire après une lyse complète à l'aide de microbilles, est déposé sur un gradient de sucrose préalablement préparé. Puis une ultracentrifugation permet de récupérer les fractions ribosomales (40s et 60s). Le gradient est composé de cinq phases de différents pourcentages en sucrose (7% - 17% - 27% - 37% - 47%), déposées l'une sous l'autre afin de former, par une différence de densité, un gradient. Le gradient sera ensuite uniforme après une nuit de stabilisation.

Le gradient contenant les protéines séparées par l'ultracentrifugation est fractionné en prélevant, à partir du fond du tube, des fractions de volume constant, à débit constant, en utilisant une pompe péristaltique. Ces fractions de 200µl sont déposées sur une microplaque de 96 puits. Ensuite, une lecture de l'absorbance est réalisée, au spectrophotomètre à longueur d'onde de 354nm, afin de mesurer la concentration en ARN. Celle-ci peut être corrélée à la concentration des protéines ribosomiques. Ainsi sur le profil suivant, les différentes fractions contenues dans les puits annotés sont récupérées car contenant les ribosomes. Un dosage des protéines dans les puits sélectionnés est effectué afin de calculer les concentrations correctes en protéines. Ces mesures permettent de réaliser un gel SDS-Page, mettant en évidence la gamme de masse des protéines présentes dans l'échantillon (Figure 74).

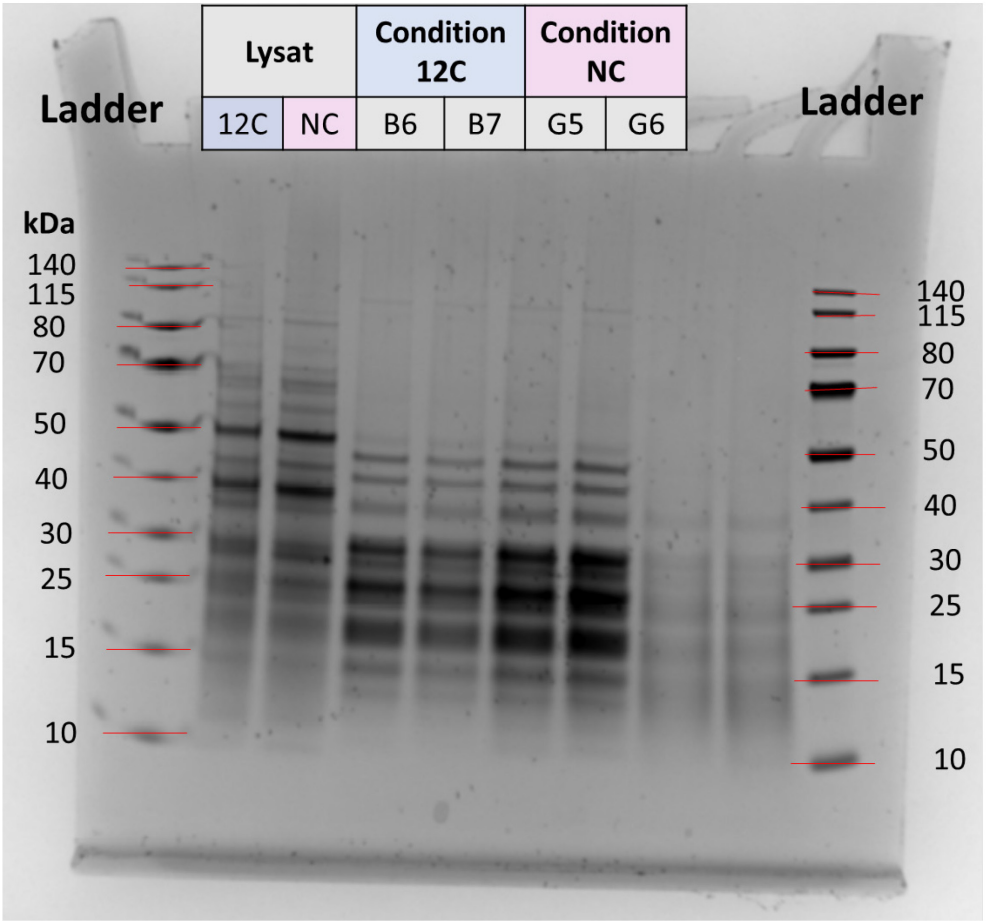


Figure 74 : SDS-PAGE des lysats cellulaires et des puits sélectionnés (B6, B7 et G5, G6) représentant les échantillons purifiés de protéines ribosomales en répliquât biologique et technique.

Les échantillons ont été dessalés à l'aide d'un concentrateur sur membrane semi-perméable, par centrifugation (Amicon) avec un seuil de rétention théorique de 10kDa. Puis 5 lavages sont réalisés par l'ajout de 500µl (25mM NH_4COO_3) puis une concentration à 50µl consécutif par centrifugation. Des mélanges tSLIM de quantités identiques de ribosomes ont été préparés pour les conditions 12C et NC (1:0, 0.75:0.25, 0.5:0.5, 0.25:0.75 et 0:1 vol/vol). Chaque mélange a été analysé par chromatographie en phase liquide, en tandem avec un spectromètre de masse en approche intacte (Top-down LC-MS/MS).

III.4 Acquisition des données en Top-down

Chromatographie :

L'HPLC était un appareil Proxeon U3000 équipé d'un module capillaire, le débit était de 6.00µl/min avec un gradient de 5-80% en 70mins, consistant à atteindre 5% d'ACN en 10min, 25% en 40min, 45% en 10min puis finalement 80% d'ACN. La composition des tampons A et B était respectivement de 2% d'ACN, 0.1% DFA et de 80% d'ACN, 0.1% DFA. La colonne BioResolve RP mAb Polyphenyl (2.7 mm, 150 mm * 2.1 mm, 450 Å) Polyphenyl de Waters a été maintenue à une température de 60°C.

MS :

Les protéines ont été analysées à l'aide d'un spectromètre de masse QexactivePlus de Thermo Fisher en mode positif, Orbitrap/Orbitrap. La tension de la source CID est réglée à 15.0 eV, un état de charge par défaut de 20, 2 microscans et une résolution de 140 000, l'AGC target est réglée à 1e6. L'intensité maximum est de 500 ms et la gamme d'analyse de 500 à 4500 m/z, les données ont été analysées en mode profil.

MS/MS :

Les ions ont été fragmentés en utilisant la technologie CID, 3 microscans, à une résolution de 70 000, AGC target de 5e5 avec un minimum à 1e-4 et intensité maximale à 250 ms. Une fenêtre d'isolement de 2.0 m/z, une gamme d'analyse de 200 à 2000 m/z. La première masse fixe est de 100.0 m/z, et le seuil d'intensité a été fixé à 1.0e4, toutes les charges multiples, cependant avec une exclusion des isotopes et une exclusion dynamique de 30.0 s avec exclusion des ions monochargés.

III.5 Reproductibilité des injections

Les échantillons analysés proviennent de deux cultures cellulaires liquides en milieu YNB (Yeast Nitrogen base) la seule différence entre les échantillons étant la source de carbone, avec ou sans glucose isotopiquement marqué ^{12}C (condition NC et 12C) ce qui n'influe pas sur la biologie fonctionnelle. Les deux cultures ont été traitées strictement à l'identique (lyse cellulaire et purification des protéines ribosomiques) ainsi, les échantillons ont été démontrés comme reproductibles biologiquement c'est-à-dire qu'ils ont la même quantité et composition en protéines ribosomiques.

La colonne de séparation chromatographique utilisée apporte de la robustesse et de reproductibilité dans les analyses, produisant des profils d'élution similaires pour les différents échantillons préparés (voir Figure 75 et Figure 76). Ceci implique que l'on peut utiliser le même temps de rétention cible pour extraire les mêmes protéines identifiées dans la condition naturelle pour les autres injections des mélanges. Cela permet de répondre à la problématique intrinsèque aux logiciels d'identifications qui ne fonctionnent pas si les ions analysés ont des compositions isotopiques non-naturelles.

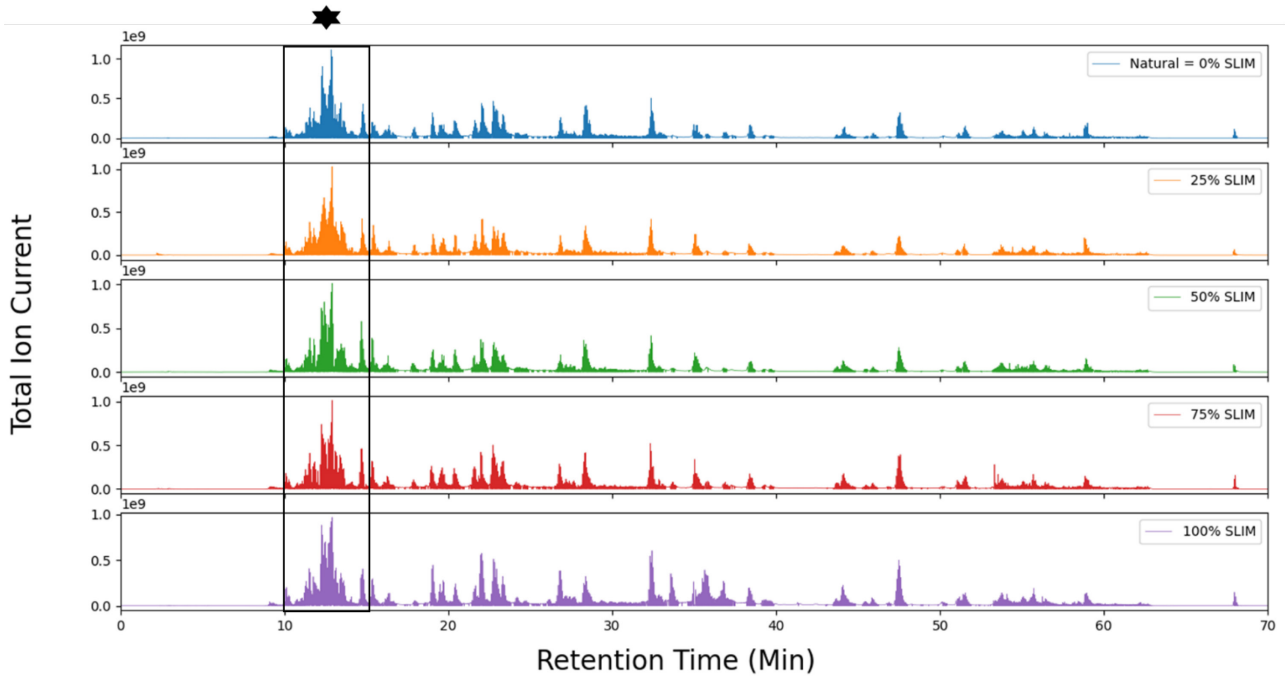


Figure 75: Chromatogramme des 5 mixtures en fonction du temps de rétention

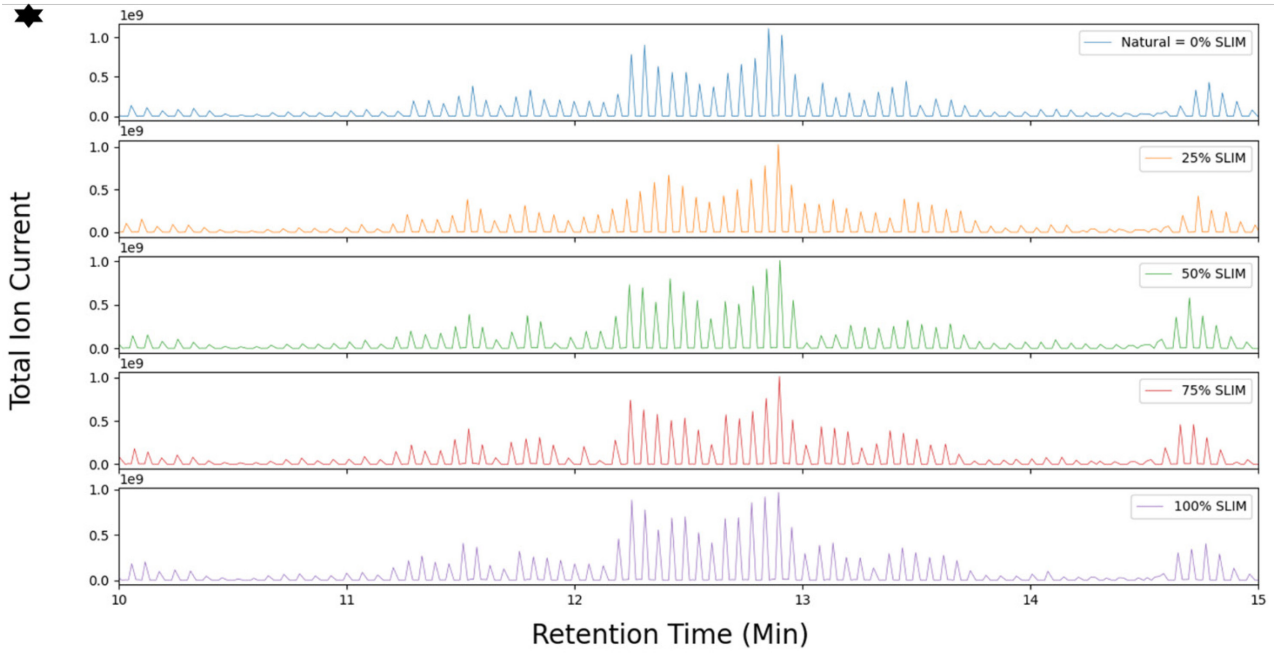


Figure 76: Chromatogramme des 5 mixtures entre les 10 et 15 minutes du temps de rétention

III.6 Tentatives d'identification des protéines ribosomiques en Top-down

Les données spectrales Top-down des protéines ribosomiques purifiées provenant de la condition naturelle (présentées ci-dessus), ont été interrogées avec les deux logiciels de recherche en banque de séquences protéiques *TopPIC* et *Prosight*. Malgré des données expérimentales identiques (Figure 77), les moteurs de recherche identifient des protéines uniques mais également différentes entre les deux logiciels décrits précédemment. Cela s'explique premièrement par le fait que les algorithmes traitent les données différemment, même si tous deux utilisent les données de fragmentations. Deuxièmement, les calculs statistiques permettant l'obtention des scores et la sélection des identifications significativement justes introduites dans ces logiciels ne sont pas comparables. De plus, les identifications résultantes sont dépendantes de la taille de la banque de séquences protéiques. En effet, plus le nombre d'entrées est grand, moins les identifications seront exhaustives. Cela s'explique par la combinatoire des multiples positions possibles des PTMs sur un grand nombre d'acides aminés cibles d'une part, et par les statistiques significatives strictement

dépendantes du nombre d'entrées d'autre part. Une augmentation du nombre de résidus possédant des masses proches rend ainsi la discrimination moins sélective pour l'identification.

Une analyse du même échantillon a été fait en Bottom-up sur l'appareil « timsTOF Pro 2 ». Très sensible, les données spectrales de cet appareil ont révélé la présence de 613 protéines. Outre la grande majorité des protéines ribosomiques, d'autres, associées à la synthèse du ribosome, ont été identifiées. Par ailleurs, d'autre protéines « contaminantes » possédant des physiologies proches ont été purifiées parallèlement au complexe ribosomique. Cette analyse nous a permis de générer une banque de données de séquences protéiques plus restreinte et issue de véritables identifications contenues dans l'échantillon. Cette banque a pu être utilisée pour le logiciel *intact* par exemple. Les résultats issus des interrogations des acquisitions Top-down sur les deux logiciels et la banque du protéome totale de *S. cerevisiae*, permettent d'identifier 89 protéines au total.

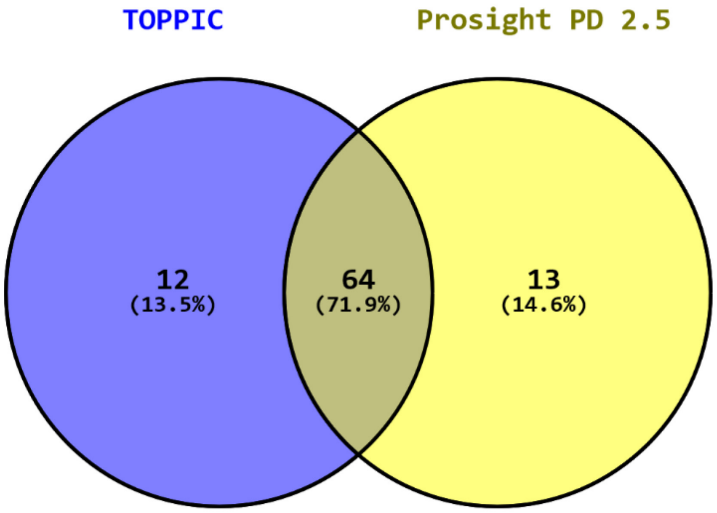


Figure 77: Comparaison du nombre et de la redondance des protéines identifiées selon le moteur de recherche utilisé, la banque est le protéome total de S. cerevisiae.

Cette liste de protéines a été analysée en utilisant le site StringDB, qui permet de visualiser le réseau d'interaction protéine-protéine par une analyse d'enrichissement fonctionnel. Cette représentation en Figure 78 nous permet d'observer la très riche composition des identifications Top-Down dans l'échantillon purifié de protéines ribosomiques.

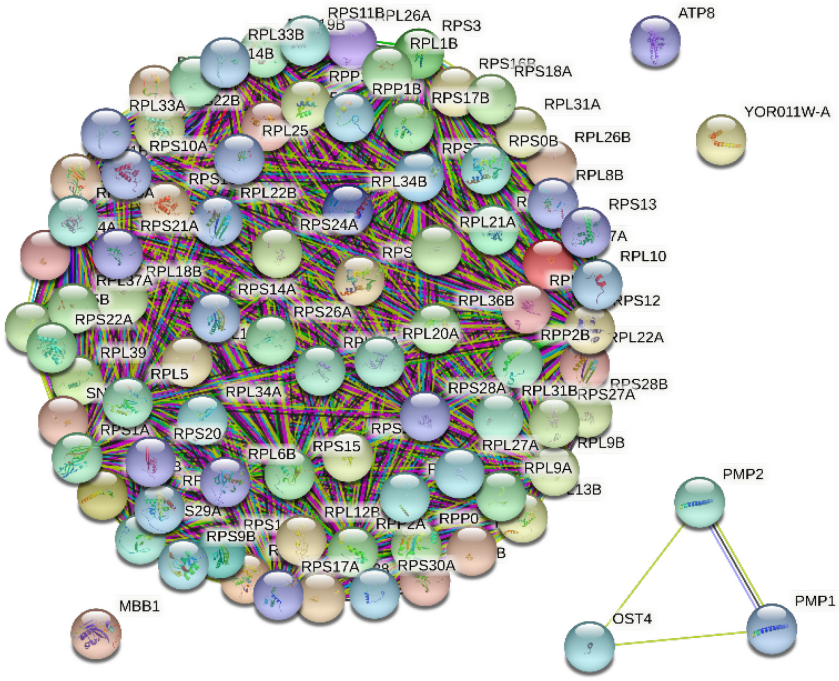


Figure 78: Réseaux d'interaction protéine-protéine par analyse d'enrichissement fonctionnel des protéines identifiées par l'analyse Top-down de l'échantillon ribosome. Cette figure a été générée sur le site StringDB[1]

III.7 Optimisation des protocoles de préfractionnement des échantillons protéique.

Actuellement, nous travaillons à deux grandes avancées permettant l'application de la stratégies Top-down sur des protéomes de hautes complexités.

Le développement du protocole pour la séparation chromatographique d'échantillons protéiques complexes est résolutive et reproductible. Il permet de résoudre une difficulté majeure des analyses par spectrométrie de masse en approche Top-down. Néanmoins, dans le but d'analyser des protéomes bien plus importants, comme un lysat cellulaire total, un préfractionnement biochimique des échantillons est nécessaire. Cela reste un défi pour la communauté scientifique. En particulier, la méthode de préfractionnement comme le Gel-free a été considérée comme le protocole standard de préparation d'échantillons. Cependant cette méthode pose des difficultés

[1] https://string-db.org/

et des limitations dans la reproductibilité. De plus, cette stratégie n'est plus commercialisée de nos jours.

A cet effet, nous avons proposé des stratégies nouvelles permettant l'enrichissement d'échantillons complexes et une réduction en petits jeux de protéines de plus faible poids moléculaire. Ces modes de fractionnement sont reproductibles car fondés sur une séparation purement biochimique, c'est-à-dire utilisant les propriétés physiques et chimiques propres des protéines. Pour cela, j'ai préparé un échantillon de lysat cellulaire total de la levure *Saccharomyces cerevisiae* puis nous avons évalué la pertinence des protocoles de préfractionnement par la mise en application de complexités progressives.

La première méthode repose sur une précipitation de l'échantillon par un ajout de 6 volumes d'Acétonitrile. Ce solvant polaire permet la précipitation des protéines hydrophobes, permettant donc d'enrichir l'échantillon final en protéines de faible poids moléculaire.

La deuxième méthode mise en œuvre, est un fractionnement en utilisant une colonne de type Sep-Pak C18. La phase alkyle permettant la rétention des protéines très hydrophobes et donc leur enrichissement dans l'échantillon.

Enfin, l'échantillon sans préparation initiale du lysat cellulaire total a été analysé en spectrométrie de masse, la séparation chromatographique permettant tout de même une séparation correcte de l'ensemble du protéome sur le temps de chromatographique (Figure 79).

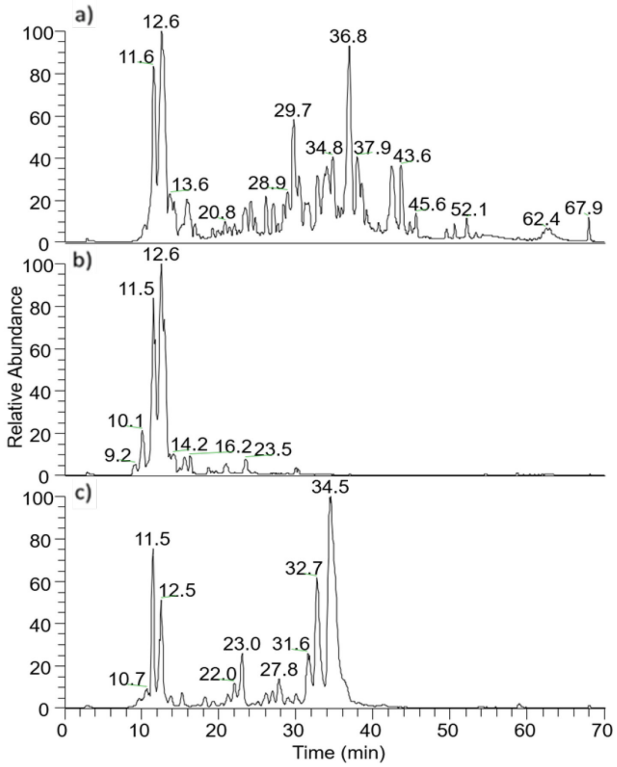


Figure 79: Profil d'analyse LC-MS/MS d'un lysat cellulaire de Saccharomyces cerevisiae. (a) chromatogramme (TIC) du lysat cellulaire total, (b) chromatogramme (TIC) du préfractionnement Sep-Pak C18, (c) chromatogramme (TIC) du préfractionnement à l'acétonitrile.

Nous avons effectué une recherche en banque de données de ces différentes analyses en spectrométrie de masse. Pour cela le logiciel TopPIC a été utilisé en utilisant une banque de données de séquence du protéome total de *S. cerevisiae*. Nous avons identifié 65 protéines pour le lysat total, 44 pour la séparation Sep-Pak, et 117 protéines par la précipitation acétonitrile. Ces résultats d'identifications plus faibles en Sep-Pak s'expliquent par la forte affinité des protéines avec la phase, l'élution bien que réalisée en acétonitrile pure, ne permet pas une récupération optimum de l'échantillon. Néanmoins, le préfractionnement permet d'étudier de manière fine les protéines en interaction plus faible avec la phase alkyle donc de plus faible poids moléculaire. L'observation d'un nombre plus grand d'identifications en préfractionnement à l'acétonitrile s'explique par le fait que ce protocole permet un enrichissement en protéines hydrophiles, par rapport aux autres protéines contenues dans le lysat total. Ainsi, lors de la séparation chromatographique et durant l'analyse

en masse, chaque spectre de fragmentation est de meilleure qualité. En particulier, ces informations spectrales sont critiques pour l'identification par le logiciel TopPIC.

III.8 Tentatives d'identification des protéines ribosomiques dans nos expériences tSLIM en Top-down

Si les données de la condition naturelle présentent une telle divergence dans les identifications, cela est d'autant plus critique en tSLIM. En effet, le tableau suivant (Tableau 10) illustre le biais induit par l'utilisation du modèle d'abondance naturelle des isotopes par l'algorithme TopPIC. Cet algorithme s'avère être dans l'incapacité d'analyser correctement des données de type tSLIM. Une interrogation des différentes acquisitions de la gamme tSLIM de ribosomes a été réalisée à l'aide d'une banque de séquences du protéome totale de *S. cerevisiae*.

Échantillon	NC	25% SLIM	50% SLIM	75% SLIM	100% SLIM
Nombre d'identifications	74	54	31	9	7

Tableau 10 : La diminution du nombre d'indentifications est fonction de la quantité de conditions 12C dans les mélanges. (Donnée d'identification issue de la gamme tSLIM des ribosomes et une banque de séquence du protéome complet de S. cerevisiae)

Partie 4 : Conclusions et Perspectives de la thèse

La méthode SLIM initialement développée au laboratoire proposait une voie originale de quantification du protéome par spectrométrie de masse en approche Bottom-up. Cette méthode était fondée sur le marquage métabolique *in vivo* à partir d'une source de carbone uniformément marquée au ^{12}C. Cela permettait une réduction de la complexité des isotopes des protéines, ayant pour conséquence une augmentation de l'intensité de l'ion monoisotopique et donc une meilleure précision dans la mesure de la masse des peptides. Néanmoins, plusieurs limitations ont été soulignées, elles serviront de point de départ pour ma thèse.

Celle-ci avait trois objectifs : Le développement d'un workflow robuste d'analyse de données bSLIM, l'extension de la méthode sur des organismes supérieurs (auxotrophes), et le développement de la méthode tSLIM pour l'analyse des protéines intactes.

Le premier objectif a permis de rendre la méthode SLIM opérationnelle dans sa version bSLIM par l'introduction de nouvelles méthodes de calcul et l'obtention de valeurs quantitatives. En particulier, nous avons procédé à des simulations numériques dans l'objectif d'observer et de définir des formules algébriques robustes permettant le suivi de l'incorporation de ^{12}C dans les peptides en fonction du rapport des intensités expérimentales des isotopologues M_1 et M_0. De même, la quantification des clusters expérimentaux est rendue possible par un calcul comportant des données expérimentales et théoriques. Les valeurs théoriques sont issues d'une modélisation fine des intensités des isotopologues dans les clusters isotopiques. Pour cela nous avons utilisé l'identification des peptides permettant la résolution des expressions algébriques polynomiales à partir des formules chimiques. L'extraction des valeurs

expérimentales d'intensité des isotopologues dans les clusters isotopiques est une étape critique. La détermination de la fraction molaire des peptides non-marqués est une quantification relative qui permet de décrire les variations du protéome de manière fiable. Si les valeurs quantitatives obtenues par le bSLIM s'avèrent être robustes, une méthode statistique a été néanmoins développée pour évaluer statistiquement le taux de fausse découverte dans les résultats obtenus.

Les méthodes algébriques bSLIM ont été implémentées dans un workflow de retraitement des données. Le pipeline d'analyse est composé en particulier d'un traitement des spectres de masse, d'un module d'extraction des signaux expérimentaux et de modélisation théorique des clusters isotopiques. Il répond à l'objectif final de fournir des valeurs quantitatives au niveau des peptides et des protéines. Les statistiques descriptives ont également été implémentées à l'aide d'outils bio-informatiques. Ces workflows sont rapides, simples d'utilisation et aisément paramétrables pour une adaptation à chaque nouvelle expérience.

A l'issue du développement de la méthode bSLIM, nous avons procédé d'une part à une validation expérimentale des méthodes de calcul par un suivi de valeurs biologiques attendues. D'autre part, nous avons évalué la résolution et la sensibilité de notre méthode en étudiant les variations biologiques de deux protéomes issus de souches de levure proches. Les résultats ont confirmé une sensibilité qui nous a permis d'observer des variations fines du protéome.

De façon similaire nous avons développé les méthodes de calculs pour déterminer les valeurs quantitatives issues de marquages incomplets par des organismes présentant des auxotrophies. Cela s'est traduit par une adaptation des calculs algébriques permettant la prise en compte de la présence d'acides aminés exogènes non-marqués dans les peptides. Le workflow précédemment développé a été adapté afin de répondre en conséquence. Cela nous a permis d'aborder le deuxième objectif qui vise à étendre la méthode bSLIM à des organismes eucaryotes supérieurs, d'une part dans une lignée cellulaire humaine "HEK", nécessitant un apport d'acides aminés essentiels exogènes et d'autre part dans un organisme multicellulaire, le nématode *C. elegans,* nécessitant un marquage indirect. La réussite du marquage a été observée par le suivi de l'incorporation de ^{12}C dans les peptides issus de ces deux organismes. La réponse à cet objectif ouvre la porte à des applications de quantification purement biologique dans des questionnements biomédicaux.

L'approche Top-down permet de procéder à des études plus informatives sur la structure des protéines, en particulier l'étude des modifications post-traductionnelles des protéines. Cependant, cette approche est plus complexe pour de nombreuses raisons, notamment la masse élevée des objets et la physiologie propre des protéines.

Le développement au préalable du bSLIM en Bottom-up nous a permis d'aborder les difficultés de l'approche Top-down et d'étudier les caractéristiques pour le développement des analyses en protéine intacte. Ainsi, différentes méthodes fondées sur des approches biochimiques et technologiques ont été développées en parallèle de la réalisation du bSLIM.

Les approches biochimiques ont consisté à la préparation d'échantillons reproductibles et de faible complexité comme les protéines ribosomiques. Les approches technologiques ont conduit au développement de méthodes de séparation d'échantillons de complexité graduelle, par chromatographie en phase liquide et d'acquisition optimale en spectrométrie de masse.

Le dernier objectif a permis de conceptualiser et d'appliquer la méthode tSLIM à l'analyse de protéines intactes. Dans cette approche, les masses des objets étudiés étant plus élevées qu'en Bottom-up, les spectres de masse expérimentaux observés présentent une intensité du monoisotopique très peu intense. Un calcul fondé sur l'intensité des isotopologues de rang faible n'est pas possible. La quantification des clusters isotopiques reste cependant possible par la prise en compte de la totalité de l'intensité des isotopologues. Pour cela la modélisation des clusters isotopiques théoriques indépendamment de leurs identifications associées est un objectif essentiel. La simulation numérique nous a permis de confirmer que l'approximation des clusters isotopiques par une distribution de Poisson est juste. Des calculs algébriques nous ont permis de déterminer les paramètres répondant à cette distribution tant en condition naturelle qu'en condition marquée.

Comme cela était attendu, les identifications des protéoformes sur les données issues d'acquisition en Top-down se sont révélées limitées. En effet, les différents algorithmes nécessitent une mesure de masse précise ainsi que des données spectrales issues de fragmentation de protéines. De plus les logiciels du domaine s'avèrent incapables de traiter des données dans le cadre d'une expérience tSLIM. Ces différentes observations nous ont montré la pertinence de développer des outils propres, aptes à quantifier des données tSLIM. Ainsi d'une méthode originale utilisant ces résultats de

quantification pour procéder à une identification. La méthode innovante utilise une recherche de type *de novo* et tire parti de l'utilisation des informations contenues dans les spectres comme la fragmentation des ions et les multiples états de charge observés, pour y apporter une puissance statistique.

Les activités liées à ces trois objectifs ont donné lieu à trois publications disponibles en annexes. Les travaux développés pendant cette thèse permettent d'aborder de nouveaux défis méthodologiques et biologiques.

La méthode bSLIM peut désormais être utilisée dans le but de répondre à des questionnements biologiques pertinents.

La méthode tSLIM a été conceptualisée et une preuve de faisabilité a été réalisée avec des résultats prometteurs. Ses applications biologiques futures permettront la caractérisation exhaustive des protéoformes. Actuellement, nous sommes en train de travailler sur un exemple concret : l'étude de la régulation épigénétique de la filamentation chez *C. albicans*.

À l'avenir, il est très envisageable qu'une combinaison des deux approches bSLIM et tSLIM permettra une étude approfondie de protéomes complexes en alliant la puissance de l'identification du bSLIM à la caractérisation fine des protéoformes par le tSLIM.

L'arrivée de nouveaux équipements, encore plus performants et disposant de technologies nouvelles de séparation, devrait renforcer les bénéfices de la méthode de quantification SLIM. Il s'agit des appareils "tims-TOF Pro 2" de Bruker pour une approche Bottom-up et "Select Series Cyclic IMS" de Waters pour une approche Top-down.

REFERENCES

Achcar, F., Camadro, J. michel, & Mestivier, D. (2009). AutoClass@IJM: A powerful tool for Bayesian classification of heterogeneous data in biology. *Nucleic Acids Research*, *37*(SUPPL. 2), W63. https://doi.org/10.1093/nar/gkp430

Almén, M. S., Nordström, K. J. V., Fredriksson, R., & Schiöth, H. B. (2009). Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology*, *7*, 50. https://doi.org/10.1186/1741-7007-7-50

Alves, G., Ogurtsov, A. Y., & Yu, Y. K. (2010). RAId_aPS: MS/MS analysis with multiple scoring functions and spectrum-specific statistics. *PLoS ONE*, *5*(11). https://doi.org/10.1371/journal.pone.0015438

Alves, G., Ogurtsov, A. Y., & Yu, Y. K. (2014). Molecular Isotopic Distribution Analysis (MIDAs) with adjustable mass accuracy. *Journal of the American Society for Mass Spectrometry*, *25*(1), 57–70. https://doi.org/10.1007/s13361-013-0733-7

Alves, G., & Yu, Y. K. (2005). Robust accurate identification of peptides (RAId): Deciphering MS2 data using a structured library search with de novo based statistics. *Bioinformatics*, *21*(19), 3726–3732. https://doi.org/10.1093/bioinformatics/bti620

Andreeva, A., Kulesha, E., Gough, J., & Murzin, A. G. (2020). The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, *48*(D1), D376–D382. https://doi.org/10.1093/nar/gkz1064

Anfinsen, C. B. (1972). Studies on the principles that govern the folding of protein chains. *Les Prix Nobel*, *Figure 1*, 103–119.

Audi, G., & Wapstra, A. H. (1993). The 1993 atomic mass evaluation. (I) Atomic mass table. *Nuclear Physics, Section A*, *565*(1), 1–65. https://doi.org/10.1016/0375-9474(93)90024-R

Audi, G., & Wapstra, A. H. (1995). The 1995 update to the atomic mass evaluation.

Nuclear Physics, Section A, *595*(4), 409–480. https://doi.org/10.1016/0375-9474(95)00445-9

Azzaz, F., & Fantini, J. (2022). The epigenetic dimension of protein structure. *Biomolecular Concepts*, *13*(1), 55–60. https://doi.org/10.1515/bmc-2022-0006

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Bantscheff, M., Lemeer, S., Savitski, M. M., & Kuster, B. (2012). Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*, *404*(4), 939–965. https://doi.org/10.1007/s00216-012-6203-4

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, *389*(4), 1017–1031. https://doi.org/10.1007/s00216-007-1486-6

Ben-Bassat, A., Bauer, K., Chang, S. Y., Myambo, K., & Boosman, A. (1987). Processing of the initiation methionine from proteins: Properties of the Escherichia coli methionine aminopeptidase and its gene structure. *Journal of Bacteriology*, *169*(2), 751–757. https://doi.org/10.1128/jb.169.2.751-757.1987

Benjamin, D. R., Robinson, C. V., Hendrick, J. P., Hartl, F. U., & Dobson, C. M. (1998). Mass spectrometry of ribosomes and ribosomal subunits. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(13), 7391–7395. https://doi.org/10.1073/pnas.95.13.7391

Berglund, M., & Wieser, M. E. (2011). Isotopic compositions of the elements 2009 (IUPAC Technical Report)*. *Pure Appl. Chem*, *83*(2), 397–410. https://doi.org/10.1351/PAC-REP-10-06-02

Biemann, K. (1990). Nomenclature for peptide fragment ions (positive ions). *Methods in Enzymology*, *193*(C), 886–887. https://doi.org/10.1016/0076-6879(90)93460-3

Bobály, B., Lauber, M., Beck, A., Guillarme, D., & Fekete, S. (2018). Utility of a high

coverage phenyl-bonding and wide-pore superficially porous particle for the analysis of monoclonal antibodies and related products. *Journal of Chromatography A*, *1549*, 63–76. https://doi.org/10.1016/j.chroma.2018.03.043

Borzou, A., Sadygov, V. R., Zhang, W., & Sadygov, R. G. (2019). Proteome dynamics from heavy water metabolic labeling and peptide tandem mass spectrometry. *International Journal of Mass Spectrometry*, *445*, 116194. https://doi.org/10.1016/j.ijms.2019.116194

Breen, E. J., Hopwood, F. G., Williams, K. L., & Wilkins, M. R. (2000). Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis*, *21*(11), 2243–2251. https://doi.org/10.1002/1522-2683(20000601)21:11<2243::AID-ELPS2243>3.0.CO;2-K

Carpentier, M., Chomilier, J., & Valencia, A. (2019). Protein multiple alignments: Sequence-based versus structure-based programs. *Bioinformatics*, *35*(20), 3970–3980. https://doi.org/10.1093/bioinformatics/btz236

Casey, J. R., Grinstein, S., & Orlowski, J. (2010). Sensors and regulators of intracellular pH. *Nature Reviews Molecular Cell Biology*, *11*(1), 50–61. https://doi.org/10.1038/nrm2820

Catherman, A. D., Skinner, O. S., & Kelleher, N. L. (2014). Top Down proteomics: Facts and perspectives. *Biochemical and Biophysical Research Communications*, *445*(4), 683–693. https://doi.org/10.1016/j.bbrc.2014.02.041

Chen, A. Y., Lee, J., Damjanovic, A., & Brooks, B. R. (2022). Protein p K a Prediction by Tree-Based Machine Learning. *Journal of Chemical Theory and Computation*, acs.jctc.1c01257. https://doi.org/10.1021/acs.jctc.1c01257

Chen, X., Sun, Y., Zhang, T., Shu, L., Roepstorff, P., & Yang, F. (2021). Quantitative Proteomics Using Isobaric Labeling: A Practical Guide. *Genomics, Proteomics & Bioinformatics*, *19*(5), 689–706. https://doi.org/10.1016/j.gpb.2021.08.012

Chen, X., Wei, S., Ji, Y., Guo, X., & Yang, F. (2015). Quantitative proteomics using SILAC: Principles, applications, and developments. *Proteomics*, *15*(18), 3175–3192. https://doi.org/10.1002/pmic.201500108

Chevreux, G., Tilly, N., & Bihoreau, N. (2011). Fast analysis of recombinant monoclonal antibodies using IdeS proteolytic digestion and electrospray mass spectrometry. *Analytical Biochemistry*, *415*(2), 212–214.

https://doi.org/10.1016/j.ab.2011.04.030

Choe, L., D'Ascenzo, M., Relkin, N. R., Pappin, D., Ross, P., Williamson, B., Guertin, S., Pribil, P., & Lee, K. H. (2007). 8-Plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics*, *7*(20), 3651–3660. https://doi.org/10.1002/pmic.200700316

Cody, C. W., Prasher, D. C., Westler, W. M., Prendergast, F. G., & Ward, W. W. (1993). Chemical Structure of the Hexapeptide Chromophore of the Aequorea Green-Fluorescent Protein. *Biochemistry*, *32*(5), 1212–1218. https://doi.org/10.1021/bi00056a003

Compton, P. D., Zamdborg, L., Thomas, P. M., & Kelleher, N. L. (2011). On the scalability and requirements of whole protein mass spectrometry. *Analytical Chemistry*, *83*(17), 6868–6874. https://doi.org/10.1021/ac2010795

Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F., Burkhard, P. R., & Sanchez, J. C. (2008). Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Analytical Chemistry*, *80*(8), 2921–2931. https://doi.org/10.1021/ac702422x

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The Protein Folding Problem. *Annual Review of Biophysics*, *37*(1), 289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558

Dimitrov, L. N., Brem, R. B., Kruglyak, L., & Gottschling, D. E. (2009). Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. *Genetics*, *183*(1), 365–383. https://doi.org/10.1534/genetics.109.104497

Donnelly, D. P., Rawlins, C. M., DeHart, C. J., Fornelli, L., Schachner, L. F., Lin, Z., Lippens, J. L., Aluri, K. C., Sarin, R., Chen, B., Lantz, C., Jung, W., Johnson, K. R., Koller, A., Wolff, J. J., Campuzano, I. D. G., Auclair, J. R., Ivanov, A. R., Whitelegge, J. P., … Agar, J. N. (2019). Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nature Methods*, *16*(7), 587–594. https://doi.org/10.1038/s41592-019-0457-0

Duncan, D. T., Craig, R., & Link, A. J. (2005). Parallel Tandem: A Program for Parallel Processing of Tandem Mass Spectra Using PVM or MPI and X!Tandem. *Journal*

of Proteome Research, *4*(5), 1842–1847. https://doi.org/10.1021/pr050058i

Eidhammer, I., Barsnes, H., Eide, G. E., & Martens, L. (2013). Computational and Statistical Methods for Protein Quantification by Mass Spectrometry. In *Computational and Statistical Methods for Protein Quantification by Mass Spectrometry*. https://doi.org/10.1002/9781118494042

Ellis, S. R., Iv, J. B. M., Farrar, J. E., Arceci, R. J., & Liu, J. M. (2010). Distinct ribosome maturation defects in yeast models of Diamond-Blackfan anemia and Shwachman-Diamond syndrome. *Haematologica*, *95*, 57–64. https://doi.org/10.3324/haematol.2009.012450

Erve, J. C. L., Gu, M., Wang, Y., DeMaio, W., & Talaat, R. E. (2009). Spectral Accuracy of Molecular Ions in an LTQ/Orbitrap Mass Spectrometer and Implications for Elemental Composition Determination. *Journal of the American Society for Mass Spectrometry*, *20*(11), 2058–2069. https://doi.org/10.1016/j.jasms.2009.07.014

Esvelt, K. M., Carlson, J. C., & Liu, D. R. (2011). A system for the continuous directed evolution of biomolecules. *Nature*, *472*(7344), 499–503. https://doi.org/10.1038/nature09929

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Andrew Senior, T. G., Augustin Žídek, R., Bates, S. B., Jason Yim, O., Ronneberger, S., Bodenstein, M., Zielinski, A. B., Anna Potapenko, A., & Cowie, K, D. H. (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv*. https://doi.org/10.1101/2021.10.04.463034

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). KNIME for reproducible cross-domain analysis of life science data. *Journal of Biotechnology*, *261*, 149–156. https://doi.org/10.1016/j.jbiotec.2017.07.028

Gouw, J. W., Krijgsveld, J., & Heck, A. J. R. (2010). Quantitative proteomics by metabolic labeling of model organisms. *Molecular and Cellular Proteomics*, *9*(1), 11–24. https://doi.org/10.1074/mcp.R900001-MCP200

Graves, P. R., & Haystead, T. A. J. (2002). Molecular Biologist's Guide to Proteomics. *Microbiology and Molecular Biology Reviews*, *66*(1), 39–63. https://doi.org/10.1128/mmbr.66.1.39-63.2002

Greer, J. B., Early, B. P., Durbin, K. R., Patrie, S. M., Thomas, P. M., Kelleher, N. L., LeDuc, R. D., & Fellers, R. T. (2022). ProSight Annotator: Complete control and customization of protein entries in UniProt XML files. *PROTEOMICS*, 2100209.

https://doi.org/10.1002/pmic.202100209

Guillarme, D., & Fekete, S. (2013). Advantages and Applications of Revolutionary Superficially Porous Particle Columns in Liquid Chromatography. *Analytical Pharmaceutical Chemistry*.

Hall, D. A., Ptacek, J., & Snyder, M. (2007). Protein microarray technology. *Mechanisms of Ageing and Development*, *128*(1), 161–167. https://doi.org/10.1016/j.mad.2006.11.021

Haver, T. O. (2022). A Pragmatic Introduction to Signal Processing. In *University of Maryland at College Park* (Issue April). http://terpconnect.umd.edu/~toh/spectrum/TOC.html

Ho, B., Baryshnikova, A., & Brown, G. W. (2018). Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Systems*, *6*(2), 192-205.e3. https://doi.org/10.1016/j.cels.2017.12.004

Hogan, M., Higdon, R., Kolker, N., Kolker, E., & Al, H. E. T. (2005). Spectrometry Proteomics. In *OMICS: A Journal of Integrative Biology* (Vol. 9, Issue 3). http://www.amazon.com/Computational-Methods-Mass-Spectrometry-Proteomics/dp/0470512970%3FSubscriptionId%3D1V7VTJ4HA4MFT9XBJ1R2%26tag%3Dmekentosjcom-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0470512970

Holman, J. D., Tabb, D. L., & Mallick, P. (2014). Employing ProteoWizard to convert raw mass spectrometry data. *Current Protocols in Bioinformatics*, *46*(SUPPL.46), 13. https://doi.org/10.1002/0471250953.bi1324s46

Horn, D. M., Ueckert, T., Fritzemeier, K., Tham, K., Paschke, C., Berg, F., Pfaff, H., Jiang, X., Li, S., & Lopez-Ferrer, D. (2016). New Method for Label-Free Quantification in the Proteome Discoverer Framework. *Available Online: Https://Planetorbitrap.Com/*, 2–4. https://tools.thermofisher.com/content/sfs/posters/PN-64792-Label-Free-Proteome-Discoverer-ASMS2016-PN64792-EN.pdf

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., & Cooks, R. G. (2005). The Orbitrap: A new mass spectrometer. *Journal of Mass Spectrometry*, *40*(4), 430–443. https://doi.org/10.1002/jms.856

Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. *Proteins: Structure, Function and Bioinformatics*, *77*(3), 499–508. https://doi.org/10.1002/prot.22458

Jeong, K., Babović, M., Gorshkov, V., Kim, J., Jensen, O. N., & Kohlbacher, O. (2021). FLASHIda: Intelligent data acquisition for top-down proteomics that doubles proteoform level identification count. *BioRxiv*, 2021.11.11.468203. https://doi.org/10.1101/2021.11.11.468203

Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X., & Tian, Y. (2021). A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports 2021 11:1*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-81279-4

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. ✉. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583. https://doi.org/10.1038/s41586-021-03819-2

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*(11), 923–925. https://doi.org/10.1038/nmeth1113

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics*, *24*(21), 2534–2536. https://doi.org/10.1093/bioinformatics/btn323

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., Baker, D., & Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(47), 18949–18953. https://doi.org/10.1073/PNAS.1115898108/-/DCSUPPLEMENTAL/PNAS.1115898108_SI.PDF

Kingdon, K. H. (1923). A method for the neutralization of electron space charge by positive ionization at very low gas pressures. *Physical Review*, *21*(4), 408–418. https://doi.org/10.1103/PhysRev.21.408

Knight, R. D. (1981). Storage of ions from laser-produced plasmas. *Applied Physics*

Letters, *38*(4), 221–223. https://doi.org/10.1063/1.92315

Kondori, N., Kurtovic, A., Piñeiro-Iglesias, B., Salvà-Serra, F., Jaén-Luchoro, D., Andersson, B., Alves, G., Ogurtsov, A., Thorsell, A., Fuchs, J., Tunovic, T., Kamenska, N., Karlsson, A., Yu, Y. K., Moore, E. R. B., & Karlsson, R. (2021). Mass Spectrometry Proteotyping-Based Detection and Identification of Staphylococcus aureus, Escherichia coli, and Candida albicans in Blood. *Frontiers in Cellular and Infection Microbiology*, *11*(July), 1–15. https://doi.org/10.3389/fcimb.2021.634215

Kou, Q., Xun, L., & Liu, X. (2016). TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics*, *32*(22), 3495–3497. https://doi.org/10.1093/BIOINFORMATICS/BTW398

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function and Bioinformatics*, *87*(12), 1011–1020. https://doi.org/10.1002/prot.25823

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins: Structure, Function and Bioinformatics*, *89*(12), 1607–1617. https://doi.org/10.1002/prot.26237

Lardeux, H., Duivelshof, B. L., Colas, O., Beck, A., McCalley, D. V., Guillarme, D., & D'Atri, V. (2021). Alternative mobile phase additives for the characterization of protein biopharmaceuticals in liquid chromatography – Mass spectrometry. *Analytica Chimica Acta*, *1156*, 338347. https://doi.org/10.1016/j.aca.2021.338347

Le Faouder, J., Laouirem, S., Alexandrov, T., Ben-Harzallah, S., Léger, T., Albuquerque, M., Bedossa, P., & Paradis, V. (2014). Tumoral heterogeneity of hepatic cholangiocarcinomas revealed by MALDI imaging mass spectrometry. *Proteomics*, *14*(7–8), 965–972. https://doi.org/10.1002/pmic.201300463

Le, Z. (2004). Maximum Entropy Modeling Toolkit for Python and C ++. *Natural Language Processing Lab, Northeastern University, China.*, *December*, 8–11.

Léger, T., Garcia, C., Collomb, L., & Camadro, J. M. (2017). A simple light isotope

metabolic labeling (SLIM-labeling) strategy: A powerful tool to address the dynamics of proteome variations in vivo. *Molecular and Cellular Proteomics*, *16*(11), 2017–2031. https://doi.org/10.1074/mcp.M117.066936

Léger, T., Garcia, C., Videlier, M., & Camadro, J. M. (2016). Label-free quantitative proteomics in yeast. In Frédéric Devaux (Ed.), *Methods in Molecular Biology* (Vol. 1361, pp. 289–307). Springer New York. https://doi.org/10.1007/978-1-4939-3079-1_16

Lelandais, G., Denecker, T., Garcia, C., Danila, N., Léger, T., & Camadro, J. M. (2019). Label-free quantitative proteomics in Candida yeast species: Technical and biological replicates to assess data reproducibility. *BMC Research Notes*, *12*(1), 10–12. https://doi.org/10.1186/s13104-019-4505-8

Lermyte, F., Tsybin, Y. O., O'Connor, P. B., & Loo, J. A. (2019). Top or Middle? Up or Down? Toward a Standard Lexicon for Protein Top-Down and Allied Mass Spectrometry Approaches. *Journal of the American Society for Mass Spectrometry*, *30*(7), 1149–1157. https://doi.org/10.1007/s13361-019-02201-x

Liebermeister, W., Noor, E., Flamholz, A., Davidi, D., Bernhardt, J., & Milo, R. (2014). Visual account of protein investment in cellular functions. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8488–8493. https://doi.org/10.1073/pnas.1314810111

Lin, X., Li, X., & Lin, X. (2020). A review on applications of computational methods in drug screening and design. *Molecules*, *25*(6), 1–17. https://doi.org/10.3390/molecules25061375

Liu, T., Belov, M. E., Jaitly, N., Qian, W. J., & Smith, R. D. (2007). Accurate mass measurements in proteomics. *Chemical Reviews*, *107*(8), 3621–3653. https://doi.org/10.1021/cr068288j

Liu, X., Inbar, Y., Dorrestein, P. C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J. P., Bafna, V., & Pevzner, P. A. (2010). Deconvolution and database search of complex tandem mass spectra of intact proteins: A combinatorial approach. *Molecular and Cellular Proteomics*, *9*(12), 2772–2782. https://doi.org/10.1074/MCP.M110.002766/ATTACHMENT/EB60E78C-5929-4321-BDE8-5A8C8AD46ED0/MMC1.PDF

Ljungdahl, P. O., & Daignan-Fornier, B. (2012). Regulation of amino acid, nucleotide,

and phosphate metabolism in Saccharomyces cerevisiae. *Genetics*, *190*(3), 885–929. https://doi.org/10.1534/genetics.111.133306

Maguire, B. A., & Zimmermann, R. A. (2001). The Ribosome in Focus. *Cell*, *104*(6), 813–816. https://doi.org/10.1016/S0092-8674(01)00278-1

Makarov, A. (2000). Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry*, *72*(6), 1156–1162.

Makarov, A. A. (1999). *Mass spectrometer* (Patent No. 5 886 346).

Makarov, A., Grinfeld, D., & Ayzikov, K. (2019). Fundamentals of Orbitrap analyzer. In *Fundamentals and Applications of Fourier Transform Mass Spectrometry*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-814013-0.00002-8

Makarov, A., Hardman, M. E., Schwartz, J. C., & Senko, M. W. (2005). *Mass spectrometry method and apparatus* (Patent No. 6 872 938 B2).

Mamyrin, B. A., Karataev, V. I., Shmikk, D. V., & Zagulin, V. A. (1973). The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Sov Phys JETP*, *37*(1), 45–48. http://www.google.com/search?client=safari&rls=en-us&q=The+mass-reflectron,+a+new+nonmagnetic+time-of-flight+mass+spectrometer+with+high+resolution&ie=UTF-8&oe=UTF-8

Maráková, K., Rai, A. J., & Schug, K. A. (2020). Effect of difluoroacetic acid and biological matrices on the development of a liquid chromatography–triple quadrupole mass spectrometry method for determination of intact growth factor proteins. *Journal of Separation Science*, *43*(9–10), 1663–1677. https://doi.org/10.1002/jssc.201901254

Marshall, A. G., Senko, M. W., Li, W., Li, M., Dillon, S., Guan, S., & Logan, T. M. (1997). Protein molecular mass to 1 Da by 13C, 15N double-depletion and FT-ICR mass spectrometry. *Journal of the American Chemical Society*, *119*(2), 433–434. https://doi.org/10.1021/ja9630046

Matthiesen, R. (2020). Methods in Molecular Biollogy. In R. Matthiesen (Ed.), *Mass Spectrometry Data Analysis in Proteomics* (3rd ed., Vol. 2051). Springer New York. https://doi.org/10.1007/978-1-4939-9744-2

McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C.,

Kuhn, K., Pike, I., Grothe, R. A., Blethrow, J. D., & Gygi, S. P. (2012). Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Analytical Chemistry*, *84*(17), 7469–7478. https://doi.org/10.1021/ac301572t

McInnes, C. (2007). Virtual screening strategies in drug discovery. *Current Opinion in Chemical Biology*, *11*(5), 494–502. https://doi.org/10.1016/j.cbpa.2007.08.033

Melby, J. A., Roberts, D. S., Larson, E. J., Brown, K. A., Bayne, E. F., Jin, S., & Ge, Y. (2021). Novel Strategies to Address the Challenges in Top-Down Proteomics. *Journal of the American Society for Mass Spectrometry*, *32*(6), 1278–1294. https://doi.org/10.1021/jasms.1c00099

Miyagi, M., & Rao, K. C. S. (2007). Proteolytic 18O-labeling strategies for quantitative proteomics. *Mass Spectrometry Reviews*, *26*(1), 121–136. https://doi.org/10.1002/mas.20116

Mosimann, S., Meleshko, R., & James, M. N. G. (1995). A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins: Structure, Function, and Genetics*, *23*(3), 301–317. https://doi.org/10.1002/prot.340230305

Nagornov, K. O., Kozhinov, A. N., Gasilova, N., Menin, L., & Tsybin, Y. O. (2020). Transient-Mediated Simulations of FTMS Isotopic Distributions and Mass Spectra to Guide Experiment Design and Data Analysis. *Journal of the American Society for Mass Spectrometry*, *31*(9), 1927–1942. https://doi.org/10.1021/jasms.0c00190

Nakao, A., Yoshihama, M., & Kenmochi, N. (2004). RPG: The Ribosomal Protein Gene database. *Nucleic Acids Research*, *32*(DATABASE ISS.), 168–170. https://doi.org/10.1093/nar/gkh004

Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., Mariani, M., Assadourian, G., Lee, A., Van Sluyter, S. C., & Haynes, P. A. (2011). Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics*, *11*(4), 535–553. https://doi.org/10.1002/pmic.201000553

Nguyen, J. M., Smith, J., Rzewuski, S., Legido-Quigley, C., & Lauber, M. A. (2019). High sensitivity LC-MS profiling of antibody-drug conjugates with difluoroacetic acid ion pairing. *MAbs*, *11*(8), 1358–1366.

https://doi.org/10.1080/19420862.2019.1658492

Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., & Ahn, N. G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Molecular and Cellular Proteomics*, *4*(10), 1487–1502. https://doi.org/10.1074/mcp.M500084-MCP200

Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics : MCP*, *1*(5), 376–386. https://doi.org/10.1074/mcp.M200025-MCP200

Ovchinnikov, S., Park, H., Kim, D. E., DiMaio, F., & Baker, D. (2018). Protein structure prediction using Rosetta in CASP12. *Proteins: Structure, Function and Bioinformatics*, *86*(September 2017), 113–121. https://doi.org/10.1002/prot.25390

Panasenko, O. (2012). Ribosome Fractionation in Yeast. *BIO-PROTOCOL*, *2*(16), 1–6. https://doi.org/10.21769/BioProtoc.251

Pearce, R., & Zhang, Y. (2021a). Deep learning techniques have significantly impacted protein structure prediction and protein design. *Current Opinion in Structural Biology*, *68*, 194–207. https://doi.org/10.1016/j.sbi.2021.01.007

Pearce, R., & Zhang, Y. (2021b). Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry*, *297*(1), 100870. https://doi.org/10.1016/j.jbc.2021.100870

Pfeuffer, J., Sachsenberg, T., Dijkstra, T. M. H., Serang, O., Reinert, K., & Kohlbacher, O. (2019). EPIFANY – A method for efficient high-confidence protein inference. *BioRxiv*, *19*, 1060. https://doi.org/10.1101/734327

Planta, R. J., & Mager, W. H. (1998). The list of cytoplasmic ribosomal proteins ofSaccharomyces cerevisiae. *Yeast*, *14*(5), 471–477. https://doi.org/10.1002/(SICI)1097-0061(19980330)14:5<471::AID-YEA241>3.0.CO;2-U

Protein Metrics Inc. (2021). *PMI Intact Analysis (Intact Mass TM) User ' s Manual* (Issue March).

Rak, M., Zeng, X., Brière, J. J., & Tzagoloff, A. (2009). Assembly of Fo in

Saccharomyces cerevisiae. *Biochimica et Biophysica Acta - Molecular Cell Research*, *1793*(1), 108–116. https://doi.org/10.1016/j.bbamcr.2008.07.001

Ramakrishnan, V., & Moore, P. B. (2001). Atomic structures at last: the ribosome in 2000. *Current Opinion in Structural Biology*, *11*(2), 144–154. https://doi.org/10.1016/S0959-440X(00)00184-6

Reid, G. E., & McLuckey, S. A. (2002). 'Top down' protein characterization via tandem mass spectrometry. *Journal of Mass Spectrometry*, *37*(7), 663–675. https://doi.org/10.1002/jms.346

Rockwood, A. L., & Palmblad, M. (2020). *Isotopic Distributions* (pp. 79–114). https://doi.org/10.1007/978-1-4939-9744-2_3

Rodgers, R. P., Blumer, E. N., Hendrickson, C. L., & Marshall, A. G. (2000). Stable isotope incorporation triples the upper mass limit for determination of elemental composition by accurate mass measurement. *Journal of the American Society for Mass Spectrometry*, *11*(10), 835–840. https://doi.org/10.1016/S1044-0305(00)00158-6

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., & Pappin, D. J. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular and Cellular Proteomics*, *3*(12), 1154–1169. https://doi.org/10.1074/mcp.M400129-MCP200

Rožman, M., & Gaskell, S. J. (2012). Charge state dependent top-down characterisation using electron transfer dissociation. *Rapid Communications in Mass Spectrometry*, *26*(3), 282–286. https://doi.org/10.1002/rcm.5330

Sadygov, R. G. (2018). Poisson Model to Generate Isotope Distribution for Biomolecules. *Journal of Proteome Research*, *17*(1), 751–758. https://doi.org/10.1021/acs.jproteome.7b00807

Sadygov, R. G. (2021). Using Heavy Mass Isotopomers for Protein Turnover in Heavy Water Metabolic Labeling. *Journal of Proteome Research*, *20*(4), 2035–2041. https://doi.org/10.1021/acs.jproteome.0c00873

Sanyal, S. C., & Liljas, A. (2000). The end of the beginning: Structural studies of ribosomal proteins. *Current Opinion in Structural Biology*, *10*(6), 633–636.

https://doi.org/10.1016/S0959-440X(00)00143-3

Schaffer, L. V., Millikin, R. J., Miller, R. M., Anderson, L. C., Fellers, R. T., Ge, Y., Kelleher, N. L., LeDuc, R. D., Liu, X., Payne, S. H., Sun, L., Thomas, P. M., Tucholski, T., Wang, Z., Wu, S., Wu, Z., Yu, D., Shortreed, M. R., & Smith, L. M. (2019). Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics*, *19*(10), 1800361. https://doi.org/10.1002/pmic.201800361

Scheraga, H. A., Liwo, A., Oldziej, S., Czaplewski, C., Pillardy, J., Lee, J., Ripoll, D. R., Vila, J. A., Kazmierkiewicz, R., Saunders, J. A., Arnautova, Y. A., Gibson, K. D., Jagielska, A., Khalili, M., Chinchio, M., Nanias, M., Kang, Y. K., Schafroth, H., Ghosh, A., … Makowski, M. (2006). The Protein Folding Problem. In *Lecture Notes in Computational Science and Engineering* (Vol. 49, Issue 1, pp. 90–100). Springer-Verlag. https://doi.org/10.1007/3-540-31618-3_6

Selvi, B. R., & Kundu, T. K. (2009). Reversible acetylation of chromatin: Implication in regulation of gene expression, disease and therapeutics. *Biotechnology Journal*, *4*(3), 375–390. https://doi.org/10.1002/biot.200900032

Sénécaut, N., Alves, G., Weisser, H., Lignières, L., Terrier, S., Yang-Crosson, L., Poulain, P., Lelandais, G., Yu, Y. K., & Camadro, J. M. (2021). Novel Insights into Quantitative Proteomics from an Innovative Bottom-Up Simple Light Isotope Metabolic (bSLIM) Labeling Data Processing Strategy. *Journal of Proteome Research*, *20*(3), 1476–1487. https://doi.org/10.1021/acs.jproteome.0c00478

Sénécaut, N., Poulain, P., Lignières, L., Terrier, S., Legros, V., Chevreux, G., Lelandais, G., & Camadro, J.-M. (2022). Quantitative Proteomics in Yeast: From bSLIM and Proteome Discoverer Outputs to Graphical Assessment of the Significance of Protein Quantification Scores. In Frederic Devaux (Ed.), *Yeast Functional Genomics: Methods and Protocols* (2nd ed., pp. 275–292). https://doi.org/10.1007/978-1-0716-2257-5_16

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Senko, M. W., Beu, S. C., & McLaffertycor, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, *6*(4), 229–233. https://doi.org/10.1016/1044-0305(95)00017-8

Shi, S. D. H., Hendrickson, C. L., & Marshall, A. G. (1998). Counting individual sulfur atoms in a protein by ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry: Experimental resolution of isotopic fine structure in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(20), 11532–11537. https://doi.org/10.1073/pnas.95.20.11532

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Sleator, R. D. (2012). Functional Genomics. In M. Kaufmann & C. Klinger (Eds.), *Genome Research* (Vol. 815, Issue 2). Springer New York. https://doi.org/10.1007/978-1-61779-424-7

Smith, L., Agar, J., Chamot-Rooke, J., Danis, P., Ge, Y., Loo, J., Pasa-Tolic, L., Tsybin, Y., & Kelleher, N. (2020). The Human Proteoform Project: A Plan to Define the Human Proteome. *Preprint*, 1–18. https://www.preprints.org/manuscript/202010.0368/v1

Smith, L. M., Agar, J. N., Chamot-rooke, J., Danis, P. O., Ge, Y., & Loo, J. A. (2021). The Human Proteoform Project : Defining the Human Proteome. *Science Advances*, *2021*(November), 1–12. https://doi.org/10.1126/sciadv.abk0734

Spitz, O., Erenburg, I. N., Kanonenberg, K., Peherstorfer, S., Lenders, M. H. H., Reiners, J., Ma, M., Luisi, B. F., Smits, S. H. J., & Schmitt, L. (2022). Identity Determinants of the Translocation Signal for a Type 1 Secretion System. *Frontiers in Physiology*, *12*(February), 1–13. https://doi.org/10.3389/fphys.2021.804646

Steen, H., & Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, *5*(9), 699–711. https://doi.org/10.1038/nrm1468

Steffen, K. K., McCormick, M. A., Pham, K. M., Mackay, V. L., Delaney, J. R., Murakami, C. J., Kaeberlein, M., & Kennedy, B. K. (2012). Ribosome Deficiency Protects Against ER Stress in Saccharomyces cerevisiae. *Genetics*, *191*(1), 107–118. https://doi.org/10.1534/GENETICS.111.136549

Stewart, I. I., Thomson, T., & Figeys, D. (2001). O labeling: A tool for proteomics. *Rapid Communications in Mass Spectrometry*, *15*(24), 2456–2465. https://doi.org/10.1002/rcm.525

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., & Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, *75*(8), 1895–1904. https://doi.org/10.1021/ac0262560

Todd, A. E., Orengo, C. A., & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, *307*(4), 1113–1143. https://doi.org/10.1006/jmbi.2001.4513

Tsakou, F., Jersie-Christensen, R., Jenssen, H., & Mojsoska, B. (2020). The role of proteomics in bacterial response to antibiotics. *Pharmaceuticals*, *13*(9), 1–27. https://doi.org/10.3390/ph13090214

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I. M., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419

Valkenborg, D., Assam, P., Thomas, G., Krols, L., Kas, K., & Burzykowski, T. (2007). Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. *Rapid Communications in Mass Spectrometry*, *21*(20), 3387–3391.

https://doi.org/10.1002/rcm.3237

Van De Waterbeemd, M., Tamara, S., Fort, K. L., Damoc, E., Franc, V., Bieri, P., Itten, M., Makarov, A., Ban, N., & Heck, A. J. R. (2018). Dissecting ribosomal particles throughout the kingdoms of life using advanced hybrid mass spectrometry methods. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04853-x

Videler, H., Ilag, L. L., McKay, A. R. C., Hanson, C. L., & Robinson, C. V. (2005). Mass spectrometry of intact ribosomes. *FEBS Letters*, *579*(4 SPEC. ISS.), 943–947. https://doi.org/10.1016/j.febslet.2004.12.003

Waanders, L. F., Hanke, S., & Mann, M. (2007). Top-down quantitation and characterization of SILAC-labeled proteins. *Journal of the American Society for Mass Spectrometry*, *18*(11), 2058–2064. https://doi.org/10.1016/j.jasms.2007.09.001

Wagner, G. R., & Payne, R. M. (2013). Widespread and enzyme-independent Nε-acetylation and Nε-succinylation of proteins in the chemical conditions of the mitochondrial matrix. *Journal of Biological Chemistry*, *288*(40), 29036–29045. https://doi.org/10.1074/jbc.M113.486753

Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., & von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, *15*(18), 3163–3168. https://doi.org/10.1002/pmic.201400441

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., & Von Mering, C. (2012). PaxDb, a database of protein abundance averages across all three domains of life. *Molecular and Cellular Proteomics*, *11*(8), 492–500. https://doi.org/10.1074/mcp.O111.014704

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*, *24*(11), 437–440. https://doi.org/10.1016/S0968-0004(99)01460-7

Weber, R. J. M., Southam, A. D., Sommer, U., & Viant, M. R. (2011). Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification. *Analytical Chemistry*, *83*(10), 3737–3743. https://doi.org/10.1021/ac2001803

Weisser, H., & Choudhary, J. S. (2017). Targeted Feature Detection for Data-Dependent Shotgun Proteomics. *Journal of Proteome Research*, *16*(8), 2964–2974. https://doi.org/10.1021/acs.jproteome.7b00248

Winkler, R. (2010). ESIprot: A universal tool for charge state determination and molecular weight calculation of proteins from electrospray ionization mass spectrometry data. *Rapid Communications in Mass Spectrometry*, *24*(3), 285–294. https://doi.org/10.1002/rcm.4384

Woolford, J. L., & Baserga, S. J. (2013). Ribosome Biogenesis in the Yeast Saccharomyces cerevisiae. *Genetics*, *195*(3), 643–681. https://doi.org/10.1534/genetics.113.153197

Yamashita, M., & Fenn, J. B. (1984). Electrospray ion source. Another variation on the free-jet theme. *Journal of Physical Chemistry*, *88*(20), 4451–4459. https://doi.org/10.1021/j150664a002

Ye, X., Luke, B., Andresson, T., & Blonder, J. (2009). 18O stable isotope labeling in MS-based proteomics. *Briefings in Functional Genomics and Proteomics*, *8*(2), 136–144. https://doi.org/10.1093/bfgp/eln055

Young, M. J., & Court, D. A. (2008). Effects of the S288c genetic background and common auxotrophic markers on mitochondrial DNA function in Saccharomyces cerevisiae. *Yeast*, *25*(12), 903–912. https://doi.org/10.1002/yea.1644

Zentner, G. E., & Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nature Structural and Molecular Biology*, *20*(3), 259–266. https://doi.org/10.1038/nsmb.2470

Zhong, Y., Kanagaratham, C., & Radzioch, D. (2013). Chromatin Remodelling During Host-Bacterial Pathogen Interaction. In *Chromatin Remodelling*. InTech. https://doi.org/10.5772/55977

ANNEXES

Annexe 1 : La notion d'averagine est discutable

La notion d'acide aminé moyen de composition chimique moyenne, l'averagine, a été définie en 1995 (Senko et al., 1995). Cette notion permet de rapidement calculer une approximation de la masse des protéines sans connaissance précise de la séquence.

L'averagine, moyenne des compositions existantes, a été calculée sur toutes les ORFs connues au moment de sa présentation, il y a 20 ans. Or depuis cette époque, de très nombreux génomes ont été entièrement séquencés (humain notamment et de multiples autres organismes). Le coût du séquençage a également considérablement diminué et les techniques utilisées sont tout à fait éprouvées. Notamment, si une première version de la séquence du génome humain complet avait nécessité des millions de dollars, aujourd'hui, quelques centaines de dollars seulement sont nécessaires pour séquencer les gènes composant le génome de multiples organismes.

Au début des années 2000, quand a été introduit et calculé la notion d'averagine, la banque de données des protéomes PIR3 possédait 1 341 021 entrées issues de 50 organismes vivants (Release 1.31, 22-Sep-2003). Actuellement, Uniprot (successeur de PIR3) regroupe actuellement les protéomes de plus de 1 274 389 espèces décrites représentant 230 328 648 séquences protéiques (Release 2022_01).

En une vingtaine d'année, le nombre d'entrées a donc été multiplié par un facteur 230. C'est la raison pour laquelle la notion d'Averagine est à revoir ou à adapter selon l'organisme étudié. En effet, chaque organisme possède une histoire évolutive et un environnement de vie différent. Il existe donc une grande diversité de composition des protéines en acides aminés, car les propriétés de ces derniers servent à maintenir la structure protéique ou à capter certains éléments dans des conditions extrêmes (température, humidité, oxygénation, pH, etc ..).

	C	H	O	N	S
ARATH	4.560010696	7.1945215	1.36147511	1.25206936	0.04238189
BACSU	4.772682481	7.53373911	1.39218914	1.25694606	0.03703325
CANAL	4.640068407	7.28855262	1.41418312	1.24497055	0.02743061
CANGA	4.647695013	7.33199785	1.41166284	1.26299045	0.03304459
CAEEL	4.656604209	7.3038392	1.37890767	1.26516446	0.04674678
CHLRE	4.259640415	6.77134953	1.30113084	1.25898435	0.03401394
DEIRA	4.579326956	7.27986713	1.3413029	1.30949613	0.02630568
DROME	4.560914497	7.19118461	1.37839769	1.28091467	0.04187464
ECOLI	4.701732548	7.43506463	1.35976587	1.2850687	0.04074632
ENCCU	4.705904721	7.48244883	1.38107274	1.28705564	0.04769996
GIAIC	4.520264022	7.20461377	1.34351326	1.26688794	0.05126335
HUMAN	4.569797195	7.21206151	1.35862806	1.27750183	0.04557979
KLULA	4.641893544	7.3062401	1.40678655	1.25260324	0.03158927
METJA	4.970413265	7.95131182	1.40986896	1.26935558	0.03752456
MOUSE	4.569468641	7.20977915	1.35882426	1.27137568	0.04617647
OSTTA	4.478199425	7.14557419	1.37523468	1.31327263	0.0385003
PFAL7	4.941536486	7.74260473	1.47569242	1.32336859	0.03847066
PHATC	4.537046854	7.15387946	1.37282977	1.26978294	0.03684301
PLABA	4.92212344	7.72157364	1.45799908	1.30418688	0.0358981
PLAVS	4.716252583	7.41669501	1.41696974	1.30093467	0.03820485
RAT	4.591507934	7.23744203	1.36028508	1.27036895	0.04472668
SACS2	4.86464135	7.72867021	1.37347758	1.25355753	0.02894994
SCHPO	4.633089579	7.28535314	1.377658	1.25440521	0.03450801
SYNE7	4.657951454	7.37739022	1.35714738	1.29539787	0.02771491
YEAST	4.651917516	7.32608569	1.39446861	1.26042614	0.03369665
Mean_Proteomes	4.654027329	7.35327359	1.38237885	1.27548344	0.03787697
Senko et al (1995)	4.9384	7.7583	1.4773	1.3577	0.0417

Figure S1 : Composition chimique moyenne calculée à partir du protéome de différents organismes. Ces données non-publiés ont été produites au laboratoire.

Annexe 2 : Mesure de l'intensité des isotopologues par le spectromètre de masse

Si la précision de la mesure en masse est le critère le plus décrit et déterminant pour un spectromètre de masse, la mesure correcte de l'intensité des ions au sein d'un cluster isotopique est tout aussi importante. Cela est d'autant plus essentiel dans le cas de notre méthode de quantification qui repose sur l'interprétation et l'analyse stricte de données expérimentales produite par la machine. Lors du développement de la méthode bSLIM nous avons remarqué une trop grande dispersion de chaque valeur d'intensité des clusters isotopiques. L'origine de ces incertitudes de mesure réside dans l'acquisition du signal par l'appareil mais surtout lors du traitement *in situ* des signaux durant l'analyse. Déjà discutée dans la littérature, il a été démontré par diverses expériences que les Orbitrap de par leur conception sont biaisés dans leurs mesures.

En effet, l'Orbitrap est un outil de mesure de masse performant puisque que la précision de mesure des masses est de l'ordre de l'unité de ppm pour une masse supérieure à 600 Da (Weber et al., 2011). Mais également, la résolution des spectres obtenue est très élevée. Cependant les incertitudes sur la mesure de l'intensité des isotopologues au sein d'un cluster isotopique ont été grandement reporté. Ces valeurs représentant l'abondance des molécules sont d'autant plus critiques en métabolomique. Dans ce contexte, il a été observé dans un LTQ Orbitrap que l'erreur de mesure des intensités augmentait en fonction de la résolution. Notamment pour une résolution de 7 500, les spectres présentent une incertitude de 3%, qui peut aller jusqu'à 10% dans le cas d'une résolution à 100 000. Ce biais a notamment été étudié par (Erve et al., 2009) au travers un objectif de mesure en métabolomique, la détermination du « relative isotope abundance » (RIA).

L'origine des incertitudes de mesure réside dans la physique de l'instrument mais s'explique certainement d'avantage lors du traitement des signaux bruts et la production des spectres masse/charge. C'est la raison pour laquelle, l'un des plus grands handicaps est le traitement du signal post-traitement fait par l'algorithme

constructeur. Le signal brut produit par l'analyseur est massif, en effet chaque transient est bien trop volumineux pour être stocké simplement. Aussi un premier traitement est appliqué à l'issue de la transformé de Fourier, un niveau seuil est appliquée d'environ 5 sigmas par rapport aux variations du signal. Toutes les valeurs en dessous de ce seuil auront une intensité nulle. Les valeurs nulles étant écartées, cela réduit significativement la taille des données à stocker. Généralement, ce genre de compromis ne perturbe pas les utilisateurs désirant une mesure de masse. Cependant dans le cas d'un usage des signaux bruts bien plus exigeant tel que le nôtre, cela représente une difficulté. L'une des solutions commerciales de plus en plus utilisée est un boitier électronique de l'entreprise Spectro-swiss. Il permet d'accéder aux signaux et d'enregistrer les données brutes. De plus, cette entreprise propose un algorithme défini afin de traiter le signal différemment et d'extraire des données plus exhaustives sur les transients. Cela garantit un accès d'excellente qualité à la source des signaux, bien supérieure à l'outil (spectromètre de masse) vendu originalement. Cependant le système Spectro-swiss reste onéreux et son implémentation n'est pas aisée pour un usage courant.