Advances in analytical methodologies for the characterization and quantification in proteomic analysis

Diego Bertaccini

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

UMR7178-IPHC

THÈSE

Présentée par Diego Bertaccini

Soutenue le 30 septembre 2014 pour obtenir le grade de

Docteur de l'Université de Strasbourg

Discipline / Spécialité: Chimie analytique

Analyse protéomique:

progrès en caractérisation et en quantification

Advances in analytical methodologies

for the characterization and quantification

in proteomic analysis

THÈSE dirigée par :

Dr. Alain VAN DORSSELAER — CNRS, Université de Strasbourg
Dr. Christine SCHAEFFER-REISS — CNRS, Université de Strasbourg

RAPPORTEURS :

Dr. Philippe MARIN — CNRS Université de Montpellier
Pr. Andreas THOLEY — Professeur, University of Kiel

MEMBRES DU JURY:

Pr. Jacques HAIECH — Université de Strasbourg
Dr. Virginie REDEKER — INSERM CNRS, Gif-sur-Yvette

To my wife Federica

To my family

«Genius is one percent inspiration, ninety-nine percent perspiration. »

Thomas Edison.

Thanks

I would like to thank Alain Van Dorsselaer for having welcomed me in his group during these four years. Thank you for giving me the possibilities of learning and working in idyllic conditions.

Thanks to Christine Schaeffer-Reiss for having accepted the challenge of making of a biologist a good mass spectrometrist, and moreover thank you for your empathy and 24/7 availability even during the holidays! It has been pleasure learning from you.

Thanks to Philippe Marin, Andreas Tholey, Jacques Haiech, and Virginie Redeker for evaluating my thesis.

Thanks to Merck Serono for founding my Ph.D. and my line of research, especially to Carlo Giartosio, Mara Rossi e Horst Bierau for making this happen.

Many thanks to the entire LSMBO: management, researcher, post-doc, student and stagier for all the good time we had together.

Thanks to Marine and Sarah for all what we shared: day night and weekend in lab (sometimes also outside the lab). Thanks you for having never thrown anything harder than an anti stress “nouage”.

Thanks to the many friends that shared with me the office, the joy and sometimes the sorrow of a ph.D.: Anna, Eli Magali Nina in R5 and when I moved in R2 Giles Benoit Lesly (LSA was also again there..)

Thanks Daniel for her precious help in the Bio-lab, thanks to Jan-Marc for letting me use the G1.

Thanks to the SRM club and the label free lovers club guided by Christinette and Agnes (double thanks to Agnes she also read all my chapters).

Thanks to Francois to be always ready to help also outside the lab in the everyday life. Thanks to Fabrice, Alex, Alex 2 and Patrick for having always realized my exotic informatics needs. Thanks to Kevin for the huge help in solving my French and Italian bureaucracy issues and thank to Laurance for her warm welcome in R2.

Thanks to Amandine her guide during my targeted problems with the TSQ, obviously thanks to the “non-cov” friend Sarah C. Francoise Juellien Guillaume Johann for responding to my non sense questions about antibody.

Thanks to the “bureux de precieux” Allison, Giorgio, Sebastien, Luc and to who is starting a new adventure in the LSMBO Gauthier, Margaux and Charlotte.

An ISO thanks to Hellen and Veronique and not certified one for Fabrice B. also thanks to who left the lab. Cyril, Nico, Cédric and to I might have forgot.

Thanks to my wife for her infinite support during these years, and to my family

Diego

Table des matières (Contents)

Résumé de thèse en Français

Analyse protéomique: progrès en caractérisation et en quantification

Résumé de thèse en Français

Analyse protéomique: progrès en caractérisation et en quantification

INTRODUCTION GENERALE: importance et limitations en analyse protéomique appliquée à la biologie.

L'analyse protéomique, née en 1995, n'a pas encore tenu toutes les promesses que de nombreux auteurs lui prêtaient, et ceci principalement pour les deux raisons suivantes :

- *Elle ne permet, en général, qu'une simple identification des protéines, et non pas leur caractérisation complète*. L'analyse protéomique de routine permet aujourd'hui de collecter, sur des milliers de protéines, suffisamment d'informations structurales pour les « reconnaitre ». On parle alors d'identification de protéines présentes dans l'extrait analysé. En général, moins de 10 % de la séquence de chaque protéine est ainsi déterminée, et c'est seulement par extrapolation que chaque protéine est dite « identifiée ». Mais on ne peut en aucun cas parler d'une « caractérisation » au sens chimique du terme. Ce ne serait le cas que si la totalité de la structure pouvait être déduite, de façon univoque, des données expérimentales. Une corrélation fiable entre structure et activité biologique, ne peut reposer que sur une connaissance totale de la structure de la molécule, et donc sur sa caractérisation totale.

- *Elle ne donne souvent que des données de quantification très approximatives*. Il est admis qu'entre les protéines majeures et les protéines mineures, il peut y avoir une différence de concentration d'un facteur 10^{12}. Dans ces conditions, il est compréhensible qu'un résultat d'analyse protéomique qui se limiterait à dire « telle protéine est présente » sans aucune donnée quantitative, serait d'un intérêt limité pour un biologiste.

Notre travail de thèse avait pour objectif de progresser sur ces deux points (caractérisation et quantification). Nous avons donc développé de nouvelles méthodologies pour obtenir sur chaque protéine déclarée « identifiée » (et donc présente) dans un extrait protéique, d'avantage de précisions sur sa structure, y compris sur les éventuelles modifications post-traductionnelles. Nous avons également développé des approches qui permettent d'obtenir des données de quantification plus précises pour permettre au biologiste de donner un sens aux identifications de protéines réalisées par analyse protéomique.

Nous avons voulu développer des méthodes compatibles avec l'analyse à haut débit, puisque les protéomes sont dynamiques, et ceci contrairement aux séquences génomiques.

L'objectif d'une caractérisation complète de toutes les protéines présentes dans un extrait cellulaire semble aujourd'hui hors d'atteinte. Mais nous pensons qu'il le sera un jour, comme l'objectif de séquencer le génome d'un individu en une journée est atteint aujourd'hui.

Plutôt que de développer et de tester nos méthodes sur des extraits protéiques standards, nous avons tenu à travailler sur échantillons de la « vraie vie » pour lesquels de nombreux problèmes sont rencontrés : faible quantité, tampon d'extraction non optimisé pour l'analyse protéomique, quantité de matériel difficile à évaluer, par exemple. C'est pour cela que mes travaux ont donné lieu à des collaborations avec des laboratoires de biologie.

Mon manuscrit de thèse s'articule autour de deux parties.

La première partie est consacrée à la détermination exacte de la position N-terminale des protéines. J'aborde d'abord l'étude des clivages protéolytiques des protéines exportées par le parasite *Plasmodium falciparum* dans le globule rouge (chapitre I) grâce à la stratégie de marquage « in-gel » au TMPP par la méthodologie N-TOP (N-terminal Oriented Proteomics). Je décris ensuite une amélioration importante de la méthodologie N-TOP permettant la détermination des positions N-terminales et qui ouvre la porte à une exploitation automatique à haut débit des données MS/MS (chapitre II).

La seconde partie de mon travail est dédiée aux développements de différentes stratégies analytiques de quantification appliquées à une série de problématiques biologiques. Celles-ci portent sur l'étude des complexes protéiques avec une méthode de comptage des spectres (chapitre I), sur le dosage de prion par deux méthodes de quantification (chapitre II), sur la détermination des glycations d'anticorps monoclonaux thérapeutiques (chapitre III), sur le dosage de l'hémoglobine A2 par LC-MS (chapitre IV), et enfin sur l'optimisation d'une nouvelle méthode de balayage MS/MS en mode « Data Independant Acquisition » (chapitre V).

Le centre de gravité de mon travail est donc le développement de méthodologies analytiques, plutôt que la biologie elle-même.

Partie I Caractérisation de la position N-terminale des protéines

Pour améliorer la caractérisation des protéines, nous avons fait porter nos efforts sur la détermination des positions N-terminales. Leur connaissance est nécessaire pour l'étude des processus protéolytiques qui affectent les protéines en modifiant leur activité biologique. Ces processus protéolytiques ne peuvent pas toujours être déterminés par analyse *in silico* du génome.

Une méthodologie avait donc été développée au laboratoire pour la caractérisation des extrémités N-terminales. Cette stratégie, appelée N-TOP, est basée sur un marquage spécifique des amines N-terminales des protéines (1) par un réactif chimique, le TMPP-AcOSu. Cette méthode a été développée de telle manière qu'elle puisse être intégrée de façon simple dans une analyse protéomique classique avant l'étape de fractionnement des protéines par gel SDS-PAGE.

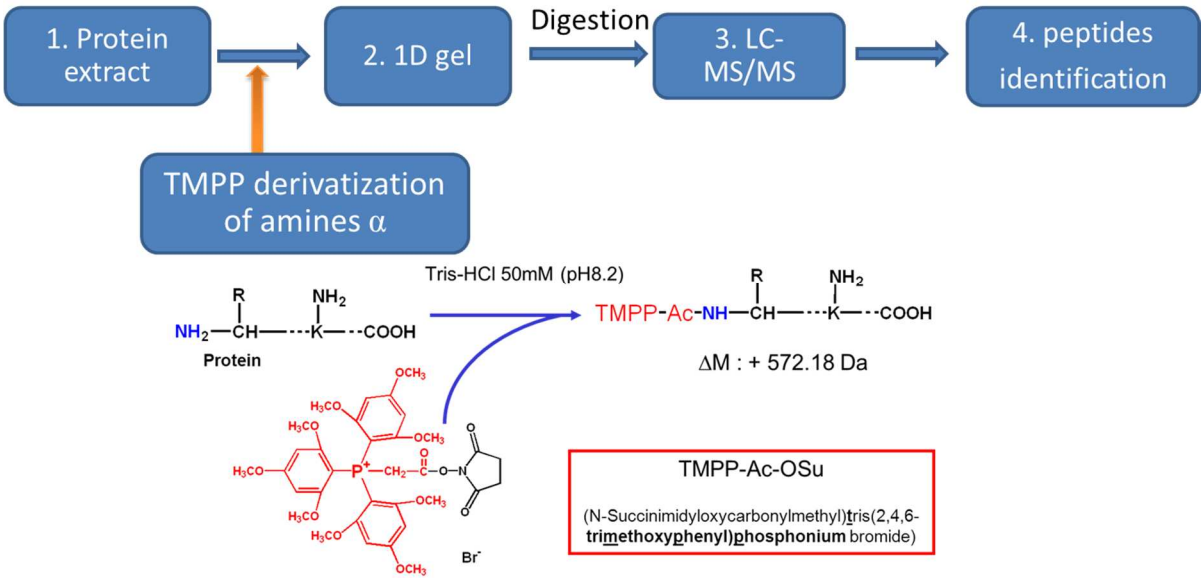


Figure 1 Représentation schématique de la stratégie N-TOP. Une étape supplémentaire de dérivation chimique au TMPP est introduite dans un workflow d'analyse protéomique

Au cours de mon travail de thèse, je me suis focalisée sur deux points d'amélioration de cette approche N-TOP, à savoir le marquage chimique de protéines immobilisées dans le gel (**chapitre I**) et une automatisation de la validation des données grâce à l'utilisation d'un mélange de réactifs TMPP lourd et léger (**chapitre II**).

Partie I - Chapitre I - Caractérisation des extrémités N-terminales des protéines exportées de *Plasmodium falciparum*

Plasmodium falciparum, agent pathogène de la malaria, infecte les érythrocytes humains et, pour survivre dans cet environnement hostile, il doit y installer des fonctions vitales en exportant plus de 300 protéines dans sa cellule hôte. La compréhension de ce mécanisme sophistiqué d'export des protéines (trafficking) constitue donc une clef pour identifier de nouvelles cibles thérapeutiques.

L'objectif analytique de ce projet était de déterminer précisément quels sont les acides aminés de ces séquences consensus conservées (formé d'un motif pentapeptidique **R x L x (Q/E/D)** appelé Pexel pour ***P**lasmodium* **EX**port **EL**ement, voir figure 2) qui sont essentiels pour l'exportation de certaines protéines dans la cellule hôte.

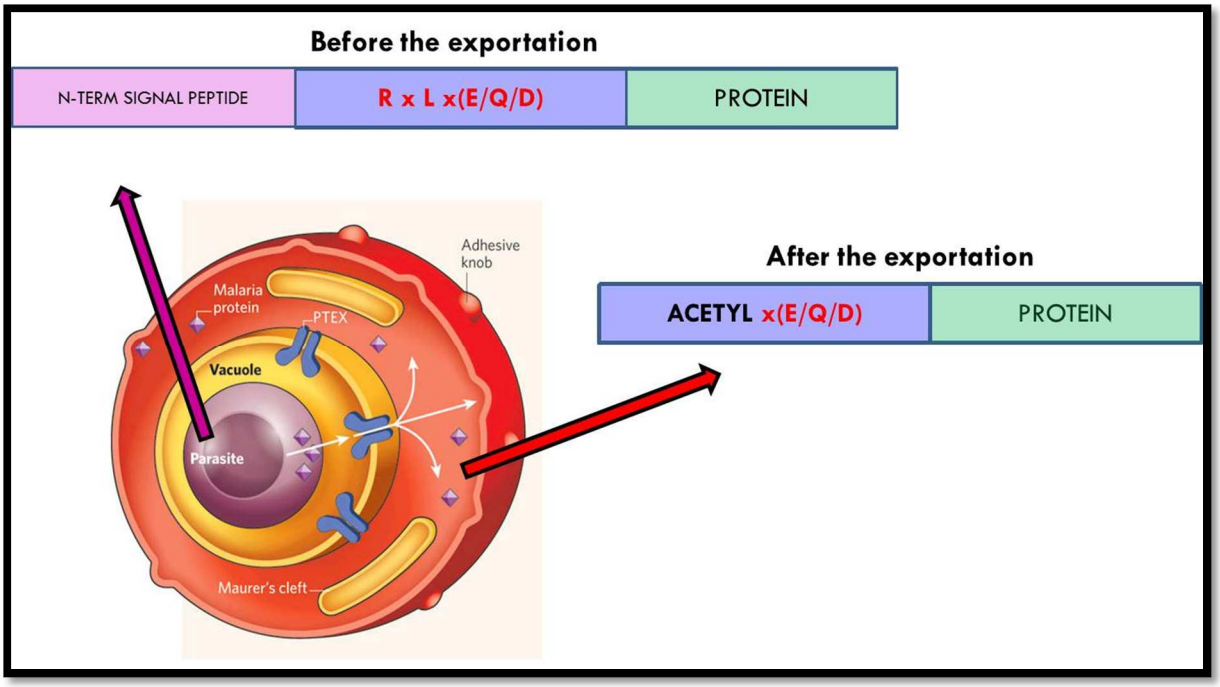


Figure 2: représentation schématique d'un globule rouge infecté. La membrane plasmatique du parasite est entourée d'une vacuole parasitophore dont sa membrane forme une barrière entre le pathogène et le cytosol de l'érythrocyte. Une protéine synthétisée dans le parasite est composée de 3 parties : l'extrémité N-terminale (en violet), le motif Pexel (en bleu) et la partie C-terminale (en vert). Pour être exportée, le motif Pexel doit être reconnu et clivé par des protéases spécifiques et acétylés (représentation schématique d'une protéine Pexel avant et après exportation).

La stratégie "Ingel N-TOP", développée par D. Ayoub au cours de sa thèse réalisée au LSMBO, représente la stratégie analytique optimale pour atteindre cet objectif dans la mesure où la dérivation à l'aide de TMPP réalisée après l'étape de fractionnement par gel 1D permet :

- d'éliminer les détergents et tampons incompatibles avec une analyse LC-MS/MS
- d'augmenter la sensibilité de détection des protéines du fait du pré-fractionnement de l'extrait protéique

Nous avons donc optimisé le marquage chimique au TMPP de la position N-terminale des protéines immobilisées dans un gel SDS PAGE pour une meilleure détection de ces peptides N-terminaux souvent peu abondants.

Le contrôle du pH était particulièrement critique pour limiter le nombre de réactions secondaires sur les chaines latérales des lysines et des tyrosines.

La quantité de matériel disponible très faible et la présence de détergents dans l'extrait protéique immunoprécipité ne permettait pas de mesurer le contenu de ces échantillons de *P. falciparum*. Pour déterminer la quantité de réactif à rajouter dans le gel, nous avons chargé sur le même gel différentes quantités de protéines afin de corréler l'intensité de la coloration au bleu de Coomassie des bandes des protéines chimériques et celle de l'hémoglobine. Cette approche nous a permis de minimiser les réactions secondaires liées à un ratio trop élevé de réactif.

Le second point délicat consistait à améliorer l'étape d'élimination de l'excès de TMPP, qui est habituellement réalisée lors de l'étape de gel 1D lorsque la dérivatisation est menée en solution.

Pour chaque mutant de ce motif Pexel, j'ai mis en place un protocole spécifique de digestion afin de générer un peptide N-terminal spécifique qui permette de générer des spectres de fragmentation. A titre d'exemple, la figure 3 présente un spectre MS/MS qui a permis l'identification du peptide N-terminal acétylé dans le cas du construit chimérique Q52N.

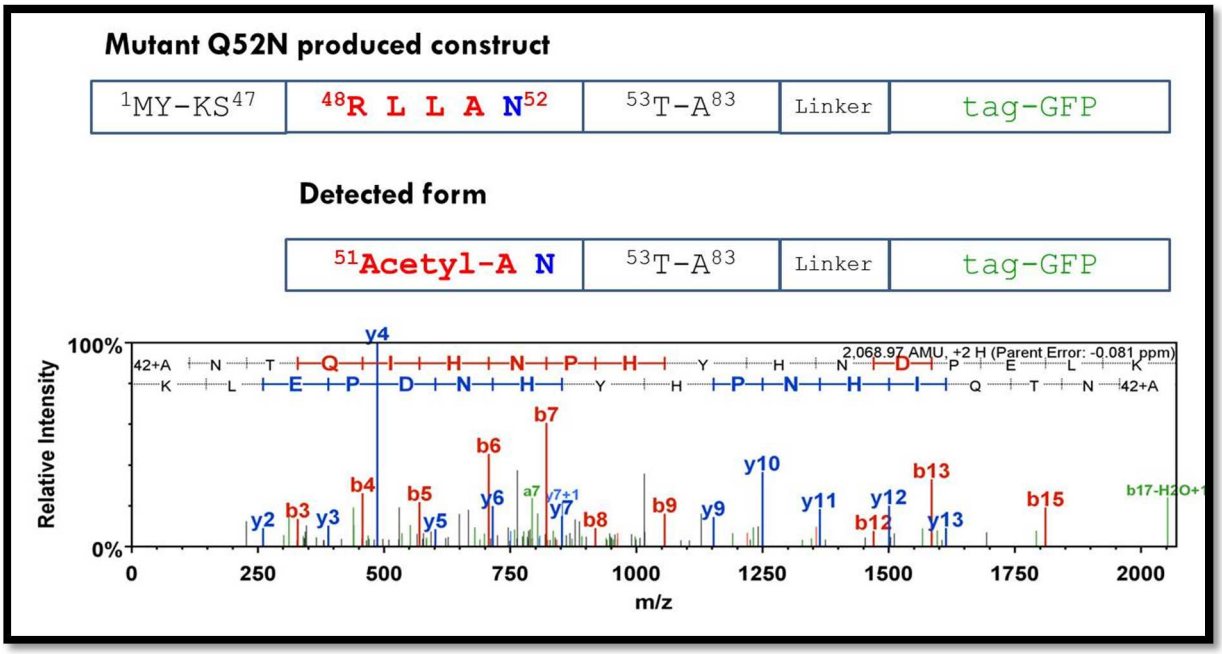


Figure 3 : représentation schématique du construit chimérique Q52N avant et après exportation et spectre MS/MS du peptide N-terminal acétylé.

Pour caractériser la forme exportée d'un des mutants étudiés (mutation d'un acide aminé essentiel pour cet adressage), nous avons également eu recours à une méthode de détection plus

sensible en combinant l'utilisation du peptide de référence marqué aux isotopes stables et d'une méthode de SRM.

Les études complémentaires de biologie moléculaire pour valider ces résultats de protéomique sont en cours de finalisation. Ces résultats de caractérisation des sites de clivage protéolytique sont actuellement en cours de rédaction.

Ces développements de marquage chimique dans le gel ont également permis de caractériser les extrémités N-terminales d'anticorps monoclonaux dans le cadre d'un projet mené en collaboration avec Pierre Fabre. Ces résultats vont être soumis à publication dans le journal Mabs.

Partie I - Chapitre II - Amélioration de la stratégie N-TOP grâce à l'utilisation d'un mélange de réactif TMPP lourd et léger.

Le point faible de la stratégie N-TOP développée précédemment au laboratoire concernait l'étape d'interprétation manuelle et donc fastidieuse des données de spectres de fragmentation MS/MS. En effet, la charge permanente apportée par le groupement phosphonium du TMPP conduit à une fragmentation atypique qui va générer des scores souvent faibles mais qui ne sont pas directement corrélés à une mauvaise « qualité » de spectres MS/MS. En fait, les algorithmes de recherche utilisés en analyse protéomique ne sont pas adaptés pour ces peptides modifiés par du TMPP, ce qui nécessitait de vérifier manuellement la qualité des spectres. Lorsqu'une analyse protéomique génère plusieurs centaines de spectres MS/MS, cette étape de vérification manuelle constituait le point faible de cette stratégie qui rendait cette approche difficilement compatible avec des analyses à haut débit.

Nous avons donc développé une approche qui consiste à dériver les protéines avec un mélange équimolaire de réactifs TMPP lourd (TMPP portant 9 atomes de ^{13}C) et TMPP léger (figure 4). L'identification des protéines, réalisée avec les moteurs de recherche utilisés dans notre stratégie d'analyse protéomique classique, permet ainsi de générer une liste de peptides N-terminaux lourds et légers (identifiés quel que soit leur score). Les peptides N-terminaux seront ensuite validés sur la base de leur séquence et temps de rétention identiques.

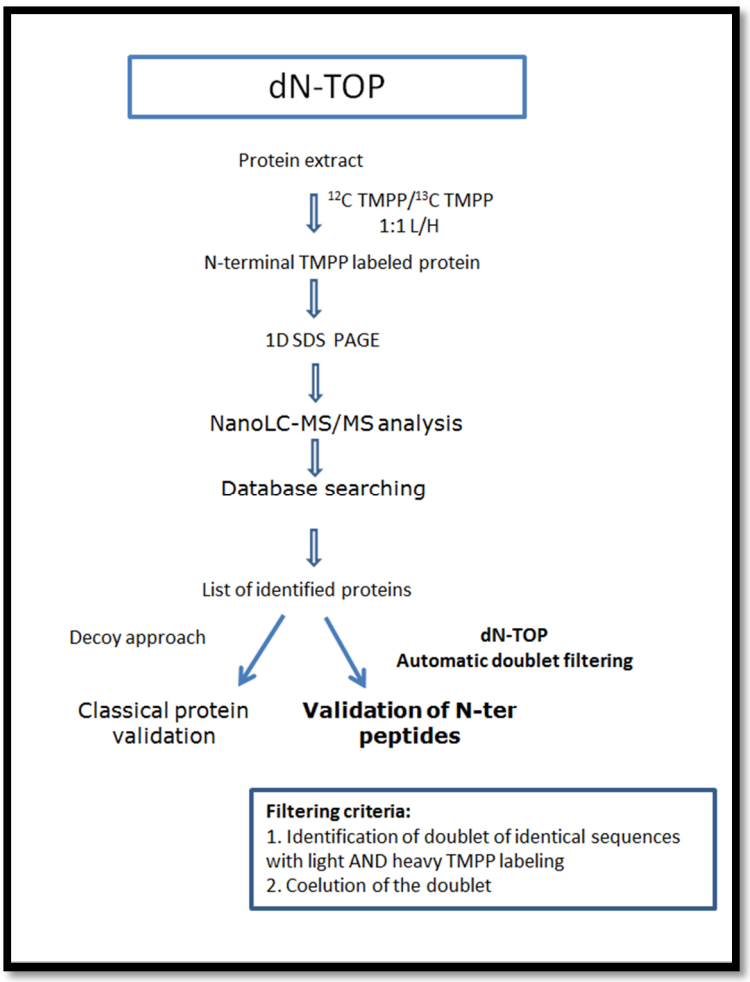


Figure 4 : Représentation schématique de la stratégie dN-TOP (doublet N-TOP).

L'extrait protéique est dérivé avec un mélange de TMPP lourd/léger (1 :1). Le mélange est ensuite soumis au protocole analytique standard. L'analyse par LC-MS/MS donne ainsi une liste de peptides N-terminaux lourds et légers. De cette liste sont ensuite extraits les peptides qui présenteront la même séquence peptidique avec une étiquette TMPP lourd et léger et qui co-élueront parfaitement en LC.

Cette approche a été appliquée à l'étude d'une bactérie *Herminiimonas arsenicoxydans*, bactérie impliquée dans la résistance à l'arsenic, déjà étudiée par le passé au laboratoire par une approche d'analyse protéomique classique.

Les résultats du N-terminome de cette bactérie nous ont permis de valider automatiquement une centaine de peptides N-terminaux, de corriger 13 sites « start » mal prédit, de mettre en évidence des sites de clivages protéolytiques et d'identifier plus de 500 protéines.

A titre d'exemple, la figure 5 présente le cas d'un peptide N-terminal expérimental qui montre clairement que le start de la protéine est mal prédit.

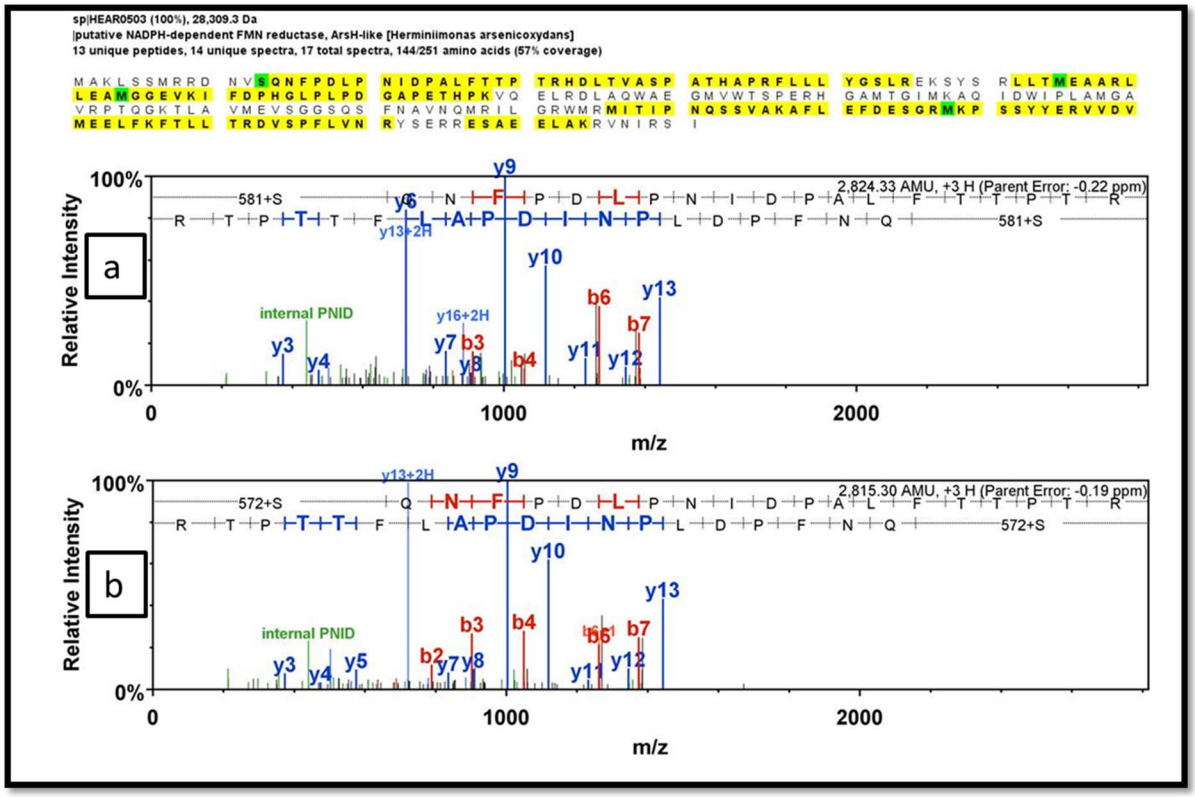


Figure 5 : Spectres MS/MS du peptide N-terminal de la protéine putative NADPH-dependent FMN reductase, ArsH-like protein. A) spectre du peptide marqué au TMPP lourd. B) spectre du peptide marqué au TMPP léger

En haut : la séquence protéique avec les acides aminés identifiés en jaune, les acides aminés modifiés en vert. L'extrémité N-terminale est validée en position 13.

Ces résultats ont donné lieu à une publication dans Journal of Proteome Research, 2013.

Cette nouvelle approche dN-TOP va permettre d'aborder des études pour réaliser l'identification massive des N-terminomes, notamment dans le cadre de l'étude des processus protéolytiques dans des globules rouges infectés par *Plasmodium falciparum*. Une étude du protéome combinée à celle du N-terminome des 3 compartiments (parasite, vacuole parasitophore et le cytosol du globule rouge infecté) permettra de fournir une vue globale des processus protéolytiques et devrait montrer l'importance de l'acétylation en N-terminale des protéines parasitaires exportées.

Partie II : Vers une meilleure quantification des protéines

L'analyse protéomique quantitative souffre encore d'un certain nombre de limitations telles que la gamme dynamique, la reproductibilité et la précision. Ces limitations ont pour origine, entre autre, une cadence de génération de spectres MS/MS insuffisante par les spectromètres, mais aussi l'absence d'outils bio-informatiques capables de traiter les volumes importants de données

spectrales produites. C'est dans ce contexte que je me suis attaché à mettre au point une série de stratégies de quantification relative et absolue pour contourner en partie ces limitations et ceci dans le cadre de plusieurs problématiques biologiques.

Partie II - Chapitre I - Quantification de sous-unités de complexes protéiques par comptage de spectres MS/MS

Nous avons développé une stratégie pour réaliser des quantifications différentielles sur des sous-unités de complexes protéiques. Nous nous sommes orientés vers une méthode de quantification sans marquage et basée sur le comptage du nombre de spectres acquis sur les peptides identifiant une protéine. Avec cette approche de quantification (appelée « spectral counting ou spectral count »), l'abondance d'une même protéine présente dans différents échantillons est comparée à partir du nombre de spectres acquis.

Dans notre stratégie, nous avons élaboré des séquences de balayage en MS et en MS/MS pour optimiser le nombre et la qualité des spectres acquis. Ceci permet d'établir un meilleur lien entre le nombre de spectres MS/MS acquis après détection d'un ion moléculaire, et la quantité totale d'un peptide élué dans la source du spectromètre de masse. La quantification est alors améliorée du point de vue de la sensibilité et de la gamme dynamique.

Nous avons appliqué cette stratégie à l'étude de l'impact des modifications post-traductionnelles, portées par l'histone H2B, sur la maturation du RNA et l'export nucléaire. Cette étude s'est faite dans le cadre d'une collaboration avec le Pr C. Dargemont de l'Institut Monod à Paris.

L'objectif de ce travail était d'identifier, dans un premier temps, toutes les protéines présentes dans un complexe ribonucléoprotéique (mRNP). La première étape de ce projet consistait donc à établir une cartographie des protéines présentes dans ces immunopurifications (TAP-Tag).

Une première série d'expériences a été menée pour déterminer la complexité des échantillons en terme de nombre de protéines et de déterminer la gamme dynamique de ces immunopurifications.

Ces analyses protéomiques ont permis de mettre en évidence la présence de protéines appartenant aux complexes U1 et U2 (machine d'épissage), la machinerie d'export et le complexe THO (qui promouvoit le couplage entre la transcription et le processing des mRNA).

Cette étape exploratoire a permis de montrer que les protéines du mRNP sont très faiblement abondantes en se basant sur le nombre de peptides détectés pour chacune des protéines présentes dans le complexe.

Ces premiers résultats ont ainsi permis d'ajuster les conditions de lavage de l'étape d'immunopurification et de montrer la présence des protéines du complexe.

La seconde étape consistait à mettre en évidence les variations de ce complexe afin de déterminer comment la composition relative de ces mRNP est affectée par l'ubiquitylation de la lysine en position 123 de l'histone H2B. Nous avons comparé la composition des mRNPs provenant dun souche mutées (mutation de la lysine en arginine pour empêcher l'ubiquitylation de H2B) du point de vue de leur composition mais également du point de vue quantitatif.

Ces études protéomiques quantitatives ont permis de mettre en évidence le rôle clé de certains facteurs protéiques majeurs pour la formation d'une mRNP fonctionnelle et ainsi de montrer que l'absence d'ubiquitylation de H2B diminue l'association avec la machinerie d'export nucléaire, résultats en accord avec les tests biologiques.

Ce travail a donné lieu à une publication dans Biology of the Cell.

D'un point de vue analytique, le développement de méthodes réalisé dans ce chapitre a permis de détecter les régulations majeures de protéines faiblement abondantes dans une matrice complexe. Les nouvelles capacités de balayage des dernières générations de spectromètres de masse, notamment avec le développement de la MS/MS de type DIA (Data Independent Acquisition) qui devrait permettre d'obtenir des spectres MS/MS sur tous les peptides contenus dans un échantillon, devraient fournir des informations quantitatives plus fines.

Partie II - Chapitre II - Semi quantification des glycations d'anticorps monoclonaux thérapeutiques

Ce projet a été réalisé en collaboration avec Merck Serono (Rome, Italie).

La glycation, ou réaction de Maillard, correspond à un ensemble de réactions non enzymatiques qui débute par l'addition d'un sucre réducteur sur des acides aminés basiques (arginine, lysine). Cette première étape est ensuite suivie par une cascade de réactions de dégradation plus ou moins avancées de ce sucre (Advanced Glycation End products: AGE products) pour aboutir à des protéines glyquées de structures très hétérogènes dont certaines vont présenter des propriétés antigéniques (figure 6). Il s'agit donc d'un problème d'analyse particulièrement complexe puisque ces glycations sont minoritaires et de structures très variées.

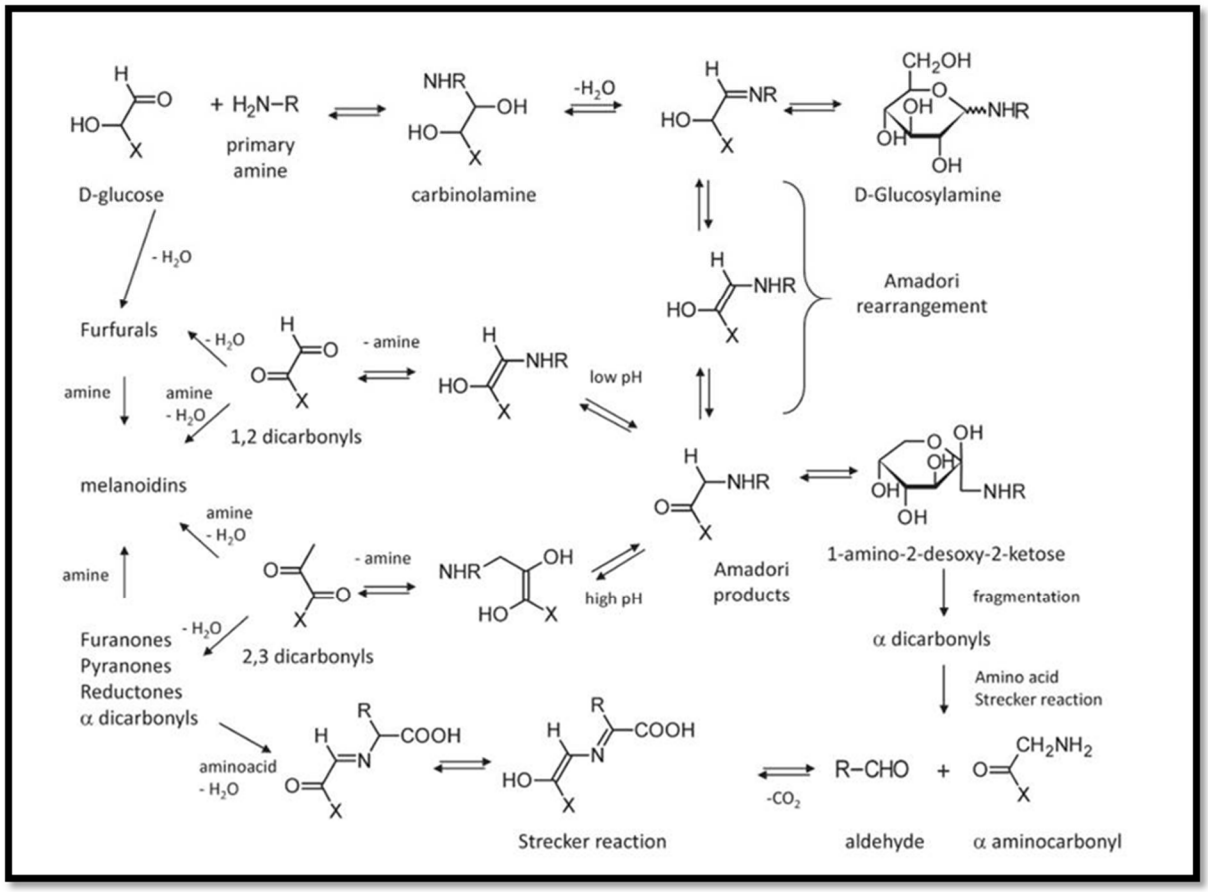


Figure 6 : représentation schématique de la réaction de Maillard (adapté de Vistoli et al, 2013).

Dans le cadre de la caractérisation des anticorps monoclonaux à usage thérapeutique (mAbs), il est donc indispensable de disposer d'outils analytiques pour déterminer la présence de ces formes glyquées. Or actuellement, les méthodes de routine (test ELISA ou enrichissement au boronate) ne permettent pas de fournir une image complète de tous les produits de glycation potentiels présents dans un échantillon protéique. La méthode de dosage utilisée par Merck-Serono est une quantification par LC-UV de la fraction enrichie, obtenue après une colonne d'affinité de boronate. Or cette colonne ne permet de retenir que les peptides dont les structures glyquées comportent des fonctions cis-diols, laissant ainsi échapper à l'analyse une fraction des formes glyquées.

Pour remédier à ce problème, j'ai donc développé une méthode pour caractériser l'ensemble des différentes formes glyquées et déterminer leurs variations relatives par spectrométrie de masse MS et MS/MS. Les difficultés analytiques étaient principalement liées à la concentration faible de ces produits de dégradation et à l'hétérogénéité chimique de ces modifications donnant souvent lieu à des fragmentations peu informatives en MS/MS.

La première étape de ce développement de méthode s'est focalisée sur l'étape de préparation d'échantillons. Les différents essais (conditions de dénaturation différentes, digestion liquide ou

in-gel) ont montré que la dénaturation à l'aide de SDS suivie d'une étape de gel stacking permet d'obtenir le meilleur recouvrement de séquence et le plus grand nombre de spectres MS/MS assignés à des peptides glyqués.

La deuxième étape s'est focalisée sur le traitement des données de MS/MS. Nous avons défini un index de glycation (GI) qui permet de comparer l'intensité d'un même peptide glyqué à travers différents lots d'anticorps.

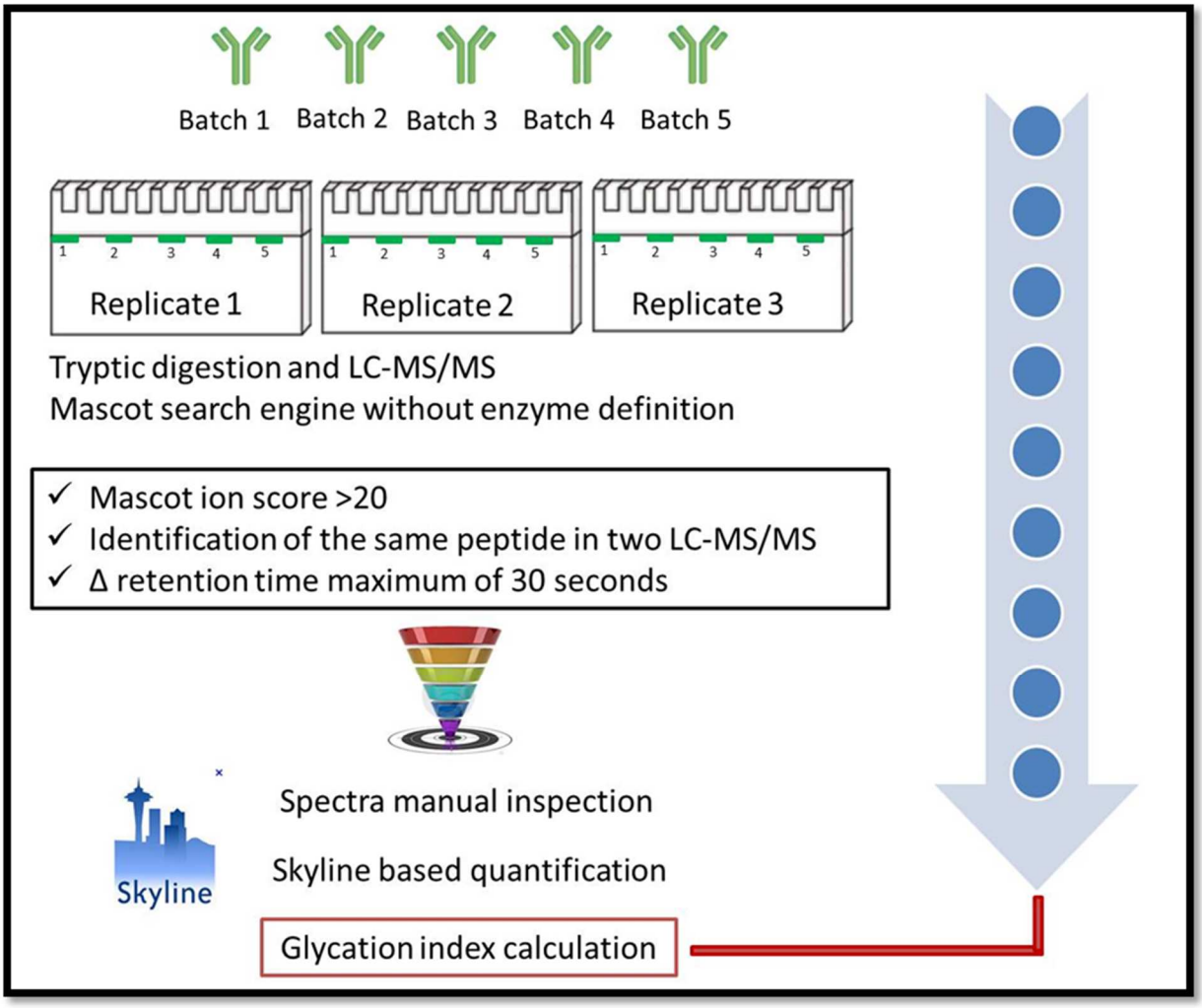


Figure 7: representation schématique de la méthode développée pour la quantification relative des glycations entre les différents lots

Le protocole développé pour l'analyse de la glycation des mAb (figure 7) a ainsi permis d'identifier 17 peptides glyqués. L'application de cette méthode à différents lots d'anticorps a permis de montrer la reproductibilité de production de ces mAbs.

Partie I - Chapitre III - Quantification de la protéine prion par approche ciblée SRM (Selected Reaction Monitoring)

Les maladies à prions, tels que les encéphalopathies spongiformes transmissibles, sont des maladies neurodégénératives mortelles causées par la conversion de la protéine prion PrP^{c} en forme « scrapie », qui elle est insoluble PrP^{sc}. Nous avons réalisé deux études qui nécessitaient des analyses protéomiques quantitatives de protéine prion par approche ciblée SRM.

La **première étude** a été réalisée en collaboration avec la compagnie pharmaceutique Merck-Serono (Rome, Italie) et le Pr N. Cashman (University of Vancouver, Canada).

Dans ce contexte, le LSMBO avait développé en 2008-2010 une approche ciblée par LC-SRM pour détecter et quantifier du prion dans des extraits urinaires d'hormones de fertilité, les gonadotrophines (FSH, LH, hCG). Cette méthode avait été mise en place suite à une étude protéomique de produits urinaires purifiés qui avait permis d'identifier, de façon totalement inattendue, la protéine prion humaine parmi une série d'autres contaminants.

L'objectif de mon travail consistait à améliorer d'un facteur 10 au moins, la sensibilité de détection du prion à partir de la méthode d'origine déjà développée au laboratoire.

Mon travail de thèse s'est focalisé sur l'amélioration de la spécificité et de la sélectivité opérée par le spectromètre de masse durant les analyses (utilisation de temps de scan ou « dwell time » plus bas sans perdre en sensibilité, une meilleure résolution du quadripôle de sélection qui a permis d'utiliser une fenêtre d'isolation plus faible). Ces diverses optimisations ont permis d'aboutir à une diminution de la limite de détection (LOD) d'un facteur 10. Cette amélioration de la méthode a ainsi permis, de détecter du prion dans des produits pharmaceutiques qui étaient considérés comme « prion free » avec la précédente méthode.

La **deuxième étude** porte sur le dosage relatif de deux variants polymorphiques du prion (collaboration avec le Pr. L. Gambetti, University of Cleveland, USA). Les phénotypes observés chez des patients affectés par la maladie du prion, pourraient être liés à la formation préférentielle de PrP^{sc} d'un individu hétérozygote pour un variant polymorphique de la protéine prion (129 Met et 129 Val).

Afin de déterminer la quantité de chaque isoforme présente, nous avons développé une méthode de quantification relative basée sur la SRM. Nous avons été confronté à un problème analytique supplémentaire, lié à la séquence de ces protéines, qui a nécessité la mise au point d'un clivage chimique à l'aide de bromure de cyanogène (CNBr). En effet aucune digestion enzymatique classique ne permettait de générer des peptides de taille adéquate pour l'analyse par SRM.

Cette coupure chimique va transformer la méthionine en homoserine mais également en une forme majeure cyclisée appelée homoserine lactone. A ces deux formes se rajoutent le problème de l'oxydation des méthionines. Or, le clivage chimique au CNBr présente l'inconvénient de ne

pas couper après une méthionine oxydée et génère ainsi plusieurs formes du peptide discriminant de chaque protéine avec 0, 1 et 2 sites de coupure manqués.

Au vue des premiers résultats, nous avons rajouté aux 2 peptides discriminants les peptides qui ont donné lieu à un miss cleavage mais seulement sous la forme majoritaire (lactone) afin d'augmenter la robustesse de la méthode de quantification.

Les résultats préliminaires ainsi obtenus sur un extrait de cerveau affecté par la maladie du prion montrent que la stratégie de quantification mise au point permet de déterminer le rapport entre les deux variants de la protéine prion présents.

Part I - Chapitre IV - Quantification de protéines entières par LC-MS, le cas de HbA_2

Le diagnostic de la β-thalassémie repose sur une quantification très précise (intervalle de confiance ±0,3% et coefficient de variation maximum de 5%) de l'hémoglobine A2 (composée de α2δ2) dans le sang, qui représente environ 3% chez un sujet sain et entre 3,6 à 6,0 % chez un sujet beta-thalassémique.

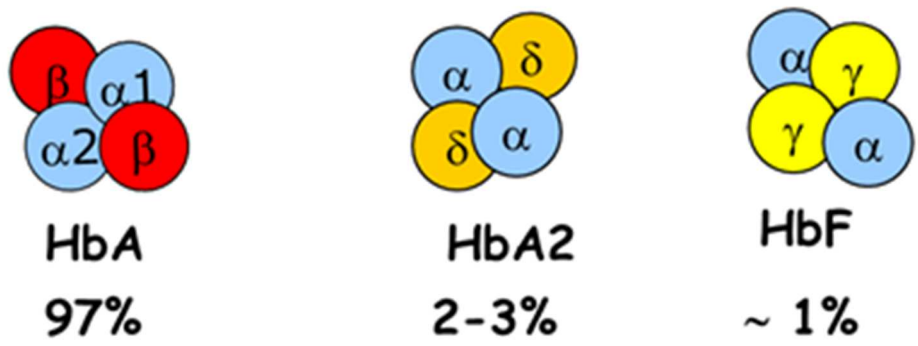


Figure 8 : présence de trois hémoglobines dans les érythrocytes d'un adulte en bonne santé.

Dans le contexte du développement d'une méthode de référence par LC-MS pour la quantification de l'hémoglobine A2 dans le sang (Collaboration avec l'International Federation of Clinical Chemistry and Laboratory Medicine), nous avons développé une méthodologie basée sur la quantification de protéines intactes (et non pas des peptides de digestion) par LC-MS.

L'objectif de ce projet est de développer une méthode permettant de garantir le contenu de HbA_2 par rapport à Hb_0 avec un coefficient de variation proche de 5% et une précision de 0,3 %, pour des standards de référence allant de 1 à 10 % de HbA_2.

Mon travail a consisté à mettre au point une méthode secondaire de quantification de l'hémoglobine A2, en parallèle de celle développée par le groupe de travail de l'IFCC qui consiste à doser des peptides de digestion des chaines alpha et delta par LC-SRM.

Après une série d'optimisations de la séparation LC et de l'utilisation de différentes configurations de spectromètre de masse, nous avons pu améliorer de façon significative la quantification de l'HbA_2, notamment en tirant partie de la précision de mesure de masse délivrée par un spectromètre de masse de type Q-TOF.

Nous avons montré que seule une interface du spectromètre de masse parfaitement propre pour une transmission des ions, avec la meilleure sensibilité possible, permettait d'obtenir des mesures reproductibles.

Cette méthode de LC-MS offre la possibilité de détecter des pics mineurs dans le chromatogramme qui pourraient correspondre à la présence de chaines de globines mutées.

Une approche quantitative par LC-MS des protéines entières, qui ne nécessitent donc pas de digestion enzymatique, représente un intérêt pour les applications dans le domaine biopharmaceutique et biothérapeutique, notamment avec l'essor des protéines recombinantes médicaments.

Partie I - Chapitre V - Optimisation de méthodes d'acquisition en mode DIA

Nous avons évalué un nouveau mode d'acquisition des spectres MS/MS, le mode DIA (Data Independant Acquisition), du point de vue de la protéomique quantitative. Le mode DIA permet de fragmenter plusieurs peptides simultanément, tandis qu'en mode d'acquisition classique DDA (Data Dependant Acquisition) seuls les peptides les plus intenses sont généralement fragmentés et de façon stochastique.

Les possibilités du mode DIA sont encore peu connues et nécessiteront beaucoup de mises au point pour permettre une meilleure quantification.

Nous avons évalué ce mode d'acquisition sur un instrument d'architecture Q-TOF de dernière génération, un Impact HD (Bruker), qui propose ce mode de balayage sous la dénomination « Middle band ».

Comme dans ce mode de balayage un nombre élevé d'ions moléculaires est admis simultanément dans la cellule de collision, il nous a fallu déterminer les meilleurs compromis pour la largeur des fenêtres d'isolation, pour le temps de cycle total requis pour couvrir toute la gamme de m/z et pour le nombre de transitions à utiliser pour limiter les interférences de la matrice.

Pour évaluer ce mode d'acquisition DIA, par rapport à la méthode d'acquisition classique DDA avec extraction des courants d'ions MS (appelée MS1), nous avons utilisé un échantillon représentatif des échantillons analysés en protéomique. Cet échantillon de référence est composé d'un extrait protéique de levure dans lequel des quantités connues d'un mélange de 48 protéines humaines (mélange équimolaire UPS1 commercial) ont été rajoutés (de 250 attomoles à 25 femtomoles dans 1µg de matrice).

Après avoir optimisé les paramètres d'acquisition, nous nous sommes focalisés sur le traitement des données, avons évalué la nécessité de réaliser une étape de normalisation, ainsi que les outils statistiques pour valider les variations d'abondance des peptides.

La comparaison de la méthode Middle band DIA and MS1 DDA montre une meilleure sensibilité de détection des peptides d'UPS1 et globalement une meilleure robustesse. Néanmoins, il est à noter que l'approche DIA nécessite au préalable la génération de librairies de spectres.

En résumé, ces résultats préliminaires montrent que le mode DIA est très prometteur pour les aspects de quantification en analyse protéomique. Il permet d'accéder à des protéines minoritaires ainsi qu'à une quantification plus reproductible et plus précise.

CONCLUSION

L'objectif de ma thèse était de développer et d'optimiser de nouvelles approches méthodologiques et analytiques afin d'améliorer les performances de l'analyse protéomique basée sur la spectrométrie de masse.

L'objectif à long terme est le développement de méthodes qui permettraient d'accéder à une caractérisation complète de toutes les protéines présentes dans un protéome. Cet objectif, encore hors d'atteinte à l'heure actuelle, permettra dans le futur d'ouvrir la voie vers des comparaisons détaillées de protéomes avec une quantification précise et robuste pour chaque protéine présente.

Dans la partie 1, qui porte sur la caractérisation des protéines en analyse protéomique à haut débit, nous avons montré qu'elle pouvait être améliorée de façon générale grâce à la détermination quasi systématique des séquences N-terminales (N-terminomics). Ceci peut se faire grâce la méthode de marquage N-terminale par le réactif TMPP utilisé sous forme de mélange lourd/léger, ce qui permet une exploitation des données MS/MS automatisée à haut débit. La nouvelle méthodologie développée a été appliquée avec succès à l'étude du N-terminome de la bactérie *H. arsenicoxydans.* Nous avons ainsi confirmé 90 positions N-terminales, corrigé 13 erreurs de prédiction et mis en évidence des clivages protéolytiques nouveaux.
La stratégie N-TOP, optimisée pour être compatible avec un marquage chimique dans le gel, a également permis de caractériser les acides aminés essentiels du motif Pexel, de protéines chimériques de *Plasmodium falciparum*, pour leur exportation vers la cellule hôte.

Dans la partie 2, qui porte sur des améliorations en analyse protéomique quantitative, nous avons obtenu des données de quantification qui ont permis de répondre, en partie, à plusieurs problématiques biologiques. Ces données de quantification ont pu être obtenues grâce à des développements de la méthode ciblée par SRM, et de la méthode basée sur le comptage du nombre de spectres acquis pour les peptides d'identification d'une protéine.
Dans le cadre de l'étude de l'impact des modifications post-traductionnelles, portées par l'histone H2B, sur la maturation du RNA et l'export nucléaire, une série d'améliorations de la méthode de comptage des spectres a permis de mesurer les variations d'abondance de protéines présentes en faible quantité et dans une matrice complexe.
Nous avons développé une méthode semi-quantitative pour la détermination des glycations d'anticorps monoclonaux thérapeutiques.
L'approche ciblée par SRM a permis de détecter la présence de protéine prion humaine dans des produits pharmaceutiques d'hormones de fertilité d'origine urinaire.
Une seconde méthode de quantification par SRM, mettant en jeu une digestion chimique dans le gel à l'aide de CNBr, va permettre de vérifier l'hypothèse de l'accumulation préférentielle de l'un des 2 variants polymorphiques du prion dans des biopsies de cerveaux humains affectés par cette maladie.
Nous avons mis en place une quantification de protéines purifiée par LC-MS, pour doser une forme mineure de l'hémoglobine HbA_2. Cette méthode n'implique pas de cliver les protéines en peptides, étape de digestion qui non seulement induit des biais mais rajoute une étape dans le processus analytique quantitatif.
Et finalement, nous avons également exploré le potentiel d'un nouveau mode d'acquisition (DIA) qui est très prometteur pour les aspects de quantification en analyse protéomique.

Les méthodologies que nous avons mises au point sont compatibles avec des analyses protéomiques à haut débit, ce qui est essentiel pour les applications en biologie. En effet, nous les avons appliquées avec succès dans des collaborations avec des laboratoires académiques de renommée internationale (Institut Jacques Monot, Université d'Heidelberg, University Hospitals of Cleveland) et industriels (Merck-Serono).
Nos travaux ont donc permis d'améliorer d'une part, la caractérisation des protéines, et d'autre part, d'obtenir de meilleures précisions en analyse protéomique quantitative.

Notre contribution reste bien sûr modeste par rapport aux attentes des biologistes. En effet, ceux-ci souhaiteraient dans l'avenir avoir une caractérisation complète de toutes les protéines d'un protéome (y compris les PTMs) ainsi que leur quantification. De plus, ces analyses devraient pouvoir se faire à haut débit afin de suivre toutes les fluctuations qui affectent les protéomes. Nous pensons que cet objectif pourra être atteint dans un futur plus ou moins lointain, comme cela a été le cas pour le séquençage des génomes. En effet, le séquençage du génome humain a pris de nombreuses années, alors qu'aujourd'hui cela peut être fait en moins d'un jour.

C'est en persévérant dans la mise au point de nouvelles méthodologies et de nouvelles approches instrumentales, comme cela a été le cas pour le séquençage des génomes, que l'analyse protéomique pourra apporter aux biologistes l'énorme masse d'information nécessaire à une meilleure compréhension du rôle des protéines dans les mécanismes biologiques.

General introduction

General introduction

The word proteome has been introduced by Marc Wilkins in 1994, to define «all the proteins expressed by a genome, cell or tissue» (2, 3). Proteomic analysis aims at a global characterization of a proteome, which is an enormous task. From an analytic point of view the proteomic analysis, widely based on mass spectrometry, is still far from being a mature analytical method. Proteomics still aim at developing methods capable of achieving a confident, high-throughput and inter-laboratory reproducible, full characterization and quantification of each protein of a proteome.

When such methods will be available for the scientific community, they will make possible the characterization of each protein, including all isoforms. But we are far from the full description of a proteome, even if it is claimed that proteomic analysis has covered approximately 50% of the proteome theoretically expected in a human cell line (4). Most likely the expression of our genome is much more complex to decipher repect what is believed today, and the number of proteins "theoretically expressed " by a human cell is probably much higher.

An exhaustive analysis of all proteins from a proteome will certainly permit major progresses in the understanding of structure/function relationship. Results have been obtained in this direction, some of them considered unthinkable, just few years ago. Nevertheless, we are far from being able to compare one proteome to another, with a sharp quantification of each protein. There is obviously a huge gap between the needs and the dreams of biologists and what with today's proteomics it can be achieved.

Very often, the identification of a protein is obtained with only a few digestion peptides analyzed by nano liquid chromatography coupled with mass spectrometry (nanoLC-MS-MS), resulting in poor sequence coverage. This does not allow, for example, discriminating between proteins produced by homologues genes and does not give any information about possible PTMs. N- and C-terminal amino acids are seldom identified and in fact most PTMs are ignored. Virtually all proteins expressed are subjected to modifications (5) to modulate, and in some cases, radically change the biologic function. More than 300 PTMs (6) have been described up to now, among them, the proteolytic cleavage is crucial in the context of protein processing and activation. Since proteolytic cleavages cannot be predicted from the DNA, this type of information can only been determined by direct protein analysis. About 900 human proteases (7) have been catalogued up to now. These proteases are particularly important in the cell homeostasis and if deregulated are often related to pathologic state (8). The proteolytic cleavage is an example of a common modification which affects many proteins and which is not often described in detail in a proteomic study.

Clearly, today the description of a proteome is still very inaccurate despite enormous improvements during the last years. There is a real need for new methodologies and analytical approaches to improve the sequence coverage of all identified proteins including the PTMs and to quantify them on a high-throughput base. This was the goal of my Ph.D., being able, in a far future, of replacing "the identification" by "the full characterization" of all proteins from a proteome. For example, I tried to bring a better knowledge of N-termini position and hence of the cleavage position during processing and activation. I have also tried to improve methods for the quantification of proteins.

The first part of my manuscript will address the analytical efforts produced during my Ph.D. in order to develop and optimize analytical methods to characterize the N-terminal position in the context of proteolytic events.

1. Characterization of the PEXEL export mechanism in *P. falciparum* infected erythrocytes.
   In the first chapter, I will present the study of proteolytic cleavages of the exported proteins in *P. falciparum*, parasite responsible for the malaria, performed using the strategy of chemical labeling of the methodology "Ingel N-TOP" (In gel N-termini Oriented Proteomics).
2. Development of a stable isotope labeling method to validate N-terminal TMPP derivatized peptides.
   In the second chapter, I will describe an important improvement of the N-TOP strategy which opens the door to an automated MS/MS data processing allowing high-throughput N-terminal determination. This strategy has been applied to perform the deep characterization of *H. arsenicoxydans* proteome, with a special focus on its N-terminome.

In the second part, I will address a crucial aspect of the mass spectrometry based proteomics that needs to be improved: the quantitative information. This is essential to get from proteomic analysis its full potential as a major tool in biology. The human proteome spans across ten orders of magnitude, making virtually impossible to obtain the complete quantitative information from a single experiment. A panel of analytical approaches is available to the scientific community, (9, 10) each of one presenting peculiar advantages and consequential withdrawals. Current quantitative proteomics methods must make significant progress to gain in sensitivity, dynamic, precision and robustness, in order to be able to perform the complete quantification of a biologic sample.

I devoted a part of my Ph.D. to the optimization of five different protein quantification approaches, four at the peptide level and one at the protein level.

1. Relative quantification of affinity-purified proteins complex.
   In this chapter, I will present the optimization performed on a series of instrumental parameters and the sample preparation that allowed gaining in sensitivity in the differential quantification of proteins complex affinity-purified. The biologic context was the study of the yeast histone H2B ubiquitylation effect on the RNA nuclear exportation.

2. Targeted Selected Reaction Monitoring (SRM) based Prion protein quantification.
   Prion diseases (PDs), also known as transmissible spongiform encephalopathies (TSEs), are fatal neurodegenerative diseases caused by conversion of the prion protein PrP^{c} in a so called scrapie form, PrP^{sc} an insoluble isoform. We have developed quantification of prion protein in two different biological contexts:
   - Urine derived injectable fertility products, represent an unusual source of prion protein. I have improved the LC-SRM method, already used in our laboratory in a first study, where an absolute quantification of the prion protein was performed in pharmaceutical products.
   - Biopsies of human brain TSEs affected: I have developed an in gel chemical digestion using cyanogen bromide (CNBr) to relatively quantify prion polymorphic variants which differ only from one amino acid.

3. Monoclonal IgG glycation batch to batch quantification.
   Glycation is a not enzymatic addition of a reducing sugar to the residue of basic amino acids that can be followed by a rearrangement of sugar producing Advanced Glycation Endproducts (AGE). This reaction represents an important source of variability in therapeutic recombinant protein. The analysis of this chemically heterogeneous family of PTM is still challenging. Therefore, I have developed a mass spectrometry based label free method capable of characterizing this modification and quantifying it in different production batches.

4. Intact protein label free quantification: future developments.
   This chapter will present the results of a method development for a rapid and simple method to quantify the level of HbA_2 in blood samples. The accurate quantitative determination of hemoglobin A2 ($\alpha 2 \delta 2$) is a key marker for the correct diagnosis of the beta-thalassemia. We developed a methodology based on quantitation of intact (non digested) proteins by LC-MS which is usually considered as non-compatible with accurate quantification.

5. Optimization of a Data Independent Acquisition (DIA) method and MS1 comparison.
   In the frame of the continuous effort for better protein quantification, I have performed a comparative study (on a Q-TOF Impact HD, Bruker) to evaluate the benefit of DIA acquisition over DDA-HrXIC (Data-Dependent Analysis with MS/MS for ID coupled to the integration of highly specific MS traces for quantitation purposes) for label free quantification in complex protein mixture. The optimization performed on this technique allowed increasing the sensitivity of the technique and the confidence of the results.

Part I
N-Termini characterization

Introduction
The N-Termini characterization

Chapter I
Characterization of the PEXEL export mechanism in *P. falciparum* red blood

Chapter II
Development of a stable isotope labeling method to validate N-terminal TMPP derivatized peptide

Part I – Introduction - The N-Termini characterization

State of the art in the N-Terminomics

The experimental determination of protein N-termini is an analytical challenge, but is of major importance for different reasons:

- The protein sequence databases are the results of an "in silico" translation of genome sequence which are experimentally obtained. The bio-informatics determination of the start codon is not an error free process, expecially in prokaryote organism with genome rich in GC. An error at this level could have a dramatic effect on the subsequent biologic study. This topic will be detailed in the chapter II.
- More of 1% of the human proteome is represented by proteases (approximately 500 proteins). Then, it can be expected that a proteome could be continuously modified by these enzymes. It has been widely shown that proteases are not only responsible for the proteins turnover and degradation, but also for the regulation and modification of the biologic functions through the alteration of the primary protein sequences.
  Recently the N-terminomic has been proposed as a specific pathology biomarker signature (11).

In 2002 the family of "omics sciences" was enlarged when Lopez-Otin and Overall (12) introduced the terms "degradomics". Degradomics describes the global characterization of proteases substrate, of inhibitors and how the proteolytic activity influences the cells, tissues and organs. Proteases, thanks to the capabilities of hydrolyzing the peptidic linkages, can modify the destiny of a protein, realizing a complete degradative proteolysis associated with the protein catabolism, or performing a limited proteolysis.

Two different effects (13) can result from a limited proteolysis. In a first scenario a sequential maturation is performed, where a peptide is excised from a protein. This is the case of the Chemochin protein family, often regulated by N-terminus truncation (14).

A second scenario is illustrated by the case of the HARP protein. In this case, after the proteolytic cleavage by the metalloproteinase MMP2, two proteins are generated with opposite biologic functions in the cellular cycle control (15).

In the recent years, many different proteomics approaches to characterize the "degradome" have been presented. In this chapter, I will briefly present the most common techniques for the precise determination of the N-terminus position of a protein (16). Two orthogonal approaches are presented in the literature and will be addressed here: positive selection based on a direct targeting of the N-terminal peptide and negative selection based on the depletion of internal peptides.

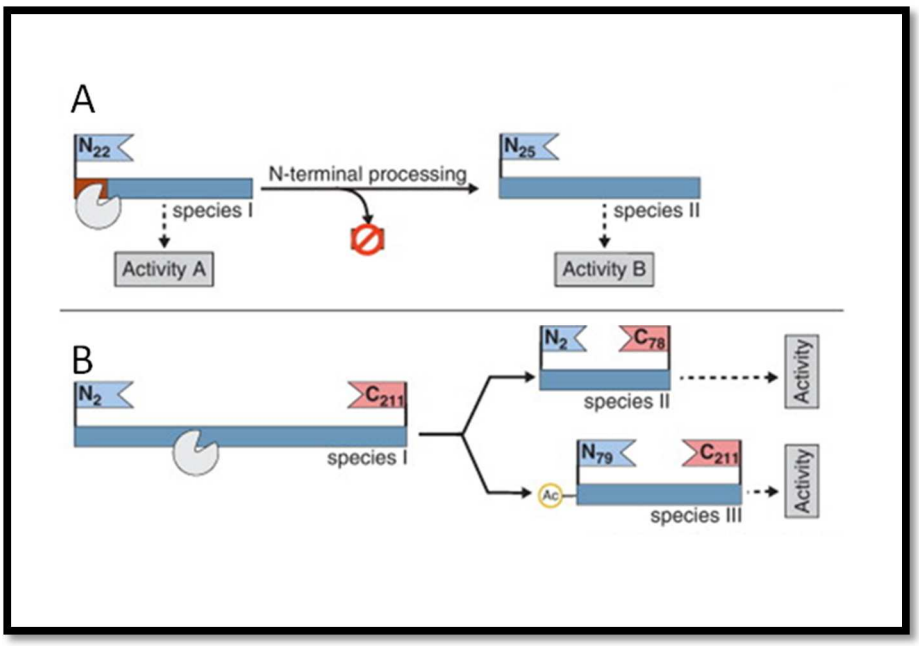


Figure 3: adapted from (13). The picture A describes two effects of proteolytic event: case A the biologic function of the protein is modified by the cleavage, case B two different proteins are generated by the protease.

Positive selection of N-terminal peptide

Enrichment of N-terminal peptides based on the streptavidin biotin interaction

Typically the ε-amines (amines of the side chain of lysine) are blocked using a guanidination reaction and the α-amines (N-terminal) are derivatized with biotin. This technique takes advantage of a milestone discovery in molecular biology: the Streptavidin, a bacterial protein isolated from the actinobacterium *Streptomyces avidinii* capable of binding biotin with one of the strongest, non-covalent, biological interaction described in literature (17). Biotin is a small water soluble enzymatic cofactor that does not impair the structure and biologic function of the protein. The addition of this tag to the α-amines can be performed with activated biotin or enzymatically mediated. The second approach offers a higher specificity of binding. Once the sample is submitted to tryptic proteolysis the N-terminal peptides are positively selected using commercial available cartridges coated with streptavidin, the N-terminal peptides are then released with different techniques depending on what has been choose to interposed the linkage biotin peptide:

- with dithiotreitol in the case of N-Hydroxysuccinimide-biotin linkage (18).
- with the protease TEV in the case of a viral cleavage site cassette (19).

The limitation of this approach lies in the impossibilities of derivatizing the blocked α-amines. Depending on the organism choose for the study this can be really limiting, as in eukaryotic cell, up to 80-85% of the α-amines are acetylated (20) and consequently escape this enrichment technique. Side reaction of the guanidination on the aminoacidic residues of serine, threonin and histidine are responsible for the contamination of N-terminal enriched peptide with internal one.

Edman sequencing reactive based derivatization

The "N-terminalomics" by chemical labeling of the α-amine of proteins" (N-CLAP) (21, 22) is a two steps enrichment technique, that uses the Edman chemistry (23) to solve the issue of the side reaction generated by the guanidination. The phenyl isothiocyanate (PITC), is used to derivative both the α and ε amines. The addition of acid Trifluoroacetic induces the cyclization and the release of the N-terminal amino acids derivatized with a phenyl thiocyanate (PTC). This reaction determines the formation of a new N-terminal α amines that is derivatized with biotin. The proteins are then digested into peptides and the N-terminal derivatized peptides enriched using Streptavidin and submitted to LC-MS/MS. Like in the method described above, the not free α amines escape to the initial derivatization.

Proteomics identification of cleavage sites (PICS)

This strategy has been developed to characterize the specificity of a protease, and it has the peculiarity of being peptide based and not protein based as the two precedent described approaches (24). Starting from an entire proteome, a peptide library is produced after incubation with a trypsin or endoproteinase GluC, then after reduction and alkylation, both ε and α amines are blocked thought a methylation reaction. The peptides are incubated with the protease selected to be studied and this results in the formation of new α amines that are subjected to biotin derivatization, enriched by affinities capture and submitted to LC-MS/MS.

This technique, since it is based on a peptide library, presents the advantage of not having the protease specificity and efficiency being affected by the 2nd and 3rd proteins structure conformation.

Negative selection N-terminal peptide

Combined fraction diagonal chromatography (COFRADIC)

In this technique, as in all the negative N-terminal selection, all the amine groups are blocked and then the proteins are digested into peptides. Specifically in the COFRADIC approach (25), all the amine groups are acetylated. After the digestion, a first run of reverse phase liquid chromatography is used to pre-fractionate the peptides and a series of fractions are collected across the chromatographic gradient. The new α-amines, generated by the trypsin are blocked with 2,4,6-trinitrobenzensulfonic acid (TNBS). This modification increases the retention time in the reverse phase thanks to an increased hydrophobicity. Each fraction is then submitted to a second reverse phase LC. At this point, all the internal peptides are eluted with a shifted retention time, due to the TNBS derivatization, when the N-terminal peptides keep their original retention time. This allows a fraction collection of the peaks containing the N-terminal peptides (acetylated or blocked) which can be analyzed by LC-MS/MS for identification.

COFRADIC represents the first choice in case of high level of protein acetylation, due to the intrinsic advantage of performing a negative selection. On the other hand, one experiment can result in hundreds of chromatographic runs (26). This high number of manipulation steps increases the percentage of sample loss and requires a considerable amount of starting material. The original method described by J. Vandekerckhove (25), presents the limitation of not being capable of discriminate between endogenous N-termini acetylation and the one induced artificially during the sample preparation. Six years after, the same group, (27) solved this limitation performing the acetylation with isotopically stable labelled reagent. Another interesting improvement, is the introducing of an ion exchange purification step, performed with

a cartridge after the tryptic digestion (28). Since at pH3 the N-terminal blocked peptides present no charge, they are not retained, when peptides with an arginine missed cleavage are retained. This approach increase the purity of the peptide submitted to the further two COFRADIC separation. Recently the TMPP-Ac-OSu (N-Succinimidyloxycarbonylmethyl)tris(2,4,6 trimethoxyphenyl)phosphonium bromide (from now on referred as TMPP) derivatization was integrated in the COFRADIC workflow, the first step is the TMPP derivatization of the N-termini followed by protection of the free amino group through an acetylation step and classic combined fraction diagonal chromatography with TNBS based negative selection (29).

Biotinylation of internal peptide as negative selection

In this strategy, all the primary free amines are blocked with an acetylation, and the trypsin digestion generates new free α-amine that are then blocked with biotin. In this way the internal peptides can be removed from the sample using Streptavidin (30). This approach is fairly simple and easily to reproduce but, as COFRADIC, does not allow differential study and it does not give quantitative information.

iTRAQ-TAILS based termini amine isotope labelling

iTRAQ is an isobaric labelling that has been developed to tag all the free α-amines at the peptide level, the advantage of this technique is that those commercially available kit are designed to allow an high degree of multiplexing thanks to the different reporter ions generated during the fragmentations. Prudova (24) has adapted this quantitative approach to the N-termini protein characterization, performing the labelling at the protein level and submitting each sample to separate tryptic digestion. This allows, once the sample has been mixed, to negatively select the free N-termini generated by the trypsin using polyaldehyde dendritic polymer. In such a way, the homologue peptides labelled with different iTRAQ will perfectly coelute in the same chromatographic peak and the different reporter ions will allow quantifying the proteolytic events occurred in the different samples. This approach allows a high degree of multiplexing and is one of the few offering quantitative information.

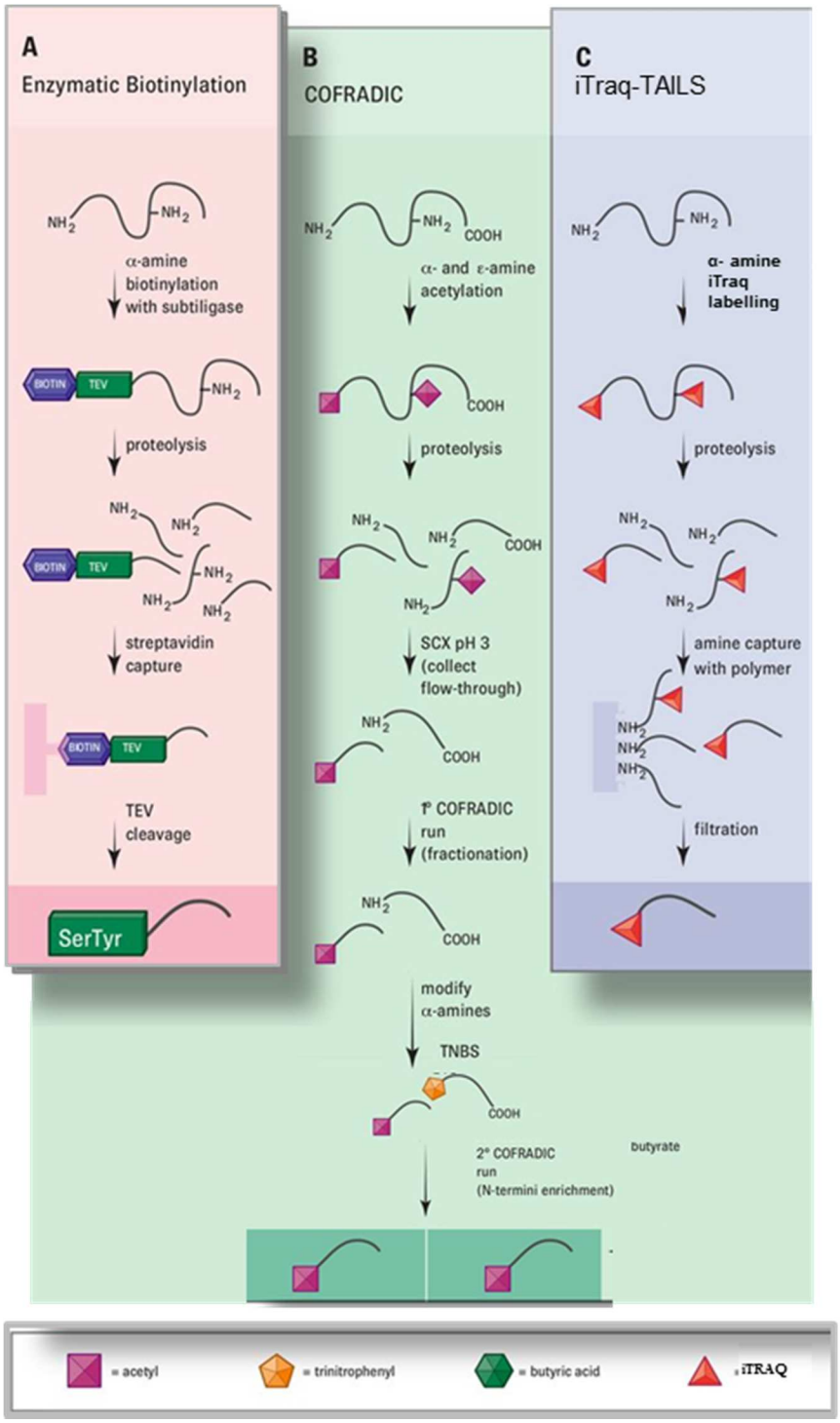


Figure 4: adapted from (26). The figure schematizes the workflow of : A) enzymatic biotinylation, B) COFRADID C) iTraq-TAILS

TMPP-Ac-OSu based N-termini characterization

Introduction

The use of phosphonium based compound as derivation agent in mass spectrometry was described the first time by Watson and coworkers in 1991 (31). Thanks to the N-terminal or C-terminal addition, at the peptide level, of a molecule carrying a positive charge, Watson achieved an increased efficiency of fast atom bombardment ionization mode.

In 1997, again the group of Watson introduced the derivatization of the N-terminal amine peptides with TMPP-Ac-OSu (N-Succinimidyloxycarbonylmethyl)tris(2,4,6 trimethoxyphenyl)phosphonium bromide (TMPP) in MALDI Post Source Decay (PSD). The TMPP reagent was selected since the fixed N-terminal charge allowed generating less complex spectra, easier to manually interpret.

In 2009 our laboratory published the first study that described the derivatization of proteins complex mixture using TMPP and this strategy was called N-TOP. Gallien *et al.* (1) took advantage of the pK difference between the α– and ε– amines for the TMPP derivatization, to specifically target the N-terminal protein amines. The detection, in an MS/MS peptide spectrum, of a TMPP group on its N-terminal amino acid allows an unambiguous identification of the N-termini of the protein. This technique was applied successfully to the correction of hundreds of the-start codons in the annotation database of *M. Smegmatis.*

The TMPP derivatization and its advantage

The N-TOP strategy is based on the derivatization of the N-termini amines at the protein level. Performing the reaction at pH 8.2 allows minimizing the side reaction of the TMPP on the hydroxyls of the tyrosins and on the amine side chain of the lysins, thanks to more acidic pK of the N-termini amines. In Figure 5 is schematized the TMPP derivatization.

Once the proteins are derivatized, the sample is submitted to SDS-PAGE for pre-fractionation which allows also removing the excess of TMPP. The gel bands are then extruded and submitted to classic proteomic workflow. The N-TOP strategy allows detecting in the same LC-MS/MS analysis internal peptides and N-terminal derivatized peptides.

The derivatization performed with TMPP-Ac-OSu presents two main advantages due to its peculiar chemical characteristics (Figure 6).

- The positive charge carried by the phosphonium allows an increased ionization efficiency of the derivatized peptide.
- The hydrophobic nature of the TMPP derivatized peptide results in a late gradient elution with respect to the internal peptides. In such a way, the high response factor in ESI of the TMPP derivatized peptides does not suppress the ionization of the internal one.

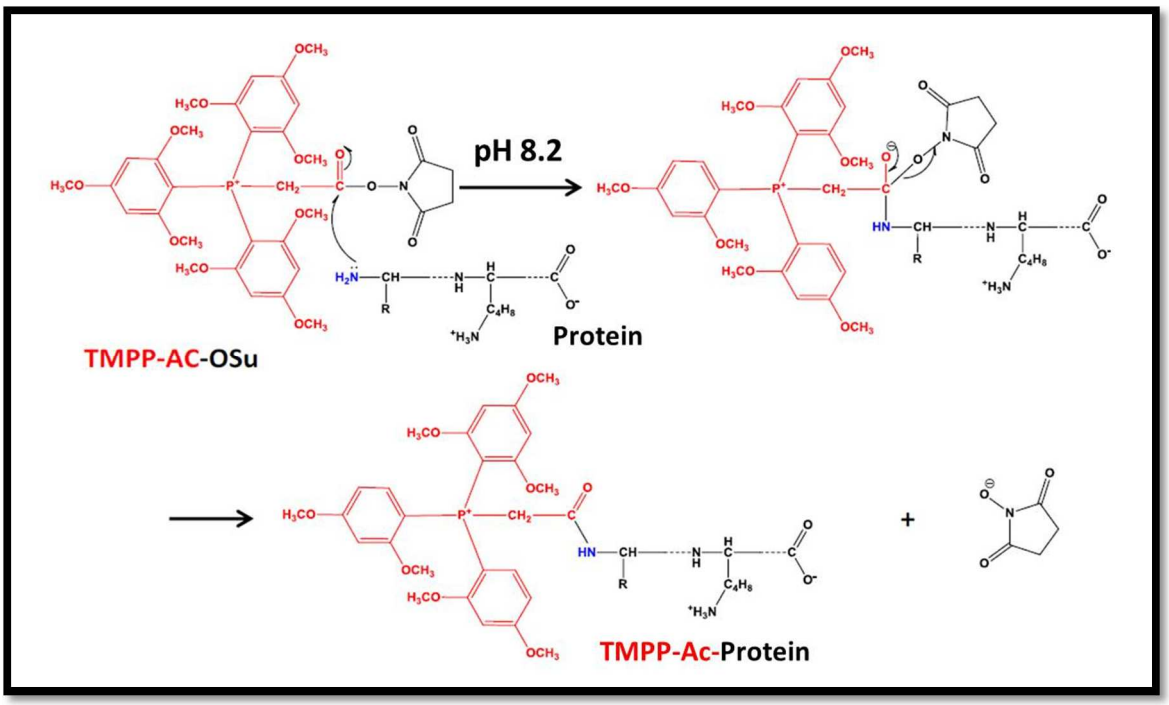


Figure 5: Schematic representation of the TMPP reaction performed at pH 8.2 allowing to specifically targeting the N-termini amines. Adapted from Daniel Ayoub Ph.D. thesis (32).

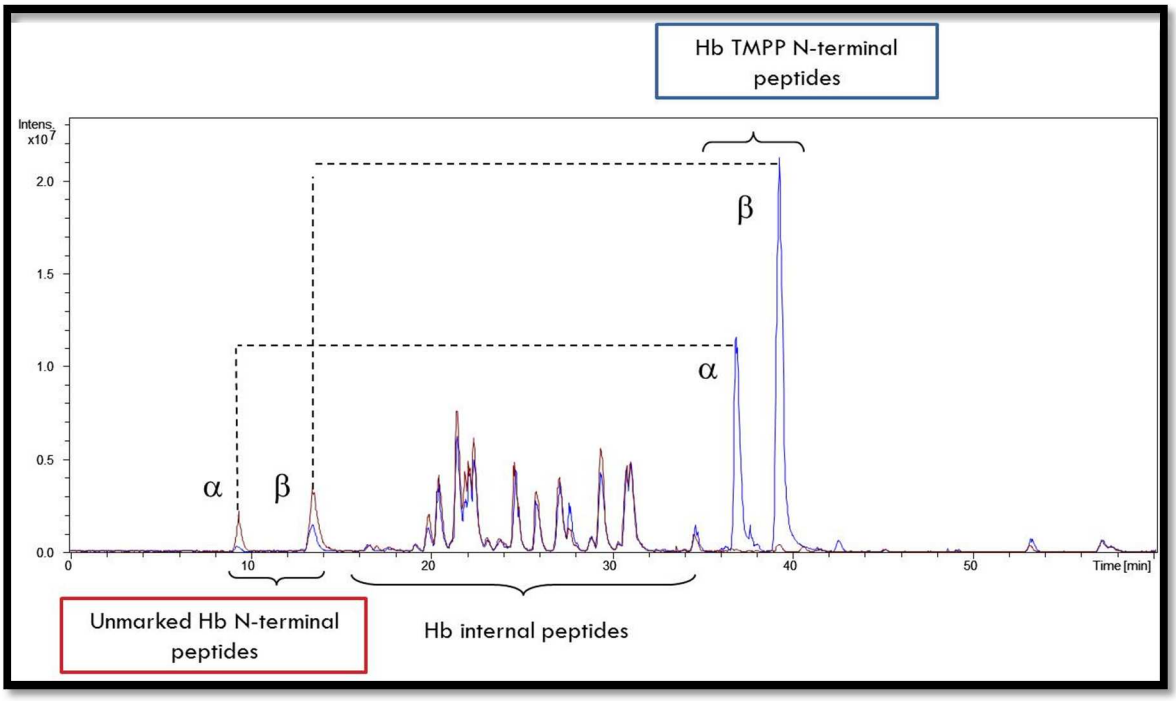


Figure 6: Overlay of the BPC obtained injecting a tryptic digest of hemoglobin TMPP derivatized in blue and not in red. On the left the peaks of the two N-terminal unmarked peptides. In the middle the internal peptides and on the late elution area the two N-terminal peptides derivatized.

Conclusion

In this chapter we have presented the most common proteins N-termini determination approaches described in literature. This has not the presumption of being an exhaustive review of all possible approachesand, many variants of the main school of thought were not presented.

Generally the positive enrichment based techniques do not allow characterizing the endogenously blocked N-termini.

The negative selection instead is based on a depletion or segregation of the internal peptides and allows characterizing in the meantime the N-terminus acetylation and pyro glutamic acid formation. The Overall group used this approach to characterize the N-terminome of the human erythrocytes (33).

Both positive and negative enrichment based strategies present complex workflows with multiple sample manipulations that can not be easily integrated in a classic high throughputs proteomics. They also often require important instrumental and bioinformatics resources.

The TMPP based strategy is the only approach that can provide a comprehensive N-terminomic characterization, of free and blocked α-amine and in the same time provides classic proteomic identification of the internal peptide in the same LC-MS/MS analysis.

In chapter I, I will present an improvement of the TMPP derivatization made “in gel” (Ingel N-TOP), first developed by Ayoub during his Ph.D. (32). I have optimized this approach for the characterization of the PEXEL export mechanism in *P. falciparum* red blood infected cell.

In chapter II, I will present the dN-TOP approach, an evolution of the N-TOP strategy that allowed automating the validation of the N-termini position.

Part I - Chapter I –

Characterization of the PEXEL export mechanism in *P. falciparum* infected erythrocytes

Part I - Chapter I - Characterization of the PEXEL export mechanism in P. falciparum infected erythrocytes

Introduction

P. falciparum is a parasite responsible for one of the most severe malaria diseases. In 2010 it caused 216 million new cases of infection and 655 000 deaths (34).

P. falciparum has a complex life cycle based on three main host steps invasions (35), during which sexual and asexual stages alternate. The primary host of the parasite is the *Anopheles* mosquitoes, during the *Anopheles* bites, the parasites are transmitted to the secondary host, the Human. The second step of the life cycle *of P. falciparum,* takes place in the Human liver, where each parasite infects one hepatocyte and multiplies inside it. One week after the infection, one successful sporozoite can reproduce up to 30 000 merozoites. The third step of the life cycle is the Human red blood cell invasion. During this phase the parasite deeply alters the structure and the physiology of host cell in order to survive in it. These alterations are responsible for severe symptoms of the malaria disease.

The life cycle of the parasite restarts when human infected red blood cells are aspired and transferred to new secondary host during mosquitoes bites.

This project was carried out in collaboration with Prof. Lanzer and coworkers of the Heidelberg University, Germany.

Biologic context

The invasion of the red blood cells is a turning point of the malaria disease. *P. falciparum* penetrates the red blood cell, generating an invagination of the erythrocyte lipid bilayer (36, 37). This results in the formation of a protective barrier, called *P. falciparum* Membrane Vacuole (PMV), between the parasite and the red blood cytoplasm, as displayed in Figure 7.

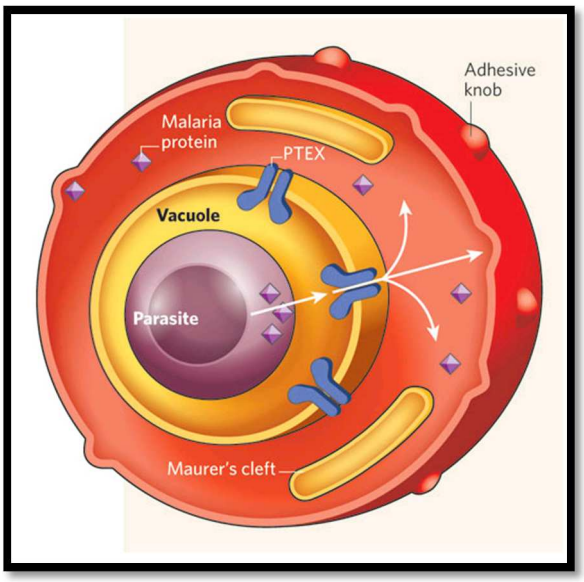


Figure 7: schematic representation of a red blood cell infected by *P. falciparum*.

The parasite, in order to ensure its homeostasis, exports many proteins outsides its membrane. Some are exported in the vacuole and others are exported in the cytoplasm of the erythrocyte. Due to the need of differentiate the destiny of the exported protein (the PMV, or the erythrocyte cytoplasm), *P. falciparum* has developed a complex exportation system based on a Host Cell Targeting signal (HCT).

The HCT is like a "code bar sequence" added to every protein that need to be exported, coding the destination. The group of Marti (38) identified the HCT signal as a pentameric amino acidic sequence RxLxE/Q/D and called it PEXEL motif (Plasmodium Export Element). The bioinformatics analysis of the *P. falciparum* genome allowed predicting the exportation of more than 300 proteins, thanks to the detection of the PEXEL sequence.

Site specific mutation experiments, validated the importance in the PEXEL motif of the residue R, L and E/D/Q (38-40), but the minimal functional size of the motif and the role of the N-terminus acetylation (41) during the cytoplasmic exportation remains unclear.

Transport vesicles system and transmembrane proteins complexes, recognizing the PEXEL sequence can mediate the export of the *P. falciparum* proteins, the complete understanding of this mechanism represents the first step to the development of an efficacy therapy against this deadly disease (42).

In Figure 8 is schematized the structure of a *P. falciparum*'s protein carrying the PEXEL motif. The conserved PEXEL sequence, is placed after an N-terminus hydrophobic peptide stretch, necessary to the protein to be inserted in the Rough Endoplasmic Reticulum (RER) (42). The Golgi vesicles apparatus, in concert with the RER, is responsible for the transfer of the protein across the lipid bilayer of the parasite in the parasitophorus vacuole.

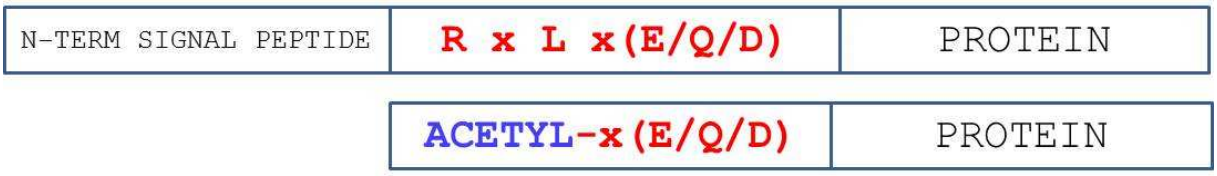


Figure 8: schematic representation of a PEXEL protein, respectively on the top before the exportation, below after the erythrocyte cytoplasm exportation.

On the vacuole lipid bilayer, a translocon proteins complex recognizes the PEXEL motif and cleaves the protein after the third PEXEL amino acid. An acetyl transferase, associated to the translocon, is responsible for the blocking of the new protein N-terminus before the flip of the protein in the erythrocyte.

The analytical task

The goal of this project was:

- To understand if the fourth amino acid position in the PEXEL motif, considered as not influent in export process, might instead has a role in the process.
- To validate the hypothesis that other amino aside E/Q/D on the last PEXEL position are recognized by the translocon.

To investigate such hypothesis, the group of Prof. Lanzer produced a series of erythrocyte cell lines, infected by *P. falciparum*, carrying chimeric constructs genomically GFP tagged. Each

construct, see Figure 9, is composed of: the N-terminus hydrophobic sequence, the PEXEL motif with one amino acid mutated and an aminoacidic sequence fused with the Green Fluorescent Protein (GFP).

The fluorescent microscopy, allowed confirming the exportation of the construct, but mass spectrometry was required in order to confirm the exact N-terminus position after the cleavage and the possible N-terminus acetylation performed during the export.

Figure 9: schematic representation of the WT construct.

The constructs were purified using anti-GFP antibody and eluted from the affinity column using Radio-Immunoprecipitation Assay buffer (RIPA), containing three detergents: IGEPAL®, sodium deoxycholate and SDS. This was done by the group of Prof Lanzer.

The "Ingel N-TOP" strategy, first developed by Ayoub during his Ph.D. (32), represents the optimal analytical strategy for this task, since the TMPP derivatization performed after 1D SDS PAGE allows:

- Removing the detergents, not compatible with LC-MS/MS.
- Increasing the sensitivity due to the sample pre-fractionation.

The "Ingel N-TOP" technique, like the original in solution derivatization called "N-TOP"(1) (described in the part I introduction) is fully compatible with blocked (acetylated) protein N-termini and allows characterizing the protein N-termini and the internal peptides in the same LC-MS/MS run.

Figure 10 describes the three possible destinations of the construct in terms of exportation and the results of our "Ingel N-TOP" study:

- The mutation on the PEXEL is performed on an essential amino acid, therefore the construct is not cleaved and not exported. The protein N-terminus (N-start 1) will be detected as TMPP derivatized.
- The mutation on the PEXEL is performed on an amino acid not essential for the recognition of the cleavage sequence, but essential for the acetylation of the new N-terminus. The protein is cleaved but not exported. The new protein N-terminus, N-start 51, is detected TMPP derivatized.
- The mutation on the PEXEL is performed on an amino acid not essential for the processing and not essential fro the acetylation of the new N-terminus. The protein is cleaved and exported to the erythrocyte cytoplasm. The protein N-terminus, N-start 51, is detected acetylated.

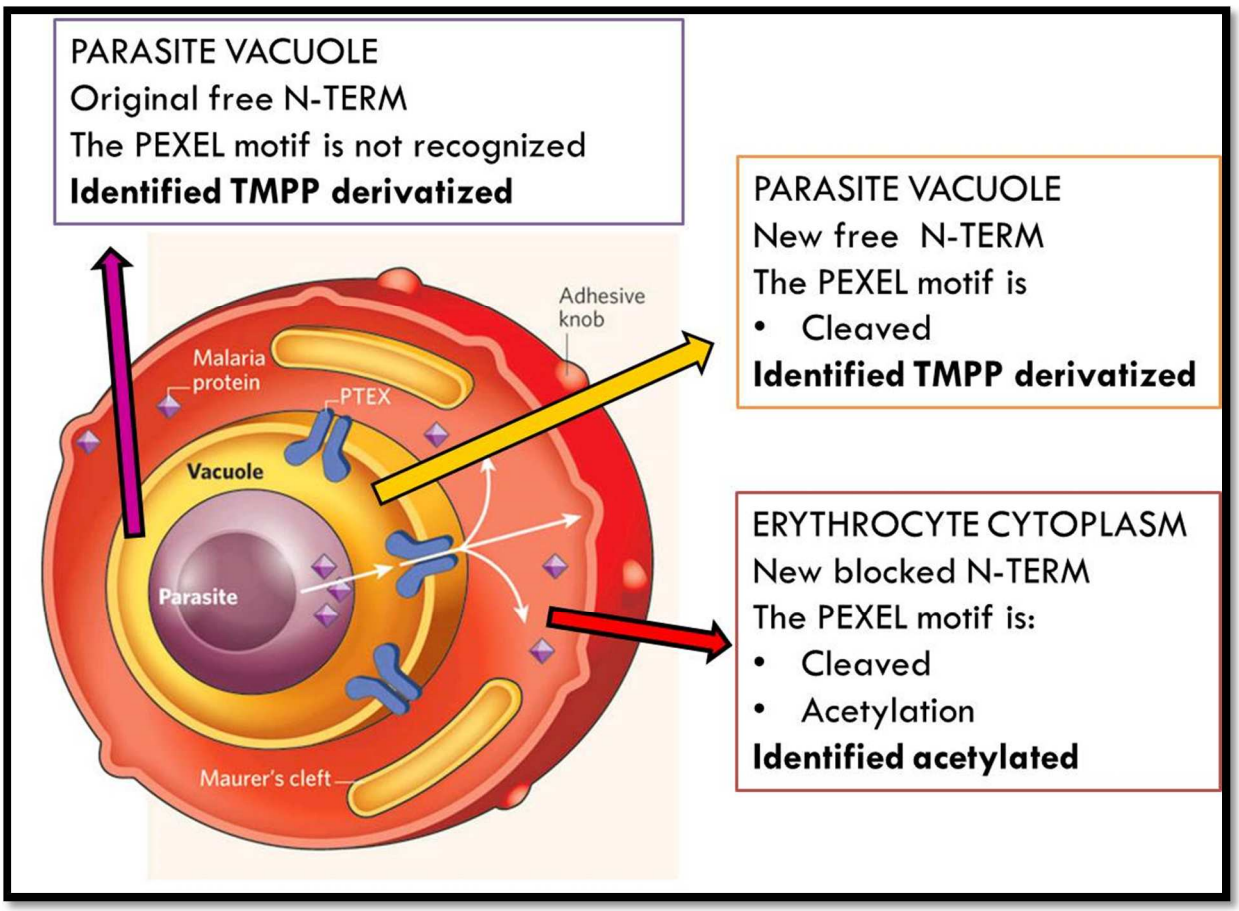


Figure 10: the three possible destinations of the construct produced.

Optimization of the Ingel N-TOP

The first step of my work was to optimize the "Ingel N-TOP" to the characterization of the chimeric constructs, so that it could be used on really tiny amounts. The original method required the optimizations of the amount of TMMP used to perform the derivatization.

The "Ingel N-TOP" technique, like the N-TOP, is based on a strictly controlled pH 8.2 reaction condition. This pH represents optimal compromise between the derivatization yield and the lysine and tyrosine side reactions. The selectivity is obtained thanks to the more acid pK of the N-termini amines. A more basic pH would result in an increased N-terminus derivatization, but also in an exponentially increased side reaction.

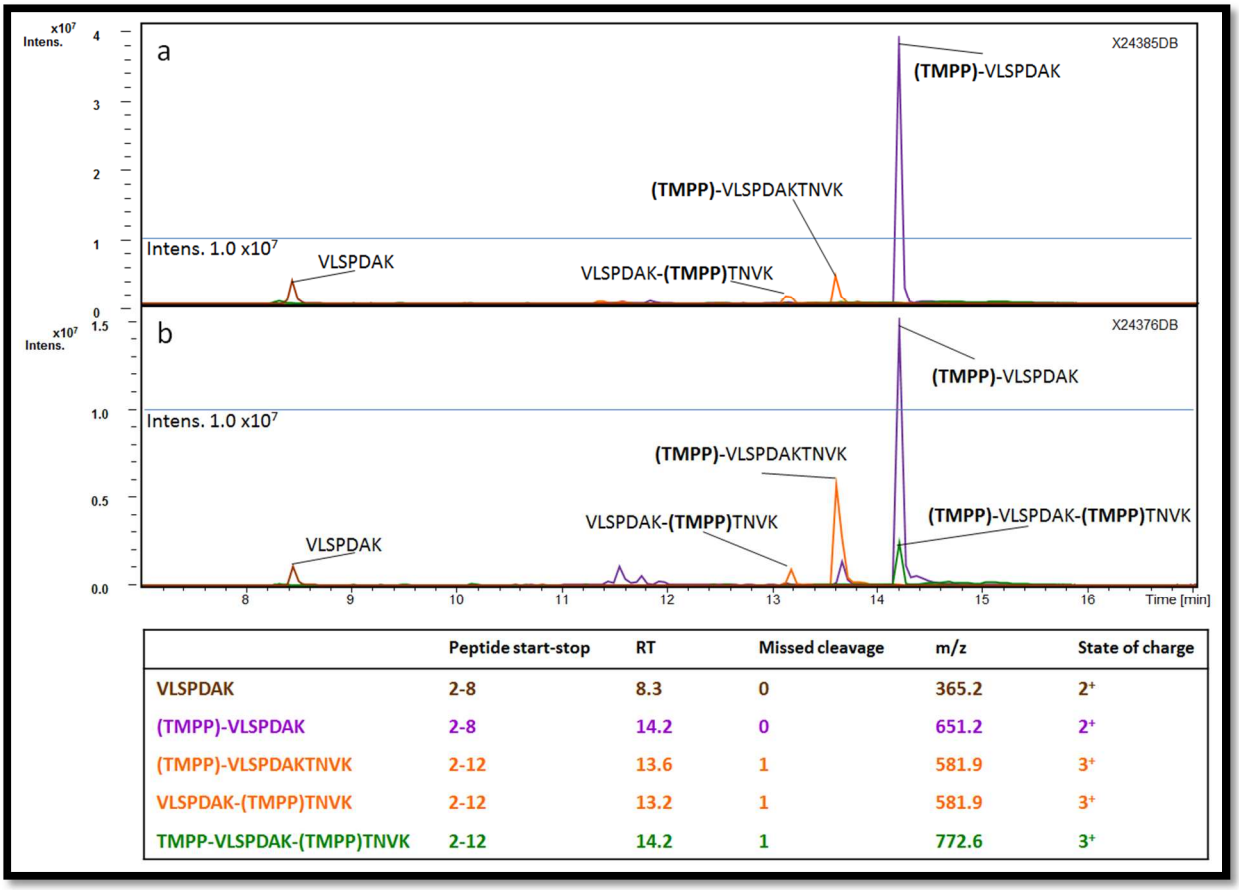
In the "Ingel N-TOP", the sample is pre-fractionated into an SDS PAGE. Then the bands are extruded, destained, dehydrated and incubated with an excess of TMPP reagent for 1 hour. The addition of an excess of hydroxylamine allows quenching the residual TMMP.

In a classical N-TOP experiment performed on a complex sample, the amount of proteins submitted to derivatization is measured performing Bradford colorimetric assay. The TMPP is then added with a controlled 200 fold molar excess. Due to presence of detergents used to elute the construct from the affinity column and to the really tiny amount of proteins available, no Bradford could be performed resulting in unknown ratio TMPP/proteins.

To evaluate the effect of an excess TMMP proteins ratio during the gel derivatization, a series of assays has been performed using different amount TMPP protein ratio. In Figure 11, two IEC of Hemoglobin α chain submitted to "Ingel N-TOP" are presented.

When the derivatization was performed with a 200 fold TMPP excess,(Figure 11 chromatogram a), the most intense peak is the expected N-terminal TMPP derivatized peptide ions, the same peptide is also detected with trypsin missed cleavage (m/z 581.9 at RT 13.6 min). The peptide resulting from the side reaction of the TMPP with a lysine is also detected but with a low intensity (m/z 581.9 at RT 13.2 min).

When the derivatization was performed with a 400 fold TMPP excess (Figure 11 chromatogram b). Like in the previous case, the N-terminal TMPP derivatized peptide ions, with and without missed cleavage, are the two most intense. But globally the yield of the reaction is reduced of about 50%. The reason for this significant yield reduction is the increase of side reaction (on ε-lysines) and the consequential generation of peptides with two missed cleavages. A form with N-terminal and side reaction derivatization, is also detected (m/z 772.6).



	Peptide start-stop	RT	Missed cleavage	m/z	State of charge
VLSPDAK	2-8	8.3	0	365.2	2⁺
(TMPP)-VLSPDAK	2-8	14.2	0	651.2	2⁺
(TMPP)-VLSPDAKTNVK	2-12	13.6	1	581.9	3⁺
VLSPDAK-(TMPP)TNVK	2-12	13.2	1	581.9	3⁺
TMPP-VLSPDAK-(TMPP)TNVK	2-12	14.2	1	772.6	3⁺

Figure 11: IEC of two Hemoglobin samples submitted to Ingel N-TOP. A) The Hemoglobin gel slice was incubated with 200 folds molar excess TMPP. B) The Hemoglobin gel slice was incubated with 400 folds molar excess TMPP.

The results displayed in Figure 11 shows that a too large excess of TMPP results in the global decrease of the yield due to the increase side reaction on the lateral chain of lysine and tyrosine.

For this reason, each chimeric construct was loaded on a lane for SDS PAGE and different known amount of Hemoglobin were also loaded in order to relate the intensity of the construct bands with the intensity of the Hemoglobin after Coomassie revelation. This approach allowed minimizing the side reaction of the TMPP due to unbalanced ratio.

Results

Five different chimeric constructs carrying a site mutation of the PEXEL motif were produced and submitted to the "Ingel N-TOP" approach we have settle down.

The mutant A51S

The processed N-terminal peptide was detected as acetylated. This result was expected since the position submitted to mutation is described as non relevant for the export. Figure 12 displays the results obtained for A1S. The confocal microscopy experiment confirmed the localization in the cytoplasm of the erythrocyte

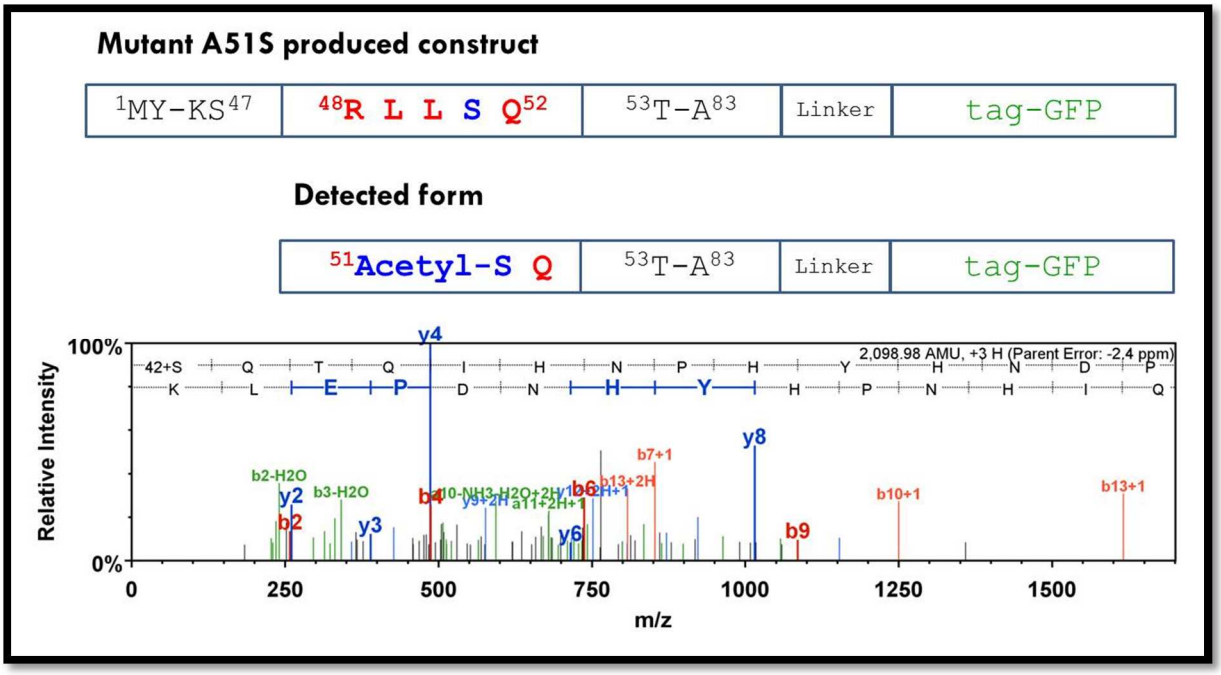


Figure 12: on the top is displayed the chimeric construct A51S, below the scheme and the MS/MS spectrum of the peptide detected.

The mutants A51D and A51R

The processed N-terminal peptide was detected as TMPP derivatized, with a low intensity MS/MS spectrum (Figure 13). This result was surprising since the position submitted to mutation is described as non relevant for the export. Due to the low quality spectra and to the unexpected biologic results, in order to validate the N-terminal derivatization of the A51D mutant a not quantitative targeted SRM method was developed.

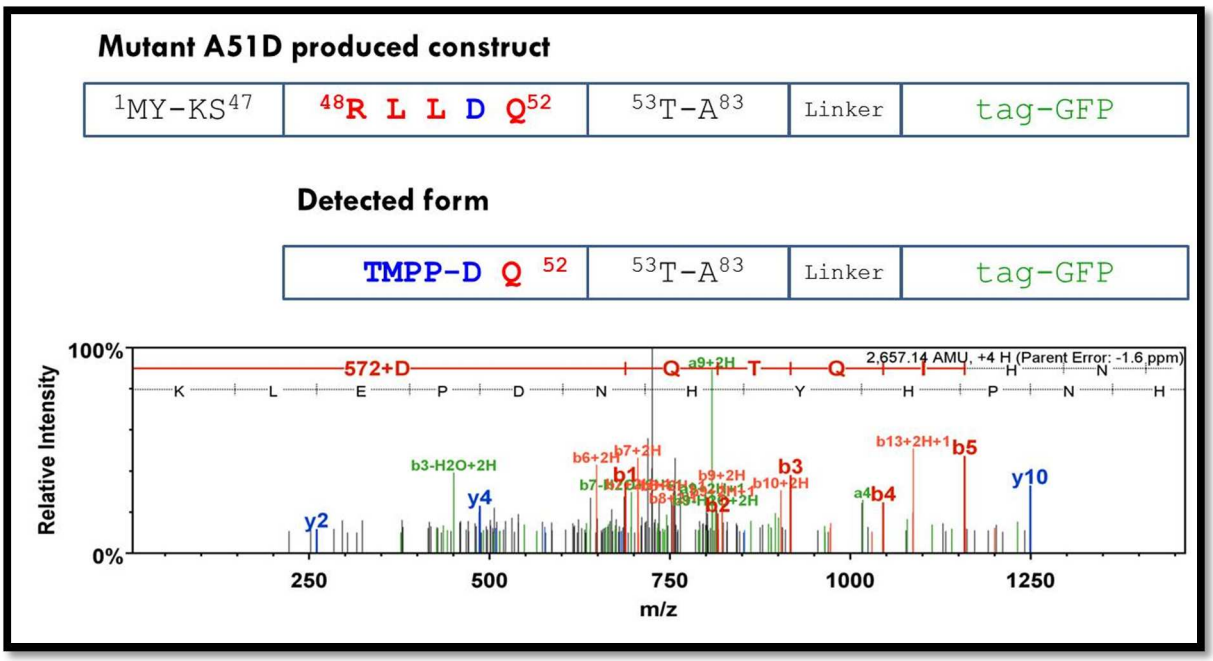


Figure 13: on the top is displayed the chimeric construct A51D, below the scheme and the MS/MS spectrum of the peptide detected.

The SRM approach, described in part II chapter IV, was selected since thanks to the physical characteristics of the QqQ. It achieves a superior sensitivity compared to DDA based experiment.

The tryptic peptide A51D, sequence TMPP-DQTQIHNPHYHNDPEL**K,** was synthetized isotopically labelled at the C-terminal Lys. The difficulties of performing a TMPP derivatization at the peptide level, resulted in the impossibilities of achieving a pure product, as displayed in Figure 14. The LC-SRM analysis of the isotopically labelled peptide produced a large and irregular peak due to TMPP degradation product, (confirmed by ESI-Q-Tof analysis) normally removed during the SDS PAGE.

The LC-SRM analysis of the endogenous TMPP derivatized peptide produced three transitions presenting a perfect peak shape and the expected relative intensities ratio.

The SRM experiment results, allowed confirming the "Ingel N-TOP" results. The combination of these orthogonal approaches, showed that the fourth amino acids of the PEXEL motif, RxL**x**E/Q/D, is also responsible for the process of exportation, since the processed N-terminal peptide, of the mutant A51D, was detected TMPP derivatized. The confocal microscopy confirmed these results detecting the GFP fluorescence in the parasite vacuole.

This unexpected result was confirmed by the processed N-terminal peptide of the **mutant A51R** that was also detected TMPP derivatized, as displayed in Figure 15.

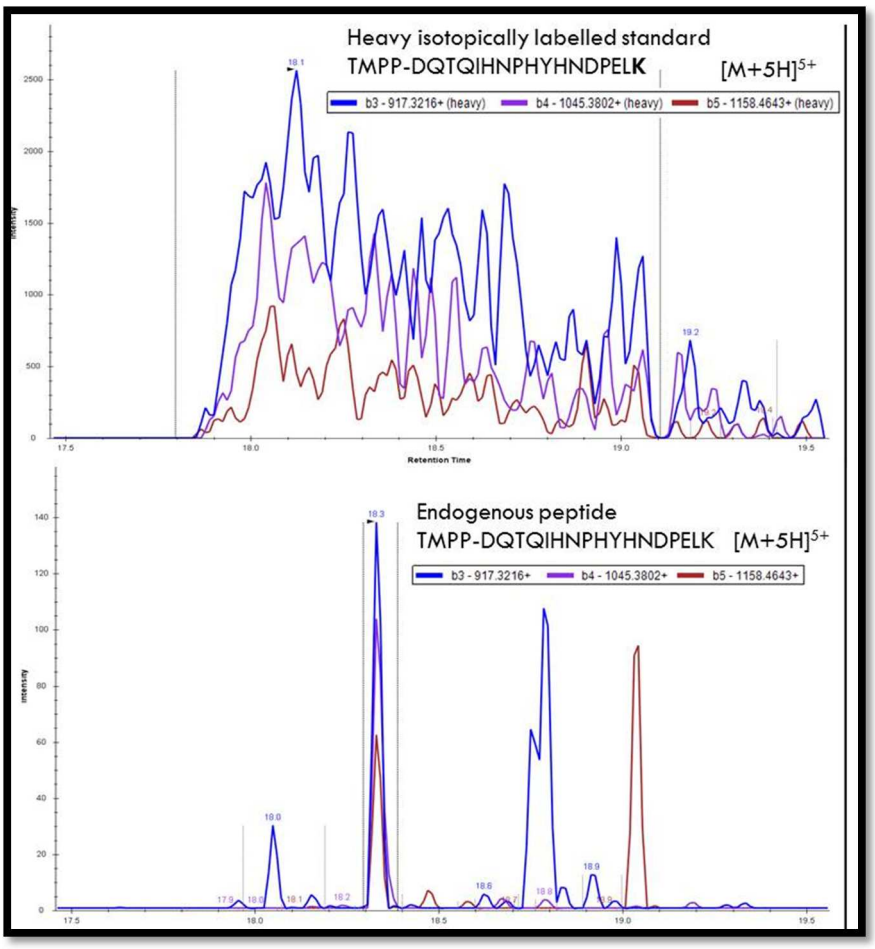


Figure 14: SRM transitions of the heavy and light peptide. The N-terminal fixed charge carryied by the TMPP required monitoring b_n series ions.

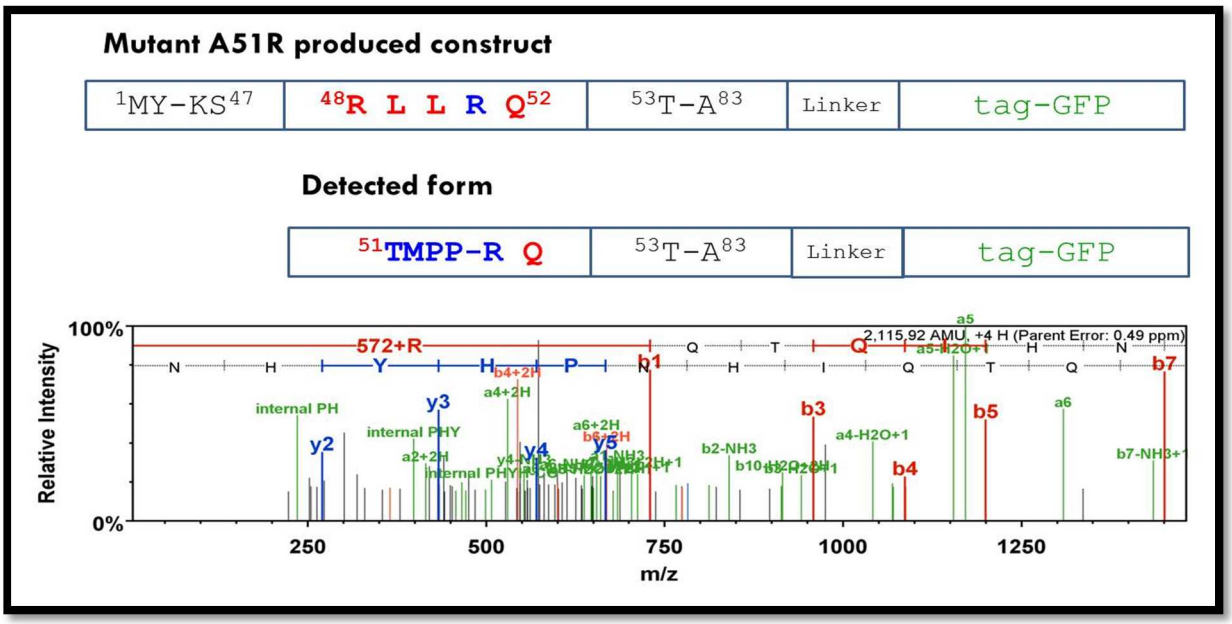


Figure 15: on the top is displayed the chimeric construct A51R, below the scheme and the MS/MS spectrum of the peptide detected.

The mutants Q52R and Q52N

To validate the hypothesis that other amino acids aside E/Q/D on the last PEXEL position are recognized by the translocon, two constructs were produced and analyzed: Q52R and Q52N.

In both cases the construct were detected with the new N-terminal acetylated (Figure 16 and Figure 17). The confocal microscopy experiment confirmed the exportation in erythrocyte despite mutation on the last PEXEL amino acid.

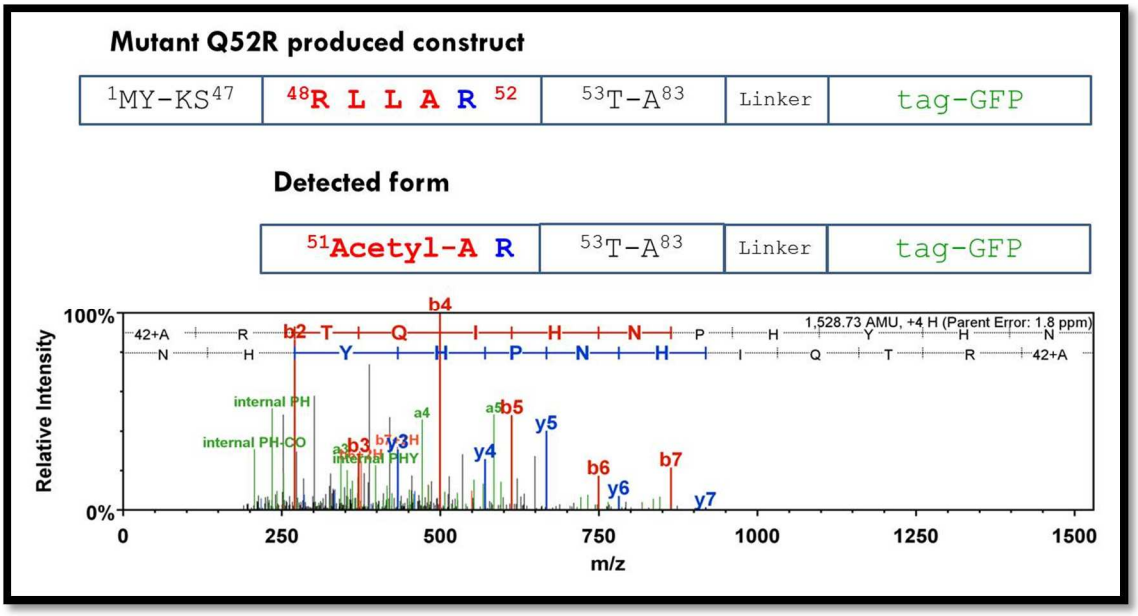


Figure 16: on the top is displayed the chimeric construct Q52R, below the scheme and the MS/MS spectrum of the peptide detected.

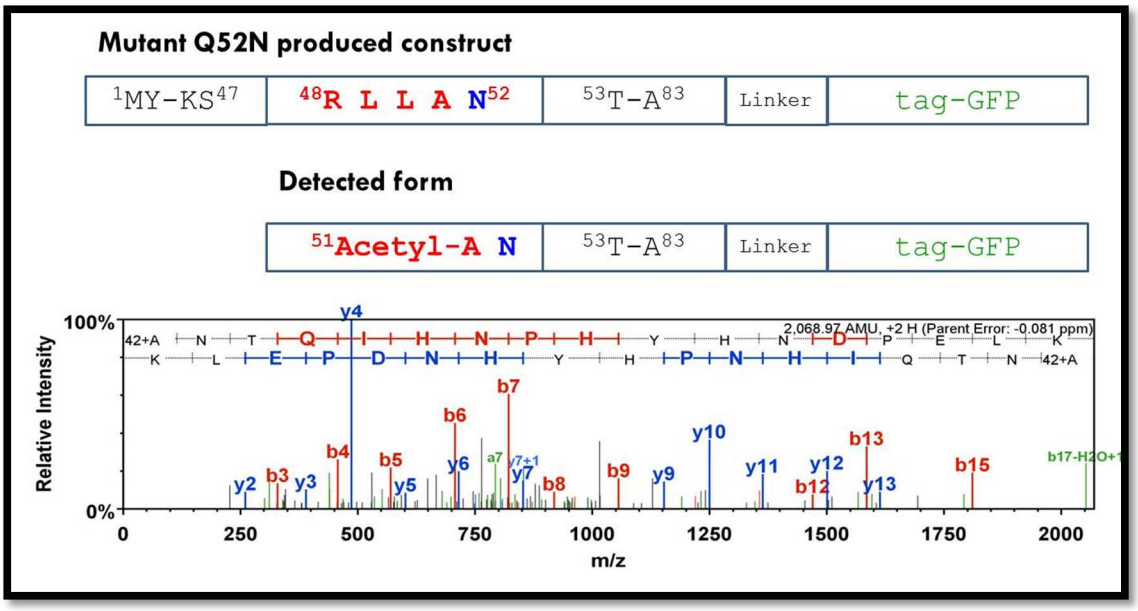


Figure 17 on the top is displayed the chimeric construct Q52N, below the scheme and the MS/MS spectrum of the peptide detected.

Conclusion

In this chapter, I have presented a method optimization of the "Ingel N-TOP" approach, allowing characterizing the N-termini position on unknown amount of low concentrated proteins prefractionated on a SDS PAGE.

The key point was the use of an "internal standard of staining intensity" when running the SDS-PAGE. In such way, intensities of the chimeric protein constructs were related to the intensity of the standard, allowing minimizing the side reactions of the TMPP due to unbalanced ratio protein/TMPP.

This optimized method, was applied to the study of the Host Cell Targeting signal (HCT) proteins exportation system in *P. falciparum.* For each amino acidic substitution on the *P. falciparum* chimeric PEXEL construct, the site of cleavage and the new N-termini position were determined.

The results obtained by orthogonal approaches like the "Ingel N-TOP" and fluorescent confocal microscopy confirmed the two initial hypotheses:

- The fourth amino acid position in the PEXEL motif considered as not influent in export process instead showed a role in the process of exportation.
- Also the amino acid R and N, aside the described E/Q/D, on the last PEXEL position are recognized by the translocon and allows an efficient export to the erythrocyte cytoplasm.

The writing of a publication paper presenting these results is ongoing.

Part I - Chapter II -

Development of a stable isotope labeling method to validate N-terminal TMPP derivatized peptides

Part I - Chapter II - Development of a stable isotope labeling method to validate N-terminal TMPP derivatized peptides

Introduction

The modern high-throughput DNA sequencing technologies allowed to decode the entire Human genome in 2004 (43). Since then, every year new species are sequenced and 180 entire genomes are up to now completed. The availability for the scientific community of the genome sequence, boosted the development of proteomics. It also laid the foundation stone to a new "omics" approach called proteogenomics. This new scientific effort is based on the "integration of transcriptomics and proteomics data for a better understanding of the genome" (44).

Despite the progress of DNA sequencing the process of gene finding, which has as goal to localize the position of the coding sequence in the genome, is a task far to be solved.

The majorities of the protein databases is obtained from in silico prediction and are translation of the coding sequence. For example, UniProtKB/SwissProt, in the July 2014 release, contains 69.9% of protein sequences inferred by homology and just 36.8% with proteomics or transcriptomics evidences (45).

Proteomics search engines compare the experimental MS/MS spectra with the theoretical obtained by the *in silico* digestion and fragmentation of the protein database. This workflow relies on the idea of a database free of error.

Experimentally determine the protein N-start allows to:

- Annotating an alternative protein N-terminus.
- Characterizing protein N-termini processing and protease cleavage sites.
- Correcting protein N-start in databases today almost exclusively the result of automatic bioinformatic start prediction.

The method development and the results obtained in this study have been published in Journal of Proteome Research (46). This publication is enclosed in the annexes.

Analytical task

The N-TOP strategy, developed by Gallien (1), is based on a selective derivatization of the N-termini amines with TMPP. The chemical reaction and the advantage offered are detailed in the introduction of the Part I.

The fixed charge carried by the phosphonium group in the TMPP, responsible for the increase electrospray ionization efficiency, is also responsible for an unexpected behavior during CID fragmentation. In tryptic peptides, the presence of two fixed charges (the N-terminal TMPP and the C-terminal lysine or arginine) increase the thermodynamic competiveness of the remote fragmentation model (47). This type of fragmentation produces mainly ions from the a_n and b_n series instead of the b_n and y_n series produced by the mobile proton fragmentation (48). Figure 18 displays two MS/MS spectra obtained by the same peptide, on the left without derivatization, on the right TMPP derivatized. Highlighted in red, the b_n and y_n series are the most intense fragments in *the TMPP* derivatized peptide spectrum.

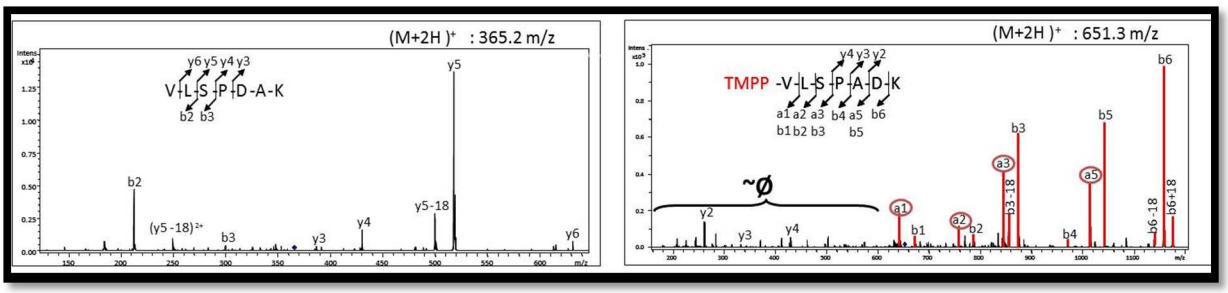


Figure 18: two MS/MS spectra obtained by the same peptide, on the left without derivatization and on the right TMPP derivatized.

Search engines are trained to score classic MS/MS spectra, for such reason, TMPP derivatized peptides are often not optimally scored despite a sufficient spectra quality. In Figure 19 is displayed an MS/MS spectrum of a TMPP derivatized peptide. Due to the prevalence of a_n and b_n ions series the spectrum has been scored by Mascot with an ion score of 12, resulting in its discard during the FDR validation process, despite the relevant fragmentation quality.

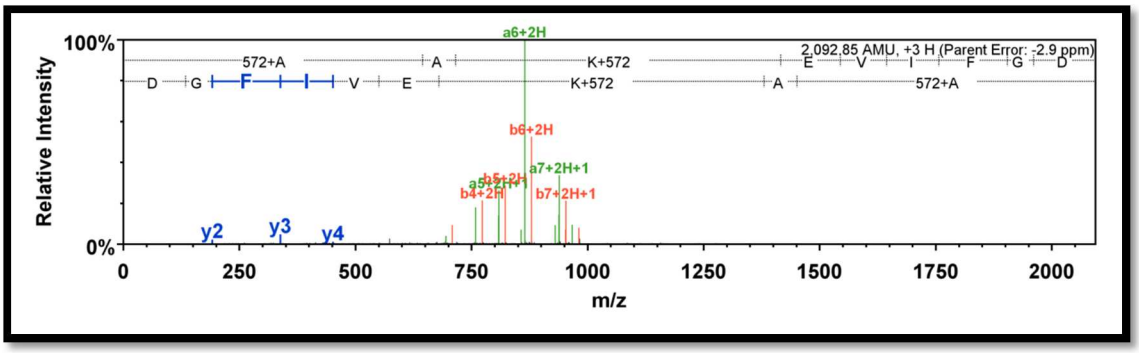


Figure 19: MS/MS spectra of a TMPP peptide identified by Mascot with a low ion score despite the decent signal over noise and the good fragmentation quality.

The not optimal scoring of the derivatized peptide represents the main bottleneck of the N-TOP approach. Up to now, a manual inspection of the spectra was required in order to validate them.

The goal of this project was to overcome the scoring problem by developing an automated validation method for TMPP derivatized peptides.

Method development results

The strategy developed to overcome the not optimal scoring of the TMPP derivatized peptides, is based on the double derivatization of the sample with an equimolar mixture of TMPP heavy isotopically labelled and light TMPP. This approach was called dN-TOP and relies on the search engine double identification of the same peptide derivatized with the two TMPP.

For this purpose, we have designed a TMPP carrying $^{13}C_9$. Since the 9 Da are easily detectable by mass spectrometry. The two peptides derivatized (one with the heavy TMPP and one with the light) will exhibit the same:

- Chromatographic properties.
- Ionization and fragmentation efficiency.

In Figure 20-I is displayed the ion chromatogram extraction of a peptide derivatized with the TMPP mixture. Both the TMPP forms exhibit the same derivatization efficiency, retention time and ionization efficiency. Below in Figure 20-II is displayed the isotopic envelop of two derivatized peptides separated by 4.5 m/z (double charged)

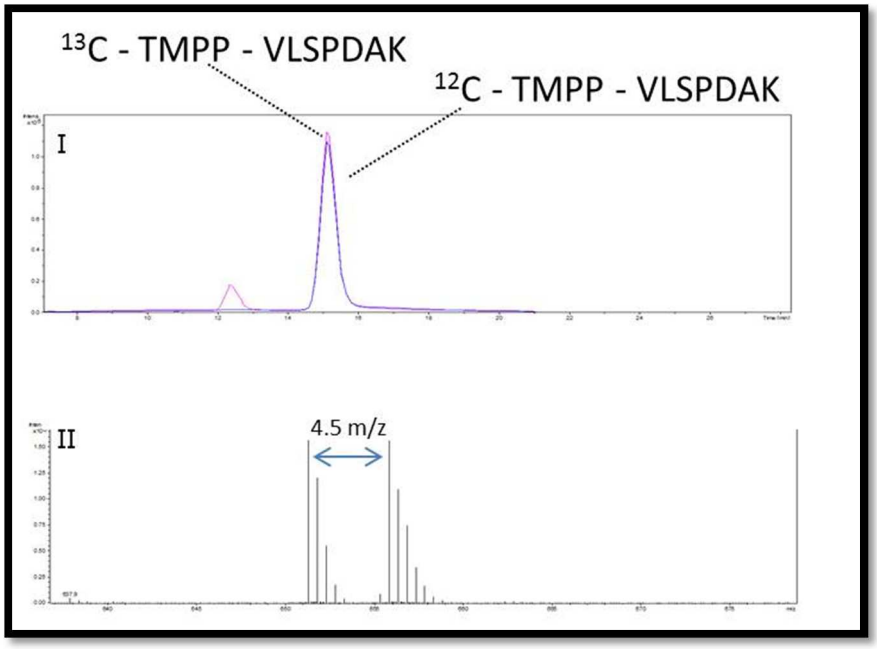


Figure 20-I: ion chromatogram extraction of two peptides derivatized, one with the heavy TMPP and one with light. Figure 18-II: the isotopic envelope of the two derivatized peptides.

In Figure 21, two MS/MS spectra obtained derivatizing with the TMPP heavy/light mixture, display a similar fragmentation pattern.

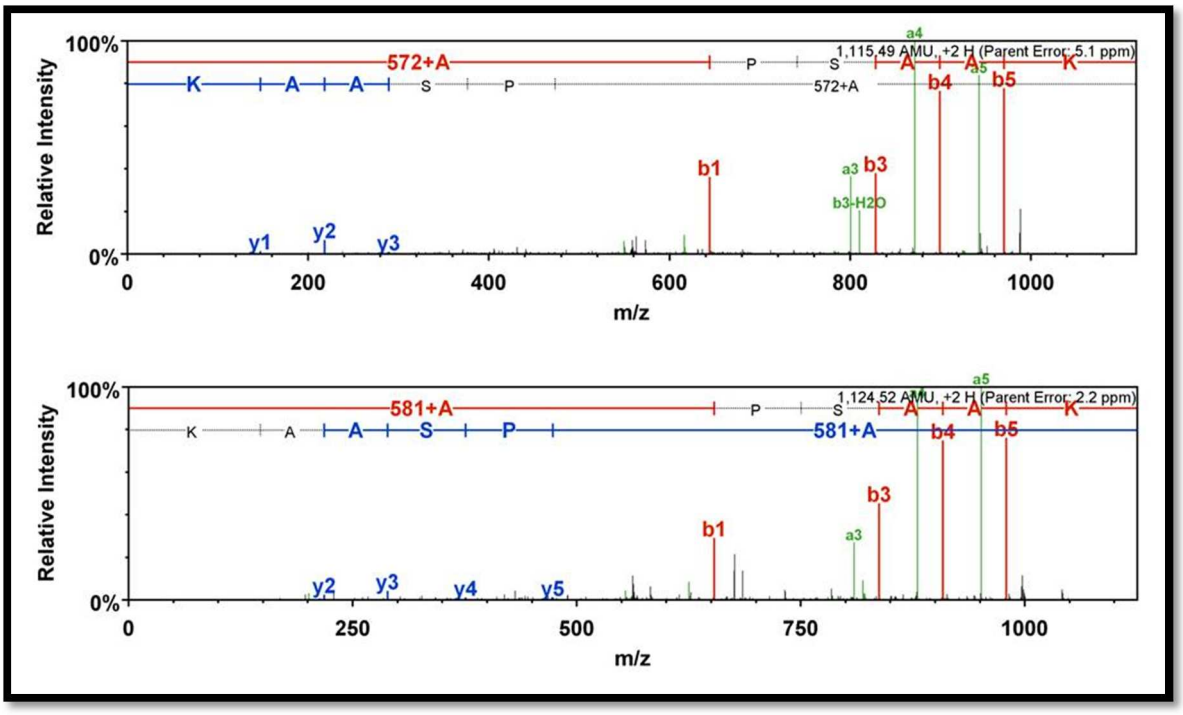


Figure 21: MS/MS fragmentation pattern of a peptide derivatized with light and heavy TMPP.

In Figure 22, the new dN-TOP workflow is compared with the classic N-TOP. The general workflow is common for the two approaches, in black bold are highlighted the differences.

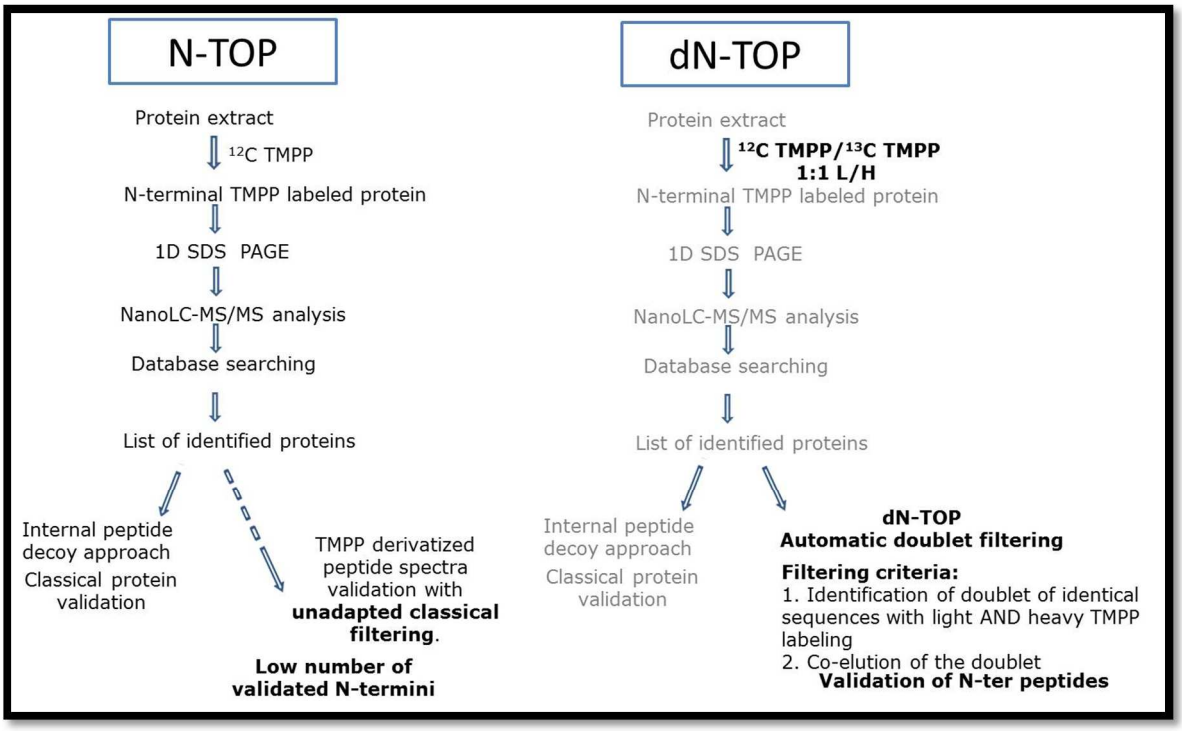


Figure 22: schematic comparison of the N-TOP original method and the optimized dN-TOP. In black bold are highlighted the differences between the two methods.

The protein extract is labelled the two TMPP isotopologues (L and H) and the excess of TMMP is removed during the 1D SDS PAGE prefractionation. The gel bands are extruded and submitted to classic proteomic workflow. The internal peptides are validated using the decoy approach. In parallel, the list of derivatized peptides is exported to excel without regards to the search engine ion score. We have developed an in house macro which scans for all the doublet identification, monitoring the following characteristics:

- The same peptide sequence.
- The same modification (oxidation, carbamidomethylation, methylation).
- Derivatized once with light once with heavy TMPP.
- Δ retention time of the two peptides <30 seconds.

The automatic validation of the derivatized peptide is performed if all the criteria are satisfied.

Characterization of the *Herminiimonas arsenicoxydans* N-terminome

As a proof of concept the dN-TOP method was applied to the characterization of the *Herminiimonas arsenicoxydans* N-terminome, since GC rich prokaryotes genomes results in a less confident bioinformatics N-termini start prediction. The results are detailed in the annexed publication. Briefly 504 unique proteins were identified, 90 unique N-terminal peptides were validated, among which 13 were wrongly predicted by the genome annotation.

In Figure 23 is displayed the result of the N-terminus determination experiment for the protein "putative NADPH-dependent FMN reductase, ArsH-like protein". The N-terminus experimentally

validated is at the position 13. The dN-TOP strategy allowed us to correct the protein start originally predicted by the software AMIGene (49). The MS/MS spectra of the doublet present a good spectra quality. It should be noticed the detection of an unusual ion fragment, composed by four internal amino acids PNID 20-23 caused by the TMPP fixed charge.

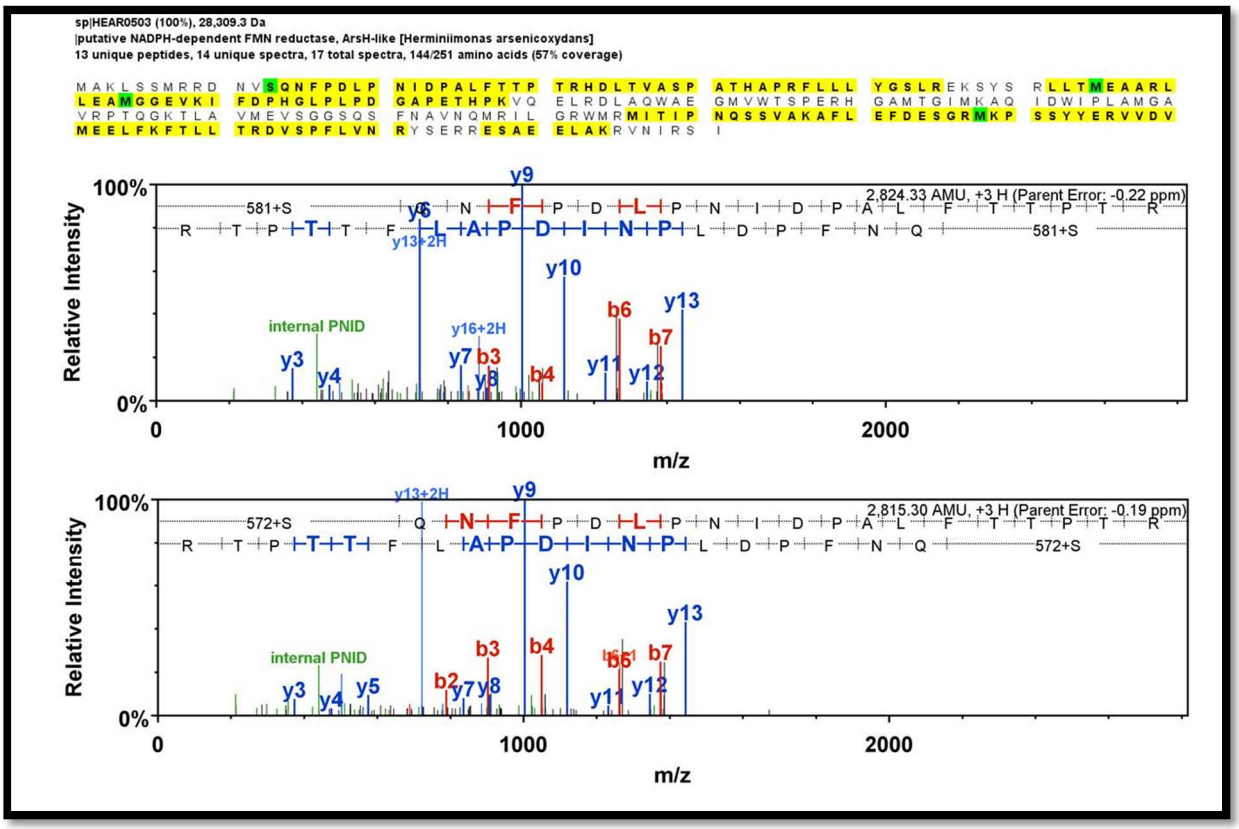


Figure 23: MS/MS fragmentation of doublet of the putative NADPH-dependent FMN reductase, ArsH-like protein. On the top is displayed the protein sequence, highlighted in yellow the sequence of the peptide detected. In green the amino acids identified modified. The N-terminus experimentally validated is the position 13.

Conclusion

In this chapter, we have shown that the dN-TOP approach allowed overcoming the time consuming manual validation step. This process was up to now needed because of the not optimally scored TMPP derivatized peptides. The identification of the TMPP doublet (light and heavy TMPP) in the same chromatographic peak permits validating the identification of the N-terminal peptide without ambiguity. This new approach is now fully automatized, reproducible and at least 10 time faster.

Thanks to the dN-TOP development, the application of the TMPP derivatization can now be integrated in high throuput study, making of proteogenomics an every day application. For example, dN-TOP strategy is now used in our laboratory in an collaborative project with Prof Lanzer on the proteome of Human red blood cells infected by *P. falciparum.*

Part II

Introduction:
The quantification in mass spectrometry based proteomics

Chapter I
Relative quantification of affinity-purified proteins complex

Chapter II
Targeted Selected Reaction Monitoring based Prion protein quantification

Chapter III
Monoclonal IgG glycation batch to batch quantification

Chapter IV
HbA_2 quantification of intact proteins in human hemoglobin by LC-MS

Chapter V
Optimization of a Data Independent Acquisition method and MS1 comparison

Part II – Introduction -

The quantification in mass spectrometry based proteomics

Part II – Introduction - The quantification in mass spectrometry based proteomics

Providing exhaustive quantitative information is one of the main goals of the modern proteomic (10, 50), but the most common soft ionization processes: Electro Spray Ionization (ESI) and Matrix Assisted Laser Desorption Ionization (MALDI) are not inherently quantitative approaches.

The ionization efficiency of each peptide is determined by: size, amino acidic compositions and polarity, characteristics that are specific for each peptide.

For such reasons, the intensity recorded by the mass spectrometer can't be straightforward related to the concentration of the analyte. In Figure 24, is presented, as example, a chromatogram obtained extracting the ions masses of an equimolar mixture of peptides analyzed with an LC-ESI-MS. Clearly, the intensity detected does not reflect the equimolar concentration. In addition to the limitation of the ionization step, also the transmission of the ions in the mass spectrometer often favor specific m/z, resulting to be not linear across all the m/z range.

To overcome these limitations, four orthogonal approaches (schematized in Table 1) have been developed (51): gel based quantification, label free, label based and targeted quantification, each of these approaches presents drawbacks and has specific advantages that will be summarized in this chapter.

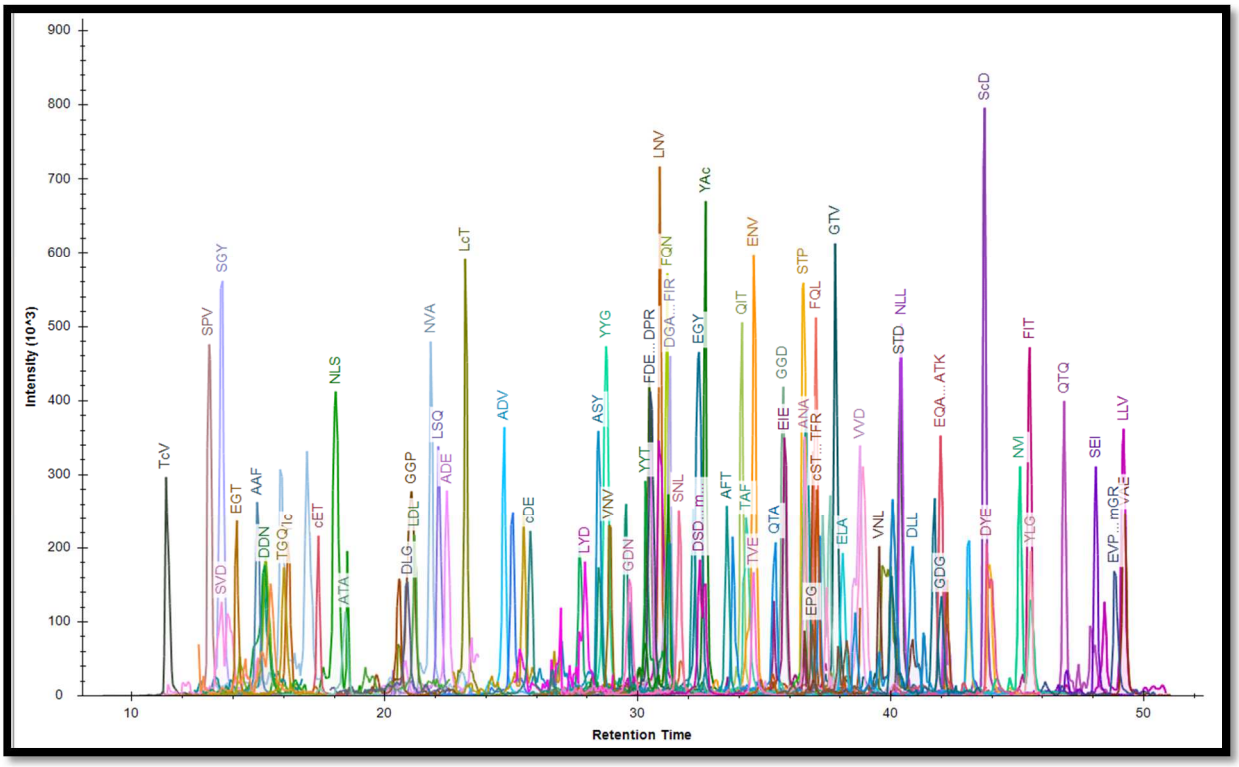


Figure 24: extracted ion chromatogram of the parent masses of tryptic peptides generated by the digestion of an equimolar mixture of proteins has been extracted in an LC-MS analysis. The ionization efficiency spans across a factor ten due to chemical properties of the peptides.

2DE Gel based quantification

The 2DE gel quantification is based on the generation of two dimensional gels proteomics maps, that are differentially compared using the intensity of the coloration as quantification information.

The two dimensions separation are orthogonal; the first is the electro-focalization, that allows to separate the proteins according to the Isoelectric point pI; the second dimension is a classic SDS PAGE separation, that allows a separation based on the molecular weight.

Specifically developed software (PDQuest or SameSpot) can perform alignment of the same spot in different gels and determine an eventual differential intensity of the spots, that are then extruded from the gels and submitted to mass spectrometry based identification.

A powerful technical improvement of the 2D gel map was the introduction of the DiGE (Differential Gel Electrophoresis). This approach is based on the labelling of two biologic states of a sample with two fluorescent cyanines, responding to different wavelengths. The two samples can then be mixed in equal amount and separated on a 2D gel. The DiGE (52-54) allows detecting low level protein regulation and offers the possibilities of multiplexing, thanks to specifically developed system running four gels in the same time. The identification of the spots of performed by MS based proteomic analysis despite the derivatization.

Despite the 2DE resolving power, it's possible to detect, during the mass spectrometry analysis, a spot composed by more than one protein. Since the quantification is performed measuring the absorbance of the spot, complementary analysis are required to highlight how many proteins are regulated.

The main bottlenecks of this technique are represented by the relevant expertise required to minimize the analytical variability produced during analysis of the replicates. In addition, this approach is usually considered as time consuming.

Label free

Despite the not quantitative nature of mass spectrometry, quantitative information can be obtained even without using isotopically labelled standard, or performing derivatization, such approaches are called label free. Using these techniques, differential relative quantification can be obtained, based on the comparison of two biologic states of the same sample. Classically, the intensity of an analyte detected in an MS or MS/MS analysis is compared against the intensity of the same analyte in a different biologic state (healthy/pathologic or wild type/mutant).

The results that can be expected from this type of quantification are to highlight main biologic regulation but not to detect fine regulation (55) as it will be illustrated in chapter 1. Different label free techniques are now available, as described in Table 1, classified respect to the MS acquisition methods used (56, 57):

- In the Data Dependent Acquisition (DDA), the mass spectrometer performs an MS scan, also called survey scan and depending on the tuning and the performances of the instrument the most intense peptides are selected to be submitted one by one to MS/MS fragmentation.

- In the Data Independent Acquisition (DIA), the mass spectrometer selects and sequences all the peptides in one or multiple consecutives m/z windows, without any *a priori* regards to the parent ions.

Label free DDA mode

The classic proteomic identification is performed acquiring in DDA mode (56, 58-60). Performing a fine optimization of some instrument parameters, such as cycle time or exclusion time, permits to extract quantitative information from DDA acquired data. These instrumental adjustments, affect just marginally the number of identifications.

The quantitative information can be obtained: from the intensity of the parent ion masses in the case of the MS1 method (61), or from the counts of MS/MS spectra identified for each protein in the case of the spectral counting (62).

The MS1 label free has been used since a long time in small molecule mass spectrometry. This workflow, revisited for proteomics applications, just recently became applicable to it, thanks to the technological advance of the mass spectrometers in term of MS/MS acquisition rate. The MS1, will be presented in detail in chapters 1 and 3, has the advantage of not being restricted by the stochastic selection of the peptide submitted to MS/MS in DDA mode. This means that even if a peptide has not been selected and therefore not identified in a specific injection (due to under-sampling), it can be differentially quantified if it was present with sufficient intensity during the survey scan.

Two types of software allow generating MS1 data(63):

- Features detection like: MaxQuant (64) or Progenesis (65), based on the chromatogram alignment and detection of all the ions masses, defined as features, that present intensity variation. This approach allows virtually detecting all the peptides regulated, even those that have not been identified by the search engine.
- Ion chromatogram extraction of spectral library (66) enlisted peptides, like Skyline, developed by the lab of MacCoss (67) or Proteinscape (68), are instead based on the extraction of the parent mass identified by the search engine. In this approach just the peptides identified at least once are quantified.

The features detection approaches, due to the way of conception, based on the chromatographic alignment, are more robust against issue of chromatographic retention time reproducibility.

The MS2 spectral counts, which will be detailed in chapter 1, is the most flexible label free method to set up. Software like Scaffold (69) allows to rapidly extract the number of spectra identified for each protein. The counts of spectra reflect the number of times that all the peptides composing a protein are sequenced and identified by the search engine. The limitation of this method, is represented by the not easy to find equilibrium between number of protein identified and solidity of the results. The accuracy of the spectral count is increased by the redundancy of the peptides identified; on the other hand, maximizing the redundancy, increase the under-sampling effect for the less abundance proteins that risk to not be identified.

Label free DIA mode

An alternative way, of circumventing the under-sampling in DDA label free experiment, is to do not select ions peptides one by one, but to fragment all the ions present in a defined m/z that are

eluted from the chromatographic system. Many original DIA acquisition methods have been presented in the last 20 years, starting from the shotgun CID DIA described by Purvine (70), but search algorithms capable of interpreting MS/MS spectra composed by multiple peptides, are not yet available. This limits the DIA, to a quantitative method not capable of performing discovery proteomics, since spectral library are needed.

The only exception is represented by PLGS®, a WATERS® proprietary search algorithm, developed for MS^e data. This acquisition method represents the evolution of the Purvine's shotgun CID, where all the peptides were submitted to in source fragmentation. In the MS^e (71) instead an MS scan is alternate to an MS/MS fragmentation performed in the collision cell, like in the above described approach, no peptide ion selection is performed by the quadrupole. The possibilities of performing a label free study on an extended dynamic range and the availability of search engine made of the MS^e a method ready to be used in routine acquisition method (72-74).

A second DIA acquisition method has recently gaining popularity in the scientific community as label free technique: the SWATH™. Originally developed by the group of Aebersold (75), is based on consecutive selections of broad m/z windows submitted to MS/MS. In this approach, the quantitative information is obtained at the MS/MS level by extracting a transition from a parent mass to a product ion. The advantage of this technique is represented by an increased dynamic range and a lower limit of detection respect to the DDA MS1, as presented in the chapter 1.

As described above, no search engine has been released for DIA method and it is therefore necessary to use a file called spectral library (generated acquiring in DDA mode) listing the transition that will be extracted. This additional step, results to be time consuming and it represents the limitation of the SWATH™.

SWATH™ is now available for different mass spectrometers (AB Sciex, ThermoScientific, Bruker Daltonics) the neutral vendor characteristic of this acquisition method, make it the perfect candidate for the development of a dedicated and performant search engine, that could be used for all the mass spectrometers platforms.

Label based quantification

In this paragraph are briefly summarized some of the differential quantitative methods based on the sample labeling, two main categories are described: metabolic isotope incorporation (SILAC) and chemical derivatization (ICAT, iTRAQ, TMT, …).

Among the metabolic labeling, the SILAC (Stable Isotopic Labeling with Amino acids in Cell) culture is the most widely used. Two or more biologic samples can be differentially analyzed: the cell culture of the control sample is performed in classic condition and the different biologic state is grown in a media containing one amino acid, generally Leucine, Lysine (76, 77) or Arginine (78) isotopically labelled. After minimum 5 cellular mitotic events the cell proteins present 100% of the selected amino acid isotopically labelled. The cellular proteins extracts are then mixed and differentially quantified using the intensity of the peptide ions acquired in the MS scan.

Conceptually the SILAC method represents the best option to obtain not targeted robust quantitative results, since the labelling is total and performed at the early stage of the experiment. and the technical bias are minimized due to the identical workflows used for both

labelled and not labelled samples. The most relevant and obvious limitation is the not applicability of the metabolic labelling to complex organism, aside few proofs of concept experiments (79, 80) and the relevant cost and time required (59).

The ICAT (Isotope Affinity Tag) (81) is based on the chemical derivatization of cysteines residues with a reagent composed of a iodoacetamide linked to a biotin tag through a space linker. The two samples are derivatized separately, one with an isotopically labelled tag and another with a light one. After the derivatization, the samples are mixed, digested and submitted to purification thanks to the biotin tag. The differential intensities of the light and heavy peptides are used to quantify the two samples. The original tag described by the Aebersold group had the limitation of inducing a chromatographic shift between the light and heavy labelling therefore an alternative tag based on ^{13}C has been developed to circumvent this limitation (82).

An alternative strategy, based on the chemical derivatization of the free amine at peptidic level, is the TMT Tandem Mass Tags (83). In the original workflow, four samples are labelled with isobaric reagents and then mixed. Once the TMT labelled peptide are selected and fragmented each plex generates a different ion reporter. The intensities of the ions reporter are used to perform the quantitation.

This approach presents two limitations: first, on the contrary of the methods based on the MS quantitation, the labelled peptide needs to be fragmented in order to be quantitated, second the workflow requires relevant amount of proteins (84) not always available.

Selected Reaction Monitoring

The SRM (85, 86) Selected Reaction Monitoring (the gold standard in terms of mass spectrometry based quantification (87)) is a targeted quantitative approach, based on the spike of synthetic peptides isotopically labelled. The SRM is mainly performed, but not necessarily, on a mass spectrometer triple quadrupole QqQ, where the first quadrupole selects the peptide mass, the second is responsible for the fragmentation and the third one selects the expected product ions. The peculiar architecture of this mass spectrometer is responsible for highest cycle duty time among the mass spectrometer configuration, resulting in an increased sensitivity. Two main limitations are characteristic of this approach: the limited number of peptides that can be quantified in a single LC-SRM run, approximately 100 endogenous peptides and the time consuming method development. In chapter 2 two methods development and applications of the SRM will be presented.

Conclusion

As described in Table 1, two main orthogonal approaches are used in routine in proteomics, targeted and shotgun proteomics. For each of them are described in literature advantage and drawback but neither of the two is capable of answering all the different biologic questions, since specific analytical limitation are presented in all the techniques. For such reasons in some cases the combination of different approaches can be useful, but due to availability of the sample and of the mass spectrometer this is not always possible.

SRM represents the sharpest tool in terms of results robustness and sensitivity, but due to the targeted approach that requires a hypothesis driven experiment, and due to the limited number of peptides quantifiable in one injection, it is not applicable to all the projects.

The labelled based techniques present an increased robustness against the analytical variability induced by the chromatographic system (stability of the retention time and enlargement of the peak width) and induced by the mass spectrometer (decreasing sensibility across the experiment) thanks to multiplexing.

The label free approaches are extremely flexible and easy to set up, but results are affected by the analytical variability since multiplexing is not possible. Despite this limitation label free has the potential of becoming a valid alternative to the label based techniques, as long as the scientific community will keep developing and optimizing it.

In conclusion all the methods summarized in this introduction of the part II are valid options in specific analytic task, but any of them can be pointed as the best.

In the next chapters five different **quantitative projects** will be presented, each of them requiring a careful development in order to achieve the required results.

Category	Method	Labeling stage	Labeling site	Quantification basis	Nature of samples	Type of quantification
Chemical labeling	ICAT/ cICAT	Protein level prior to enzyme digestion	Cysteine residues	Relative area of MS peak	Cell lysate or tissue lysate (post-extraction)	Accurate relative
	iTRAQ/ TMT	Peptide level	Primary amines at peptide N-terminali and Lys-side chains	Relative area of signature ion peaks in each MS/MS spectrum	Cell lysate or tissue lysate (post-extraction)	Accurate relative
Metabolic labeling	SILAC	cell culture	Depending on the labelled amino acids used, peptides can be labelled at Leu, Lys and Arg	Area of MS peaks	Actively dividing cells auxotrophic for the labelled amino acid	Accurate relative
Synthetic labeling	AQUA	synthetic labelled peptides generated for use as reference	Labelled residues used	Relative area of MRM traces	Can be spiked into any sample	Absolute
Label free	Spectral counts	-	-	Number of search engine assigned spectra for protein	Any	Relative
	MS1	-	-	Relative area of MS peak	Any	Relative
	DIA Swath™ (Broadband)	-	-	Relative area of pseudo MRM traces	Any	Relative
	MS^e	-	-	Relative area of MS peak	Any	Relative

Table 1 Adapted from (55). The Table summarizes the quantitative approaches presented in the chapter.

Part II - Chapter I -

Relative quantification of affinity-purified proteins complex

Part II - Chapter I – Relative quantification of affinity-purified proteins complex

Introduction

Proteins are the biologic effector of the DNA coded information. They interact with DNA, RNA and other proteins in a fast and dynamic way in order to maintain the cellular homeostasis.

The characterization of this type of interaction is far to be a routine technique. The two step affinity-purification called TAP-TAG (88, 89) represents one of the best approach capable of purifying with a genomically tagged protein, all the proteic partners exhibiting a not covalent interaction (Figure 25). The purified proteins are then submitted to proteomic workflow in order to be identified.

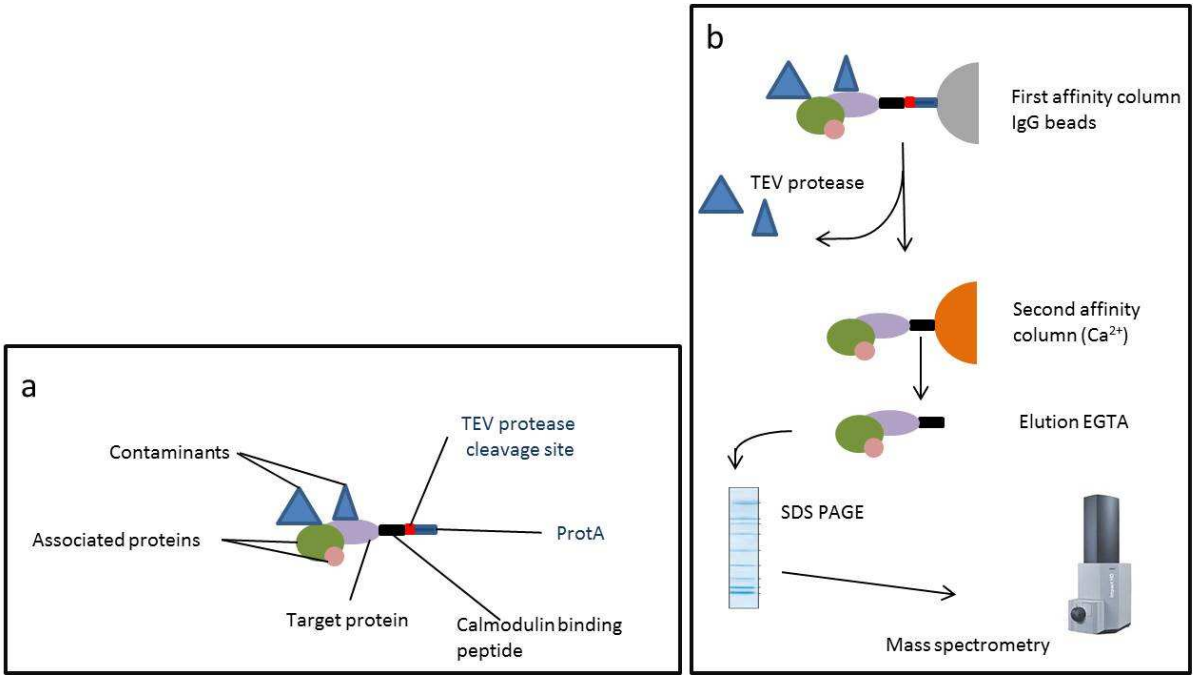


Figure 25-a: scheme of the construct of the TAP-tag purification. The target protein is expressed fused at the N- or C-Terminus with a construct composed by calmodulin binding peptide, TEV protease cleavage site motif and protein A.

Figure 23-b: schematic workflow of the AP-Mass Spectrometry: during the first step of purification the IgG binds the protein A allowing the purification of the targeted associated proteins. The complex yet carrying a high level of contaminants is released using the TEV, a protease recognizing the specific sequence inserted. The contaminants are removed during the second purification step, where the calmodulin binding protein recognizes the Ca^{2+} ions exposed on the column. The complex so purified is then eluted from the column using ethylene glycol tetra acetic acid (EGTA) and then submitted to mass spectrometry characterization and quantification.

This technique, thanks to the two steps purification, allows strongly reducing the not specific proteins that would be retained by a single step process.

For this reason, this approach is also called "Affinity purification Mass Spectrometry" (AP-MS) became one of the most widely used molecular biology approach to determine and quantify the stoichiometry of an in vivo protein-protein interaction, Figure 25-b (90-92)

In this chapter, it will be presented the method optimization performed to adapt the classic mass spectrometry workflow, to the label free stoichiometric quantification of the proteins spliceosome complex in Saccharomyces *cerevisiae*.

This project was carried out in collaboration with the group "Ubiquitin and dynamic of molecular assembling" guided by Dr. Dargemont at the Monod Institut, Paris. The results of this collaboration have been published in Biology of the Cell (93).

The biologic context

The DNA in Eukaryote is stored in the nucleus as chromatin, a DNA-protein structure. The smallest unit of the chromatin is the nucleosome: a protein octamer containing two copies of each histone: H2A, H2B, H3 and H4 wrapped by two turns of DNA (Figure 26).

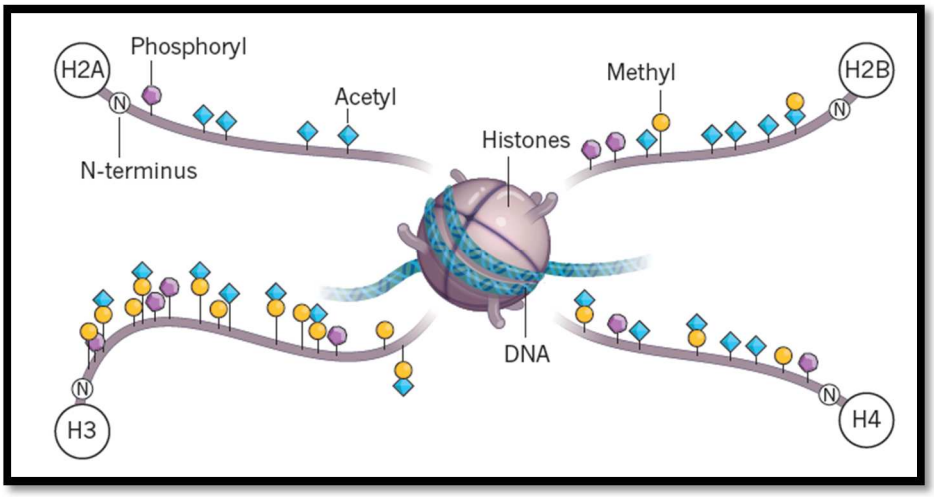


Figure 26: schematic representation of a nucleosome, adapted from (94). At the center is displayed the histone octamer, in the four corner are zoomed the N-terminus histone protein carrying different PTMs responsible for the transcription activation/repression of the protein coded in the close DNA region.

The N-termini of histones protrude outside the globular structure and are exposed to reversible PTMs like: acetylation, phosphorylation, methylation and ubiquitylation. These modifications act like molecular switches capable of modulating the gene expression, altering the electrostatic repulsion between protein and DNA double helix (94-96).

The ubiquitylation of the H2B histone (Ub-H2B), on the K123, has recently been proved to positively regulate the messenger RNA export from the nucleus to the cytoplasm, thanks to an increased recruiting efficiency of the nuclear export machinery (97, 98).

The goal of this project was to establish if the ubiquitylation on the K123 has also an influence on the recruiting efficacy of the splicing machine components. To investigate such theory, a yeast strain was mutated in order to prevent the ubiquitylation on the Ub-H2B (the Lysine 123 has

been substituted with a Tyrosine K123Y). Since the ubiquitylation is a PTM specific for K, the amino acidic substitution performed in position K123 prevents the modification.

In such way the recruiting efficiency of the splicing machine was compared among the two strains: the wt and the mutated where the ubiquitylation was prevented.

The protein nuclear cap-binding complex small subunit (STO1) was genomically tagged due to its biologic function of recognizing the 5' end of the RNA and recalling the splicing machine proteins components.

If preventing the ubiquitylation in position K123 will results in a decrease of the splicing machine efficiency recruitment, a decrease of such related proteins would be observed in the sample carrying the mutation respect to the wild type (WT).

The analytical task

To validate the hypothesis just described, a mass spectrometry based quantitative method has been developed, in order to detect differential protein quantity recovered with TAP-tag.

In all the experiments three samples have been analyzed:

- **WT:** cell lysate of Saccharomyces *cerevisiae* with STO1 TAP-tagged submitted to AP-MS.
- **Mutant:** cell lysate of Saccharomyces *cerevisiae* with STO1 TAP-tagged carrying a genomic substitution K123Y submitted to AF.
- **Negative control:** cell lysate of Saccharomyces *cerevisiae* with no genomically TAP-tag construct, but submitted to AP-MS to evaluate eventual false positive results.

Explorative proteomics study

As explained in the introduction of the part II, no quantitative approach can be pointed out as capable of solving all the different analytical tasks. One of the imperative analytical goals was to have a flexible approach, eventually capable of being re-oriented in case of new evidences obtained by one of the three approaches used in the study: molecular biology, transcriptomics and proteomics. For such reason the targeted SRM approach could not be selected.

A first identification experiment was performed in order to evaluate:

1. **The sample complexity**, determined by how many proteins are present in the sample.
2. **Dynamic range**, defined as the concentration difference from the most and the less abundance proteins.

The samples were separated on a mini SDS PAGE gel, the gel lanes were systematically cut and analyzed by nanoLC-MS/MS on the Q-TOf maXis Bruker.

Sample complexity results: the number of proteins identified, deeply exceeded the expected one for a TAP-tag experiment directed against a protein known to have just a few partners. Almost 700 proteins were identified for the WT, whereas the sample carrying the mutation and the negative control resulted in a lower complexity (Table 2). To evaluate the overlap of the proteins identified in this preliminary experiment a Venn diagram is displayed in Figure 27. Globally, the analysis of the protein list, showed a relevant amount of contaminant proteins that can be categorized in three main classes: chaperone proteins, ribosome related proteins,

metabolic proteins. Such low specificity in terms of the TAP-tag purification was due to the need of preserving low strength interaction.

Dynamic range results: the number of peptide identified for each protein was considered as a quantitative index. The Cap binding proteins (subunit 1 and 2) were identified, as expected, in really high amount, due to TAP-tag directed against the subunit 1 and due to the high reciprocal affinity of the subunit 1 and 2. Any other protein was identified as high abundance.

The dynamic range results mitigate the analytical difficulties of developing a quantitative method on sample presenting a deep complexity. The eventual presence of extended dynamic range would have required, a not easy to set up, multi steps pre-fractionation (99).

Sample	N° of proteins identified
WT	692
Mutant	476
Negative control	229

Table 2: the identification results obtained for the three samples.

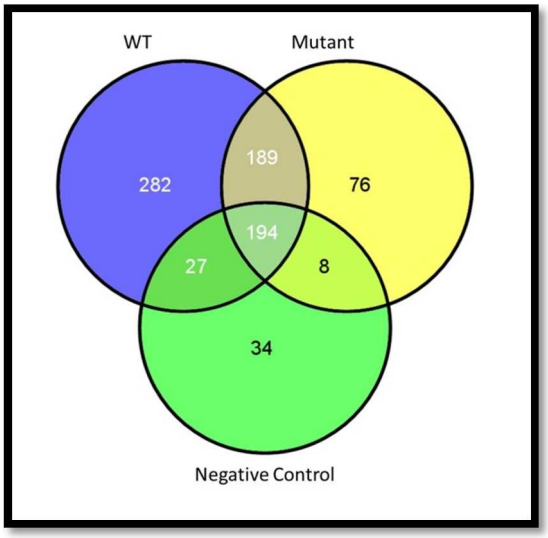


Figure 27: venny diagram presenting the proteins identified in common in the three samples. The 76 proteins identified just in the mutant were non specific; all the splicing related proteins were identified in WT and mutant. Any splicing related protein was identified.

Results of the explorative experiment

Despite a not completely suitable experiment design (no replicate was performed), quantitative information has been extracted from this data set. The number of peptides identified for each protein has been considered as an index of relative abundance.

Any inference should be done regarding the absolute abundance of a protein when considering the number of peptides identified. Since a low molecular weight protein, even if presents in high amount, would generate just few tryptic peptide, resulting in a low number of peptide identified.

The tryptic digestion generates peptide with different size depending upon the distribution of K and R, but peptide comprised in a the range of 5-30 amino acids are more likely to be identified by LC-MS/MS

The two proteins of the cap binding complex are an example of above expressed concept:

- Cbc2 is composed by just 208 amino acids and the in silico tryptic digestion generates just approximately 15 peptide detectable by LC-MS/MS.
- Sto1 is composed by 861 amino acids and the in silico tryptic digestion generates approximately 60 peptide detectable by LC-MS/MS.

Table 3: displays the results of the peptide counts obtained during characterization experiment. It also presents the number of peptides identified in the three samples belonging to: U1 and U2 complexes (splicing machines), the export machinery and the THO complex (promotes coupling between transcription and mRNA processing).

COMPLEX	PROTEIN	N° OF PEPTIDES WT	N° OF PEPTIDES MUTANT	NEGATIVE CONTROL
Cap binding complex	Cbc2	11	11	0
	Sto1	49	47	3
Export machinery	Mex67	13	10	0
	Mtr2	2	0	0
	Nab2	13	5	0
THO Complex	Hpr1	11	11	0
	Tho2	44	43	0
	Thp2	4	4	0
	Mft1	1	5	0
U1 snRNP	Prp42	19	10	0
	Prp40	16	7	0
	Prp39	16	13	0
	Luc7	5	3	0
	Mud1	11	1	0
	Nam8	5	2	0
	Snu56	10	2	0
	Snu71	15	10	0
	Yhc1	6	2	0
	Prp5	2	0	0
	Snp1	13	4	0
U2 snRNP	Lea1	5	1	0
	Msl1	2	1	0
	Prp9	6	0	0
	Prp21	2	0	0
	Cus1	4	0	0
	Rse1	4	0	0
	Prp11	1	1	0

Table 3: displays the results of the peptide counts obtained during characterization experiment.

Despite a not stringent wash during the TAP-tag, any of the proteins related to STO1 (aside cbc2) has been identified in the negative control, proving that high level of not specific protein detected did not affect the experiment.

The results of this explorative proteomics study were also used to optimize the wash condition of the TAP-tag in order to decrease the number of not specific proteins for the following experiments.

This early quantitative information was in agreement with the transcriptomic results: the WT proteome was found enriched in splicing components (UsnRNPs) respect to the mutant.

Quantitative study

Selection of the analytics approach

To improve the confidence of the quantitative information obtained in the preliminary peptides counts, two quantitative approaches have been tested:

- Method 1: SDS-PAGE pre-fractionation followed by systematic cut of the gel, digestion and LC-MS/MS based spectral counts.
- Method 2: Liquid digestion and 3 hours LC-MS/MS followed by the manual chromatogram extraction of the interesting features from the DDA-MS1.

A new set of samples were prepared performing a new TAP-tag purification taking into account the results of first set of data in terms of specificity of the purification.

At the time when this project was going on, no software was available in LSMBO to perform automatic DDA-MS1 label free quantification like Skyline (see Part II chapter V), so the combination of gel separation and manual chromatogram extraction of the parent mass MS1 has been discarded.

In Table 4 are presented the identification results obtained with the two methods.

	Method 1 (SDS-PAGE)		Method 2 (3hrs LC)	
	Total proteins identified	STO1 related proteins identified	Total proteins identified	STO1 related proteins identified
WT	403	27	217	5
Mutant	313	25	158	4
Negative control	149	0	80	0

Table 4: results of the methods comparison

The results obtained injecting the samples with a 3 hours LC gradient were not satisfying. As displayed in Table 4 the comparison of the two tests showed a loss of sensitivity inacceptable therefore the method 1 was selected for the further experiment.

It should not be neglected that gel based approach required 45 minutes gradient for each slice, resulting in a time analysis of 12 hours for each sample. The method 2 based on a liquid digestion, instead, required just three hours of time machine for each sample. Globally, a lower sample complexity was observed due to a stronger wash condition performed during the purification in this set of samples.

Development of the spectral counts based methods.

In the spectral counts method, a quantitative index is obtained summing all the spectra assigned to a protein. This technique represents a flexible widely used way of performing discovery proteomics and achieving quantitative information.

This approach is based on the concept that the peptides of a more abundant protein have more possibilities of being selected for CID fragmentation in DDA acquisition method, respect to peptide originated by a less abundance protein.

The robustness of the spectral counts results it increases proportionally with the number of spectra identified for each protein, number defined by:

1. **The number of peptides having size compatible** with LC-MS/MS generated by enzymatic digestion.
2. **The heterogeneity of the proteins in terms of modifications**, as example: identification of the same peptide carrying a cysteine carbamidomethylated and not, is a classic situation that allows to increase the spectral counts.
3. **Multiple state of charge detection of a peptide**: depending on the geometry of the source and on the competitive effect of the ionization process, peptides carrying missed cleavage, or highly abundant one, can be identified multiple times due to the different m/z detected by the mass spectrometer.
4. **The redundancy of the peptide selection**: in the DDA acquisition method to increase the possibilities of selecting the peptides less abundant, once a peptide is selected to be fragmented, its parent mass is excluded for a user defined time.
   In Figure 28 is displayed the effects of exclusion time respect to a chromatographic peak, this parameter needs to be carefully adjusted in order to maximize the redundancy without reducing too much the number of identified proteins.

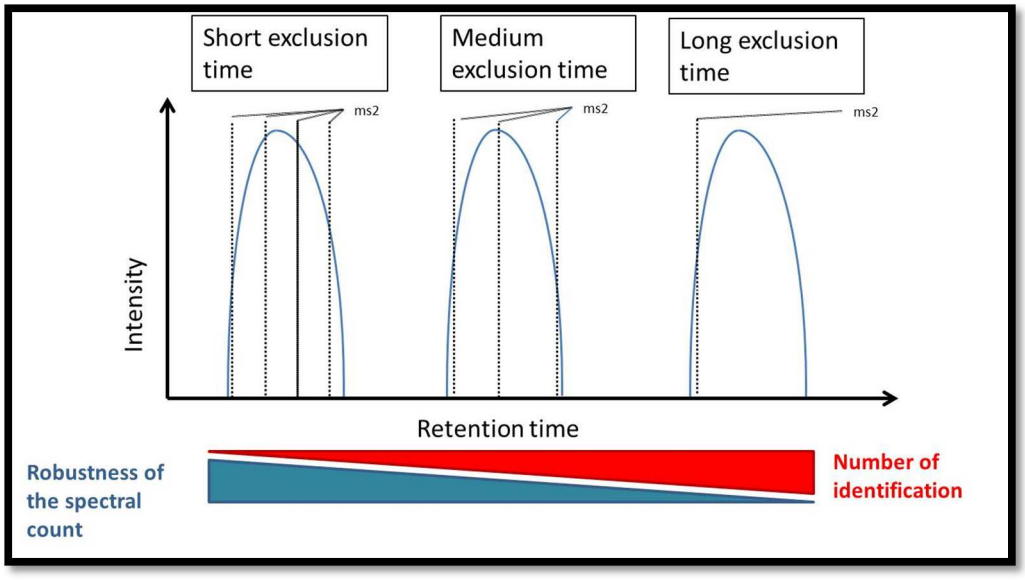


Figure 28: the effect of the dynamic exclusion time on the number of MS/MS spectra.

In Figure 29 are displayed the results obtained optimizing the exclusion time in a WT yeast model sample. These results showed that for a sample of medium complexity generating chromatographic peaks of 30 seconds at the base, the optimal exclusion time is 0.2 minutes, since it represents the compromise between sensitivity and redundancy.

The results obtained in this method development should not be considered as directly transferable to other project. Each coupling LC mass spectrometer presents peculiar parameters, like the scan speed, or the intensity threshold selection for the MS/MS, that require to be adjusted to the chromatographic performances and to sample characteristics described above.

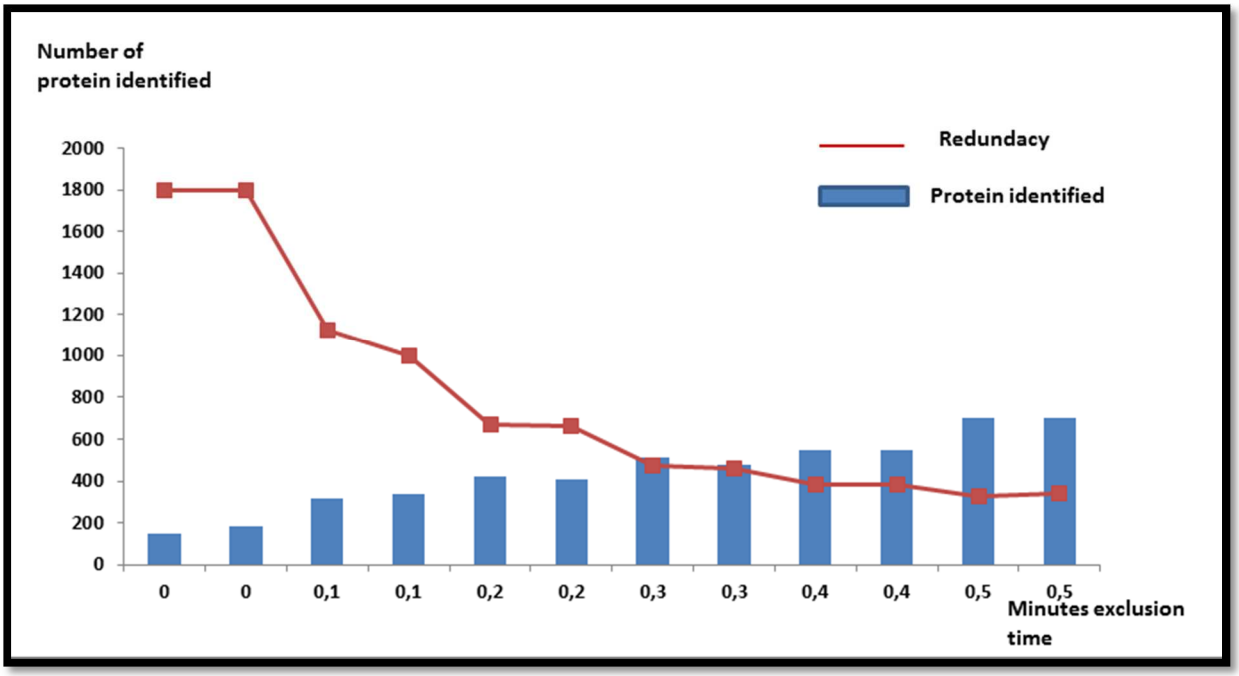


Figure 29: plot of the number of proteins identified and redundancy obtained injecting in duplicate 100 ng of yeast with different exclusion times. The blue histogram represents the number of proteins identified, the red plot represents the redundancy

Results

The optimized SDS-PAGE spectral counts method has been applied to a second series of TAP-tag purification. The results obtained during the explorative experiment helped the collaborator in adjusting the the affinity purification stringency of the wash. This resulted in the identification of roughly 50% less contaminant respect to the explorative study.

Table 5 shows the spectral counts results obtained for the subunits 1 and 2 of the nuclear cap binding protein (Cbc 1-2). The sample mutant presents a lower protein quantity, respect to the WT and a relevant difference is observed across the triplicate of injection for both samples.

During the immune-purification, the same amount of protein was loaded into the affinity column for the three samples. Assuming an equimolar amount of Cbc 1 in the three samples, a horizontal normalization was performed on this protein. The results of this normalization type (Table 5) strongly decreased the artificial variability between the two samples and the variability across the triplicate.

To evaluate the results of the normalization, the spectral counts of Cbc 2 was compared across the samples. Cbc2 is a not tagged protein, which exhibits a high affinity for Cbc1 and it is presents in a ratio 1/1 respect to Cbc1 in the cell.

	WT	WT	WT	MUT	MUT	MUT
	Injection 1	Injection 2	Injection 3	Injection 1	Injection 2	Injection 3
	not normalized N° of spectra			not normalized N° of spectra		
TAP-tagged Nuclear cap-binding protein subunit 1	401	429	377	251	263	269
Nuclear cap-binding protein subunit 2	119	152	125	92	95	100
	normalized N° of spectra			normalized N° of spectra		
TAP-taged Nuclear cap-binding protein subunit 1	429	429	429	429	429	429
Nuclear cap-binding protein subunit 2	157	155	159	152	127	142

Table 5: displays the effect of the horizontal normalization. On the top the not treated results presents a global lower amounts for the sample Mutant. Below the horizontal normalization performed on the subunit 1 can be appreciated on the stable amount of the subunit 2.

Complex	Protein	Number of spectra identified for WT	Number of spectra identified for Mutant	N° of spectra identified for negative control
U1 snRNP	Prp42	11	2	0
	Prp39	12	5	0
	Luc7	5	3	0
	Mud1	13	4	0
	Nam8	3	1	0
	Snu56	5	2	0
	Snu71	15	10	0
	Yhc1	3	2	0
U2 snRNP	Lea1	5	0	0
	Prp9	3	0	0
	Prp21	2	1	0
	Rse1	4	1	0
	Prp11	4	0	0

Table 6: displays the spectral counts results obtained with the optimized method.

Conclusion

A spectral count method was optimized and sucessfully used to detect major fold-change of low concentrated proteins in complex matrix. This optimization was based on a fine tuning of a series of parameters, among these the most important were the exclusion time and the redundancy of the spectral counts.

From a biologic point of view, the spectral counts combined with the transcriptomic study showed that preventing Ub-H2B ubiquitylation impaired the association of the nuclear export machinery, but maintained wild type levels of other factors known to bind mRNAs.

This experiment was carried out three years ago, with the best machine available at that time in LSMBO. Having the possibilities of re-performing these experiments today, with more performant mass spectrometers, would result in the selection of different analytical methods. Thanks to the increased MS/MS scan speed capabilities and new software, more and more quantitative information is obtained performing data independent acquisition (100) (part II chapter V) or MS1 (part II chapter III) permitting to generate more accurate and confident results.

Part II - Chapter II -

Targeted Selected Reaction Monitoring based Prion protein quantification

Intoduction

Chapter II part A

Sensitivity improvement of prion protein quantification in urine derivate fertility hormone

Chapter II part B

Method development for a relative quantification of two polymorphic variants of prion protein in human biopsies.

Part II - Chapter II - Targeted Selected Reaction Monitoring based Prion protein quantification

Introduction

Selected Reaction Monitoring (SRM) is a targeted quantitation method, which combines the Stable Isotope Dilution (SID), with an acquisition usually performed on a triple quadrupole (QqQ) mass spectrometer.

The SID approach has been developed in the early 30s. It was applied for the first time to mass spectrometry, in the 50s, coupled to Thermal Ionization to quantify small molecules. The SID is based on the spike of a known amount of the target molecule (the analyte) isotopically labelled. In such way the not quantitative ionization efficiency problem is circumvented.

In the context of proteomics, the target analyte to be quantified (light form) and the corresponding isotopically labelled standard (heavy form), are expected to exhibit the same physical and chemical properties; in terms of chromatographic or mass spectrometry behavior (ionization and fragmentation), at least when using stable isotopes such as ^{13}C, ^{15}N or ^{18}O as heavy isotopes. In all cases one should verify that the chromatographic behaviors (retention time, peak shape) are strictly identical, particularly when using deuterium. The difference in mass between the light and the heavy form should be large enough to be easily detected and without interference between the naturally occurring isotopical peaks (P+1, P+2,…P+n) of the light form and the molecular ions cluster from the heavy form. In general a mass difference of at least 6 Da is recommended. Then, an absolute or differential quantification can be made, measuring the ratio of corresponding ions (101, 102).

SRM is very often, but not always, performed on QqQ architecture mass spectrometer, where the first resolving quadrupole (Q1) selects a peptide parent ion, which is fragmented in a collision cell (q) to produce daughter ions, which will be measured in the second resolving quadrupole (Q3) that will select one by one the daughter ions of interest. For simplicity, this fragmentation process of a selected molecular ion leading to a selected fragment, is usually called “transition”.

In the SRM mode, no survey scan is performed. The QqQ instrument scans in cycles for all the transitions chosen in a step mode. The spectra acquired are displayed as a chromatogram, giving the intensity measured for each transition from the parent peptide molecular ion to each ion fragment alternatively selected by MS2. Figure 30 presents the schematic functioning of the QqQ.

QqQ architecture mass spectrometers are often used in SRM due to the duty cycle efficiency of the quadrupole, theoretical close to 100% (103), defined as the percentage of time during which ions are effectively transmitted analyzed and detected. Such physical characteristics are responsible for the superior sensitivity of the QqQ compared to other mass spectrometers architectures.

SRM approach has been used and mainly developed for the quantification of small pharmaceutical molecules during the 15 last years (104). It was often named Multiple Reaction Monitoring (MRM), but since its use in the context of proteomics because of minor methodological differences, the name SRM was preferred. SRM is now used in proteomics quantitative studies to perform accurate differential or absolute quantitation in different context

like biomarker validation (105-109) or in some cases even as alternative to ELISA assay (110-113).

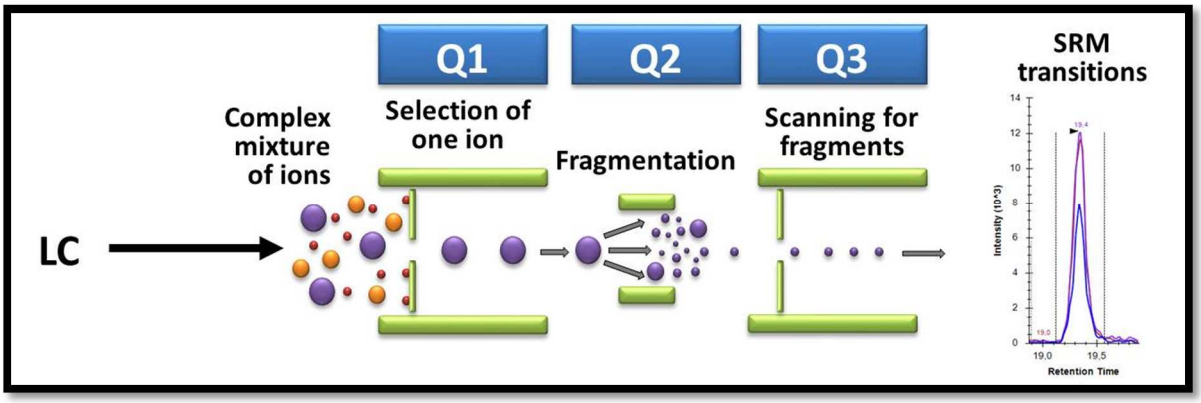


Figure 30: the schematic functioning of the QqQ in SRM.

Performing SRM on a QqQ grants the best sensitivity but the lack of resolution of the SRM remains a relevant problem (114). Even with the resolution improvement of the last QqQ mass spectrometer generation, the possibilities of interferences due to isobaric or close mass peptides is relevant. This is due to the fact that the number of digestion peptides obtained from a protein extract is usually so high (several hundred thousands) that there are always several different peptides eluted simultaneously from the chromatographic column in the mass spectrometer. Some of these peptides could have very close, if not identical molecular clusters, which will obviously induce interferences.

The quality and the robustness of the chromatographic system (stability of the retention time) and a critical validation process of each transition allow overcoming this limitation. Even though not yet considered as a high-throughput analytical method, SRM represents the gold standard of mass spectrometry based quantification (87).

General workflow of an SRM method development

Due to its targeted characteristics SRM requires more than other approaches a careful method development since, once the method is set up, the QqQ instrument will exclusively measure ions corresponding to the transitions selected and will be "blind" to all the others ions.

In Figure 31 are presented the main steps of an SRM set up experiment, which are detailed below:

Digestion peptides selection

- How many peptides for one protein?
  The quantification process is performed at the peptide level assuming a theoretical protein digestion efficiency of 100%. Monitoring multiples peptides for each protein, ideally distributed across the whole length of the protein sequence, allow possibly detecting region of the protein less accessible to the digestion enzyme or subjected to proteolysis and to discard such peptides results.

- Are all peptides univocally generated by the target protein?
  The digestion enzyme converts a protein in series of peptides of variable length. Due to the presence of conserved domains (even in proteins with different biologic function) some peptide sequence could be shared by many proteins.
  For such reason, it is advisable to select only prototypic peptides (115) (unique in the organisms studied and of size compatible with LC-MS/MS). This selection can be performed using software such as PeptideSieve (115) ESPredictor (116).
- Are all the forms of the target peptide monitored
  During the peptide selection, should also be avoided peptides carrying amino acids that could be modified by a PTM like: glycosylation or phosphorylation or oxidation etc. Due to the heterogeneity in terms of presence / absence of the PTM, the results could be distorted.

Selection of the transitions

- Global overview of all transitions
  Depending on the capabilities in terms of dwell time (milliseconds needed to scan a transition of the QqQ), a limited numbers of transitions can be monitored.
  Assuming a minimum of ten data points across a chromatographic peak, it is crucial to have a cycle time (duration required to acquire all the transitions once) close to 1/10 of the chromatographic peaks width.
  The most performant QqQ mass spectrometers, commercially available in 2014, have an optimal dwell time of 20-10ms, to which should be added the interscan (milliseconds required to change the voltage on the quadrupoles and to quench the residual ions of the previous fragmentation) of about 1-3 ms. Adding up these numbers, a maximum of 250 transitions can be monitored in cycle time of approximately 3 seconds.
- Best transition to be monitored
  Because the number of transitions which could be acquired is limited, the selection of the best transitions to be followed is a pivotal step of the SRM development.
  For each parent ion a series of products (daughter) ions to be measured is selected.
  The y_n ions having a higher m/z value than the parent ion should be privileged. This is due to the higher sequence specificity for such fragment ions (9, 117), taking also in account the fragmentation efficiency.
  When experimental data are not available, public repository like Peptide atlas, or predictive software like MRMaid (118) or TIQAM (119) can be used to select the best transition to be monitored.
  Guidelines are not yet available, but it is highly recommended to monitor three transitions for each peptide to increase the confidence of the results (85).

Validation of the transitions

In SRM, the validation of the measured transitions is often still manually performed and is based on the following criteria:

- The perfect coelution of heavy and light peptides. The molecular ion chromatograms of both the light and heavy peptides must present the same retention time
- The presence of a comparable peak shape across the different transitions. If two different endogenous peptides have very close molecular ion clusters (the series of

isotopic molecular ions) and a slight different in their retention times, the presence of this interference could be highlighted by non symmetrical chromatogram trace for the transition (shoulder).

- The presence of a comparable ratio between the intensity for the three transitions of the light form and for the three transition of the heavy form.

In the context of high-throughput SRM analysis the validation of the transitions can be performed using algorithm specifically developed like mProphet (120), capable scoring and validating the transitions.

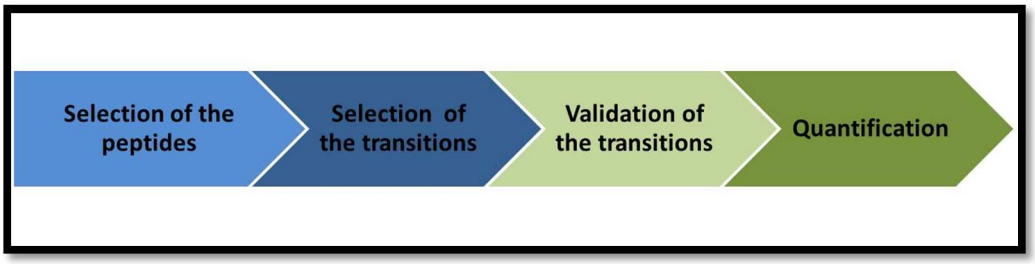


Figure 31: schematization of the SRM method development workflow.

Part II - Chapter II A -

Sensitivity improvement of prion protein quantification in urine derivate fertility hormone

Part II - Chapter II A - Sensitivity improvement of prion protein quantification in urine derivate fertility hormone

The biologic context

The fertility products commercially available are mixtures of proteins having hormone functions. They were for a long time purification products obtained from urine given by donors. During the two last decades they are more and more produced by biotechnology as recombinants proteins.

In the context of a collaboration with the pharmaceutical company Merck Serono, the LSMBO has developed in 2008-2010 a targeted SRM approach to detect and quantify human prion protein, in urine derived Injectable fertility products (121). This method was developed during a proteomic study on urine purified product where human prion protein was unexpectedly identified among different contaminants.

Because this first work showed that the results obtained by SRM were not sensitive enough, I have undergone in 2011 to improve this approach. My goal was to obtain an indisputable quantification of the prion protein on a series of urinary products for which the presence of Prion was not clearly established with the first method developed in the laboratory.

The Prion protein (PrP^{C}) is an endogenous cellular protein that, because of a conformational change, can be responsible for Transmissible Spongiform Encephalopathies (TSEs), a neurodegenerative disorders (122-125). TSEs have been described in the antique Greece as affecting sheep and were named scrapie. In TSEs the not infectious prion form has undergone misfolding, resulting in an increased content of β-sheet structure that alters the solubility and the tendency to aggregate. This misfolded pathogenic form of the PrP^{C} is named PrP^{SC}. It slowly aggregates in the brain and causes the Scrapie disease. PrP^{SC} is extremely resistant to proteolysis and heating, even at more than 150 degrees. A field contaminated by fecal matter and urine from scrapie sheep could induce the disease in healthy sheep after more than 10 years. The reason is that PrP^{SC} acts in a catalytic way to transform normal endogenous PrP^{C} in the pathognic PrP^{SC} form. It has been shown that within the same species (sheep, bovine, human), ingestion of $PrPS^{C}$ results after many years in the presence of PrP^{SC} in the brain. There are several cases of human to human transmission of prion related diseases by ingestion (Kuru). Iatrogenic infection is also described. For example, Creutzfeldt-Jakob disease (CJD) was transmitted by injection of growth hormone (hGH) extracted from cadaveric pituitary glands. In this case, a batch used to treat about 300 patients was infectious because of a single pituitary gland from a donor affected by CJD (out of several hundred glands). This was enough to infected at least 100 patients who died from CJD on a period of time of 25 years. Most of the transmission mechanism is still unknown. This illustrate that the catalytic process converting non-pathogenic PrP^{C} in pathogenic PrP^{SC} can occur during the purification process. Also, CJD was transmitted by cornea grafts (126). The delay for the development of this transmitted CJD could be very long (up to 30 year) (126).

In conclusion, from these well know and very media examples, it is clear that a batch of urinary fertility hormone corresponding to at least one thousand donors (400 000 litters of urine from postmenopausal women are used for one batch of u-hGH), has a significant rate of been infectious, because of the occurrence of CJD in the population is approximately $1/10^{6}$.

For this the reason, the laboratory has developed the SRM based method which should enable the detection of traces of prion protein in fertility products commercially available and extracted from urine. The method developed should detect any trace of prion protein which could have been co-purified with the therapeutic protein.

The analytical task

An analytical method is expected to detect and quantify a target molecule with known limitation. Therefore a Limit Of Detection (LOD) and a Limit Of Quantification (LOQ) must be determined. These limits are the result not only of the intrinsic instrumental sensitivity, but also depend on the matrix of the sample. This matrix is specific for each sample.

In the first study performed in the laboratory in 2010 (121), for two commercial gonadotropins urine derivate traces of prion were detected (above LOD), but could not be quantified (below LOQ) (121). I have therefore tried to optimize the SRM protocol and I have used a new generation instrument, expected to have a better sensitivity.

In this chapter, I will present in a first part the mass spectrometry method optimization performed in order to decrease the LOQ of the original SRM method. In a second part, I will present the results obtained applying this optimized method to the two samples in which the content of prion were detected, but not measured < LOQ.

Method 1 brief description

For each sample of this study, the matrix could be different. Developing an SRM method specific for a series of pharmaceutical products from different manufacturers is a challenging goal, since despite containing the therapeutic protein (for the same therapeutic indication) these products originating from different purification protocol can be deeply different in terms of:

- **Total protein content per vial:** injectable hormones are formulated and sold in vials containing a known amount of International Units (IU).This IU amount corresponds to the required biological effect and not to a fixed amount (in micrograms for example) of the active protein. This is because proteins purified from human extract and recombinant proteins could present differences in terms of folding and/or PTMs which could modulate the biological activity. Also, the level of contamination by other proteins, copurified with the therapeutic protein, is often very different as it will be illustrated by our study.
- **Small molecule content**: injectable products contain the active principle and a mixture of small molecules, called formulation, to ensure the stability and the adsorption of the drugs. The formulation is specific for each commercial product.

Such differences explain that the matrix (all proteins and/or small molecules which are not the therapeutic protein of interest) could be very different from one sample to the other. As a consequence, one can expect that the LOD and the LOQ can vary from one sample to the other, even using strictly the same SRM method. We have observed such a variation with a factor of 10. In Figure 32 is presented the workflow of the SRM prion quantification. All the steps are identical to the method described in reference 24. The only different concerns the mass spectrometer which was more recent and the instrumental parameter tuning.

Sample preparation. Bradford assays were performed in order to digest known amount of protein material and to inject the same amount of sample for all the samples. Before this assay, the sample was treated by ultrafiltration for excipient removal before analysis. The sample was denatured with urea, spiked with isotopically labelled standard peptides and submitted to tryptic digestion using a ratio protein/enzyme of 1/100.

SRM method. Figure 33 presents the amino acid sequence of the human prion protein ($hPrP^C$). The mature form is in bold letters. The peptides we have selected for SRM monitoring are underlined. The Table below presents for each selected peptide the transitions that we optimized. For each peptide (heavy and light form) the most intense transition was used for quantification and the other two transitions as qualifier. Selection of the peptides was performed considering the peptide identified in the first proteomics study (121) where the Prion protein was identified (discovery step). All transitions were optimized using the automatic software Peptide Optimizer tool.

Method for LOD determination. The following criteria were used to decide that a peptide was detected:

- Coelution and same peak shape for heavy and light peptides for molecular ions.
- Coelution and same peak shape of the three light transitions.

Due to the "unit" resolution of QqQ instrument, the presence of partially resolved interference is quite common and can be detected by unusual peak shape. For this reason the retention time and the peak shape of the heavy peptides is used as reference.

Method for LOQ determination. The following criteria were used to decide that a peptide can be quantified with confidence:

- Detection of transitions made with a signal over noise ratio >3, considering a window of 4 minutes across the transitions retention time.
- Intensity ratio of the three transitions conserved between the standard peptide (heavy form) and the endogenous peptide (light form) within ±30%.

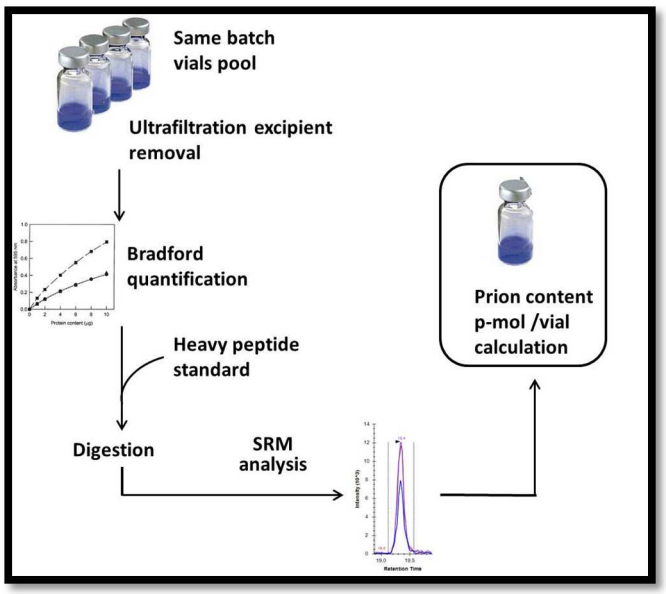


Figure 32: the schematic workflow.

```
        10         20         30         40         50         60
MANLGCWMLV LFVATWSDLG LCKKRPKPGG WNTGGSRYPG QGSPGGNRYP PQGGGGWGQP

        70         80         90        100        110        120
HGGGWGQPHG GGWGQPHGGG WGQPHGGGWG QGGGTHSQWN KPSKPKTNMK HMAGAAAAGA

       130        140        150        160        170        180
VVGGLGGYML GSAMSRPIIH FGSDYEDRYY RENMHRYPNQ VYYRPMDEYS NQNNFVHDC

       190        200        210        220        230        240
NITIKQHTVT TTTKGENFTE TDVKMMERVV EQMCITQYER ESQAYYQRGS SMVLFSSPPV

       250
ILLISFLIFL IVG
```

Peptides	Precursor ion m/z charge state	Quantifier charge state fragment state, Fragment type	Qualifier charge state fragment state	Qualifier charge state fragment state
Native 209-220	778.4 (2+)	696.4 (1+, y5)	1228.6 (1+, y9)	1105.5 (1+, y8)
Heavy 209-220	783.4 (2+)	706.4 (1+, y5)	1238.6 (1+, y9)	1115.5 (1+, y8)
Native 221-228	522.7 (2+)	692.3 (1+, y4)	466.3 (1+, y3)	700.3 (1+, y5)
Heavy 221-228	527.8 (2+)	693.4 (1+, y4)	476.3 (1+, y3)	710.4 (1+, y5)

Figure 33: the human prion protein, in bold the mature form. Underlined the two peptides 209-220 and 221-228 monitored in the SRM quantification. Below the Table presents the three transitions selected for each peptide.

Results obtained with method 1

In 2012, I first applied the method 1 to three urinary pharmaceutical products, using the Agilent 6410B QqQ mass spectrometer. Results are summarized below on Table 7. Peptide 221-228 was always quantified as less abundant than the 209-220. These results could be explained by a C-terminus degradation of the endogenous prion protein *in vivo*.

The quantification of the prion protein was therefore performed using peptide 209-220 and peptide 221-228 was considered as an additional detection criteria.

Products from manufacturers B and C did not satisfy our quantification criteria (LOQ).Sample B was characterized by a LOQ (p-mole/Vial) particularly high due to a strong matrix effect observed. The influence of the matrix on the LOQ is clearly illustrated by these results where exactly the same SRM method was used on the three samples.

Samples	Detection	Peptide 209-220 (VVEQMCITQYER)			Peptide 221-228 ESQAYYQR
		p-moles/Vial	% CV	LOQ p-mole/Vial	Detected
Product A batch 90302	Yes	5.56 Per vial	3%	0.34 Per vial	YES
Product B batch CE02714K	Yes	<LOQ	-	0.048 Per vial	NO
Product C batch100721	Yes	< LOQ	-	0.69 Per vial	NO

Table 7: summaries the results obtained with the method 1.

Development of the Method 2

Importance of the dwell time. The method 2 was developed on a TSQ Vantage QqQ Thermo, capable of acquiring a transition in a dwell time of 20 msec. This parameter allows increasing the number of transition that can be acquired by a factor of 4 compared with the QQQ 6410B from Agilent (used in method 1) that presents an optimal dwell time of 50-100 msec. In fact, in this specific project the multiplexing capacity of the machine were not at all exploited since only 24 transitions were acquired.

Importance of the quadrupoles resolution. The TSQ allows a fine tuning of the Full With at Half Maximum (FWHM) operated by the quadrupole during the selection (four widths are independently available for each quadrupole). Decreasing the width could be wrongly considered as responsible for a loss of sensitivity; since in term of absolute intensity a smaller isolation width results in a loss of signal. In method 1, due to a less performant quadrupole installed on the Agilent 6410B instrument, just three FWHM settings were available: unit: 0.7 u; wide: 1.2 u; widest: 2.5 u. After SRM optimization the analyses have been performed selecting wide settings on both Q1 and Q3 quadrupoles. To evaluate the possible benefit on specificity of a peak width reduction, sample A was analyzed on the TSQ instrument using differential width. Results are presented in the Table 8.

In order to calculate the signal to noise ratio, the noise has to be evaluated in the elution zone of each peptide. This is because the noise level is not constant during the elution. Because of that, a

manual evaluation of the noise in 4 minutes windows across the RT of the peaks has been performed, in order to calculate a signal over noise ratio for each peptide.

Starting from a width of 0.7 (the same used in method 1), the absolute signal decreases proportionally with the more stringent selection performed on the quadrupole. It is interesting to notice that the signal over noise ratio in the first 6 tests does not change significantly, but that when scanning with a width of 0.2 for both quadrupoles the signal to noise ratio was doubled.

These results showed that sensitivity can be increased thanks to the selectivity and that it is not only related to the absolute signal intensity.

Injection numbers	Width Q1	Width Q3	SIGNAL	Noise	Signal /noise 1
N01323	0.7	0.7	45000	500	90
N01319	0.7	0.4	25000	300	83
N01317	0.7	0.2	20000	200	100
N01324	0.4	0.4	17000	200	85
N01320	0.4	0.2	12000	150	80
N01321	0.4	0.1	8000	70	106
N01318	0.2	0.2	10000	50	200
N01322	0.1	0.1	5000	750	66

Table 8: test on the possible benefit on the specificity of a peak width reduction. The signal over noise is calculated for each combination of scans tested.

In Table 9, the transitions monitored are displayed. The optimization of the transitions was performed using the software skyline. Three transitions per peptide were monitored and acquired at 0.2 peak width for both the quadrupoles Q1 and Q3 in order to maximize the specificity. Dynamic SRM method setup was used with a dwell time of 10 msec for transition.

Peptides	Precursor ion m/z charge state	Quantifier charge state fragment state, Fragment type	Qualifier charge state fragment state	Qualifier charge state fragment state,
Native 209-220	778.4 (2+)	696.4 (1+, y5)	809.4 (1+, y6)	969.4 (1+, y7)
AQUA 209-220	783.4 (2+)	706.4 (1+, y5)	819.4 (1+, y6)	979.4 (1+, y7)
Native 221-228	522.7 (2+)	692.3 (1+, y4)	466.3 (1+, y3)	700.3 (1+, y5)
AQUA 221-228	527.8 (2+)	693.4 (1+, y4)	476.3 (1+, y3)	710.4(1+, y5)

Table 9: Heavy and light transitions monitored in the method 2.

Results obtained with Method 2

The three samples, originally analyzed with the method 1, were analyzed with the method 2 specifically developed for this peculiar type of sample acquiring the transitions using a width of 0.2 for both quadrupoles, MS1 and MS2.

Below are presented in Table 10 the results obtained. It is also displays the amount (picomoles) of prion obtained for sample A with method 1, as a comparison for the two experiments. It

appears that method 2 allowed a quantification of human prion protein thanks to the decreased LOQ.

A minor difference between the two results is observed probably due to small but inevitable error introduced in each step of the multi-step workflow requiring ultrafiltration and Bradford quantification in addition to the LC–SRM experiment itself.

Samples	Method used	Detection	Peptide 209-220 (VVEQMCITQYER)			Peptide 221-228 ESQAYYQR
			p-moles/Vial	% CV	LOQ p-mole/Vial	Detected
Product A batch n° 90302	1	Yes	5.56 Per vial	3%	0.34 Per vial	Yes
Product A batch n° 90302	2	Yes	7.13 Per vial	8%	0.15 Per vial	Yes
Product B batch CE02714K	2	Yes	0.036 Per vial	3%	0.032 Per vial	No
Product C Batch 100721	2	Yes	0.017 Per vial	18%	0.003 Per vial	No

Table 10: Quantification of peptide 209-220 from hPrP in different commercially available urinary fertility hormones

Conclusion

In this first part of chapter II, I have presented the results obtained with an optimization method specifically aiming at the decrease of the LOQ in a SRM experiments. The resolution of the 2 quadrupoles appears to be a key factor. This optimization enabled to decrease the LOD by a factor of 10, allowing quantifying the prion protein in drugs that were originally considered having prion protein content lower than the LOQ.

Part II - Chapter II B –

Method development for a relative quantification of two polymorphic variants of prion protein in human biopsies

Part II - Chapter II B - Method development for a relative quantification of two polymorphic variants of prion protein in human biopsies

The biologic context

Human prion diseases are linked to multiple diverse phenotypes. This name is referring to many differently categorized pathology (127), having in common the conversion of the normal prion protein, PrP^{C}, into the scrapie state PrP^{SC}.

The human prion protein exists in two sequence variants, due to a genetic polymorphism, differing for one amino acid residue: 129 Methionine and 129 Valine. This polymorphism results in three possible combinations: homozygote for one of the two genotypes or heterozygote.

This project was carried out collaboration with Prof. Gambetti of the USA National Prion Disease Pathology Surveillance Center. Prof Gambetti has hypothesized that a specific phenotype could linked with a preferential conversion of one of the two forms into the PrP^{SC}.

My contribution was to develop an SRM method capable of differentially quantify the PrP^{SC} 129M and PrP^{SC} 129V in prion protein purified from of human brain tissue.

The analytical task

This project presents an unusual difficulty: quantifying the relative abundance ratio of two proteins differing just by one amino acid (Figure 35).

The in silico digestion analysis reported a second problem: the tryptic digestion of the prion protein generates a 26 amino acids length peptide (110-136) which does not displays proteotypic properties (too long for informative fragmentation).

In addition this tryptic 26 amino acid peptide has a proline in the C-terminal position. The presence of this proline next to the tryptic cleavage site (position 136-137) is responsible for a diminished the cleavage efficiency. In case of missed cleavage in position 136-137 (since the next tryptic site is located in position 148), the peptide produced would be a 38 amino acids peptide (110-148). Such a 38 amino acid peptide is inadequate for an LC-MS/MS application.

The cleavage selection

After an in silico evaluation of the cleavage performed by the common enzyme and chemical reagents, the cyanogen bromide (CNBr) has been selected. This chemical way of generating peptide if often used in analysis of membrane associated protein lacking of tryptic site (128). CNBr reaction in acidic condition cleaves at the C-terminal site of not oxidized methionine, converting it in homoserine or homoserine lactone in C-terminal position. These two not conventional amino acids present respectively a mass difference of -30 Da and -48 Da with respect to the methionine. Figure 34 presents the cleavage reaction.

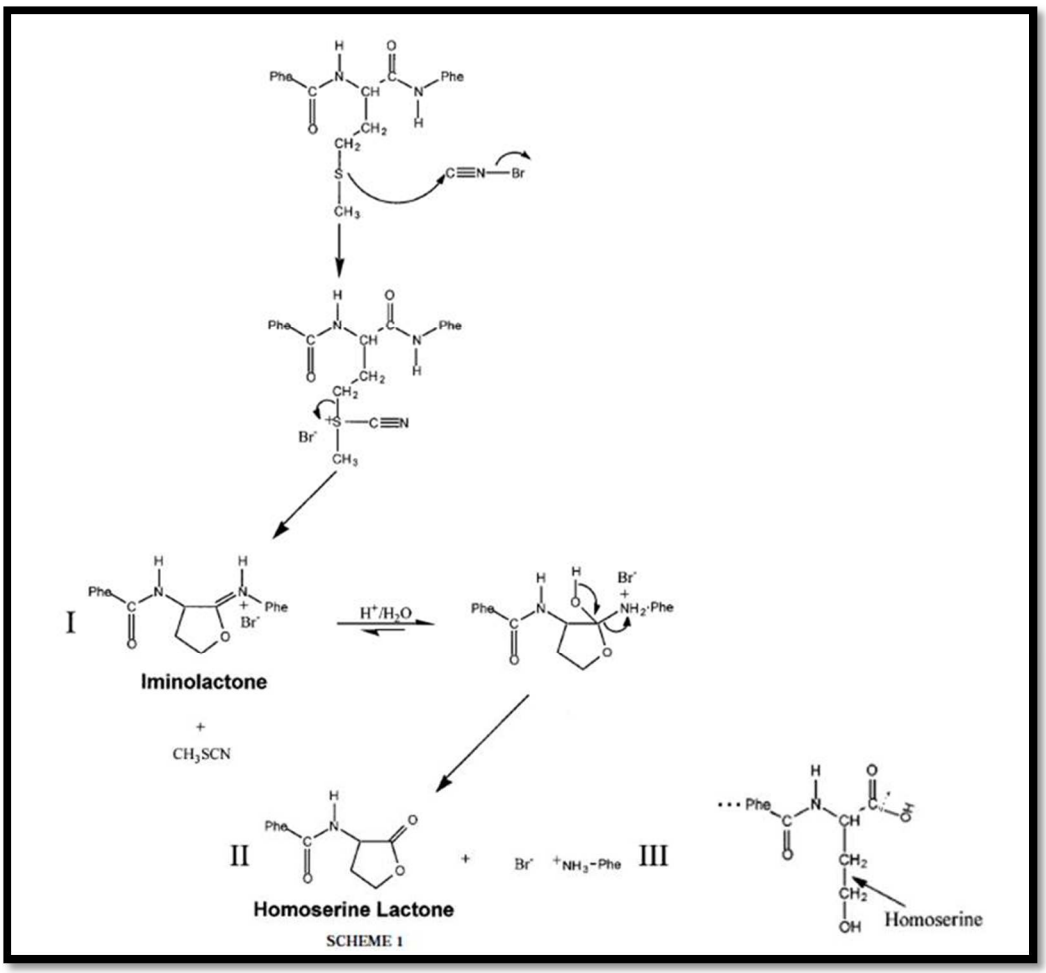


Figure 34: CNBr cleavage mechanism. Adapted from (129).

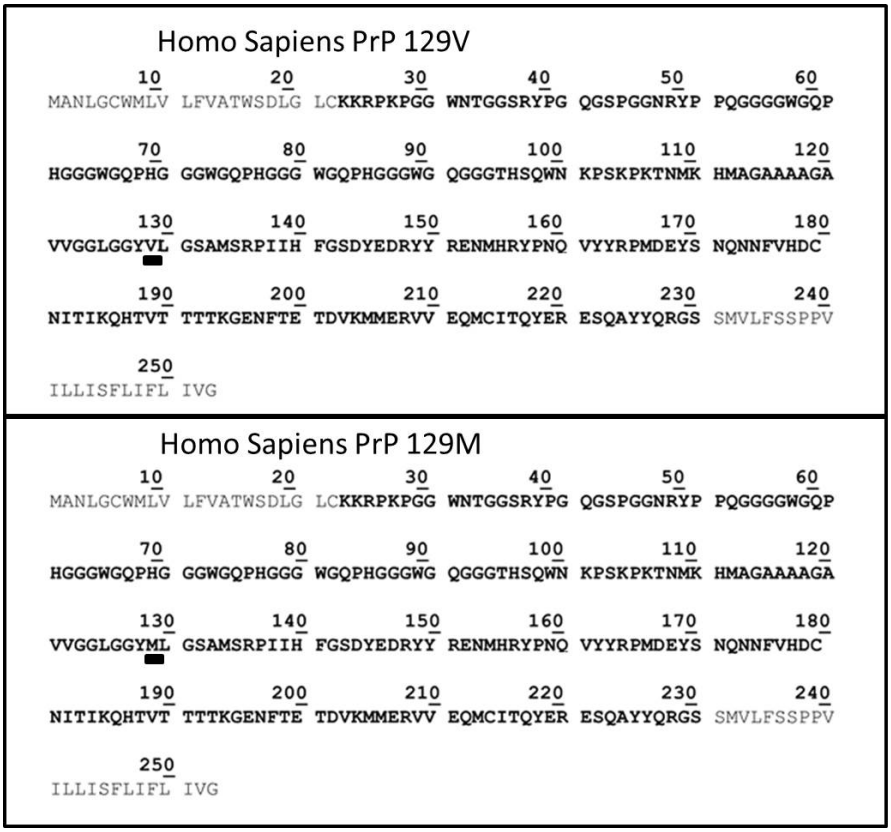


Figure 35: the sequences PrP 129M and PrP129V. Underlined in black the position 129 representing the unique discriminating amino acid.

In gel CNBr digestion optimization

In order to avoid artificial methionine oxidation during the sample shipment from USA, which would prevent subsequent CNBr cleavage, the samples have been separated in a SDS-PAGE to ensure the protein stability.

Model study for in gel CNBr cleavage. In the literature different in gel CNBr cleavage conditions are described. The CNBr is added with a variable molar excess (20 to 100 fold) with respect to the methionine residue. The acidic condition described may also vary substantially from 30% to 90% of HCl, FA or TFA (130-133). For optimization, we therefore tested these different conditions in a model study performed on two model proteins: cytochrome-C and myoglobin.

Protocol used for the optimization of the in gel CNBr cleavage. After "in gel" reduction and alkylation performed with 10 mM DTT and 55 mM, the 1D gel bands of cytochrome-C or myoglobin were incubated overnight at 5 C° with 200 mM CNBr dissolved respectively in 50% TFA and 30% FA to test the efficacy of the digestion in the presence of different acids. The reaction was quenched with 50 µl water and the peptides were extracted with 60% ACN in 0.1% FA. The peptides were concentrated and injected on a NanoAcquity (Waters) LC system coupled with a Q-TOF Maxis, (Bruker) mass spectrometer. We have observed, as expected, that the cleavage peptides are generated under two forms at the C-terminal position: homoserine or homoserine lactone. We tested the possibility of generate only one of the two forms to facilitate the quantification. We incubated the mixture of cleavage peptides after reaction with CNBr in basic or acidic medium to favor the formation of the lactone or of the free acid respectively as described in (134). These results could not be reproduced and the methionine cleavage peptides were identified by LC-MS/MS as converted mainly in lactone. Only a small percentage of acidic

form was detected. The acidic condition of the reverse phase might be responsible for restore the original ratio lactone acid.

Conclusion. The results obtained by the two CNBr cleavage conditions (TFA or FA) in term of protein coverage were equivalent: all the peptides having a size compatible with LC-MS/MS were identified. The digestion performed with TFA was selected since generated a higher number of spectra and did not produce adducts like in the case of the FA.

Selection of the isotopically labelled peptides for quantification

Prediction of peptides generated by CNBr cleavage. Usually the quantification of a protein using the SRM strategy of quantification o one digest peptide is relatively easy. A proteotypic peptide has to be selected with known criteria (described in general workflow of an SRM method development). In the case of this study, were peptides will be generated by a chemical cleavage by CNBr, the selection of peptides to be quantified by LMC-SRM is made very complex for the following reasons:

- The oxydation of Met prevent the CNBr cleavage. This open the door to the generation of a series of possible peptides containing the 129Met (see Table 11 below). Quantification should take in account all these peptide forms containing 129Met. Indeed, in the sequence of the prion protein 129M, in the 25 amino acids across the position 129 four methionines are presents. Assuming up to 2 missed cleavages due to possible methionine oxidation and taking in account the generation of homoserin lacton (hsl) and homoserin (hse), 16 different peptides could be generated by the CNBr digestion.
- The polymorphism of interest is Met/Val and therefore the Val containing will not be cleaved at position 129. The protein with 129V carries three methionine residues resulting in six possible peptides containing 129V when considering all the possible combinations.

Table 11 summarizes all the peptides which could be predicted from CNBr cleavage taking in account the fact that four Met could be oxidized. (3 first columns, left). Another degree of complexity could be introduced by the possibility of generating for each peptide an homoserin or an homoserin lactone.

Protein	Peptide	Peptide sequence	Met oxidized	Missed cleavage
129M	Peptide 113-129	AGAAAAGAVVGGLGGY**M**	0	0
129M	Peptide 110-129	KHMoxAGAAAAGAVVGGLGGY**M**	1	1
129M	Peptide 110-129	KHMAGAAAAGAVVGGLGGY**M**	0	1
129M	Peptide 113-134	AGAAAAGAVVGGLGGY**M**oxLGSAM	1	1
129M	Peptide 113-134	AGAAAAGAVVGGLGGY**M**LGSAM	0	1
129M	Peptide 110-134	KHMoxAGAAAAGAVVGGLGGY**M**oxLGSAM	2	2
129M	Peptide 110-134	KHMAGAAAAGAVVGGLGGY**M**oxLGSAM	1	2
129M	Peptide 110-134	KHMoxAGAAAAGAVVGGLGGY**M**LGSAM	1	2
129 V	Peptide 110-134	KHMoxAGAAAAGAVVGGLGGY**V**LGSAM	1	1
129 V	Peptide 110-134	KHMAGAAAAGAVVGGLGGY**V**LGSAM	0	1
129 V	Peptide 113-134	AGAAAAGAVVGGLGGY**V**LGSAM	0	0

Table 11: all the possible CNBr generated peptides carrying the amino acid 129. For each peptide is indicated the number of methionine oxidized and the number of missed cleavages.

First study on Prion preparation from CJD Human brain

At this level of the model study it was necessary to test our approach on a real sample. We were provided by Prof. Gambetti with a purified sample of Prion protein from an of human brain patient who died from CJD. The purification protocol was based on ultracentrifugation sedimentation in sucrose and tube SDS gel. Then multiple fractions were collected and tested for prion content with antibody. Fractions positive to the antibody were pooled. The pooled sample was loaded on 1D SDS-PAGE and revealed with Coomassie bleu. The band around 27 kDa was extruded and sent to our laboratory for analysis.

The in gel CNBr cleavage was performed accordingly to the method presented above. A nano LC-MS/MS was then performed on a MaXis (Bruker) in order to identify the different peptides. Interpretations of the data were made with Mascot against the SwissProt generated human database. All identified peptides identified corresponded to the human prion sequence which was an indication of good purity for the sample. Since only one protein is identified (Prion protein), no Decoy approach could be performed.

In Table 12 are presented our results obtained after in gel CNBr digestion of a Prion protein affinity-purified from human brain tissue.

Protein	Peptide	Peptide sequence	Met oxy	Missed cleavage	Met converted in	Number of spectra identified in Human brain biopsy
129M	Peptide 113-129	AGAAAAGAVVGGLGGY**M**	0	0	hsl	32
					hse	7
129M	Peptide 110-129	KHMoxAGAAAAGAVVGGLGGY**M**	1	1	hsl	5
					hse	0
129M	Peptide 110-129	KHMAGAAAAGAVVGGLGGY**M**	0	1	hsl	3
					hse	0
129M	Peptide 113-134	AGAAAAGAVVGGLGGY**M**oxLGSAM	1	1	hsl	6
					hse	0
129M	Peptide 113-134	AGAAAAGAVVGGLGGY**M**LGSAM	0	1	hsl	4
					hse	0
129M	Peptide 110-134	KHMoxAGAAAAGAVVGGLGGY**M**oxLGSAM	2	2	hsl	0
					hse	0
129M	Peptide 110-134	KHMAGAAAAGAVVGGLGGY**M**oxLGSAM	1	2	hsl	0
					hse	0
129M	Peptide 110-134	KHMoxAGAAAAGAVVGGLGGY**M**LGSAM	1	2	hsl	0
					hse	0
129 V	Peptide 110-134	KHMoxAGAAAAGAVVGGLGGY**V**LGSAM	1	1	hsl	0
					hse	0
129 V	Peptide 110-134	KHMAGAAAAGAVVGGLGGY**V**LGSAM	0	1	hsl	0
					hse	0
129 V	Peptide 113-134	AGAAAAGAVVGGLGGY**V**LGSAM	0	0	hsl	7
					hse	1

Table 12: results obtained after in gel CNBr digestion of a Prion protein affinity-purified from human brain tissue. All the possible CNBr generated peptides carrying the amino acid 129M/V are displayed. For each peptide is indicated the number of methionine oxidized, the number of missed cleavages and the number of spectra identified. Also is mentioned for each peptide the amount of homeserin and homeserin lacton (evaluated by spectral counting).

Discussion. Both proteins 129M and 129V were identified and in both cases the highest number of spectra was obtained for the peptides not presenting missed cleavage and terminating with homoserine lacton like expected.

The discriminant peptide of PrP 129M was identified with 1 missed cleavage in position 113 or 129 (peptide 110-129 and 113-134), the peptide with 2 missed cleavages, 110-134, was not detected. Roughly 70% of the missed cleavages were due to a methionine oxidized and 30% to incomplete reaction despite the relevant fold excess of CNBr used.

The discriminant peptide of PrP129V was only identified with the form 113-134 with no missed cleavage and mainly in the lactone form.

The result of this experiment has been used to select the heavy isotopically labelled peptide that will be used in the SRM quantification. The two peptides from the protein 129M and 129V

without missed cleavage has been ordered in both form lactone and acid since they represent they expected form and were identified as major form.

To increase the robustness of the quantification were also ordered all the synthetic peptides with one missed cleavage but just in the lactone form displayed in Table 12, due to the identification as minor form.

In Figure 36 two MS/MS spectra of the peptide 113-129 of the protein PrP 129M are presented. On the top is displayed the spectrum of the lactone form, below the spectrum obtained by the peptide with the opened ring. In both cases the b_n series is predominant respect to the y_n series. Despite the absence of basic amino acid on the N-terminal capable of accepting the mobile charge, the fragmentation quality is fine.

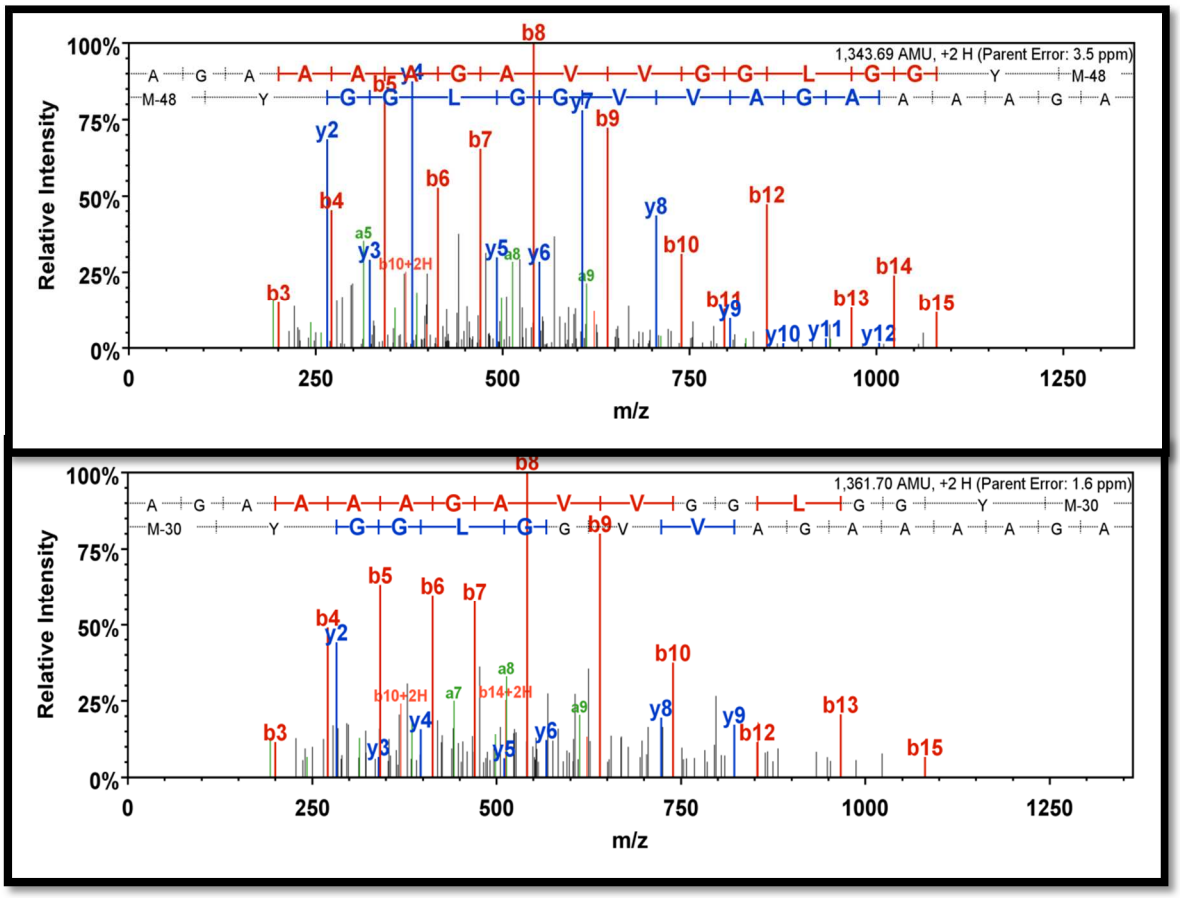


Figure 36: two MS/MS spectra acquired of the protein PrP129M. On the top is displayed the spectrum generated by the peptide 113-129 with Homoserin lactone. Below is presented the spectrum generated by the peptide 113-129 with Homoserin.

Conclusion

In this chapter it has been presented a method development for the quantification of two proteins not presenting any tryptic proteotypic peptide and discriminated by just one amino acid. For this reason I have performed a model study for in gel CNBr cleavage using cytochrome-C and myoglobin. The method developed was applied on a Prion protein sample purified from human brain tissue. The results allowed selecting the heavy isotopically labelled peptides that will be used in the SRM quantification, such experiments will be carried out after the redaction of my thesis.

Thanks to the availability of the heavy labelled peptide PrP 209-220, used for the prion protein quantification in urine derivate fertility hormone, it could be tested a double digestion performing CNBr and then tryptic digestion. This would lead to the possibilities of performing the differential quantification between PrP 129M and 129V, but would also allow differentially quantifying, across the samples, the tryptic peptide 209-220 common to both the proteins.

In this way the quantification on the tryptic peptide would be representative of both the proteins and could be used as additional level of confidence to rule out bias in the 129M/129V ratio due to extensive missed cleavage.

Part II - Chapter III -

Monoclonal IgG glycation batch to batch quantification

Part II - Chapter III - Monoclonal IgG glycation batch to batch quantification

Introduction

Luis Camille Maillard described for the first time the glycation process in 1912. This chemical reaction between proteins and sugar (Figure 37) is responsible for the characteristics browning process of many cooked dishes.

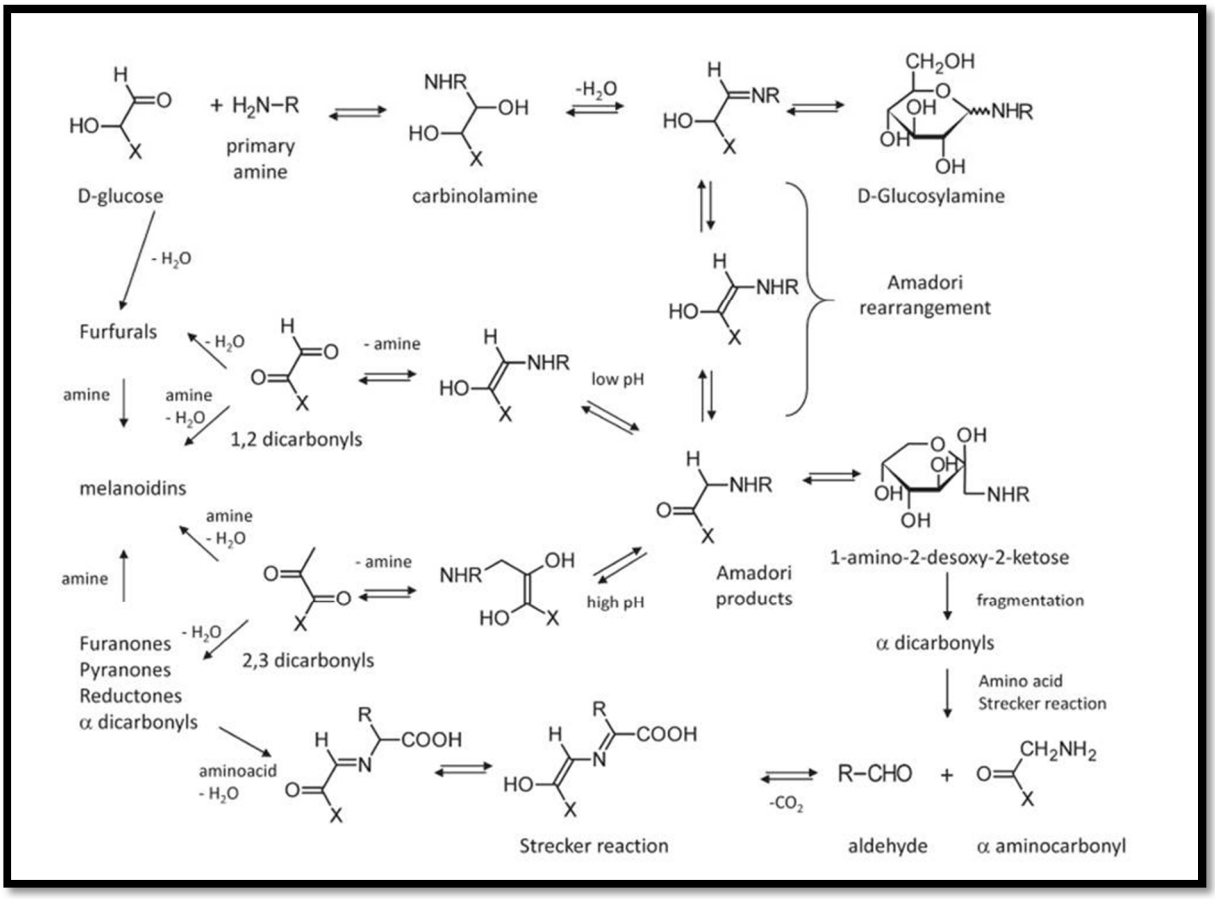


Figure 37: schematic representation of Maillard reaction following the Hodge scheme [5]. Adapted from [1].

Protein glycation is a non enzymatic post-translational modification, resulting of the condensation of a reducing sugar such as glucose and the amino groups of: lysine, arginine or with the protein N-termini.

This glycation reaction results in the formation of an unstable carbinolamine (Schiff base), that will spontaneously rearrange in a stable ketoamine, also called an Amadori product. The latter compounds are then degraded through various pathways leading to the formation of the advanced glycation products.

On the contrary of the glycosylation, glycation has no essential or accessory function in the cell homeostasis. Glycation results in an impaired biologic function of the modified protein by alteration of the pI of the protein solubility and 3D structure (135). The formation of AGEs (Advanced Glycation End products) in living system has been described by Kunkell and Vallenius in 1955 (136), when they discovered the glycated hemoglobin (HbA1c) in Human blood.

The quantification of the glycated hemoglobin, is currently a diagnostic test, used as long term monitor of the glycemic index in patient affected by diabetes (137). AGEs are also immunogenic, and showed to be able to trigger IgG response in pathologies such as alcoholic liver disease, diabetes mellitus and rheumatoid arthritis (138).

Glycated proteins analysis is still challenging due to the complexity of the protein-carbohydrate products that results in the modification of arginine and lysine in dozens of different structures (135, 139-144).

The analytical task

This project was carried out in collaboration with the pharmaceutical industry Merck Serono.

Recombinant proteins are produced in medium containing glucose (a reducing sugar) and are therefore exposed at the glycation process. Monitoring this degradation is crucial since glycated IgG might trigger immunogenic response in the patient and might have an altered half-life (145-150).

Due to the chemical heterogeneity of the AGEs, no derivatization process is capable of targeting all the products in one experiment. Recently, ELISA test are commercially available (151, 152) but no guarantee is given about the equal specificity of the polyclonal antibodys against all the AGEs products.

In Merck Serono, it has been developed a LC-UV quantification method based on a Boronate Affinity Column (BAC) separation. The boronate column has a specific affinity for ligands exposing a cis-diol (exposed by some glycation and glycosylation products). The capture is performed thanks to the formation of cyclic esters between boronic-acid ligands and the cis-diol analytes under alkaline conditions, (Figure 38) (153).

Figure 38: representation of the chemical reaction of the boronic acid group with cis-diol used in the enrichment step

The quantification method is based on the boronate separation of the monoclonal IgG in two fractions: the glycated and glycosylated fraction is retained on the BAC column, the non-modified fraction is eluted in the flow-through. The quantification is performed integrating the area under the curve detected with UV at 220 nm.

Since not all the glycated products contain a cis-diol, the BAC quantitative method presents the limitation of not targeting the complete population of the AGEs products. In Figure 39, Figure 40 and Figure 41 are presented the glycation products specific for arginine and lysine, whose carrying a cis-diol are highlighted with a blue star.

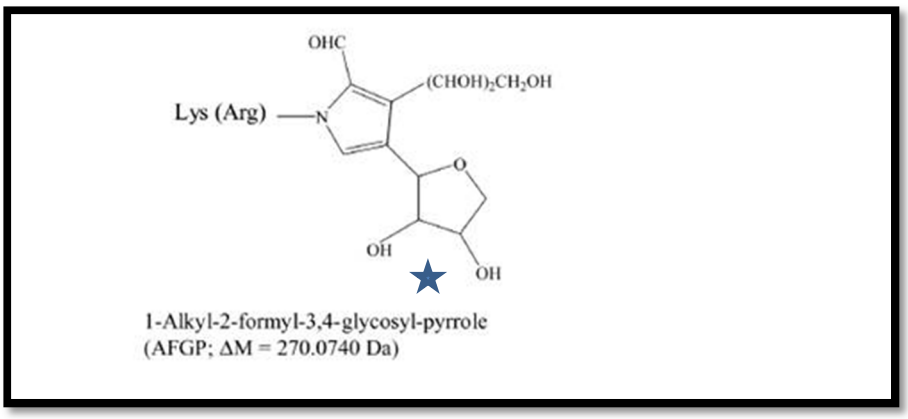


Figure 39: glycation modification specific of arginine and lysine adapted from [7]. The glycation products carrying a cis-diol are highlighted with a blue star.

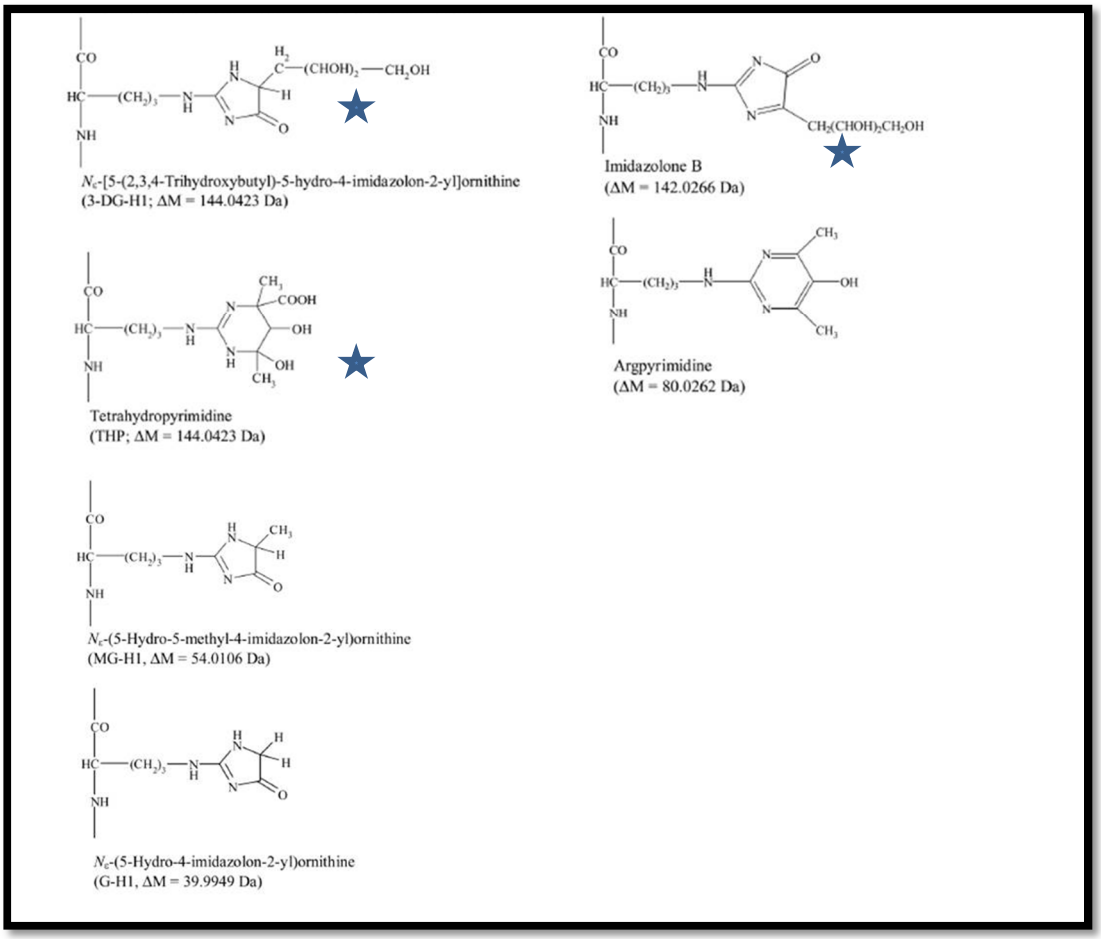


Figure 40: glycation modification specific of arginine adapted from [7]. The glycation products carrying a cis-diol are highlighted with a blue star.

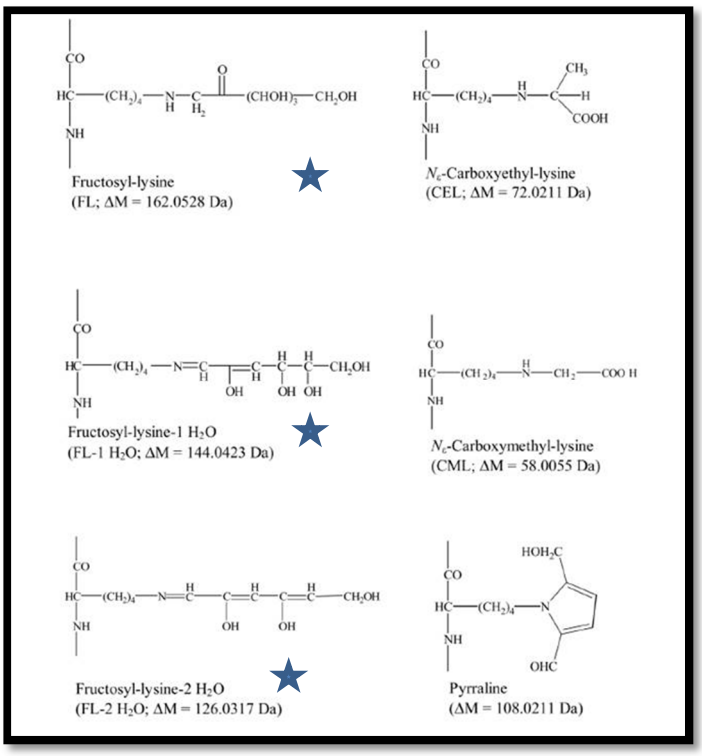


Figure 41: Glycation modification specific of lysine adapted from [7]. The glycation products carrying a cis-diol are highlighted with a blue star.

The goal of this project was to develop a method capable of:

- Identify of all the glycation sites and characterizing all the glycated structures.
- Measuring differences in the glycation levels of different monoclonal IgG batches.

Mass spectrometry based proteomics represents a valid approach to characterize and quantify AGEs due to the high ΔM carried by such modifications that can be easily detected.

Achiving the necessary sensitivity to quantify the glycation represented a challenging task, since the glycated fraction quantified with the BAC method represents less than 1% in the reference IgG. Should not be negletted that the presence of sugar reisdue linked to the peptidic backbone can decrease the ionization efficency.

I will present in the first part the initial method development performed on a reference standard batch of a monoclonal antibody produced by Merck Serono. In the second part I wil present the application of this method to other four batches in order to evaluate the presence of different glycation patterns and eventually quantify it.

The analytical method development

Denaturation and digestion optimization step

The optimization of the enzymatic digestion condition was a crucial step in this project, since a not complete denaturation would have impaired a completely efficient enzymatic digestion. Such scenario would have precluded the possibility of detecting and quantifying glycated modifications that affect just 1% of the reference IgG.

Three different denaturation conditions were tested:

- Chaotropic agents, like urea and thiourea, allow disrupting the hydrogen bonds, dipole-dipole interactions and hydrophobic interactions, thereby facilitating protein denaturation.
- Sodium dodecyl sulfate SDS is an anionic detergent that allows breaking even ionic interaction.
- The RapiGest™ is also an ionic detergent.

RapiGest™ presents the advantage of being cleavable in acidic condition and therefore easily removable. On the contrary, chaotropic agents need to be removed through solid phase extraction to avoid ionization suppression effect.

Protocols to remove SDS prior to mass spectrometry analysis are described in literature (154), but such results were never reproduced in our hands. Therefore, we used a gel stacking method which consists in performing a partial migration in SDS PAGE without protein pre-fractionation. In a stacking gel proteins are concentrated into a tight band which limits the number of analysis for quantification.

In Figure 42, is presented the image of the gel stacking performed for the control mAb batch. In this approach the gel preparation is the same of a classic SDS-PAGE gel, but the migration is stopped as soon as the front migration passes from the concentration gel to the separation gel.

Figure 42: Gel stacking of the mAb batch reference.

To evaluate the different denaturation conditions tested the parameters considered were:

- The achievement of the complete enzymatic digestion was considered as the first criteria. A manual inspection of the late part of the chromatographic gradient was performed to evaluate the presence of polypeptide containing many missed cleavage.
- The sequence coverage was evaluated to ensure that the sample cleanup process or the peptide extraction, were not responsible for the retaining of peptides with particularly hydrophobic or hydrophilic characteristics.

In Table 13 are summarized the results obtained for each of the condition tested in terms of sequence coverage, eventual detection of not completely digested polypeptide and the detection of glycated MS/MS spectra.

Condition	Sequence coverage %		Complete digestion	Detection of glycated MS/MS spectra
	Heavy chain	Low chain		
Liquid digestion 8M urea	90%	87%	Yes	Yes
Liquid digestion 6M urea 2M thiourea	90%	87%	Yes	Yes
Liquid digestion Rapigest™	90%	87%	No	No
SDS Gel stacking	96%	100%	Yes	Yes

Table 13: presents the denaturation test. Each experiment has been performed in duplicate and the results are the average of these duplicates

The Rapigest™ based protocol, despite an enzyme protein ratio 1/20, did not allow obtaining a complete digestion. In the frame of a classic peptide mapping of a purified recombinant protein this would have not be a problem, but in this project the complete digestion was required in order to detect low abundant glycated peptides. In Figure 43 is presented the chromatograms overlay of the tryptic digest obtained with the RapiGest™ protocol (in blue) and the digest obtained with 6M urea and 2M thiourea (in yellow). Globally, the signal of the two BPCs are similar, suggesting that no loss is observed due to the additional solid phase extraction steps performed on the chaotropic denatured sample. On the contrary, the chromatogram trace of the sample treated with RapiGest™ presents two peaks at 13 minutes and one peak at 22.5 minutes less intense than the peaks generated by the chaotropic denatured sample.

The results obtained for the denaturation conditions with urea and with gel stacking are quite similar, but the gel stacking followed by in gel digestion raised to an higher sequence coverage and moreover to an higher number of spectra assigned to glycated peptides.

These test showed that for the detection of glycated peptide the SDS-PAGE stacking represents the best option and for such reason this approach was selected for the next method development phase.

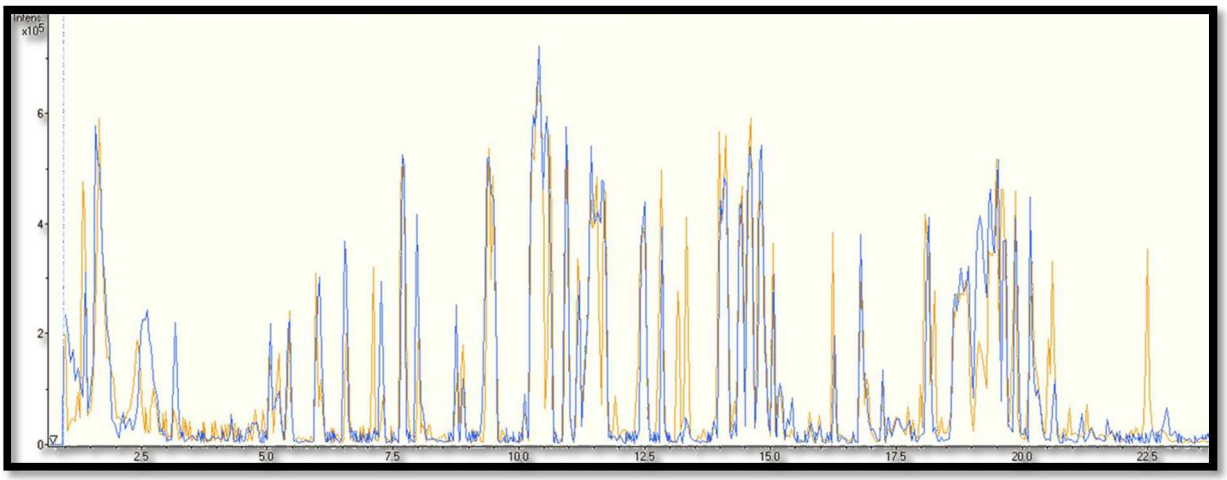


Figure 43: overlay of the chromatogram BPC tryptic digest obtained with the Rapigest™ protocol, in blue and the digest obtained with 6M urea 2M thiourea in yellow.

Selection of proteolytic enzymes

Trypsin is the most widely used enzymes in proteomics, due to the specific cleavage operated after arginine and lysine residues. The distribution of Arg and Lys in the proteome ensures to generate an average peptide size of ten amino acids (155). The cleavage performed at the C-terminal of the only two positively charged amino acids, results in a superior ionization efficiency and in a optimal fragmentation behavior in CID (156).

The glycation process specifically affects arginine and lysine residues, exactly the same amino acids recognized by the trypsin enzyme. For this reason, trypsin might not represent the first choice as digestion enzymes, since missed cleavage are expected for all the glycated peptides.

To maximize the detection of the glycated peptide in parallel to trypsin, alternative digestion enzymes were tested:

- Chymotrypsin
- AspN
- GluC
- Pepsin

To assess the results, at this early point of the method development, all the spectra identified by mascot with ions score >20 were considered as positive match. Due to the presence of only two proteins in the sample the decoy approach could not be used. More stringent criteria, detailed later on, for the spectra validation were used once the method was set up.

The results obtained during the enzymes test were unexpected. No glycated peptides were detected after digestion with AspN, GluC, or Pepsin and only six MS/MS spectra of glycated peptides, were identified with chymotrypsin. Indeed, the best results were produced by the trypsin digestion with 10 glycated peptides spectra.

In Figure 44 is presented one MS/MS spectrum obtained with trypsin digestion, the peptide was identified, with a good MS/MS quality, carrying a carboxymethyl (+58) on the lysine in C-terminal.

This spectrum particularly retained my attention, because such peptide should not have been cleaved by the trypsin, since the N-terminal Lys is modified in a carboxymethyl lysine.

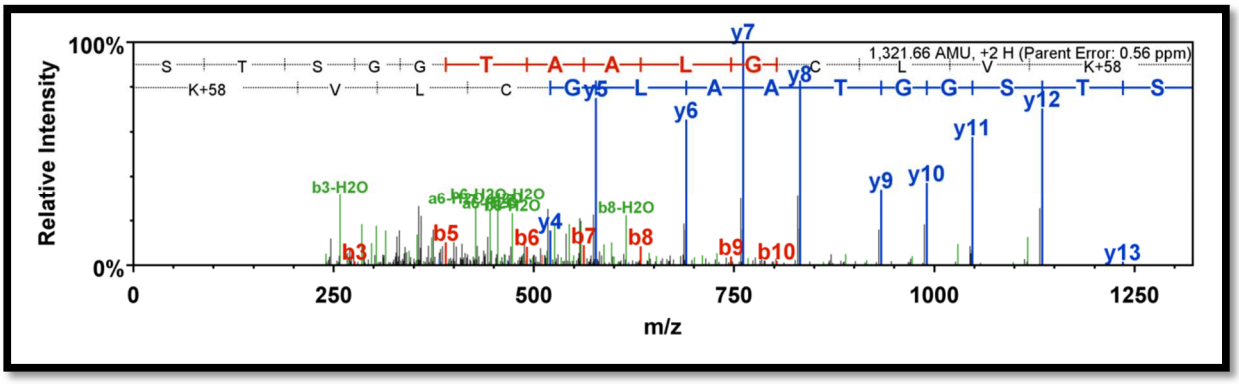


Figure 44: MS/MS spectrum of a peptide carrying a carboxymethyl lysine.

An unspecific behavior of the trypsin has already been described by Aebersold and coworkers in 2007 (156). They have performed a study on the specificity of the trypsin digestion using model proteins. In this paper, they showed that using a targeted approach, despite different denaturation conditions and enzyme ratios, 117 unique peptides are generated by the tryptic digestion of the Lactoglobulin. This small protein is composed of 162 amino acids and if submitted to in silico digestion only 12 peptides with full tryptic specificity are generated.

Test on the trypsin digestion specificity

During the proteomic analysis of proteins complex mixture, in order to reduce the possibilities of a false identification, the specificity of the protease used, is added as criteria for the search engine process. Therefore, peptides generated by an unspecific cleavage are *a priori* not identified in a classical database search.

The data set obtained with the digestion enzymes test, was re-submitted to search with Mascot, without adding enzymatic specificity. The results obtained with such parameters allowed the increase of the identified spectra carrying a glycation of a factor 100.

Clearly, among the hundreds spectra identified, the majority resulted to be originated by not enough informative and low signal over noise MS/MS, or false positive identification.

Due to the lack of reproducibility of the manual inspection of hundreds of spectra, I selected a series of criteria allowing the pre-filter the spectra to be validated:

- Identification of a modified peptide with a Mascot ion score >20. Using such filter low qualities spectra unlikely to be validated were eliminated.
- Identification of the same peptide carrying the same modification (glycation, oxidation, or carbamidomethylation) in at least two LC-MS/MS analysis (samples were injected in triplicate). The repeated identification of the same peptide highlights the peptides presenting intensity sufficient to be quantified.
- Δ retention time maximum of 30 seconds, for the identification of the same peptide in two LC-MS/MS analyses. This additional criteria rule out false positive identification.

Applying these filters, the numbers of spectra referring to glycated peptides to be manually inspected were drastically reduced at 36 for the trypsin digestion and 6 for the chymotrypsin digestion.

The manual inspection of the spectra leaded to a validation rate of approximately 50%. Clearly the trypsin digestion combined with the gel stacking gave the best results and for such reason this workflow was selected to be used in the quantification method.

Discussion

Thanks to this test, it was possible to show that when analyzing not complex sample also peptides generated by unspecific cleavage, if manually inspected and validated, are a resourceful source of data. Among the 36 glycated peptides, produced by trypsin, just seven were fully tryptic, one completely unspecific and 28 were semi tryptic.

It is interesting to note that all the 28 semi-tryptic peptides presented the C-terminal side tryptic, having as the last amino acids as Arg of Lys.

Such results are not due to a particular selectivity of the enzyme, but to the increased ionization efficiency of such peptides, that allows them to have an easier detection (156).

Glycation characterization and label free quantification of five IgG batches

In Figure 45 is presented the workflow developed for the quantification of the glycation in different batches. Five different batches, of the same monoclonal antibody, were analyzed in triplicate. Each sample was loaded on a stacking gel, the band were extruded, reduced and alkylated. Each band was cut in two parts and submitted to parallel trypsin and chymotrypsin digestion. The peptides were extracted and injected on Q-tof Impact HD (Bruker) coupled with a nanoACQUITY (Waters).

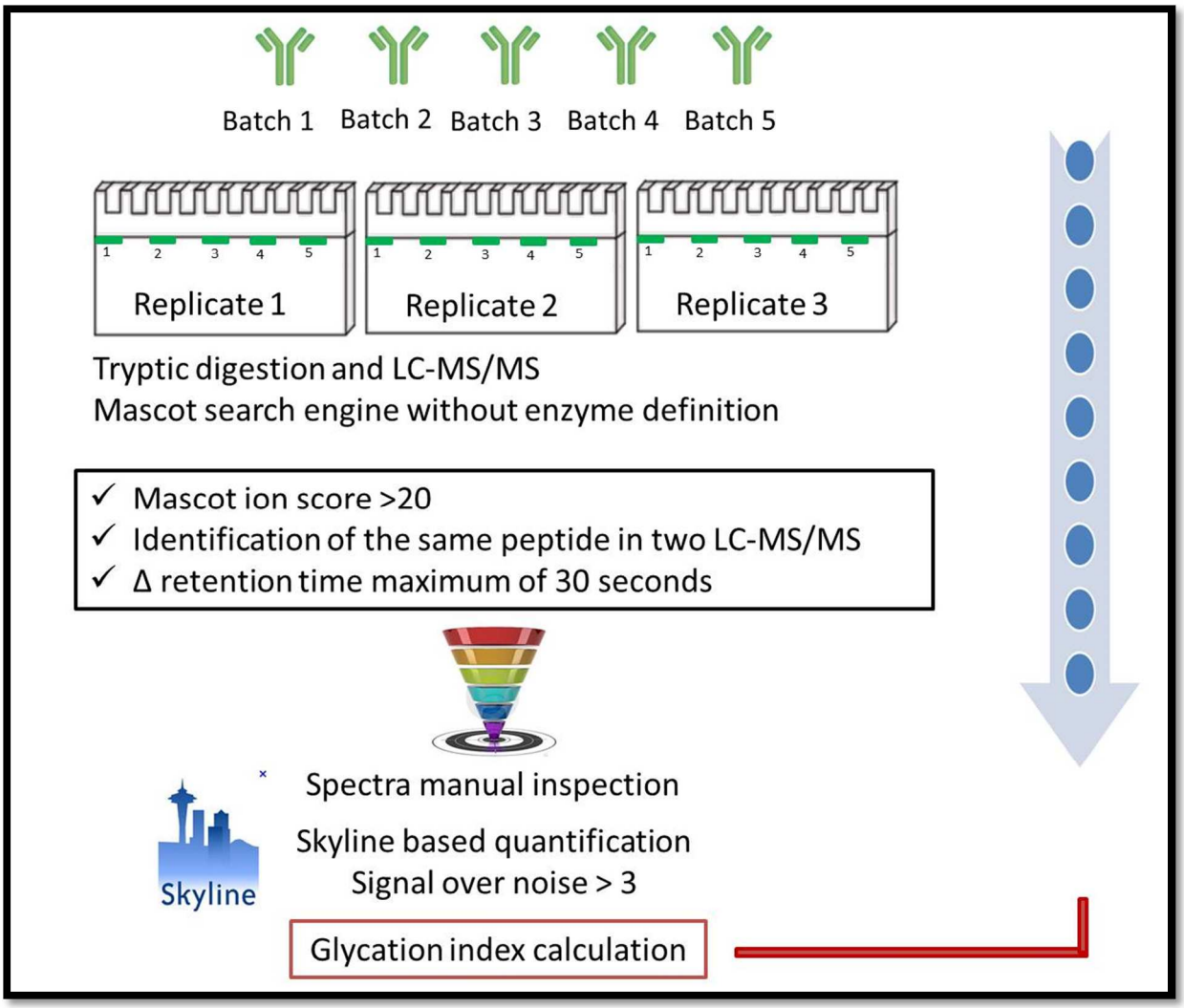


Figure 45: schematic representation of the workflow developed for the glycation batch to batch quantification.

The peptides were separated on 60 minutes linear gradient from 3% to 35% of acetonitrile, in order to maximize the possibilities of detecting low abundance peptides carrying glycation modification. The Q-tof was tuned in order to acquire MS/MS at a variable rate from 2 Hz to 4 Hz depending on the intensity of the parent ion. The ion parents selected were excluded for 20 second after fragmentation.

Such low rate MS/MS method (the Impact HD can efficiently scan up to 25 Hz) was selected in order to generate high spectra quality even from low intense glycated peptides. The peak list generated was submitted to Mascot search using as variable modification all the glycation PTMs described above, no enzyme specificity was defined.

The results obtained for the two enzymes digestion were filtrated using the criteria described above. Briefly, only spectra with ion score >20 were selected, originated by peptides identified in at least two different injections out of the three, with a Δ retention time < 30 sec.

The spectra that passed the validation were submitted to manual inspection, in order to evaluate if the MS/MS fragmentation quality was enough to confirm with out doubts the amino acidic position of the PTM. In Figure 46 is presented a high quality MS/MS spectrum of a peptide identified with a fructosyl Lys - 1 H_2O (ΔM +144).

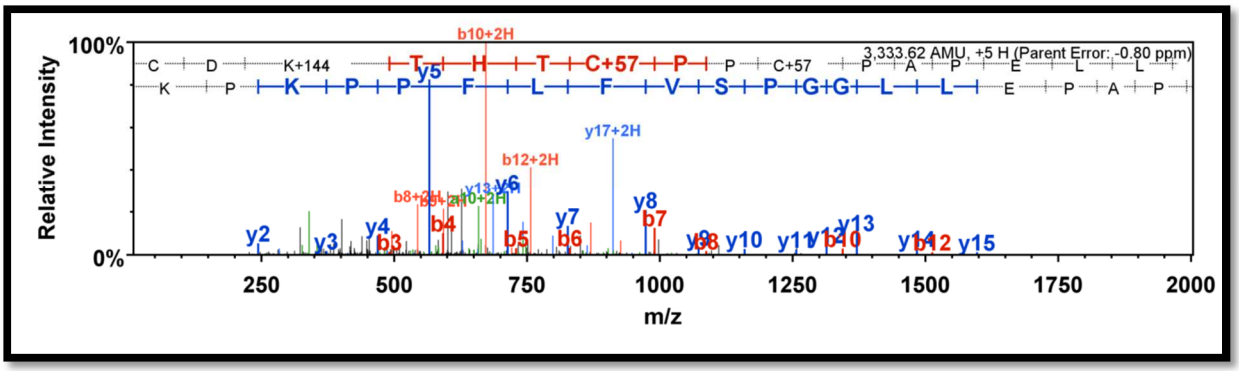


Figure 46: MS/MS spectrum of a peptide identified with a fructosyl Lys - 1 H_2O (ΔM +144).

After this process, 33 spectra of glycated peptides were validated out of the hundreds initially identified by Mascot database search.

The quantification was performed using the software Skyline, the detailed description of a general quantification workflow performed with Skyline is presented in part II chapter I paragraph “MS1 label free workflow” and paragraph “Peak detection in MS1”.

Briefly, the results of the search engine were used to generate the spectra library imported into Skyline. The ions mass peptides were extracted in a four minutes retention time window, across the retention time corresponding to the mascot identification.

Applyng the criteria for the peak detection of a minimal signal over noise >3 just 17 glycated peptides, all produced by trypsin digestion, were validated for the quantification step.

During the MS1 ion chromatogram extraction all the 17 peptides were detected with confidence thanks to:

- Δ retention time < 1 minute across the 15 injection (5 batches analyzed in triplicate).
- The expected intensity ratio between the mono isotope (P), the P+1 and P+2.

In Figure 47 is presented the distribution of the glycated peptides quantified respectively for the heavy and light chains. A glycation hot spot has been detected in the heavy chain from the position 200 to the position 250. In the light chain all the five modified peptides were identified

from the position 140 to 200. These identifications suggest an exposition to the solvents of these two regions of the protein.

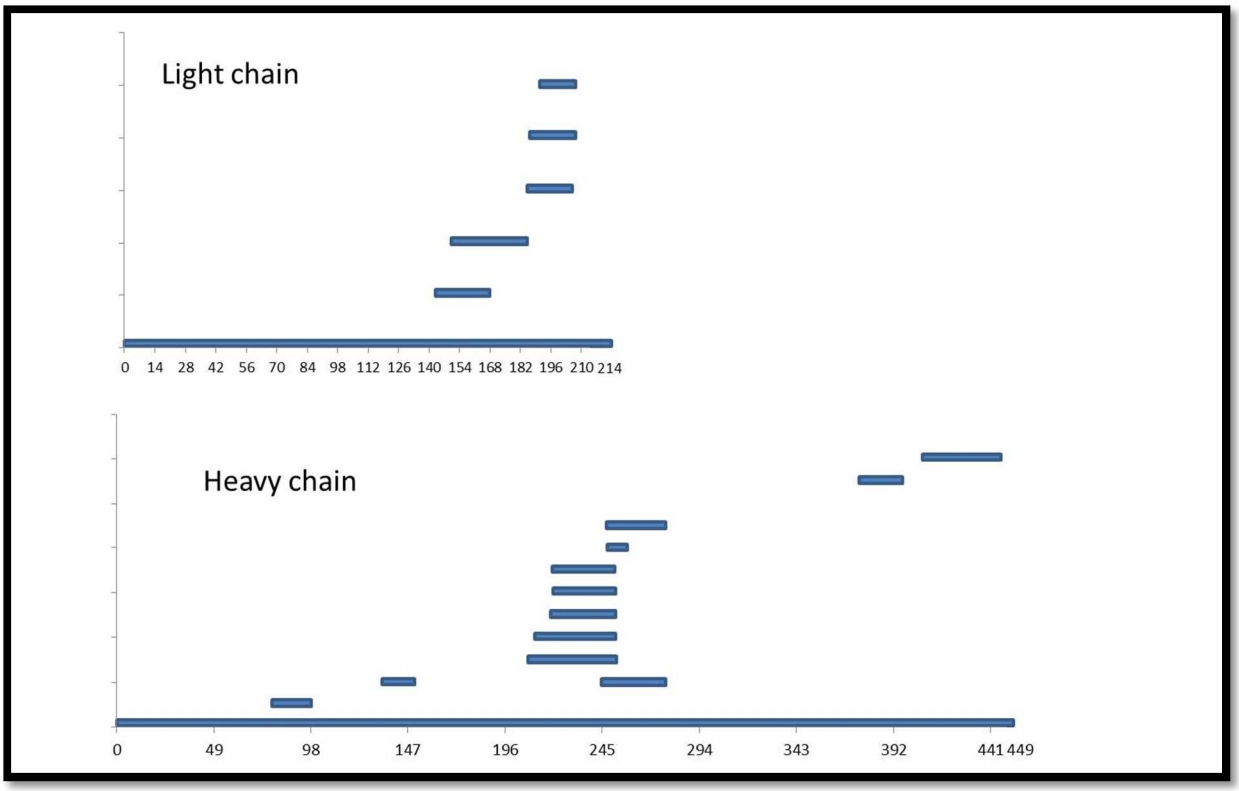


Figure 47: map of the peptides quantified with glycation modification in the light and heavy chain.

In order to minimize the artificial intensity variation induced by the sample preparation, the area under the curve, obtained for each glycated peptide, was normalized against the sum of the area under the curve obtained for all the fully tryptic peptides detected in each analysis.

In such way an a-dimensional Glycation Index (GI) for each modified peptide was calculated. The GI should not be considered as absolute value representing the amount of a modified peptide, but as an index to compare the abundance of the same modified peptide across different batches.

GI= Area glycated peptide / ∑ Areas not modified peptide.

In Table 14 are presented the results obtained applying the method developed below, no major difference in terms of presence/absence of specific glycated peptide has been detected. Some minor fluctuations are detected like for the heavy chain peptide 80-97 that presents a Carboxymethyl Lys (Δm +58.01) quantified in higher amount for the batch 5 respects to the others. Such increase is compensated by the decrease in batch 5 of the light chain peptide 187-207 that carries two Carboxymethyl Lys (Δm +58.01).

Conclusion

In this chapter it has been described the method optimization performed to develop a quantitative approach capable of detecting glycation differences in monoclonal IgG batches, without enrichment step despite the chemical diversity of the glycated products.

The key steps that allowed the achievemnt of the presented results were the optimization of the denaturation and digestion conditions, and the development of a robust data treatment permitting the validation of not fully tryptic peptides.

It was shown that also peptides generated by unspecific cleavage, if manually inspected and validated, are a resourceful source of data. Among the 17 glycated peptides, quantified, just 7 were fully tryptic, one completely unspecific and 9 were semi tryptic with the C-terminal side tryptic.

This approach has been successfully used to decipher the pattern of glycation of mAbs, accomplishing the sensitivity required to quantify the glycated peptide representing approximately 1% of the IgG.

A glycation index, named GI was produced, wich can be used to compare the intensity of the same glycated peptide across different mAbs batches.

As perspective in order to increase the ionization efficiency of the glycated peptide it will be tested the enrichment of the nebulizer gas with vaporized acetonitrile, or other organic solvents in order to increase the sensitivity against such modified peptides.

							BATCH 1		BATC 2		BATCH 3		BATCH 4		BATCH 5	
m/z	Modified Sequence	Chain	Mascot Ion score	Spectrum charge	Pep. start	Pep. stop	Area avg	CV%	Area avg	CV%	Areaavg	CV%	Area avg	CV%	Area avg	CV%
679.14	C[+57]xK[+144]xxxC[+57]xxC[+57]xxxxxxxxxxxxxxxxKxK	HC	63.3	5	222	250	2581	20	2453	12	2375	1	2936	4	2308	20
667.73	CxK[+144]xxxC[+57]xxC[+57]xxxxxxxxxxxxxxxxKxK	HC	86.2	5	222	250	313	9	311	15	360	8	398	3	260	7
1128.51 752.67	xK[+58]x[+16]xxxxxxxxxxxxC[+57]xR	HC	72.3	2, 3	80	97	404	1	422	4	313	9	451	9	839	8
658.48	K[+72]xxK[+72]RxxxK[+108]xCxK[+108]xxxCxxCxxxxxxxxxxxxxxxxxKxK	HC	42.5	7	212	250	499	2	563	4	566	6	533	7	530	8
570.90	xxxK[+162.1]xxxxK[+162.1]xR[+144]xxxxxxxxC[+57]xxxxxxxxxxxxxK	HC	33	8	408	441	78	5	88	7	80	7	90	5	90	14
413.55	xK[+162.1]xxxx[+16]xxR	HC	25.5	3	249	257	4419	12	5232	17	3989	16	3916	16	6592	18
640.32	xK[+58]xxxx[+16]xxRxxxxxCxxxxxxxxxxxK	HC	67.1	5	249	276	209	2	255	6	279	2	246	4	263	5
764.77	xKxK[+270x1]xxxxxxR[+144]xxxxxC[+57]xxxxxxxxxxxxK	HC	38	6	247	276	108	5	133	7	134	1	144	6	134	9
847.71	xxxxxxxxxxxxxxxxxK[+162.1]xx	HC	22.9	3	376	396	904	10	948	16	916	17	1026	12	1001	15
700.14	xC[+57]xK[+162.1]xxxC[+57]xxC[+57]xxxxxxxxxxxxxxxxKxK	HC	66.4	5	221	250	1053	13	1216	13	967	4	1199	4	1393	20
441.56 661.83	xxxxxxxxxCxxK[+58]	HC	57.2	2, 3	136	149	6	5	6	3	6	5	6	10	6	17
696.06	xKxxKR[+270x1]xxxK[+126]xCxK[+126]xxxCxxCxxxxxxxxxxxxxxxxKxK	HC	35.2	7	211	250	733	9	813	18	1253	5	1037	7	1068	7
1261.26 946.20	xxxxxxxxxxxxxxxxxxK[+162.1]xxxxxxxxxxxxxKx	LC	91.6	3, 4	150	183	1386	9	1556	3	945	3	1208	10	1493	10
947.11	xxxK[+162.1]xxxxxxxxxxxxxxxxxxK	LC	111	3	146	169	1911	12	2322	6	1446	2	2121	15	2312	17
645.32	xxxC[+57]xxxxxxxxxxxxK[+58]	LC	31	3	191	207	1520	14	1593	7	1196	16	1194	13	1208	4
696.84	xxxKxK[+108]xxxC[+57]xxxxxxxxxxxxK	LC	31.4	4	185	207	2948	8	3564	6	2680	6	3909	24	4150	20
629.31 503.65	xK[+58]xK[+58]xxxC[+57]xxxxxxxxxxxxKx	LC	48.5	4, 5	187	207	344	3	367	4	266	7	375	6	98	7

Table 14: results of the quantitative glycation batch analysis.

Part II - Chapter IV -

Intact protein label free quantification: future developments

Part II - Chapter IV - Intact protein label free quantification: future developments

Introduction: biological context and analytical task

Hemoglobin (Hb) is a tetrameric protein, which surrounds the heme molecule responsible for the oxygen and carbonic dioxide exchange in the vertebrate blood. Hb is composed of two β-like chains (δ, $^{g}\gamma$, $^{a}\gamma$ or β) and two α-like chains (ξ or α). Different combinations of chains are generated during three main stages of the human development: embryonic development, fetal development and normal adult in order to modulate the affinity of Hb for oxygen.

In normal adults, the major hemoglobin is composed by approximately 95% of HbA ($\alpha_2\beta_2$) while the rest is mainly HbA_2 ($\alpha_2\delta_2$) and a low level HbF ($\alpha_2\gamma_2$), Figure 48

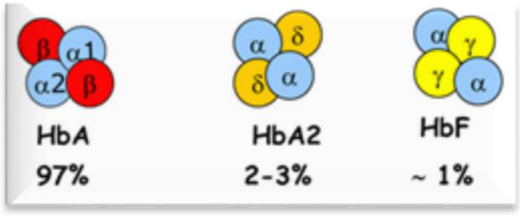


Figure 48: presence of three forms of hemoglobin in the human erythrocytes

Thalassemia is a family of genetically based pathology, caused by the unbalance, or mutation of α and β like chains in human blood. The diagnosis of Hb related pathology (157, 158) requires the combination of different assays, like gap polymerase chain reaction (gap-PCR), to evaluate the presence of genetic mutation and relative quantitative Hb chain analyses.

Beta-thalassemia is characterized by a decrease or absence in the synthesis of β-globin chains of human hemoglobin. In the screening for classical beta-thalassemia trait, the hallmark is the presence of an elevated level of hemoglobin A2 (α2δ2). In normal subjects, HbA_2 is present between 2.0 to 3.2 % of total hemoglobin and in beta-thalassemia carriers is usually between 3.8 to 7.0 % of total hemoglobin. Therefore, an accurate quantitative determination of HbA2 is a key marker for the correct diagnostic of beta-thalassemia.

Several routine laboratory tests are available to quantify this minor hemoglobin (159-161) but due to lack of reference material, diagnostic sensitivity is variable.

Currently, the first test screening performed in patient suspected to be carrier of Thalassemia, is a relative quantification of the Hb chain performed using cation exchange liquid chromatography (CE-HPLC). The CE-HPLC is a high-throughput technique used for routine clinical measurements (162-164), systems commercially available can automatically analyze and generate a report in less than ten minutes (165). Due to the impossibilities of discriminating proteins coelution in the same peak, this method requires the complete chromatographic separation of the tetrameric Hb.

In the event of a genetic mutation responsible for the production of a Hb variant, differing usually just for one amino acid in the sequence, the CE-HPLC method is not capable of detecting co-eluting species and for such reason the quantification of HbA_2 could be distorted (165).

To standardize the measurement of hemoglobin A2, a project of developing a new reference system has been started under the auspices of the International Federation of Clinical Chemistry (IFCC) in the last years. In this context, a first method which provides quantification of HbA_2 based on enzymatic cleavage and LC/MS was developed during the Ph.D. thesis of Hovasse (166). Despite the reproducible results obtained in term of CV% this method is time consuming and due to analytical skills required, is unlikely to be selected as routine test. Therefore, besides this approach, we decided to explore the feasibility of a quantification approach at the protein level.

In this work, I have developed a secondary reference measurement method based on the quantification of intact globin chains by LC-ESI/MS which eliminates the need for a digestion step prior to LC-MS/MS analysis. The goal was to develop a mass spectrometry based method on the intact globins, capable of achieving accuracy and a reproducibility close to the one obtained by the classical routine measurement methods and which offers the possibilities of detecting the presence of Hb carrying mutation.

LC-MS/MS method development

Samples description

Briefly, blood samples from patients were collected in EDTA coated tube and centrifuged to remove the plasma. The erythrocytes were washed 3 times with isotonic saline to remove the leucocytes and platelets and lysed by addition of distilled water and carbon tetrachloride. After centrifugation, the hemoglobin solution forming the surnatant was collected. These hemolysates were directly analyzed by LC-MS.

A total of 38 samples were purified and analyzed using the actual reference method CE-HPLC.

Two series of 38 samples in total were analyzed in triplicate. For each series, a set of four samples are used as calibrators. These calibrators were prepared by mixing a precise amount of HbA_2 in pure HbA0. The different concentrations were chosen in order to have an HbA_2 concentration spanning across the expected range of measurement.

Series 1: 4 calibrators plus 12 samples to be analysed in triplicate

- Calibrator 1: HbA_2 = 2.5 %
- Calibrator 3: HbA_2 = 3.4 %
- Calibrator 2: HbA_2 = 5.6 %
- Calibrator 4: HbA_2 = 6.2 %

Series 2: 4 calibrators plus 18 samples to be analysed in triplicate

- Calibrator 13: HbA_2 = 2.5 %
- Calibrator 14: HbA_2 = 3.6 %
- Calibrator 15: HbA_2 = 4.9 %
- Calibrator 16: HbA_2 = 6.3 %

Measurement (calibration) method

The HbA_2% value was calculated performing an ion chromatogram extraction of the masses corresponding to the states of charge ionized with the highest intensity. The values so obtained for the δ chain were divided by the value of the α chain, producing a value defined as ratio δ/α.

Calibration was performed against calibrators made of a mixture of HbA0 + HbA_2. The ratio δ/α of each calibrator was plotted against the known HbA_2 % value and fitted using a linear equation in order to calculate the HbA_2 % of the unknown samples.

Chromatographic method development

The δ and β chains differ only from 11 amino acids over 147 and among the 11 differences, 8 positions present amino acidic exchange between the same categories (basic, acid, polar or hydrophobic residues) as displayed in Figure 49. Such high degree of similarity between the chains results in a not easily achievable chromatographic separation.

```
1    MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPK   60  P02042     HBD_HUMAN
1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK   60  P68871     HBB_HUMAN
     *********:**.********* ***************************:*********

61   VKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFG  120  P02042     HBD_HUMAN
61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120  P68871     HBB_HUMAN
     **************************: ****************************:.**

121  KEFTPQMQAAYQKVVAGVANALAHKYH  147  P02042     HBD_HUMAN
121  KEFTPPVQAAYQKVVAGVANALAHKYH  147  P68871     HBB_HUMAN
     ***** :********************
```

Figure 49 displays the sequence alignment of the Human δ and β chains. The stars represent the complete identity of the two aminoacids.

In the reverse phase-HPLC separation coupled to mass spectrometry, the formic acid (FA) is the most common ion-pairing agent used. The addition of low percentage of FA, generally 0.1%, allows increasing retention on the RP chromatography, thanks to ionic interaction generated with the peptides and thanks to the hydrophobic interaction with the stationary phase.

The trifluoroacetic acid (TFA) is definitely a stronger ion-pairing agent (167) that allows to obtain thinner chromatographic peaks. It is well known that TFA is responsible for a relevant decrease of the Electrospray ionization efficiency (168), since the ionic pairs between peptides and TFA are maintained even during the spray formation in the mass spectrometer source (169).

In a first series of experiments, different reversed phase columns were tested using formic acid in order to achieve the best possible separation between the three globin chains (α, δ and β) with the highest detection sensitivity possible. Among them, the X Bridge C4 BEH300 (2.1mm X 250mm) from Waters is the one that gave the best results, as illustrated in Figure 50-a.

The gradient optimized to achieve the separation displayed a slow slope, from 30% of acetonitrile to 35% in 10 minutes, with the drawback of enlarging the peaks (2 minutes at base peak width instead of 1.5).

In order to improve the peak width, TFA was tested as alternative ion-pairing agent, in Figure 50-b is presented the BPC obtained injecting the same amount of purified Hb. Thanks to the TFA, when using the same gradient slope, the peak width at the base is reduced of 50% and is close to 1 minute for the peak of the β chain, the most intense of the three.

On the other hand, as expected, the TFA is responsible for a drop of sensitivity really drastic, the peaks containing the δ chain does not present intensity and a shape allowing a robust integration. Since the amount of samples available for the method development was limited, a first series of quantification was performed using FA as ion-pairing agent.

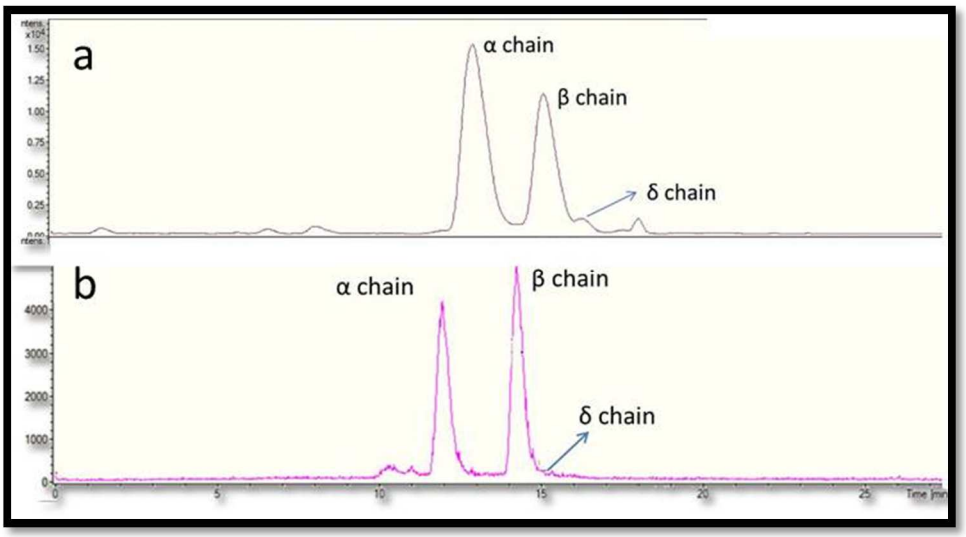


Figure 50: LC-MS chromatogram obtained with the Waters X Bridge C4, two ion pairing are compared: 61-a FA, 61-b TFA. The TFA allowed reducing of 50% the peak width, but the sensitivity is drastically reduced.

Stability of the sensitivity across the experiment

In a quantitative experiment, monitoring the stability of the chromatographic performances and the sensitivity of the mass spectrometer is crucial in order to have stable instrumental conditions across all the measurements. The scheduled cleaning of mass spectrometer source part allows the preservation of the performances for a longer time, decreasing the rate of the generation of the charge effect. This process is due to the accumulation of solvent and sample residue on the lenses, resulting in a local alteration of the voltage applied.

The first series of sample has been injected on the ESI-Q-Tof (micrOTOF-Q, Bruker). This mass spectrometer is equipped with a 30 degrees angled sprayer respect to the entrance (Figure 51). The sprayer shield and the cap sprayer are two metallic parts that can be removed and cleaned without breaking the vacuum. These parts are placed on the top of the glass capillary entrance and they function as protection for the mass spectrometer, blocking a big part of the not ionized solvent.

The glass capillary is the entrance of the Q-Tof and is placed slightly out of the axes respect to the dual stage ion funnel. Since the neutral ions are not deviated by the voltages applied on the ion funnel they collide on it. The cleaning of the glass capillary is a procedure that requires venting the mass spectrometer, but since removing the capillary and replacing it with a clean one, is a procedure that requires just few seconds, the amount of atmospheric humidity that enters in the Q-Tof is limited and the pumping procedure requires approximately four hours, allowing to rapidly restart the analysis.

The cleaning of the funnels cartridge instead is a longer procedure that results in 24 hours stop of the Q-Tof and for such reason is performed less often respect to the weekly changing of the glass capillary.

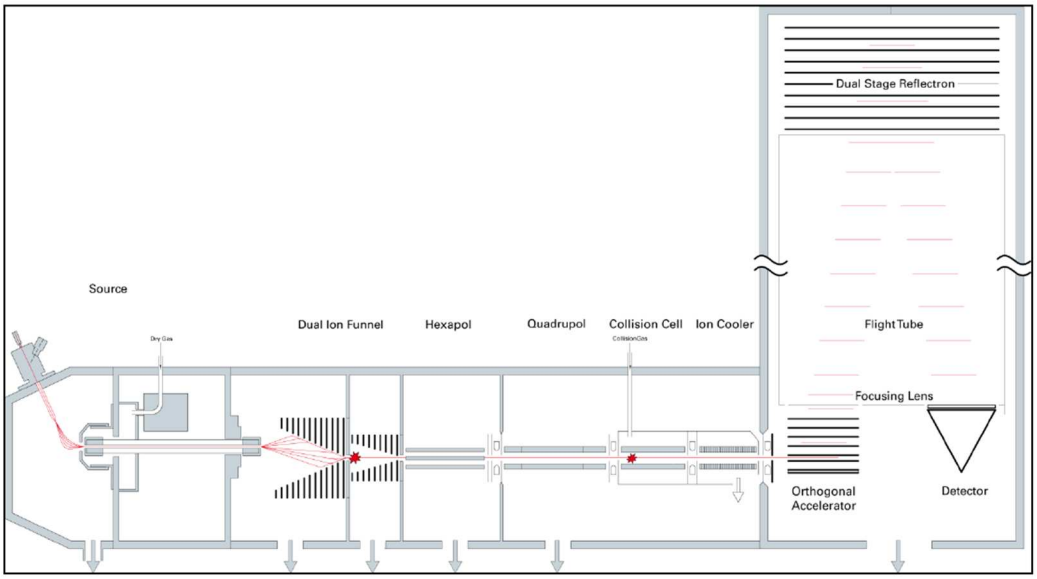


Figure 51: schematic representation of the micrOTOF-Q, Bruker

The measurement of HbA_2 was done in triplicate to evaluate the CV%, with the analytical sequence described below:

Calibrator set 1 - first run
Patient samples (1-12)- first run
Calibrator set 2 - second run
Patient samples (1-12) - second run
Calibrator set 3- third run
Patient samples (1-12)- third run

For the calibration curves calculations, the ratios δ/α were plotted against the values of HbA_2.

Each sample injection was followed by a blank injection to avoid a carryover of Hb across sequential samples. After each set of samples the sprayer shield and the cap sprayer were sonicated in organic solvent to clean them. The transmission capillary was changed before the start of experiment.

In Figure 52 the three charts obtained plotting the ratio δ/α of the calibrators against the known HbA_2 % values are displayed. In the first two calibrations curves, the R squared is fairly close to 1 and stable, respectively 0.998 and 0.997. Carefully observing the chart of series 2, it is possible to detect the first sign of the sensitivity degradation. Indeed, the slope of the trend equation is decreased from 0.15 to 0.010, because the ratio δ/α is less accurate since the loss of sensitivity has a higher impact on the δ peak, the smallest one.

In the third chart the loss of sensitivity is relevant, the slope 0.008 and R squared 0.972. This set of data was discarded due to the not stable sensitivity across the 3 days experiment.

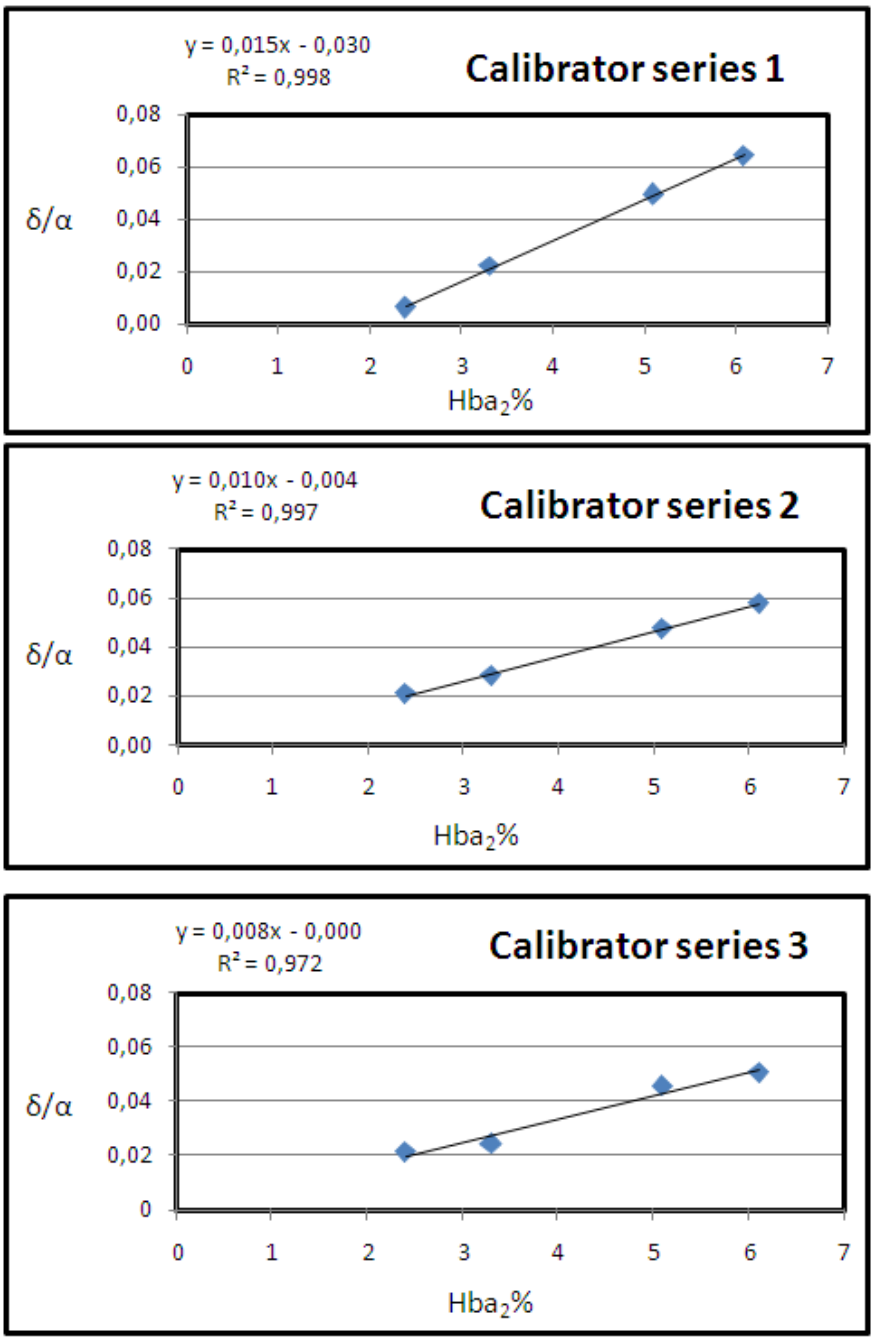


Figure 52: the three charts obtained plotting the ratio δ/α of the calibrators against the known HbA_2% value. In the series three the degradation of the mass spectrometer sensitivity results in a less precise measurement of the ratio δ/α.

The series of injection experiment was reproduced on the same LC-MS system. To ensure a better instrumental stability, after each set of samples the Q-Tof was vented and the transmission glass capillary replaced. This procedure allowed the preservation of sensitivity and for such reason the R squared are reproducible and the slope perfectly constant across the whole experiment.

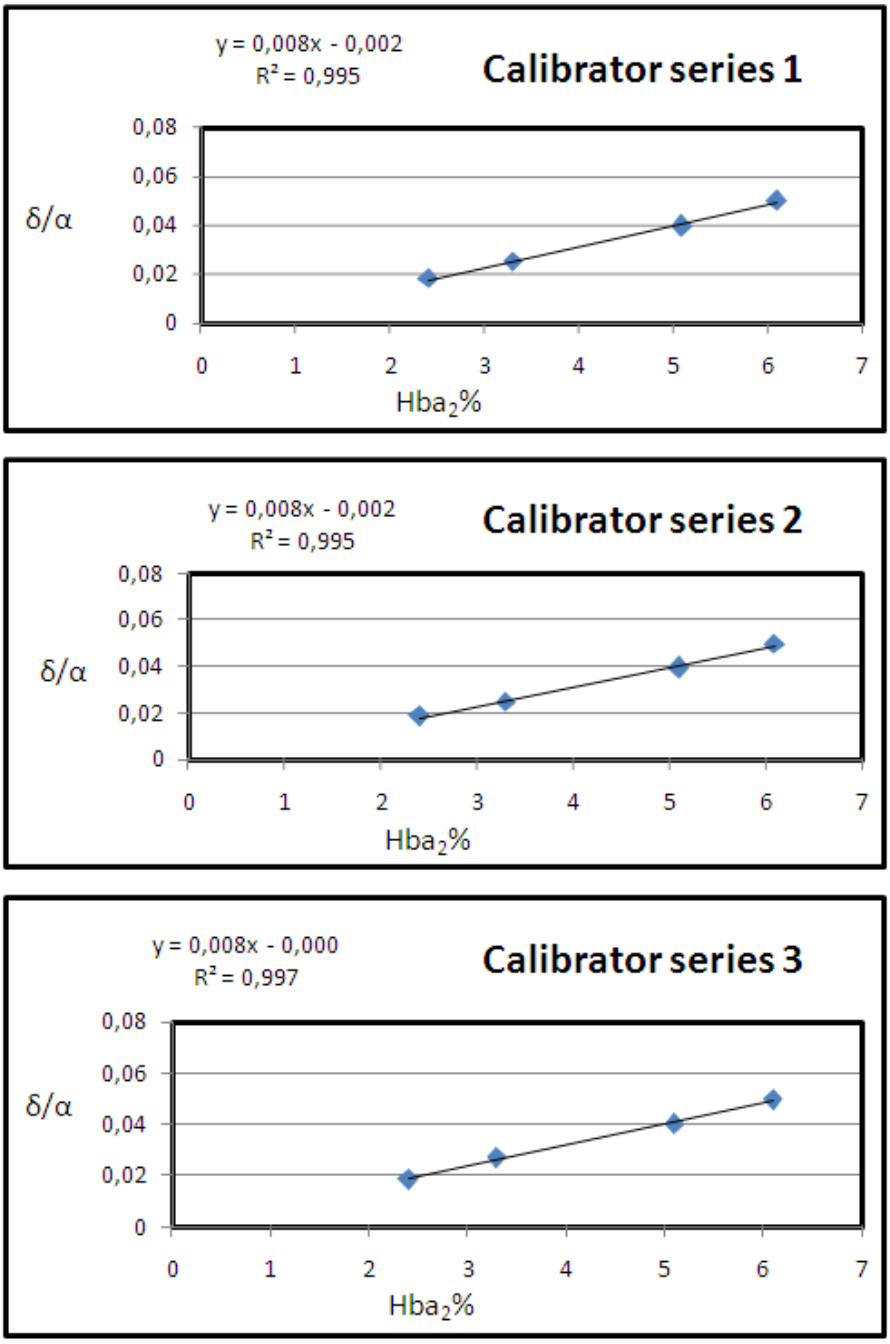


Figure 53: the three charts obtained plotting the ratio δ/α of the calibrators against the known HbA_2% value. No degradation of the mass spectrometer sensitivity is observed.

Results

The two series of 38 samples have been analyzed by LC-MS using the optimized method described above.

In Figure 54 is presented a global chart where the results of the two series of samples are plotted against the results obtained with the reference CE-HPLC method. Globally the majority of the measurements are included in the 99% confidence, aside few outliners.

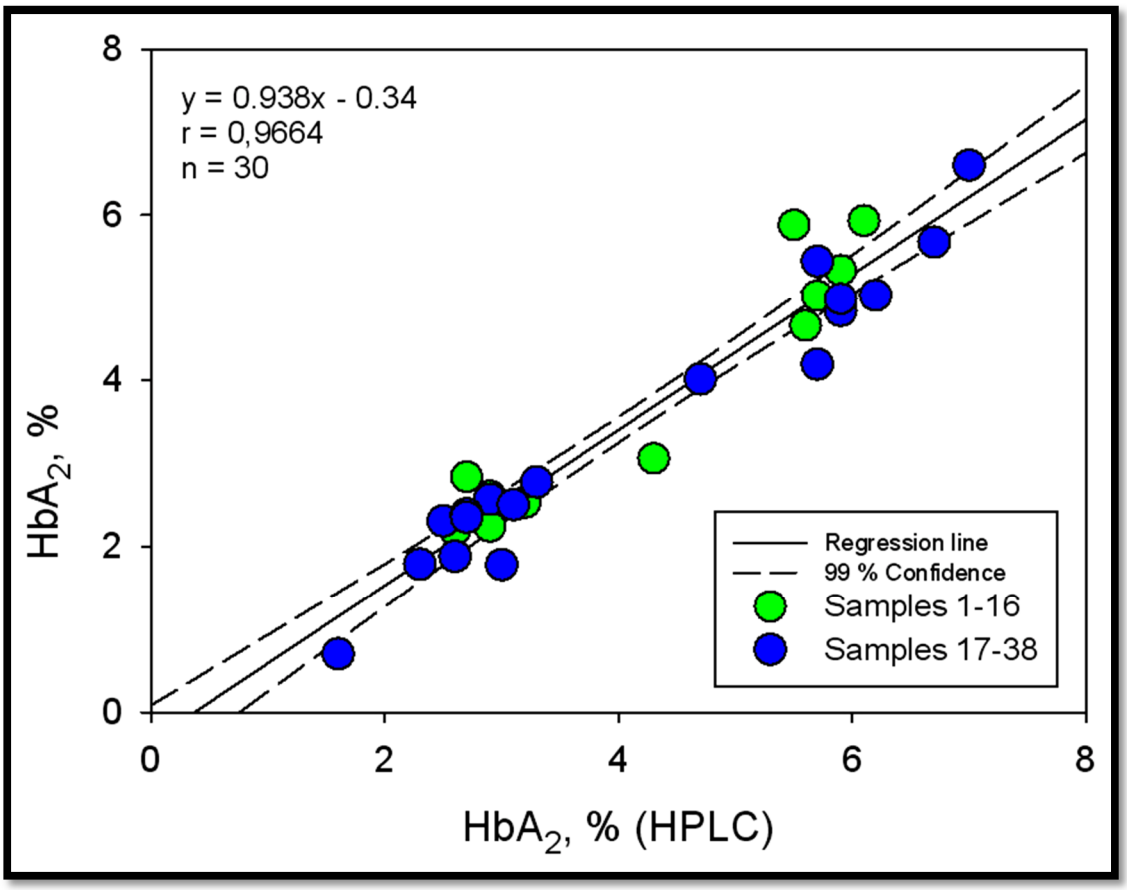


Figure 54 charts presenting the plot of the HbA_2 measurement obtained by mass spectrometry against the measurement performed with CE-HPLC.

It should not be neglected that these results have been obtained submitting the Q-Tof to a relevant mechanical stress given by the vacuum venting process performed five times in one week.

The next step will be to analyze a new series of samples with an improved LC method. We have already achieved a complete separation of the three globin chains using a Vydac TP C4 250mmX1 mm column with TFA as ion-pairing Figure 55.

We have also planned to use other mass spectrometer architectures (Q-Exactive Thermo and triple TOF AB SCIEX) to evaluate the robustness of this LC-MS method.

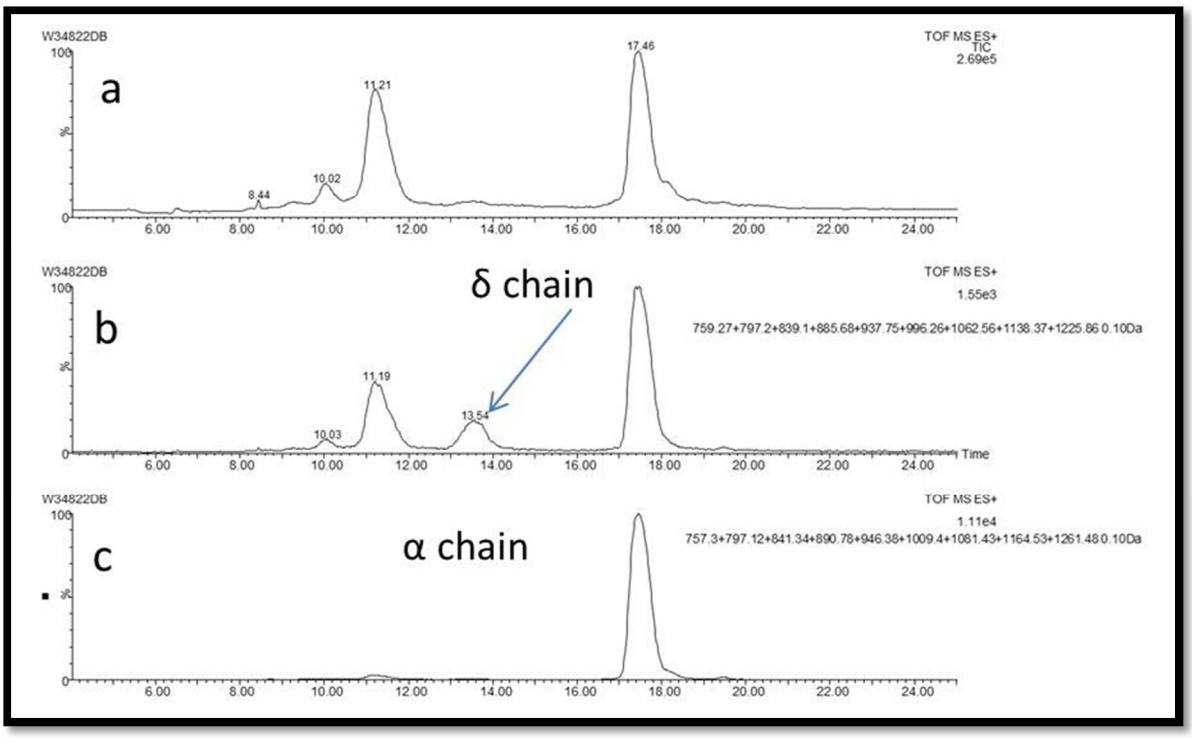


Figure 55: LC-MS chromatogram obtained with the Vydac TP C4 column. b) and c) present respectively the IEC of the δ and α chains.

Conclusion

In this work, I have presented a methodological development of a measurement method for HbA_2 based on the quantification of intact globin chains by LC-ESI/MS which eliminates the need for a digestion step prior to LC-MS/MS analysis. This work is part of an interlaboratories effort driven by the International Federation of Clinical Chemistry (IFCC) to provide tools for β-Thalassemia diagnosis.

The mass spectrometry presents the unique advantage of being able of performing accurate δ/α intensity ratio as the CE-HPLC based on an UV detection, but also allows a fast detection of eventual sequence variations. The structural information so obtained, can be used to direct the further time consuming PCR test, allowing a faster diagnosis. This LC-MS method presents the potential of detection of mutated globin chains in addition of the quantitation information.

We have shown that to achieve the high accuracy required for this quantitation (CV lower than 5% while usually CV<20% is given by the bioanalytical guidelines) a particular attention had to be paid on the mass spectrometer to ensure long-term sensitivity throughout extended sample batches.

This quantitative method should now be validated by an inter-laboratory assay and by analyzing more samples. This validation step is under current investigation.

Also, a quantitative approach by LC-MS based on intact proteins represents an interest for biopharmaceutical applications and could provide in the future a high-throughput method for quantitation for biotherapeutic applications.

Part II - Chapter V -

Optimization of a Data Independent Acquisition method and MS1 comparison

Part II - Chapter V - Optimization of a Data Independent Acquisition method and MS1 comparison

Introduction

In the 2004 the group of Yates in California(170) published a paper describing a Data Independent Acquisition method "DIA", this technique is based on the repetition of two steps:

1. Accumulation of ions across a narrow window (10 m/z) and the subsequent fragmentation of all the peptide selected.
2. Sliding the selection window of each cycle respect to the precedent in order to sample all the m/z.

With this revolutionary approach the authors demonstrated an increased sensitivity, respect to other DIA method already described, but based on the selection of all the ions present in a chromatographic peak and subsequent MS/MS fragmentation(70). This technique is radically different from the DDA, where after an MS scan the most intense peptides are selected one by one to be submitted to fragmentation.

Recently, modifications of the original workflow has been proposed, like in the case of the PACIFC (171).

Despite the DIA proved better performance in the quantitative results, respect to the MS1 (where the quantification is based on the Ion Chromatogram extraction of the parent mass(170)) the DIA did not gain popularity(172). The lack of publically available user friendly software and the need of a high resolution mass spectrometry capable of performing MS/MS at least at 10 HZ, hampered the development of this technique.

In 2012, Gilet (75) with the so called SWATH™, revolutionized the DIA acquisition method without changing the mass spectrometry way of acquiring, but modifying the way of analyzing the data acquired. Up to then, the raw data were treated with search algorithm developed for DDA, or by searching pseudo MS/MS spectra, considering the coeluting peak profile of the peptide in the MS scan and the fragment ions. Gilet proposed to extract, in targeted way like in SRM, the chromatograms traces of the ions expected to be produced by the fragmentation of a peptide.

The recent increased popularity of this approach is also due on the public availability of free open source and user friendly software for SWATH™ data treatment like Open MS(173) and Skyline (174) that allow to treat SWATH™ data generated by different mass spectrometry vendors.

The goal of this chapter was to develop a SWATH™ (from now on referred as Middle band) based workflow, for Bruker Impact HD q-tof, able to perform differential quantification in complex mixtures. In a second step to evaluate the interest of such technique respect to the MS1 label free quantification. Such comparison was performed using the software Skyline.

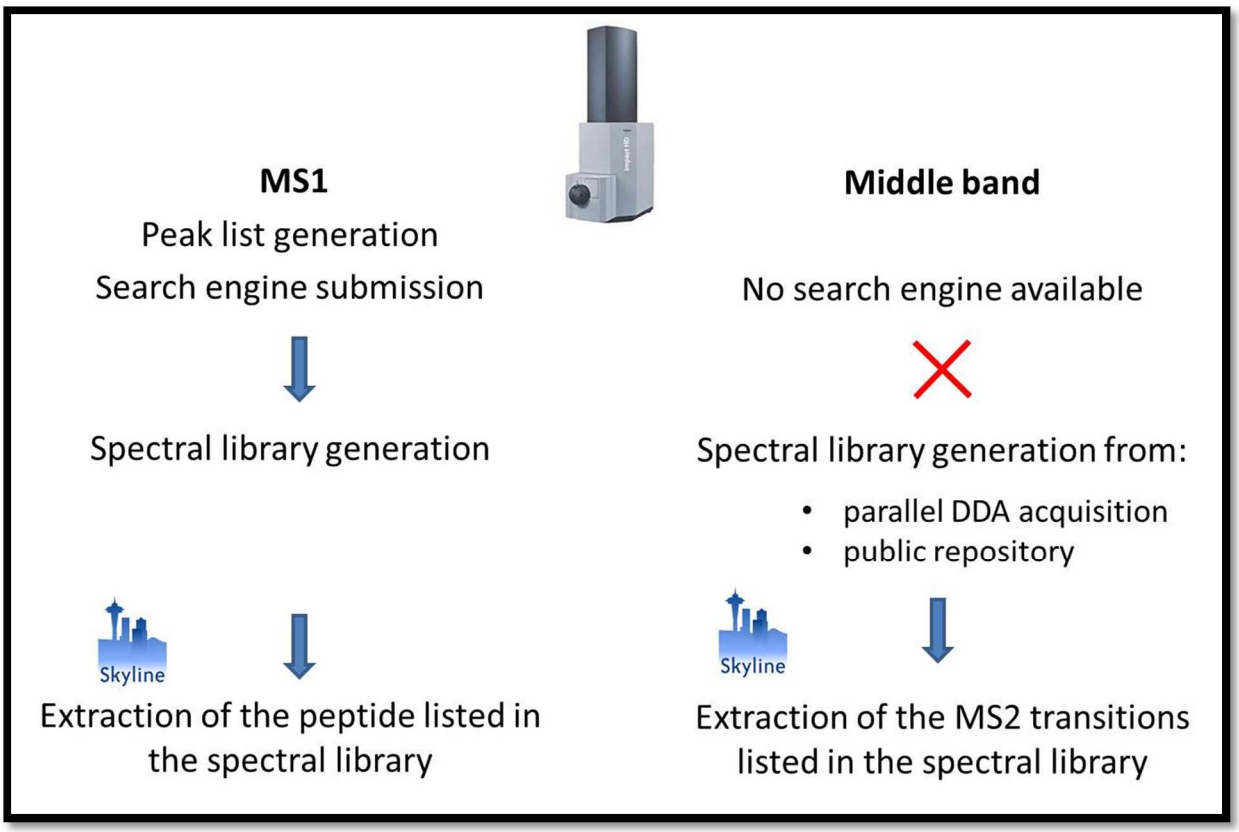


Figure 56: comparison of the workflow DDA and Middle band. In the Middle band the absence of search engine capable of interpreting the data requires an additional step to generate the spectral library.

1 MS1 label free workflow

The MS1 quantification is based on the ion chromatogram extraction (XIC) of the peptides mass from a classic Data Dependent Acquisition (DDA) acquired raw data and the consecutive integration of the area under the peak. The mass of the peptides to extract are listed in text files called spectral libraries, these extensions files, are obtained loading into Skyline the results of the search engine. An important feature of this workflow is the possibilities of loading a virtually unlimited number of search engine results (obtained from different injections) and to merge them in a single library. Extracting a library so generated, allows detecting and quantifying peptides that have not been identified all in the same injection, due to the stochastic behavior of the parent selection.

2 Middle band label free workflow

As described in Figure 56, the Middle band presents a similar workflow to the one of MS1, but with the limitation of not having a search engine capable of interpreting the data. For this reason the quantitative information is obtained extracting the transitions included in the library of spectra generated by others injections acquired with a classic DDA workflow. In the approach Middle band, like in the MS1, the quantification is performed integrating the signal under the curve, but in this case the chromatogram is generated extracting a transition from a parent mass to a product ion like in the SRM.

3 Peak detection in MS1

The Impact HD q-tof achieves the considerable resolving power of 40000 and a low range ppm error (after recalibration with external lockmass). Despite which the ion chromatogram

extraction (XIC) of a peptide mass, often results in the generation of a chromatogram presenting multiple peaks. The presence of peptides with small Δ mass is responsible for the generation of multiples peaks in the XIC generation.

In order to increase the robustness of the peak selection process, the software Skyline extracts for each peptide the mass of the monoisotopic peaks and also the P+1 and the P+2. The algorithm of peak peaking, selects the peak that presents the best coelution of the three isotopic peaks and the closest intensity ratio of the three peaks (P, P + 1, P + 2) respect to the ratio observed during the spectral library generation. The software allows extracting the peptide mass in a defined time window centered on the retention time at witch the peptide has been identified, this results in relevant decrease of the possible wrong peak selection.

4 Peak detection in Middle band

In the Middle band workflow the quantification is based on the MS/MS intensity, a user defined numbers of transitions are extracted starting from the most intense ions fragments registered in the spectral library. The algorithm of peak peaking, selects the peptide that presents the transition ratio closest to the one described in the library.

Like in the MS1 the software allows extracting the peptide mass in a defined time window centered on the retention time at witch the peptide has been identified during the spectral library generation.

In Bruker Impact HD q-tof it is possible to add an MS scan every cycle of selection, resulting in the possibilities of extracting from the raw data also the signal of the parent mass and to generate a chromatographic peak representing the combination of the intensity obtained from the MS scan and the MS/MS.

Development of the Middle band acquisition method

Scan rate and cycle time choosen

In order to develop the most appropriate Middle band method for the Impact Q-tof Bruker the balance of two parameters has been carefully evaluated: the scan speed of each Middle band windows and the global cycle time needed to cover all the m/z.

1. The scan speed at which the transition is acquired must be appropriate in order to allow also the less intense peptide to generate a sufficient signal. During the classic DDA analysis, in the Impact HD, the MS/MS is performed at a variable scan speed from 2 to 25 Hz depending upon the intensity of the peptide selected.
   In the Middle band, such kind of intensity related scan speed can't be performed since the Middle band cycling is performed without any regards to the intensity of the peptides. For this reason 15 Hz has been choose since it represent a good compromise between the sensitivity obtained when scanning at lower Hz and the efficacy of an higher scan speed.
2. The cycle time represent a crucial parameter to achieve robust quantification. Having a cycle time longer than 1/10 of the chromatographic peak width would result as showed in the Figure 57 in a not precise definition of the peak and consequently in not robust quantification results.

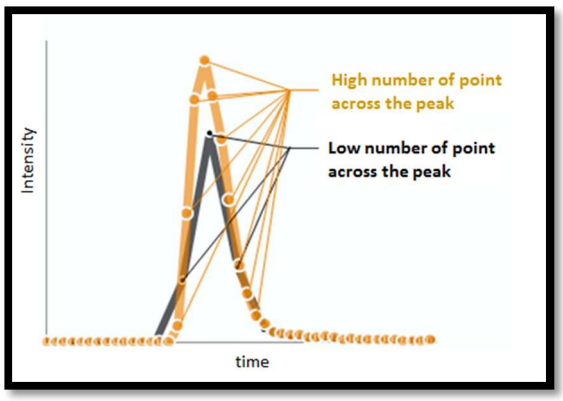


Figure 57: illustration of the effect of the sampling rate. In yellow a high numbers of points across the peak result in a good peak definition, in black the peak is defined by not enough points.

Once the two parameters described above have been defined, the calculation of the width of each window of selection was obtained considering:

- **The chromatographic peaks of approximately 30 seconds width at the base.**
- **A targeted m/z to scan of 900 m/z (from 400 m/z to 1300 m/z).**
- **A global cycle time of 3 seconds requested.**
- **15Hz of scan speed.**

The combination of these parameters, result in the generation of 45 consecutives Middle band scan of 20 m/z plus an additional MS scan that can be eventually used for the quantification.

Evaluation of the quadrupole efficiency transmission efficiency

In the Middle band the isolation performed by the quadrupole is not centered on the monoisotopic peak of the peptide, on the contrary of the classic DDA. For such reason it is crucial to verify that each peptide across the m/z has the possibility of being selected with the complete isotopic envelope in at least one window (see Figure 58).

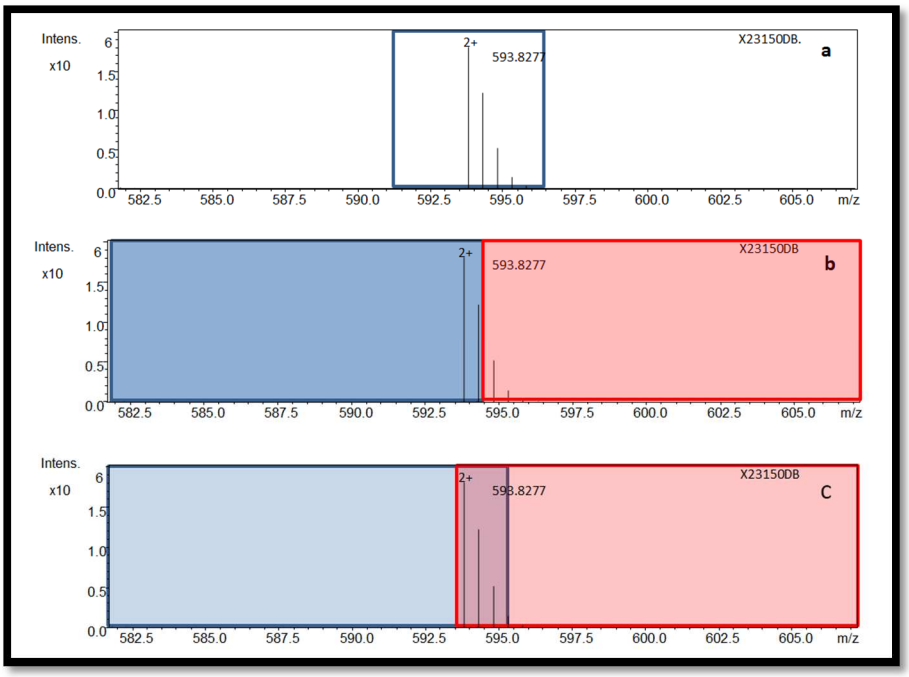


Figure 58: comparison of the DDA and the Middle band peptide selection performed by the quadrupole.

The Figure 58a represents a **DDA** peptide selection: the isolation window (blue square) of the quadrupole is **centered** on the mono isotopic peak of the peptide, this allows to the complete isotopic envelope to be always selected.

The Figure 58b represents the results of two consecutives Middle band windows (blue and red square) acquired with a **theoretical quadrupole** having 100% efficiency and perfectly sharped boundary. An hypothetic peptide having the isotopic envelope across the two windows of selection would not be properly quantified in neither of the two consecutives selection.

The Figure 58c represents the results obtained on the same **theoretical quadrupole** described in the image (b), but this time the Middle band acquisition has been performed using an **overlay** of 2.5 m/z. The overlay allows to all the peptide to be completely selected at least in one Middle band selection window.

In order to evaluate the necessity of performing overlapping windows, the transmission of the quadrupole efficiency has been tested performing an infusion of BSA digest and evaluating the efficacy of transmission when changing the width of the selection.

As displayed in Figure 59, during the infusion of a Bovine Serum Albumin BSA digest, four peptides have been selected and multiple acquisitions have been performed with decreasing width of selection.

On the left side of the selection boundary, the efficiency of transmission proved to be constant across the entire window and the drop of intensity is observed at 2 m/z after the boundary.

On the right side of the selection boundary instead a drop of 50% of the intensity of is observed already inside the last 25% of the isolation width.

Those unexpected results can be summarized in a transmission profile that present sharped boundary on the left side of the isolation windows and a tailing shape on the right side.

Since an overlap of the Middle band windows of 25%, would result in an important increase of the complexity of the MS/MS spectra and consequently in a loss of sensitivity, it has been decided to do not perform any Middle band overlap.

The asymmetric performance of the quadrupole results in a not constant sensibility and dynamic range for all the peptides measured. Depending upon how the m/z of the peptide is placed in the Middle band selection window respect to the boundary, the efficiency of transmission will be differently affected. This issue is mitigated by the fact that the Middle band goal is to detect relevant differential regulation in protein abundance across different biologic sample and not to compare different peptides in a same sample.

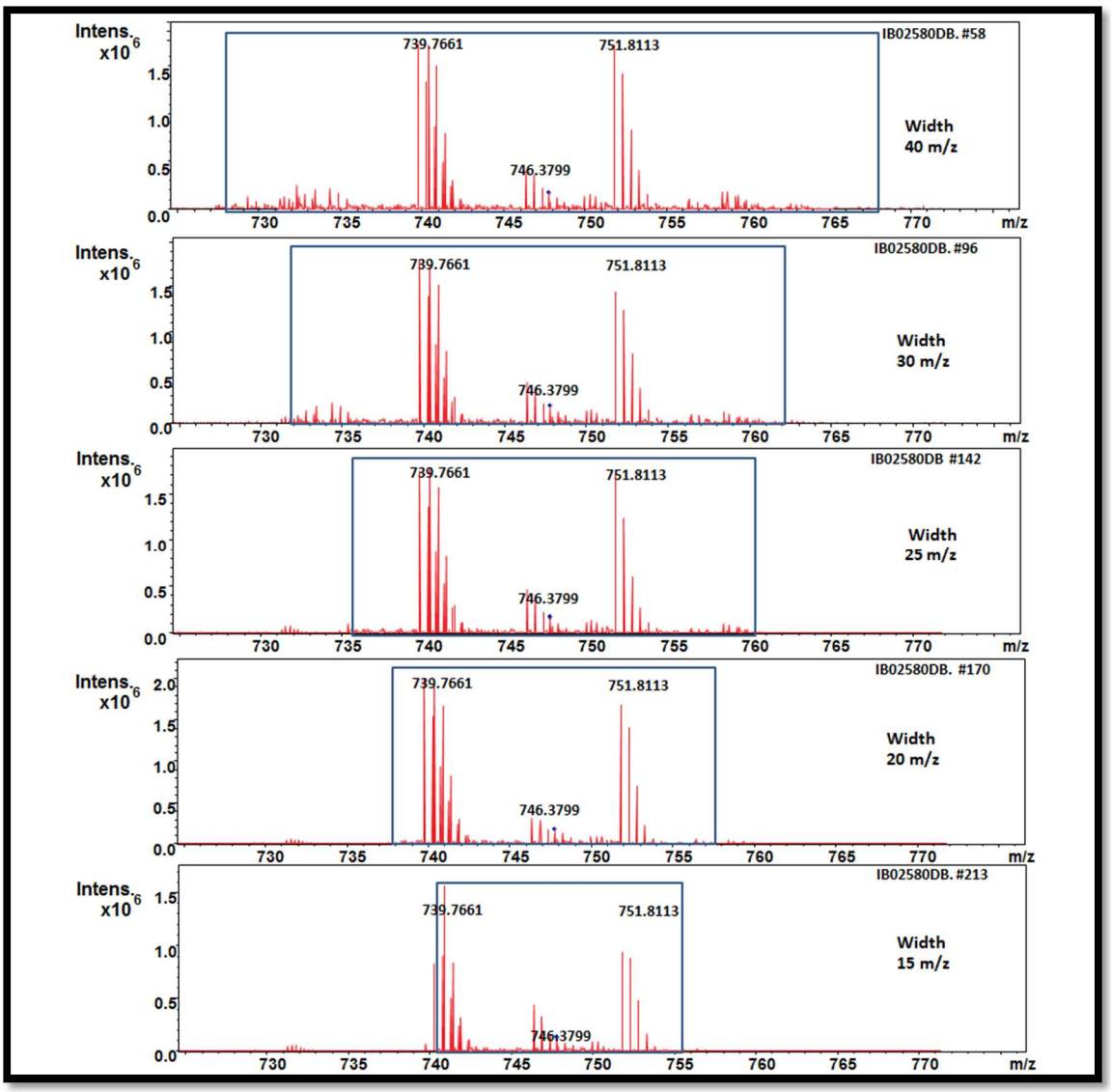


Figure 59: isolation and transmission test efficiency performed infusing a Bovine Serum Albumin (BSA) dioptgest.

The intensity of 4 peptides having m/z of: 739, 740, 746 and 751 is monitored in order to evaluate the efficiency across the isolation window.

The isolation is centered on 748 m/z in all the acquisitions displayed, starting from the top, the test is performed with a decreasing width from 40 m/z to 15 m/z and is graphically represented by the blue square.

The intensity of the four peptides are constant from the isolation width 40 m/z to the isolation width of 20 m/z, this means that the central part of the isolation window present constant efficiency.

At a width of 15 m/z both the monoisotopic peaks of the peptides 739 m/z and 740 m/z are outside the windows of selection, the first is transmitted with an intensity decreased of 50% instead the second do not present any decrease, this means that the left side of the isolation window presents the expected profile with an important drop outside the window of selection.

On the right side instead a decrease of signal is observed in the last 25% of the isolation width, as the peptide 751 m/z already present a drop of 50 % even if the boundary of selection is placed at 755 m/z.

The sample used for the test

UPS1 (Universal Proteomics Standard) is commercial equimolar mixture of 48 proteins, ranging in molecular weight from 6,000 to 83,000 daltons. To evaluate the MS1 and the Middle band capabilities of detecting proteins abundance variations in complex mixture, different amounts of UPS1 has been spiked in protein cell lysate of *Saccharomyces cerevisiae* and submitted to tryptic digest, Table 15 summarize the amount injected.

1 µg yeast +	250 attomol UPS1
1 µg yeast +	500 attomol UPS1
1 µg yeast +	1 fentomol UPS1
1 µg yeast +	5 fentomol UPS1
1 µg yeast +	10 fentomol UPS1
1 µg yeast +	25 fentomol UPS1

Table 15 the six different amount of UPS1 injected in triplicate

Optimization of the Middle band data treatment workflow

Assessment of the test

Assessing the results of the label free test just measuring the ratios of the UPS1 proteins and comparing them with the expected would have been a reductive approach. Since one of the goals of this chapter was to develop a Middle band method capable of detecting proteins regulations in complex mixture. For this reason the results have been evaluated with a statistical approach fitting a not linear mixed model.(175-177) This model, developed by Olga Vitek, is specifically adapted for the area under the curve based LC-MS quantification approaches, like SRM, Middle band and MS1 but it is not adapted for spectral counting data. The particularity of the model is to consider all the peptides measured for each protein as a replicate measurement of the same protein and to be able to take in account such redundancy information.

Applying a classic T-student test would require to average or sum the intensities of all the peptides measured for each protein, in order to have a single measurement for each protein to be submitted to the test. In the case of a peptide not precisely quantified due to interference or an artefact, the error would propagate to the protein quantification.

The linear mixed model instead it calculates the probability of having a significant variation for each peptide composing a protein. Then it estimates a global p value for each protein describing the possibilities of facing not random event. MSstat is open source software available as package for the R software at the website http://www.msstats.org/ and is also available as external tool in Skyline.

The effects of the parameters optimization has been analyzed evaluating four parameters:

1. The numbers of UPS1 proteins detected as significant regulated, considering as threshold a $\log_2$ of fold-change of >2 or <-2 and a minimal p value of 0.05.
2. The number of yeast proteins wrongly detected as regulated considering as threshold a $\log_2$ of fold-change of >2 or <-2 and a minimal p value of 0.05.
3. The global dispersion of the yeast protein considering an expected a $\log_2$ of fold-change equal to zero.

4. The whisker box plots overlap across the 18 injection.

Evaluation of the need of normalization

In a quantification experiments, it is crucial to minimize all the effects of intensity variations that are not the results of a biologic regulation. An example is the case of a signal decreasing across the replicates of injections due to a partial loss of some hydrophobic peptides, or like in the case of an artificial increase of the intensity, generated by an evaporation of the solution in the vials and a consequential concentration of the samples.

In Figure 60 are presented the 18 wishers boxes plot obtained for the six amount of UPS1 spiked and injected in triplicate, this plot is really helpful during the process of evaluation of the need of normalization. This distribution representation plots the $\log_2$ of the intensity measured for each peptide. The horizontal line in each box is the median value of the distribution, the bottom and the top of the box represent the 25^{th} and 75^{th} quartile also called Q1 and Q3. The knots represent the outliners measurements that are considerably far from the rest of the data. Defining the Interquartile Range IQR, as Q3-Q1, a value is considered outlier when is bigger than Q3 by at least 1.5 times the IQR, or smaller than Q1 by at least 1.5 times the IQR.

As displayed in Figure 60, the samples 250 attomol, 1 femtomol and 25 femtomol present an increase of the median across the three replicates that could be due to a concentration effect. The sample 5 femtomol in the third injection is globally less intense. The impact of this intensity instability across the injection should not be underestimated, especially in this data set where the 97.5% of the proteins (the yeast background) are present at exactly the same concentration.

In Figure 61 is presented the volcano plot of the comparison between the 25 femtomol (defined from now on as sample 1) against 1 femtomol (defined as sample 2 from now on). The p-values, obtained applying the linear mixed model to the 3 replicates of injection of both the samples, are plotted against the $\log_2$ of the fold-change. This type of representation allows to easily evaluate the number of proteins detected regulated and the global distribution considering as expected a $\log_2$ of the fold-change for the yeast proteins. The volcano plot so obtained displays the majority of the yeast proteins having a $\log_2$ of the fold-change negative. This results in the wrong detection of a global decrease of the intensity when comparing the sample 1 against the sample 2.

In Figure 62 and in Figure 64 and are presented the whisker box plot obtained applying respectively two type of normalization (178): the total and the quantile normalization.

- The total normalization is performed summing up all the proteins intensity in each injection and normalizing across the injection that produced the highest count.
- The quantile normalization splits each data set in four quantile and then normalizes separately each of them, this approach requires an important number of measurements and for this reason perfectly fits with this SWATH dataset where more than 80000 transition where extracted in each injection.

Both the normalization approaches succeed in obtaining a stable median across the replicates and in centering the fold-change onto the zero. But observing the volcano plot generated comparing the 25 femtomol against the 1 femtomol (Figure 64) the quantile normalization generated a less dispersed data set. For this reason from now on all the data that will be presented have been quantile normalized.

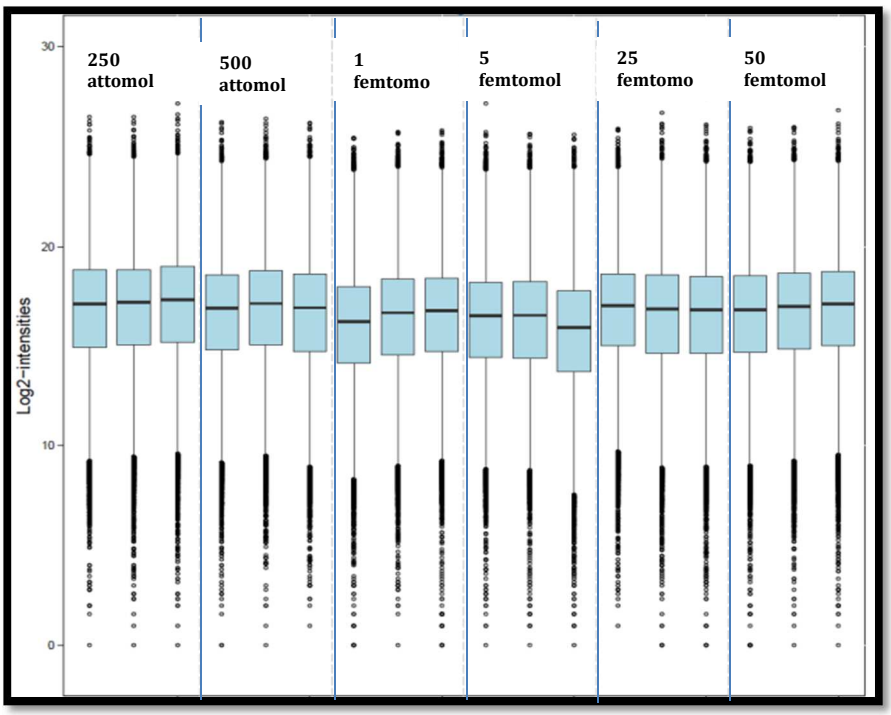


Figure 60: whisker box plot obtained without normalization for the 18 injection.

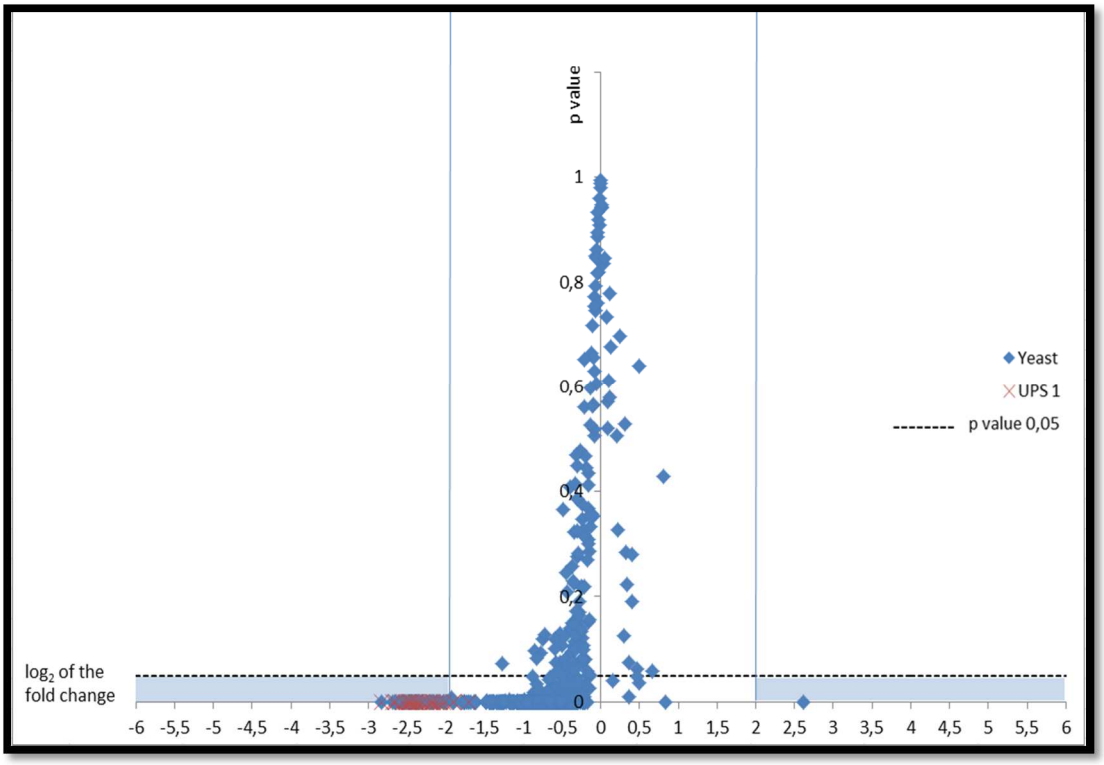


Figure 61: volcano plot obtained without normalization of the data set comparing the 25 femtomol against 1 femtomol. The p values obtained applying the linear mixed model to the 3 replicates of injection of both the samples, are plotted against the log2 of the fold-change. The proteins detected as significantly regulated are enclosed in the two blues squares defined by a log2 of fold-change of >2 or <-2 and a minimal p value of 0.05.

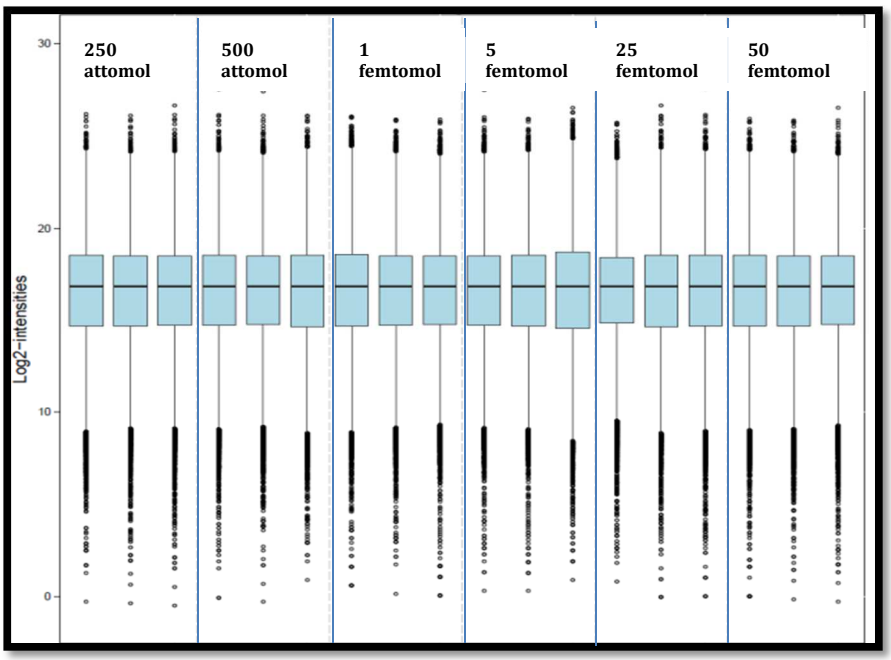


Figure 62 whisker box plot obtained applying a total normalization for the 18 injection.

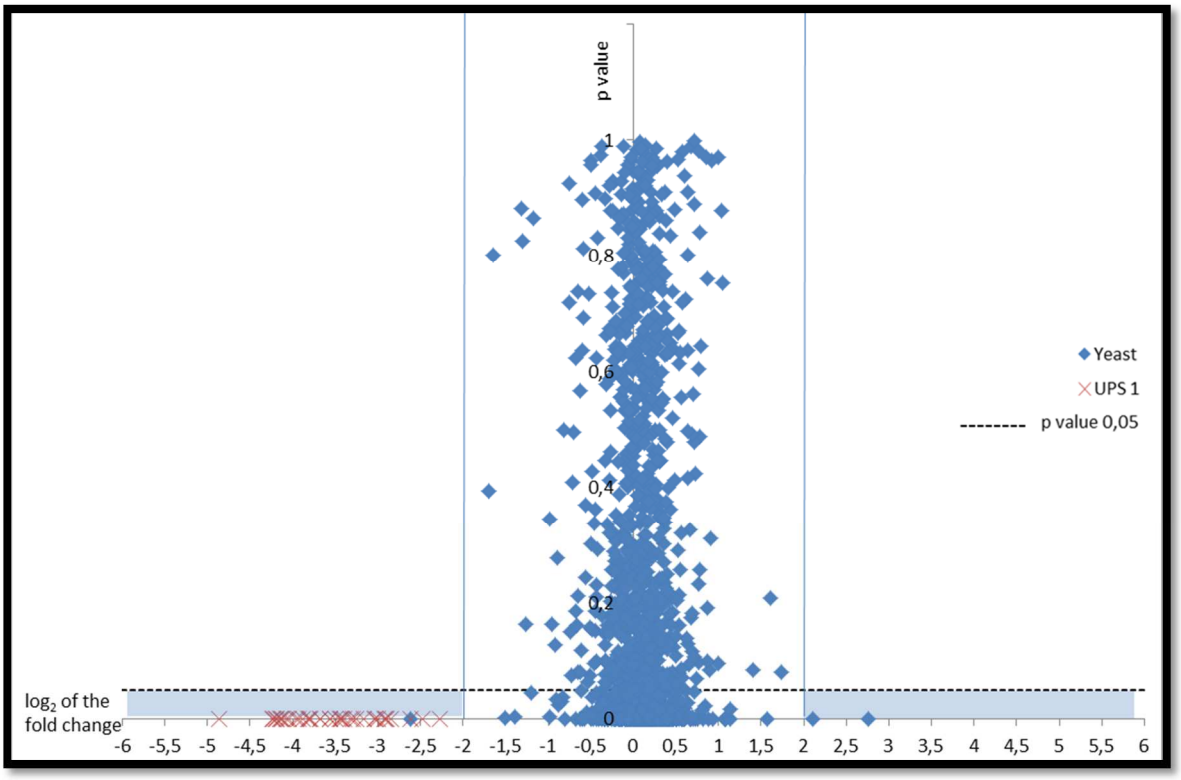


Figure 63: volcano plot of the comparison between the 25 femtomol against 1 femtomol after total normalization. The p values obtained applying the linear mixed model to the 3 replicates of injection of both the samples, are plotted against the log2 of the fold-change. The proteins detected as significantly regulated are enclosed in the two blues squares defined by a log2 of fold-change of >2 or <-2 and a minimal p value of 0.05.

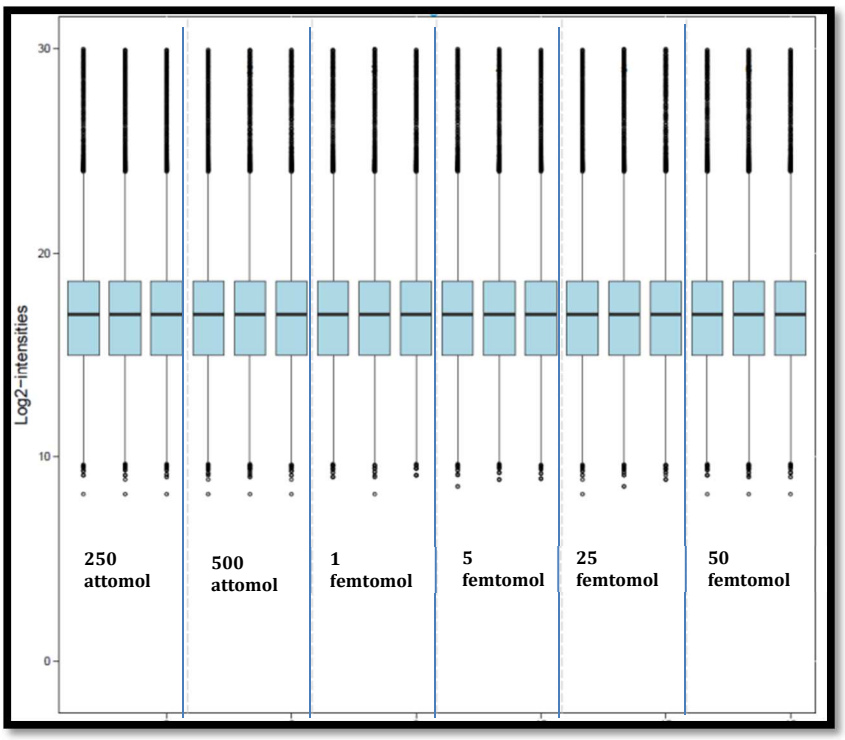


Figure 64: whisker box plot obtained applying a quantile normalization for the 18 injection.

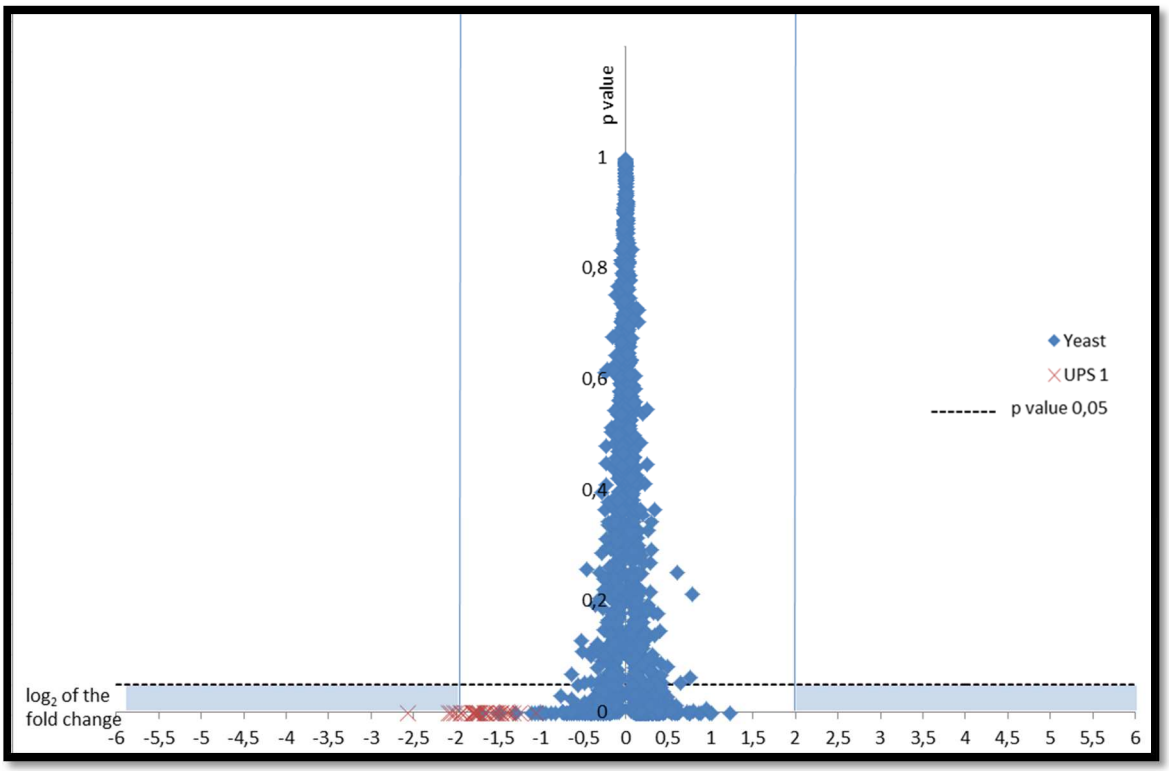


Figure 65: volcano plot of the comparison between the 25 femtomol against 1 femtomol after quantile normalization. The p values obtained applying the linear mixed model to the 3 replicates of injection of both the samples, are plotted against the log2 of the fold-change. The proteins detected as significantly regulated are enclosed in the two blues squares defined by a log2 of fold-change of >2 or <-2 and a minimal p value of 0.05.

Optimization of number of transition extracted

New task are more and more faced in the high-throughput proteomics, performant mass spectrometry are now capable of generating up to 50 MS/MS scan in one second, identifying thousand of proteins for hour. Such technological improvements also have drawbacks: strong bioinformatics skill are now imperative in proteomics, handle data set that can easily exceed one terabytes require hardware and tool demanding relevant economic investment.

On this frame the results of this paragraph have been evaluated not just considering the ability of discriminating the regulated proteins, but also the computation time required to process the data has been taken in account since it can represent an important limitation.

Despite the important parallel calculation capabilities of the server used (12 cores cpu, 24Gb RAM) the time needed to extract the chromatogram (in this fairly small test data set) is comparable to the chromatographic analysis time.

The Middle band workflow is really flexible in terms of data treatment; Skyline allows to extract a user defined numbers of transition, from one to six and to also extract the P, P+1, P+2. The results presented up to know were generated extracting 6 ions and the parent mass, in this paragraph will be presented the results of the test performed extracting and quantifying:

- 6 ions and P, P+1, P+2 Figure 65
- 6 ions Figure 66
- 3 ions Figure 67

  When comparing Figure 65 and Figure 66 it's clear that including the parent extraction in the quantification results in a worst discrimination of the UPS1 proteins form the yeast, due to a less accurate fold-change detected for the UPS1 proteins. The reason for a worst accuracy is in the less specific information carried by the parent ion chromatogram extraction respect to MS/MS based quantification.

Figure 66 and Figure 67 present the test performed extracting respectively 6 and 3 ions, in both cases without extracting the parent mass. The results obtained extracting just the most 3 intense ions presents globally a slightly increased dispersion, but UPS1 fold-change detection is improved. This plus a dramatically shortened computation time required for the chromatogram extraction made of the 3 most intense ions extraction the selected method.

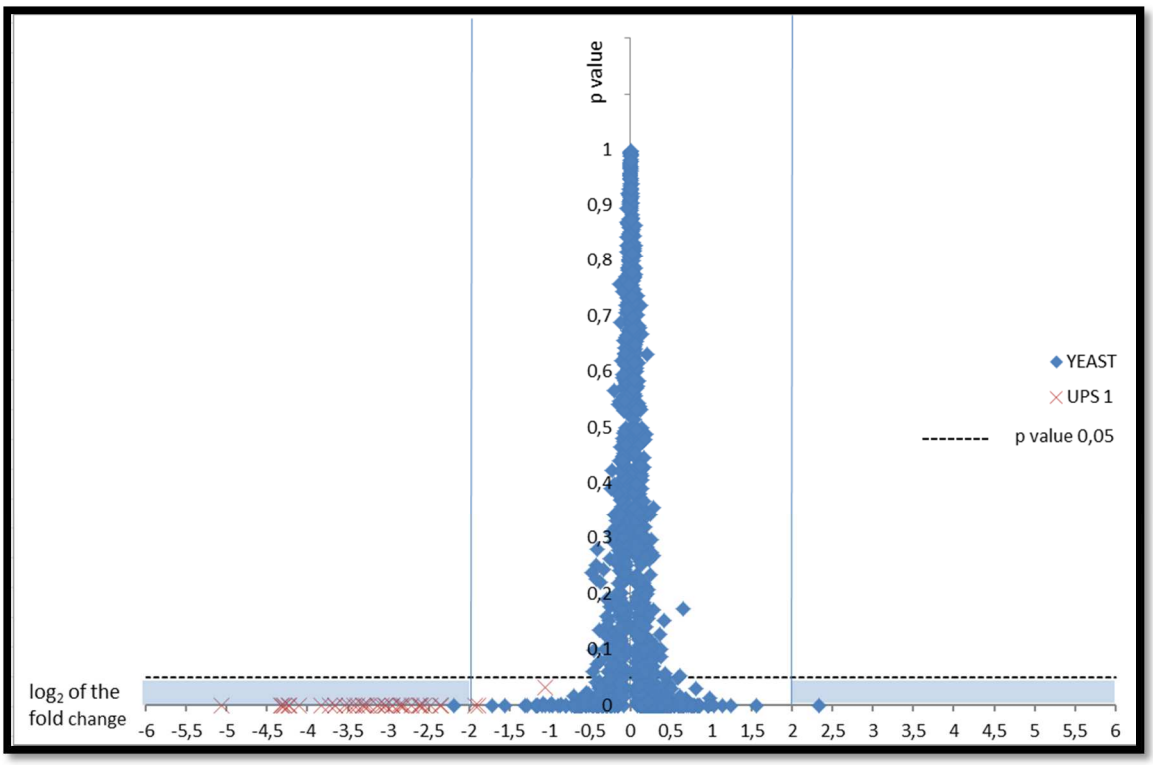


Figure 66: volcano plot of the comparison between the 25 femtomol against 1 femtomol extracting 6 transitions for peptide.

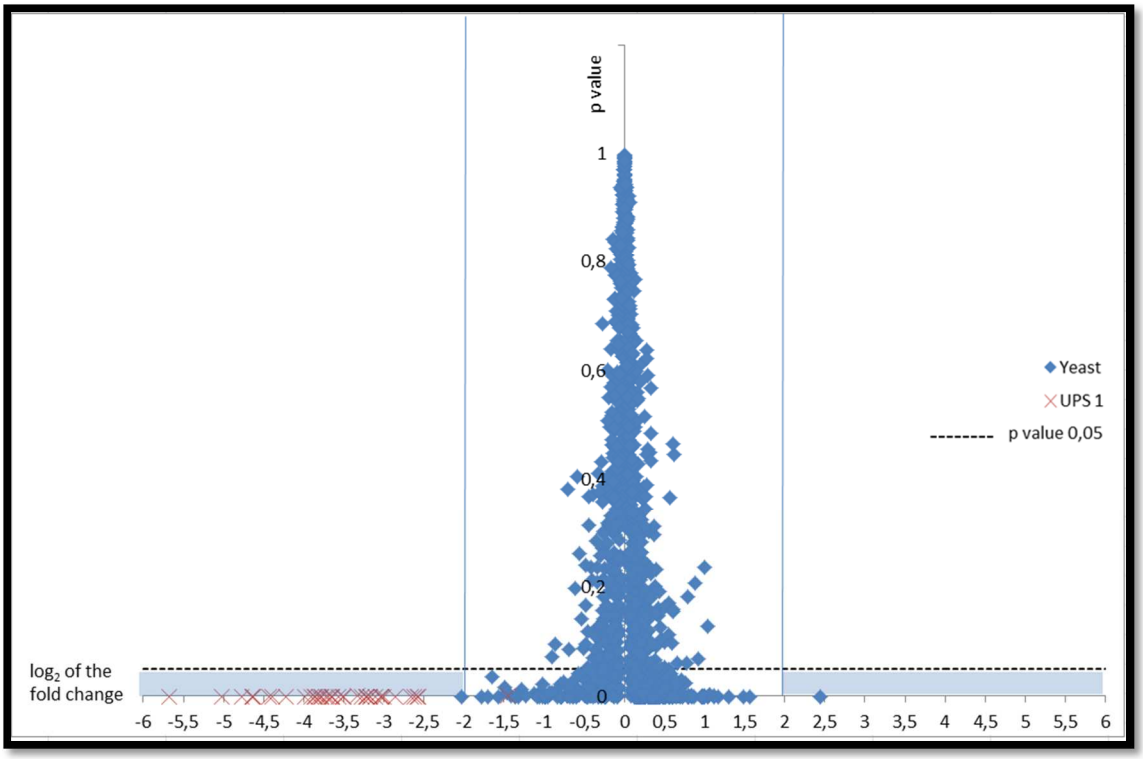


Figure 67: volcano plot of the comparison between the 25 femtomol against 1 femtomol extracting 3 transitions for peptide.

Results of the comparison Middle band and MS1

To benchmark the two techniques an additional parameter has been considered: the manual inspection of the UPS1 peptides peaks at the three lowest concentrations points. The presence in all the three replicates of injection of the peaks P, P+1 and P+2, with a signal over noise >3 has been considered as fulfilling criteria. The manual inspection represents an efficient way to assess the limit of detection. Since the algorithm of peak detection always integrate a signal in the expected retention time, even when just noise is presents, resulting in some cases in an artificially increased count of peptides detected at low concentration.

	1 µg yeast + 250 attomol UPS 1	1 µg yeast + 500 attomol UPS 1	1 µg yeast + 1 femtomol UPS 1
Number of peptide manually validated in Middle band	9	11	39
Number of peptide manually validated in DDA	9	10	19

Table 16: the two acquisitions method are compared in terms of peptide detected with signal over noise >3, all the chromatogram of the UPS 1 peptides has been manually inspected.

In Table 16 the numbers of peaks manually validated are presented, surprisingly both the approaches detect the same number of peptides at 250 and 500 attomol but at 1 femtomol the SWATH workflow detects the double of the peptides respect to the MS1 proving to be more sensitive.

Comparing the two volcano plots (Figure 67 Middle band, Figure 68 MS1) we can appreciate that both approaches present the majority of the proteins fold-change centered on the zero, but the SWATH approach can discriminate slightly better the population of the yeast proteins and the one of the UPS1.

Analyzing in detail the Middle band volcano plot all the UPS1 proteins, except one, present a p value <0.05. Appling a threshold of 2 for the $\log_2$ of the fold-change, just three UPS1 proteins are not detected as differential. Concerning the yeast proteins just two are wrongly detected as regulated.

The volcano plot obtained from the DDA instead shows 9 proteins with a fold-change < -2 and so not detected as regulated, two of them present also a p vale >0.05 and for this reason not significance. Among the protein detected as regulated in the DDA experiment also one of yeast is presents. Those errors has been manually inspected and are due to wrong peak selection of the software, caused by not resolved interference peak in the zone of the integration.

The peak selection could have been manually corrected, but the goal of this experiment was to simulate the workflow and the related issue of real label free differential quantification experiment, where no a priori information is given on the expected fold-change. Globally the Middle band performs better in terms of pure sensitivity. The higher specificity obtained by the MS/MS quantification results in a better discriminated population as displayed in the volcano plot. On the other hand the Middle band workflow requires a higher time machine due to the

need of generating the library in a classic DDA experiment. This means that in the context of maximization of the productivity of the mass spectrometer the Middle band is disadvantaged respect to the MS1.

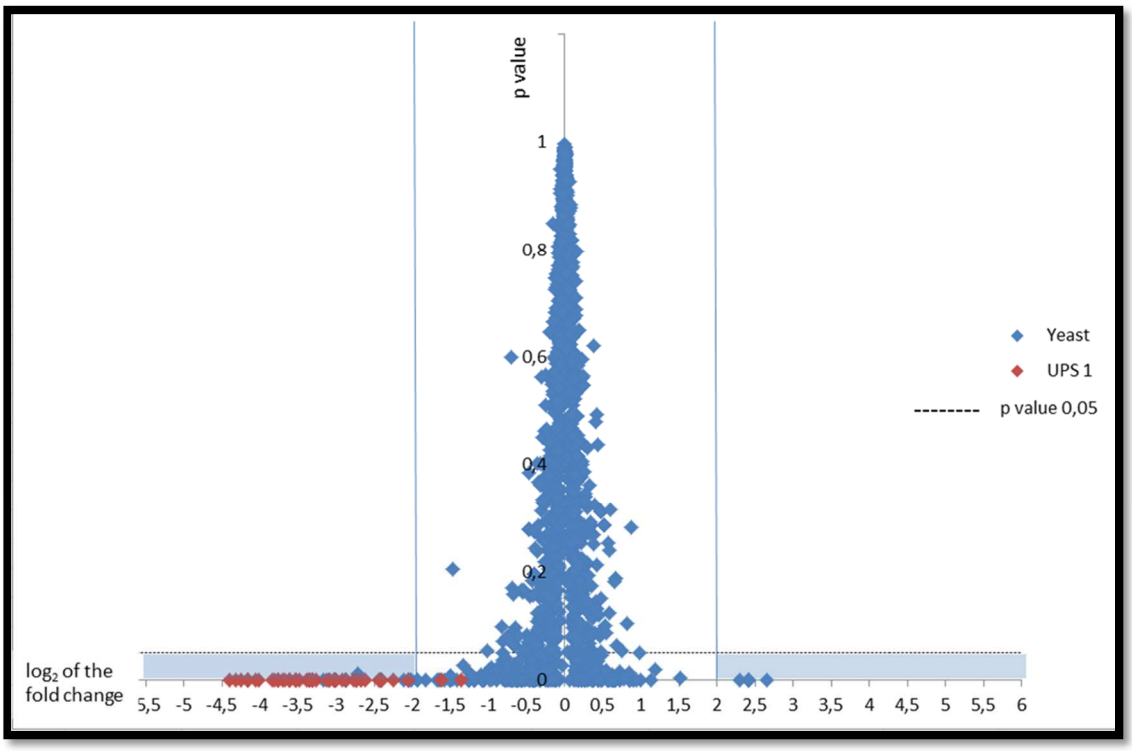


Figure 68: MS1 volcano plot of the DDA comparison between the 25 femtomol against 1 femtomol. The p values obtained applying the linear mixed model to the 3 replicates of injection of both the samples, are plotted against the log2 of the fold-change. The proteins detected as significantly regulated are enclosed in the two blues squares defined by a log2 of fold-change of >2 or <-2 and a minimal p value of 0.05.

Conclusion

In this chapter we have explored the possibilities of the DIA (Middle band) acquisition mode, and compared with the classic DDA (MS1) extensively optimized mode. We have in particular tested the DIA acquisition method on the Bruker Impact HD Q-TOF. We have shown that the key parameters for this acquisition method are: the isolation width, the scan spedd and the number of transitions extracted during the data treatment.

To evaluate the DDA and the DIA capabilities of detecting proteins abundance variations in complex mixtures, different amounts of UPS1 have been spiked in protein cell lysate of *Saccharomyces cerevisiae*

The results of the DIA workflow method development, presented in the first part of the chapter, proved that this approach is a valid label free tool. DIA is capable of discriminating, with statistical confidence and without requiring manual inspection of the chromatogram, relevant protein regulation, like in the case presented in Figure 67, where the sample spiked with 25 femtomols was compared to the one spiked with 1 femtomol.

Concerning the other amount spiked (Table 15) the DIA approach is not capable of completely discriminating the two populations of proteins and it requires the manual intervention and critical validation of a reasonable number of proteins candidate.

Despite the recent advances the label free approach can highlight just major differences and it represents a complementary technique to the absolute quantification (selected reaction monitoring SRM).

In the second part of the chapter it has been proposed a comparison of DIA and MS1 label free. As discussed in the results paragraph, the DIA obtained an higher sensibility in terms of UPS1 peptides detected at the three low UPS1 spike points. DIA also generated more robust results thanks to the higher specificity brought by the MS/MS based quantification. The results presented in the comparison do not completely justify the additional time required to generate spectral library in DIA. For this reason until no search engine will be available for this approach, the DDA label free will represent the optimal choise in a differential label free experiment.

General conclusion

General conclusion

The objective of my Ph.D. thesis was to develop and optimize new methodologies and analytical approaches to improve the performances of the mass spectrometry based proteomics. The long term goal is the development of methods capable of achieving a high confidence full characterization of all proteins present in a proteome. Such an achievement, still out of grasp today, will permit to open the door to detailed comparison of different proteomes with an accurate quantification of each protein.

The first part of my thesis has focused on the development of the N-termini proteomics, in the context of a better characterization of the proteome, which was presented in the two first chapters.

1. N-termini characterization of *Plasmodium falciparum* exported protein.

In the first chapter, I presented the method optimization and the results, obtained in the study of the Host Cell Targeting signal (HCT) proteins exportation system in *P. falciparum.* The optimization performed, allowed characterizing the N-termini position of unknown amount of low concentrated proteins prefractionated on a SDS PAGE.

Thanks to the use of an "internal standard of staining intensity" when running the SDS-PAGE the intensities of the chimeric constructs were related to the intensity of the standard, allowing minimizing the side reaction of the TMPP due to unbalanced ratio protein/TMPP.

2. Development of a stable isotope labeling method to validate N-terminal TMPP derivatized peptide.

In the second chapter, I presented the dN-TOP approach which is an improvement of the original N-TOP strategy. Based on two TMPP isotopologues (light and heavy) derivatization, the process of the N-terminal peptides validation is now automatized. This fast and robust data validation of the dN-TOP opens the door to high-throughput N-termini characterization study. This strategy has been successfully applied to the characterization of *H. arsenicoxydans* proteome and its N-terminome.

In the second part of my thesis, I presented the results of the method optimization focused on a crucial mass spectrometry aspect, the quantification. Five different protein quantification techniques were addressed in five chapters, among which four at the peptide level and one at the protein level.

1. Label free quantification of proteins complex subunits using spectral count approach.

In this chapter, proteomic analysis was used to study the effect of the yeast histone H2B ubiquitylation on the RNA nuclear exportation.

Thanks to a series of optimization performed on the mass spectrometry acquisition methods (the exclusion time and the redundancy of the spectral counts), it was possible to achieve the sensitivity necessary to differentially quantitate proteins complex present in low amount and in complex matrix.

2. Targeted Selected Reaction Monitoring (SRM) based prion protein quantification

In this chapter, two projects on SRM based quantification were presented, both targeting the human prion protein, which in case of miss folding cause the transmissible spongiform encephalopathy (TSEs).

- I have improved the LC-SRM method, already used in our laboratory in a first study, where an absolute quantification of the prion protein was performed in pharmaceutical products urine derived injectable fertility products.
  I presented the method development performed to decrease the Limit Of Detection (LOD) of the original approach. Thanks to this improvement it has been possible to quantify the prion protein in pharmaceutical products defined as prion free with the first method.
- I have developed an "in gel" chemical digestion using cyanogen bromide (CNBr) based relative quantification of prion polymorphic variants, which differ only from one amino acid and do not present tryptic proteotypic peptide. This method will allow investigating the hypothesis of the preferential accumulation of one of the two prion polymorphic variants in biopsies of human brain TSEs affected.

3. Monoclonal IgG glycation batch to batch quantification.

In this chapter, it is described the development of a quantitation method of protein glycation. Glycation is a process which can affect therapeutic recombinant products. The extent of glycation must be accurately quantified, because glycated proteins could be immunogenic. The method development, particularly focused on the sample preparation and on the data treatment, allowed us to differentially quantify a chemically heterogeneous family of PTM present at contaminant level. The application of this method in different production batches, allowed assessing the reproducibility of the batch production.

4. Intact protein label free quantification: future developments.

The accurate quantitative determination of the minor form hemoglobin A2 ($\alpha 2 \delta 2$) is required for the diagnosis of the beta-thalassemia. In this frame I developed a fast and precise label free method, based on the relative quantification of the hemoglobin chains at protein level, approach often considered not suitable for accurate quantification.

5. Optimization of a Data Independent Acquisition (DIA) method.

I have improved the data treatment of a DIA method and increased the sensitivity and the specificity of this approach. To assess the performance of the optimized DIA method, I have performed a comparative study (on a Q-TOF Impact HD, Bruker) to evaluate the benefit of DIA over DDA MS1 for label free quantification in complex proteins mixture.

The proteomic analysis is, in essence, much more complex that the analysis of genomes. The quantitative analysis of a mixture of thousands of proteins, corresponding to a defined state, is obviously tremendously more challenging than the determination of the sequence of a unique DNA, molecule which does not change with time.

Therefore, despite the constant effort of the scientific community, proteomics is far from being a "mature" analytic method. Nevertheless, we can reasonably expect many other future developments which will allow achieving the sensitivity and the dynamic range needed to fully characterize and quantify all the proteins present in a proteome.

Twenty years ago, nobody would have imagined that a human genome could be sequenced in less than one day, as it is the case today. If the efforts for the development of new analytical methods and strategies are maintained, proteomics will certainly bring answers to many questions about the role of proteins in biology which seems today too complex to be addressed.

References

1. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, Van Dorsselaer A, Poch O, Schaeffer C and Lecompte O: Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res 19: 128-35, 2009.
2. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF and Williams KL: Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev 13: 19-50, 1996.
3. Huber LA: Is proteomics heading in the wrong direction? Nat Rev Mol Cell Biol 4: 74-80, 2003.
4. Pirmoradian M, Budamgunta H, Chingin K, Zhang B, Astorga-Wells J and Zubarev RA: Rapid and deep human proteome analysis by single-dimension shotgun proteomics. Mol Cell Proteomics 12: 3330-8, 2013.
5. Karve TM and Cheema AK: Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. J Amino Acids 2011: 207691, 2011.
6. Zhao Y and Jensen ON: Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. Proteomics 9: 4632-41, 2009.
7. Rawlings ND, Waller M, Barrett AJ and Bateman A: MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 42: D503-9, 2014.
8. Lopez-Otin C and Hunter T: The regulatory crosstalk between kinases and proteases in cancer. Nat Rev Cancer 10: 278-292, 2010.
9. Picotti P and Aebersold R: Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods 9: 555-66, 2012.
10. Ong SE and Mann M: Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 1: 252-62, 2005.
11. Huesgen PF, Lange PF and Overall CM: Ensembles of protein termini and specific proteolytic signatures as candidate biomarkers of disease. Proteomics Clin Appl 8: 338-50, 2014.
12. Lopez-Otin C and Overall CM: Protease degradomics: a new challenge for proteomics. Nat Rev Mol Cell Biol 3: 509-19, 2002.
13. Lange PF and Overall CM: Protein TAILS: when termini tell tales of proteolysis and function. Curr Opin Chem Biol 17: 73-82, 2013.
14. Tam EM, Morrison CJ, Wu YI, Stack MS and Overall CM: Membrane protease proteomics: Isotope-coded affinity tag MS identification of undescribed MT1-matrix metalloproteinase substrates. Proc Natl Acad Sci U S A 101: 6917-22, 2004.
15. Dean RA, Butler GS, Hamma-Kourbali Y, Delbe J, Brigstock DR, Courty J and Overall CM: Identification of candidate angiogenic inhibitors processed by matrix metalloproteinase 2 (MMP-2) in cell-based proteomic screens: disruption of vascular endothelial growth factor (VEGF)/heparin affin regulatory peptide (pleiotrophin) and VEGF/Connective tissue growth factor angiogenic inhibitory complexes by MMP-2 proteolysis. Mol Cell Biol 27: 8454-65, 2007.
16. van den Berg BH and Tholey A: Mass spectrometry-based proteomics strategies for protease cleavage site identification. Proteomics 12: 516-29, 2012.
17. Holmberg A, Blomstergren A, Nord O, Lukacs M, Lundeberg J and Uhlen M: The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. Electrophoresis 26: 501-10, 2005.
18. Timmer JC, Enoksson M, Wildfang E, Zhu W, Igarashi Y, Denault JB, Ma Y, Dummitt B, Chang YH, Mast AE, Eroshkin A, Smith JW, Tao WA and Salvesen GS: Profiling constitutive proteolytic events in vivo. Biochem J 407: 41-8, 2007.
19. Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL and Wells JA: Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. Cell 134: 866-76, 2008.

20. Polevoda B and Sherman F: Nalpha -terminal acetylation of eukaryotic proteins. J Biol Chem 275: 36479-82, 2000.
21. Xu G, Shin SB and Jaffrey SR: Global profiling of protease cleavage sites by chemoselective labeling of protein N-termini. Proc Natl Acad Sci U S A 106: 19310-5, 2009.
22. Xu G and Jaffrey SR: N-CLAP: global profiling of N-termini by chemoselective labeling of the alpha-amine of proteins. Cold Spring Harb Protoc 2010: pdb prot5528, 2010.
23. Edman P: A method for the determination of amino acid sequence in peptides. Arch Biochem 22: 475, 1949.
24. Prudova A, auf dem Keller U, Butler GS and Overall CM: Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. Mol Cell Proteomics 9: 894-911, 2010.
25. Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR and Vandekerckhove J: Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol 21: 566-9, 2003.
26. Rogers LD and Overall CM: Proteolytic post-translational modification of proteins: proteomic tools and methodology. Mol Cell Proteomics 12: 3532-42, 2013.
27. Arnesen T, Van Damme P, Polevoda B, Helsens K, Evjenth R, Colaert N, Varhaug JE, Vandekerckhove J, Lillehaug JR, Sherman F and Gevaert K: Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. Proc Natl Acad Sci U S A 106: 8157-62, 2009.
28. Staes A, Van Damme P, Helsens K, Demol H, Vandekerckhove J and Gevaert K: Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). Proteomics 8: 1362-70, 2008.
29. Bland C, Hartmann EM, Christie-Oleza JA, Fernandez B and Armengaud J: N-Terminal-oriented proteogenomics of the marine bacterium roseobacter denitrificans Och114 using N-Succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP) labeling and diagonal chromatography. Mol Cell Proteomics 13: 1369-81, 2014.
30. McDonald L, Robertson DH, Hurst JL and Beynon RJ: Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. Nat Methods 2: 955-7, 2005.
31. Wagner DS, Salari A, Gage DA, Leykam J, Fetter J, Hollingsworth R and Watson JT: Derivatization of peptides to enhance ionization efficiency and control fragmentation during analysis by fast atom bombardment tandem mass spectrometry. Biol Mass Spectrom 20: 419-25, 1991.
32. Ayoub D: Vers une étude approfondie des protéomes: caractérisation des extrémités N-terminales des protéines. Ph.D. thesis Ecole doctorales des sciences chimiques UMR7 178-IPHC: 2012.
33. Lange PF, Huesgen PF, Nguyen K and Overall CM: Annotating N termini for the human proteome project: N termini and Nalpha-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. J Proteome Res 13: 2028-44, 2014.
34. World Health Organization (WHO): world malaria report 2011 http://www.who.int/mediacentre/news/releases/2011/malaria_report_20111213/en/.
35. White NJ, Pukrittayakamee S, Hien TT, Faiz MA, Mokuolu OA and Dondorp AM: Malaria. Lancet 383: 723-35, 2014.
36. Dluzewski AR, Mitchell GH, Fryer PR, Griffiths S, Wilson RJ and Gratzer WB: Origins of the parasitophorous vacuole membrane of the malaria parasite, Plasmodium falciparum, in human red blood cells. J Cell Sci 102 (Pt 3): 527-32, 1992.
37. McPherson RA, Donald DR, Sawyer WH and Tilley L: Proteolytic digestion of band 3 at an external site alters the erythrocyte membrane organisation and may facilitate malarial invasion. Mol Biochem Parasitol 62: 233-42, 1993.
38. Marti M, Good RT, Rug M, Knuepfer E and Cowman AF: Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 306: 1930-3, 2004.

39. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, Lopez-Estrano C and Haldar K: A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science 306: 1934-7, 2004.
40. Przyborski JM, Miller SK, Pfahler JM, Henrich PP, Rohrbach P, Crabb BS and Lanzer M: Trafficking of STEVOR to the Maurer's clefts in Plasmodium falciparum-infected erythrocytes. EMBO J 24: 2306-17, 2005.
41. Joannin N, Kallberg Y, Wahlgren M and Persson B: RSpred, a set of Hidden Markov Models to detect and classify the RIFIN and STEVOR proteins of Plasmodium falciparum. BMC Genomics 12: 119, 2011.
42. Crabb BS, de Koning-Ward TF and Gilson PR: Protein export in Plasmodium parasites: from the endoplasmic reticulum to the vacuolar export machine. Int J Parasitol 40: 509-13, 2010.
43. Finishing the euchromatic sequence of the human genome. Nature 431: 931-45, 2004.
44. Renuse S, Chaerkady R and Pandey A: Proteogenomics. Proteomics 11: 620-30, 2011.
45. Release 2014_07 of 09-Jul-14 of UniProtKB/Swiss-Prot.
46. Bertaccini D, Vaca S, Carapito C, Arsene-Ploetze F, Van Dorsselaer A and Schaeffer-Reiss C: An improved stable isotope N-terminal labeling approach with light/heavy TMPP to automate proteogenomics data validation: dN-TOP. J Proteome Res 12: 3063-70, 2013.
47. He Y, Parthasarathi R, Raghavachari K and Reilly JP: Photodissociation of charge tagged peptides. J Am Soc Mass Spectrom 23: 1182-90, 2012.
48. Gucinski AC, Dodds ED, Li W and Wysocki VH: Understanding and exploiting Peptide fragment ion intensities using experimental and informatic approaches. Methods Mol Biol 604: 73-94, 2010.
49. Bocs S, Cruveiller S, Vallenet D, Nuel G and Medigue C: AMIGene: Annotation of MIcrobial Genes. Nucleic Acids Res 31: 3723-6, 2003.
50. Vaudel M, Sickmann A and Martens L: Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics. Biochim Biophys Acta 1844: 12-20, 2014.
51. Domon B and Aebersold R: Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol 28: 710-21, 2010.
52. Unlu M, Morgan ME and Minden JS: Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis 18: 2071-7, 1997.
53. Burniston JG, Kenyani J, Gray D, Guadagnin E, Jarman IH, Cobley JN, Cuthbertson DJ, Chen YW, Wastling JM, Lisboa PJ, Koch LG and Britton SL: Conditional independence mapping of DIGE data reveals PDIA3 protein species as key nodes associated with muscle aerobic capacity. J Proteomics: 2014.
54. Lihong H, Linlin G, Yiping G, Yang S, Xiaoyu Q, Zhuzhu G, Xiaohan Y, Xin Z, Liyan X and Shujuan S: Proteomics Approaches for Identification of Tumor Relevant Protein Targets in Pulmonary Squamous Cell Carcinoma by 2D-DIGE-MS. PLoS One 9: e95121, 2014.
55. DeSouza LV and Siu KW: Mass spectrometry-based quantification. Clin Biochem 46: 421-31, 2013.
56. Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, Lee A, van Sluyter SC and Haynes PA: Less label, more free: approaches in label-free quantitative mass spectrometry. Proteomics 11: 535-53, 2011.
57. Nilsson T, Mann M, Aebersold R, Yates JR, 3rd, Bairoch A and Bergeron JJ: Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods 7: 681-5, 2010.
58. Bantscheff M, Lemeer S, Savitski MM and Kuster B: Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem 404: 939-65, 2012.
59. Bantscheff M, Schirle M, Sweetman G, Rick J and Kuster B: Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389: 1017-31, 2007.
60. Nahnsen S, Bielow C, Reinert K and Kohlbacher O: Tools for label-free peptide quantification. Mol Cell Proteomics 12: 549-56, 2013.
61. America AH and Cordewener JH: Comparative LC-MS: a landscape of peaks and valleys. Proteomics 8: 731-49, 2008.

62. Lundgren DH, Hwang SI, Wu L and Han DK: Role of spectral counting in quantitative proteomics. Expert Rev Proteomics 7: 39-53, 2010.
63. Sandin M, Teleman J, Malmstrom J and Levander F: Data processing methods and quality control strategies for label-free LC-MS protein quantification. Biochim Biophys Acta 1844: 29-41, 2014.
64. Cox J and Mann M: MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26: 1367-72, 2008.
65. Patel VJ, Thalassinos K, Slade SE, Connolly JB, Crombie A, Murrell JC and Scrivens JH: A comparison of labeling and label-free mass spectrometry-based proteomics approaches. J Proteome Res 8: 3752-9, 2009.
66. Perez-Riverol Y, Wang R, Hermjakob H, Muller M, Vesada V and Vizcaino JA: Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective. Biochim Biophys Acta 1844: 63-76, 2014.
67. Schilling B, Rardin MJ, MacLean BX, Zawadzka AM, Frewen BE, Cusack MP, Sorensen DJ, Bereman MS, Jing E, Wu CC, Verdin E, Kahn CR, Maccoss MJ and Gibson BW: Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. Mol Cell Proteomics 11: 202-14, 2012.
68. Thiele H, Glandorf J and Hufnagel P: Bioinformatics strategies in life sciences: from data processing and data warehousing to biological knowledge extraction. J Integr Bioinform 7: 141, 2010.
69. Searle BC: Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics 10: 1265-9, 2010.
70. Purvine S, Eppel JT, Yi EC and Goodlett DR: Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. Proteomics 3: 847-50, 2003.
71. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P and Geromanos S: Quantitative proteomic analysis by accurate mass retention time pairs. Anal Chem 77: 2187-200, 2005.
72. Hummel M, Cordewener JH, de Groot JC, Smeekens S, America AH and Hanson J: Dynamic protein composition of Arabidopsis thaliana cytosolic ribosomes in response to sucrose feeding as revealed by label free MSE proteomics. Proteomics 12: 1024-38, 2012.
73. Pizzatti L, Panis C, Lemos G, Rocha M, Cecchini R, Souza GH and Abdelhay E: Label-free MSE proteomic analysis of chronic myeloid leukemia bone marrow plasma: disclosing new insights from therapy resistance. Proteomics 12: 2618-31, 2012.
74. Dator RP, Gaston KW and Limbach PA: Multiple Enzymatic Digestions and Ion Mobility Separation Improve Quantification of Bacterial Ribosomal Proteins by Data Independent Acquisition Liquid Chromatography-Mass Spectrometry. Anal Chem: 2014.
75. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R and Aebersold R: Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 11: O111 016717, 2012.
76. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A and Mann M: Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1: 376-86, 2002.
77. Ong SE, Foster LJ and Mann M: Mass spectrometric-based approaches in quantitative proteomics. Methods 29: 124-30, 2003.
78. Ong SE, Kratchmarova I and Mann M: Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). J Proteome Res 2: 173-81, 2003.
79. Scholten A, Mohammed S, Low TY, Zanivan S, van Veen TA, Delanghe B and Heck AJ: In-depth quantitative cardiac proteomics combining electron transfer dissociation and the metalloendopeptidase Lys-N with the SILAC mouse. Mol Cell Proteomics 10: O111 008474, 2011.

80. Sury MD, Chen JX and Selbach M: The SILAC fly allows for accurate protein quantification in vivo. Mol Cell Proteomics 9: 2173-83, 2010.
81. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R: Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17: 994-9, 1999.
82. Li J, Steen H and Gygi SP: Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response. Mol Cell Proteomics 2: 1198-204, 2003.
83. Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK and Hamon C: Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75: 1895-904, 2003.
84. Mertins P, Udeshi ND, Clauser KR, Mani DR, Patel J, Ong SE, Jaffe JD and Carr SA: iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. Mol Cell Proteomics 11: M111 014423, 2012.
85. Carr SA, Abbatiello SE, Ackermann BL, Borchers C, Domon B, Deutsch EW, Grant RP, Hoofnagle AN, Huttenhain R, Koomen JM, Liebler DC, Liu T, MacLean B, Mani DR, Mansfield E, Neubert H, Paulovich AG, Reiter L, Vitek O, Aebersold R, Anderson L, Bethem R, Blonder J, Boja E, Botelho J, Boyne M, Bradshaw RA, Burlingame AL, Chan D, Keshishian H, Kuhn E, Kinsinger C, Lee JS, Lee SW, Moritz R, Oses-Prieto J, Rifai N, Ritchie J, Rodriguez H, Srinivas PR, Townsend RR, Van Eyk J, Whiteley G, Wiita A and Weintraub S: Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. Mol Cell Proteomics 13: 907-17, 2014.
86. Domon B: Considerations on selected reaction monitoring experiments: implications for the selectivity and accuracy of measurements. Proteomics Clin Appl 6: 609-14, 2012.
87. Method of the Year 2012. Nat Meth 10: 1-1, 2013.
88. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B: A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17: 1030-2, 1999.
89. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M and Seraphin B: The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods 24: 218-29, 2001.
90. Gingras AC, Gstaiger M, Raught B and Aebersold R: Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol 8: 645-54, 2007.
91. Lambert J-P, Ivosev G, Couzens AL, Larsen B, Taipale M, Lin Z-Y, Zhong Q, Lindquist S, Vidal M, Aebersold R, Pawson T, Bonner R, Tate S and Gingras A-C: Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. Nat Meth 10: 1239-1245, 2013.
92. Collins BC, Gillet LC, Rosenberger G, Rost HL, Vichalkovski A, Gstaiger M and Aebersold R: Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nat Meth 10: 1246-1253, 2013.
93. Herissant L, Moehle EA, Bertaccini D, Van Dorsselaer A, Schaeffer-Reiss C, Guthrie C and Dargemont C: H2B ubiquitylation modulates spliceosome assembly and function in budding yeast. Biol Cell 106: 126-38, 2014.
94. Helin K and Dhanak D: Chromatin proteins and modifications as drug targets. Nature 502: 480-8, 2013.
95. Soldi M and Bonaldi T: The proteomic investigation of chromatin functional domains reveals novel synergisms among distinct heterochromatin components. Mol Cell Proteomics 12: 764-80, 2013.
96. Berger SL: The complex language of chromatin regulation during transcription. Nature 447: 407-12, 2007.
97. Vitaliano-Prunier A, Menant A, Hobeika M, Geli V, Gwizdek C and Dargemont C: Ubiquitylation of the COMPASS component Swd2 links H2B ubiquitylation to H3K4 trimethylation. Nat Cell Biol 10: 1365-71, 2008.
98. Vitaliano-Prunier A, Babour A, Herissant L, Apponi L, Margaritis T, Holstege FC, Corbett AH, Gwizdek C and Dargemont C: H2B ubiquitylation controls the formation of export-competent mRNP. Mol Cell 45: 132-9, 2012.

99. Anderson NL and Anderson NG: The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 1: 845-67, 2002.
100. Collins BC, Gillet LC, Rosenberger G, Rost HL, Vichalkovski A, Gstaiger M and Aebersold R: Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nat Methods 10: 1246-53, 2013.
101. Heumann KG: Isotope dilution mass spectrometry (IDMS) of the elements. Mass Spectrometry Reviews 11: 41-67, 1992.
102. Zakett D, Flynn RGA and Cooks RG: Chlorine isotope effects in mass spectrometry by multiple reaction monitoring. The Journal of Physical Chemistry 82: 2359-2362, 1978.
103. Canas B, Lopez-Ferrer D, Ramos-Fernandez A, Camafeita E and Calvo E: Mass spectrometry technologies for proteomics. Brief Funct Genomic Proteomic 4: 295-320, 2006.
104. Magiera S: Fast, simultaneous quantification of three novel cardiac drugs in human urine by MEPS-UHPLC-MS/MS for therapeutic drug monitoring. J Chromatogr B Analyt Technol Biomed Life Sci 938: 86-95, 2013.
105. Surinova S, Schiess R, Huttenhain R, Cerciello F, Wollscheid B and Aebersold R: On the development of plasma protein biomarkers. J Proteome Res 10: 5-16, 2011.
106. Huillet C, Adrait A, Lebert D, Picard G, Trauchessec M, Louwagie M, Dupuis A, Hittinger L, Ghaleb B, Le Corvoisier P, Jaquinod M, Garin J, Bruley C and Brun V: Accurate quantification of cardiovascular biomarkers in serum using Protein Standard Absolute Quantification (PSAQ) and selected reaction monitoring. Mol Cell Proteomics 11: M111 008235, 2012.
107. Ressom HW, Xiao JF, Tuli L, Varghese RS, Zhou B, Tsai TH, Ranjbar MR, Zhao Y, Wang J, Di Poto C, Cheema AK, Tadesse MG, Goldman R and Shetty K: Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis. Anal Chim Acta 743: 90-100, 2012.
108. Gillette MA and Carr SA: Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. Nat Methods 10: 28-34, 2013.
109. Castro-Gamero AM, Izumi C and Rosa JC: Biomarker verification using selected reaction monitoring and shotgun proteomics. Methods Mol Biol 1156: 295-306, 2014.
110. Kuzyk MA, Smith D, Yang J, Cross TJ, Jackson AM, Hardie DB, Anderson NL and Borchers CH: Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. Mol Cell Proteomics 8: 1860-77, 2009.
111. Kim YJ, Zaidi-Ainouch Z, Gallien S and Domon B: Mass spectrometry-based detection and quantification of plasma glycoproteins using selective reaction monitoring. Nat Protoc 7: 859-71, 2012.
112. Lin D, Alborn WE, Slebos RJ and Liebler DC: Comparison of protein immunoprecipitation-multiple reaction monitoring with ELISA for assay of biomarker candidates in plasma. J Proteome Res 12: 5996-6003, 2013.
113. Wei J, Ding C, Zhang J, Mi W, Zhao Y, Liu M, Fu T, Zhang Y, Ying W, Cai Y, Qin J and Qian X: High-throughput absolute quantification of proteins using an improved two-dimensional reversed-phase separation and quantification concatemer (QconCAT) approach. Anal Bioanal Chem: 2014.
114. Gallien S, Bourmaud A, Kim SY and Domon B: Technical considerations for large-scale parallel reaction monitoring analysis. J Proteomics 100: 147-59, 2014.
115. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B and Aebersold R: Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 25: 125-31, 2007.
116. Fusaro VA, Mani DR, Mesirov JP and Carr SA: Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol 27: 190-8, 2009.
117. Gallien S, Duriez E and Domon B: Selected reaction monitoring applied to proteomics. J Mass Spectrom 46: 298-312, 2011.
118. Mead JA, Bianco L, Ottone V, Barton C, Kay RG, Lilley KS, Bond NJ and Bessant C: MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. Mol Cell Proteomics 8: 696-705, 2009.

119. Lange V, Malmstrom JA, Didion J, King NL, Johansson BP, Schafer J, Rameseder J, Wong CH, Deutsch EW, Brusniak MY, Buhlmann P, Bjorck L, Domon B and Aebersold R: Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. Mol Cell Proteomics 7: 1489-500, 2008.
120. Reiter L, Rinner O, Picotti P, Huttenhain R, Beck M, Brusniak MY, Hengartner MO and Aebersold R: mProphet: automated data processing and statistical validation for large-scale SRM experiments. Nat Methods 8: 430-5, 2011.
121. Van Dorsselaer A, Carapito C, Delalande F, Schaeffer-Reiss C, Thierse D, Diemer H, McNair DS, Krewski D and Cashman NR: Detection of prion protein in urine-derived injectable fertility products by a targeted proteomic approach. PLoS One 6: e17815, 2011.
122. Prusiner SB: Novel proteinaceous infectious particles cause scrapie. Science 216: 136-44, 1982.
123. Prusiner SB: Prions. Proc Natl Acad Sci U S A 95: 13363-83, 1998.
124. Prusiner SB: The prion diseases. Brain Pathol 8: 499-513, 1998.
125. Prusiner SB, Scott MR, DeArmond SJ and Cohen FE: Prion protein biology. Cell 93: 337-48, 1998.
126. Will RG: Acquired prion disease: iatrogenic CJD, variant CJD, kuru. Br Med Bull 66: 255-65, 2003.
127. Gambetti P, Kong Q, Zou W, Parchi P and Chen SG: Sporadic and familial CJD: classification and characterisation. Br Med Bull 66: 213-39, 2003.
128. Lee JE, Kwon J and Baek MC: A combination method of chemical with enzyme reactions for identification of membrane proteins. Biochim Biophys Acta 1814: 397-404, 2011.
129. Kaiser R and Metzka L: Enhancement of cyanogen bromide cleavage yields for methionyl-serine and methionyl-threonine peptide bonds. Anal Biochem 266: 1-8, 1999.
130. Nika H, Hawke DH and Angeletti RH: C-terminal protein characterization by mass spectrometry: isolation of C-terminal fragments from cyanogen bromide-cleaved protein. J Biomol Tech 25: 1-18, 2014.
131. Quach TT, Li N, Richards DP, Zheng J, Keller BO and Li L: Development and applications of in-gel CNBr/tryptic digestion combined with mass spectrometry for the analysis of membrane proteins. J Proteome Res 2: 543-52, 2003.
132. Goodlett DR, Armstrong FB, Creech RJ and van Breemen RB: Formylated peptides from cyanogen bromide digests identified by fast atom bombardment mass spectrometry. Anal Biochem 186: 116-20, 1990.
133. Murphy CM and Fenselau C: Recognition of the Carboxy-Terminal Peptide in Cyanogen Bromide Digests of Proteins. Analytical Chemistry 67: 1644-1645, 1995.
134. Samyn B, Sergeant K, Castanheira P, Faro C and Van Beeumen J: A new method for C-terminal sequence analysis in the proteomic era. Nat Methods 2: 193-200, 2005.
135. Vistoli G, De Maddis D, Cipak A, Zarkovic N, Carini M and Aldini G: Advanced glycoxidation and lipoxidation end products (AGEs and ALEs): an overview of their mechanisms of formation. Free Radic Res 47 Suppl 1: 3-27, 2013.
136. Kunkel HG and Wallenius G: New hemoglobin in normal adult blood. Science 122: 288, 1955.
137. Sato A: [Indicators of glycemic control --hemoglobin A1c (HbA1c), glycated albumin (GA), and 1,5-anhydroglucitol (1,5-AG)]. Rinsho Byori 62: 45-52, 2014.
138. Kurien BT, Hensley K, Bachmann M and Scofield RH: Oxidatively modified autoantigens in autoimmune diseases. Free Radic Biol Med 41: 549-56, 2006.
139. Priego-Capote F, Ramirez-Boo M, Hoogland C, Scherl A, Mueller M, Lisacek F and Sanchez JC: Human hemolysate glycated proteome. Anal Chem 83: 5673-80, 2011.
140. Rabbani N and Thornalley PJ: Methylglyoxal, glyoxalase 1 and the dicarbonyl proteome. Amino Acids 42: 1133-42, 2012.
141. Ramirez-Boo M, Priego-Capote F, Hainard A, Gluck F, Burkhard P and Sanchez JC: Characterization of the glycated human cerebrospinal fluid proteome. J Proteomics 75: 4766-82, 2012.

142. Loziuk PL, Wang J, Li Q, Sederoff RR, Chiang VL and Muddiman DC: Understanding the role of proteolytic digestion on discovery and targeted proteomic measurements using liquid chromatography tandem mass spectrometry and design of experiments. J Proteome Res 12: 5820-9, 2013.
143. Arena S, Salzano AM, Renzone G, D'Ambrosio C and Scaloni A: Non-enzymatic glycation and glycoxidation protein products in foods and diseases: an interconnected, complex scenario fully open to innovative proteomic studies. Mass Spectrom Rev 33: 49-77, 2014.
144. Priego-Capote F, Ramirez-Boo M, Finamore F, Gluck F and Sanchez JC: Quantitative Analysis of Glycated Proteins. J Proteome Res: 2014.
145. Brady LJ, Martinez T and Balland A: Characterization of nonenzymatic glycation on a monoclonal antibody. Anal Chem 79: 9403-13, 2007.
146. Miller AK, Hambly DM, Kerwin BA, Treuheit MJ and Gadgil HS: Characterization of site-specific glycation during process development of a human therapeutic monoclonal antibody. J Pharm Sci 100: 2543-50, 2011.
147. Yuk IH, Zhang B, Yang Y, Dutina G, Leach KD, Vijayasankaran N, Shen AY, Andersen DC, Snedecor BR and Joly JC: Controlling glycation of recombinant antibody in fed-batch cell cultures. Biotechnol Bioeng 108: 2600-10, 2011.
148. Gadgil HS, Bondarenko PV, Pipes G, Rehder D, McAuley A, Perico N, Dillon T, Ricci M and Treuheit M: The LC/MS analysis of glycation of IgG molecules in sucrose containing formulations. J Pharm Sci 96: 2607-21, 2007.
149. Quan C, Alcala E, Petkovska I, Matthews D, Canova-Davis E, Taticek R and Ma S: A study in glycation of a therapeutic recombinant humanized monoclonal antibody: where it is, how it got there, and how it affects charge-based behavior. Anal Biochem 373: 179-91, 2008.
150. Zhang B, Yang Y, Yuk I, Pai R, McKay P, Eigenbrot C, Dennis M, Katta V and Francissen KC: Unveiling a glycation hot spot in a recombinant humanized monoclonal antibody. Anal Chem 80: 2379-90, 2008.
151. Ono Y, Aoki S, Ohnishi K, Yasuda T, Katsumi K and Tsukada Y: Increased serum levels of advanced glycation end-products and diabetic complications. Diabetes Research and Clinical Practice 41: 131-137, 1998.
152. Huang QF, Sheng CS, Liu M, Li FH, Li Y and Wang JG: Arterial stiffness and wave reflections in relation to plasma advanced glycation end products in a Chinese population. Am J Hypertens 26: 754-61, 2013.
153. Li H and Liu Z: Recent advances in monolithic column-based boronate-affinity chromatography. TrAC Trends in Analytical Chemistry 37: 148-161, 2012.
154. Bereman MS, Egertson JD and MacCoss MJ: Comparison between procedures using SDS for shotgun proteomic analyses of complex samples. Proteomics 11: 2931-5, 2011.
155. Swaney DL, Wenger CD and Coon JJ: Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 9: 1323-9, 2010.
156. Picotti P, Aebersold R and Domon B: The implications of proteolytic background for shotgun proteomics. Mol Cell Proteomics 6: 1589-98, 2007.
157. Traeger-Synodinos J and Harteveld CL: Advances in technologies for screening and diagnosis of hemoglobinopathies. Biomark Med 8: 119-31, 2014.
158. Kutlar F: Diagnostic approach to hemoglobinopathies. Hemoglobin 31: 243-50, 2007.
159. Daniel YA, Turner C, Haynes RM, Hunt BJ and Dalton RN: Quantification of hemoglobin A2 by tandem mass spectrometry. Clin Chem 53: 1448-54, 2007.
160. Edwards RL, Griffiths P, Bunch J and Cooper HJ: Compound heterozygotes and beta-thalassemia: top-down mass spectrometry for detection of hemoglobinopathies. Proteomics 14: 1232-8, 2014.
161. Acosta-Martin AE, Graca DC, Antinori P, Clerici L, Hartmer R, Meyer M, Hochstrasser D, Samii K, Lescuyer P and Scherl A: Quantitative mass spectrometry analysis of intact hemoglobin A2 by precursor ion isolation and detection. Anal Chem 85: 7971-5, 2013.
162. Pornprasert S, Kasemrad C and Sukunthamala K: Diagnosis of thalassemia on dried blood spot samples by high performance liquid chromatography. Hemoglobin 34: 486-94, 2010.

163. Clarke GM and Higgins TN: Laboratory investigation of hemoglobinopathies and thalassemias: review and update. Clin Chem 46: 1284-90, 2000.
164. Aguilar-Martinez P, Badens C, Bonello-Palot N, Cadet E, Couque N, Ducrocq R, Elion J, Francina A, Joly P, Pissard S and Rochette J: [Flowcharts for the diagnosis and the molecular characterization of hemoglobinopathies]. Ann Biol Clin (Paris) 68: 455-64, 2010.
165. Galanello R, Barella S, Gasperini D, Perseu L, Paglietti E, Sollaino C, Paderi L, Pirroni MG, Maccioni L and Mosca A: Evaluation of an automatic HPLC analyser for thalassemia and haemoglobin variants screening. J Automat Chem 17: 73-6, 1995.
166. Hovasse A: Protéomique: maitrise de l'instrumentation, applications et étude des glycosylations. Thèses de doctorat, Université de Strasbourg. : 2010.
167. Huber CG and Premstaller A: Evaluation of volatile eluents and electrolytes for high-performance liquid chromatography-electrospray ionization mass spectrometry and capillary electrophoresis-electrospray ionization mass spectrometry of proteins. I. Liquid chromatography. J Chromatogr A 849: 161-73, 1999.
168. Garcia MC: The effect of the mobile phase additives on sensitivity in the analysis of peptides and proteins by high-performance liquid chromatography-electrospray mass spectrometry. J Chromatogr B Analyt Technol Biomed Life Sci 825: 111-23, 2005.
169. McCalley DV: Comparison of an organic polymeric column and a silica-based reversed-phase for the analysis of basic peptides by high-performance liquid chromatography. J Chromatogr A 1073: 137-45, 2005.
170. Venable JD, Dong MQ, Wohlschlegel J, Dillin A and Yates JR: Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods 1: 39-45, 2004.
171. Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, Miller SI and Goodlett DR: Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. Anal Chem 81: 6481-8, 2009.
172. Chapman JD, Goodlett DR and Masselon CD: Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom Rev: 2013.
173. Rost HL, Schmitt U, Aebersold R and Malmstrom L: pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. Proteomics 14: 74-7, 2014.
174. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC and MacCoss MJ: Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26: 966-8, 2010.
175. Surinova S, Huttenhain R, Chang CY, Espona L, Vitek O and Aebersold R: Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies. Nat Protoc 8: 1602-19, 2013.
176. Chang CY, Picotti P, Huttenhain R, Heinzelmann-Schwarz V, Jovanovic M, Aebersold R and Vitek O: Protein significance analysis in selected reaction monitoring (SRM) measurements. Mol Cell Proteomics 11: M111 014662, 2012.
177. Clough T, Thaminy S, Ragg S, Aebersold R and Vitek O: Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. BMC Bioinformatics 13 Suppl 16: S6, 2012.
178. Bolstad BM, Irizarry RA, Astrand M and Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-93, 2003.

Annexes

Publications

Journal of proteome research

Technical Note

pubs.acs.org/jpr

An Improved Stable Isotope N-Terminal Labeling Approach with Light/Heavy TMPP To Automate Proteogenomics Data Validation: dN-TOP

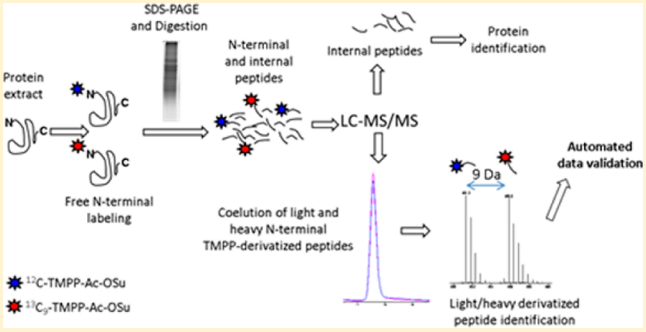
Diego Bertaccini,[†] Sebastian Vaca,[†] Christine Carapito,[†] Florence Arsène-Ploetze,[‡] Alain Van Dorsselaer,[†] and Christine Schaeffer-Reiss[†,*]

[†]Laboratoire de Spectrométrie de Masse BioOrganique, IPHC, Université de Strasbourg, CNRS, UMR7178, Strasbourg, France

[‡]Laboratoire de Génétique Moléculaire, Génomique et Microbiologie, Université de Strasbourg, CNRS UMR7156, Strasbourg, France

Supporting Information

ABSTRACT: *In silico* gene prediction has proven to be prone to errors, especially regarding precise localization of start codons that spread in subsequent biological studies. Therefore, the high throughput characterization of protein N-termini is becoming an emerging challenge in the proteomics and especially proteogenomics fields. The trimethoxyphenyl phosphonium (TMPP) labeling approach (N-TOP) is an efficient N-terminomic approach that allows the characterization of both N-terminal and internal peptides in a single experiment. Due to its permanent positive charge, TMPP labeling strongly affects MS/MS fragmentation resulting in unadapted scoring of TMPP-derivatized peptide spectra by classical search engines. This behavior has led to difficulties in validating TMPP-derivatized peptide identifications with usual score filtering and thus to low/underestimated numbers of identified N-termini. We present herein a new strategy (dN-TOP) that overwhelmed the previous limitation allowing a confident and automated N-terminal peptide validation thanks to a combined labeling with light and heavy TMPP reagents. We show how this double labeling allows increasing the number of validated N-terminal peptides. This strategy represents a considerable improvement to the well-established N-TOP method with an enhanced and accelerated data processing making it now fully compatible with high-throughput proteogenomics studies.



KEYWORDS: *N-terminome analysis, proteogenomics, TMPP derivatization, automated data validation*

INTRODUCTION

In a mass spectrometry-based proteomic discovery experiment, protein identifications are achieved by matching the experimentally obtained spectra with the theoretical mass lists obtained by *in silico* digestion and fragmentation of the protein sequences available in databases. This well-known workflow relies on the assumption that the database is an error free, exhaustive list of all proteins coded by a genome. The reality is far from this assumption since protein databases are mainly obtained by *in silico* translation of the genome sequence. An approach that has proven to be prone to errors and to generate incomplete protein data sets.[1−3] The consequence of this can be dramatic as it can affect any biological experiment that has been based on it.

UniProtKB/SwissProt[4] is a curated protein sequence database in which each entry gets thoroughly analyzed and annotated by expert curators ensuring a high standard of annotation and maintaining the quality of the database.[5] When considering the last reported global statistics of UniProtKB/SwissProt, only 14.2% of all entries have evidence at protein level; 70.1% are inferred from homology; 12.7% have evidence at the transcription level; the predicted entries represent 2.7% and the uncertain entries 0.4% (released the 31-Oct-12).

Among the common errors introduced by *in silico* predictions and propagated by ortholog alignments, the incorrect prediction of initiation codon is particularly pervasive as, up to now, any bioinformatics tool is able to properly estimate with high confidence all initiation sites of the mature proteins;[6,7] this is especially the case for prokaryotic genomes with high GC content since they are characterized by many long open reading frames that are not genic.[8] Based on this evidence, the necessity to collect experimental data to assess and refine the quality of the genome annotation becomes obvious and the development of proteogenomic approaches urgent. In this context, proteomics data are unique resources and can improve many of the problematic areas of genome annotation, like the start site assignment.

To maximize the number of identified N-terminal sequences, the classical high-throughput proteomic workflow has been

Received: April 3, 2013
Published: May 6, 2013

dx.doi.org/10.1021/pr4002993 | *J. Proteome Res.* 2013, 12, 3063−3070

implemented with many complementary approaches able to specifically target the protein N-termini, based on chemical derivatization of the N-terminal function.[9]

The TMPP labeling approach (N-TOP approach) is an efficient N-terminomic approach, that allows the characterization of both N-terminal and internal peptides in a single experiment and has been applied very successfully to various proteomes such as *Mycobacterium smegmatis* and *Sterolibacterium denitrificans* in our laboratory[10,11] and by others.[12] This now well-established N-TOP method is based on a N-terminal protein labeling performed with (*N*-succinimidyloxycarbonylmethyl) tris (2,4,6-trimethoxyphenyl) phosphonium bromide (TMPP-Ac-OSu) on a total biological extract and is fully compatible with all standard detergents, chaotropic agents, and reduction conditions used for protein extraction in proteomics. Two characteristics of this labeling reagent promote the sensitivity of the method: (i) TMPP labeling introduces a permanent positive charge resulting in an enhanced ionization efficiency and thus a better detection of low-abundance proteins; (ii) the hydrophobic TMPP group shifts the retention time of derivatized peptides in reversed phase chromatography toward a less complex part of the chromatogram, therefore, increasing the sensitivity of detection (including the possibility to detect short N-terminal peptides that otherwise would not be retained on the column).

Besides the fact that this approach allows maintaining intact all internal proteolytic peptides, its easy experimental design is the major advantage of this approach: a single chemical derivatization step, performed at the protein level, that can easily be integrated in a classical 1D SDS-PAGE/LC-MSMS proteomics workflow, without requiring any other additional step like immune capture or multidimensional chromatography. Nevertheless, one limitation of the approach resides so far in the validation of labeled peptides, as they present unusual fragmentation patterns. It is well-known that low energy peptide fragmentation (post source decay (PSD) or collision induced dissociation (CID)) is obtained thanks to the delocalization of a proton on the peptidic backbone generating mainly y- and b-type ions.[13]

Alternatively, a peptide labeled with a chemical tag that carries a fixed charge, a permanent positive charge in the case of a TMPP derivatization, behaves in the mass spectrometer in a completely different way; all fragments are generated with a charge remote mechanism that results in a massive production of uncommon ions.[14] The TMPP labeling significantly enhances a- and b-type ions that are usually missing in tryptic peptide MS/MS spectra. Since the search algorithms have been developed and educated for classical fragmentation patterns, a TMPP-derivatized peptide will not be assessed with an optimal score, resulting in a too stringent filtration when operated by the target/decoy approach with 1% FDR and thus in underestimated validation of N-terminal peptides.

To overcome those difficulties, we have developed a new method allowing an easy, reliable and automated TMPP-derivatized peptides' validation based on a stable isotope labeling experiment, a widely applied method in quantitative proteomics.[15−17]

For this purpose, a ^{13}C-labeled analog of the TMPP reagent was designed and a double labeling was performed (1:1 light and heavy TMPP) allowing to identify doublets of identical N-terminal peptide sequences. We designate this labeling strategy as doublet N-terminal oriented proteomics (dN-TOP).

As proof of concept, we applied this method to a cellular lysate of *Herminiimonas arsenicoxydans*. The 3.4 Mbp single chromosome of this arsenite-oxidizing bacterium has already been sequenced and carefully annotated.[18,19] A previously generated proteome map allowed us to characterize 447 proteins among which 365 proteins are in the soluble fraction, representing 13.6% of the total proteome predicted from the genome sequence for this bacterium. For 5 proteins, proteomic data had allowed correcting 5 start codons, even if no N-terminal labeling strategy was applied.[20] To evaluate the specificity and the labeling kinetics of the new isotopically labeled TMPP compared to the light reagent, we present here the comparison of N-TOP to dN-TOP applied to our model organism *H. arsenicoxydans*.

EXPERIMENTAL PROCEDURES

Unless otherwise specified, all chemicals were obtained from Sigma Aldrich (St. Louis, MO).

Growing Conditions and Cell Lysis

H. arsenicoxydans was cultivated in a chemically defined medium (CDM) containing 2.66 mM of As(III) ($NaAsO_2$) in the same conditions as previously described.[20] Late exponential phase cells (100 mL) were disrupted as previously described,[20] and the soluble extract was further analyzed.

Protein Labeling and 1D SDS-PAGE

The protocol used here was carried out according to the original reference paper by Gallien et al. with slight modifications.[10] A batch of heavy labeled (*N*-succinimidyloxycarbonyl-methyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide ($13C_9$ TMPP-Ac-OSu) was synthetized in collaboration with Alsachim. This $13C_9$-TMPP induces a mass increase of 581.21 Da instead of 572.18 Da for light TMPP on labeled peptides. After reduction and alkylation, an equimolar solution of 0.1 M of ^{12}C-TMPP-Ac-OSu and $^{13}C_9$-TMPP-Ac-OSu in CH_3CN:water (2:8; v/v) was added at a molar ratio of 200:1 to 50 μg of *H. arsenicoxydans* protein extract solubilized in labeling buffer (50 mM Tris-HCl, 6 M urea, 2 M thiourea, pH 8.2, 1 mM phenylmethylsulfonyl fluoride, 1 mM EDTA, 5 mM TBP (Bio-Rad Laboratories)). Selective N-terminal TMPP derivatization is achieved by a careful control of reaction pH at 8.2, exploiting the weaker basicity of the N-terminal amine relative to the ε-amino group of the lysine side chain. After a short mix, the reaction was maintained at room temperature for 1 h. Residual derivatizing reagent was quenched by adding a solution of 0.1 M hydroxylamine at room temperature for 1 h, in order to minimize derivatization of tyrosine residues. N-terminal labeled protein extract was finally supplemented with glycerol at a concentration of 10%. Proteins were then separated on a 12% 1D SDS-PAGE (10.1 cm × 7.3 cm) on a mini PROTEAN (Bio-Rad) apparatus at 10 mA for 20 min and 100 mA until the complete migration of the blue front. After electrophoresis, gels were stained with colloidal Coomassie Blue (BioSafe coomassie stain; Bio-Rad) and whole lanes were systematically cut into 28 bands (5 × 2 mm) using a disposable grid-cutter (The Gel-Company, Tübingen, Germany). Bands were cut into three pieces and in-gel digestion using trypsin (Promega, Madison, WI) was performed overnight at 37 °C after in-gel reduction and alkylation using the MassPrep Station (Waters, Milford, MA). Tryptic peptides were extracted using 60% CH_3CN in 0.1% formic acid for 1 h at room temperature. The volume was reduced in a vacuum centrifuge and adjusted to 10 μL using 0.1% formic acid in water before nanoLC-MS/

dx.doi.org/10.1021/pr4002993 | J. Proteome Res. 2013, 12, 3063−3070

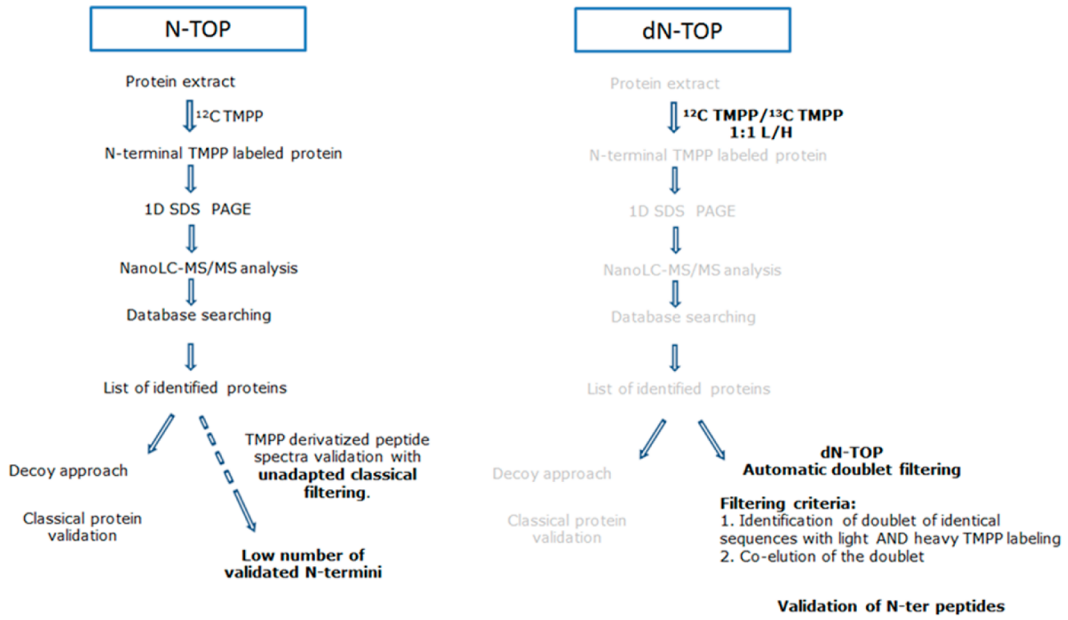


Figure 1. Schematic overview of the dN-TOP approach and its improvement steps (in black) when compared to N-TOP, the steps in common are presented in gray in the dN-TOP workflow.

MS (nanoliquid chromatography coupled to tandem mass spectrometry) analysis.

LC-MS/MS and Data Analysis

NanoLC-MS/MS analyses were performed on a NanoAcquity-LC coupled with a QToF mass spectrometer (maXis 4G, Bruker Daltonics, Bremen, Germany). The UPLC system was equipped with a Symmetry C18 precolumn (0.18 × 20 mm, 5 μm particle size, Waters, Milford, MA) and an ACQUITY UPLC BEH130 C18 separation column (75 μm × 200 mm, 1.7 μm particle size, Waters). The solvent system consisted of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). Of each sample 3 μL was injected. Peptides were trapped during 1 min at 15 μL/min with 99% A and 1% B. Elution was performed at 60 °C at a flow rate of 450 nL/min, using a linear gradient from 6 to 50% B over 50 min. The mass spectrometer was operating in positive mode, with the following settings: source temperature was set to 160 °C while dry gas flow was at 5 L/min. The nanoelectrospray voltage was optimized to −5000 V. External mass calibration of the TOF was achieved before each set of analyses using Tuning Mix (Agilent Technologies, Paolo Alto, CA) in the mass range of 322−2722 m/z. Mass correction was achieved by recalibration of acquired spectra to the applied lock masses (methylstearate ([M + H] + 299.2945 m/z) and hexakis-(2,2,3,3,-tetrafluoropropoxy)phosphazine ([M + H] + 922.0098 m/z)). For tandem MS experiments, the system was operated with automatic switching between MS and MS/MS modes in the range of 100−2500 m/z (MS acquisition time of 0.4 s), MS/MS acquisition time between 0.05 s (intensity >250 000) and 1.25 s (intensity <5000). The 6 most abundant peptides (absolute intensity threshold of 1500) were selected from each MS spectrum for further isolation and CID fragmentation using nitrogen as collision gas. Ions were excluded after acquisition of one MS/MS spectrum and the exclusion was released after 0.25 min.

Peak lists in mascot generic format (.mgf) were generated using Data Analysis (version 4.0; Bruker Daltonics) and merged for each lane using an in-house developed tool available at https://msda.unistra.fr.

Internal Peptide Data Processing

MS and the MS/MS data were analyzed using a local Mascot server (version 2.4.1, Matrix Science, London, England). The search were performed against a *H. arsenicoxydans* database composed of all the original entries (created 2013-02-22 and containing 3400 sequences) downloaded from the public available repository (http://www.genoscope.cns.fr/agc/mage/). The reverse sequences of all entries and common contaminants (keratins, trypsin) were added using our in-house toolbox (Mass Spectrometry Data Analysis, MSDA) freely available after registration at https://msda.unistra.fr. Full trypsin enzyme specificity was fixed, carbamidomethylation of Cysteine (+57 Da) and of oxidation of Methionine (+16 Da) were set as variable modifications and mass tolerances on precursor and fragment ions of 10 ppm and 0.02 Da were used, respectively. Mascot results files (.dat files) were uploaded into the Scaffold software (version 3.6.5; Proteome Software Inc., Portland, USA) for identification validation.

The following filtering criteria based on probability-based scoring of the identified peptides were applied in order to obtain a false discovery rate (FDR) <1% based on the number of decoy hits. Peptides having a Mascot Ion scores higher than Mascot's threshold score of identity (95% confidence level) and absolute Mascot Ion scores >25 were validated.

N-Terminal Labeled Peptide Data Processing

A second Mascot search was performed using semitrypsin enzyme specificity and adding the different TMPP modifications (TMPP N-ter (+572.18 Da), ^{13}C-TMPP N-ter (+581.21 Da), TMPP derivatization of Tyr and Lys (+572.18 Da) and ^{13}C-TMPP derivatization of Tyr and Lys (+581.21 Da)) as variable modifications when compared to the first search.

The Mascot results files (.dat files) were uploaded into the Scaffold software and directly exported, without ion score filtration, in an excel file. An automation tool, Validor freely available at https://msda.unistra.fr in the download software

dx.doi.org/10.1021/pr4002993 | *J. Proteome Res.* 2013, 12, 3063−3070

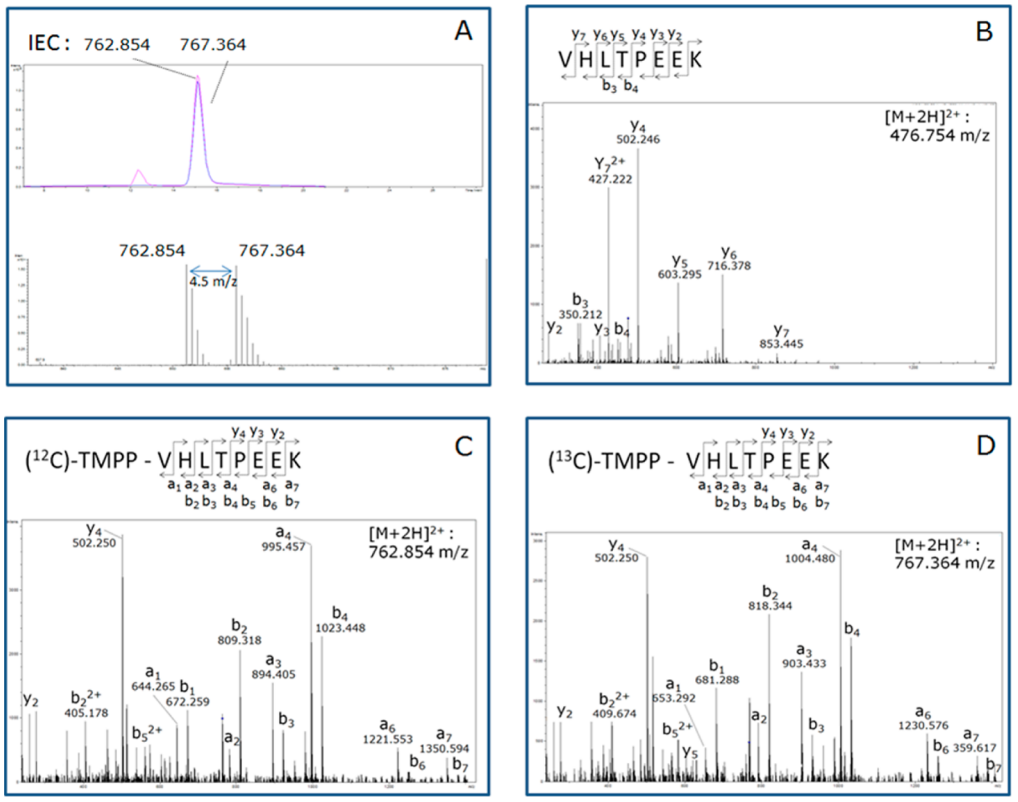


Figure 2. Detailed characterization of the tryptic N-terminal peptide VHLTPEEK of the alpha chain of hemoglobin: (A) Ion extracted chromatogram (EIC) from an LC-MS/MS analysis of the peptide derivatized with ^{13}C-TMPP and ^{12}C-TMPP that clearly shows the perfect coelution of the two peptides. Below the MS spectrum of the two peptides on which the two isotopic profiles are separated by 9 Da is presented. (B) The underivatized peptide produced the expected fragmentation generating mainly y_n- or b_n-type ions. (C and D) The comparison of MS/MS spectra of the peptide derivatized with light TMPP and with heavy TMPP, respectively. The derivatized peptides present similar fragmentation patterns with predominant a_n- and b_n-type ion series when compared to the predominant y_n-type ion in the nonderivatized spectrum (B). As expected, the mass difference of 9 Da affects all a_n-and b_n-type fragments, when comparing the MS/MS spectra of the light and heavy TMPP derivatized peptides.

section, was in-house developed to automate N-terminal peptides validation. In a first step, identical peptide sequences are detected and retained if both the ^{12}C-TMPP-derivatized and the ^{13}C-TMPP-derivatized peptides are identified. In a second step, retention times are extracted for every doublet and a user defined tolerance window is applied to ensure coelution of both heavy and light forms. Validor requires an excel file with the following information present in separate columns: accession number, peptide sequence, retention time, peptide modification (see Supporting Information and Method for details).

RESULTS AND DISCUSSION

A general schematic overview of the N-TOP and dN-TOP strategies is depicted in Figure 1. Both experimental workflows are comparable except for the use of the ^{12}C-TMPP/^{13}C-TMPP mixture for the labeling reaction instead of using only the ^{12}C-TMPP reagent. This allowed us to significantly improve the so far limiting step of N-terminal peptide validation and to significantly increase the number of the validated protein starts, while maintaining the strength of the approach, i.e. preserving intact all internal peptides.

dN-TOP Identification and Validation of Internal Peptides

The workflow starts with the denaturation of proteins by reduction and alkylation of cysteine residues to enhance the accessibility of N-termini for chemical derivatization. After treatment of the protein extract with a 1:1 mixture of light and heavy TMPP, a 1D gel separation followed. 1D SDS-PAGE step was shown to be ideal to remove TMPP excess and had the additional advantage not only of being compatible with strong detergents but also reducing the complexity of protein extracts prior to LC-MS/MS analysis. After systematic band cutting, tryptic in gel digestion and nanoLC-MS/MS of each band, all files are merged to generate a global peak list for each lane. This global peak list is then submitted to database searches using Mascot (with full enzyme specificity). Proteins are identified thanks to internal peptides and validated in a usual way (using Scaffold software and score filtering for significant identification at a false discovery rate of 1% with a target/decoy database), since internal peptides are not chemically affected by the TMPP labeling.[10]

dN-TOP Identification and Validation of N-Terminal Peptides

A second database search is then performed with semitrypsin specificity and TMPP modifications to identify N-terminal peptides. Semitrypsin specificity, which allows a one peptidic termini to be aspecific, is required for this search in order to identify also unpredicted protein starts (downstream of the predicted protein start) that would be missed with full trypsin searches. Indeed, a full trypsin search only allows identifying the N-terminal peptides of the proteins as predicted and present in the database. During this search, both light and heavy forms of

dx.doi.org/10.1021/pr4002993 | J. Proteome Res. 2013, 12, 3063−3070

TMPP are set as variable modifications. The $^{13}C_9$-TMPP modification has been added to the UNIMOD database (http://www.unimod.org). The list of identified peptides contains a series of peptides modified on their N-termini by light or by heavy TMPP. As suspected, those identifications have assigned scores non representative of the spectral quality due to their unusual fragmentation patterns. Figure 2b–d illustrates, in the case of a mixture of model proteins (alpha and beta chains of hemoglobin), the unusual fragmentation pattern of TMPP-derivatized peptides compared to the nonmodified peptides. For TMPP-derivatized peptides, a- and b-type fragmentation ions are dominant due to the permanent charge introduced by the TMPP reagent while the nonmodified peptide produces the expected tryptic peptide fragmentation (mostly y-type ions and a few b-type ions).

Perfect coelution of N-terminal tryptic peptides derivatized by light or heavy TMPP is verified on all doublet peptides. Figure 2a shows that intensities between light and heavy labeled peptides are in close agreement with the initial 1:1 ratio of ^{12}C- and ^{13}C-TMPP reagent. The MS spectrum shows that the difference of mass-to-charge values (m/z) of the doublet monoisotopic peaks is 4.5 for this doubly charged peptide, corresponding to a mass increase of 9 Da which is adapted to separate the light and heavy peptides' isotope envelopes.

As described in the Experimental Procedures section, Validor allows an automatic validation of the N-terminal peptides based on 2 criteria: the identification of both the ^{12}C-TMPP-modified and the ^{13}C-TMPP-modified peptide sequence and a perfect coelution of both forms.

Application of the Workflow to *H. arsenicoxydans* Proteome

H. arsenicoxydans is a β-proteobacteria which uses organic compounds as an electron donor, oxidizes As(III) and can resist to up to 6 mM As(III) and 200 mM As(V).[18] *H. arsenicoxydans* is the first arsenite-oxidizing bacterium whose genome has been sequenced in 2007 and is rather well annotated.[18,20] However, start site assignment has not yet been validated by experimental proteomics data. Therefore, we have applied our N-TOP and dN-TOP strategies to this organism and we present here a deeper characterization of its proteome, with a special focus on its N-terminome.

Comparison of N-TOP versus dN-TOP

To verify that the doublet dN-TOP strategy allows identifying a maximum of N-terminal peptides, we have first performed two separate experiments using a single TMPP isotopologue. One protein extract of *H. arsenicoxydans* lysate was treated with ^{12}C-TMPP while the other one with the ^{13}C-TMPP, and both derivatized protein mixtures were subjected to the classical N-TOP workflow as described in Figure 1. The two MS/MS data sets were validated using the classical target/decoy approach with a FDR $\leq$1%. These experiments yielded 50 and 74 N-terminal peptides with ^{12}C-TMPP and ^{13}C-TMPP labeling, respectively, when using classical validation criteria (Table 1).

Then, the dN-TOP strategy was applied to the same lysate of *H. arsenicoxydans*.

Except for the labeling with a 1:1 mixture of ^{12}C-TMPP and ^{13}C-TMPP reagents, the same experimental workflows was applied, i.e., derivatization, separation on 1D SDS-PAGE, in gel digestion and LC-MS/MS analysis of the extracted peptides. In total, Validor allowed the automatic validation of 90 N-terminal peptides thanks to the light and heavy doublet identification (Supporting Information Table S2). This experiment illustrates the significantly underestimated validation of N-termini when using the N-TOP strategy and proves the major advantage of the dN-TOP approach as half of the peptides have been discarded in the 2 individual N-TOP experiments (Table 1).

Table 1. Results Obtained with the Classical N-TOP Method Compared to the dN-TOP Approach with Validor

	N-TOP ^{12}C TMPP	N-TOP ^{13}C TMPP	dN-TOP
Number of N-ter validated with FDR < 1%	55	78	n.d
Number of N-ter validated with Validor	n.d	n.d	90
(Not expected N-ter)			(13)
Number of proteins identified with FDR < 1%	566	588	504

Concerning total protein identifications in these three separate experiments, comparable numbers of proteins have been identified (Table 1 and supplemental Table S1). This proves that the doublet labeling does not affect the global identification rate (even if sample complexity is slightly increased by the labeling with the 2 TMPP forms). This is also due to the fact that TMPP labeling shifts N-terminal peptides' elution times toward a less complex part of the chromatogram, out of the eluting area of internal peptides.

The *H. arsenicoxydans* N-Terminome with dN-TOP

In total, 504 unique proteins were identified from internal digestion peptides (Table 1). When combining the lists of unique identified proteins over the three experiments (^{12}C TMPP labeling N-TOP, ^{13}C TMPP labeling N-TOP and dN-TOP, Supporting Information Table S1), the total number of proteins raises to 650, increasing the previously published proteome with 384 additional proteins.[20]

From the same data set, 90 unique N-terminal peptides were identified with Validor among which 77 were correctly predicted by the genome annotation (Supporting Information Table S2). The 13 remaining N-terminal peptides did not match to the predicted starts annotated in the *H. arsenicoxydans* database (Table 2). We carefully analyzed these N-terminal peptides in order to highlight possible annotation errors or proteolytic events.

In the case of Flavoprotein HEAR0503, we have identified an N-terminal derivatized peptide presenting a wrongly annotated start site. As illustrated in Figure 3, the identified N-terminal peptide of protein HEAR0503 showed clearly that the start site was experimentally detected 13 amino acids downstream of the annotated translational start site. We checked further if this new start may be in agreement with alternative start prediction algorithms, and if this start would fit to alignments with other known proteins. This identification provides the experimental evidence of remaining incorrectly predicted start sites even after expert manual annotation.[18]

Besides start site annotation errors, we have identified six TMPP-derivatized peptides corresponding to signal peptide cleavage sites (Table 2), allowing to experimentally validate those cleavage sites as predicted by the SignalP 4 algorithm.[21] Interestingly, in one case, identification of the N-terminal TMPP labeled peptide FDFNDVAK supports the predicted cleavage site of protein Glucan Biosynthesis G HEAR3286, and indirectly the prediction of an alternative start codon (Figure 4A). Indeed, in the case of this periplasmic protein involved in the synthesis of membrane-derived oligosaccharides (MDO), two possible starts were predicted according to two different

dx.doi.org/10.1021/pr4002993 | J. Proteome Res. 2013, 12, 3063–3070

Table 2. List of Identified N-Terminal Peptides That Do Not Match with the Annotated Protein N-Termini

protein accession numbers	peptide sequence	peptide start index	SignalP prediction
sp|HEAR0005	AIPNDNTPQSPSTLSAAYGASSIQILEGLEAVR	3	Between amino acid 27 and 28
sp|HEAR0225	TNSIAR	114	
sp|HEAR0310	ATVLK	28	
sp|HEAR0348	AWEPTKPVEFVVPAGTGGGADQMAR	34	Between amino acid 33 and 34
sp|HEAR0415	DAAYPNK	23	Between amino acid 22 and 23
sp|HEAR0503	SQNFPDLPNIDPALFTTPTR	13	
sp|HEAR1107	ADITGAGATFPYPIFSK	26	Between amino acid 25 and 26
sp|HEAR1195	APSAAK	30	Between amino acid 29 and 30
sp|HEAR1337	TTPAYK	28	
sp|HEAR2797	TMLGFMATDAK	196	
sp|HEAR3286	FDFNDVAK	31	Between amino acid 30 and 31
sp|HEAR3424	TTTFR	95	
sp|HEAR3468	MLLTR	97	

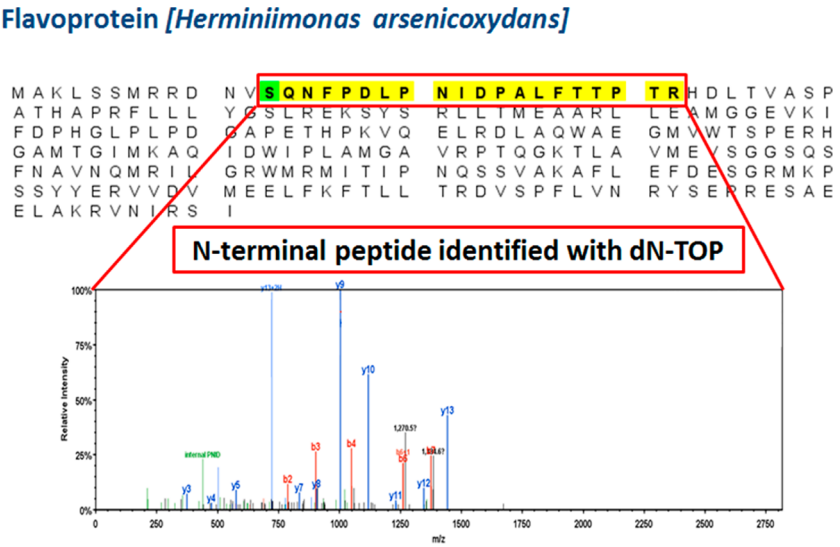


Figure 3. Example of a *H. arsenicoxydans* protein, Flavoprotein (HEAR0503), with an experimental start codon correction (13 amino acids after to the currently annotated translation start site).

algorithms (AMIGene and Yuko-Makita), with a good prediction obtained only for the second algorithm. Thus, identification of this TMPP-derivatized peptide allowed to experimentally validate the signal peptide prediction after protein reannotation.

An interesting case of post-translational proteolytic cleavage is illustrated in figure 4B. An N-terminal derivatized peptide was identified starting at position 196 on HEAR 2797, Bifunctional glutamate *N*-acetyltransferase/amino-acid acetyltransferase. A sequence alignment with *C. crenatum* Arginine biosynthesis bifunctional protein showed a high degree of similarity. In this organism, the protein undergoes a proteolytic autolysis between the amino acids 182 and 183, corresponding to residues 195 and 196 in *H. arsenicoxydans*, which generates two chains, the α and the β chains.[22] Figure 4 thus shows that the dN-TOP provides a useful tool to identify proteolytic events such as cleavage sites and signal peptide processing.

Five additional TMPP-derivatized peptides were identified with N-termini that could correspond to proteolytic cleavage sites. However, no proteolytic fragments for these proteins identified *in vivo* are yet reported in the literature. Therefore, no biological interpretation can be given to those proteolytic events without additional experiments.

CONCLUSION AND OUTLOOK

In conclusion, our proof-of-concept experiment on *H. arsenicoxydans* allowed confirming predicted N-termini, correcting wrong start site predictions, and identifying proteolytic events, such as signal peptide cleavages and a proteolytic cleavage sites. We have also demonstrated that dN-TOP presents a significant improvement over the N-TOP approach, for which the labeled peptide validation step was limiting. This improvement makes this methodology compatible with high-throughput and large-scale proteomics studies. This opens also the door to the possibility of performing large-scale experimental validations of predicted genome annotations and dN-TOP reveals to be a powerful proteogenomics tool.

Additionally, the availability of a $^{13}C_9$ TMPP reagent offers the possibility to perform quantitative N-terminomics. It will indeed be possible to compare the N-terminome of two different samples by labeling them with light or heavy TMPP, respectively. The identification and validation method presented here will be useful for a fast detection of the mass spectrum of interest for determining the ratio of the two molecular ions.

dx.doi.org/10.1021/pr4002993 | *J. Proteome Res.* 2013, 12, 3063−3070

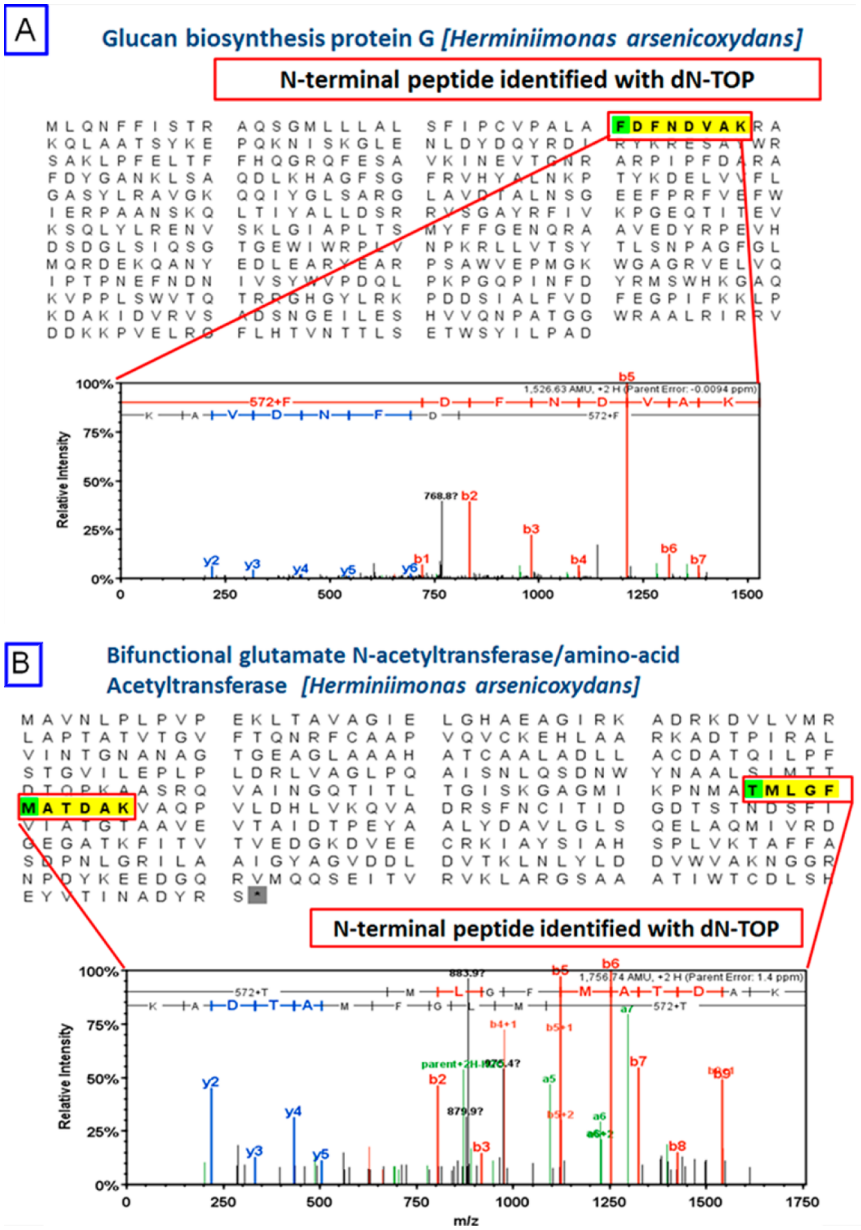


Figure 4. (A) Example of an N-terminal processing validated with the dN-TOP approach: the protein Glucan biosynthesis protein G is synthetized with a signal peptide and processed in the ER. The experimental evidence of a mature form (after signal peptide cleavage) with the N-start at position 31 is supported by the *in silico* signal peptide prediction performed with the software SignalP 4.0[21] that predicts a cleavage between amino acid 30 and 31. (B) Example of an endoproteolytic cleavage on protein HEAR 2797. The position 196 has been detected as alternative N-terminal in Bifunctional glutamate N-acetyltransferase/amino-acid acetyltransferase. The closest sequence identity with a *Corynebacterium crenatum* protein which exhibits a proteolytic autolysis confirms a proteolytic cleavage site in this region of the sequence.

■ ASSOCIATED CONTENT

Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: christine.schaeffer@unistra.fr. Phone: (+33) 3.68.85.27.79. Fax: (+33) 3.68.85.27.81.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the CNRS, the "Agence National de la Recherche" (ANR), the Proteomic French Infrastructure (ProFI; ANR-10-INSB-08-03) and Fondation Pour La Recherche Médicale FRM.

■ REFERENCES

(1) Schrimpe-Rutledge, A. C.; Jones, M. B.; Chauhan, S.; Purvine, S. O.; Sanford, J. A.; Monroe, M. E.; Brewer, H. M.; Payne, S. H.; Ansong, C.; Frank, B. C.; Smith, R. D.; Peterson, S. N.; Motin, V. L.; Adkins, J. N. Comparative omics-driven genome annotation refinement: application across Yersiniae. *PLoS One* **2012**, *7* (3), e33903.

(2) Delalande, F.; Carapito, C.; Brizard, J. P.; Brugidou, C.; Van Dorsselaer, A. Multigenic families and proteomics: extended protein

characterization as a tool for paralog gene identification. *Proteomics* **2005**, *5* (2), 450−60.
(3) Blakeley, P.; Overton, I. M.; Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **2012**, *11* (11), 5221−34.
(4) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **1999**, *27* (1), 49−54.
(5) Magrane, M.; Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, *2011*, bar009.
(6) Aivaliotis, M.; Gevaert, K.; Falb, M.; Tebbe, A.; Konstantinidis, K.; Bisle, B.; Klein, C.; Martens, L.; Staes, A.; Timmerman, E.; Van Damme, J.; Siedler, F.; Pfeiffer, F.; Vandekerckhove, J.; Oesterhelt, D. Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis. *J. Proteome Res.* **2007**, *6* (6), 2195−204.
(7) Bonissone, S.; Gupta, N.; Romine, M.; Bradshaw, R. A.; Pevzner, P. A. N-terminal protein processing: a comparative proteogenomic analysis. *Mol. Cell. Proteomics* **2013**, *12* (1), 14−28.
(8) Venter, E.; Smith, R. D.; Payne, S. H. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS One* **2011**, *6* (11), e27587.
(9) Armengaud, J. A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr. Opin. Microbiol.* **2009**, *12* (3), 292−300.
(10) Gallien, S.; Perrodou, E.; Carapito, C.; Deshayes, C.; Reyrat, J. M.; Van Dorsselaer, A.; Poch, O.; Schaeffer, C.; Lecompte, O. Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res.* **2009**, *19* (1), 128−35.
(11) Chiang, Y. R.; Ismail, W.; Gallien, S.; Heintz, D.; Van Dorsselaer, A.; Fuchs, G. Cholest-4-en-3-one-delta 1-dehydrogenase, a flavoprotein catalyzing the second step in anoxic cholesterol metabolism. *Appl. Environ. Microbiol.* **2008**, *74* (1), 107−13.
(12) Baudet, M.; Ortet, P.; Gaillard, J. C.; Fernandez, B.; Guerin, P.; Enjalbal, C.; Subra, G.; de Groot, A.; Barakat, M.; Dedieu, A.; Armengaud, J. Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Mol. Cell. Proteomics* **2010**, *9* (2), 415−26.
(13) Gucinski, A. C.; Dodds, E. D.; Li, W.; Wysocki, V. H. Understanding and exploiting Peptide fragment ion intensities using experimental and informatic approaches. *Methods Mol. Biol.* **2010**, *604*, 73−94.
(14) He, Y.; Parthasarathi, R.; Raghavachari, K.; Reilly, J. P. Photodissociation of charge tagged peptides. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (7), 1182−90.
(15) Altelaar, A. F.; Frese, C. K.; Preisinger, C.; Hennrich, M. L.; Schram, A. W.; Timmers, H. T.; Heck, A. J.; Mohammed, S. Mohammed, S., Benchmarking stable isotope labeling based quantitative proteomics. *J. Proteomics* **2012**, DOI: 10.1016/j.jprot.2012.10.009.
(16) Kline, K. G.; Sussman, M. R. Protein quantitation using isotope-assisted mass spectrometry. *Annu. Rev. Biophys.* **2010**, *39*, 291−308.
(17) Evans, C.; Noirel, J.; Ow, S. Y.; Salim, M.; Pereira-Medrano, A. G.; Couto, N.; Pandhal, J.; Smith, D.; Pham, T. K.; Karunakaran, E.; Zou, X.; Biggs, C. A.; Wright, P. C. An insight into iTRAQ: where do we stand now? *Anal. Bioanal. Chem.* **2012**, *404* (4), 1011−27.
(18) Muller, D.; Medigue, C.; Koechler, S.; Barbe, V.; Barakat, M.; Talla, E.; Bonnefoy, V.; Krin, E.; Arsene-Ploetze, F.; Carapito, C.; Chandler, M.; Cournoyer, B.; Cruveiller, S.; Dossat, C.; Duval, S.; Heymann, M.; Leize, E.; Lieutaud, A.; Lievremont, D.; Makita, Y.; Mangenot, S.; Nitschke, W.; Ortet, P.; Perdrial, N.; Schoepp, B.; Siguier, P.; Simeonova, D. D.; Rouy, Z.; Segurens, B.; Turlin, E.; Vallenet, D.; Van Dorsselaer, A.; Weiss, S.; Weissenbach, J.; Lett, M. C.; Danchin, A.; Bertin, P. N. A tale of two oxidation states: bacterial colonization of arsenic-rich environments. *PLoS Genet.* **2007**, *3* (4), e53.
(19) Muller, D.; Simeonova, D. D.; Riegel, P.; Mangenot, S.; Koechler, S.; Lievremont, D.; Bertin, P. N.; Lett, M. C. Herminiimonas arsenicoxydans sp. nov., a metalloresistant bacterium. *Int. J. Syst. Evol. Microbiol.* **2006**, *56* (Pt 8), 1765−9.
(20) Weiss, S.; Carapito, C.; Cleiss, J.; Koechler, S.; Turlin, E.; Coppee, J. Y.; Heymann, M.; Kugler, V.; Stauffert, M.; Cruveiller, S.; Medigue, C.; Van Dorsselaer, A.; Bertin, P. N.; Arsene-Ploetze, F. Enhanced structural and functional genome elucidation of the arsenite-oxidizing strain Herminiimonas arsenicoxydans by proteomics data. *Biochimie* **2009**, *91* (2), 192−203.
(21) Petersen, T. N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8* (10), 785−6.
(22) Dou, W.; Xu, M.; Cai, D.; Zhang, X.; Rao, Z.; Xu, Z. Improvement of L-arginine production by overexpression of a bifunctional ornithine acetyltransferase in Corynebacterium crenatum. *Appl. Biochem. Biotechnol.* **2011**, *165* (3−4), 845−55.

Biol. Cell (2014) **106,** 126–138 DOI: 10.1111/boc.201400003

Research article

H2B ubiquitylation modulates spliceosome assembly and function in budding yeast

Lucas Hérissant*[1], Erica A. Moehle†[1], Diego Bertaccini‡, Alain Van Dorsselaer‡, Christine Schaeffer-Reiss‡, Christine Guthrie† and Catherine Dargemont*[2]

*Pathologie Cellulaire, University Paris Diderot, Sorbonne Paris Cité, INSERM U944, CNRS UMR7212, Equipe labellisée Ligue contre le cancer, Hôpital Saint Louis, Paris Cedex 10, France, †Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA, and ‡Institut Plurisdisciplinaire Hubert Curien, University of Strasbourg, CNRS, Strasbourg, France

Background information. Commitment to splicing occurs co-transcriptionally, but a major unanswered question is the extent to which various modifications of chromatin, the template for transcription *in vivo*, contribute to the regulation of splicing.

Results. Here, we perform genome-wide analyses showing that inhibition of specific marks – H2B ubiquitylation, H3K4 methylation and H3K36 methylation – perturbs splicing in budding yeast, with each modification exerting gene-specific effects. Furthermore, semi-quantitative mass spectrometry on purified nuclear mRNPs and chromatin immunoprecipitation analysis on intron-containing genes indicated that H2B ubiquitylation, but not Set1-, Set2- or Dot1-dependent H3 methylation, stimulates recruitment of the early splicing factors, namely U1 and U2 snRNPs, onto nascent RNAs.

Conclusions. These results suggest that histone modifications impact splicing of distinct subsets of genes using distinct pathways.

Additional supporting information may be found in the online version of this article at the publisher's web-site

Introduction

Transcriptional control of gene expression has long been thought to require the coordinated modification of histones (Suganuma and Workman, 2011) but recent evidence suggests an additional role for these modifications in controlling the eventual fate of an mRNA after it is transcribed (Carrillo Oesterreich et al., 2011; Hnilicova and Stanek, 2011; Luco et al., 2011; Nino et al., 2013). A number of correlative studies have shown that nucleosomes at DNA encoding exons and introns bear distinct covalent modifications in metazoa (reviewed in Carrillo Oesterreich et al., 2011) and yeast (Shieh et al., 2011). These observations raise the possibility that epigenetic marks may have a widespread role in providing direct regulatory input into co-transcriptional splicing decisions.

In *Saccharomyces cerevisiae*, splicing occurs co-transcriptionally for the vast majority of intron-containing genes (Carrillo Oesterreich et al., 2010), and consistently, splicing factors are recruited to nascent transcripts (Kotovic et al., 2003; Gornemann et al., 2005; Lacadie and Rosbash, 2005; Moore et al., 2006; Tardiff et al., 2006; Aitken et al., 2011). Interestingly, the histone acetyltransferase catalytic subunit of the SAGA complex, Gcn5, has been shown to control the co-transcriptional recruitment of the U2 small nuclear ribonucleoprotein (snRNP)

[1]These authors contributed equally to this work.
[2]To whom correspondence should be addressed (email dargemont@gmail.com)
Key words: H2B ubiquitylation, Histone marks, Pre-mRNA Splicing, snRNP.
Abbreviations used: CBC, cap-binding complex; ChIP, chromatin immunoprecipitation; CTD, C-terminal domain; H3K4me, lysine 4 methylation of histone H3; mRNP, messenger ribonucleoprotein; RNAPII, RNA polymerase II; RPG, ribosomal protein gene; snRNP, small nuclear ribonucleoprotein; Ub-H2B, histone H2B mono-ubiquitylation; WT, wild-type

(Gunderson and Johnson, 2009). However, the extent to which additional histone modifications of the chromatin landscape regulate the co-transcriptional recruitment of the spliceosome is still unclear.

Transcription-associated histone H2B mono-ubiquitylation (Ub-H2B) and the downstream histone H3 methylation events have established roles in transcription activation and nucleosome dynamics (Robzyk et al., 2000; Sun and Allis, 2002; Wood et al., 2003; Vitaliano-Prunier et al., 2008). In addition, we recently showed that Ub-H2B influences export of messenger ribonucleoproteins (mRNPs) by promoting the recruitment of the nuclear export machinery to nascent transcripts (Babour et al., 2012; Vitaliano-Prunier et al., 2012). Furthermore, we recently reported genetic and functional interactions between the Ub-H2B machinery and the SR-like spliceosome-associated factor Npl3 (Moehle et al., 2012), suggesting that the role of Ub-H2B in gene expression is not limited to directing transcription itself. This result prompted us to determine the contribution of transcription-dependent chromatin marks, and in particular Ub-H2B, on spliceosome assembly and function on nascent transcripts.

Results

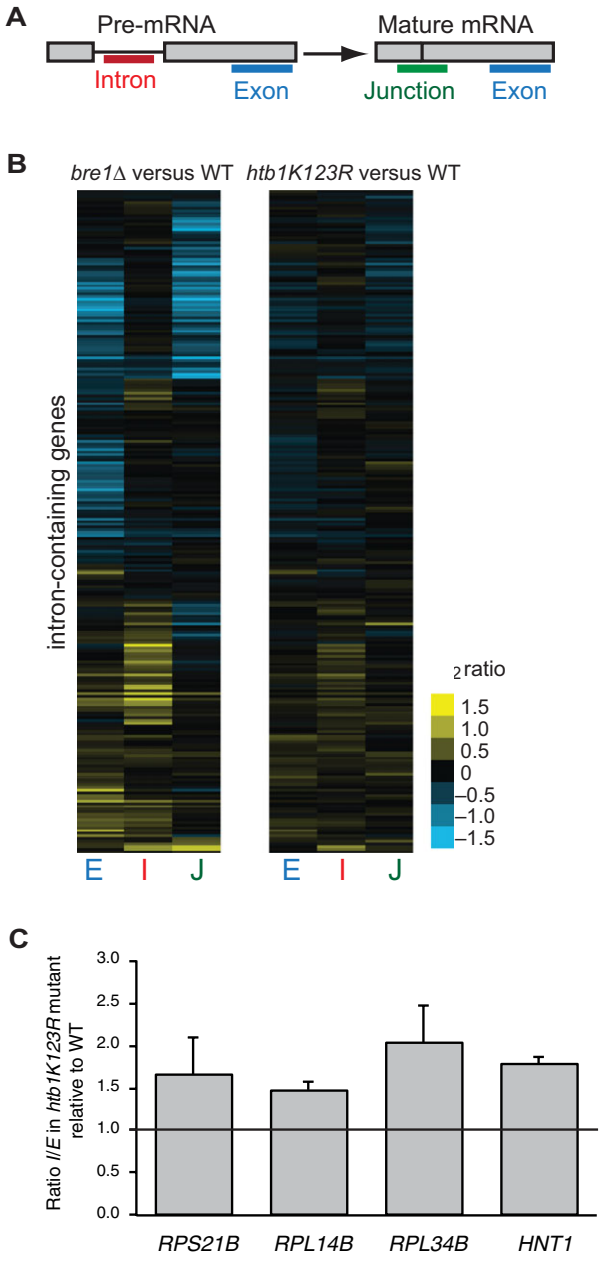
Defects in Ub-H2B, H3K4me, H3K36me cause introns to accumulate for distinct subsets of transcripts

Transcripts in *S. cerevisiae* do not generally undergo alternative splicing, but the constitutive splicing reaction is sensitive to a number of environmental perturbations (Pleiss et al., 2007; Munding et al., 2010; Bergkessel et al., 2011). While relatively few genes are spliced, intron-containing genes account for nearly one third of total cellular transcription (Ares et al., 1999), so it is critical for yeast to appropriately control the efficiency of this step in gene expression. We recently reported that in a genetic background sensitised by loss of Npl3, a protein known to promote splicing of a subset of genes, a short 37°C temperature shift revealed a modest dependence of pre-mRNA splicing on Ub-H2B (Moehle et al., 2012).

Here, we further explored the potential connection between chromatin modification and splicing by capitalising on the observation that nuclear export factor assembly onto nascent mRNPs is very tightly regulated by Ub-H2B during a 3-h shift to 39°C, an experimental condition that challenges mRNA biogenesis without affecting genome-wide expression (Babour et al., 2012; Vitaliano-Prunier et al., 2012). Indeed, using splicing-sensitive microarrays (Figure 1A), we see that at 39°C, abrogating Ub-H2B by deleting the H2B E3 ligase, *BRE1*, or mutating the targetted residue in H2B (*htb1K123R*) led to increases in the levels of intron for many genes, consistent with a defect in the splicing of those transcripts (Figures 1B and 2C and Supplementary Table S3). To more easily compare these datasets, we calculated intron/exon ratios, an established approach to normalise for differences in transcription (Clark et al., 2002). We observed that, importantly, genes affected by H2B mutation extensively overlapped with genes affected by *BRE1* deletion (Figure 1B). While the ribosomal protein genes (RPGs) are a category of spliced genes often regulated together (Pleiss et al., 2007; Bergkessel et al., 2011), Ub-H2B-dependent effects on splicing were not enriched for RPG transcripts. Validation of the microarray data by using RT-qPCR to measure relative intron and exon abundance of several transcripts confirmed Ub-H2B-mediated changes with respect to a wild-type (WT) strain (Figure 1C). Taken together, our data show that loss of Ub-H2B has clear gene-specific effects on intron accumulation and, thus, prompt the conclusion that Bre1-dependent Ub-H2B is important for splicing at 39°C.

Ub-H2B is strictly required for other histone marks such as the trimethylation of histone H3 on both lysine 4 by the Set1-containing COMPASS complex (Sun and Allis, 2002) and lysine 79 by Dot1 (Briggs et al., 2002; Ng et al., 2002), and facilitates the Set2-mediated methylation of H3K36 on some intron-containing genes (Shieh et al., 2011) (not shown). Surprisingly, we found that deletion of *SET2* also causes accumulation of intron and an increase in the intron/exon ratio for many transcripts (Figures 2A and 2C), but the observation that 83% of Set2-dependent genes are not also dependent on Ub-H2B (Figure 2D) suggests Set2 is working separately from the Ub-H2B pathway. Microarray results from a strain lacking *SET1* were consistent with a mild splicing defect as gauged both by intron accumulation and intron/exon ratios (Figures 2B and 2C). However, only a small number of genes (13) overlap with those affected in *htb1K123R* (Figure 2D), indicating that the effects of Ub-H2B on intron/exon ratios

Figure 1 | Defects in Ub-H2B promote splicing defects
(**A**) Schematic of probes contained on the microarray for each intron-containing gene. (**B**) Heat maps of $\log_2$ ratios of each gene feature in *bre1*Δ or *htb1K123R* compared with isogenic WT strains after a 3-h shift to 39°C. Gene order is the same for both arrays. (**C**) RT-qPCR measurements of unspliced mRNAs using single-locus RT-qPCR. Percent unspliced RNA is represented as fold change compared with WT.



were not strictly mediated by H3K4 methylation. We observe a comparatively larger effect from *BRE1* deletion than mutation of the H2B target residue, which is consistent with an additional Bre1 target or function that also promotes splicing. Bre1 targets many of the same genes as COMPASS component Set1, but whether ubiquitylation of the Swd2 component of the COMPASS complex by Bre1 might be involved in this process remains to be determined (Vitaliano-Prunier et al., 2008).

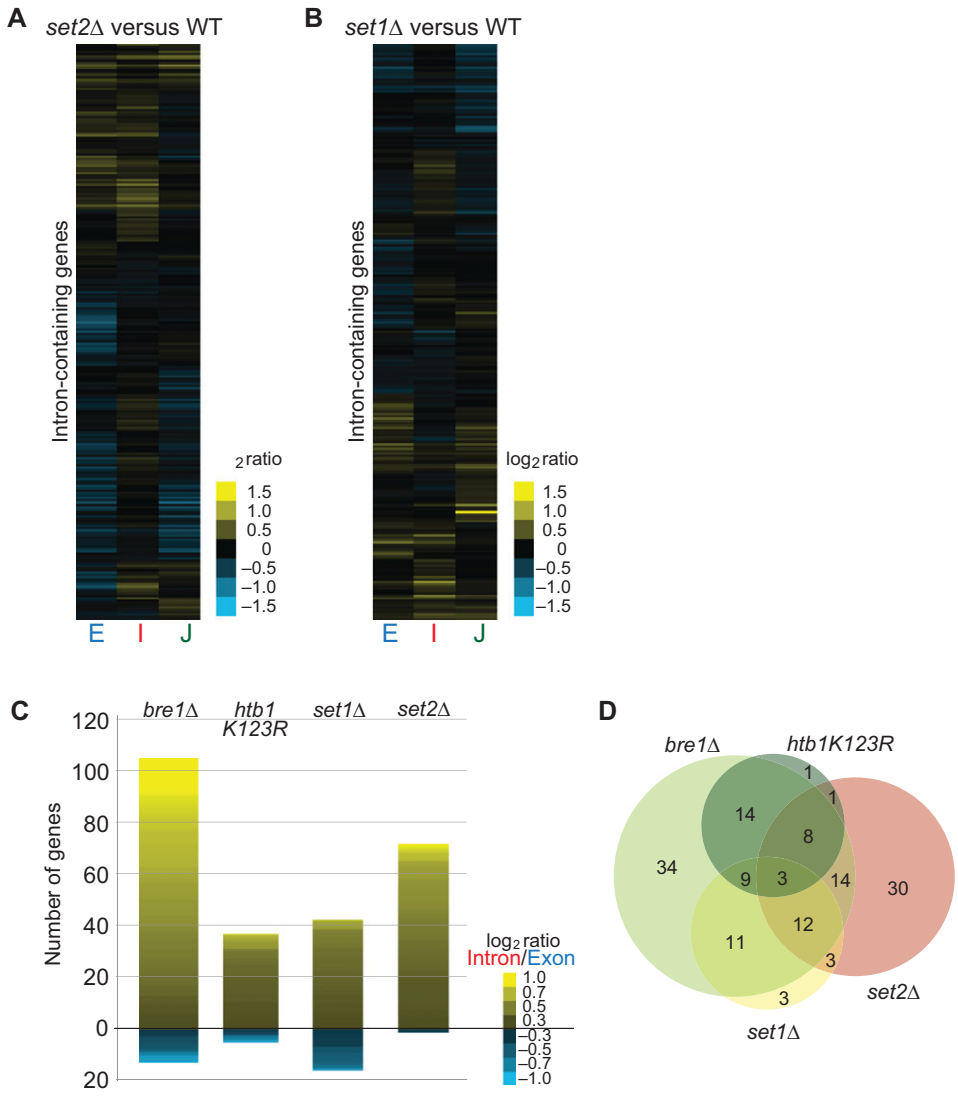
Importantly, no global changes of gene expression were observed upon inhibition of Ub-H2B or downstream H3 methylations either at 30°C (Lenstra et al., 2011; Margaritis et al., 2012) or 39°C (Vitaliano-Prunier et al., 2012). Only a minority of genes exhibited altered expression in the different mutant strains, none of which encoded components of the splicing machinery. This argues against direct transcriptional control of the splicing machinery expression by Ub-H2B.

Together, these results suggest that splicing efficiencies – inferred by changes in pre-mRNA and total mRNA levels – are dependent on contributions from multiple transcription-coupled histone marks, with the relative contribution being different from one intron-containing gene to another. We reasoned that the decrease in splicing efficiency we observed was unlikely to be caused by a wholesale block in spliceosome function, but rather could relate to a delay in the onset of the splicing reaction. We therefore sought to determine whether Ub-H2B might influence the ability of splicing factors to associate with transcripts.

Preventing ubiquitylation of H2B alters the recruitment of early splicing factors to Cap-binding complex-associated mRNPs

Since Ub-H2B-mediated splicing is likely mechanistically separated from that driven by downstream histone methylations, what step in splicing is impacted by loss of Ub-H2B? To address this question, nuclear mRNPs were purified from temperature-shifted WT and *htb1K123R* cells using a genomically TAP-tagged Cbc2 of the nuclear cap-binding complex (CBC), and their proteomes were analysed by tandem mass spectrometry (Oeffinger et al., 2007). Because the CBC is associated with nuclear mRNPs from early synthesis to nuclear exit, the composition of purified mRNPs reflects the sum of all biogenesis

Figure 2 | Defects in Ub-H2B, H3K4me, H3K36me cause introns to accumulate for distinct subsets of transcripts
(**A and B**) Heat maps of $\log_2$ ratios of each gene feature in *set2*Δ or *set1*Δ strains compared with isogenic WT strains after a 3-h shift to 39°C. Gene order is different for each sub-panel. (**C**) Histogram of number of genes exhibiting $\log_2$ (intron/exon) ratio greater than 0.3 or less than −0.3. Heat map within bar shows degree of splicing change of those genes. (**D**) Venn diagram of genes from histograms for comparison of each genotype directly. Note: *set1*Δ and *htb1K123R* have 1 gene overlap which cannot be shown. The circle sizes are representative of the numbers of genes with $\log_2 I/E > 0.3$ but the overlaps are not to scale.



events, including transcription, mRNA processing and mRNA packaging that lead to their formation (Oeffinger et al., 2007). The proteome of CBC-associated mRNPs was enriched in splicing factors (snRNPs), 3′ end processing machinery, mRNA export factors and other factors that are recruited during transcription elongation such as the THO complex (Supplementary Table S4; Oeffinger et al., 2007). A recent transcriptome-wide analysis of CBC-associated mRNAs reveals a clear enrichment of unspliced versus spliced mRNA, consistent with an interaction occuring at an early step of transcription (Tuck and Tollervey, 2013). In addition, multiple subunits of RNA Polymerase II (RNAPII) were detected in

CBC-interacting mRNPs, confirming that some nascent transcripts were associated with TAP-tagged Cbc2 (Supplementary Table S4).

This global approach showed that preventing Ub-H2B impaired the association of the nuclear export machinery as previously described (Vitaliano-Prunier et al., 2012), but maintained WT levels of other factors known to bind mRNAs, including the transcription elongation THO complex (Figure 3A and Supplementary Table S4). Early spliceosome assembly onto pre-mRNAs entails binding of the U1 and U2 snRNPs: consistent with the observed splicing defect, the number of peptides corresponding to U1 and U2 was significantly lower in the *htb1K123R* strain compared with WT, whereas the overall protein level of these factors was similar in both strains (Figures 3A and 3C). This defect was also seen using a semi-quantitative analysis based on a spectral counting approach (Heintz et al., 2009) (Figure 3B and Tables S5 and S6). To confirm this decrease in CBC-associated U1 and U2, we directly assayed the co-immunoprecipitation of an endogenously HA-tagged version of either Prp42 (U1 snRNP) or Lea1 (U2 snRNP) with TAP-tagged Cbc2 and observed a reproducible decrease in both Prp42-HA and Lea1-HA (Figure 3C). Although we cannot rule out that some factors pulled down by TAP-tagged Cbc2 might be directly bound to the CBC, as has been observed for the tri-snRNP in mammals (Pabis et al., 2013), the decreased association of U1 and U2 proteins in the pull-down reported here is consistent with the splicing defect seen in strains lacking Ub-H2B. In contrast, this defect was not phenocopied by deletion of *SET1* (Figure 4), further arguing that the control of U1 and U2 recruitment by Ub-H2B is not mediated by downstream methylation by Set1.

Loss of H2B ubiquitylation impairs recruitment of early splicing factors to transcribing genes

It has been suggested that the splicing activity can be facilitated by the co-transcriptional recruitment of the splicing machinery (Tardiff et al., 2006; Aitken et al., 2011); thus, we reasoned that if the defect seen in recruitment of U1 and U2 to mRNPs occurs co-transcriptionally, this may further account for the intron accumulation phenotype of a strain lacking Ub-H2B. To ask this, we used chromatin immunoprecipitation (ChIP) to study early splicing factor association with on four intron-containing genes whose splicing is sensitive to loss of Ub-H2B: three ribosomal protein genes *RPS21B*, *RPL14B* and *RPL34B* and the non-ribosomal protein gene *HNT1* (Supplementary Table S3). Of note, *HNT1* splicing was also sensitive to deletion of *SET1* or *SET2* (Supplementary Table S3). We found a marked decrease in Prp42 association at these genes in the *htb1K123R* strain. Importantly, this decrease in U1 association occurred at genes where RNAPII levels and resulting mRNA expression remained unchanged (Figure 5A and data not shown), and does not occur in cells lacking the *SET1*, *SET2* and *DOT1* methyltransferases (Figure 5B). Furthermore, in agreement with data from the CBC pulldown (Figure 3), Lea1 exhibited weaker association with these genes in the *htb1K123R* strain, which was revealed primarily at the 3′ ends (Figure 5A). Our data reveal that loss of Ub-H2B adversely affects recruitment of the early splicing machinery to nascent transcripts.

Loss of H2B ubiquitylation marginally affects recruitment of Npl3

We previously showed that the SR protein Npl3 promotes U1 and U2 association with nascent transcripts and physically interacts with the Ub-H2B machinery (Kress et al., 2008; Moehle et al., 2012). Although we do see a modest decrease in Npl3 ChIP (Figure 6A) at genes whose splicing is promoted by Npl3 (Kress et al., 2008), this cannot account for the broad consequences of losing Ub-H2B on pre-mRNA splicing shown here, in particular on non-RPGs.

Similar to SR proteins in metazoa, Npl3 is known to associate with the C-terminal domain (CTD) of RNAPII upon Serine 2 phosphorylation (Dermody et al., 2008) but we only observed a weak decrease in Serine 2 phosphorylation on the RNAPII CTD at *RPS21B* and *RPL34B* in the *htb1K123R* strain (Figure 6B).

Discussion

Early splicing factors associate with nascent pre-mRNAs (Kotovic et al., 2003; Gornemann et al., 2005; Lacadie and Rosbash, 2005; Moore et al., 2006; Tardiff et al., 2006), and, therefore, it is critical to understand how the chromatin landscape contributes to co-transcriptional spliceosome assembly and function. Our results indicate that in budding yeast, an organism with relatively simple intron/exon architecture and limited splice site variation, multiple histone

Figure 3 | Preventing ubiquitylation of H2B alters the recruitment of early splicing factors to CBC-associated mRNPs
(**A**) Nuclear mRNPs were purified from Cbc2-TAP-tagged WT or *htb1K123R* cells after a 3-h shift to 39°C and associated proteins were analysed by MS–MS after SDS-PAGE. In order to highlight relative differences in the protein composition, the number of unique peptides for each indicated protein has been considered as index of abundance (Merz et al., 2007). Complete identification of the CBC proteomes is shown in Supplementary Table S4. (**B**) Components of the U1 and U2 snRNPs associated with nuclear mRNPs were semi-quantified using a spectral counting approach. The mean ± SD of spectral counts corresponding to three injections are indicated. Significance of the differences observed between both strains was evaluated using Student *t*-test (**P* 0.01–0.05; ****P* < 0.001). Significant differences are indicated in bold. (**C**) Nuclear mRNPs were purified from WT or *htb1K123R* cells expressing Cbc2-TAP and Prp42-HA or Lea1-HA. Co-purifying proteins were detected by Western blot with anti-HA or anti-Mex67 antibodies (left panel). The ratio of mRNP-associated proteins relative to the WT cells and to the immuno-purified Cbc2-TAP was determined from at least three independent experiments (mean ± SD) (right panel). Significance of the differences observed between both strains was evaluated using Student *t*-test (***P* 0.001–0.01).

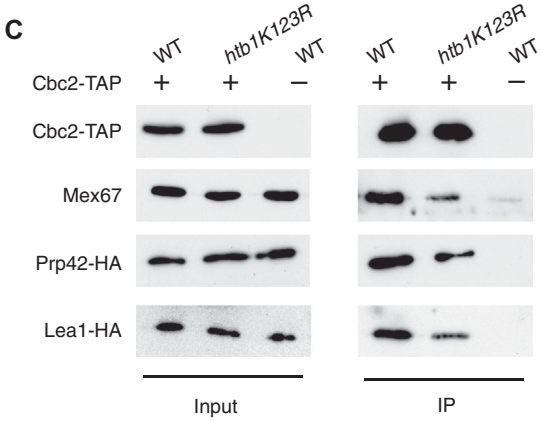
A

Complex	Protein	Number of peptides WT	*htb1K123R*
Cap binding complex	Cbc2	11	11
	Sto1	49	47
Export machinery	Mex67	13	10
	Mtr2	2	0
	Nab2	13	5
THO complex	Hpr1	11	11
	Tho2	44	43
	Thp2	4	4
	Mft1	1	5
U1 snRNP	Prp42	19	10
	Prp40	16	7
	Prp39	16	13
	Luc7	5	3
	Mud1	11	1
	Nam8	5	2
	Snu56	10	2
	Snu71	15	10
	Yhc1	6	2
	Prp5	2	0
	Snp1	13	4
U2 snRNP	Lea1	5	1
	Msl1	2	1
	Prp9	6	0
	Prp21	2	0
	Cus1	4	0
	Rse1	4	0
	Prp11	1	1

B

Complex	Protein	Spectral counts WT	*htb1K123R*	*t*-test
U1 snRNP	**Prp42**	11 ± 2	2 ± 0	*
	Prp39	12 ± 3	5 ± 0	*
	Luc7	5 ± 1	3 ± 1	
	Mud1	13 ± 2	4 ± 1	*
	Nam8	3 ± 0	1 ± 1	*
	Snu56	5 ± 2	2 ± 2	
	Snu71	15 ± 2	10 ± 1	*
	Yhc1	3 ± 1	2 ± 0	
U2 snRNP	**Lea1**	5 ± 0	0 ± 0	***
	Prp9	3 ± 1	0 ± 0	*
	Prp21	2 ± 1	1 ± 1	
	Rse1	4 ± 1	1 ± 1	
	Prp11	4 ± 1	0 ± 0	*

C



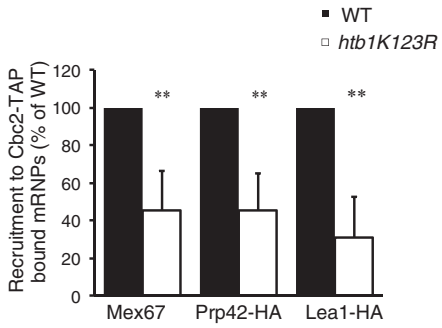
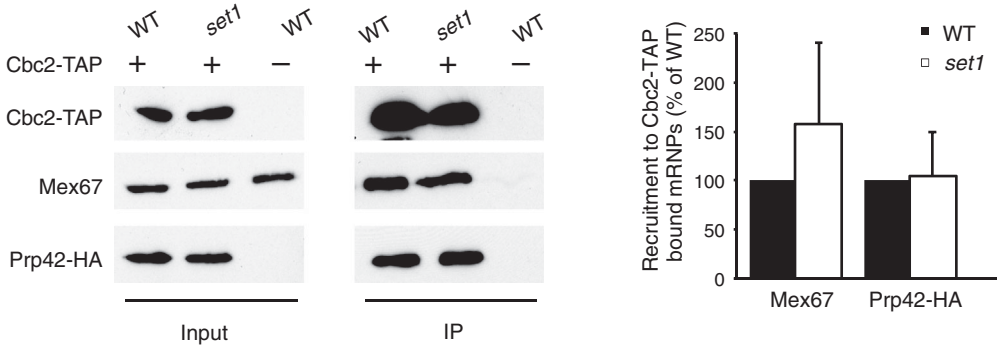


Figure 4 | Loss of Set1-dependent methylation does not affect the recruitment of U1 snRNP to CBC-associated mRNPs
Nuclear mRNPs were purified from WT or *set1*Δ cells expressing Cbc2-TAP and Prp42-HA. Co-purifying proteins were detected by Western blot with anti-HA or anti-Mex67 antibodies (left panel). The ratio of mRNP-associated proteins relative to the WT cells and to the immuno-purified Cbc2-TAP was determined from at least three independent experiments (mean ± SD) (right panel). Significance of the differences observed between both strains was evaluated using Student *t*-test (**P 0.001–0.01).



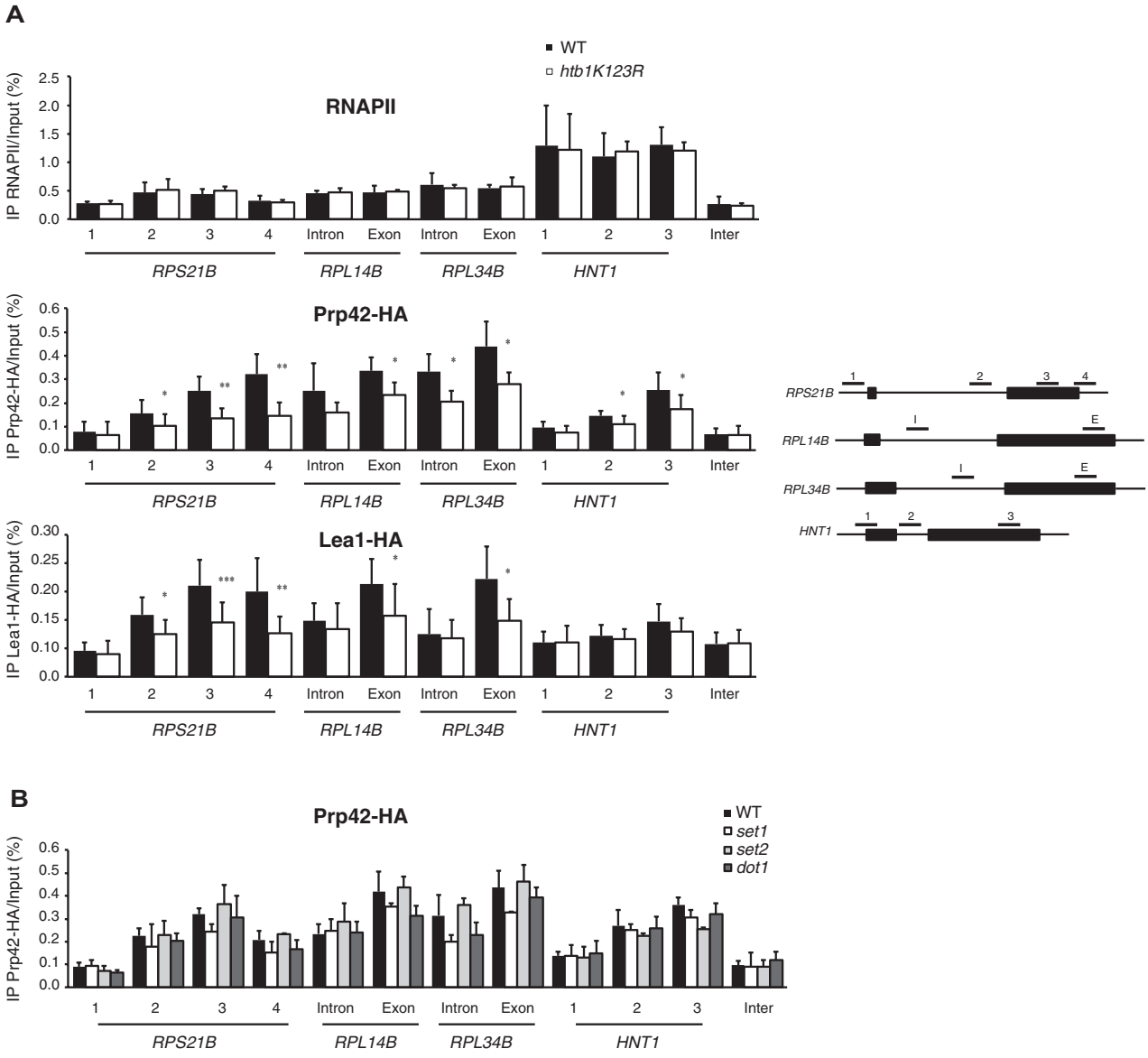
modifications (Ub-H2B, H3K4me and H3K36me) are required for optimal spliceosome function at distinct subsets of genes. In agreement with this, we also show that Ub-H2B facilitates the association of the early splicing machinery to mRNPs while the RNA is still being transcribed, in a molecular pathway that is distinct from Set1- or Set2-mediated splicing. Thus, our data support the idea that the chromatin landscape of a locus, as defined by these modifications, can impact the fate of a transcript, reflecting the capacity of the spliceosome to interpret and integrate multiple inputs from the chromatin landscape. This idea is further supported by evidence that histone acetylation by Gcn5 and deacetylation by Hos2/3 (Gunderson and Johnson, 2009; Gunderson et al., 2011), as well as incorporation of the variant histone H2A.Z (Albulescu et al., 2012) can also influence spliceosome assembly onto nascent transcripts and splicing efficiency of those transcripts.

In metazoa, where the recognition of and discrimination between alternative, degenerate splice sites must be tightly regulated, histone modifications, including Ub-H2B, have been shown to reflect intron/exon structure (Kolasinska-Zwierz et al., 2009; Schwartz et al., 2009; Spies et al., 2009; Dhami et al., 2010; Huff et al., 2010; Jung et al., 2012). In fact, modulation of the levels of the Ub-H2B machinery in human cells has revealed that this mark promotes recognition of splice sites, and importantly, as we show here in budding yeast, does so in context-dependent ways (Figures 1–3; Jung et al., 2012; Zhang et al., 2013). While Ub-H2B also reflects intron/exon structure on budding yeast genes (Shieh et al., 2011), it is remarkable that the one to two differentially marked nucleosomes associated with yeast introns, typically 100–400 bp in budding yeast (Spingola et al., 1999), can impact splicing. The conservation of the connection between Ub-H2B and splicing from yeast to humans suggests that this is a universal strategy for spliceosome regulation.

How does Ub-H2B promote splicing factor association? It is possible that Ub-H2B-mediated spliceosome association is, in part, influenced by a minor impairment in the recruitment of the CBC to some genes (not shown), as the CBC and early splicing factors physically interact in yeast (Colot et al., 1996). In addition, it has been recently shown that CBC depletion in mammalian cells led to defects in co-transcriptional spliceosome assembly via both RNA-dependent interaction with U1 and U2 snRNPs and RNA-independent interactions with U4/U6 and U5 snRNPs (Pabis et al., 2013). We could barely detect components of the U4/U6 and U5 snRNPs in our yeast CBC proteome, suggesting that this interaction may not occur in yeast. However, Pabis et al. (2013) hypothesised that CBC, U1 and U2 bind cooperatively onto RNA to promote splicing efficiency, a process that could be influenced by Ub-H2B. Conversely, Ub-H2B and H3K36 methylation levels are highly dependent on

Figure 5 | Loss of H2B ubiquitylation, but not Set1-, Set2- or Dot1-dependent methylation of H3, impairs recruitment of early splicing factors to transcribing genes

ChIP experiments were performed on extracts prepared from the indicated HA-tagged strains after a 3-h shift to 39°C, using anti-RNAPII or -HA antibodies. Four intron-containing genes were considered, three ribosomal protein genes *RPS21B*, *RPL14B* and *RPL34B* and the non-ribosomal protein gene *HNT1* (Supplementary Table S1). Histograms depict the mean and standard deviations of at least three independent experiments. The significance of the differences of recruitment observed between WT and mutant cells was evaluated using Student *t*-test (*P 0.01–0.05; **P 0.001–0.01).

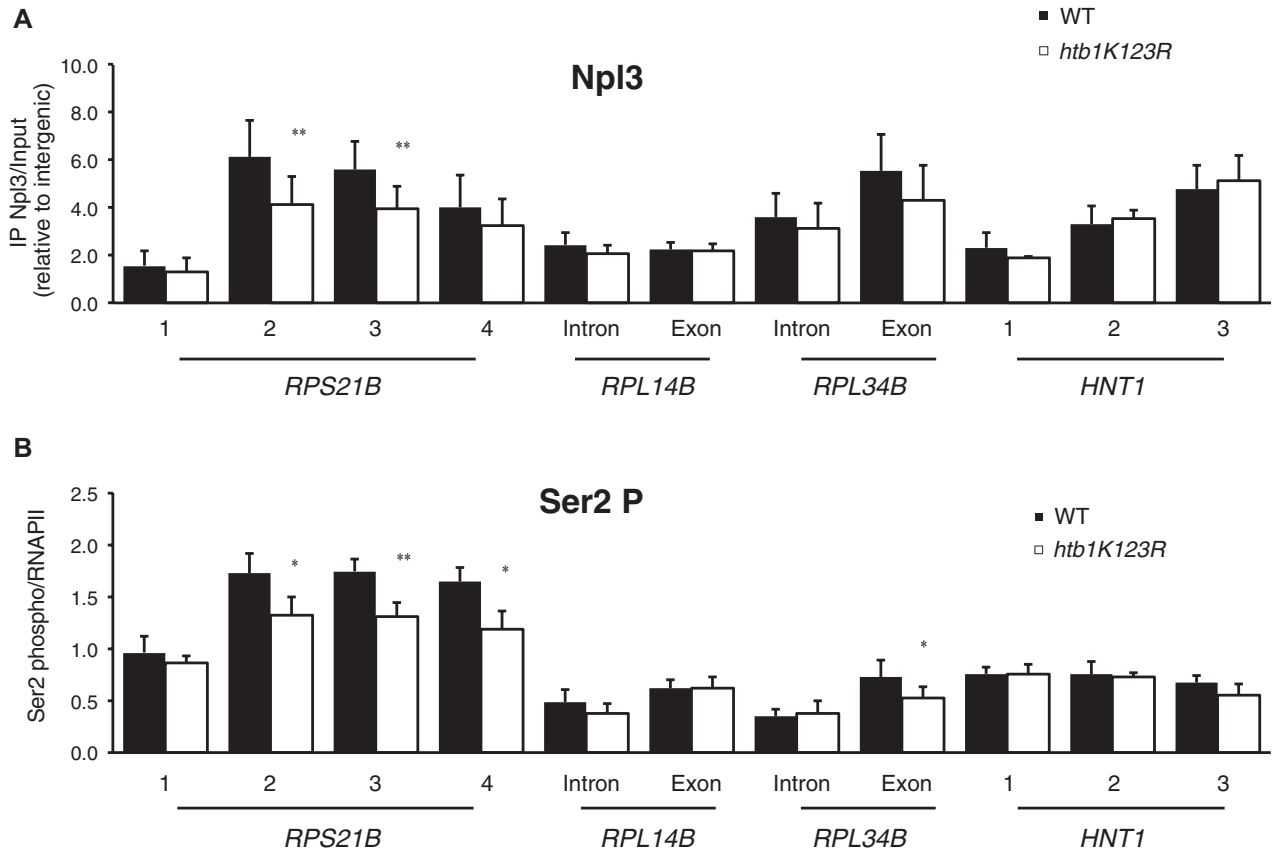


an intact CBC (Hossain et al., 2009; 2013), highlighting two examples of the high degree of coupling between mRNA processing and chromatin structure.

In metazoans, it has been suggested that histone modifications may directly recruit splicing factors via specific histone mark-specific adaptors (Sims et al., 2007; Luco et al., 2010); this has been previously

Figure 6 | H2B ubiquitylation promotes to Npl3 recruitment and RNAPII Ser2 phosphorylation on a subset of intron-containing genes

(**A**) Recruitment of Npl3 on indicated intron-containing genes was analysed by ChIP assay in WT and *htb1K123R* cells using anti Npl3 antibodies and normalised to the intergenic region. (**B**) Phosphorylation of RNAPII Ser2 on indicated intron-containing genes was analysed by ChIP assay in WT and *htb1K123R* cells using antibodies to RNAPII and phosphoSer2 specific antibodies. The RNAPII Ser2P/RNAPII ratio is shown. Significance of the differences observed between both strains was evaluated using Student *t*-test (*P 0.01–0.05; **P 0.001–0.01).



proposed to occur in yeast as well (Gunderson et al., 2011). Thus, it is possible that ubiquitylated H2B also directly or indirectly recruits a splicing factor in budding yeast. Alternatively, as histone modifications can influence RNAPII dynamics, it is possible that changes in elongation speed may explain the splicing defects seen in the chromatin mutants tested here (Howe et al., 2003; Braberg et al., 2013; Dujardin et al., 2013). Therefore, it is interesting that multiple reports have suggested that the ubiquitin modification stabilises nucleosomes during elongation (Chandrasekharan et al., 2009); given the splicing defects that can occur when RNA polymerase elongates too quickly (Braberg et al., 2013), perhaps uncontrolled RNAPII elongation in the *htb1K123R* strain accounts for the decrease in splicing efficiency reported here.

An intriguing aspect of chromatin-based splicing regulation is the complexity of the chromatin template; as our microarrays revealed, Ub-H2B, H3K4me and H3K36me influence splicing at somewhat overlapping subsets of genes. As the field moves forward, a challenge will be to understand how the spliceosome integrates the signals from individual histone modifications to achieve the appropriate splicing outcomes for the needs of the cell. The many histone post-translational modifications and their combined effects on transcription, RNAPII CTD

phosphorylation state and mRNA processing perhaps provide each emerging transcript with a specific mRNA 'identity' that allows dynamic modulation of splicing, mRNA export and quality control.

Materials and methods

Yeast strains and culture

Strains used in this study are listed in Supplementary Table 1. The derivative strains were obtained using PCR based homologous recombination as described in Longtine et al. (1998). Yeast cells were grown overnight at 30°C in yeast extract, peptone and dextrose (YPD) medium. To perform the 3-h shift at 39°C, when the 30°C overnight culture reached $OD_{600} = 1$, one volume of medium preheated to 48°C was added and cultures were incubated at 39°C for 3 h.

Microarray analysis

Strains were grown for microarray analysis as described in the Yeast strains and culture paragraph with a 3-h shift to 39°C, at which point they were collected by centrifugation. Microarrays were performed as in Moehle et al. (2012). For each microarray shown, results from two biological replicates were averaged. In addition, each biological replicate contains six technical replicates per probe, as well as dye-flipped replicates. The heat maps in Figures 1 and 2 were created using Java Treeview (Saldanha, 2004) and show the $\log_2$-based fold change in the indicated strain as compared with an isogenic WT. Intron/Exon ratios were calculated for each intron-containing gene [$\log_2$ (Intron/Exon) = $\log_2$ ($Intron_{mutant}/Intron_{WT}$) − $\log_2$ ($Exon_{mutant}/Exon_{WT}$)]; these values were converted into a histogram for any gene with Intron/Exon value <-0.3 or >0.3.

RNA isolation and RT-qPCR

Total RNA isolation was performed by the hot acid phenol method (Sigma–Aldrich). cDNA from total RNAs were obtained by retro-transcription with random oligonucleotides (Roche) using the SuperScriptTM II reverse Transcriptase (Invitrogen). Real time qPCR was then performed using the SYBR Green mix (Roche) and the Light Cycler 480 system (Roche) with gene specific primers described in Supplementary Table 2.

Antibodies

Commercial antibodies used in this study were anti-RNA polymerase II (RNAPII ChIP 12 µg, 8WG16, MMS126R; Covance), anti-phosphorylated Ser 2 of RNAPII CTD (ChIP 12 µg, clone 3E10, 04–1571; Millipore), anti-HA (ChIP 8 µg, WB 1:2000, clone HA-11, MMS-101R; Covance) anti-H3 (ChIP 4 µg, ab1791; Abcam), anti-H3K36me3 (ChIP 4 µg, ab9050; Abcam). Polyclonal anti-Mex67 (WB 1:20,000) and anti-Npl3 (ChIP 1.5 µl, WB 1:10,000) antibodies were previously described (Siebel and Guthrie, 1996; Gwizdek et al., 2005). TAP-tagged proteins were immunoprecipitated using IgG sepharose beads (50 µl for ChIP analysis). Western blot analyses were performed using appropriate horseradish peroxidase-coupled secondary antibodies and chemiluminescence protein immunoblotting reagents (Pierce).

Immunoaffinity purification of nuclear mRNPs from frozen cell grindate

Cells shifted at 39°C for 3 h in YPD were rapidly frozen in liquid nitrogen before cryolysis (Ossareh-Nazari et al., 2010). Immunoaffinity purification of mRNPs was performed as described in Oeffinger et al. (2007) with minor modification. Frozen cell grindates were rapidly thawed into nine volumes of RNP buffer [20 mM Hepes, pH 7.4, 110 mM KOAc, 0.5% Triton, 0.1% Tween 20, 1:100 Solution P, 1:5000 RNasin (Promega), 1:5000 Antifoam B (Sigma); 1:1000 DTT]. The resulting extracts were centrifuged 5 min at low speed to eliminate the large cellular debris before clarification by filtration. Immunoprecipitation was performed using magnetic beads (Dynal) conjugated with rabbit IgG (Sigma) (Alber et al., 2007). Resulting eluates were lyophilised, resuspended in SDS-PAGE sample buffer, separated by SDS-PAGE on a 4–12% NuPAGE Novex Bis-Tris precast gel (Invitrogen) according to the manufacturer's specifications and visualised by Coomassie blue staining.

Alternatively, nuclear mRNPs were purified from glass-beads lysed cells as previously described (Iglesias et al., 2010; Vitaliano-Prunier et al., 2012).

SDS-PAGE and in-gel digestion

After electrophoresis, gels were stained by colloidal Coomassie Blue (BioSafe coomassie stain; Bio-Rad) and whole lanes were cut into 14 5 × 4 mm^2 slices using a disposable grid-cutter (The Gel-Company). Slices were divided into three and in-gel digestion using trypsin (Promega) was performed overnight at 37°C, after in-gel reduction and alkylation using the MassPrep Station (Waters). Tryptic peptides were extracted using 60% ACN in 0.1% formic acid for 1 h at room temperature. The volume was reduced in a vacuum centrifuge and adjusted to 10 µl using 0.1% formic acid in water before nanoLC–MS/MS (nanoliquid chromatography coupled to tandem mass spectrometry) analysis (see Supplementary information).

Liquid chromatography/mass spectrometry

NanoLC–MS/MS was performed using a nanoACQUITY ultra performance liquid chromatography (UPLC®) system (Waters) coupled to a maXis 4G Q-TOF mass spectrometer (BrukerDaltonics). Detailed procedure is presented in Supplementary information.

Peptide counts

The peptide count of unique peptides was performed using Scaffold 3 software (version 3.6.5; Proteome Software Inc.) after having filtered the results at 1% < False Discovery Rate (FDR).

Spectral counts

The label-free semi-quantitation was performed using Scaffold 3 software (version 3.6.5; Proteome Software Inc.) exporting the un-weighed spectral count into an excel file after having filtered the result at 1% < False Discovery Rate as described in the Supplementary data. The three replicates of each biological sample shown in Supplementary Table S6 (except the negative control) were horizontally normalised to the highest spectral count value obtained for nuclear cap-binding protein complex

subunit (WT injection 1:429 un-weighed spectra counts). The following procedure as described in Gokce et al. (2011) and Miguet et al. (2013) has been applied for the six independent injections: the number of spectra obtained for the nuclear CBC has been divided by 429 and the ratio obtained was multiplied independently by the number of spectral counts for each protein, in order to reduce the variance observed between samples and replicates.

ChIP and qPCR

ChIPs were performed as described previously (Gwizdek et al., 2006). Eight OD units of cell lysate were immunoprecipitated with the amount of antibodies indicated in the Antibodies paragraph. Immunoprecipitated DNA was analysed by quantitative PCR using primers referenced in Supplementary Table S2. Non-specific signals were assessed by analysing immunoprecipitated DNA with primers against an intergenic region. The specificity of the signal was also evaluated using untagged strains. The resulting amplifications were similar to those observed for the intergenic region (not shown). Results correspond to the mean of at least three independent experiments.

Author contribution

L.H., E.A.M., D.B. performed experiments; L.H., E.A.M., D.B., A.V.D., C.S., C.G., C.D. designed the experiments and analysed the data; C.D., E.A.M. and C.G. wrote the manuscript.

Funding

This study was supported by the National Research Agency (grant 2010 BLAN1227-01 to C.D.), the Who am I? laboratory of excellence (ANR-11-LABX-0071 to C.D.) funded by the 'Investments for the Future' program (ANR-11-IDEX-0005-01), the NIH grant NIHGM21119 to C.G., the Proteomic French Infrastructure (ProFI to A.V.D. and C.S.R.) and the Fondation pour la Recherche Médicale (to A.V.D. and C.S.R.). L.H. is supported by the University Paris Descartes and the Association de Recherche contre le Cancer (ARC).

Acknowledgements

We are grateful to Thanasis Margaritis and Frank Holstege for helpful advice. C.G. is an American Cancer Society Research Professor of Molecular Genetics.

Conflict of interest statement

The authors have declared no conflict of interest.

References

Aitken, S., Alexander, R.D. and Beggs, J.D. (2011) Modelling reveals kinetic advantages of co-transcriptional splicing. PLoS Comput. Biol. **7**, e1002215

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., Sali, A. and Rout, M.P. (2007) The molecular architecture of the nuclear pore complex. Nature **450**, 695–701

Albulescu, L.O., Sabet, N., Gudipati, M., Stepankiw, N., Bergman, Z.J., Huffaker, T.C. and Pleiss, J.A. (2012) A quantitative, high-throughput reverse genetic screen reveals novel connections between pre-mRNA splicing and 5′ and 3′ end transcript determinants. PLoS Genet. **8**, e1002530

Ares, M., Jr., Grate, L. and Pauling, M.H. (1999) A handful of intron-containing genes produces the lion's share of yeast mRNA. RNA **5**, 1138–1139

Babour, A., Dargemont, C. and Stutz, F. (2012) Ubiquitin and assembly of export competent mRNP. Biochim. Biophys. Acta **1819**, 521–530

Bergkessel, M., Whitworth, G.B. and Guthrie, C. (2011) Diverse environmental stresses elicit distinct responses at the level of pre-mRNA processing in yeast. RNA **17**, 1461–1478

Braberg, H., Jin, H., Moehle, E.A., Chan, Y.A., Wang, S., Shales, M., Benschop, J.J., Morris, J.H., Qiu, C., Hu, F., Tang, L.K., Fraser, J.S., Holstege, F.C., Hieter, P., Guthrie, C., Kaplan, C.D. and Krogan, N.J. (2013) From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell **154**, 775–788

Briggs, S.D., Xiao, T., Sun, Z.W., Caldwell, J.A., Shabanowitz, J., Hunt, D.F., Allis, C.D. and Strahl, B.D. (2002) Gene silencing: trans-histone regulatory pathway in chromatin. Nature **418**, 498

Carrillo Oesterreich, F., Bieberstein, N. and Neugebauer, K.M. (2011) Pause locally, splice globally. Trends Cell Biol. **21**, 328–335

Carrillo Oesterreich, F., Preibisch, S. and Neugebauer, K.M. (2010) Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol. Cell **40**, 571–581

Chandrasekharan, M.B., Huang, F. and Sun, Z.W. (2009) Ubiquitination of histone H2B regulates chromatin dynamics by enhancing nucleosome stability. Proc. Natl. Acad. Sci. U.S.A. **106**, 16686–16691

Clark, T.A., Sugnet, C.W. and Ares, M., Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science **296**, 907–910

Colot, H.V., Stutz, F. and Rosbash, M. (1996) The yeast splicing factor Mud13p is a commitment complex component and corresponds to CBP20, the small subunit of the nuclear cap-binding complex. Genes Dev. **10**, 1699–1708

Dermody, J.L., Dreyfuss, J.M., Villen, J., Ogundipe, B., Gygi, S.P., Park, P.J., Ponticelli, A.S., Moore, C.L., Buratowski, S. and Bucheli, M.E. (2008) Unphosphorylated SR-like protein Npl3 stimulates RNA polymerase II elongation. PLoS One **3**, e3273

Dhami, P., Saffrey, P., Bruce, A.W., Dillon, S.C., Chiang, K., Bonhoure, N., Koch, C.M., Bye, J., James, K., Foad, N.S., Ellis, P., Watkins, N.A., Ouwehand, W.H., Langford, C., Andrews, R.M., Dunham, I. and Vetrie, D. (2010) Complex exon–intron marking by histone modifications is not determined solely by nucleosome distribution. PLoS One **5**, e12339

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gomez Acuna, L.I., Fiszbein, A., Godoy Herz, M.A., Nieto Moreno, N., Munoz, M.J., Allo, M., Schor, I.E. and Kornblihtt, A.R. (2013) Transcriptional elongation and alternative splicing. Biochim. Biophys. Acta **1829**, 134–140

Gokce, E., Shuford, C.M., Franck, W.L., Dean, R.A. and Muddiman, D.C. (2011) Evaluation of normalization methods on GeLC-MS/MS label-free spectral counting data to correct for variation during proteomic workflows. J. Am. Soc. Mass. Spectrom. **22**, 2199–2208

Gornemann, J., Kotovic, K.M., Hujer, K. and Neugebauer, K.M. (2005) Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. Mol. Cell **19**, 53–63

Gunderson, F.Q. and Johnson, T.L. (2009) Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. PLoS Genet. **5**, e1000682

Gunderson, F.Q., Merkhofer, E.C. and Johnson, T.L. (2011) Dynamic histone acetylation is critical for cotranscriptional spliceosome assembly and spliceosomal rearrangements. Proc. Natl. Acad. Sci. U.S.A. **108**, 2004–2009

Gwizdek, C., Hobeika, M., Kus, B., Ossareh-Nazari, B., Dargemont, C. and Rodriguez, M.S. (2005) The mRNA nuclear export factor Hpr1 is regulated by Rsp5-mediated ubiquitylation. J. Biol. Chem. **280**, 13401–13405

Gwizdek, C., Iglesias, N., Rodriguez, M.S., Ossareh-Nazari, B., Hobeika, M., Divita, G., Stutz, F. and Dargemont, C. (2006) Ubiquitin-associated domain of Mex67 synchronizes recruitment of the mRNA export machinery with transcription. Proc. Natl. Acad. Sci. U.S.A. **103**, 16376–16381

Heintz, D., Gallien, S., Wischgoll, S., Ullmann, A.K., Schaeffer, C., Kretzschmar, A.K., van Dorsselaer, A. and Boll, M. (2009) Differential membrane proteome analysis reveals novel proteins involved in the degradation of aromatic compounds in Geobacter metallireducens. Mol. Cell. Proteomics **8**, 2159–2169

Hnilicova, J. and Stanek, D. (2011) Where splicing joins chromatin. Nucleus **2**, 182–188

Hossain, M.A., Chung, C., Pradhan, S.K. and Johnson, T.L. (2013) The yeast cap binding complex modulates transcription factor recruitment and establishes proper histone H3K36 trimethylation during active transcription. Mol. Cell. Biol. **33**, 785–799

Hossain, M.A., Claggett, J.M., Nguyen, T. and Johnson, T.L. (2009) The cap binding complex influences H2B ubiquitination by facilitating splicing of the SUS1 pre-mRNA. RNA **15**, 1515–1527

Howe, K.J., Kane, C.M. and Ares, M., Jr. (2003) Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. RNA **9**, 993–1006

Huff, J.T., Plocik, A.M., Guthrie, C. and Yamamoto, K.R. (2010) Reciprocal intronic and exonic histone modification regions in humans. Nat. Struct. Mol. Biol. **17**, 1495–1499

Iglesias, N., Tutucci, E., Gwizdek, C., Vinciguerra, P., Von Dach, E., Corbett, A.H., Dargemont, C. and Stutz, F. (2010) Ubiquitin-mediated mRNP dynamics and surveillance prior to budding yeast mRNA export. Genes Dev. **24**, 1927–1938

Jung, I., Kim, S.K., Kim, M., Han, Y.M., Kim, Y.S., Kim, D. and Lee, D. (2012) H2B monoubiquitylation is a 5'-enriched active transcription mark and correlates with exon–intron structure in human cells. Genome Res. **22**, 1026–1035

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S. and Ahringer, J. (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. Nat. Genet. **41**, 376–381

Kotovic, K.M., Lockshon, D., Boric, L. and Neugebauer, K.M. (2003) Cotranscriptional recruitment of the U1 snRNP to intron-containing genes in yeast. Mol. Cell. Biol. **23**, 5768–5779

Kress, T.L., Krogan, N.J. and Guthrie, C. (2008) A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. Mol. Cell **32**, 727–734

Lacadie, S.A. and Rosbash, M. (2005) Cotranscriptional spliceosome assembly dynamics and the role of U1 snRNA:5′ss base pairing in yeast. Mol. Cell **19**, 65–75

Lenstra, T.L., Benschop, J.J., Kim, T., Schulze, J.M., Brabers, N.A., Margaritis, T., van de Pasch, L.A., van Heesch, S.A., Brok, M.O., Groot Koerkamp, M.J., Ko, C.W., van Leenen, D., Sameith, K., van Hooff, S.R., Lijnzaad, P., Kemmeren, P., Hentrich, T., Kobor, M.S., Buratowski, S. and Holstege, F.C. (2011) The specificity and topology of chromatin interaction pathways in yeast. Mol. Cell **42**, 536–549

Longtine, M.S., McKenzie, A., 3rd, Demarini, D.J., Shah, N.G., Wach, A., Brachat, A., Philippsen, P. and Pringle, J.R. (1998) Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. Yeast **14**, 953–961

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R. and Misteli, T. (2011) Epigenetics in alternative pre-mRNA splicing. Cell **144**, 16–26

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M. and Misteli, T. (2010) Regulation of alternative splicing by histone modifications. Science **327**, 996–1000

Margaritis, T., Oreal, V., Brabers, N., Maestroni, L., Vitaliano-Prunier, A., Benschop, J.J., van Hooff, S., van Leenen, D., Dargemont, C., Geli, V. and Holstege, F.C. (2012) Two distinct repressive mechanisms for histone 3 lysine 4 methylation through promoting 3′-end antisense transcription. PLoS Genet. **8**, e1002952

Merz, C., Urlaub, H., Will, C.L. and Luhrmann, R. (2007) Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. RNA **13**, 116–128

Miguet, L., Lennon, S., Baseggio, L., Traverse-Glehen, A., Berger, F., Perrusson, N., Chenard, M.P., Galoisy, A.C., Eischen, A., Mayeur-Rousse, C., Maar, A., Fornecker, L., Herbrecht, R., Felman, P., Van Dorsselaer, A., Carapito, C., Cianferani, S. and Mauvieux, L. (2013) Cell-surface expression of the TLR homolog CD180 in circulating cells from splenic and nodal marginal zone lymphomas. Leukemia **27**, 1748–1750

Moehle, E.A., Ryan, C.J., Krogan, N.J., Kress, T.L. and Guthrie, C. (2012) The yeast SR-like protein Npl3 links chromatin modification to mRNA processing. PLoS Genet. **8**, e1003101

Moore, M.J., Schwartzfarb, E.M., Silver, P.A. and Yu, M.C. (2006) Differential recruitment of the splicing machinery during transcription predicts genome-wide patterns of mRNA splicing. Mol. Cell **24**, 903–915

Munding, E.M., Igel, A.H., Shiue, L., Dorighi, K.M., Trevino, L.R. and Ares, M., Jr. (2010) Integration of a splicing regulatory network within the meiotic gene expression program of *Saccharomyces cerevisiae*. Genes Dev. **24**, 2693–2704

Ng, H.H., Xu, R.M., Zhang, Y. and Struhl, K. (2002) Ubiquitination of histone H2B by Rad6 is required for efficient Dot1-mediated methylation of histone H3 lysine 79. J. Biol. Chem. **277**, 34655–34657

Nino, C.A., Herissant, L., Babour, A. and Dargemont, C. (2013) mRNA nuclear export in yeast. Chem. Rev. **113**, 8523–8545

Oeffinger, M., Wei, K.E., Rogers, R., DeGrasse, J.A., Chait, B.T., Aitchison, J.D. and Rout, M.P. (2007) Comprehensive analysis of diverse ribonucleoprotein complexes. Nat. Methods **4**, 951–956

Ossareh-Nazari, B., Bonizec, M., Cohen, M., Dokudovskaya, S., Delalande, F., Schaeffer, C., Van Dorsselaer, A. and Dargemont, C. (2010) Cdc48 and Ufd3, new partners of the ubiquitin protease Ubp3, are required for ribophagy. EMBO Rep. **11**, 548–554

Pabis, M., Neufeld, N., Steiner, M.C., Bojic, T., Shav-Tal, Y. and Neugebauer, K.M. (2013) The nuclear cap-binding complex interacts with the U4/U6.U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. RNA **19**, 1054–1063

Pleiss, J.A., Whitworth, G.B., Bergkessel, M. and Guthrie, C. (2007) Rapid, transcript-specific changes in splicing in response to environmental stress. Mol. Cell **27**, 928–937

Robzyk, K., Recht, J. and Osley, M.A. (2000) Rad6-dependent ubiquitination of histone H2B in yeast. Science **287**, 501–504

Saldanha, A.J. (2004) Java Treeview—extensible visualization of microarray data. Bioinformatics **20**, 3246–3248

Schwartz, S., Meshorer, E. and Ast, G. (2009) Chromatin organization marks exon–intron structure. Nat. Struct. Mol. Biol. **16**, 990–995

Shieh, G.S., Pan, C.H., Wu, J.H., Sun, Y.J., Wang, C.C., Hsiao, W.C., Lin, C.Y., Tung, L., Chang, T.H., Fleming, A.B., Hillyer, C., Lo, Y.C., Berger, S.L., Osley, M.A. and Kao, C.F. (2011) H2B ubiquitylation is part of chromatin architecture that marks exon-intron structure in budding yeast. BMC Genomics **12**, 627

Siebel, C.W. and Guthrie, C. (1996) The essential yeast RNA binding protein Np13p is methylated. Proc. Natl. Acad. Sci. U.S.A. **93**, 13641–13646

Sims, R.J., 3rd, Millhouse, S., Chen, C.F., Lewis, B.A., Erdjument-Bromage, H., Tempst, P., Manley, J.L. and Reinberg, D. (2007) Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol. Cell **28**, 665–676

Spies, N., Nielsen, C.B., Padgett, R.A. and Burge, C.B. (2009) Biased chromatin signatures around polyadenylation sites and exons. Mol. Cell **36**, 245–254

Spingola, M., Grate, L., Haussler, D. and Ares, M., Jr. (1999) Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. RNA **5**, 221–234

Suganuma, T. and Workman, J.L. (2011) Signals and combinatorial functions of histone modifications. Annu. Rev. Biochem. **80**, 473–499

Sun, Z.W. and Allis, C.D. (2002) Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. Nature **418**, 104–108

Tardiff, D.F., Lacadie, S.A. and Rosbash, M. (2006) A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. Mol. Cell **24**, 917–929

Tuck, A.C. and Tollervey, D. (2013) A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. Cell **154**, 996–1009

Vitaliano-Prunier, A., Babour, A., Herissant, L., Apponi, L., Margaritis, T., Holstege, F.C., Corbett, A.H., Gwizdek, C. and Dargemont, C. (2012) H2B ubiquitylation controls the formation of export-competent mRNP. Mol. Cell **45**, 132–139

Vitaliano-Prunier, A., Menant, A., Hobeika, M., Geli, V., Gwizdek, C. and Dargemont, C. (2008) Ubiquitylation of the COMPASS component Swd2 links H2B ubiquitylation to H3K4 trimethylation. Nat. Cell Biol. **10**, 1365–1371

Wood, A., Krogan, N.J., Dover, J., Schneider, J., Heidt, J., Boateng, M.A., Dean, K., Golshani, A., Zhang, Y., Greenblatt, J.F., Johnston, M. and Shilatifard, A. (2003) Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter. Mol. Cell **11**, 267–274

Zhang, Z., Jones, A., Joo, H.Y., Zhou, D., Cao, Y., Chen, S., Erdjument-Bromage, H., Renfrow, M., He, H., Tempst, P., Townes, T.M., Giles, K.E., Ma, L. and Wang, H. (2013) USP49 deubiquitinates histone H2B and regulates cotranscriptional pre-mRNA splicing. Genes Dev. **27**, 1581–1595

Received: 17 January 2014; Accepted: 24 January 2014; Accepted article online: 29 January 2014

Analyse protéomique: progrès en caractérisation et en quantification

Advances in analytical methodologies for the characterization and quantification in proteomic analysis

Résumé

L'objectif de cette thèse était de développer et d'optimiser de nouvelles méthodologies et approches analytiques afin d'améliorer le potentiel de l'analyse protéomique pour les études biologiques.
La première partie de ce travail est consacrée à la détermination massive et exacte de la position N-terminale des protéines (N-terminome). Pour cela, nous avons utilisé et développé une approche basée sur une dérivation N-terminale au TMPP. Cette méthodologie de marquage de la position N-terminale a permis d'aborder l'étude des clivages protéolytiques des protéines exportées par le parasite P. falciparum (pathogène de la malaria) dans le globule rouge.
Afin de permettre une exploitation automatique à haut débit des données de MS/MS, nous avons élaboré une nouvelle méthodologie (dénommée dN-TOP). Celle-ci repose sur l'utilisation de TMPP portant des isotopes stables et permet ainsi d'accéder à la détermination des positions N-terminales pour des études de N-terminome à large échelle.
La seconde partie est dédiée aux développements de différentes stratégies analytiques de quantification, aussi bien au niveau peptidique qu'au niveau protéique, appliquées à une série de problématiques biologiques. Ces optimisations ont été réalisées dans le contexte de l'étude des complexes protéiques, du dosage de prion par SRM, de quantification des glycations d'anticorps monoclonaux thérapeutiques et de l'hémoglobine HbA2 pour la standardisation des méthodes de référence.

Abstract

The objective of this Ph.D. thesis was to develop and optimize new methodologies and analytical approaches to improve the potential of the mass spectrometry based proteomics.
The first part of this work focused on the development of the N-termini proteomics. This topic was addressed with a specific N-termini chemical derivatization based on TMPP. We have shown that our method allowed both specific N-terminomics and classical proteomics studies in the same experiment.
This N-terminus methodology was applied to study the proteolytic cleavages of the exported proteins in P. falciparum, a parasite responsible for the malaria.
In order to automatize the complex and tedious informatics processsing of the MS/SM data of our TMPP based N-terminomics method, we have introduced a new approach (named dN-TOP), based on the use of a stable isotope labeled TMPP which made now N-terminome proteomics compatible with high throughput studies.
The second part addresses quantitative aspects of proteomics. It describes the optimization of quantitative methods at the peptide level or at the protein level for five different proteomic studies in the context of protein complex subunits, targeted SRM based prion, quantification of monoclonal antibodies glycation and hemoglobin HbA2 for reference measurement methods standardization.

Mots-clès key words :
Proteogenomics, protein N-terminus, proteolytic cleavage, TMPP, label free quantitation, DDA, DIA.