THESE

Présentée pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITE DE STRASBOURG

par

Nicolas BARTHÉLEMY

Protéomique qualitative et quantitative, une passerelle pour relier l'expression génomique à la construction des édifices biologiques. Application à la compréhension de la structure moléculaire du cheveu humain.

Soutenue le 22 juin 2011 devant la commission d'examen :

Dr Alain VAN DORSSELAER	Directeur de thèse
Dr Odile SCHILTZ	Rapporteur
Dr Thierry RABILLOUD	Rapporteur
Dr Olivier POCH	Examinateur
Dr Bruno BERNARD	Examinateur

A ma famille,

A la mémoire de mon grand-père Pierre Tabarant,

« On ne trouve que ce qu'on a idée de chercher. »

Cours de chimie analytique d'Alain Tchapla, IUT d'Orsay

Remerciements

Ce travail de thèse a été réalisé au Laboratoire de Spectrométrie de Masse Bio-Organique, au sein du Département Sciences Analytiques de l'Institut Pluridisciplinaire Hubert Curien de Strasbourg (IPHC, CNRS-UdS 7178).

Tout d'abord, je tiens à remercier Alain Van Dorsselaer pour m'avoir permis de bénéficier des moyens techniques et des compétences humaines et scientifiques de son laboratoire.

Je tiens également à adresser ma reconnaissance à L'ORÉAL et personnellement à Dominique Julien pour avoir soutenu financièrement ce travail. Mes remerciements s'adressent particulièrement à Nükhet Cavusoglu et Georges Hussler pour avoir suivi mon travail tout au long de ces années. J'ai une pensée pour Franck Zerbib grâce à qui j'ai mis les doigts dans l'engrenage.

Je suis profondément reconnaissant à Olivier Poch, président du jury, Odile Schiltz et Thierry Rabilloud, rapporteurs, et Bruno Bernard, examinateur, pour m'avoir fait l'honneur d'évaluer ce travail de thèse.

Merci à Christine Schaeffer-Reiss pour le temps consacré à mon encadrement au laboratoire.

Un remerciement spécial à Christine Carapito pour son aide, son écoute et ses encouragements tout au long de mon périple docto «râle».

Merci tout particulièrement à Cyril Colas pour m'avoir fait bénéficier de ses inestimables connaissances entre autres des couplages LC-MS et pour toutes les manips faites jour et surtout nuit pour entrevoir la rationalité du séquençage en protéomique.

Je remercie René Brennetot pour m'avoir permis de faire une collaboration extra-capillaire et pour ses précieux conseils sur les plans d'expérience.

Merci aux cadres du laboratoire : Jean Marc Strub (qui est le plus fort : l'Alien, la Mammoet PTC 35 DS ou le Prédator ?), Danièle Thierse (Grrr… encore un coup des Snockeles), François Delalande (il est à tout le monde le canap'…), Agnès Hovasse, Hélène Diemer, Fabrice Varrier, Fabrice Bertile, Sarah Sanglier-Cianférani et Laurence Sabatier. Merci à Alexandre Burel et Patrick Guterl (toc, toc, toc, c'est ici pour les macros ?) et aussi à Véronique Trimbour et Kévin Jeanpert pour leur travail qui nous a simplifié la vie et les départs en congrès.

Mes sincères remerciements aux deux équipes ~~de grouillots~~ d'étudiants, post-docs, contractuels qui ont accompagné mon séjour au LSMBO rythmé par la bonne humeur et les échanges d'idées et d'expériences scientifiques. Courage, encore un peu de négociations et il y aura des lits pliants à la cave (avec planning de réservation intégré au CEI).

Les *Pouilleux* : Sébastien Gallien (dit le grand chef, guide spirituel dans mon parcours initiatique à la protéomique), Audrey Bednarczyck, Cédric Atmanène (maître suprême non-covalentiste), Thierry Wasselin (697 heures de PD-Quest et 18 palettes de REM au compteur), Daniel Ayoub, Céline Heng, Stéphanie Petiot-Becard et François Debaene.

Les *Précieux* : Guillaume Béchade (précieux parmi les précieux), Véronique Delval-Dubois, Laëtitia Fouillen, Jean-Michel Saliou, Christel Husser, Antoine Zieger, Amandine Bœuf (l'exilée), Tchilabalo Dilezitoko ''Elie'' Alayi (une question ?), Sarah Lennon et Magali Rompais (56 blagues et mises en scène de bureau à leur actif, série en cours), Diego Bertaccini et Marine Plumel.

Merci à l'ensemble des membres de la section volley de la Fraternelle pour tous les bons moments extra labo sportifs et non sportifs (principalement au MB…).

Enfin, j'ai une pensée pour mes professeurs de la Licence Professionnelle de Chimie Analytique de l'IUT d'Orsay et du Master Instrumentation et Méthodes d'Analyse Moléculaire de l'Université Paris Sud pour la qualité de leurs formations. Je pense également à tous ceux qui ont pris de leur temps pour me transmettre de leurs connaissances et de leur savoir-faire pendant mes stages et particulièrement à Michel Tabarant, Jean-Marie Duda, Patrick Dupuis et Laurent Verreman.

Plan du manuscrit

Partie II Mise au point de nouvelles stratégies protéomiques pour la caractérisation de protéines issues des familles multigéniques pour l'analyse des constituants du cheveu

Partie IV Perspectives à l'analyse des protéomes des kératinocytes : la compréhension du rôle et de l'origine des protéines dans la formation des structures du cheveu

Liste des principales abréviations

2D-GE	Electrophorèse bidimensionnelle
ACN	Acétonitrile
ADC	Analog to digital converter
ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ARNc	ARN complémentaire
ARNm	ARN messager
BLAST	Basic local alignment search tool
CE	Collision energy
CID	Collision-induced dissociation
CMC	Complexe de membrane cellulaire
ESI	Ionisation électrospray
GGEL	γ-glutamyl-ε-lysine
HILIC	Chromatographie d'interaction hydrophile
HS	High sulfur
HGT	High glycine and tyrosine
IF	Filament intermédiaire
KAP	Protéine associée aux kératines
KIF	Filament intermédiaire de kératine
KRT	Gène de kératine
KRTAP	Gène de KAP
LC	Chromatographie liquide
LC 2D	Chromatographie liquide bidimensionnelle
MALDI	Désorption / ionisation laser assistée par matrice
MCP	Multichannel plate
MS	Spectrométrie de masse
MS/MS	Spectrométrie de masse en tandem
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
Q	Quadrupôle
SCX	Strong cation exchange
SDS	Sodium dodécyl sulfate
SIM	Single ion monitoring
RP	Reverse Phase
TDC	Time-to-digital convertor
TGM	Transglutaminase
TOF	Time-of-flight
UHS	Ultra high sulfur
ULF	Unité de longueur filamentaire

Introduction générale

La fibre capillaire est une structure biologique commune à toutes les espèces de mammifères et a incontestablement contribué à leur colonisation même dans les biotopes les plus hostiles de la planète. Cet édifice moléculaire complexe et original joue de multiples fonctions allant de la fourrure pour protéger l'organisme contre le froid et l'humidité jusqu'aux piquants pour dissuader la prédation. Chez l'humain, la pilosité de l'épiderme est caractérisée par une faible longueur qui, associée au mécanisme de sudation, permet une régulation efficace de la température et contribue à ses aptitudes uniques d'endurance dans un environnement chaud. Les cheveux, plus longs, ont subsisté et se retrouvent dans la population humaine sous de nombreux phénotypes. Ils sont désormais dans les sociétés des symboles de beauté et sujets aux modes.

De nombreuses études ont déjà été menées chez l'humain et d'autres mammifères pour comprendre la structure de la fibre, son organisation, sa composition ainsi que les mécanismes biologiques conduisant à sa formation. Dans ce contexte, le follicule pileux, base de la structure capillaire, apparait comme un excellent modèle pour la compréhension de la construction des structures biologiques. Cependant, différentes questions relatives à la composition et à l'arrangement des protéines dans la fibre restent posées et les connaissances actuelles ne permettent pas d'expliquer les propriétés mécaniques et la grande diversité des phénotypes de fibres observées chez l'humain et globalement chez les mammifères.

Le séquençage du génome humain au début des années 2000 a ouvert la voie à l'étude des protéines s'exprimant au cours de la différenciation des cellules du follicule pileux. Différents gènes, exprimés au niveau du transcriptome de ces cellules, ont été identifiés et mettent en évidence l'implication d'un grande diversité de familles multigéniques dans ces structures : les kératines de type I, de type II ainsi que différentes familles dites protéines associées aux kératines (KAP). La preuve de l'expression au niveau protéique de ces gènes a pu être obtenue pour un certain nombre d'entre eux grâce à l'utilisation d'anticorps monoclonaux dirigés vers des domaines spécifiques des protéines. Pourtant, la très forte homologie de séquence des protéines associées aux kératines ne permet pas de disposer d'anticorps spécifiques et d'étudier l'expression protéiques des différents gènes correspondants.

Dans ce contexte, l'utilisation des outils protéomiques parait particulièrement adaptée pour caractériser les protéines présentes et confronter ces résultats expérimentaux aux données génomiques. Elle permet d'envisager l'étude d'éventuelles modifications post traductionnelles présentes naturellement ou issues de l'exposition à des traitements de la fibre mais également d'évaluer l'implication de ces protéines dans la structure.

Pour autant, la spécificité de l'analyse de protéines homologues reste un défi puisqu'il est nécessaire de les distinguer précisément les unes des autres. Le recours aux techniques de séparation des protéines entières n'est pas adapté à l'isolement et à la détection de ces isoformes. L'analyse des extraits digérés et la recherche des peptides spécifiques à chaque isoforme constitue la meilleure alternative pour l'étude de ces protéines et va consister à cibler l'identification de séquence peptidiques spécifiques. Cette stratégie nécessite de viser le séquençage de la totalité des peptides issus de ces protéines et d'établir les conditions expérimentales adéquates.

L'**objectif de ce travail de thèse** est d'établir un état de l'art des connaissances de la structure et d'en apporter de nouvelles en développant des stratégies analytiques adaptées à la problématique particulière de caractérisation des protéines du cheveu. Ces stratégies seront basées sur l'utilisation des techniques de l'analyse protéomique. Ce travail a pour perspective de trouver de nouveaux éléments permettant de comprendre la structure de la fibre capillaire.

Le manuscrit décrivant ce travail s'articule autour de quatre parties.

La **première partie** est une étude bibliographique visant à regrouper l'ensemble des connaissances actuelles de la biologie de la fibre capillaire et l'ensemble des expériences et des résultats déjà obtenus sur la structure avec différentes stratégies analytiques. Nous décrirons les connaissances actuelles de la biologie du cheveu de l'échelle macroscopique à l'échelle moléculaire.

Cette partie permet de conclure sur les questions restant posées sur cet édifice biologique et sur la nécessité d'envisager les besoins analytiques nécessaires pour y répondre.

La **seconde partie** est introduite par un exposé des principes de la protéomique. La construction d'une stratégie expérimentale adaptée à la problématique singulière des protéines du cheveu est par la suite présentée. Nous évaluerons la pertinence de différentes techniques de séparation utilisées pour la décomplexification des protéomes puis expliciterons les choix effectués pour le génération et le traitement des futures données.

Dans une optique d'amélioration de l'obtention de données par analyse protéomique, une attention toute particulière sera accordée à l'optimisation des couplages nano-LC-MS et MS/MS utilisant des spectromètres Q-TOF.

Un chapitre est consacré à l'utilisation originale d'un plan d'expérience afin d'identifier et de quantifier les effets des principaux paramètres instrumentaux sur le séquençage des peptides par spectrométrie de masse. Parallèlement, différentes études d'optimisation montreront l'impact des paramètres chromatographiques sur les résultats et les bénéfices apportés par l'introduction de nouvelles fonctionnalités dont bénéficie la dernière génération de spectromètres de masse Q-TOF pour l'analyse protéomique.

La **troisième partie** décrit l'apport de la stratégie protéomique développée pour la caractérisation des protéines du cheveu.

Dans un premier chapitre, nous présenterons les résultats obtenus pour l'analyse du protéome du cortex du cheveu humain. Nous décrirons la caractérisation des kératines et des KAP spécifiques à ce type cellulaire et démontrons les expressions protéiques de gènes issus de famille multigéniques qui n'avaient auparavant pas été mise en évidence. A ces résultats de caractérisation, s'ajouteront des données semi quantitatives des protéines identifiées nécessaires à la compréhension des rôles de ces dernières dans les structures cellulaires étudiées.

Dans un second chapitre, nous présenterons l'étude protéomique réalisée sur les cellules cuticulaires du cheveu. Nous décrirons le développement d'une méthode d'analyse destinée à l'étude d'une modification particulière induite par l'activité d'une enzyme, la transglutaminase. Les analyses du protéome purifié de cellules cuticulaires permettront d'apporter l'évidence de l'expression de gènes dans le protéome humain dont certains étaient absents des banques protéiques utilisées. L'étude des recouvrements de séquences obtenus pour certaines familles de protéines suggère la localisation de sites substrats de la transglutaminase.

La **quatrième partie** présente les études réalisées suite aux résultats de caractérisation obtenus par analyse protéomique du cortex. Ces études ont pour but d'apporter des éléments de compréhension du rôle et de l'origine des KAP dans le cheveu.

Nous présenterons une méthode originale de quantification utilisant la combinaison des informations protéomiques et de compositions en acides aminés.

Nous montrerons comment les séquences en acides aminés des KAP peuvent être étudiées par comparaison à celles d'autres mammifères pour mettre en évidence des motifs consensuels dans les séquences primaires de ces protéines. La modélisation de la structure de certains de ces motifs montrera des propriétés d'arrangement spatiaux pouvant expliquer leur rôle dans la structure moléculaire du cheveu. Elles permettront d'envisager des mécanismes moléculaires de cornification.

Nous terminerons sur une discussion de l'origine des gènes des KAP qui soulève des questions sur l'histoire de l'évolution des mammifères.

Partie I Introduction à la biologie et à l'état de l'art des connaissances de la fibre capillaire

Pour comprendre les propriétés structurales du cheveu, il parait essentiel de connaître le résultat de sa morphogénèse et donc l'arrangement des structures finales visualisées dans la fibre mature. Par ailleurs, il convient de considérer cette structure comme un ensemble de composants moléculaires dont il faut connaître la nature, les interactions et la localisation de leur expression temporelle et spatiale.

Dans ce but, nous présenterons cette structure du point de vue de son organisation cellulaire telle qu'elle peut apparaitre à la lumière des observations microscopiques aux différentes étapes de la croissance du cheveu.

Par la suite, nous présenterons la structure comme un ensemble de molécules, essentiellement des protéines, dont l'expression et l'organisation sont définies par les mécanismes de différenciation cellulaires et gouvernées par l'expression de gènes.

Nous montrerons que la compréhension des liens existant entre l'expression des gènes et la formation d'un édifice moléculaire aussi complexe que le cheveu n'est que partielle et qu'il existe principalement des questions relatives à l'expression de familles particulières de protéines, les protéines associées aux kératines.

Nous conclurons cette partie en présentant les champs de recherches qui restent à explorer aux vues des connaissances actuelles et des développements instrumentaux et particulièrement des stratégies protéomiques.

Chapitre I Avant propos

1. Des cheveux et des hommes

L'histoire illustre la place occupée de tout temps et parmi toutes les nations au soin des cheveux. Les représentations à travers les siècles montrent les multitudes de façon que l'homme a imaginé pour apprêter ses cheveux comme un ornement permettant d'affirmer son appartenance à une communauté ou à un statut social. La possibilité de modifier la chevelure en jouant sur sa longueur, son organisation, sa frisure, de la nouer, la colorer, de l'accompagner d'une coiffe, de la cacher totalement ou partiellement, explique sans aucun doute pourquoi cette partie du corps humain a été travaillée sous autant de déclinaisons. Les coiffures sont ainsi aussi nombreuses que la diversité des peuples, des courants d'idées et des périodes que l'humanité pu traverser.

L'humain a toujours accordé une grande importance au soin et à la conservation de sa chevelure, la longueur et la densité étant souvent symbole de la force et de la vigueur de son propriétaire. Cette symbolique se retrouve bien illustrée dans l'épisode biblique de Samson et Dalila, la première enlevant la force de ce premier en lui coupant les cheveux. La scalpation décrite par Hérodote chez les guerriers Scythes vers le Vème siècle avant JC ou pratiquée par certaines tribus amérindiennes avant l'arrivée des européens et jusqu'au 19ème siècle était censée montrer la vaillance du guerrier vainqueur dans le premier cas et symboliser la prise de l'âme du guerrier vaincu dans le second. Symbole de pouvoir sous le règne des rois Francs, les cheveux longs étaient réservés aux plus hauts dignitaires de la noblesse et un prince destitué se voyait renvoyé au rôle de simple sujet par la coupe de ses cheveux (Grégoire de Tours, VIème siècle). Cette symbolique a été conservée chez les souverains jusqu'à ce que François 1er, suite à une blessure, soit contraint de se raser la tête. Ses courtisans, par respect pour leur roi, firent alors de même et les cheveux courts furent conservés jusqu'à ce que Louis XIII réintroduise la mode des cheveux longs et bouclés par la suite remplacés par les perruques parfois exubérantes.

Inversement, le crâne rasé a été associé à la pauvreté ou à la disgrâce, la tonte des cheveux pouvant même être la punition d'un crime, ou une marque infamante pour les esclaves. La tonsure pratiquée par les moines des Eglises chrétiennes dès le VIème siècle et communément jusqu'au XXème siècle symbolisait la renonciation au monde et l'effacement des péchés antérieurs. La tonte se retrouve également chez les moines d'autres religions comme le Bouddhisme, est un acte de foi dans l'Hindouisme et le dernier rituel accompli par les Musulmans lors du pèlerinage à La Mecque. Dans ces contextes, l'alopécie ou calvitie a longtemps été considérée comme dévalorisante. Jules César eu la permission de porter sa couronne de laurier de manière permanente pour la masquer (Suétone, IIème siècle) mais d'autres couvres chef ont été employés dans ce but comme les perruques portées dès l'Antiquité. Aujourd'hui dans notre société, la calvitie peut néanmoins être assumée voire même susciter des modes.

La beauté associée à la chevelure s'avère relativement subjective, la longueur et la densité de la chevelure féminine semblent avoir été de tout temps appréciées parmi les populations ne possédant pas de cheveux crépus. L'ondulation et la frisure des cheveux sont en revanche moins consensuelles aux vues des représentations des ''Vénus'' dans l'histoire des différents peuples. Toujours est-il que la chevelure a toujours joué et joue une place importante dans les mécanismes de séduction.
Aujourd'hui, les modes continuent d'évoluer en s'appuyant sur des techniques de traitements issues des développements de l'industrie cosmétique. La croissance de cette industrie s'appuie sur le besoin de consommateurs en produits et en soins à la personne. L'utilisation et le développement des connaissances scientifiques pour la compréhension et la mise au point de techniques et produits cosmétiques motivent encore aujourd'hui des efforts de recherche. Dans ce contexte, la science du cheveu a étroitement suivi les

développements des techniques analytiques pour améliorer les connaissances. La finalité de la compréhension des propriétés de cette structure réside dans la recherche de nouveaux produits permettant de la modifier originalement sans la dégrader et ainsi de compléter la palette des outils imaginés par l'homme pour continuer à faire évoluer sa chevelure.

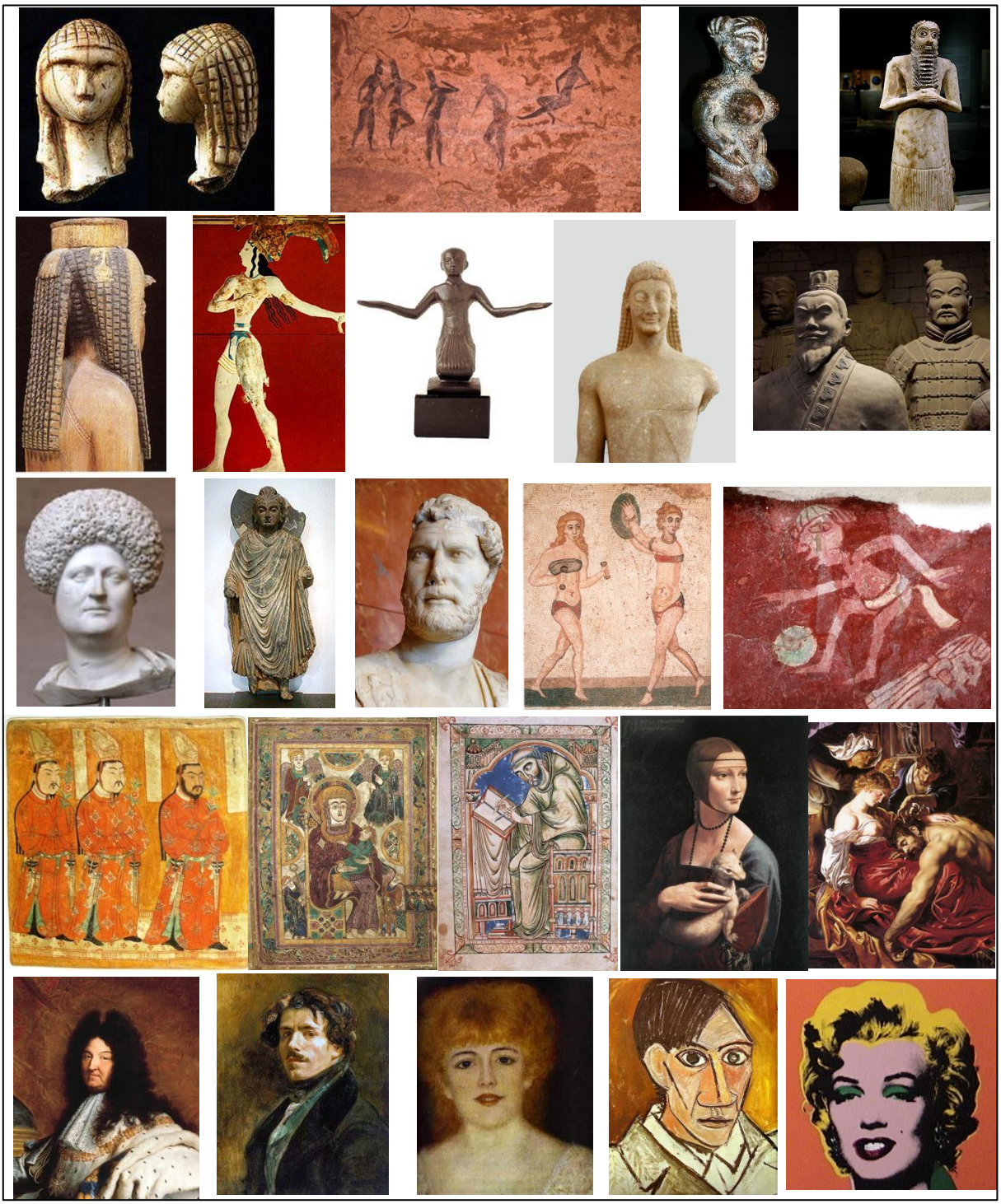


Figure 1 *: Exemples de représentations de la chevelure à travers l'art de -25000 ans av JC jusqu'au 20ème siècle.*

2. Universalité de la structure chez les mammifères

Retrouvées communément chez les **euthériens,** les **marsupiaux** et les **monotrèmes**, les trois lignées de mammifères existant encore aujourd'hui, la fibre capillaire avait donc déjà été acquise par leur ancêtre commun

vivant il y a plus de 170 millions d'années [1]. La découverte d'impressions de poil sur des fossiles datant du milieu du Jurassique [2] confirme cette acquisition considérée **antérieure à 200 millions d'années** et à l'extinction massive du Trias-Jurassique [3]. D'autres observations de fibres fossilisées datant du début du Crétacé (environ 100 millions d'années) montrent que la surface des fibres est similaire à celles retrouvées aujourd'hui [4]. La combinaison de la lactation, de l'homéothermie et de la fourrure a probablement joué un rôle crucial dans la **survie des mammifères** au cours des périodes d'extinctions d'espèces qu'ils ont pu traverser.

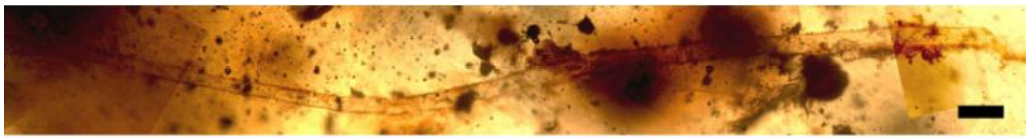


***Figure 2** : Poil de mammifère piégé dans de l'ambre du Crétacé (100 millions d'années) [4].*

Les poils des mammifères revêtent différentes fonctions qui dépendent de leur longueur, de leur forme, de leur structure et de leur implantation.

Un grand nombre d'espèces possèdent une **fourrure** composée de la combinaison de deux sortes de poils : la **bourre** constituée de poils fins souples et serrés servant à l'**isolation thermique** et la **jarre** constituée de poils plus longs et plus gros destinés à recouvrir la bourre et à l'**isoler de l'eau**. Cette fourrure peut être modifiée par mue pour s'adapter à la variation saisonnière des températures.

Pour la plupart des espèces, la **couleur du pelage** s'est adaptée à leur environnement et assure la fonction de **camouflage**. Le pelage peut également définir le caractère sexué au sein d'une espèce.

Toujours dans une optique d'adaptation environnementale, les mammifères aquatiques possèdent un poil court adapté à la nage (c*arnivores* tels que phoques, otaries, morses, éléphants de mer), voire une disparition quasi-totale à totale de pilosité (*cétartiodactyles* tels que les cétacés et les hippopotames et *siréniens* tels que les lamantins ou les dugongs). Cette faible pilosité permet également de faire face à des températures chaudes pour certains mammifères terrestres (rhinocéros, éléphants).

Les poils peuvent avoir également d'autres fonctions comme les **cils** permettant d'éviter la déposition de poussières dans les yeux, les **vibrisses** servant de détecteur sensoriels et les **épines** permettant de dissuader la prédation.

Chez l'humain, la faible pilosité corporelle est une caractéristique associée à la régulation de la température corporelle grâce à la sudation. Ces caractéristiques lui permettent une endurance unique parmi les mammifères même dans des conditions de températures chaudes. Avec l'acquisition de la bipédie, elles ont sans aucun doute joué un rôle crucial dans l'évolution des homininés en Afrique de l'Est il y a plusieurs centaines de milliers d'années et pourrait être la conséquence d'une adaptation de l'espèce au mode de vie de chasseur cueilleur [5, 6].

Par la suite, Homo Sapiens a pu utiliser la fourrure des autres mammifères pour évoluer dans les milieux plus froids de la planète.

La conservation de la chevelure parmi l'espèce reste mal appréhendée mais les hypothèses d'un rôle de protection contre l'insolation et d'ornement ayant favorisé la reproduction peuvent être imaginées.

3. Histoire de la science du cheveu

L'étude de la fibre capillaire a depuis longtemps été un sujet d'intérêt pour les chercheurs. Les pathologies associées sont décrites dès le 19ème siècle par les médecins [7-10] et vers la fin du 19ème siècle, l'extraction de la kératine, substance protéique issues des cornes et des cheveux, est décrite dans la littérature. A cette époque, son utilisation en tant que matériau a conduit à des applications allant de substance pour l'enrobage de médicament [11] à la composition de filaments pour les lampes à incandescence [12]. Parallèlement, l'utilisation de la microscopie a très tôt permis d'apporter les premières connaissances de la structure du follicule et de la tige

capillaire compartimentée en cortex, médulla et cuticule [13, 14]. L'isolement du pigment responsable de sa couleur, la mélanine, et l'observation de ses déclinaisons sont réalisées vers la même période [15, 16].
Par la suite, l'étude de cheveu a évolué de concert avec le développement des stratégies et des techniques analytiques. Nous pouvons noter que cette évolution s'est toujours faite en parallèle des travaux réalisés sur les autres fibres animales et tout particulièrement sur la laine dont l'utilisation comme matière première pour l'industrie textile a motivé un très grand nombre d'études.

L'analyse physico-chimique des hydrolysats de cheveux permet vers la fin des années 20 de connaitre l'abondance des différents acides aminés basiques et d'y montrer la présence élevée de cystéine [17-19]. Par la suite, d'autres acides aminés inhabituels comme la lanthionine [20], la citrulline [21] et l'ε-(γ-glutamyl)lysine [22] seront identifiés dans la fibre.

Le cheveu est l'une des premières structures biologiques à être étudiée par diffraction des rayons X [23-25]. Le développement de méthodes fines d'obtention de sections de cheveux [26] permet aux scientifiques d'observer l'hétérogénéité des structures de cheveu parmi les populations [27-31] et de suggérer leur utilisation comme indices pour les études criminalistiques [32] et archéologiques [33]. Les travaux de Pauling et Corey sur l'interprétation de données cristallographiques notamment de l'α-kératine du cheveu les conduisent à suggérer un modèle d'arrangement de la chaine peptidique en hélice-α passé depuis à la postérité [34, 35].

Dès 1957, l'utilisation de la microscopie électronique développée dans les années 30 permet de faire un pas de géant dans la compréhension de la structure du cheveu. Avec cette technique, les structures du follicule pileux aux différentes étapes de leur croissance sont visualisées [36-38]. L'observation du cortex révèle une régularité d'arrangement structuraux, les microfibrilles, préalablement suggérées par les études de diffractions X [39-42]. Les microfibrilles, décrites alors comme unités de base structurales, sont enrobées d'une matrice non structurée vraisemblablement riche en soufre [43, 44]. Ces analyses de microscopie électronique sont associées à des études d'auto radiographies sur le follicule dans lequel est incorporé in vivo des acides aminés radioactifs [44-47]. Dans les années 60, l'analyse d'acides aminés peut désormais être réalisée par chromatographie liquide [48]. L'utilisation d'enzymes pour réaliser des hydrolysats est citée [49-52] et permet de montrer dans la structure la présence d'acides aminés modifiés [22, 53-55]. Des techniques de précipitation adaptées des protéines permettent de diviser les protéines des extraits en différentes fractions qui peuvent alors être séparées par électrophorèse sur gel pour estimer leurs masses moléculaires [56-61]. Ces techniques montrent que les extraits de cheveu sont composés de kératines et de protéines plus petites, les protéines associées aux kératines [60-62]. L'utilisation de techniques électrophorétiques bidimensionnelles sur les extraits de cheveu et d'ongles montrent des similarités de composition et des hétérogénéités au sein des protéines associées aux kératines [59, 63-66]. La recherche des gènes correspondants à ces protéines est réalisée sur la laine grâce à l'obtention de séquence chez le mouton [63, 67-70]. Des études sont menées dans les années 90 afin d'identifier les séquences correspondantes chez l'humain [71, 72], mais il faut attendre le début des années 2000 pour que la catalogue des gènes correspondants aux protéines pouvant être exprimées dans le cheveu soit extensivement complété [3, 73-83]. L'expression de ces gènes est en partie démontrée dans la structure et les recherches actuelles portent principalement sur la compréhension de la localisation et de l'arrangement des protéines les unes par rapport aux autres [84-88]. Les techniques de protéomiques qui ont émergées au cours de la dernière décennie n'ont été à l'heure actuelle été que peu exploitées sur le cheveu [89] et les principaux travaux ont été beaucoup plus importants pour l'étude de la laine [90-98].

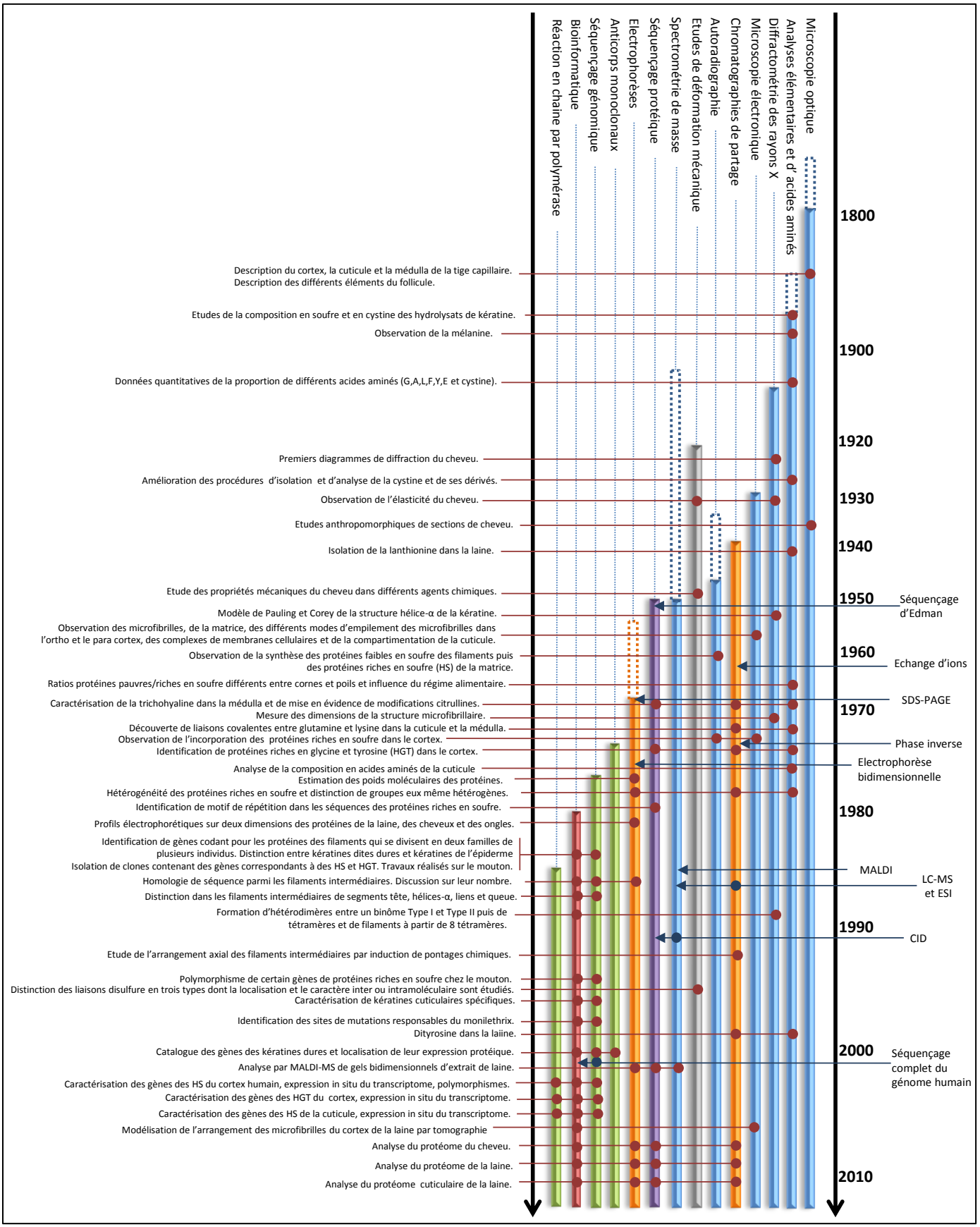


Figure 3 *: L'évolution des connaissances du cheveu en parallèle des inventions et développements des techniques analytiques au cours des deux derniers siècles.*

Chapitre II Biologie du cheveu

Dans ce chapitre, nous décrivons un panorama des informations tirées de plus d'un siècle et demi d'observations du follicule pileux au niveau cellulaire et subcellulaire.

1. Morphogénèse

a) Le follicule

La tige capillaire nait du **follicule pileux**, une structure qui se développe dans la peau au sein d'une **gaine de tissu connectif** lors de l'embryogénèse puis qui suit un cycle de renouvellement alternant phase de régression (**catagène**), d'expulsion (**télogène**) et de croissance (**anagène**) tout au long de la vie de l'individu. Ce cycle est maintenu grâce à deux réserves de cellules souches situées dans des niches du follicule disposée au dessus de la partie renouvelée tous les 3 à 5 ans.

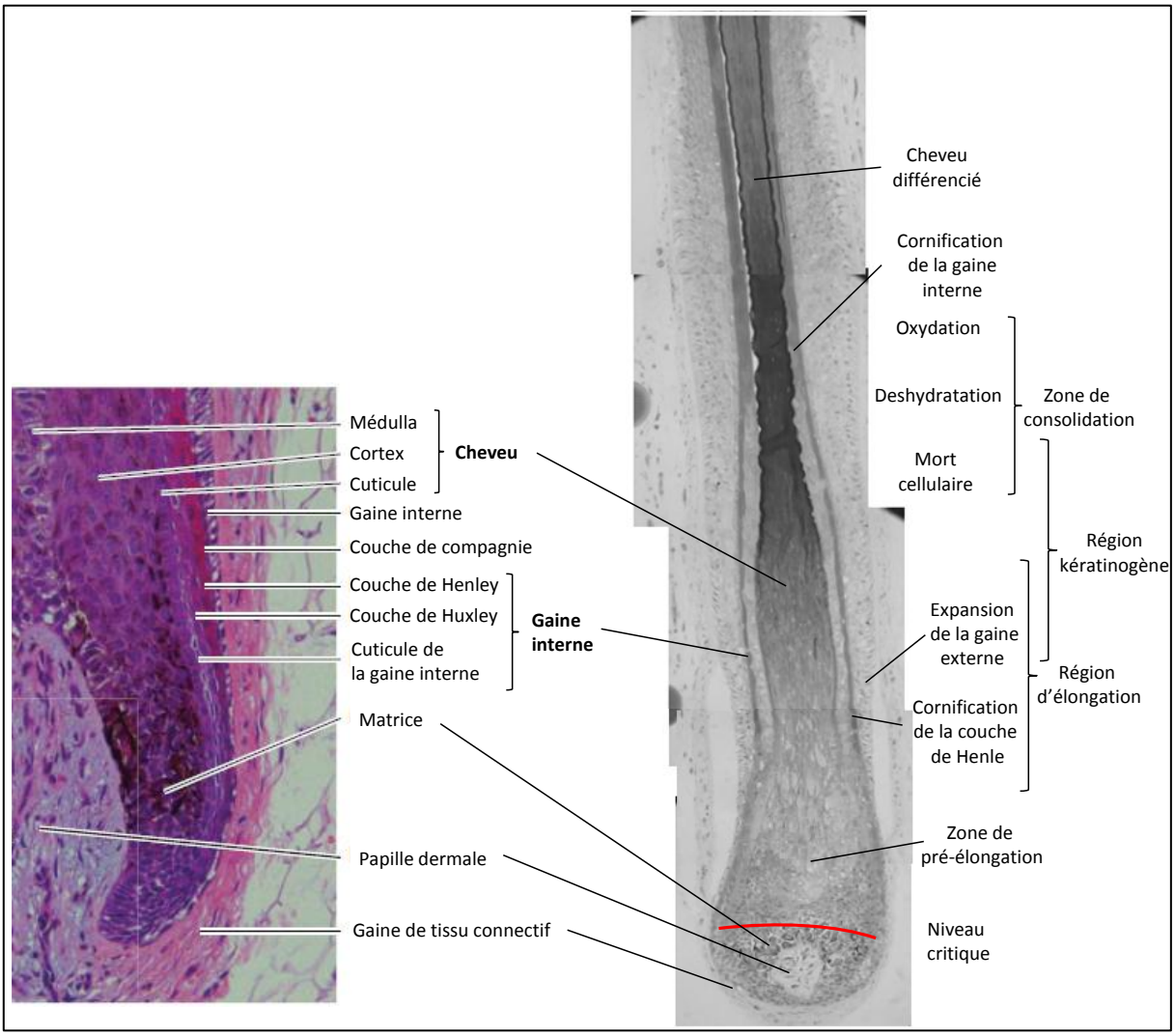


***Figure 1** : Coupes de follicule pileux. A gauche, détail d'une observation en microscopie, d'après [99], de la base d'un follicule d'une fibre médullée et dont la coloration permet de distinguer clairement les différents types cellulaires en cours de différenciation. A droite, vue en microscopie électronique globale d'un follicule sans médulla et des différentes zones de différenciation avant la différenciation définitive de la tige capillaire. Données internes.*

La complexité structurale du follicule s'illustre par sa décomposition en 8 types cellulaires répartis en couches distinctes. Cette diversité cellulaire fait du follicule une des structures les plus complexes du corps humain (Figure 1).

Leur organisation est définie au niveau du bulbe, dans la **matrice**, où les cellules se divisent et se différencient autour de la papille dermique. Ce sont trois types cellulaires qui vont donner naissance au cheveu : de multiples **cellules corticales** s'entourent de **cellules cuticulaires** et dans certains cas, des **cellules médullaires** se retrouvent insérées au centre du follicule au contact de la papille dermique.

A la périphérie de ces cellules du cheveu se retrouvent les cellules qui vont constituer la gaine racinaire avec successivement les **cellules cuticulaires de la gaine interne**, celles des **couches de Huxley** puis de **Henle**, la **couche de compagnie** et la **gaine externe** [99]. En plus de ces cellules, des **mélanocytes** présents à l'interface avec la papille dermique synthétisent des pigments de mélanines qui se retrouvent insérés au sein des cellules corticales [100].

b) Les différentes étapes de différenciation folliculaire

La prolifération cellulaire au niveau du bulbe entraine la migration permanente des cellules formées vers le haut du follicule et est à l'origine de la pousse du cheveu à une vitesse supérieure à 10 µm/heure. A partir d'une certaine zone dit **niveau critique** ou **ligne d'Auber**, les cellules, à l'exception de la gaine externe [3], interrompent le processus de division et commencent à ce différencier.

Au cours de la différenciation, les différents types de cellules vont adopter des formes spécifiques à leur fonction et exprimer des protéines au cours de plusieurs étapes. Ces modifications permettent de distinguer différentes zones de différenciation au niveau du follicule (Figure 1).

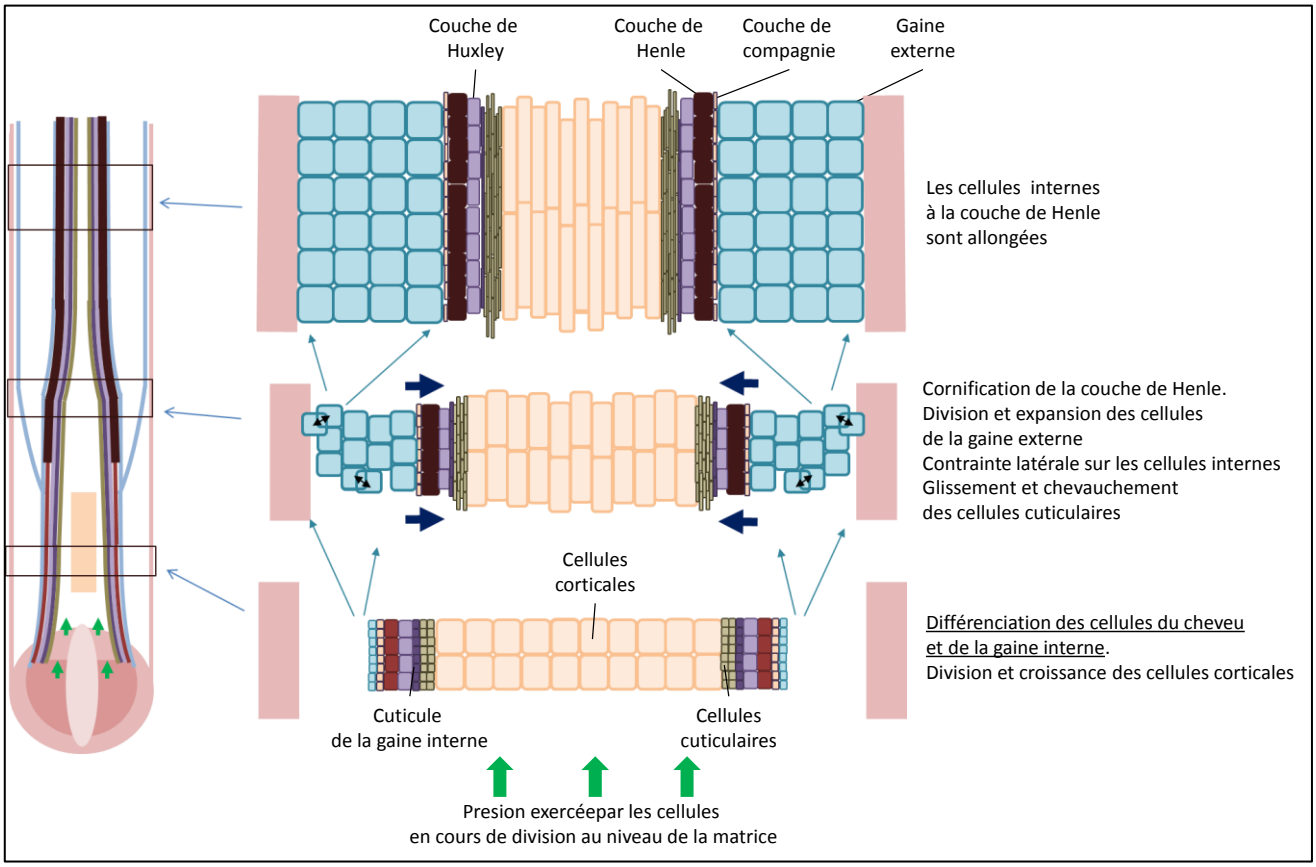


***Figure 2** : Schématisation d'une proposition de mécanisme de différenciation des cellules du follicule pour expliquer la physiologie observée après la zone d'élongation.*

Au-delà du niveau critique s'observe la zone de **pré-élongation**. Les cellules adjacentes adhèrent les unes aux autres grâce à des complexes protéiques permettant de conserver un réseau cohésif. Les cellules corticales augmentent de volume et la distinction des différentes couches cellulaires y est clairement observée.
La **région d'élongation** se distingue alors sur une zone allant de 100 à 250 µm par rapport à la base du bulbe, au cours de laquelle les cellules corticales, cuticulaires et cuticulaires de la gaine interne vont s'allonger dans le sens de la pousse et adopter leur forme définitive. L'élongation de ces cellules coïncide avec la cornification des cellules de la couche de Henle au niveau du rétrécissement de la section du bulbe. Dans cette zone, les cellules de la gaine externe se divisent puis se différencient. La combinaison de ces événements semble exercer une pression latérale sur les cellules centrales suffisante pour entrainer leur déformation. Les différences dans cette région de taille et de composition cytoplasmique entre les cellules corticales, cuticulaires et cuticulaires de la gaine interne peuvent expliquer les différences de compression et donc d'élongation observée entre les cellules corticales et les deux autres types cellulaires. Les cellules cuticulaires se retrouvent proportionnellement plus allongées que les cellules corticales (Figure 2).
La synthèse de protéines, notamment des kératines, est réalisée et se poursuit jusqu'à une distance d'environ 500 µm par rapport à la base du follicule. C'est la **région kératogène** ou **zone de kératinisation** dont la fin marque la mort de la cellule.
Au-delà, il est possible de distinguer une **zone de consolidation** longue de plusieurs centaines de microns durant laquelle les protéines synthétisées dans les cellules du cheveu sont stabilisées en un édifice rigide peu à peu déshydraté.
Au dessus de cette zone, les cellules du cheveu sont définitivement différenciées. Les cellules des couches de la racine subissent alors un mécanisme de cornification puis de dégradation, ce qui libère le cheveu. Dans la dernière zone, la fibre est libre de racine et en contact avec les sécrétions des glandes sébacées qui lubrifient la structure qui est finalement exposée à la surface de la peau et à l'environnement extérieur.

2. Structure du cheveu humain

a) Structure générale de la section du cheveu

Le cheveu est une fibre dont le diamètre de section varie entre 50 et 100 µm. La géométrie de cette section peut également varier comme nous le décrirons à la suite de ce chapitre.
Les trois types cellulaires observés au niveau du follicule se retrouvent dans la structure.
Les **cellules corticales** différentiées représentent la majorité de la structure. Suite à leur élongation et à leur kératinisation leur dimension est d'environ 100 µm sur 5 µm. La pigmentation associée aux granules de mélanine se retrouve en périphérie de la zone corticale entre les membranes des cellules corticales.

Lorsqu'elle existe, la **médulla** se retrouve sur une section de 10 à 20 µm au centre de la fibre. Elle est la superposition en colonne désorganisée de cellules de dimensions d'environ 15 µm sur 5 µm aplaties perpendiculairement à l'axe de la fibre [101, 102].

A la périphérie de la fibre, la **cuticule** est la superposition en tuiles de plusieurs couches (typiquement de 6 à 9) de cellules cuticulaires rectangulaires aplaties d'environ 40 µm de côté et d'environ 500 nm d'épaisseur. La cuticule permet d'isoler la fibre vis-à-vis de l'abrasion et de maintenir l'intégrité du cortex. Si l'eau ne semble pas vraiment pénétrer les cellules cuticulaires, la cuticule n'est cependant pas une barrière étanche. En effet, l'espace inter membranaire forme un réseau au sein duquel l'eau peut s'infiltrer jusque dans le cortex [103].

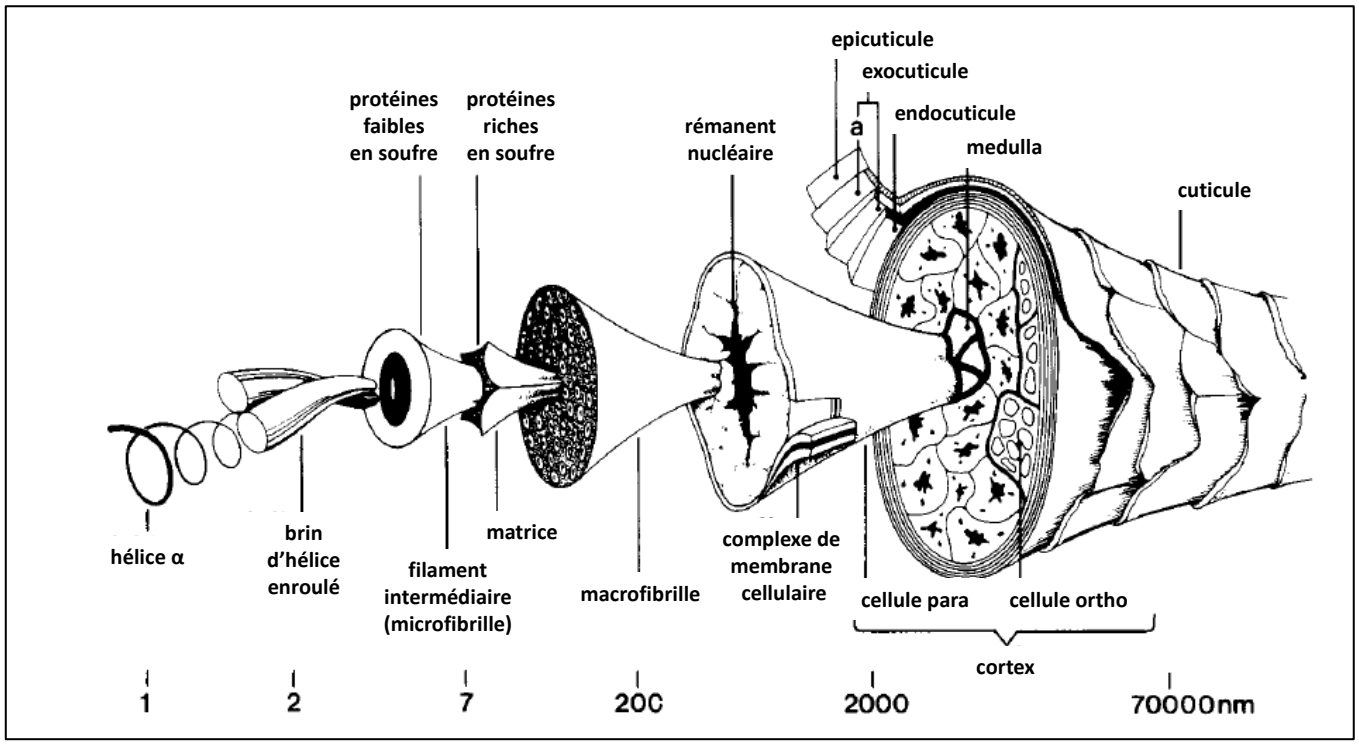


Figure 3 *: Représentation schématique des composants principaux du cheveu humain. Adaptée de [104].*

b) Ultrastructure du cortex

Les cellules corticales sont constituées des rémanents cellulaires insérés aléatoirement au sein de structures caractéristiques, les **macrofibrilles**, qui consistent en d'imposants fuseaux remplissant en quasi-totalité l'espace cellulaire.

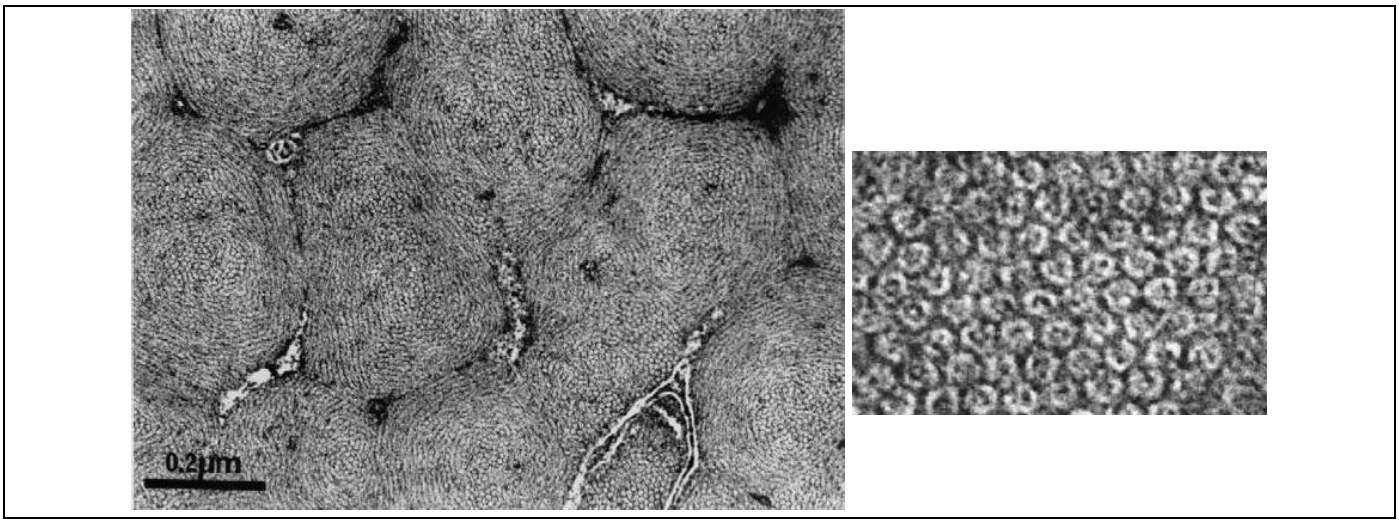


Figure 4 *: A gauche, visualisation en microscopie électronique des structures macrofibrillaires dans le cheveu. Nous noterons l'insertion entre les macrofibrilles de zones probablement constituées de rémanents cellulaires [105]. A droite, détail des microfibrilles insérées dans la matrice interfilamentaire [104].*

Les macrofibrilles peuvent être considérées comme le matériau le plus abondant au sein de la fibre. Ces fuseaux, au nombre d'une dizaine dans la cellule, se disposent parallèlement à l'axe longitudinal de la pousse. Néanmoins, elles ne sont pas l'unité structurale de base du cheveu et sont elles-mêmes composées de plusieurs centaines de sous éléments, les **microfibrilles**, observables en microscopie électronique à transmission. Les microfibrilles ou

filaments intermédiaires consistent en un empilement latéral régulier d'un assemblage de **filaments de kératine** enrobés dans une **matrice interfilamentaire** riche en soufre.

L'empilement quasi cristallin des filaments est responsable de l'activité optique de la fibre et de sa propriété de diffraction des rayons X. Cette dernière a été largement utilisée pour déterminer l'organisation des filaments au sein des microfibrilles et de mesurer la taille des filaments (7,4 nm), l'espace inter filaments (2 nm) et donc les fractions volumiques entre filaments et matrice (de l'ordre de 50%). Différents modes d'empilement distincts des microfibrilles peuvent être observés, ce qui permet de définir deux types de cellules corticales [106] :

- Les cellules contenant des macrofibrilles **paracorticales**, très majoritaires. L'empilement des microfibrilles du paracortex est proche d'un mode hexagonal compact. Les fuseaux des macrofibrilles sont fusionnés et pratiquement indistincts.
- Les cellules contenant des macrofibrilles **orthocorticales**. Les macrofibrilles sont distinctes et séparées par une matrice intermacrofibrillaire comportant vraisemblablement des rémanents cellulaires. Les macrofibrilles se torsadent le long de l'axe longitudinal. L'empilement des microfibrilles décrit un enroulement périphérique à l'axe central de la macrofibrille.

La répartition des cellules paracorticales et orthocorticales peut être à l'origine de la segmentation de la fibre en zones distinctes.

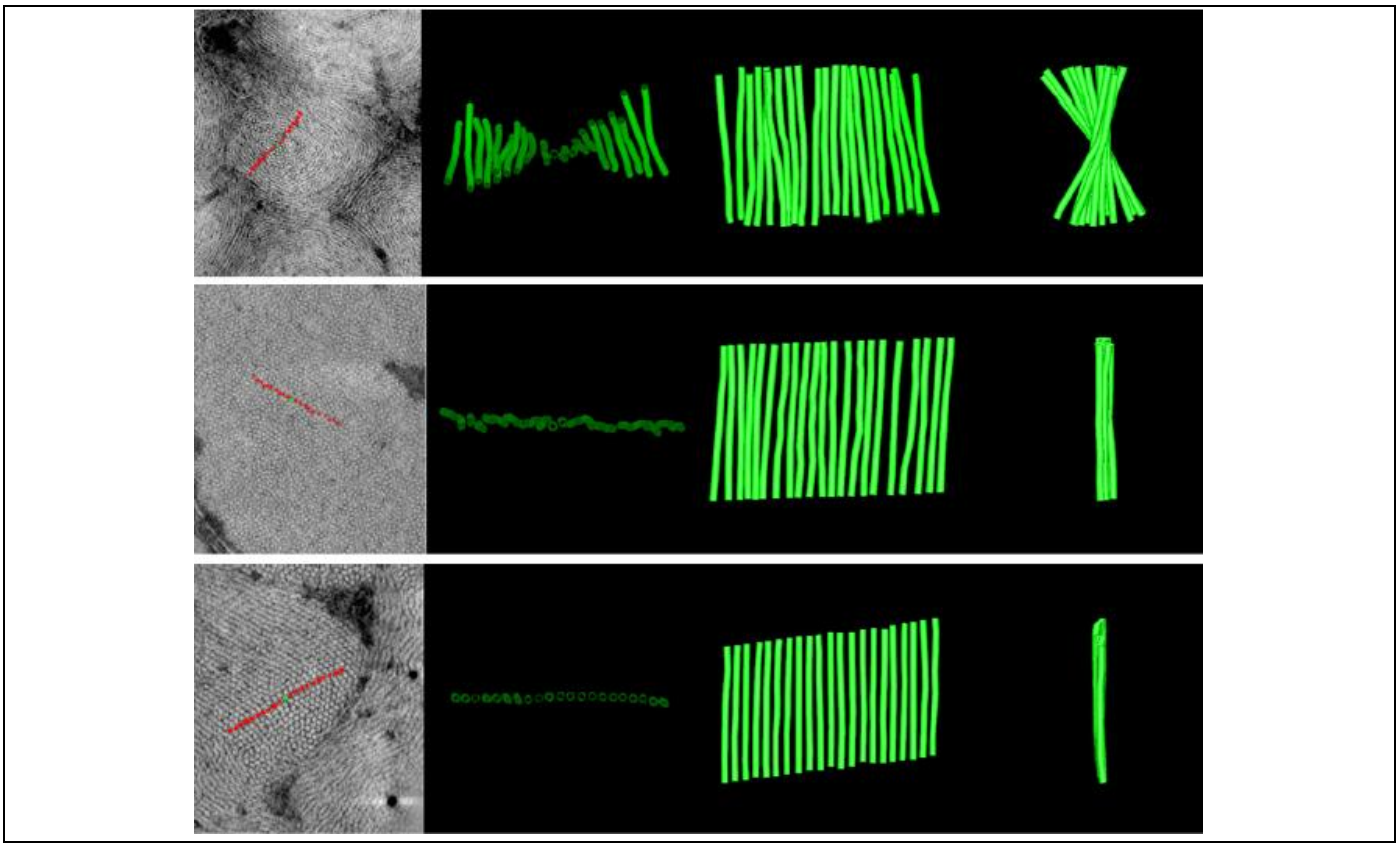


***Figure 5** : Représentation de l'empilement des microfibrilles et examens tomographiques de leur arrangement dans (de haut en bas) l'orthocortex, le méso cortex et le paracortex des structures macrofibrillaires du cortex de la laine. D'après [107].*

Certains auteurs décrivent également des macrofibrilles **mésocorticales**, intermédiaires entre les deux précédents types cellulaires. Ces structures correspondent à un mode d'empilement des microfibrilles plus strictement hexagonal compact et dont les macrofibrilles sont moins fusionnées que dans le paracortex.

Des études tomographiques des sections des différents types de cellules dans la laine ont permis d'étudier l'arrangement des microfibrilles au sein de chacune d'entre elles [107]. Dans l'orthocortex, chaque microfibrille adopte un angle avec la microfibrille adjacente située dans la direction décrite du centre vers la périphérie de la

macrofibrille. Cette disposition traduit un arrangement global torsadé des microfibrilles, très certainement en lien avec l'architecture spécifique des macrofibrilles correspondantes. Dans le paracortex et le mésocortex, les microfibrilles adoptent peu ou pas d'angle les unes par rapport aux autres.
La différence entre les types de macrofibrilles se retrouve également dans les variations de fraction volumique occupée par la matrice interfilamentaire par rapport aux microfibrilles. La matrice est en effet plus abondante dans le paracortex (de 50 à 60%) que dans l'orthocortex (environ 40%).
Nous dédierons une description des différentes étapes conduisant à la formation de ces édifices dans le chapitre suivant consacré à la biologie moléculaire du cheveu.

c) Ultrastructure de la cuticule

La cellule cuticulaire différenciée est bipolarisée et se décompose en compartiments distincts disposés en couches parallèles à l'axe d'empilement des cellules :

- Plus de la moitié de l'espace cellulaire est occupée au centre par l'**exocuticule** dont la structure relativement homogène est constituée de protéines amorphes très riches en soufre liées les unes aux autres par un dense réseau de liaisons covalentes entre les chaines latérales.
- Du côté de la membrane interne de la cellule, se trouve l'**endocuticule**. Elle est constituée de la fusion de matériel exocuticulaire avec des rémanents du noyau et des organites dégradés avant la kératinisation.
- Du côté de la membrane externe, s'observe la **couche A**, épaisse d'environ 60 nm. Comme l'exocuticule, elle est homogène mais se distingue de cette dernière par une plus forte proportion en soufre que l'exocuticule. Sa formation est le résultat de l'agrégation de granules très denses pendant la phase kératogène [108] et données internes L'Oréal].
- L'**épicuticule** constitue une couche entre la membrane externe et la couche A de 13 nm d'épaisseur, constituée de protéines riches en soufre qui sont liées à la surface externe avec une couche d'acides gras saturés. Cette couche peut être extraite sous formes de sacs générés par réaction avec de l'eau chlorée (réaction de Allwörden). Cette couche se prolonge sur toute la surface de la membrane et se nomme couche interne lorsqu'elle fait face à l'intérieur de la fibre [109, 110].

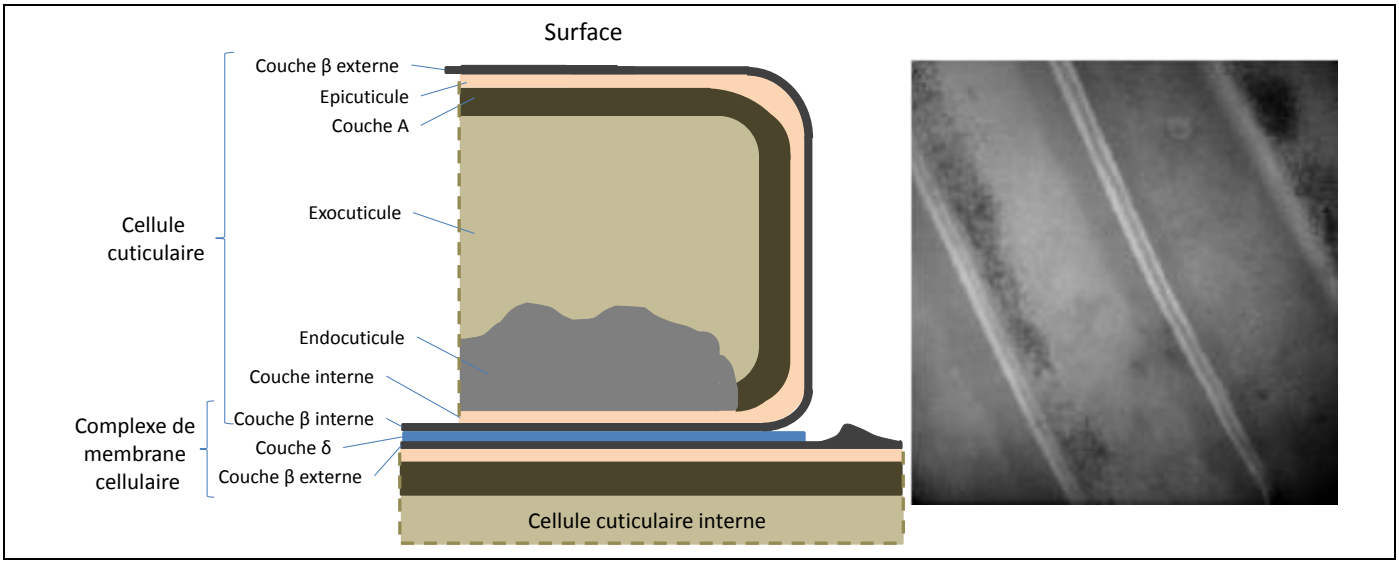


***Figure 6** : Schématisation de l'organisation en couche des cellules cuticulaires. Vue en microscopie électronique à transmission d'une coupe cuticulaire.*

- L'interface de la cellule cuticulaire avec les cellules adjacentes est constituée par le **complexe de membrane cellulaire** (CMC). Le CMC est l'association de la membrane cellulaire, appelée **couche β**, avec l'autre couche β de la cellule adjacente. Les deux membranes lipidiques, de 3,5 µm d'épaisseur chacune

sont séparées par une couche inter membranaire, la **couche δ**, épaisse de 18 nm [103, 110, 111]. Lorsque la fibre est exposée à l'eau, le gonflement de la couche δ de quelques nanomètres témoigne du passage que prend le liquide pour pénétrer dans la fibre [103].

d) Ultrastructure de la médulla

La médulla ou moelle est abondante dans les poils de barbe et aléatoirement présente dans les cheveux humains. Les cellules médullaires sont difficilement discernables les unes des autres et contiennent des structures globulaires insérées dans des structures fibreuses désorientées [102]. Ces structures globulaires sont creuses et résultent de la formation, lors de la différenciation cellulaire, de granules protéiques de trichohyaline d'environ 3 µm de diamètre. Les structures fibreuses proviennent d'un empilement de microfibrilles beaucoup plus compact et désorganisé que ce qui peut être observé dans le cortex. Des cellules corticales peuvent par ailleurs se retrouver insérées au sein des cellules médullaires ou partiellement segmenter la médulla [101].

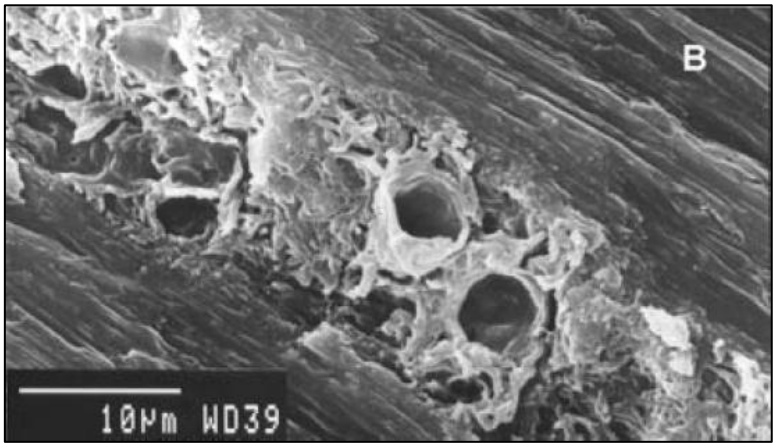


***Figure 7** : Cliché de microscopie électronique d'une coupe longitudinale d'une colonne de médulla insérée dans la structure fibreuse du cortex. D'après [102].*

3. Les origines du polymorphisme des fibres capillaires de l'espèce humaine

L'aspect de la chevelure est extrêmement variable parmi les populations humaines et est étroitement lié à la **forme des fibres**, à la **répartition des types cellulaires**, à la **densité des follicules** à la surface de la peau et bien entendu à la **couleur des fibres**.

A cette variabilité de la chevelure s'ajoute celle des autres **poils** qui à partir des mêmes principes de croissance folliculaire se différencient également en **cils**, **sourcils**, **poils corporels** ou **poils pubiens**.

La section du cheveu est rarement cylindrique mais plutôt ellipsoïdale. Cette géométrie n'est pas uniforme et des variations peuvent être observées entre les individus et au sein même d'un individu. Les études anthropomorphiques de sections de cheveux illustrent ces variations en termes de tailles, de distribution et de régularité des formes, de présence ou d'absence de médulla et d'abondance et de distribution en pigments de mélanines (Figure 8). Elles suggèrent un lien entre géométries des sections et frisure du cheveu.

L'origine de ces différences semble se trouver au niveau du bulbe. En fonction de la **taille du follicule**, la prolifération des cellules corticales et cuticulaires va être plus ou moins importante ce qui se répercutera sur le diamètre final de la fibre.

Par ailleurs, l'**orientation du bulbe** et donc de l'axe de pousse peut former un angle plus ou moins important par rapport à l'axe perpendiculaire à la surface de la peau. Il en découle une dissymétrie qui peut se répercuter au cours des différentes étapes de la différenciation des cellules.

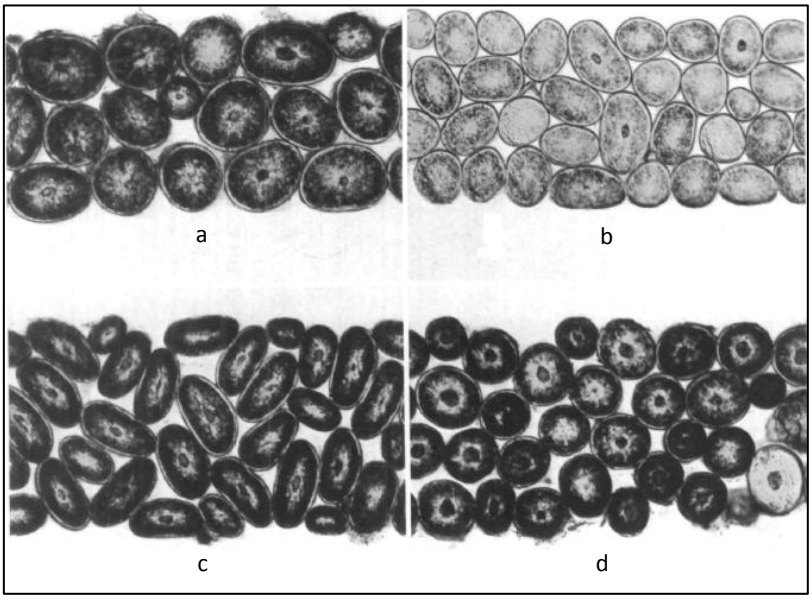


***Figure 8**. Illustration de l'hétérogénéité de sections de cheveux parmi quatre individus nord américains de différentes origines [31]. a) Amérindien navajo b) Européen hollandais c) Afro-américain d) Amérindien maya.*

Pendant la phase d'élongation, l'orientation latérale du bulbe peut entrainer une contraction de la gaine externe entre la fibre en croissance et la gaine de tissu connectif. La prolifération des cellules de la gaine externe est alors dissymétrique et une déformation de la structure interne en ellipse plus ou moins prononcée peut être observée (Figure 9) [104, 105, 112-114].
La dissymétrie de la pousse peut également entrainer une **géométrie de fibre curviligne** pendant sa kératinisation. La rigidification des cellules corticales et cuticulaires fige alors définitivement cette orientation [114].

La **répartition des types de cellules corticales** semble également jouer un rôle dans la forme des cheveux. L'observation des répartitions des types de cellules corticales comparées suggère une disposition des cellules ortho et mésocorticales parmi les cellules paracorticales différentes entre des cheveux de degré de frisure différents [112]. Par ailleurs, différentes études montrent un lien entre la segmentation du cortex en différents types cellulaires et le degré d'inclinaison du bulbe [108]. Ainsi, la répartition des cellules corticales pourrait être une conséquence de la géométrie du follicule, les différences d'empilement des microfibrilles étant causées par des contraintes mécaniques différentes lors de la formation des macrofibrilles.
L'impact de la **présence ou de l'absence de la médulla** sur la structure du cheveu humain n'est pas très bien appréhendé. Elle est couramment considérée comme peu influente sur les propriétés mécaniques du cheveu mais pourrait avoir un rôle plus important dans le caractère spécifique des poils de barbe et pubiens dans lesquels sa proportion par rapport au cortex est plus importante.
La répartition des pigments de **mélanines** s'observe en périphérie du cortex. L'**abondance** de ces pigments et leur **composition** sont variables entre les individus. La métabolisation de la tyrosine au niveau des mélanosomes permet d'aboutir à la biosynthèse de deux catégories de composés dont la polymérisation distincte produit deux classes d'oligomères : les phéomélanines (jaune et rouge) et les eumélanines (noir et brun). L'association de ces oligomères forme des granules dont la composition est responsable des différences de couleur de fibre observées [115-117].

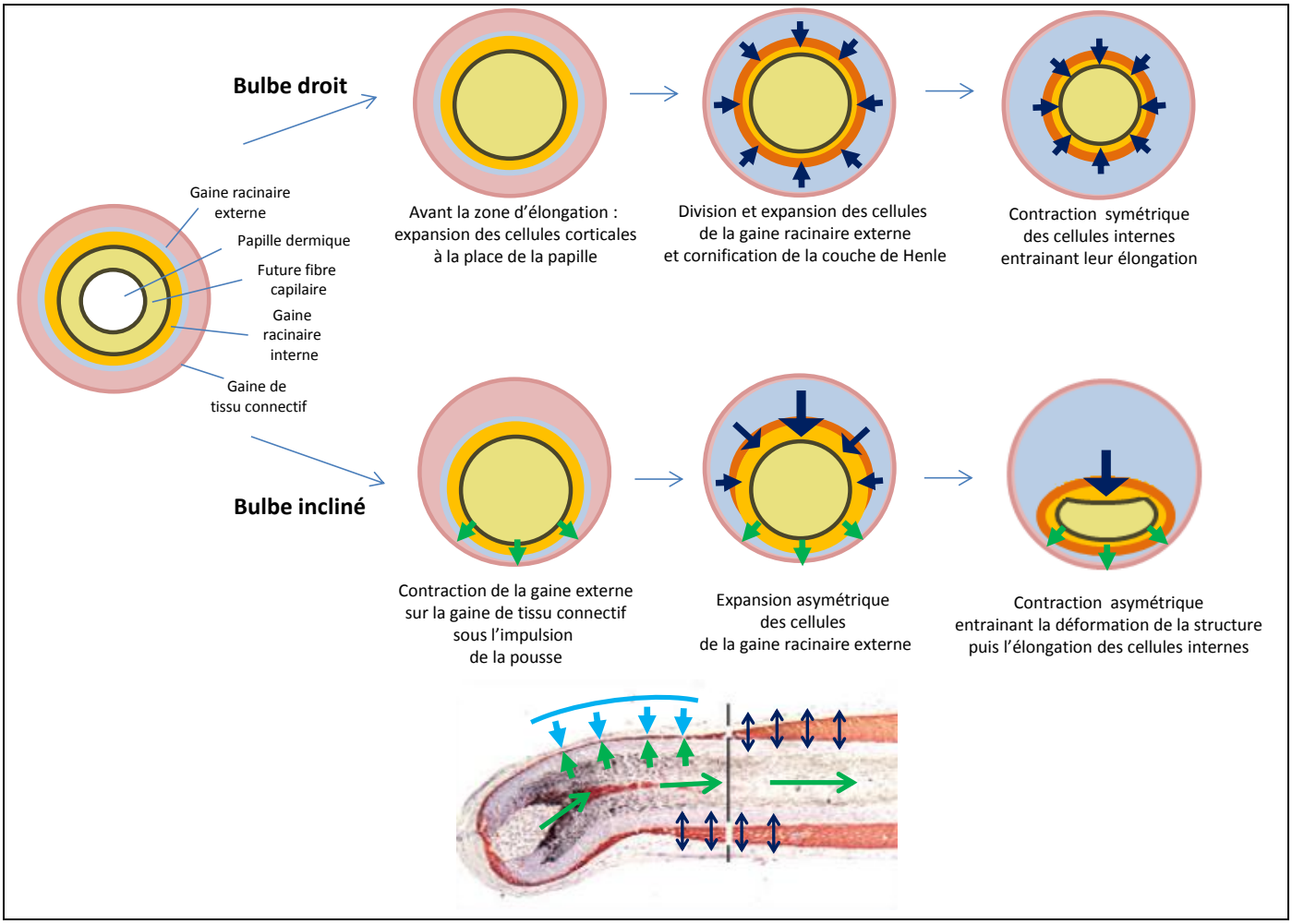


Figure 9 *: Schématisation d'une proposition de mécanisme de contraction des cellules du cheveu pendant la phase d'élongation. La contraction dissymétrique de la gaine externe sur les cellules internes lorsque le bulbe est incliné aurait pour conséquence une déformation de la structure initialement cylindrique pouvant expliquer l'obtention d'une structure rigide finale ellipsoïdale. Adapté des travaux de [114].*

4. Diversité des structures chez les mammifères

Les facteurs responsables des polymorphismes des fibres humaines sont en partie les mêmes que ceux qui expliquent la diversité des poils observés chez les mammifères. Les mécanismes de différenciation folliculaire chez des espèces bien étudiées comme le mouton ou la souris sont similaires à ceux décrits chez l'homme. Cependant, des différences de synchronisations de différenciation et de prolifération cellulaire mais également des différences d'arrangement des différentes lignées cellulaires peuvent être à l'origine de variation de structures des poils entre les espèces.

La densité de l'implantation des follicules dans la peau est également un facteur de variabilité supplémentaire. Nous pouvons noter que les mécanismes responsables de la densité folliculaire commencent à être appréhendés [118].

Parmi les mammifères, le diamètre des fibres peut aller de 10 µm à 250 µm et les sections peuvent également être plus ou moins ellipsoïdales. Tous les ratios possibles entre cuticule, cortex et médulla peuvent être envisagés. Par exemple, le nombre de couches de cellules cuticulaires peut varier de 1 pour le mouton à 35 pour le cochon et leurs dimensions peuvent également différer. L'impression de la forme des cellules cuticulaires de la gaine interne sur les cellules de la cuticule pendant l'élongation peut être à l'origine de grandes différences de géométries de surface entre les espèces [104, 108].

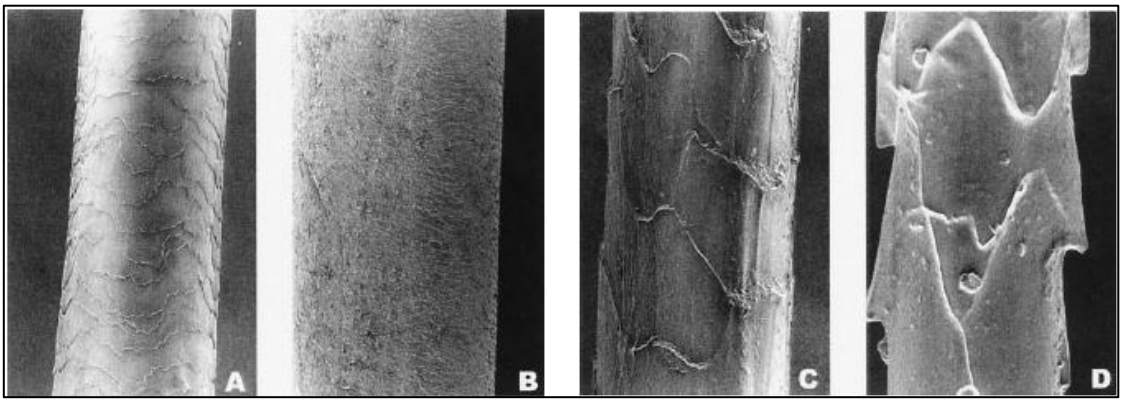


***Figure 10** : Illustration des différences de structures cuticulaires visualisées par microscopie électronique chez l'humain (A), le cochon (B), le mouton (C) et le chat (D). D'après [104].*

Les formes et la fréquence des colonnes de cellules médullaires sont très variées et souvent très éloignées de ce qui peut être observé chez l'homme. Les cellules de la médulla peuvent développer de large espaces intra ou intercellulaires permettant d'emprisonner de l'air et d'améliorer les capacités isolantes de la fibre.

Les piquants sont quasi exclusivement constitués d'organisations régulières de cellules médullaires entourées d'une fine couche corticale comme c'est le cas pour le hérisson, le porc-épic, le petit tenrec ou l'échidné australien [101, 119, 120].

La structure macrofibrillaire au sein des cellules corticales est commune à toutes les espèces de mammifères. Néanmoins, la répartition des types de cellules corticales est également source de différences. Pour un certain nombre de structures, le cortex est bipolarisé entre deux parties contenant respectivement des cellules ortho et paracorticales. La structure et la régularité des empilements peuvent également varier et des différences de ratios filament/matrice sont observés entre les espèces [121] tout comme des différences de composition des protéines de la matrice [108]. Enfin, l'orientation des filaments n'est pas toujours parallèle à l'axe de la fibre notamment lorsque les cellules corticales sont en interaction avec la médulla [108].

Chapitre III Aspects moléculaires de la structure du cheveu

Dans ce chapitre, nous décrivons l'étude du cheveu du point de vue de son organisation moléculaire et tout particulièrement de l'organisation des protéines connues pour y être exprimées. Nous aborderons les techniques ayant permis d'aboutir au catalogue de ces protéines et les informations relatives à leur arrangement dans la structure finale. Un parallèle sera réalisé entre leur expression et les différentes étapes de différenciation des cellules du cortex et de la cuticule.

1. Notion d'homologie de séquence et de famille multigénique

La majorité des protéines du cheveu que nous décrirons dans la suite de ce chapitre ont la particularité de présenter de fortes homologies de séquences permettant de les regrouper au sein de familles. Avant de présenter ces protéines et ces familles, il nous a semblé important de définir l'origine de ces homologies et le vocabulaire associé à la nomination des liens de parenté existant au sein et entre ces familles.

a) Définition et origine de l'homologie de séquence

L'homologie définit l'existence d'une relation entre les gènes qui peuvent être considérés comme possédant un gène ancestral commun.
Chaque gène a une histoire qui peut être étudiée en comparant sa séquence aux autres séquences du génome mais également à celles des génomes d'autres espèces. Les relations d'homologie existant entre les gènes peuvent alors être expliquées sur la base de quelques événements élémentaires mais fondamentaux pour la définition du processus d'évolution [122] :

- La **spéciation** est la divergence de deux génomes qui par la suite vont évoluer indépendamment.
- La **duplication** d'un gène dont les deux descendants vont être modifiés indépendamment.
- La **perte** d'un gène.
- Le **transfert de gène horizontal** qui peut intervenir entre deux espèces.
- Le **réarrangement de gènes** par fusion, fission ou d'autres mécanismes.

Parmi ces événements, les mécanismes contribuant à la majorité des homologies de séquences observées dans un génome sont la spéciation et la duplication. Les **gènes homologues** issus de ces deux événements se distinguent respectivement sous deux noms :

- Les **gènes orthologues** qui sont issus d'un gène ancestral commun mais ont été séparés suite à la spéciation.
- Les **gènes paralogues** qui ont été séparés par duplication.

Les différences de séquences pouvant exister entre les homologues s'expliquent par les mutations qui peuvent découler de la modification ponctuelle et aléatoire de l'ADN au cours de l'évolution du gène.
Il existe un lien entre le degré d'homologie de gènes homologues et la durée pendant laquelle ils ont évolué indépendamment tout en conservant une fonction biologique commune. Le degré d'homologie entre deux séquences homologues peut servir d'horloge moléculaire. Cette information permet d'estimer approximativement la date de l'événement de duplication par rapport à l'évolution de l'espèce [123].
La datation des événements de spéciation par comparaison des séquences génomiques entre les espèces repose en partie sur le principe d'horloge moléculaire. C'est une des bases de la phylogénie, science en pleine essor avec l'augmentation croissante des données génomiques, qui étudie les parentés entre les espèces. Ces informations de datation des spéciations peuvent être comparées aux informations de paléontologie et de géologie.

b) Famille multigénique et isoformes

Certains gènes ont la particularité de posséder de nombreux paralogues issus de duplications successives. Les protéines correspondantes, si elles réalisent des fonctions équivalentes au sein de l'organisme, peuvent être décrites comme des **isoformes** les unes des autres. Ce groupe de paralogues est alors associé à une **famille multigénique**.
La comparaison des séquences protéiques au sein de la famille multigénique peut révéler des différences issues des mutations ponctuelles de l'ADN (substitutions, insertions ou délétions) ayant entrainé des modifications dans la composition en acides aminés.
Ces modifications non silencieuses sont des variants post duplication et deux cas de figure peuvent être envisagés : soit la modification ne change pas le caractère physico-chimique intrinsèque du ou des résidus substitués (par exemple : modification d'un acide aminé apolaire par un autre acide aminé apolaire), soit la modification intervient dans une région peu ou pas fonctionnelle de la protéine et est conservée sans perturber la fonction de la protéine.
Au contraire, certains résidus et certaines séquences au sein d'une famille d'isoformes peuvent être systématiquement conservés et indiquer des zones spécifiques essentielles à la fonction de la protéine [124]. Dans ce cas, les mutations qui ont pu subvenir dans ces zones n'ont pas été conservées au cours de l'évolution et peuvent être considérées comme défavorables à la survie des individus mutants.

2. La découverte des protéines du cheveu et leur caractérisation

a) L'isolement des protéines

Les analyses des protéines des cheveux grâce à leur extraction et à leur séparation pendant la période du milieu des années 60 aux années 80 ont permis d'estimer leur nombre dans les tiges capillaires à une centaine environ. Ces analyses, principalement réalisées sur la laine mais transposables au cheveu, ont montré qu'il était possible de diviser ces protéines en 3 différents groupes en fonction de leurs compositions en acides aminés, de leurs propriétés physico-chimiques et de leurs masses moléculaires :

- Les protéines faibles en soufre, comprises entre 45 000 et 60 000 de masses moléculaires, qui sont les kératines composant les **filaments intermédiaires**. Les différences de leurs points isoélectriques permettent de les distinguer en deux groupes, les **kératines acides** (**type I**) et les **kératines basiques** (**type II**).
- Les protéines très riches en soufre de poids moléculaire très hétérogènes (10 000 à 30 000). Ces protéines ont été décrites comme protéines de la matrice interfilamentaire et nommées **protéines associées aux kératines** (**KAP**). L'hétérogénéité de ces protéines est due à une division en plusieurs familles distinctes constituées de plusieurs membres également hétérogènes. Il est possible entre ces familles de distinguer deux groupes en fonction de leur composition en cystéine, les **KAP riches en soufre** (**HS KAP**) et les **KAP ultra riches en soufre** (**UHS KAP**).
- Les protéines associées aux kératines, riches en glycine et tyrosine (**HGT KAP**) et décrites également dans la matrice interfilamentaire. Elles sont plutôt minoritaires dans le cheveu en comparaison des autres groupes de protéines [108]. Leurs masses moléculaires est inférieures à 10 000 et elles peuvent également être divisées en plusieurs familles sur les bases de leurs propriétés physico-chimiques.

La séparation des protéines du cheveu par ces techniques a permis d'obtenir les bases des connaissances de la composition protéique du cheveu. Néanmoins, la difficulté d'isoler individuellement ses composants très hétérogènes est restée un obstacle à la caractérisation des séquences des différents individus de ces familles de protéines [92].

b) La recherche des gènes

Les stratégies d'études des protéines du cheveu sont limitées par la résolution des techniques séparatives et ne permettent pas de distinguer clairement le nombre de membres parmi les familles de protéines d'intérêt.
En revanche, il est possible d'associer des séquences nucléotidiques à des séquences protéiques pour identifier les différents gènes correspondant au sein de ces familles. Les séquences nucléotidiques sont issues de données de séquençage d'ADN regroupées dans des banques. Les séquences protéiques proviennent de la caractérisation de séquences partielles ou totales de protéines isolées et séquencées. La comparaison des deux informations permet la recherche des sites des gènes correspondant à la famille de protéine dans la banque nucléotidique (Figure 1).
Ces recherches s'appuient sur le principe qu'il existe des homologies de séquence au sein d'une famille voire entre les familles. Cette homologie permet, à partir d'un morceau isolé de séquence d'un individu possédant des homologues, de rechercher dans la banque nucléotidique l'ensemble des zones comportant cette séquence exacte ou approchée et ainsi d'identifier les gènes homologues. Cette information permet ensuite de connaitre leur séquence nucléotidique, leur position dans le chromosome et, après conversion *in silico,* de connaitre la séquence protéique correspondante.

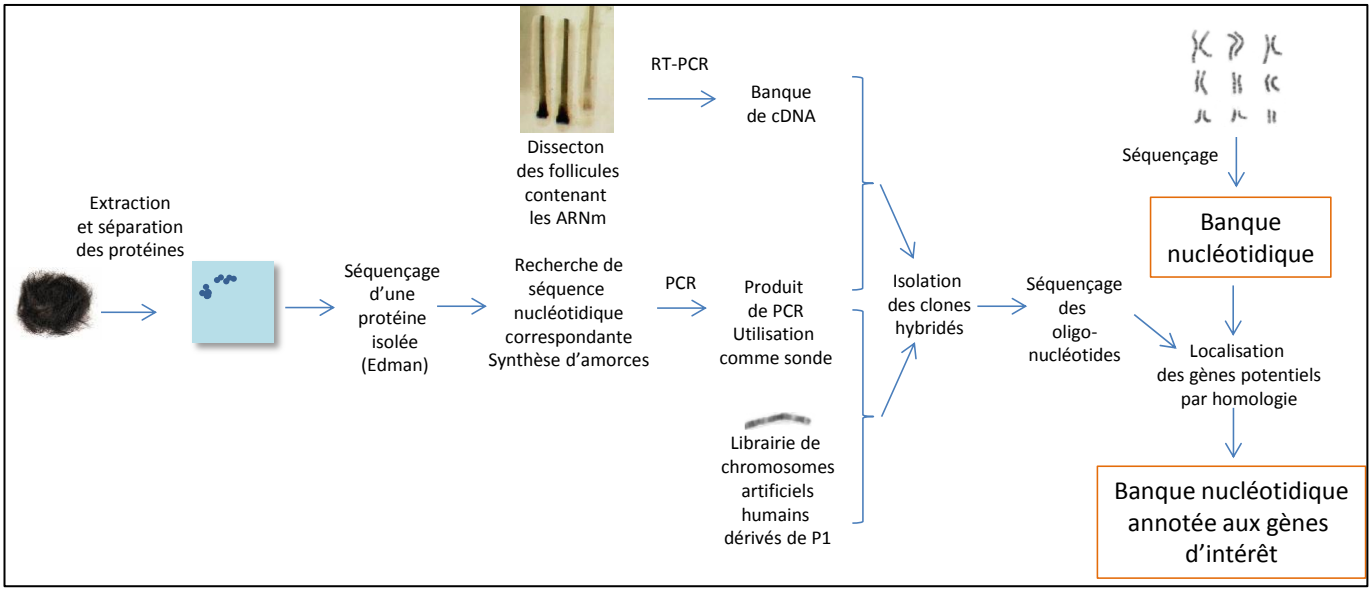


***Figure 1** : Stratégies utilisées pour la caractérisation des gènes correspondants aux protéines du cheveu humain.*

L'introduction des techniques de biologie moléculaire dans des études chez le mouton et chez la souris dans les années 80 [63, 67-70] puis chez l'humain à partir années 90 [71, 72] a permis la caractérisation des gènes pouvant y être exprimer et de remonter à l'information du nombre de familles et de membres par familles pouvant être retrouvés dans la structure.
Les données de séquences nucléotidiques disponibles associées à l'utilisation de la réaction en chaine par polymérase (PCR) et à l'isolement de séquences nucléotidiques provenant des ARN messagers (ARNm) isolés dans les follicules ont permis la caractérisation d'un certain nombre de ces gènes. Le catalogue des gènes correspondants a été extensivement complété et affiné au cours de la dernière décennie notamment grâce aux données apportées par le séquençage complet du génome humain [3, 73-83].
L'ensemble de ces données a permis de montrer l'existence de l'ensemble des gènes des individus des familles de protéines que nous décrirons dans les troisièmes et quatrièmes parties de ce chapitre.

c) L'évidence de l'expression des gènes en protéine

La caractérisation des gènes précède une seconde étape : montrer l'expression des gènes au sein d'une cellule. L'expression d'un gène peut être démontrée en mettant en évidence l'expression de son transcrit l'ARNm et de la traduction du transcrit, la protéine (Figure 2).
L'information peut être affinée en localisant ces expressions au sein des cellules en fonction de leurs différentes étapes de différenciation.
La connaissance des séquences nucléotidiques des gènes à étudier permet :

- de **localiser l'activité de transcription** en produisant des ARN complémentaires (ARNc) à la séquence ARNm à mettre en évidence. Les ARNc, radiomarqués, sont produits par PCR puis transcription et vont pouvoir s'hybrider avec les ARNm présents dans la coupe de follicule. Les hybrides peuvent alors être détectés [81].
- de **localiser les protéines** par marquage immunohistochimique. Après avoir recherché une zone spécifique dite antigénique pour la protéine issue du gène à étudier, un peptide correspondant à cette séquence est synthétisé et inoculé à un organisme pour produire un anticorps monoclonal. L'anticorps après extraction de l'organisme peut être utilisé pour identifier la protéine extraite et séparée sur gel mono ou bidimensionnel. Cette étape permet de vérifier la spécificité de l'anticorps par rapport aux autres isoformes présentes dans l'extrait. Il peut alors être dirigé sur la coupe de follicule sur laquelle la protéine est localisée [82, 83].

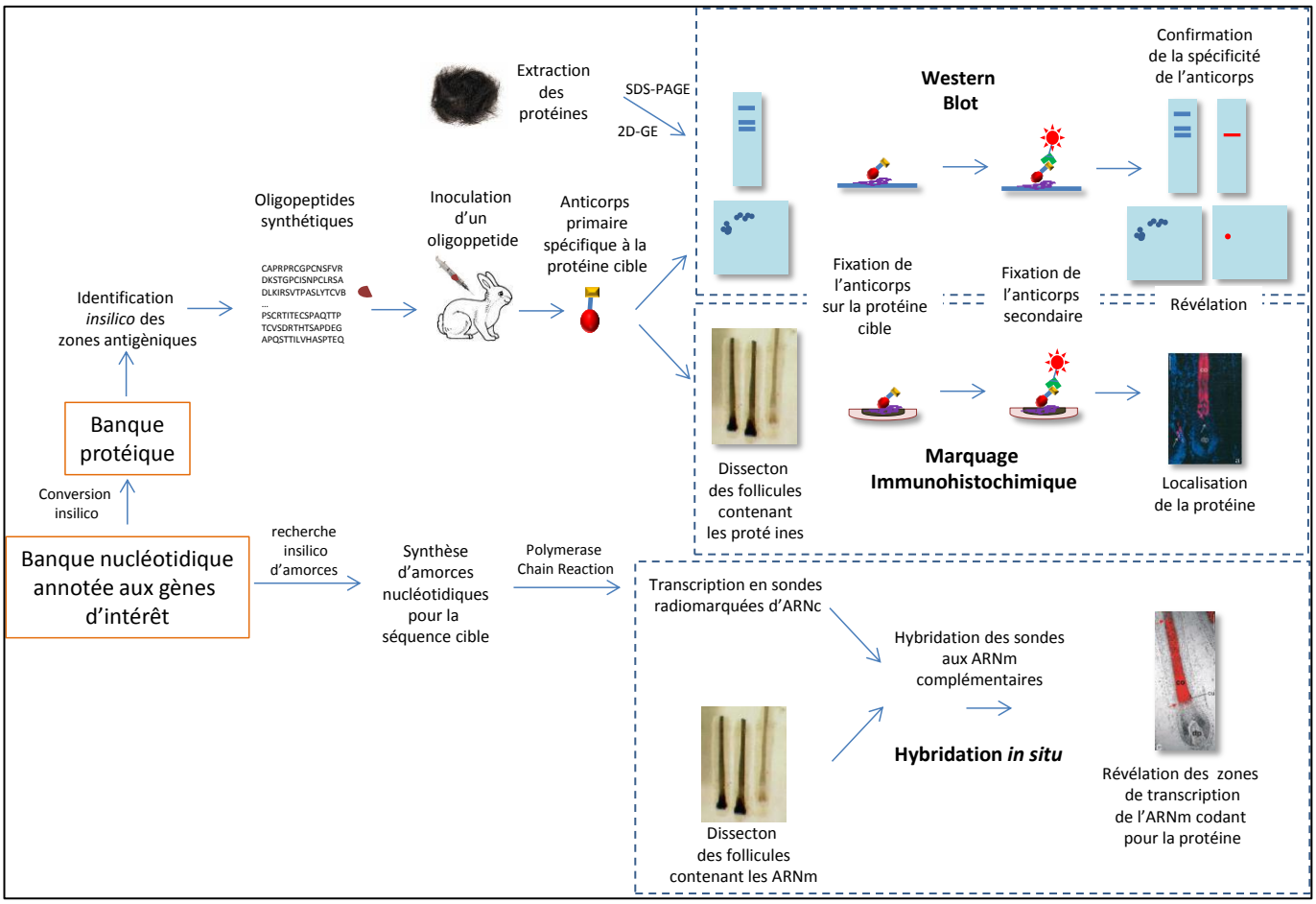


***Figure 2** : Stratégies utilisée pour la mise en évidence de l'expression de gènes dans le cheveu au niveau protéique et transcriptomique.*

Le processus de pousse du follicule rend l'étude de l'expression des protéines et des ARNm sur les sections longitudinales de follicules tout à fait originale. La localisation de ces expressions permet en un seul cliché de déduire quel type de cellule réalise la traduction ou la transcription du gène mais également à quel moment de la différenciation cellulaire ces dernières commencent et se terminent. Si l'on ajoute que l'on connait approximativement la vitesse de pousse du follicule, il est également possible de déduire la durée de ces différentes étapes de différenciation simplement en mesurant la distance sur laquelle elles se retrouvent.

3. Les kératines

a) Les kératines, des protéines des filaments intermédiaires

Les **kératines** sont les protéines qui constituent une des bases structurales du cytoplasme des **cellules épithéliales**. Elles se regroupent dans 2 des 6 sous-familles, les protéines de **type I** et de **type II**, de la superfamille des **protéines des filaments intermédiaires (IFP)** (source : Human Intermediate Filament Database). Les protéines de cette superfamille s'assemblent pour constituer des filaments dont le diamètre, de 10 à 12 nm, est intermédiaire entre les microfilaments d'actine (7-10 nm) et les microtubules (25 nm) [125].
Les kératines ont pour fonctions, entre autres, de maintenir certaines structures cellulaires et sont retrouvées chez l'ensemble des vertébrés [126, 127]. 28 gènes fonctionnels de type I et 26 de type II sont connus chez l'humain.

b) Les catégories de kératines

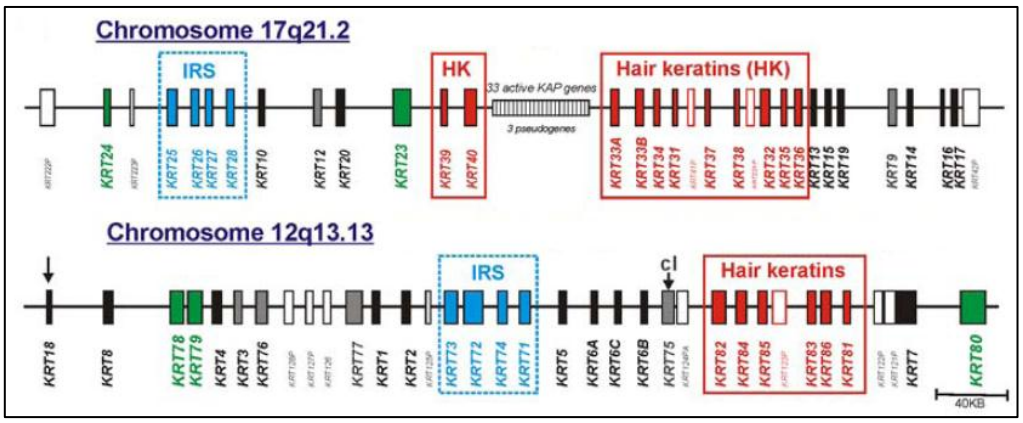


Figure 3 *: Localisation des gènes des kératines de type I et II dans les locus 17q21.2 et 12q13.13. Nous noterons l'insertion d'un segments de gènes de KAP entre des gènes des kératines du cheveu. D'après [128].*

Au cours de la dernière décennie, l'expression de ces gènes en protéines a été largement étudiée dans différents types de cellules épithéliales. Ces études ont permis la division des gènes des deux groupes en plusieurs catégories fonctions de la localisation cellulaire de leur expression.

Les gènes des deux types de sous familles sont localisés sur **deux clusters** chromosomiques distincts (Figure 3), sur le locus 17q21.2 pour les types I et 12q13.13 pour les types II [127].
Les kératines se distinguent entre les **kératines du cheveu** et les **kératines épithéliales** [128, 129]. Parmi les kératines épithéliales retrouvées dans le follicule, certaines sont retrouvées spécifiquement dans la gaine interne [130]. Les kératines épithéliales de la gaine externe sont également retrouvées dans d'autres cellules épithéliales extérieures au follicule pileux.

D'autres sont retrouvées dans d'autres types de cellules épithéliales qui peuvent être stratifiées et kératinisées/non-kératinisées ou non stratifiées (Figure 4). Ces dernières sont exprimées dans l'épiderme et dans les différentes muqueuses [131].

Les kératines co-exprimées dans chaque type de cellules différenciées du follicule ont la particularité d'être très homologues. Ces dernières peuvent, dans la plupart des cas, être considérées comme des isoformes caractéristiques du programme de différenciation de la cellule (Figure 5).

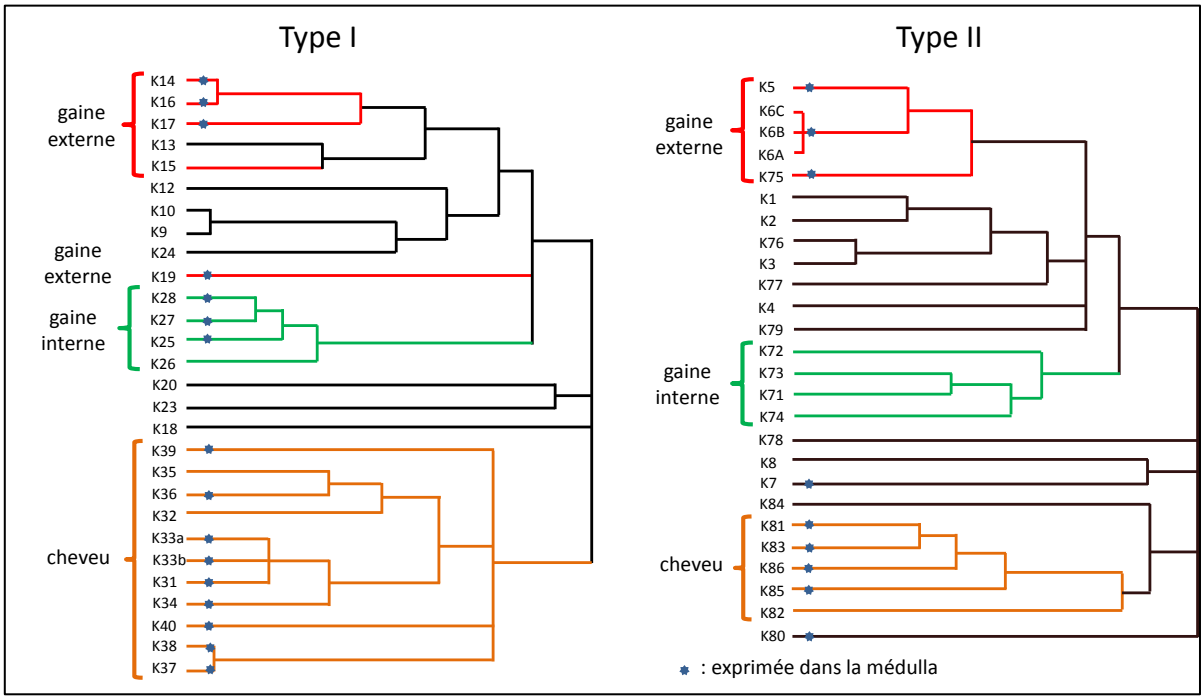


Figure 4 *: Arbre phylogénétique des kératines de type I et II et localisation de leur expression au sein des différents compartiments du follicule pileux. Obtenu d'après les clusters d'alignement trouvés avec MultAlin des séquences protéiques extraites de Swissprot (23/02/2011) et nommées en accord avec la nouvelle nomenclature. Localisation d'après les données d'immunohistochimie [101].*

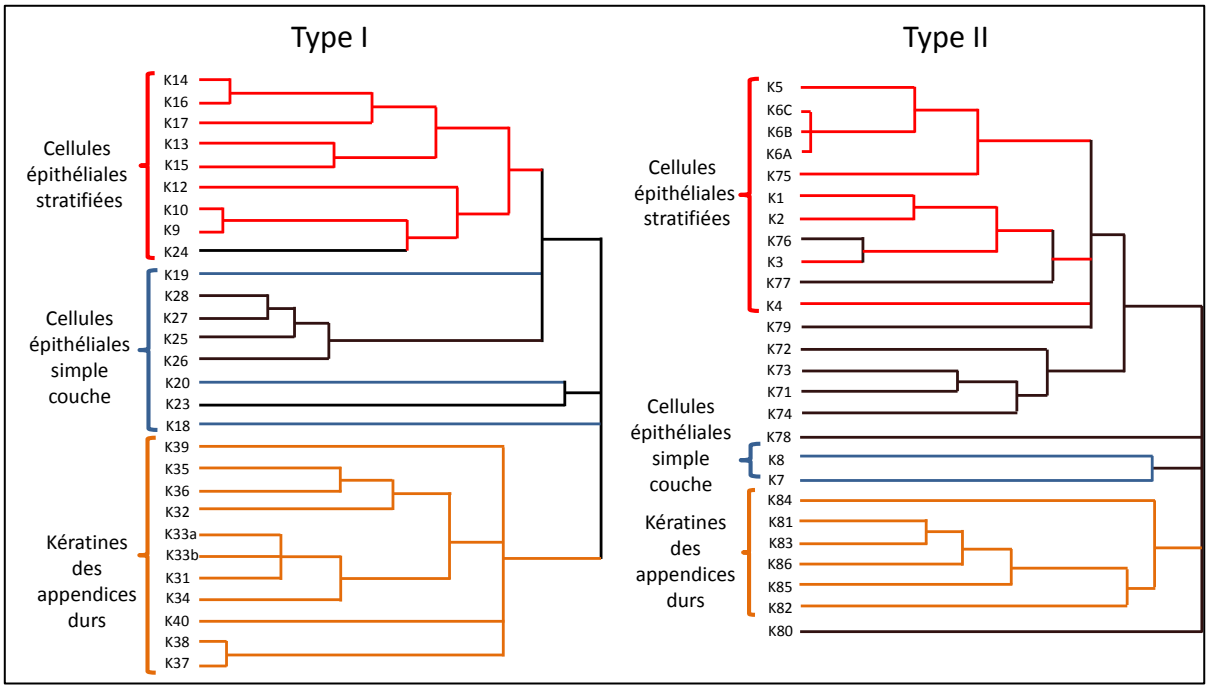


Figure 5 *: Arbre phylogénétique des kératines de type I et II et localisation de leur expression au sein de différents types de cellules épithéliales. Localisation d'après [128].*

Cette propriété de localisation des isoformes se retrouve également illustrée dans la répartition des groupes d'isoformes parmi les types de cellules épithéliales et suggère le rôle spécifique de chaque groupes sur la différenciation cellulaire.

Les kératines du cheveu et de la gaine interne se différencient donc des autres kératines des cellules épithéliales par leur séquence et leurs gènes ancestraux communs. Les liens de parenté entre les isoformes se retrouvent ainsi dans leur homologie de séquence et dans la localisation cellulaire de leur expression.

c) Structures primaires et secondaires des kératines : la programmation de l'hétérodimérisation

Toutes les kératines de type I et II ont une structure protéique commune qu'elles partagent avec les autres protéines des sous familles des filaments intermédiaires. Elles se divisent en trois parties caractéristiques incluant au centre, une **tige** constituée d'un long segment enroulé en hélice-α et, à chacune des extrémités, une **tête** et une **queue** non hélicoïdales.

La séquence en acides aminés de la tige centrale (Figure 6) révèle une périodicité de résidus heptapeptidiques de type $(abcdefg)_n$ où a et d sont préférentiellement des résidus apolaires. En solution, l'effet hydrophobe qui découle de cet enchainement est en partie responsable de l'enroulement spontané de la tige en hélice-α. [87, 125, 132]. Au sein de la structure périodique s'intercalent de courtes interruptions du motif de répétition définissant des liens flexibles (L1, L12, L2) entre 4 segments hélicoïdaux très conservés nommés de la tête à la queue 1A, 1B, 2A et 2B.

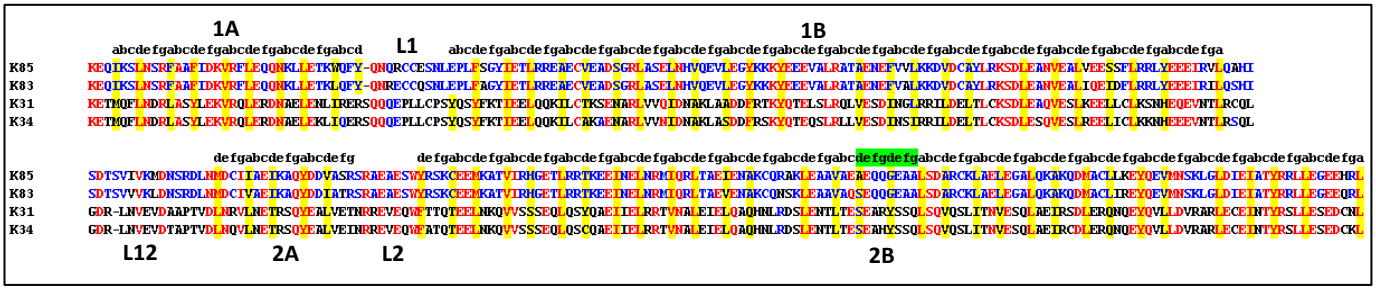


***Figure 6** : Alignement des séquences des tiges de deux couples de kératines du cheveux de type I et II. Les répétitions des motifs heptapeptidiques définissant les zones hélicoïdales et les zones de liens sont représentées. En jaune, les résidus apolaires des positions a et d. En vert, le site de bégaiement dans un des motifs de la chaine 2B.*

Les hélices constituées par les IFP ont la propriété de s'enrouler pour former un dimère constitué de l'enroulement en hélice des deux unités. Parmi la super famille des IFP, les kératines ont la particularité de s'associer très préférentiellement en hétérodimères constitués d'une kératine de type I et d'une kératine de type II. Cette association est favorisée par la complémentarité des séquences entre les deux types de kératines mais n'exclut cependant pas une formation minoritaire d'homodimères. L'hétérodimère est stabilisé par une dizaine d'interactions inter chaines de résidus acides avec des résidus basiques situés sur les positions g et e des motifs hepta peptides de chaque chaine (Figure 7). Ces interactions entrainent alors un arrangement parallèle des deux chaines. L'existence d'hétérodimères pour la formation des filaments intermédiaires de kératine pourrait être liée à la nécessité de spécificité d'interactions excluant les autres molécules de la cellule [85, 86].

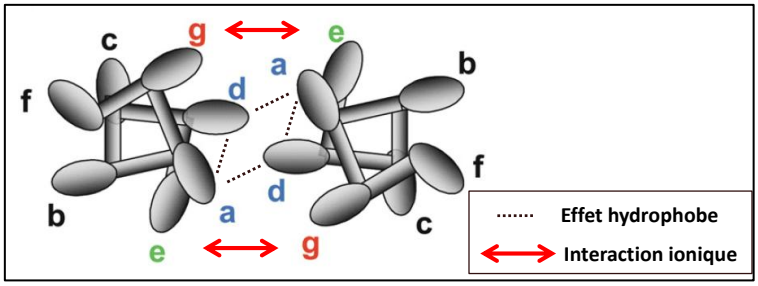


***Figure 7** : Illustration des interactions latérales pouvant exister entre les résidus de deux hélices stabilisant la formation des hétérodimères.*

Les kératines des cheveux et les différents types de kératines épithéliales se distinguent principalement par la composition et la taille des séquences des liens, des têtes et des queues. Les séquences correspondantes des têtes et des queues des kératines des cheveux contiennent une proportion importante de résidus cystéine (environ 11% pour les types II et 17% pour les type I) tandis que les kératines épithéliales sont plus riches en résidus glycine et serine et possèdent peu ou pas de cystéines [86].

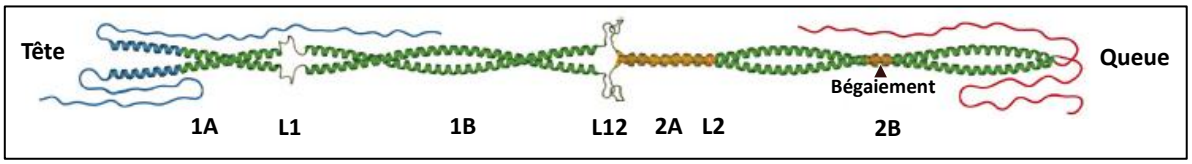


Figure 8 *: Modèle d'un homodimère d'une IFP de type III (vimentine) basé sur des données d'assemblage in vitro. [133]*

Dans les cellules du follicule pileux, la co-localisation de plusieurs isoformes de kératines pose la question des spécificités d'interaction qui pourraient exister entre elles pour la formation de dimères (Figure 8). Sur la base des homologies de séquences de la tige, seule la formation d'hétérodimères devrait être favorisée. Cependant, il ne semble pas qu'il puisse y avoir des appariements très spécifiques entre une espèce de type I et une espèce de type II particulières [86]. Néanmoins, les éventuelles complémentarités des domaines tête et queue qui pourraient favoriser certaines associations entre des kératines de type I et II homologues n'ont, à notre connaissance, pas encore été étudiées.

d) La formation des filaments intermédiaires

Suite à la formation des hétérodimères, une seconde étape d'association peut être réalisée avec l'appariement de deux dimères en tétramère ou protofilament. Des données expérimentales de pontages chimiques des structures formées *in vitro* ont permis d'envisager plusieurs modes d'association entre les segments hélicoïdaux des dimères (Figure 9). L'analyse des séquences des segments suggère que les modes d'agrégation moléculaire A_{11} et A_{22} sont les plus favorables [133].

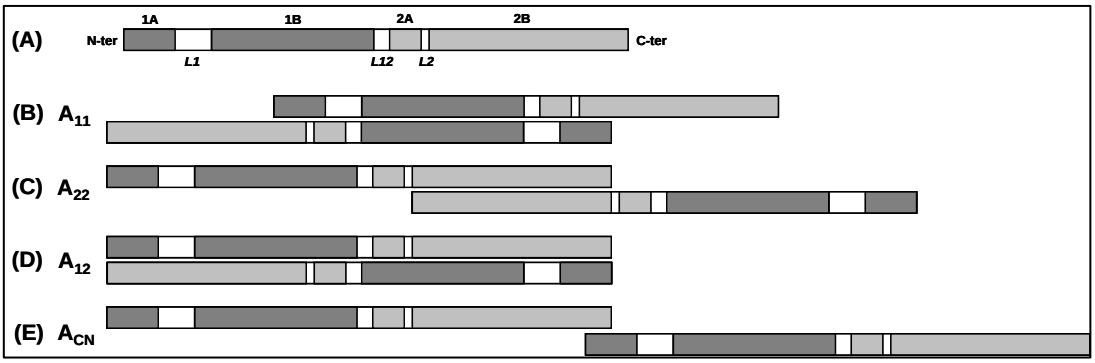


Figure 9 *: Illustration des différents modes d'arrangement tetramérique pouvant être adoptés par l'unité dimérique (A). (B), (C) et (D) les modes antiparallèles respectivement A_{11}, A_{22} et A_{12}, et (E) le mode parallèle A_{CN}.*
Adapté de [133].

Le détail atomique des interactions existent au sein des couples de dimères n'est pour l'instant pas connu, tout comme les étapes suivant la formation des tétramères et conduisant à la structure filamentaire finale constituée de 32 kératines. Des données de suivi de formation *in vitro* des filaments de vimentine par diffraction des rayons X aux petits angles suggère un modèle de formation latérale d'octamères, les protofibrilles, par la suite regroupées en quatre octamères qui définissent une **unité de longueur filamentaire (ULF)** d'environ 60 à 70 nm

[129, 133, 134]. L'assemblage latéral des ULF est réalisé de concert avec l'élongation progressive de la structure filamentaire.

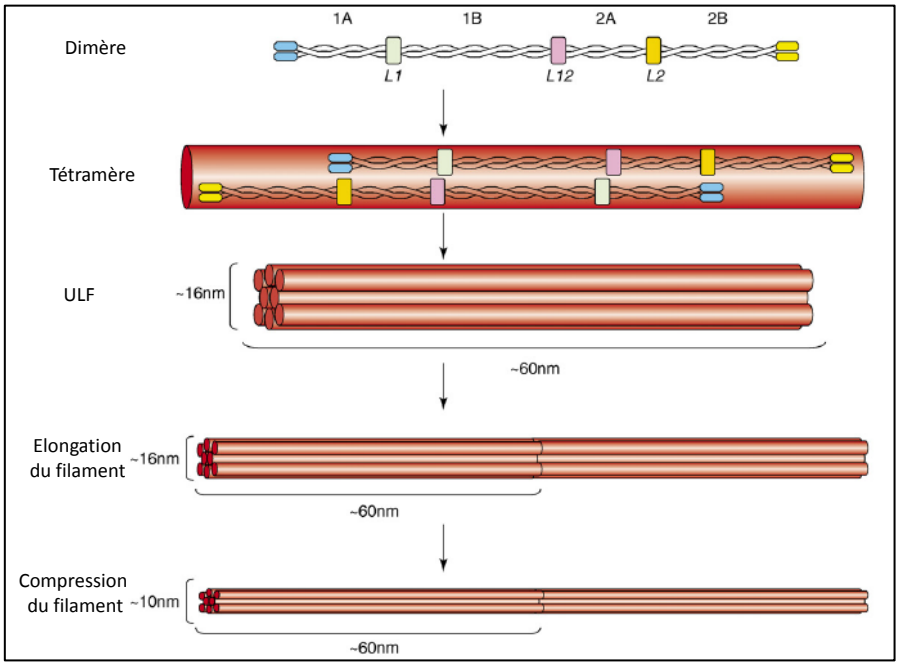


***Figure 10** : Schématisation des étapes de formation des filaments intermédiaires d'après [134].*

Un autre modèle, basé sur l'étude de la microdiffraction aux rayons X du follicule, suggère l'organisation de l'ULF en 8 tétramères organisés en cylindre par la suite compacté en un cylindre de 7 protofilaments autour d'un 8ème protofilament [135, 136]. La structure filamentaire peut donc avoir deux formes, désignées réduite ou oxydée [137].

4. Les protéines associées aux kératines (KAP)

a) Fonctions

La fonction supposée des familles de KAP dans le cortex est d'entourer les filaments intermédiaires de kératines pendant la formation des macrofibrilles. Elles se retrouvent alors au sein de la matrice interfilamentaire composant les macrofibrilles et ont très certainement un rôle dans les propriétés mécaniques de la structure finale des cellules corticales. Leur rôle est encore mal défini et leur structure tertiaire est inconnue.
Certaines familles sont fortement multigéniques (KAP 4, 5, 9, 10, 19 et 28) avec environ une dizaine de gènes décrits par famille. D'autres ne se composent que d'un unique individu (7, 8, 11, 17, 23, 24, 25, 26…) tandis que certaines ont une multigénicité intermédiaire (1, 2, 3, 6, 13, 20, 21…).

b) Les familles de KAP

Les KAP constituent plus d'une vingtaine de familles multigéniques rassemblant une centaine d'individus pouvant être communément regroupés parmi différents groupes : les HS KAP, les UHS KAP et le HGT KAP.
Sur les bases des informations génomiques et de la localisation de leur expression, établies au cours de la dernière décennie, cette division en trois groupes peut néanmoins être affinée.
Les relations phylogénétiques pouvant exister entre toutes ces familles sont difficiles à mettre en évidence au premier abord. En effet, une partie des séquences protéiques de ces familles possède des motifs de répétition des résidus suggérant que de nombreux événements de type insertions/délétions sont survenus au cours de l'évolution de ces gènes. La combinaison de ces événements, associée au fait que les relations de paralogie entre

ces familles sont très anciennes, explique la difficulté d'établir des relations phylogéniques de la même manière que pour les kératines. Différents arbres sont décrits dans la littérature mais leurs résultats ne sont pas consensuels [73, 138]. Elles permettent principalement d'exclure des liens directs de parenté entre certaines familles de HS KAP, les UHS KAP et les HGT KAP.

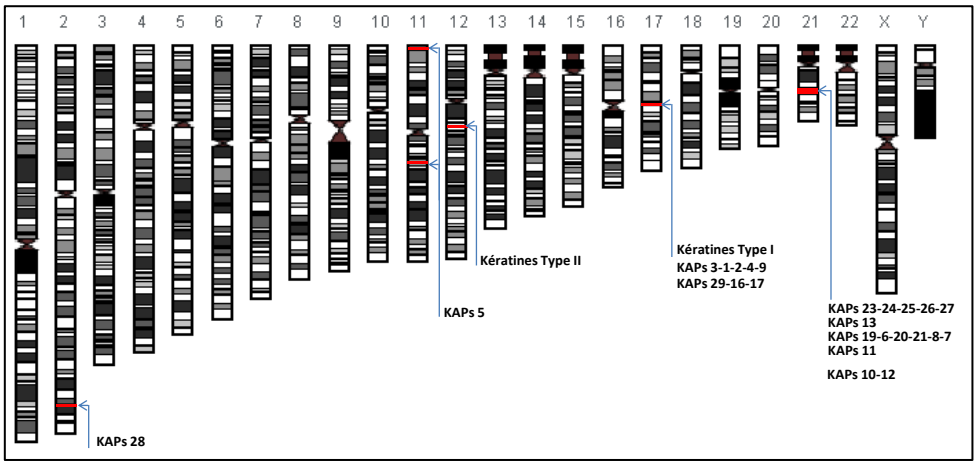


Figure 11 *: Répartitions des clusters de gènes des kératines et des KAP dans le caryotype humain. Structure du caryotype d'après HuRef Project.*

Les regroupements peuvent alors se faire sur d'autres critères : les propriétés physico-chimiques des séquences, leur localisation dans des clusters au sein du caryotype humain et la localisation de leur expression dans les cellules différenciées du follicule pileux.

L'ensemble des gènes décrits se répartis dans le génome humain sur 7 clusters chromosomiques (Figure 11) et l'expression de ces gènes peut être montrée dans le cortex et/ou la cuticule mais également dans la matrice folliculaire.

Les groupes présentés sont une adaptation de la revue sur les KAP de Rogers et al. de 2006 complétée des données issues des résultats postérieurs et de nos propres analyses des séquences (Tableau 1) [73, 138-140].

Les KAP riches en soufre du cluster 17q21.2

Parmi ce cluster inséré dans celui des kératines de type I, se distingue le groupe constitué par les KAP **3, 2, 1, 4 et 9**, riches en cystéine, proline, sérine, thréonine et proline et de tailles croissantes (respectivement entre 10,5 et 19,5 kDa en moyenne). Les KAP 4 et 9 ont en commun une très forte proportion de cystéine (>30%) et sont vraisemblablement les plus liées par un lien d'homologie. Leur expression est décrite dans les cellules corticales au niveau de la zone d'élongation.

Un second groupe correspondant à 2 gènes peut être constitué par les **KAP 16.1 et 29.1**. Ces protéines sont également riches en cystéine, sérine, proline et thréonine mais ce distingue du précédent groupe par une composition plus importante en résidus petits et hydrophobes et surtout par leur tailles beaucoup plus importante (respectivement 54 kDa et 35,2 kDa). L'expression de ces gènes en protéine n'a pas été, à notre connaissance, étudiée.

Un gène isolé correspond à la **KAP 17.1**. Cette KAP est petite (9,5 kDa) et possède une très forte proportion de cystéine (36%) et de glycine (29,5%). La transcription du gène correspondant a été localisée dans la cuticule.

Les KAP riches en soufre des clusters 11p15.5 et 11q13.5

La famille multigénique des **KAP 5** est retrouvée sur ces deux clusters. Les séquences correspondantes sont très riches en cystéine et en glycine (respectivement 33,5% et 24%) et comportent également une forte proportion de sérine (21%). La taille moyenne parmi cette famille est de 18,5 kDa et l'expression des transcrits a été localisée dans la cuticule.

KAP family	Ref	Category	Chromosomal location	Expression	Family members	SwissProt Acess Number	pI	MW	GRAVY	nAA	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
KAP 1	Rogers et al, 2001	HS	17q21.2	Cortex	4		5,95	16688,1	0,00025	162	2,3	3,7	0,0	0,8	25,3	6,1	3,9	10,4	1,9	2,2	1,5	0,0	0,2	2,0	8,5	17,2	8,4	0,6	2,3	2,8
				t	1.1	Q07627	5,55	18234,6	-0,011	176	2,3	4,0	0,0	1,1	25,6	6,3	4,0	10,2	1,1	1,7	1,1	0,0	0,0	2,8	8,5	17,0	8,5	0,6	2,3	2,8
				t	1.3	Q8IUG1	5,53	18183,6	0,026	176	1,7	3,4	0,0	0,6	26,7	6,3	4,0	9,7	1,1	1,7	1,1	0,0	0,0	2,3	8,5	17,6	9,7	0,6	2,3	2,8
				t	1.4	P0C5Y4	6,14	12323,9	-0,016	121	3,3	3,3	0,0	0,8	24,8	5,0	4,1	12,4	4,1	2,5	2,5	0,0	0,8	0,0	8,3	14,9	7,4	0,8	2,5	2,5
				t	1.5	Q9BYS1	6,59	18010,4	0,002	173	1,7	4,0	0,0	0,6	24,3	6,9	3,5	9,2	1,2	2,9	1,2	0,0	0,0	2,9	8,7	19,1	8,1	0,6	2,3	2,9
KAP 2	Rogers et al, 2001	HS	17q21.2	Cortex	5 ?		8,29	13487,1	0,108	127	1,8	8,5	0,0	1,6	27,8	4,7	1,6	5,3	0,0	1,6	2,0	0,0	0,0	0,4	14,2	10,2	11,8	0,8	0,8	7,1
				t	2.1a 2.1b	Q9BYU5	8,32	13513,9	0,086	127	1,6	8,7	0,0	1,6	27,6	4,7	1,6	5,5	0,0	1,6	1,6	0,0	0,0	0,8	14,2	10,2	11,8	0,8	0,8	7,1
				t	2.2	Q9BYT5	8,21	13460,8	0,141	127	1,6	7,9	0,0	1,6	28,3	4,7	1,6	5,5	0,0	1,6	1,6	0,0	0,0	0,8	14,2	10,2	11,8	0,8	0,8	7,1
				t	2.3	P0C7H8	8,32	13493,9	0,111	127	2,4	8,7	0,0	1,6	27,6	4,7	1,6	4,7	0,0	1,6	2,4	0,0	0,0	0,0	14,2	10,2	11,8	0,8	0,8	7,1
				t	2.4	Q9BYR9	8,32	13479,9	0,094	127	1,6	8,7	0,0	1,6	27,6	4,7	1,6	5,5	0,0	1,6	2,4	0,0	0,0	0,0	14,2	10,2	11,8	0,8	0,8	7,1
KAP 3	Rogers et al, 2001	HS	17q21.2	Cortex	3?		5,6	10437,2	0,108	98	2,0	3,1	3,1	3,1	19,0	2,7	3,1	4,8	2,0	2,7	8,8	1,4	1,0	2,0	16,0	7,8	11,2	1,0	1,0	4,1
				p	3.1	Q9BYR8	5,99	10539,3	0,065	98	2,0	3,1	3,1	3,1	18,4	2,0	3,1	5,1	2,0	3,1	8,2	2,0	1,0	2,0	15,3	8,2	10,2	1,0	3,1	4,1
				p	3.2	Q9BYR7	5,4	10407,2	0,154	98	2,0	3,1	3,1	3,1	19,4	3,1	3,1	4,1	2,0	3,1	9,2	1,0	1,0	2,0	16,3	8,2	11,2	1,0	0,0	4,1
				p	3.3	Q9BYR6	5,4	10365,1	0,105	98	2,0	3,1	3,1	3,1	19,4	3,1	3,1	5,1	2,0	2,0	9,2	1,0	1,0	2,0	16,3	7,1	12,2	1,0	0,0	4,1
KAP 4	Rogers et al, 2001	UHS	17q21.2	Cortex	11		8,33	19445,7	0,168	181	0,4	8,4	0,6	0,6	36,2	6,3	1,2	2,3	1,0	1,5	1,2	0,4	0,0	0,2	9,4	15,6	8,5	0,0	1,1	5,0
				t	4.1	Q9BYQ7	8,13	13241,5	0,153	126	0,8	7,1	0,8	1,6	34,1	7,1	0,8	4,8	0,8	0,0	2,4	0,0	0,0	0,0	7,9	15,1	11,1	0,0	0,0	5,6
				t	4.2	Q9BYR5	8,3	14401,8	0,083	135	0,0	8,1	1,5	0,7	34,8	7,4	0,7	3,7	0,0	0,0	0,7	0,0	0,0	0,7	8,1	14,8	11,9	0,0	1,5	5,2
				t	4.3	Q9BYR4	8,43	20504,3	0,27	194	0,5	7,7	0,0	0,5	36,6	3,6	0,5	2,6	1,0	4,6	0,5	1,5	0,0	1,0	9,3	20,6	7,2	0,0	0,5	1,5
				t	4.4	Q9BYR3	8,39	18023,2	-0,003	165	0,0	9,1	1,2	0,6	37,0	9,1	0,6	2,4	1,2	0,0	1,2	0,0	0,0	0,0	8,5	11,5	10,9	0,0	1,8	4,8
				t	4.5	Q9BYR2	8,24	19916,5	0,202	185	0,5	7,6	0,5	0,0	36,8	4,9	2,2	1,6	1,6	1,6	1,1	0,5	0,0	0,0	9,7	15,7	8,6	0,0	2,2	4,9
				t	4.6	Q9BYQ5	8,43	21825	0,227	204	0,5	9,8	0,0	0,5	37,3	5,9	1,0	2,0	0,5	1,5	1,5	0,0	0,0	0,0	9,8	16,2	7,8	0,0	0,5	5,4
				t	4.7	Q9BYR0	8,35	22490,8	0,221	208	0,5	8,7	0,0	0,5	37,0	6,3	1,4	1,9	1,4	1,9	1,4	0,5	0,0	0,0	10,1	15,4	6,7	0,0	1,0	5,3
				t	4.8	Q9BYQ9	8,23	20720,5	0,225	194	0,5	6,7	0,5	1,0	35,6	6,7	1,0	2,1	1,5	2,1	0,5	1,0	0,0	0,0	9,8	16,5	6,2	0,0	1,5	6,7
				t	4.9	Q9BYQ8	8,4	20572	0,093	191	0,5	9,4	0,5	0,0	36,6	5,8	2,1	0,5	1,0	1,6	0,5	0,5	0,0	0,0	11,0	15,2	8,4	0,0	1,0	5,2
				t	4.11	Q9BYQ6	8,32	20799,6	0,186	194	0,5	9,3	0,5	0,5	35,6	5,7	1,5	2,1	1,5	2,1	1,5	0,5	0,5	0,0	9,8	16,5	5,7	0,0	1,0	5,2
				t	4.12	Q9BQ66	8,37	21407,4	0,191	200	0,5	9,0	1,0	0,5	37,0	6,5	1,0	2,0	0,5	1,0	1,5	0,0	0,0	0,0	9,5	14,5	9,5	0,0	0,5	5,5
KAP 5	Yahagi et al, 2004	UHS	11p15.5 / 11q13.5	Cuticule	12?		8,27	18413,6	0,321	202	0,8	0,6	0,3	0,1	33,4	3,4	0,1	24,2	0,0	0,9	0,0	5,5	0,6	0,2	4,5	21,2	0,2	0,1	0,7	3,4
				t	5.1	Q6L8H4	8,39	24193,8	0,33	278	1,1	0,4	0,0	0,0	30,6	1,4	0,0	31,7	0,0	0,4	0,0	6,1	0,7	0,0	3,6	19,8	0,4	0,0	0,0	4,0
				t	5.2	Q701N4	8,31	16270,8	0,318	177	0,6	1,7	0,6	0,0	32,8	3,4	0,0	24,9	0,0	0,6	0,0	4,5	0,6	0,0	4,5	19,8	0,6	0,6	0,6	4,5
				t	5.3	Q6L8H2	8,32	22105,6	0,303	238	0,4	0,4	0,0	0,0	33,2	3,8	0,0	20,2	0,0	0,8	0,0	5,9	0,4	0,4	5,0	24,8	0,0	0,0	0,8	3,8
				t	5.4	Q6L8H1	8,38	25248,7	0,207	288	0,3	0,0	0,3	0,0	29,9	2,4	0,0	31,2	0,0	0,7	0,0	6,2	0,3	0,0	3,1	21,9	0,0	0,0	0,7	2,8
				t	5.5	Q701N2	8,39	21408,7	0,253	237	1,3	0,8	0,0	0,0	31,2	3,0	0,0	28,7	0,0	0,4	0,0	5,9	0,8	0,4	3,8	19,0	0,0	0,0	1,7	3,0
				t	5.6	Q6L8G9	8,18	11783,8	0,495	129	1,6	0,0	0,0	0,0	34,9	3,1	0,0	24,0	0,0	2,3	0,0	5,4	0,8	0,0	4,7	18,6	0,8	0,0	0,8	3,1
				t	5.7	Q6L8G8	8,05	15149,5	0,393	165	0,6	0,0	0,6	0,0	35,8	3,6	0,6	22,4	0,0	0,6	0,0	4,8	0,6	0,0	4,8	21,2	0,0	0,0	0,6	3,6
				t	5.8	O75690	8,24	17519,3	0,282	187	0,5	0,5	0,0	0,5	35,3	4,3	0,0	18,7	0,0	1,6	0,0	5,9	0,5	0,0	5,3	24,1	0,0	0,0	0,5	2,1
				t	5.9	P26371	8,34	16275,9	0,225	169	1,2	2,4	0,0	0,6	35,5	4,1	0,0	16,0	0,0	0,6	0,0	5,3	0,6	0,0	5,9	23,1	0,0	0,0	1,8	3,0
				t	5.10	Q6L8G5	8,19	17983,7	0,332	202	0,5	0,0	1,0	0,5	32,7	3,0	0,0	29,2	0,0	0,5	0,0	5,4	0,5	0,0	4,0	18,3	0,0	0,0	0,5	4,0
				t	5.11	Q6L8G4	8,16	14610	0,393	156	0,6	0,0	0,6	0,0	35,3	5,1	0,0	19,2	0,0	1,3	0,0	5,1	0,6	1,3	5,1	22,4	0,0	0,0	0,0	3,2
KAP 6	Rogers et al, 2002	HGT	21q22.1	Cortex	3		7,45	8114,2	-0,1767	78,7	0,0	3,4	1,6	1,2	14,2	0,0	0,5	36,4	1,7	0,0	4,6	0,0	1,3	3,3	0,5	8,3	0,5	0,0	22,0	0,3
				t	6.1	Q3LI64	8,36	7279	-0,148	71	0,0	4,2	1,4	0,0	12,7	0,0	0,0	38,0	0,0	0,0	5,6	0,0	1,4	2,8	1,4	8,5	1,4	0,0	22,5	0,0
				t	6.2	Q3LI66	6,65	6654,2	-0,332	62	0,0	3,2	1,6	1,6	14,5	0,0	1,6	32,3	3,2	0,0	3,2	0,0	1,6	3,2	0,0	9,7	0,0	0,0	24,2	0,0
				t	6.3	Q3LI67	7,35	10409,4	-0,05	103	0,0	2,9	1,9	1,9	15,5	0,0	0,0	38,8	1,9	0,0	4,9	0,0	1,0	3,9	0,0	6,8	0,0	0,0	19,4	1,0
KAP 7	Rogers et al, 2002	HGT	21q22.1	Cortex	1		8,57	9288,3	-0,12	87	1,1	3,4	8,0	0,0	6,9	0,0	0,0	21,8	1,1	2,3	5,7	0,0	1,1	8,0	6,9	11,5	5,7	1,1	12,6	2,3
				t	7.1	Q8IUC3	8,57	9288,3	-0,12	87	1,1	3,4	8,0	0,0	6,9	0,0	0,0	21,8	1,1	2,3	5,7	0,0	1,1	8,0	6,9	11,5	5,7	1,1	12,6	2,3
KAP 8	Rogers et al, 2002	HGT	21q22.1	Matrix	1		7,6	6825,6	0,01	63	4,8	3,2	3,2	1,6	6,3	0,0	0,0	23,8	0,0	0,0	4,8	0,0	1,6	7,9	7,9	7,9	1,6	1,6	19,0	4,8
				t	8.1	Q8IUC2	7,6	6825,6	0,01	63	4,8	3,2	3,2	1,6	6,3	0,0	0,0	23,8	0,0	0,0	4,8	0,0	1,6	7,9	7,9	7,9	1,6	1,6	19,0	4,8
KAP 9	Rogers et al, 2001	UHS	17q21.2	Cortex	7?		8,04	17032,2	0,07714	160	1,5	4,0	1,6	0,1	32,4	7,8	0,5	3,1	1,3	0,5	1,3	0,4	0,0	0,6	10,9	12,7	14,9	0,6	2,0	3,6
				homology	9.1	A8MXZ3	8,05	26332,8	0,054																					
				p	9.2	Q9BYQ4	8,15	18287,3	0,06	173	1,7	4,6	1,7	0,0	32,9	7,5	0,6	3,5	0,6	0,0	1,2	0,6	0,0	0,6	10,4	12,1	16,2	0,6	1,7	3,5
				p	9.3	Q9BYQ3	7,93	16853,6	0,082	158	1,3	3,8	1,9	0,0	32,9	8,2	0,6	2,5	1,9	0,6	1,3	0,0	0,0	0,6	11,4	12,0	14,6	0,6	1,9	3,8
				t	9.4	Q9BYQ2	7,95	16454,2	0,062	153	1,3	3,9	2,0	0,7	32,7	7,8	0,7	2,0	0,7	0,0	1,3	0,7	0,0	1,3	11,8	11,8	15,0	0,7	2,0	3,9
				homology	9.6	A8MVA2	7,95	16800,4	0,098	159	1,3	3,1	1,3	0,0	31,4	8,2	0,0	4,4	2,5	1,3	1,3	0,0	0,0	0,0	10,1	14,5	13,2	0,6	3,1	3,8
				homology	9.7	A8MTY7	8,01	17792,7	0,102	168	1,8	3,6	1,2	0,0	32,7	7,7	0,6	3,0	1,8	1,2	1,2	0,6	0,0	0,6	11,3	12,5	14,9	0,6	1,8	3,0
				t	9.8	Q9BYQ0	8,04	16771,5	0,076	158	1,3	3,8	1,9	0,0	32,3	8,2	0,6	3,2	1,3	0,6	1,3	0,6	0,0	0,6	11,4	12,7	14,6	0,6	1,3	3,8
				t	9.9	Q9BYP9	8,22	16265,9	0,06	153	2,0	5,2	1,3	0,0	32,0	7,2	0,7	3,3	0,7	0,0	2,0	0,7	0,0	0,7	9,8	13,1	15,7	0,7	2,0	3,3
KAP 10	Rogers et al, 2004;	UHS	21q22.3	Cuticule	12?		7,49	29708	0,33033	293	6,6	2,6	0,1	1,5	26,1	7,3	1,4	1,8	0,3	0,8	2,9	1,7	0,8	0,2	11,4	20,6	4,7	0,4	0,7	8,2
				t	10.1	P60331	7,77	28660,4	0,297	282	8,2	2,8	0,0	1,4	26,6	7,4	1,4	0,7	0,7	0,4	2,8	1,8	1,1	0,0	11,7	20,2	4,6	0,4	0,4	7,4
				t	10.2	P60368	7,52	25615,7	0,34	255	7,1	1,6	0,4	0,8	27,1	8,6	1,6	2,0	0,0	0,8	2,0	2,0	0,8	0,4	10,2	22,4	4,3	0,4	0,0	7,8
				p	10.3	P60369	7,99	22348,2	0,382	221	7,2	3,2	0,0	0,9	25,8	5,9	1,4	1,8	0,0	0,9	3,6	1,8	0,9	0,0	12,2	19,9	5,9	0,5	0,5	7,7
				t	10.4	P60372	7,12	40429	0,349	401	5,0	1,5	0,0	1,7	27,4	9,0	1,2	2,2	0,0	1,0	2,5	2,0	0,2	0,0	11,5	21,4	3,5	0,2	0,2	9,2
				t	10.5	P60370	7,5	27626,2	0,348	271	7,0	2,6	0,0	1,8	26,9	6,6	1,5	1,1	0,4	1,1	3,0	1,8	0,7	0,4	11,8	19,6	5,2	0,4	0,7	7,4
				t	10.6	P60371	6	36781,4	0,31	365	5,5	1,6	0,0	1,9	26,6	8,2	1,4	2,5	0,5	0,3	2,7	1,4	0,8	0,0	10,4	22,2	4,7	0,3	0,5	8,5
				t	10.7	P60409	7,16	37815,8	0,301	375	5,6	2,1	0,0	2,1	26,7	8,0	1,1	2,7	0,0	0,8	2,4	1,6	0,8	0,0	11,7	20,8	4,5	0,3	0,5	8,3
				t	10.8	P60410	7,82	26298,7	0,393	259	8,1	3,1	0,8	1,5	25,5	6,2	1,2	1,9	0,8	1,5	3,1	1,5	0,8	0,8	10,8	19,3	4,6	0,4	0,4	7,7
				p	10.9	P60411	7,86	29976,8	0,254	292	6,8	3,1	0,0	1,4	25,0	7,9	1,4	1,7	0,0	0,7	2,4	1,7	0,7	0,3	12,0	19,5	4,5	0,3	2,4	8,2
				p	10.10	P60014	7,35	25570,6	0,345	251	6,4	2,0	0,4	1,2	24,3	6,8	1,6	1,6	0,4	0,4	2,8	1,6	0,8	0,4	11,2	20,3	6,4	0,4	1,6	9,6
				p	10.11	P60412	7,84	30266,2	0,347	298	8,1	2,7	0,0	1,3	25,8	6,7	1,3	1,0	0,3	0,3	2,7	2,3	0,7	0,0	10,4	21,8	4,4	0,3	1,3	8,4
				p	10.12	P60413	7,94	25107,4	0,298	245	4,1	4,5	0,0	2,0	25,7	6,5	1,2	2,9	0,0	0,8	4,5	1,2	0,8	0,0	12,7	19,6	4,1	0,4	0,4	8,6
KAP 11	Rogers et al, 2004	HS	21q22.1	Matrix	1		8,32	17085,4	-0,007	163	2,5	6,1	2,5	1,8	14,1	7,4	1,2	8,0	0,6	3,7	3,7	0,0	0,6	1,8	7,4	14,7	12,9	0,6	1,8	8,6
				p	11.1	Q8IUC1	8,32	17085,4	-0,007	163	2,5	6,1	2,5	1,8	14,1	7,4	1,2	8,0	0,6	3,7	3,7	0,0	0,6	1,8	7,4	14,7	12,9	0,6	1,8	8,6
KAP 12	Rogers et al, 2004	UHS	21q22.3	Cuticule	4		8,07	11456,3	0,58325	113	5,7	2,5	0,0	0,0	22,3	5,3	0,2	3,1	1,4	2,4	2,2	0,8	1,6	1,1	14,3	17,8	5,0	0,2	2,1	12,2
	Shibuya et al, 2004			p	12.1	P59990	8,22	9737,4	0,541	96	6,2	3,1	0,0	0,0	21,9	6,2	0,0	3,1	1,0	2,1	1,0	1,0	1,0	1,0	13,5	19,8	3,1	0,0	2,1	13,5
				p	12.2	P59991	8,17	14689,3	0,719	146	6,8	2,1	0,0	0,0	21,2	6,2	0,0	2,1	0,7	2,1	2,1	1,4	1,4	1,4	13,0	21,9	1,4	0,0	1,4	15,1
				p	12.3	P60328	8,2	9965,8	0,502	96	4,2	3,1	0,0	0,0	22,9	6,2	0,0	1,0	2,1	5,2	1,0	0,0	2,1	0,0	14,6	17,7	7,3	0,0	2,1	10,4
				t	12.4	P60329	7,68	11432,6	0,571	112	5,4	1,8	0,0	0,0	23,2	2,7	0,9	6,2	1,8	0,0	4,5	0,9	1,8	1,8	16,1	11,6	8,0	0,9	2,7	9,8
KAP 13	Rogers et al, 2004	HS	21q22.1	Mx, Cuti, Cx	4?		8,78	18509,4	-0,329	170	1,4	7,2	2,7	0,7	12,4	3,7	1,2	10,7	1,3	0,8	6,7	0,6	0,9	4,5	6,0	22,1	5,5	0,9	9,3	1,7
				t	13.1	Q8IUC0	8,54	18320,2	-0,28	172	1,2	5,8	1,7	0,6	12,2	4,7	1,2	13,4	0,6	0,0	6,4	0,0	0,6	4,1	5,8	22,7	6,4	0,6	9,9	2,3
				p	13.2	Q52LG2	8,72	18726,8	-0,298	175	1,7	7,4	1,7	1,1	12,0	2,9	1,7	11,4	0,0	0,6	6,9	0,6	0,6	3,4	6,3	23,4	6,3	0,6	9,1	2,3
				t	13.3	Q3SY46	8,78	19235,5	-0,419	172	0,0	7,6	4,7	0,6	12,8	2,9	1,2	8,1	4,7	1,2	6,4	0,0	1,7	4,7	6,4	21,5	4,7	1,2	8,1	1,7
				t	13.4	Q3LI77	9,08	17755	-0,319	160	2,5	8,1	2,5	0,6	12,5	4,4	0,6	10,0	0,0	1,2	6,9	1,9	0,6	5,6	5,6	20,6	4,4	1,2	10,0	0,6
KAP 15	Rogers et al, 2004	HS	21q22.1	Cortex	1		8,15	14978,5	-0,31	137	2,2	3,6	7,3	0,7	10,9	5,1	1,5	8,8	0,7	2,2	6,6	0,7	0,7	5,8	5,1	18,2	9,5	0,0	9,5	0,7
				t	15.1	Q3LI76	8,15	14978,5	-0,31	137	2,2	3,6	7,3	0,7	10,9	5,1	1,5	8,8	0,7	2,2	6,6	0,7	0,7	5,8	5,1	18,2	9,5	0,0	9,5	0,7
KAP 16	Rogers et al, 2001	HS	17q21.2	?	1		5,37	53912	0,069	517	5,2	3,9	0,6	1,5	18,4	5,2	4,8	1,7	0,0	2,9	2,7	1,9	0,4	1,0	13,9	18,6	6,0	0,2	2,1	8,9
	Wu et al. 2008				16.1	no ref	5,37	53912	0,069	517	5,2	3,9	0,6	1,5	18,4	5,2	4,8	1,7	0,0	2,9	2,7	1,9	0,4	1,0	13,9	18,6	6,0	0,2	2,1	8,9
KAP 17	Rogers et al, 2001	UHS	17q21.2	Cuticule	1		3,83	9503,8	0,35	105	0,0	0,0	1,0	1,9	36,2	3,8	2,9	29,5	0,0	1,0	0,0	1,0	1,0	1,0	4,8	11,4	3,8	0,0	0,0	1,0
				t	17.1	Q9BYP8	3,83	9503,8	0,35	105	0,0	0,0	1,0	1,9	36,2	3,8	2,9	29,5	0,0	1,0	0,0	1,0	1,0	1,0	4,8	11,4	3,8	0,0	0,0	1,0
KAP 19	Rogers et al, 2004	HGT	21q22.1	Cuticule	7?		8,64	7443,78	-0,2911	70,4	0,2	6,6	0,8	0,7	9,7	0,0	0,8	35,4	0,6	0,5	4,0	0,3	1,5	6,5	1,6	8,9	0,5	0,4	21,0	0,2
				t	19.1	Q8IUB9	8,45	9007,6	-0,251	90	0,0	3,3	1,1	0,0	7,8	0,0	0,0	42,2	1,1	0,0	2,2	0,0	1,1	7,8	1,1	12,2	0,0	0,0	20,0	0,0
				t	19.2	Q3LHN2	8,48	5736,5	-0,25	52	0,0	11,5	0,0	1,9	23,1	0,0	1,9	26,9	1,9	0,0	1,9	0,0	1,9	5,8	1,9	3,8	1,9	0,0	15,4	0,0
				t	19.3	Q7Z4W3	8,57	8246,9	-0,217	81	0,0	3,7	0,0	0,0	6,2	0,0	0,0	43,2	0,0	0,0	3,7	0,0	1,2	7,4	1,2	9,9	0,0	0,0	23,5	0,0
				t	19.4	Q3LI73	8,93	9106,2	-0,23	84	0,0	7,1	2,4	1,2	7,1	0,0	2,4	27,4	0,0	3,6	7,1	2,4	1,2	6,0	2,4	9,5	2,4	0,0	16,7	1,2
				p	19.5	Q3LI72	8,32	7624,2	-0,467	72	0,0	5,6	2,8	2,8	5,6	0,0	0,0	38,9	0,0	0,0	4,2	0,0	1,4	6,9	1,4	6,9	0,0	0,0	23,6	0,0
				t	19.6	Q3LI70	8,97	6266,9	-0,478	58	0,0	10,3	0,0	0,0	10,3	0,0	1,7	36,2	0,0	0,0	3,4	0,0	1,7	5,2	1,7	6,9	0,0	0,0	22,4	0,0
				t	19.7	Q3SYF9	8,49	6644,3	-0,198	63	0,0	4,8	0,0	0,0	11,1	0,0	0,0	33,3	1,6	0,0	4,8	0,0	1,6	4,8	1,6	14,3	0,0	1,6	20,6	0,0
				t	19.8	Q3LI54	8,91	6917,6	-0,238	63	1,6	6,3	0,0	0,0	6,3	0,0	0,0	34,9	0,0	0,0	4,8	0,0	1,6	7,9	1,6	7,9	0,0	1,6	25,4	0,0
KAP 20	Rogers et al, 2004	HGT	21q22.1	Cortex	2?		8,39	5674,03	-0,1965	52,3	3,4	5,9	1,9	1,1	9,3	0,0	0,6	28,4	2,1	1,6	4,4	0,6	2,5	2,0	0,8	7,6	1,7	1,4	21,6	3,2
				t	20.1	Q3LI63	8,24	6201,7	-0,479	56	0,0	3,6	3,6	0,0	8,9	0,0	0,0	35,7	0,0	1,8	1,8	0,0	1,8	1,8	1,8	5,4	0,0	1,8	32,1	0,0
				t	20.2	Q3LI61	8,63	6960,7	-0,303	65	0,0	6,2	1,5	0,0	12,3	0,0	0,0	36,9	1,5	0,0	4,6	0,0	1,5	1,5	1,5	4,6	0,0	1,5	24,6	1,5
				homology	20.3	Q3LI60	7,61	4908,5	-0,077	44	2,3	4,5	2,3	4,5	9,1	0,0	0,0	25,0	0,0	4,5	4,5	2,3	2,3	4,5	0,0	6,8	2,3	2,3	20,5	2,3
				homology	putative20.4	Q3LI62	9,06	4625,2	0,073	44	11,4	9,1	0,0	0,0	6,8	0,0	2,3	15,9	6,8	0,0	6,8	0,0	4,5	0,0	0,0	13,6	4,5	0,0	9,1	9,1
KAP 21	Rogers et al, 2004	HGT	21q22.1	Cortex	2?		8,03	7647,1	-0,182	73,3	0,6	3,3	3,4	1,1	18,7	0,0	0,6	27,4	2,2	1,7	1,0	2,3	1,4	3,6	0,8	11,7	1,8	0,4	16,9	1,1
				t	21.1	Q3LI58	8,15	7936,6	-0,196	79	0,0	3,8	2,5	0,0	19,0	0,0	0,0	35,4	1,3	0,0	0,0	0,0	1,3	3,8	1,3	11,4	2,5	0,0	17,7	0,0
				t	21.2	Q3LI59	8,43	8548,4	-0,305	83	0,0	6,0	2,4	0,0	21,7	0,0	0,0	31,3	0,0	0,0	1,2	0,0	1,2	0,0	1,2	13,3	1,2	1,2	19,3	0,0
				homology	21.3	Q3LHN1	7,51	6456,3	-0,045	58	1,7	0,0	5,2	3,4	15,5	0,0	1,7	15,5	5,2	5,2	1,7	6,9	1,7	6,9	0,0	10,3	1,7	0,0	13,8	3,4
KAP 22	Rogers et al, 2004	HGT	21q22.1	?	1?		7,53	5258,3	-0,4745	46,5	4,3	3,3	5,5	2,2	10,7	1,1	2,2	18,1	2,2	0,0	5,5	3,3	2,2	6,5	3,3	10,8	0,0	2,1	17,3	0,0
				t	22.1	Q3MIV0	6,67	5298,8	-0,427	48	4,2	2,1	4,2	2,1	12,5	2,1	2,1	22,9	2,1	0,0	4,2	2,1	2,1	6,2	2,1	10,4	0,0	4,2	14,6	0,0
				homology	22.2	Q3LI68	8,39	5217,8	-0,522	45	4,4	4,4	6,7	2,2	8,9	0,0	2,2	13,3	2,2	0,0	6,7	4,4	2,2	6,7	4,4	11,1	0,0	0,0	20,0	0,0
KAP 23	Rogers et al, 2004	HS	21q22.1	Cx and Cuti	1		5,26	6891,6	-0,102	65	1,5	1,5	6,2	1,5	7,7	3,1	3,1	13,8	3,1	1,5	13,8	0,0	1,5	3,1	6,2	20,0	1,5	0,0	9,2	1,5
				t	23.1	A1A580	5,26	6891,6	-0,102	65	1,5	1,5	6,2	1,5	7,7	3,1	3,1	13,8	3,1	1,5	13,8	0,0	1,5	3,1	6,2	20,0	1,5	0,0	9,2	1,5
KAP 24	Rogers et al, 2007		21q22.1	Cuticule	1		8,71	27719,1	-0,364	254	2,0	4,3	5,5	1,2	9,1	3,9	1,6	5,1	2,0	2,0	7,9	2,4	0,8	2,8	9,8	19,3	8,7	0,4	7,5	3,9
				p	24.1	Q3LI83	8,71	27719,1	-0,364	254	2,0	4,3	5,5	1,2	9,1	3,9	1,6	5,1	2,0	2,0	7,9	2,4	0,8	2,8	9,8	19,3	8,7	0,4	7,5	3,9
KAP 25	Obayashi et al. 2002		21q22.1	?	1		9,01	11738,1	-0,234	102	1,0	5,9	5,9	1,0	6,9	7,8	0,0	4,9	4,9	3,9	3,9	0,0	2,0	12,7	5,9	20,6	2,9	1,0	4,9	3,9
				t	25.1	Q3LHN0	9,01	11738,1	-0,234	102	1,0	5,9	5,9	1,0	6,9	7,8	0,0	4,9	4,9	3,9	3,9	0,0	2,0	12,7	5,9	20,6	2,9	1,0	4,9	3,9
KAP 26	Obayashi et al. 2002		21q22.1	Cuticule	1		8,03	22554,4	-0,221	210	1,9	5,2	3,8	2,4	10,0	5,7	1,9	6,2	2,4	1,4	10,0	0,5	0,5	2,9	11,0	18,1	6,7	0,5	3,8	5,2
	Rogers et al, 2008			p	26.1	Q6PEX3	8,03	22554,4	-0,221	210	1,9	5,2	3,8	2,4	10,0	5,7	1,9	6,2	2,4	1,4	10,0	0,5	0,5	2,9	11,0	18,1	6,7	0,5	3,8	5,2
KAP 27	Obayashi et al. 2002		21q22.1	?	1		5,45	22344,7	-0,562	207	4,3	3,4	5,3	2,4	9,2	8,7	6,3	3,9	2,9	1,0	6,3	2,4	1,4	4,3	8,7	16,4	8,2	0,0	0,5	4,3
				t	27.1	Q3LI81	5,45	22344,7	-0,562	207	4,3	3,4	5,3	2,4	9,2	8,7	6,3	3,9	2,9	1,0	6,3	2,4	1,4	4,3	8,7	16,4	8,2	0,0	0,5	4,3
KAP 28	Wu et al. 2008		2q36.3	?	8		8,41	9487,7	0,32188	101	0,0	5,6	0,0	0,0	38,4	4,4	0,0	30,4	0,2	1,0	0,0	3,4	1,0	0,0	2,0	6,1	4,0	0,0	1,4	2,2
				homology	28.1	no ref	8,46	8607,2	0,311	88	0,0	6,8	0,0	0,0	39,8	4,5	0,0	25,0	0,0	1,1	0,0	3,4	1,1	0,0	2,3	6,8	4,5	0,0	2,3	2,3
				homology	28.2	no ref	8,34	10821,6	0,323	120	0,0	5,0	0,0	0,0	36,7	3,3	0,0	37,5	0,8	0,8	0,0	2,5	0,8	0,0	1,7	5,0	3,3	0,0	0,8	1,7
				homology	28.3	no ref	8,47	9511,3	0,356	100	0,0	7,0	0,0	0,0	39,0	4,0	0,0	29,0	0,0	1,0	0,0	3,0	1,0	0,0	2,0	6,0	4,0	0,0	1,0	3,0
				homology	28.4	no ref	8,41	9791,5	0,303	105	0,0	5,7	0,0	0,0	37,1	3,8	0,0	32,4	0,0	1,0	0,0	2,9	1,0	0,0	1,9	6,7	3,8	0,0	1,9	1,9
				homology	28.5	no ref	8,39	8200,8	0,4	85	0,0	5,9	0,0	0,0	41,2	4,7	0,0	25,9	0,0	1,2	0,0	3,5	1,2	0,0	2,4	5,9	4,7	0,0	1,2	2,4
				homology	28.6	no ref	8,45	9895,7	0,259	105	0,0	4,8	0,0	0,0	38,1	4,8	0,0	30,5	1,0	1,0	0,0	4,8	1,0	0,0	1,9	5,7	3,8	0,0	1,0	1,9
				homology	28.7	no ref	8,36	8950,6	0,365	96	0,0	5,2	0,0	0,0	38,5	4,2	0,0	31,2	0,0	1,0	0,0	3,1	1,0	0,0	2,1	6,2	4,2	0,0	1,0	2,1
				homology	28.8	no ref	8,39	10122,9	0,258	108	0,0	4,6	0,0	0,0	37,0	5,6	0,0	31,5	0,0	0,9	0,0	3,7	0,9	0,0	1,9	6,5	3,7	0,0	1,9	1,9
KAP 29	Wu et al. 2008		17q21.2		1		5,89	35211	-0,051	341	5,9	0,9	1,8	1,8	17,3	8,5	2,3	6,2	1,5	2,3	3,2	2,3	1,5	1,8	11,4	15,5	7,9	0,6	2,3	5,0
				homology	29.1	no ref	5,89	35211	-0,051	341	5,9	0,9	1,8	1,8	17,3	8,5	2,3	6,2	1,5	2,3	3,2	2,3	1,5	1,8	11,4	15,5	7,9	0,6	2,3	5,0
KAP family	Ref	Category	Chromosomal location	Expression	Family members	SwissProt Acess Number	pI	MW	GRAVY	nAA	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

Tableau 1 *: Fiche analytique des gènes décrits des KAP et propriétés physico-chimiques prédites des séquences primaires protéiques. Abréviations : t, exprimé au niveau du transcriptome ; p exprimé au niveau du protéome.*

Les KAP riches en soufre du cluster 2q36.3

La famille multigénique des **KAP 28**, est seule sur ce cluster. Elles ont en commun avec les KAP 5 une très grande abondance en cystéine et en glycine (respectivement 38,4% et 30,4%) mais possèdent moins de sérine (6%). Elles sont petites en taille (9,5 kDa en moyenne) et leur expression n'a pas été étudiée. La découverte de ce cluster de gènes potentiels est récente et a été obtenue par homologie avec d'autres mammifères qui possèdent tous des séquences orthologues dans leur génome [138].

Les KAP riches en soufre du cluster 21q22.1

Il est possible de distinguer un groupe composé des **KAP 13, 15.1 et 23.1** caractérisé par une forte proportion en sérine (environ 20%) et comportant environ 10% de cystéine, de glycine et de tyrosine. Leurs tailles sont relativement diffèrentes (respectivement 18,5 kDa, 15 kDa et 7 kDa). Il a été discuté de les considérer comme une unique famille parmi les KAP 13 [138].

La **KAP 11.1** se distingue du précédent groupe par une proportion moins importante de sérine (14,7%), plus de thréonine (13%) et très peu de tyrosine. Elle mesure 17 kDa.

Les transcrits correspondants à ces deux groupes s'expriment dans le cortex du niveau critique jusqu'à la zone d'élongation et sont exprimés également dans la cuticule au niveau de la zone d'élongation pour un individu de la famille 13 et les KAP 15 et 23.

Les **KAP 24.1 et 26.1** ont des compositions en acides aminés assez proches des deux groupes précédents mais sont plus grandes (respectivement 27,7 kDa et et 22,5 kDa). Leur expression protéique a été montrée dans la cuticule.

Les **KAP 25.1 et 27.1** ont également des compositions en acides aminés voisines des précédents groupes mais leur expression n'a pas été détectée chez l'humain.

Les KAP riches en glycine et tyrosine du cluster 21q22.1

En plus des gènes des KAP riches en soufre, se retrouve sur le cluster 21q22.1 l'ensemble des gènes correspondants à des protéines riches en glycine et tyrosine. Les **KAP 6, 7.1, 8.1, 19, 20, 21 et 22** ont en commun leur petite taille (entre 5 kDa et 9 kDa) et leur forte composition en glycine (de 20 à 35%) et en tyrosine (de 13 à 22%).

Les transcrits des KAP 8.1 et 21.1 se retrouvent exprimés entre le niveau critique et la fin de la zone de la zone d'élongation. Le groupe KAP 22 n'a pas été étudié. Les autres transcrits des autres individus du groupe sont exprimés dans les zones de pré-élongation et d'élongation.

Les KAP riches en soufre du cluster 21q22.3

Sur un cluster distinct de celui décrit précédemment se retrouvent les gènes des **KAP 10 et 12**. Ces deux familles ont en commun une forte proportion en cystéine (26 et 22%) et en serine et thréonine (32%). La famille 10 est plus grande (30 kDa en moyenne) que la famille 12 (11,5 kDa).

5. De l'expression des gènes à la structure finale du cheveu : la formation des structures corticales et cuticulaires

Dans le chapitre précédent, nous avons présenté les différentes étapes de différenciation des cellules du follicule pileux et décrits les ultrastructures pouvant être observées dans les cellules en fin de différenciation

Dans ce chapitre, nous présentons l'évolution des nano et des microstructures au sein des cellules corticales et cuticulaires. Nous avons mis en parallèle les données obtenues par des observations microscopiques et de diffraction X [135, 141] avec les données récentes d'expression des gènes des familles des protéines décrites précédemment. Ces données peuvent être elles-mêmes interprétées à la lumière des connaissances récentes sur

les arrangements des filaments intermédiaires décrits en partie 3 de ce chapitre. A notre connaissance, la confrontation de ces sources bibliographiques n'a par encore été réalisée dans la littérature.

a) Les étapes de l'assemblage macrofibrillaire du cortex

Les premières protéines à être exprimées après le début de la différenciation des cellules corticales au dessus du niveau critique sont les kératines K35 et K85 (Figure 14). Cette expression coïncide avec le début de la formation des filaments mais les ULF ne sont pas encore observées dans cette zone. Sachant l'aptitude de ces protéines à former des hétérodimères, il est possible d'envisager dans cette zone la formation d'hétérodimères K35/K85.
L'interruption de l'expression de K35 à la limite de la matrice coïncide avec le début de l'expression de K31. Ainsi, seule la formation d'hétérodimères K31/K85 peut être réalisée dans la zone de pré-élongation. Dans cette zone, l'expression des gènes correspondants à des individus de KAP du cluster 21q22.1 débute. Les gènes des KAP soufrés 11.1, de la famille 13, 15.1 et 23.1 et les HGT KAP 8.1 et 21.2 sont décrits pour y être exprimés [73]. Les premiers filaments intermédiaires, sans orientation particulière, peuvent être visualisés et forment les ULF qui s'allongent et s'organisent peu à peu pour former les premiers éléments de macrofibrilles [108]. Les filaments constitués sont creux, le centre étant de densité voisine de l'extérieur et ne sont pas encore empilés [135].
En début de zone d'élongation, l'expression de K85 s'interrompt. Les gènes des kératines de type II, K86, K81, K83 et de type I, K33a, K33b, K34 et K36 sont exprimées tout comme ceux des KAP des familles riches et très riches en soufre 1, 2, 3, 4 et 9 et des KAP riches en glycine et tyrosine 6.1, 7.1, 19 et 20 [3, 73]. Ces expressions vont de concert avec la disposition des filaments parallèle à l'axe d'élongation des cellules corticales et avec la continuité de la croissance des macrofibrilles [135].

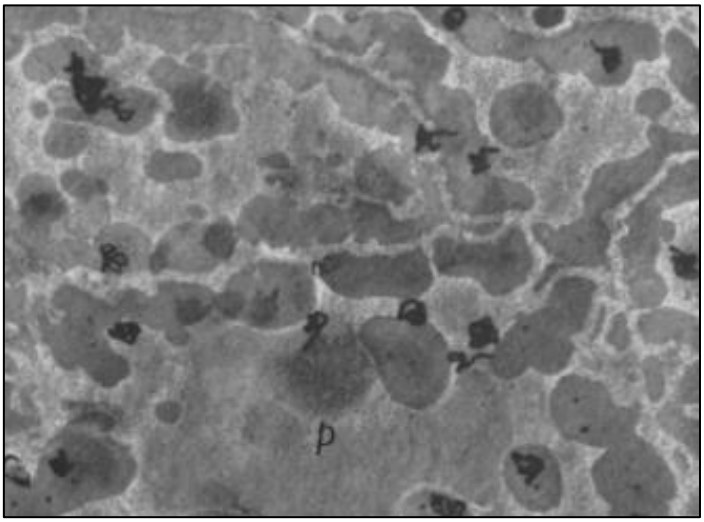


***Figure 12** : Visualisation autoradiographique d'une coupe de follicule de laine dans la zone d'élongation du cortex. De la cystéine marquée [^{35}S] est incorporée in vivo 5 heures avant l'observation. Les protéines très riches en cystéines sont visualisées sous formes de filaments plus sombres s'insérant au sein des macrofibrilles [45].*

Les observations en microscopie électronique utilisent du tétroxyde d'osmium comme agent de contraste. Ce composé a la propriété de se fixer par oxydation aux fonctions thiols des résidus cystéine et permet de visualiser l'expression des KAP riches en soufre pendant ces phases. Il apparait qu'à partir de cette étape, les HS KAP entourent les filaments intermédiaires constituant les macrofibrilles en croissance. Par ailleurs, des expériences d'autoradiographie dans la laine montrent que ces protéines se présentent sous formes de filaments incorporées aléatoirement au sein des macrofibrilles en formation (Figure 12) [45].

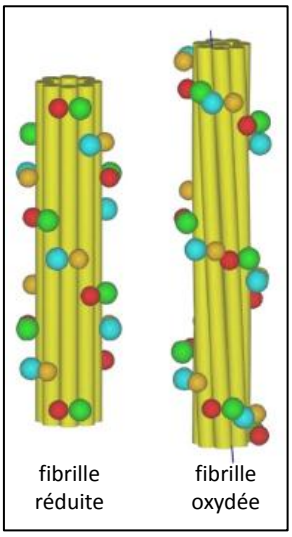


Figure 13 *: Modèle proposé par Parry et al. [133] pour la transition des protofilaments de la microfibrille entre le mode réduit et le mode oxydé.*

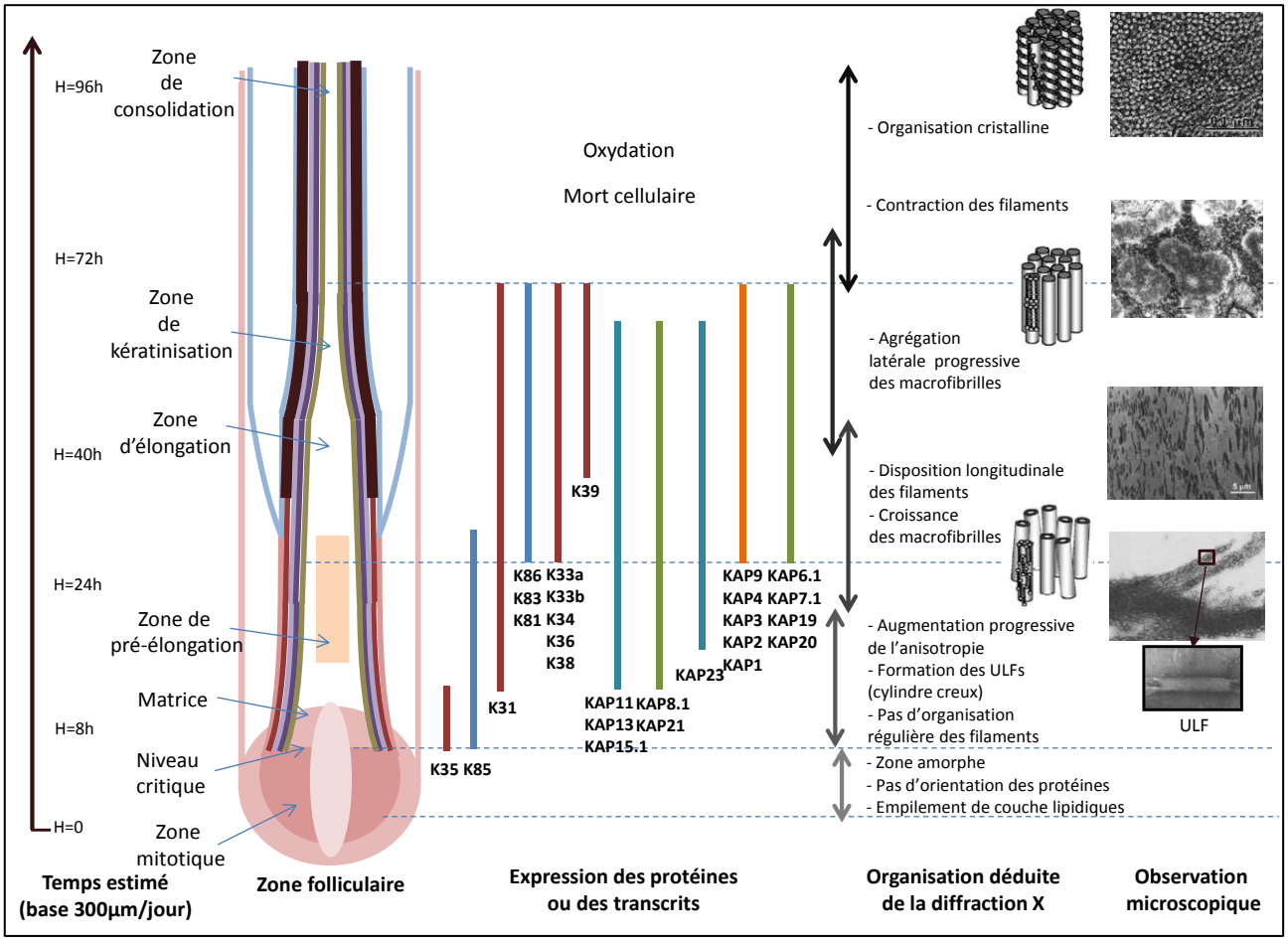


Figure 14 *: Confrontation des données de la littérature relative à la différenciation des cellules corticales et à la constitution des macrofibrilles. Ces données comprennent de gauche à droite les connaissances des zones de différenciation déduites de la microscopie, les informations d'expression des kératines au niveau du protéome [101, 142] et des ARNm des KAP au niveau du transcriptome [73, 78, 80, 81], les informations de diffraction des rayons X [135] et les visualisations par microscopie électronique de la croissance des macrofibrilles [143-145].*

L'expression de ces protéines se poursuit pendant toute la phase de kératinisation jusqu'à ce que l'espace intermacrofibrillaire (le cytoplasme) ait quasiment disparu. Les structures ne sont alors plus en solution.

L'expression protéique s'interrompt avec la mort de la cellule qui est totalement kératinisée. Les macrofibrilles sont alors totalement agrégées.

Pendant la phase de consolidation, le tétroxyde d'osmium contraste intensément sur une zone de plus de 300 µm de long indiquant que les cystéines ne sont pas encore oxydées (Figure 1 du Chapitre 2). Le contraste diminue jusqu'à disparaitre indiquant l'oxydation progressive des cystéines. Pendant cette phase d'oxydation, les filaments adoptent leur arrangement final en filaments intermédiaires oxydés contractés en un empilement des tétramères (7+1). Ce modèle d'arrangement propose un rapprochement des segments têtes et queues des filaments qui se disposent en hélice autour de la microfibrille [133].

b) La formation des structures lamellaires cuticulaires

De la même façon que dans le cortex, les premiers gènes à être exprimés au sein de la cuticule sont ceux des kératines K35 et K85 (Figure 15). Une troisième kératine spécifique à la cuticule, K32, s'exprime en même temps dans la matrice. Dans la zone d'élongation, l'expression de K35 s'interrompt et est relayée par l'expression de K82, également spécifique. Bien que des systèmes filamentaires puissent être observés pendant la phase de modification de la géométrie cellulaire, la croissance de systèmes macrofibrillaires semblables à ceux du cortex n'est pas observée.

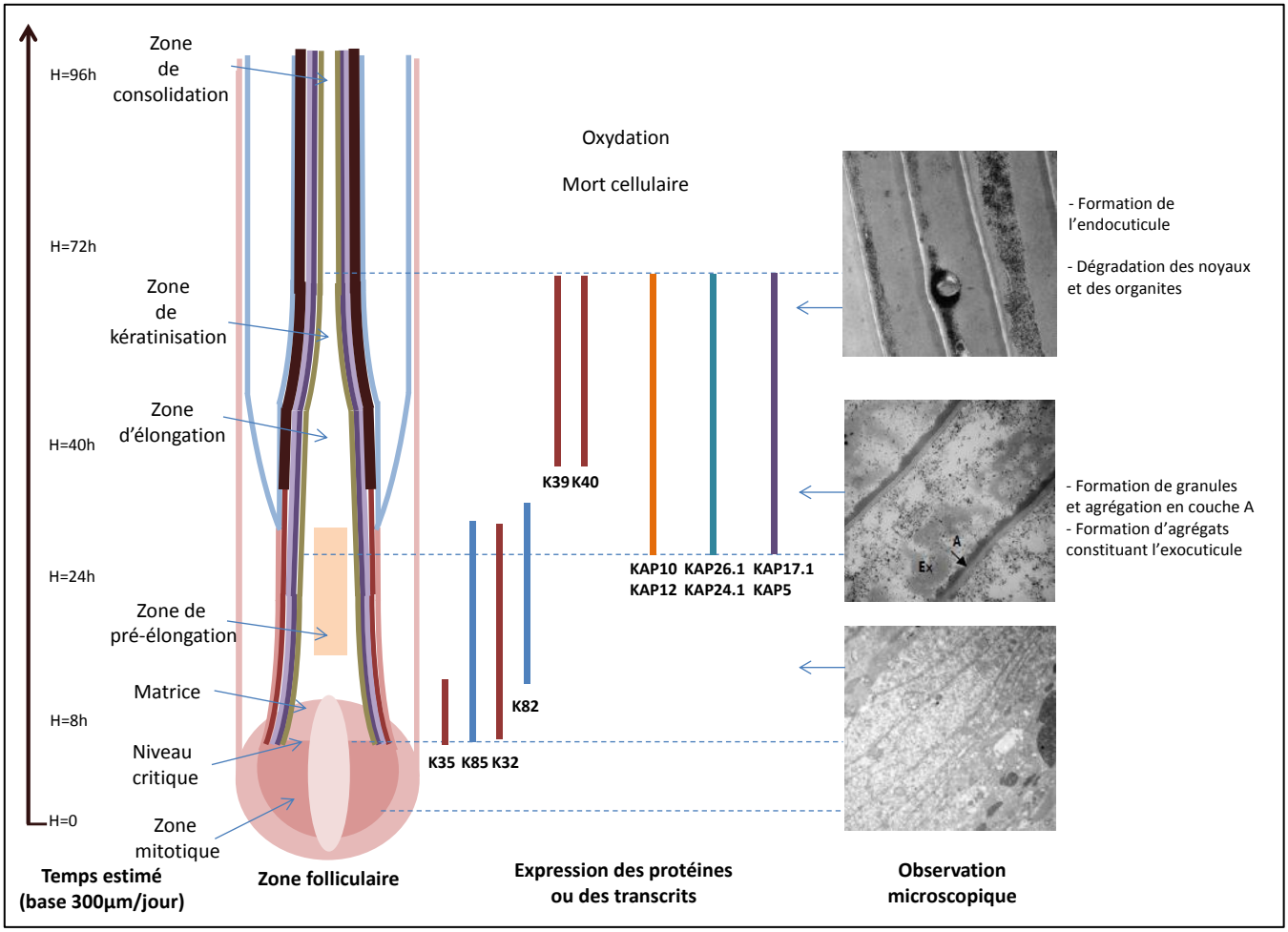


Figure 15 *: Confrontation des données de la littérature relatives à la différenciation de la cuticule. Ces données comprennent les informations relatives à l'expression des kératines [82, 83] et des KAP dans la cuticule [73, 139, 140] ainsi que les données de microscopie électronique (données internes).*

Pendant la phase d'élongation, l'ensemble des gènes correspondants aux KAP de la cuticule (10, 12, 5, 17.1, 24.1 et 26.1) commencent à être exprimés. Cette expression coïncide avec l'apparition de granules de différentes tailles. Les premières à apparaitre sont les plus petites et s'agrègent le long de la paroi périphérique pour constituer la couche A dont la forte affinité avec le tétroxyde d'osmium indique une très forte proportion de groupement thiols. Les deux autres apparaissent plus tard et s'agrègent pour constituer progressivement l'exocuticule pendant toute la phase de kératinisation.
Pendant la phase de kératinisation, les kératines précédentes ne sont plus exprimées. Les kératines K39 et K40 sont alors exprimées jusqu'en fin de kératinisation.
Lorsque l'exocuticule est constituée, les résidus du cytoplasme se condensent et se regroupent sur la paroi interne de la cellule et constituent l'endocuticule.

6. Le réseau des liaisons et des interactions dans le cheveu

Le cheveu est principalement constitué de protéines qui forment un réseau plus ou moins humidifié. Différentes interactions entre les résidus des protéines peuvent être décrites au sein de ce réseau, complété d'interactions existant également entre les cellules. Un ensemble de liaisons, faibles, fortes et covalentes est responsable des propriétés structurales et mécaniques observées dans la fibre mature.

a) Les interactions non covalentes et la solvatation

Les différentes interactions non covalentes pouvant intervenir dans le cheveu peuvent être prépondérantes ou non en fonction des phases de maturation de la fibre capillaire. Il est nécessaire de distinguer ces types de liaisons lorsque les protéines sont en solution dans le cytoplasme (entre le niveau critique et la zone de kératinisation) ou lorsque les protéines sont en interaction les unes avec les autres et libres de solvatation (pendant et après la zone de consolidation).
Lorsque les protéines sont solvatées, les résidus apolaires peuvent s'associer de manière à constituer un réseau hydrophobe permettant de diminuer leur surface de contact avec l'eau. Cet effet est un des principaux moteurs de la formation des hélices de kératines puis des associations dimériques lors de la formation des filaments intermédiaires. Il est en revanche considérablement réduit en l'absence de solvatation dans la structure finale.
Sans effet hydrophobe, la stabilisation des filaments est réalisée par un vaste réseau des liaisons hydrogènes de la chaine peptidique tels que le modèle de l'hélice de Pauling et Corey le décrit (Figure 16) [34, 35].

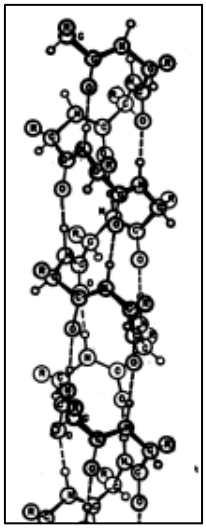


***Figure 16** : Modèle d'hélice-α stabilisée par les liaisons hydrogènes de la chaine peptidique [34].*

En l'absence d'eau, ces liaisons hydrogènes inter et intra protéines sont beaucoup plus fortes qu'en solution. Des interactions entre les résidus donneurs et accepteurs d'hydrogènes peuvent alors être favorisées et contribuer à l'arrangement de la structure kératinisée.
De la même manière, les interactions ioniques pouvant exister entre les résidus chargés en solution sont considérablement renforcées en l'absence de solvant. Les ponts salins existant entre les protéines dans le cytoplasme peuvent ainsi devenir des liaisons possédant un caractère presque covalent après kératinisation et séchage.

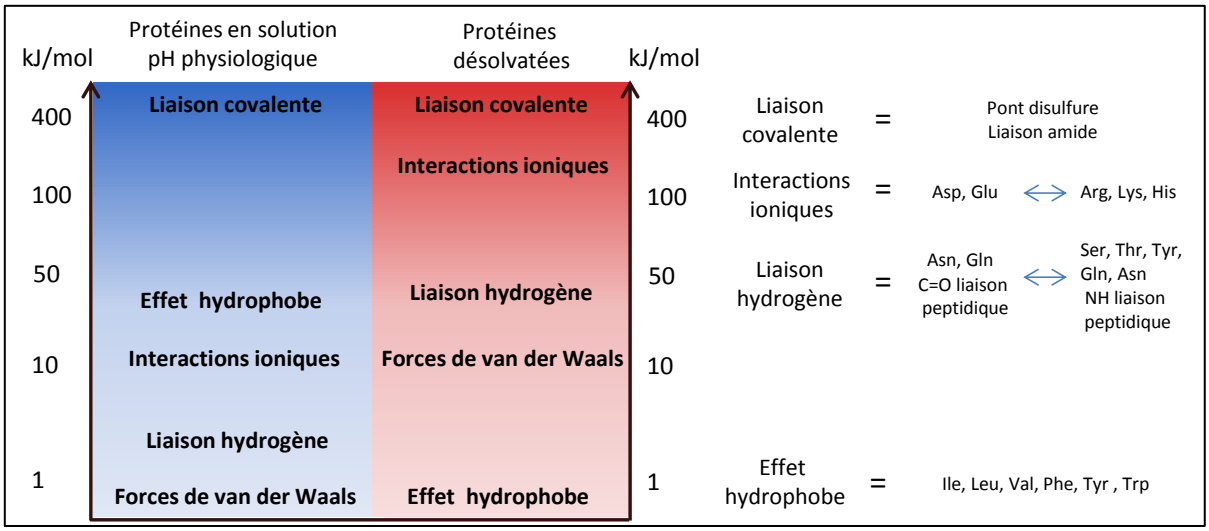


***Figure 17** : Proposition de classification des forces de liaisons existant entre les protéines solvatées et désolvatées.*

A l'inverse, lorsque la fibre est exposée à l'eau, ces interactions non covalentes fortes se retrouvent affaiblies et l'effet hydrophobe peut redevenir prépondérant. Néanmoins, la plus grande malléabilité de la chevelure mouillée suggère qu'il existe beaucoup moins d'interactions non covalentes entre les protéines dans cet état que dans la fibre sèche. Les interactions hydrophobes dans la fibre mouillée sont probablement beaucoup plus faibles que les énergies de liaisons dues aux liaisons ioniques et hydrogènes dans la fibre sèche.

b) Le réseau des ponts disulfures

L'oxydation des cystéines en pont disulfures pendant la phase de consolidation est à l'origine d'un vaste réseau tridimensionnel en grande partie responsable de la stabilité de la fibre au stress mécanique et du maintien de sa cohésion même mouillée. La structure de ce réseau de liaisons covalentes entre les protéines du cheveu n'est pas vraiment établie. Différentes études basées sur les propriétés mécaniques des fibres réduites et étirées permettent de disposer néanmoins de certaines données.
Il est possible de diviser les ponts disulfures au sein de la fibre en deux catégories [146].
La première regroupe deux catégories de **liaisons intermoléculaires**, la première SS_1 étant plus accessible à l'eau que la seconde SS_2. Les ponts SS_1 impliquent 35% du réseau et devraient se retrouver entre les régions têtes et queues des kératines. Néanmoins la quantité de cystéines dans ces régions ne permet pas à elle seule d'expliquer leur abondance et implique que des ponts intermoléculaires accessibles existent également entre les KAP dans la matrice voire entre les filaments et les KAP. Les ponts SS_2 sont moins accessibles à l'eau et sont donc dans des régions plus hydrophobes des structures. Elles représentent 18% des cystéines du réseau et devraient être localisées en partie au sein des segments hélicoïdaux des filaments intermédiaires de kératines. De la même manière que pour les SS_1, il est également possible d'envisager ce type de liaisons entre les KAP.
La seconde catégorie correspond aux liaisons **intramoléculaires** SS_3 et représente 47% des ponts entre les cystéines du cheveu. Ces liaisons intramoléculaires sont localisées dans la matrice interfilamentaire hydrophobe et concernent ainsi spécifiquement les KAP.

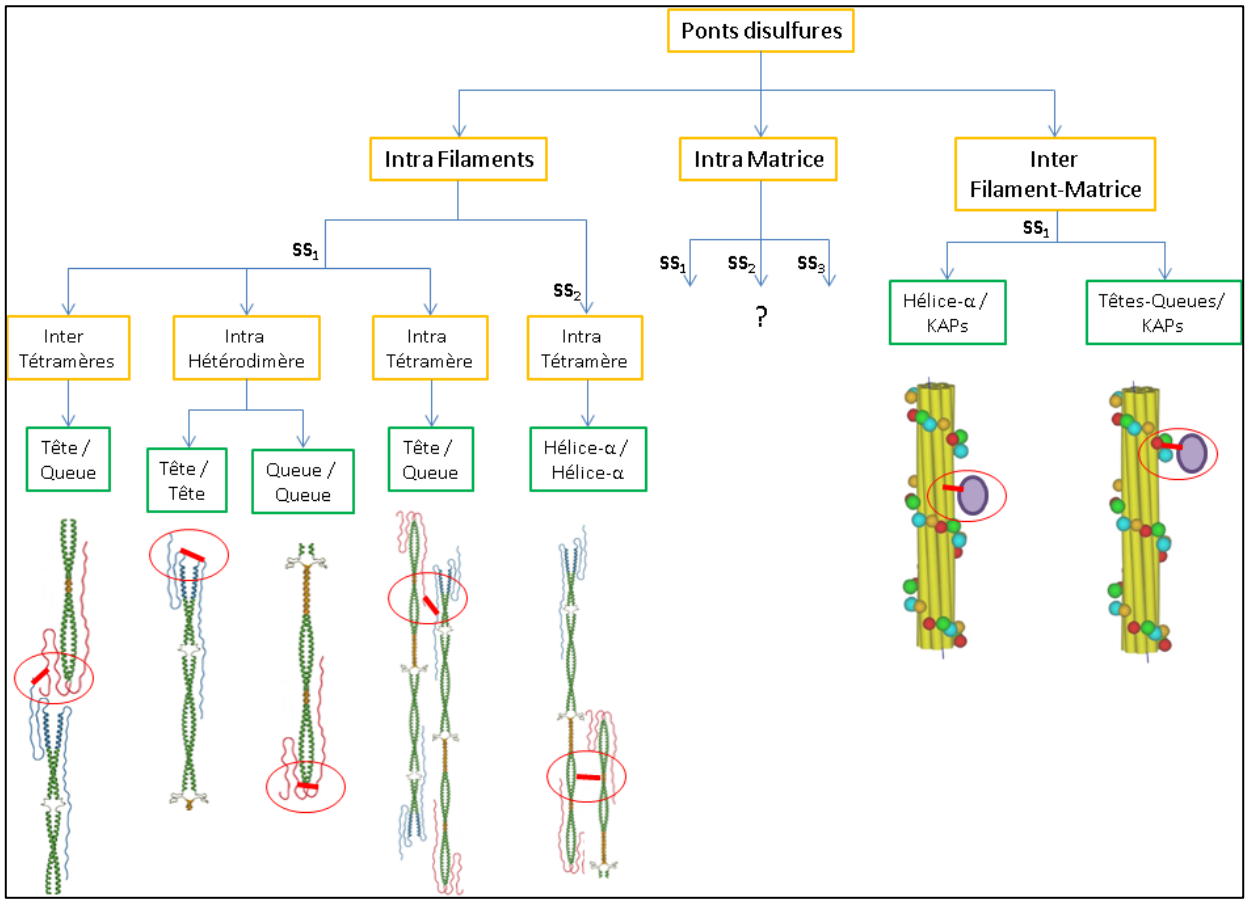


Figure 18 *: Hiérarchisation des principaux types de ponts disulfures pouvant être envisagés au sein des structures macrofibrillaires.*

La position de cystéines au sein des tiges des kératines du cheveu (également retrouvées dans les kératines épithéliales) permet d'y envisager la formation de quelques ponts disulfures qui seraient de type SS_2. Ces liaisons entre les couples de dimères permettent de consolider covalemment les interactions latérales des tétramères des édifices filamentaires et semblent [147].
Les connaissances de l'arrangement des kératines au sein des filaments intermédiaires, suggèrent la possibilité des rapprochements tête/tête et queue/queue au sein des hétérodimères mais également tête/queue entre les hétérodimères. Les proportions en cystéines entre ces segments (de 11 à 19%) permettent d'envisager la formation des ponts SS_1 permettant de consolider les hétérodimères et de les lier longitudinalement. En considérant le modèle d'arrangement de la fibrille oxydée proposé par Parry et décrit chapitre II.5.a., nous pouvons imaginer un réseau de ponts SS_1 au sein de l'hélice constituée par la disposition des segments têtes et queues entourant le filament.

Concernant les KAP et les ponts pouvant exister dans la matrice interfilamentaire, l'analyse des séquences et notamment de la distribution des cystéines des KAP suggère la présence de motifs favorables à la formation de liaisons intramoléculaires [147, 148]. Ces motifs, que nous étudierons plus en détail par la suite, pourraient être à l'origine du réseau de ponts SS_3 abondants dans la structure. L'abondance des liaisons SS_3 serait alors liée au fait que les KAP occupent environ 50% de la structure macrofibrillaire. Néanmoins, la méconnaissance de l'expression et de l'abondance des différentes familles de KAP représentées dans la matrice macrofibrillaire rend difficile l'établissement d'hypothèses concernant la structure des réseaux de ponts disulfures pouvant y exister.
La matrice peut être considérée comme un réseau pseudo élastomérique ayant la capacité de s'étirer et de reprendre sa conformation initiale [149]. Il est possible de déduire de cette propriété que le réseau de ponts disulfures ne contraint pas l'étirement de la fibre et que les protéines de la matrice ont une structure repliée à

l'état initial pouvant être étirée sous la contrainte mécanique et présentent une certaine réticulation (vraisemblablement des liaisons SS_1). La présence de ponts assurant une réticulation clairsemée des KAP est appuyée par le fait que les protéines de la matrice s'extraient relativement plus facilement que les protéines des filaments en milieu réducteur [149].

Dans la cuticule, la structure du réseau de ponts disulfures est plutôt mal appréhendée. Il est admis que l'exocuticule et la couche A contiennent un nombre plus important de ces ponts que dans le cortex. Ce réseau contribue très certainement à la très grande insolubilité des protéines qui constituent l'exocuticule ces dernières étant probablement extrêmement réticulées [50-52, 150].

c) Le réseau covalent des liaisons GGEL

Il existe dans la fibre capillaire un second type de liaisons covalentes permettant de lier deux chaines protéiques par l'intermédiaire de fonctions sur les résidus. Cette liaison est réalisée entre une glutamine et une lysine au moyen d'enzymes, les transglutaminases (TGM), et forme une liaison amide dite ε-(γ-glutamyl) lysine (GGEL) qui ne peut être ouverte que par hydrolyse. Les TGM sont activées *in vivo* par le calcium (Ca^{2+}) [151]. Certaines de ces enzymes sont décrites pour apporter une contribution essentielle à la formation des enveloppes cornifiées des cellules épithéliales et leurs substrats ont été identifiés dans le *stratum corneum*, la couche supérieure de l'épiderme [152, 153]. Dans le follicule, l'expression de différentes TGM est décrite dans les gaines membranaires interne et externe (TGM 1 et 5) mais également dans la fibre (TGM 3) [154, 155].

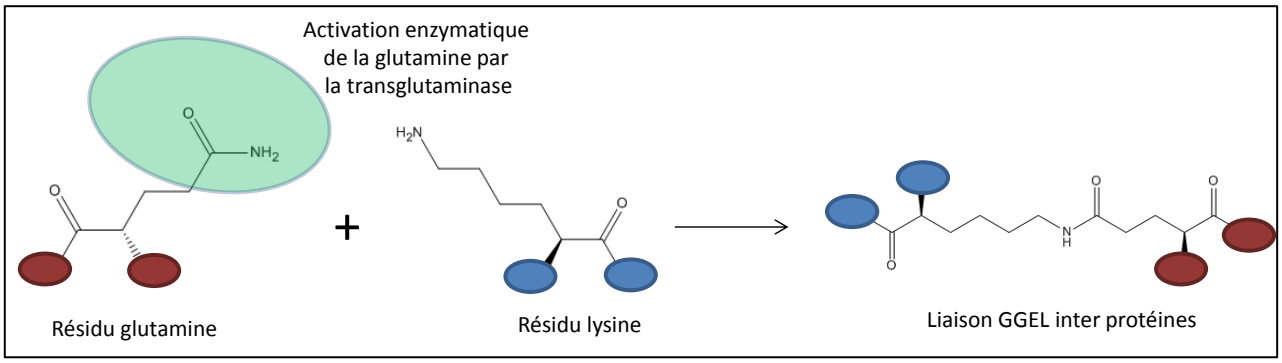


***Figure 19** : Principe de la formation de la liaison GGEL par activation enzymatique.*

A l'inverse de l'épiderme, les substrats de la transglutaminase ne sont pas totalement connus. Néanmoins, les ponts issus de cette activité enzymatique sont retrouvés très abondamment dans la cuticule et dans la médulla dans laquelle la trichohyaline est décrite comme substrat [22, 53-55]. Il est intéressant de noter que de plus fortes concentrations de calcium sont retrouvées dans la cuticule et la médulla des fibres matures [156]. Il est très probable que le réseau GGEL dans la cuticule complète la forte réticulation en pont disulfures des protéines et assure la stabilité de la cuticule même vis-à-vis de traitement réducteurs [50-52, 150].

d) Les structures des interfaces cellulaires

Les interfaces entre les cellules corticales et entre les cellules cuticulaires sont assurées par les complexes de membranes cellulaires. Ces interfaces semblent essentielles à la structure du cheveu puisqu'elles permettent de maintenir les cellules kératinisées cohésives.
Des édifices protéiques transmembranaires peuvent s'insérer au sein de ces structures afin d'y maintenir la cohésion. L'adhésion des cellules épithéliales peut être réalisées au moyen de différents complexes transmembranaires pouvant être reliées aux filaments du cytosquelette [157]. Il est possible d'y distinguer les jonctions adhérentes constituées par les desmosomes. Le desmosome est constitué par l'interaction entre une

protéine transmembranaire (desmocollines ou desmogléines) avec son homologue située dans la membrane adjacente. Le domaine intra cellulaire de la protéine transmembranaire est en interaction avec d'autres protéines intracellulaires constituant la plaque desmosomale interne (desmoplakine, plakoglobine, plakophiline) liée aux filaments intermédiaires [158]. Une étude de localisation de ces protéines dans le follicule a montré que différentes isoformes de desmogléines étaient exprimées pendant la différenciation des cellules folliculaires [158]. La présence de cystéines dans les domaines extra cellulaires des desmogléines suggère la possibilité de formation de ponts disulfures au sein du CMC.
Il est également possible d'envisager d'autres types d'interactions impliquant des composés glycosylés insérés dans les membranes cellulaires et permettant d'établir des liaisons faibles de types van der Waals au sein du CMC. La composition du CMC cuticulaire semble contenir principalement des sucres [105].

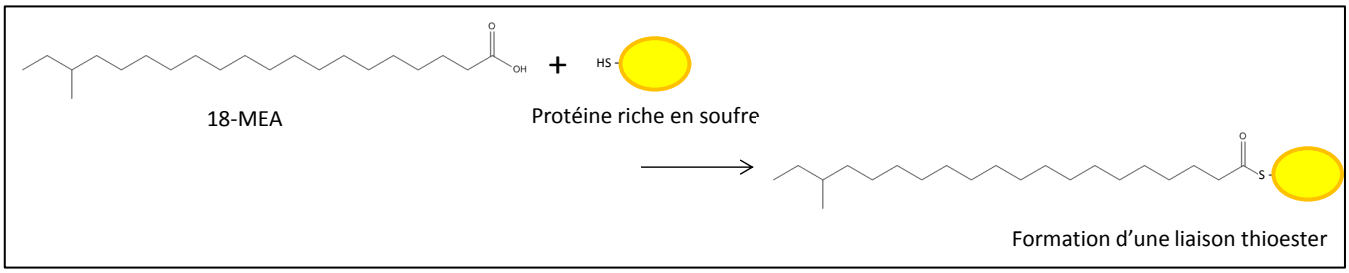


Figure 20 *: Principe de la formation de liaison thioester permettant de lier covalemment un lipide, l'acide 18-methyl-eicosanoique à une protéine riche en soufre.*

Au niveau de la cuticule, la surface externe de la fibre en contact avec l'air est constitué d'une couche de lipide particulier, l'acide 18-méthyl-eicosanoique (18-MEA), liée covalemment par des liaisons thioesters à des protéines constituant l'épicuticule. Cette couche hydrophobe et liée permet d'expliquer en partie la résistance de la fibre à la friction [159]. La nature des protéines de l'épicuticule n'est pas connue bien qu'il soit à peu prêt certain qu'elles contiennent des cystéines nécessaires à l'établissement des liaisons avec les lipides (Figure 20) [160]. Le mécanisme de formation de ce complexe n'a pas été étudié mais il est possible d'imaginer une homologie avec les mécanismes de formation des enveloppes cellulaires cornifiées des cellules épithéliales elles-mêmes constituées d'un complexe lipide-protéine [161].

e) Les modifications induites par des traitements

Les protéines des cheveux sont soumises aux conditions environnementales. Le système étant mort, elles ne bénéficient d'aucun mécanisme de réparation des modifications qui peuvent ainsi être induites sur leurs résidus. La pigmentation des tiges, conséquence de la répartition de granules de mélanine à la périphérie de la fibre, permet de minimiser les interactions entre le rayonnement UV et les résidus des protéines. Néanmoins, un certain nombre de modifications peut être observé suite à la combinaison de l'humidité, de l'oxydation, des rayonnements et éventuellement des traitements cosmétiques (Figure 21).

A la modification des résidus s'ajoutent la possibilité d'hydrolyse de la chaine peptidique des protéines suite à des réactions d'hydrolyse alcaline ou d'oxydation [162].
La photo-oxydation peut se produire sur les résidus aromatiques (tyrosine, tryptophane, phénylalanine, histidine) et va induire des modifications de ces résidus [163, 164] voire même former des liaisons covalentes inter résidus (dityrosine) [165]. Les interactions hydrophobes au sein des kératines impliquant ces résidus peuvent probablement être modifiées. Une conséquence de la photo oxydation est le jaunissement qui peut se produire principalement au sein du cortex des fibres dépourvues de mélanine [163].

Les modifications induisant la formation de liaisons inter protéines sont susceptibles de modifier les propriétés mécaniques de la fibre. L'oxydation des cystines peut également entrainer des coupures irréversibles des ponts disulfures et modifier la structure initiale de l'édifice moléculaire [163, 166, 167].

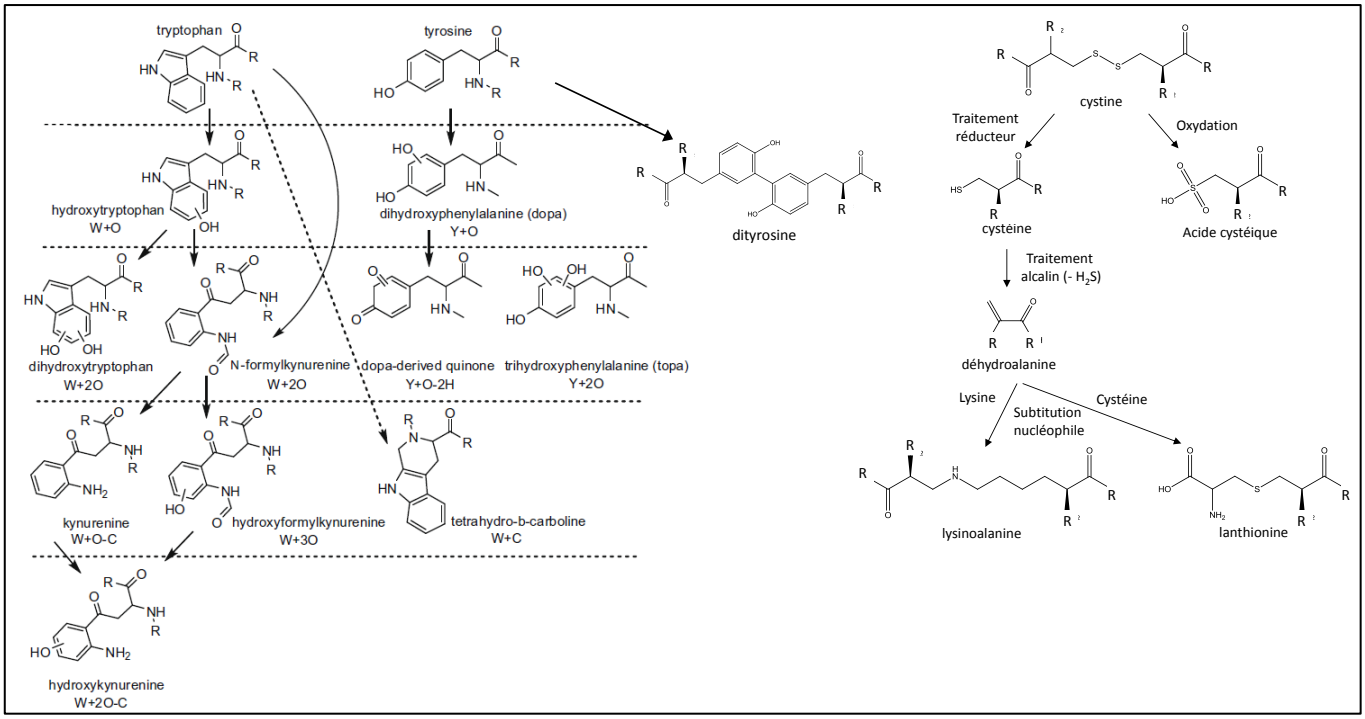


Figure 21 *: A gauche, oxydation des résidus tryptophane et tyrosine lors de l'irradiation UV [163]. A droite, mécanismes chimiques pouvant conduire à la dégradation des cystines [162].*

Les résidus glutamine et asparagine sont relativement abondants au sein des séquences des kératines et des KAP. Les fonctions amides de ces résidus sont relativement fragiles vis-à-vis des traitements et la déamidation, modification couramment rencontrée suite à la manipulation des protéines, peut être imaginée dans le cheveu [168, 169]. L'impact de cette modification, qui a pour conséquence d'augmenter les résidus porteurs de fonction acide carboxylique au sein des séquences, n'a pas été à notre connaissance évalué. Nous pouvons néanmoins imaginer que l'augmentation de l'abondance de résidus acides entraine des interactions ioniques plus fortes que les interactions électrostatiques impliquant les fonctions amides initialement présentes.

7. Les maladies associées aux protéines du cheveu

Différentes maladies génétiques sont décrites comme ayant une conséquence directe sur la structure des cheveux.

La localisation des mutations au sein du gène impliqué dans la maladie met en évidence des sites critiques dans la séquence de la protéine modifiée. La mutation situe alors un site important pour la fonction de la protéine et la conséquence de la mutation renseigne sur le rôle de la protéine dans la structure.

Ce principe est à la base des études sur les souris mutées qui permettent d'établir les rôles de chaque gène dans l'organisme par rapport aux différences de phénotypes observées avec les individus non mutés. Une partie des gènes connus comme impliqués dans la morphogénèse et la croissance folliculaire ainsi que les gènes codant les facteurs régulant la transcription des gènes des protéines de la structure du cheveu ont été mis en évidence de cette façon [3, 99].

a) Les maladies impliquant des mutations des gènes des kératines du cheveu

Deux maladies génétiques humaines sont décrites pour les kératines dures de type II du cheveu [134].
La première concerne la substitution d'une arginine en histidine sur un résidu du segment tête de K85. L'individu homozygote est alors atteint de dysplasie ectodermale caractérisée uniquement par une absence totale de tout poil et des dystrophies des ongles. Cette maladie suggère que ce site basique a très certainement un rôle important dans l'association en hétérodimère de K85 avec K35 au niveau la matrice du bulbe (voir chapitre II.5.a). Il est alors possible d'imaginer que l'absence de formation de l'hétérodimère K85/K35 conduit à l'impossibilité de l'initiation des ULF plus tard dans le pré cortex et que la structure filamentaire ne peut être correctement constituée même avec l'expression ultérieure des autres kératines [99, 170].
La seconde maladie, le monilethrix concerne les trois kératines de type II K81, K83 et K86. La transmission est autosomique dominante. Les fibres du malade se caractérisent dès l'enfance par un changement périodique de diamètre entrainant une grande susceptibilité à casse. Les ongles peuvent également présenter une dystrophie. La mutation impliquée entraine la substitution d'un acide glutamique sur une position située en fin du segment d'hélice 2B (position 407 pour K83, 402 ou 413 pour K81 et K86). La substitution implique un cluster acide particulièrement conservé parmi les séquences des filaments intermédiaires et décrit comme pouvant favoriser la répulsion des deux extrémités de chaines du dimère [99, 125, 171]. Par ailleurs, deux autres mutations sur un même site du gène KRT86 sont décrites comme pouvant être impliquées dans le monilethrix et se retrouvent en début du segment de l'hélice 1A de K86 (substitution d'une asparagine).
Une mutation est décrite chez l'humain sur la kératine K31 conduisant à la perte totale du segment queue sans pour autant entrainer une différence de phénotype [3]. La protéine tronquée est capable de s'associer *in vitro* avec K83 suggérant que le domaine queue des types I n'est pas impliqué dans le mécanisme d'appariement de ces protéines lors de leur association dans la zone de kératinisation du cortex.

b) Les maladies impliquant des mutations des gènes des kératines épithéliales du follicule

Certaines mutations des gènes codant pour les kératines K71 et K74 exprimées respectivement dans les cellules de la gaine membranaire interne et spécifiquement dans les cellules de la couche de Huxley entrainent des modifications de l'aspect des fibres [99].
Une maladie génétique autosomique dominante chez l'humain, dites du cheveu laineux, entraine chez l'individu atteint des fibres fines et très frisées. Cette maladie est liée à une substitution en début de segment de l'hélice 1A de K74 (sur l'asparagine homologue à celle décrite pour K86 pour le monilethrix) [172].
Chez la souris ou chez le chien, des substitutions impliquant des résidus situés également en début de segment 1A ou en fin de segment 2B sur la kératine K71 sont à l'origine de phénotypes entrainant une frisure de la fibre [99].
Ces maladies soulignent le rôle des kératines de type II des filaments de la gaine membranaire interne sur la texture du cheveu [99]. Néanmoins, les mutations décrites ne devraient pas être directement liées aux différents degrés de frisure du cheveu humain. En effet, la frisure du cheveu humain n'est pas décrite pour être un caractère transmis par une simple hérédité monogénique.
Il est intéressant de noter qu'aucune maladie génétique n'est associée à notre connaissance aux kératines de type I des cheveux et de la gaine membranaire interne.

c) Les maladies impliquant des mutations des gènes d'autres protéines

D'autres mutations impliquant des gènes des protéines constituant la desmosome sont décrites parmi les maladies héréditaires associées au cheveu. Des mutations sur la desmogléine 4 exprimée au niveau de la zone de kératinisation dans le cortex entrainent des formes de monilethrix et des sous développements du système pileux. D'autres mutations qui concernent des gènes des protéines du desmosomes (desmogléines, desmocollines et plakoglobines) exprimés dans la peau et dans les cheveux entrainent conjointement des anomalies sur les

différents tissus correspondants [99]. Ces observations suggèrent un rôle important des protéines du desmosome dans ces cellules pour le maintien de la cohésion des édifices.

Il n'existe pas de maladie connue impliquant directement des mutations sur les gènes des KAP. En revanche, des maladies entrainant des anomalies de la peau, des cheveux et des ongles sont identifiées et associées à des désordre pouvant influer sur l'expression des KAP. Les trichothiodystrophies, maladies autosomiques récessives sévères, sont liées à des mutations de gènes impliqués dans la réparation de l'ADN. Les malades ont, en particulier, des désordres semblant impliquer la métabolisation de la cystéine. Les cheveux des malades montrent une très faible expression d'une partie des familles des UHS KAP et une réduction de 50% de

l'abondance de la cystéine présente dans la fibre par rapport aux cheveux normaux. Les cheveux sont alors fragiles car très peu élastiques et fracturés [59, 73, 173, 174].

Conclusion : les champs de recherche qui restent à explorer

A la lumière des informations qui peuvent être extraites de la bibliographie, le cheveu est une structure biologique ayant été extensivement étudiée au cours du dernier siècle. Néanmoins, certains points ne sont, à l'heure actuelle, par encore élucidés et la recherche pour obtenir de nouvelles connaissances reste toujours d'actualité.

Compléter le catalogue des gènes exprimés en protéine

Le premier axe de recherche réside dans la détermination de l'expression des gènes dans la fibre capillaire. Si les travaux effectués au cours de la dernière décennie ont permis de mettre en évidence l'expression des kératines spécifiques du cheveu aux différentes étapes de la croissance du follicule, l'évidence de l'expression protéique des gènes correspondant aux différentes isoformes des familles de KAP est loin d'être apportée. Une partie des isoformes de ces familles a été étudiée seulement au niveau du transcriptome du follicule pileux. D'autres gènes, putatifs, ont seulement été décrits sur la base d'homologie de séquence avec des gènes déjà caractérisés.

La question de la traduction de ces gènes peut être posée et s'inscrit dans une problématique plus globale de la connaissance du génome humain qui nécessite, près de dix ans après son séquençage [175, 176], une poursuite de son annotation. Le but final est de constituer des banques de données complètes dans lesquelles l'ensemble des séquences de gènes codants serait référencé en comprenant les informations de fonctions, de localisations de l'expression, de structures, de modifications post-traductionnelles et d'interaction avec les autres molécules de l'organisme. Dans cette optique, la centaine de gènes prédits de KAP représente près de 1,5% des 7000 protéines qui n'ont peu ou pas de description d'expression parmi les 20 300 protéines humaines pouvant être attendues compte tenu des données actuelles d'annotation [177]. L'analyse du cheveu ouvre la voie à la compréhension complète de quelques uns des 230 types cellulaires de l'organisme humain [177].

Les techniques de l'analyse protéomique, que nous décrirons et utiliserons dans la suite de ce manuscrit, apparaissent comme un nouvel outil adéquat pour tenter d'apporter des éléments de réponse à cette problématique. La stratégie n'a cependant pas été considérablement exploitée pour l'étude du cheveu. A ce jour, une seule étude a été publiée sur le protéome du cheveu [89]. Les résultats de cette étude n'ont pas été interprétés dans une optique de compréhension de la structure ni dans une problématique de caractérisation des familles de KAP, qui nous le verrons, est loin d'être triviale. Par ailleurs, les annotations des banques de données et les connaissances ont entre temps évoluées et une confrontation de ces données avec des informations expérimentales est d'ores et déjà d'actualité. Des développements sont ainsi à réaliser pour améliorer les connaissances du protéome du cheveu et passent par l'utilisation des stratégies les plus récentes de la discipline analytique désormais mature qu'est la protéomique.

Obtenir des notions quantitatives de l'expression protéiques de ces gènes

Les informations quantitatives de l'expression des protéines sont une perspective à leur identification. La connaissance des quantités relatives ou absolues des protéines structurant le cellules différenciées du cheveu peut apporter des indices sur leur contribution à l'édifice macromoléculaire. L'obtention de ces données reste difficile et constitue un chalenge auquel se confronte actuellement la protéomique [178]. Des techniques de quantification différentielles peuvent plus facilement être utilisées pour évaluer les différences existant entre deux protéomes ou états de protéome. Elles peuvent être envisagées pour évaluer la spécificité de l'extraction d'un protéome ciblé et estimer le niveau de contamination de protéome adjacents mais également pour comparer des échantillons issus de différents individus ou ayant subi des traitements particuliers.

Etendre à l'étude d'autres catégories de protéines pouvant être retrouvées dans le cheveu

L'analyse des protéomes des cellules des cheveux apporte la possibilité d'identifier, en plus des kératines et des KAP pouvant être attendues, les protéines qui ne composent pas la structure mais qui ont participé aux processus de régulation de la différenciation cellulaire et qui ont pu subsister. L'utilisation de la protéomique doit ainsi permettre d'apporter des informations sur les mécanismes cellulaires ayant contribué à l'élaboration de l'édifice.

Rechercher les modifications existant dans la fibre mature

Les protéines du cheveu sont potentiellement soumises aux modifications environnementales et aux traitements. L'analyse du protéome des cheveux ouvre la voie à l'identification des protéines portant des modifications chimiques et à la localisation de celles-ci dans les séquences. Ces modifications du cheveu sont pour l'instant décrites seulement par l'analyse des compositions en acides aminés des hydrolysats protéiques. Il est également possible d'envisager l'étude de modifications labiles ne pouvant être détectées dans des analyses d'hydrolysats acides. Ces recherches n'excluent pas l'étude de modifications post traductionnelles qui pourraient intervenir au cours de la différenciation.

Comprendre le rôle de ces protéines en recherchant des informations structurales

Si la compréhension de l'arrangement des kératines du cortex pour former les macrofibrilles est de mieux en mieux appréhendée, le rôle des KAP dans la structure n'a pas été très étudié principalement du fait que leur expression n'est pas bien déterminée. De plus, les mécanismes de formation des structures de la cuticule ne sont actuellement pas théorisés. L'identification et la quantification des protéines majoritaires dans la structure permettrait d'apporter des éléments pour décrire les mécanismes moléculaires mis en jeu lors de la différenciation. Elles doivent également pouvoir être complétées par des informations de structures secondaires et tertiaires pour ces protéines.

Ces informations sont probablement des clés pour élucider la structure moléculaire de l'édifice biologique qui n'est toujours pas précisément connue et d'expliquer ses comportements mécaniques et chimiques pour envisager dans l'avenir de nouveaux traitements cosmétiques.

Partie II Mise au point de nouvelles stratégies protéomiques pour la caractérisation de protéines issues des familles multigéniques pour l'analyse des constituants du cheveu

L'analyse protéomique est une technique qui a émergé dans les années 90 suite au développement de sources d'ionisations non destructives permettant l'analyse par spectrométrie de masse de molécules en conservant leur intégrité. Ces innovations techniques amenées grâce au développement des sources de désorption et d'ionisation assistée par une matrice (MALDI) et d'ionisation électrospray (ESI), ont ouvert la voie à l'analyse des molécules biologiques par spectrométrie de masse. L'attribution à ces découvertes du Prix Nobel de Chimie en 2002 illustre l'importance de leurs champs d'application.

Parmi ces applications, le séquençage rapide des protéines par spectrométrie de masse a permis d'envisager l'étude du protéome, c'est-à-dire de l'ensemble de protéines produites dans une cellule ou un tissu suite à l'expression du génome.

Si les techniques de séquençage des protéomes semblent désormais maitrisées, la problématique de l'étude des isoformes de protéines du cheveu pose un certain nombre de difficultés. Nous expliciterons les choix nous ayant conduit à une stratégie adaptée à leur analyse. Après avoir introduit dans un premier chapitre le principe des stratégies de séquençage des protéines, nous discuterons dans un second chapitre du choix des techniques séparatives pour l'analyse d'isoformes puis nous décrirons les stratégies pouvant être envisagées aux différentes étapes de l'analyse pour améliorer l'acquisition, le traitement et la validation des résultats. Dans ce sens, une partie sera intégralement consacrée à l'optimisation du couplage nanoLC-Q-TOF, pièce maitresse pour l'obtention des informations protéomiques envisagées.

Chapitre I Introduction à la protéomique

1) Le principe fondamental : la comparaison du protéome au génome

Les séquences protéiques d'acides aminés définissent la structure de la protéine. Selon leur structure, les protéines peuvent alors avoir une grande variété de fonctions qui peuvent éventuellement être modulées par l'ajout de modifications dites post-traductionnelles induites enzymatiquement. Ces séquences en acides aminés sont spécifiques à la protéine et peuvent être utilisées pour leur identification.

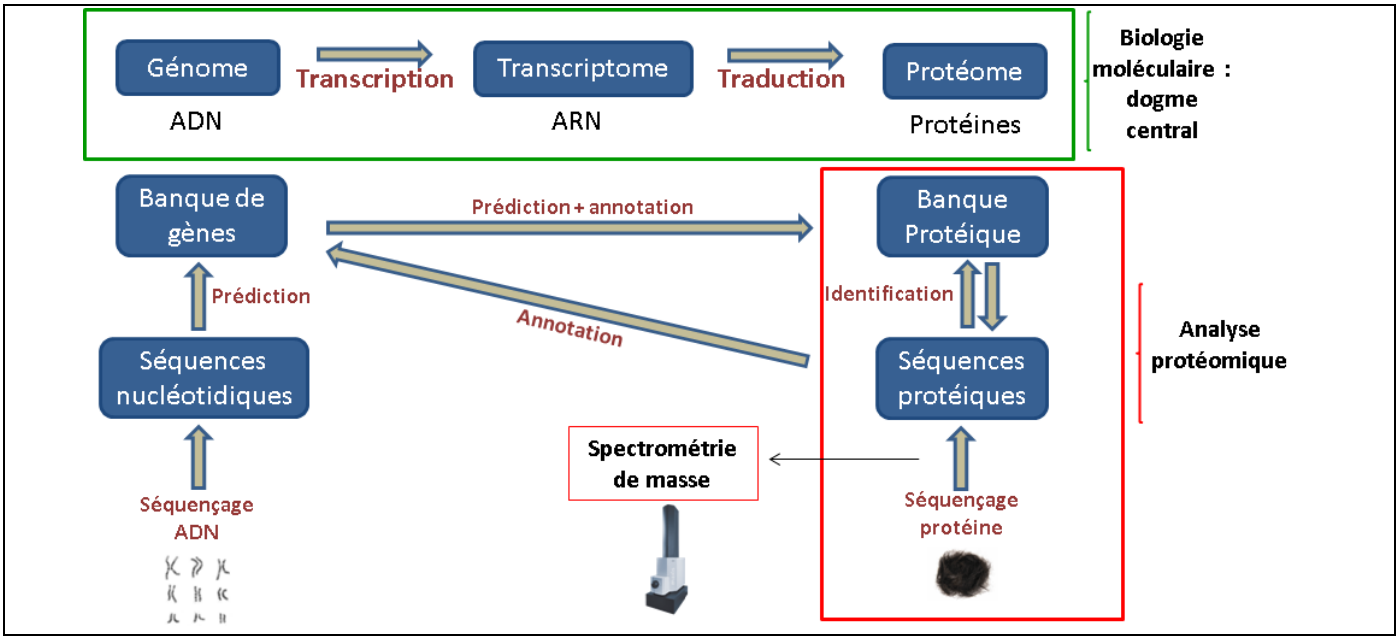


Figure 1 : *Schématisation du principe de comparaison des données protéiques aux données génomiques pour l'analyse des protéomes.*

L'étude du protéome s'appuie sur le postulat que le génome d'un organisme est transcrit en ARN puis traduit en protéines [179]. Il existe donc un lien direct entre les séquences d'ADN du génome et les séquences des protéines. En connaissant les séquences génomiques de l'organisme étudié, il est possible de les convertir en une base de données, dite banque, contenant les gènes pouvant théoriquement être exprimés dans l'organisme séquencé (Figure 1). Cette banque de gènes peut être convertie en banque protéique en s'appuyant sur les règles de traduction et de transcription du génome définies par le dogme central. Les banques de données peuvent être affinées avec les précédentes connaissances des séquences protéiques obtenues par les techniques de biologie moléculaire.

La méthode d'identification des protéines définies par la stratégie protéomique repose sur le principe de la comparaison des données des banques protéiques théoriques avec des données expérimentales [180]. Ces données expérimentales proviennent désormais de l'analyse par spectrométrie de masse des protéines réellement présentes dans le protéome [181-183]. La correspondance entre les données expérimentales et les données théoriques est réalisée au moyen d'un algorithme de calcul et permet d'identifier les protéines présentes dans le protéome analysé.

Ce principe analytique s'intègre au sein d'une succession d'étapes pouvant comprendre l'isolement du protéome, sa décomplexification, sa préparation en vue de son analyse par spectrométrie de masse, l'analyse proprement dite puis le traitement des données obtenues [182].

2) L'obtention des données de séquences protéiques expérimentales par spectrométrie de masse.

a) Architecture et principe général de fonctionnement d'un spectromètre de masse

L'analyse par spectrométrie de masse repose sur plusieurs étapes permettant la mesure de la masse d'une molécule. Cette mesure de masse est réalisée en phase gazeuse sur l'analyte préalablement ionisé.

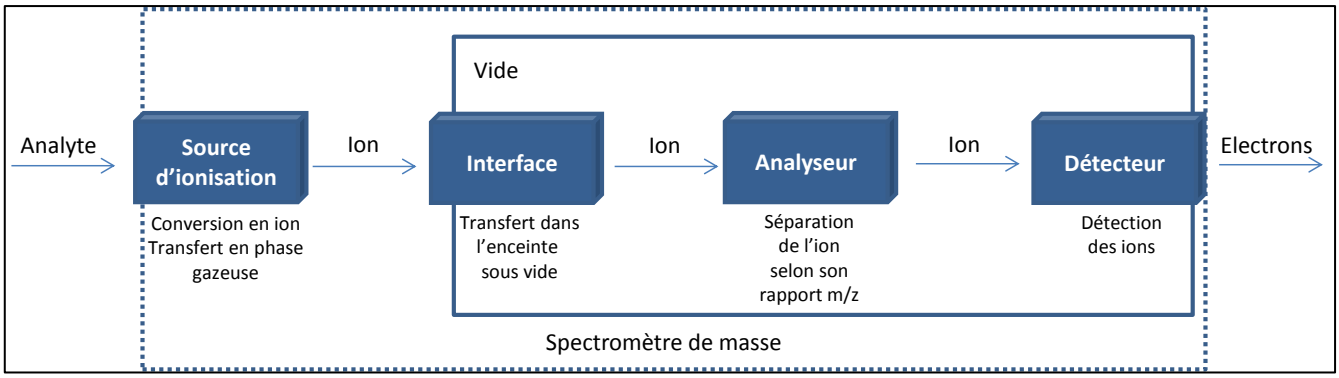


Figure 2 : *Les différents éléments composants un spectromètre de masse*

Dans ce but, le spectromètre de masse se décompose en quatre éléments principaux (Figure 2) :

- La **source d'ionisation** permet de conférer des charges à l'analyte qui va pouvoir être extrait de sa matrice solide (cas du MALDI) ou liquide (ESI) et être transféré en phase gazeuse.
- L'**interface**, entre la source et l'analyseur, permet de faire passer les composés ionisés de la pression de la source au vide de l'analyseur.
- L'**analyseur** réalise la séparation physique des ions selon leur rapport masse sur charge (m/z).
- Le **détecteur** permet de détecter les ions séparés en traduisant leur détection par un signal électrique. Ce signal électrique peut alors être amplifié puis digitalisé.

Le spectromètre de masse sépare donc des ions en fonction de leur rapport m/z et enregistre une intensité correspondante. La donnée obtenue est un spectre de masse qui est une fonction Intensité =f (m/z).

b) Principe de l'utilisation d'un spectromètre de masse pour l'obtention d'information de séquence protéique

Dans le cadre de l'analyse des séquences des acides aminés d'une protéine, l'information qui doit être obtenue est la composition en acides aminés et leur position au sein de la séquence.

Deux types de stratégies peuvent être envisagés pour accéder à ces informations par spectrométrie de masse :

- l'approche consistant à obtenir des informations de séquence directement sur la protéine entière (**approche top-down**) [184, 185]. Cette stratégie nécessite de travailler sur un mélange protéique peu complexe pouvant être résolu par l'analyseur. Les protéines analysées doivent être solubles et dans un tampon compatible avec l'analyse par spectrométrie de masse. Ces conditions sont difficilement applicables à l'analyse des protéomes issus de matrices biologiques. Ces dernières sont souvent très complexes et l'extraction des protéines peut nécessiter l'utilisation de tensio-actifs incompatibles avec l'instrumentation.
- l'approche consistant à obtenir des informations de séquence à partir de fragments de protéines obtenus après digestion enzymatique (**approche bottom-up**) [186]. L'analyse de peptides ainsi générés permet de s'affranchir des difficultés de manipulation inhérents aux protéines et notamment leur propriétés de solubilité. La plus grande facilité d'accession aux informations des séquences des peptides par spectrométrie de masse justifie que cette stratégie soit la plus utilisée dans le cadre de l'analyse protéomique haut débit.

Nous ne décrirons par la suite que le séquençage des protéines par une approche bottom-up.
La mesure de masse d'un peptide seul ne permet pas de remonter à l'identification d'une protéine. Par contre, la mesure des masses de différents peptides d'une même protéine dans un même échantillon peut permettre d'accéder à cette information (approche par empreinte peptidique massique). Elle est néanmoins peu spécifique et source d'erreur dans le cadre de l'analyse de protéomes contenant un nombre important de protéines.
En protéomique, l'approche désormais la plus courante est d'utiliser des informations supplémentaires en mesurant la masse de fragments de peptides générés grâce à des spectromètres adaptés (Figure 3). Ces spectromètres de masse dits en tandem disposent en plus des éléments classiques :

- D'un **filtre de masse** permettant d'isoler sélectivement l'ion provenant du peptide à séquencer.
- D'une **cellule de collision** permettant de fournir une énergie suffisante pour déstabiliser l'ion et générer des fragments.
- D'un **analyseur** permettant de mesurer les rapports m/z des fragments générés.

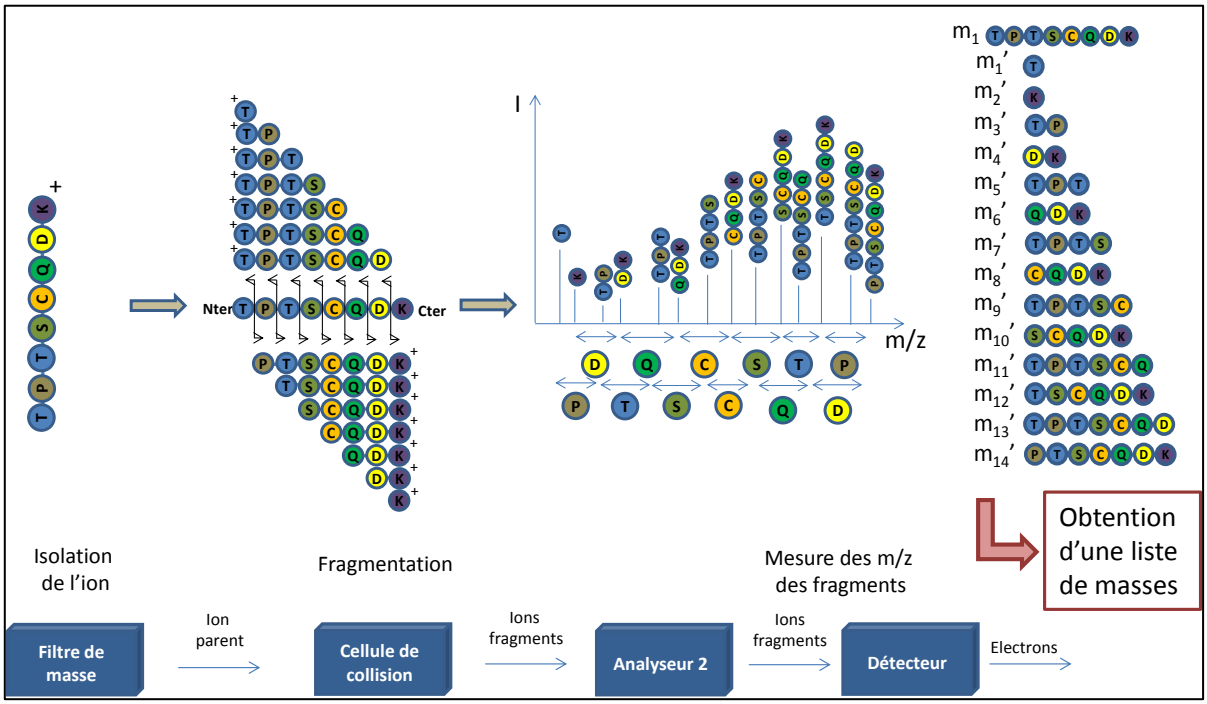


Figure 3 *: Principe de fragmentation des séquences peptidiques pour l'obtention de listes de masses de fragments nécessaires pour l'identification des protéines par spectrométrie de masse en tandem.*

Un peptide est une succession d'acides aminés liés les uns aux autres par une liaison peptidique. Lorsque le peptide est ionisé, il est possible de trouver des conditions pour favoriser des mécanismes de fragmentation affectant principalement les liaisons peptidiques. Ainsi, pour un peptide donné, plusieurs fragments dont les masses sont directement dépendantes de la séquence peptidique vont être générés. La différence de masse entre les fragments permet de remonter à l'enchainement des acides aminés dans la séquence.
L'information obtenue pour le peptide séquencé est le rapport m/z de l'ion parent et les rapports m/z des fragments issus de la fragmentation. Ces données sont facilement convertibles en une liste de masses de fragments (m'_1, m'_2, …, m'_n) issus d'un parent de masse donnée m_1. L'analyse par spectrométrie de masse d'un protéome permet d'obtenir un ensemble de listes spécifiques parent/fragments correspondant à autant de peptides effectivement séquencés.

c) Les spectromètres de masse utilisés pour le séquençage en protéomique

Plusieurs architectures d'analyseurs sont utilisées pour obtenir les informations de séquences peptidiques des analyses protéomiques [184]. Chacun des systèmes suivants permet de réaliser les étapes de mesure de masse parent/fragments, l'isolement des parents et leur fragmentation décrites précédemment. En fonction des types d'analyseurs utilisés, les performances en terme de résolution spectrale (capacité à distinguer des ions de masses voisines), de précision de mesure de masse, de gamme de mesure, de sensibilité (capacité à détecter les ions) et de gamme dynamique (ratio entre les quantités de l'ion le plus abondant et l'ion le moins abondant détectés simultanément) peuvent être différentes. La rapidité d'acquisition des données peut également différer.
Parmi ces architectures peuvent être décrits :

- L'hybride Q-q-TOF. La mesure des masses est réalisée par un analyseur à temps de vol (TOF) qui consiste à mesurer le temps de vol des ions, préalablement accélérés, parcourant une distance donnée. Le temps de parcours des ions est proportionnel à leur rapport m/z. L'ion parent à fragmenter est isolé grâce à un quadripôle (Q) puis fragmenté dans une cellule de collision située entre les deux analyseurs. Les ions fragments générés sont alors analysés par l'analyseur TOF.
  La mesure de masse est très rapide et précise mais un nombre important de cycles de mesure des ions est nécessaire pour accumuler un signal suffisant qui dépend donc du temps. Nous décrirons plus en détail ce système utilisé pour la suite de ce travail de thèse.

- Les systèmes de trappes à ions linéaires ou tridimensionnelles. Dans ces systèmes, les mesures de masse, la sélection et la fragmentation sont réalisées dans un piège quadripolaire capable de piéger puis d'éjecter sélectivement les ions. Initialement les ions sont piégés dans la trappe grâce à l'application d'une combinaison tensions/radiofréquences. Cette combinaison, associée à la présence d'un gaz de thermalisation, permet de conférer aux ions piégés une trajectoire stable dans la trappe. La mesure de masse est réalisée par balayage de combinaisons de tension et de radiofréquence permettant d'éjecter les ions en fonction de leur rapport m/z vers le détecteur. L'isolement de l'ion parent est réalisée en éjectant en deux étapes l'ensemble des ions de m/z inférieur puis l'ensemble des ions de m/z supérieur hors de la trappe. L'ion isolé est soumis à une fréquence de résonnance permettant de lui conférer de l'énergie cinétique. Sa collision avec les molécules de gaz entraine sa fragmentation, les ions générés étant conservés dans la trappe. La masse des fragments est alors mesurée.
  La mesure de masse et le cycle global d'analyse est très rapide mais cette rapidité est réalisée au détriment de la résolution et de la précision de mesure. L'accumulation préalable des ions analysés permet une très bonne sensibilité.

- Les analyseurs TOF-TOF. La mesure de masse des ions parents est réalisée par temps de vol. La sélection du parent à fragmenter est réalisée dans le tube de vol par l'intermédiaire d'une porte à ion. Elle permet de ne transmettre dans la suite du tube de vol que les ions passant dans l'intervalle des temps de vol correspondant à l'ion sélectionné. L'ion est alors fragmenté dans une cellule de collision et les fragments obtenus sont alors mesurés par temps de vol dans la suite du tube.

- Les hybrides trappes/analyseurs à transformée de Fourier (trappe à résonnance cyclotronique ionique FT-ICR et Orbitrap). L'isolement, la fragmentation et la mesure des fragments sont réalisées dans un piège quadripolaire comme décrit précédemment. Dans ces systèmes, la mesure de la masse de l'ion parent est réalisée dans un analyseur à transformée de Fourier. Le principe de ces analyseurs consiste également à placer les ions dans un piège. Les conditions de champ magnétique et/ou de champ électrique appliquées aux ions dans le piège entrainent de leur part un mouvement cohérent et périodique dont la fréquence

est fonction du rapport m/z. Cette fréquence peut être mesurée grâce à la mesure du courant induit par le mouvement des ions sur des électrodes en fonction du temps. La fonction courant induit en fonction du temps permet après transformée de Fourier de remonter à l'information de fréquence des ions oscillants et d'en déduire leur masse.
Ces analyseurs permettent une excellente résolution et des mesures précises de l'ion parent mais cette mesure nécessite un temps suffisant pour l'accumulation des ions, l'application du mouvement de cohérence aux ions et l'accumulation du courant induit à la détection. Les fragments sont donc généralement analysés dans la trappe pour gagner en rapidité d'acquisition au détriment de la précision de leur mesure.

- L'hybride trappe d'ion/TOF. Le principe de ce couplage est de combiner un analyseur apportant de la précision de mesure tel que le TOF avec une trappe qui permet de réaliser rapidement l'accumulation, la sélection, la fragmentation et le stockage des ions. Dans ce système, les ions parents et les ions fragments peuvent être mesurés avec la même précision. Le stockage des ions fragments avant l'introduction dans le TOF permet de diminuer le temps nécessaire à leur mesure de masse en comparaison de l'hybride Q-TOF.

3) L'identification des protéines

a) La comparaison des données expérimentales et théoriques

La stratégie protéomique utilise directement les données des listes de masses générées pour les comparer avec la banque protéique correspondant à l'organisme étudié [180].
A partir des données issues de la banque protéique, un algorithme génère des listes de masses théoriques correspondant aux peptides et à leurs fragments théoriques (Figure 4). L'algorithme compare alors les listes de masse théoriques générées in silico avec les listes de masse obtenues expérimentalement. L'identification du peptide est réalisée lorsque sa liste de masses expérimentale correspond à une des listes de masses théoriques calculées. La liste de masse expérimentale est ainsi rattachée à la séquence d'un peptide issu d'une protéine de la banque. Ces derniers peuvent alors être considérés comme présents dans l'échantillon analysé. La qualité de la corrélation entre la donnée théorique et la donnée expérimentale est évaluée par un score dont le calcul dépend de la version de l'algorithme utilisé.

Il est possible d'obtenir les informations de séquences directement par interprétation des spectres de fragmentations (spectres MS/MS) puis de rechercher pour chaque séquence identifiée la protéine à laquelle elle correspond (séquençage *de novo*) [187, 188]. Le séquençage s'effectue alors par mesure des différences de masse entre les fragments consécutifs qui correspondent aux masses des acides aminés. Cette stratégie nécessite néanmoins d'avoir une série de fragments non ambiguë et des spectres de très bonne qualité, ce qui n'est pas le cas pour une partie des spectres MS/MS. D'autre part, certains acides aminés possèdent des masses isobares (identiques) ou voisines ce qui peut entrainer des ambiguïtés dans l'interprétation des séquences. Cette stratégie nécessite également beaucoup plus de temps de calcul qu'une simple comparaison de listes de masses. Elle est réservée à l'étude d'organismes dont le génome n'a peu ou pas été séquencé ce qui signifie que peu de données sont présentes dans les banques protéiques. L'organisme étudié dans notre problématique est homo sapiens dont le génome a été totalement séquencé et bien annoté. Par conséquent, l'approche de novo ne sera pas considérée dans la suite de ce travail.

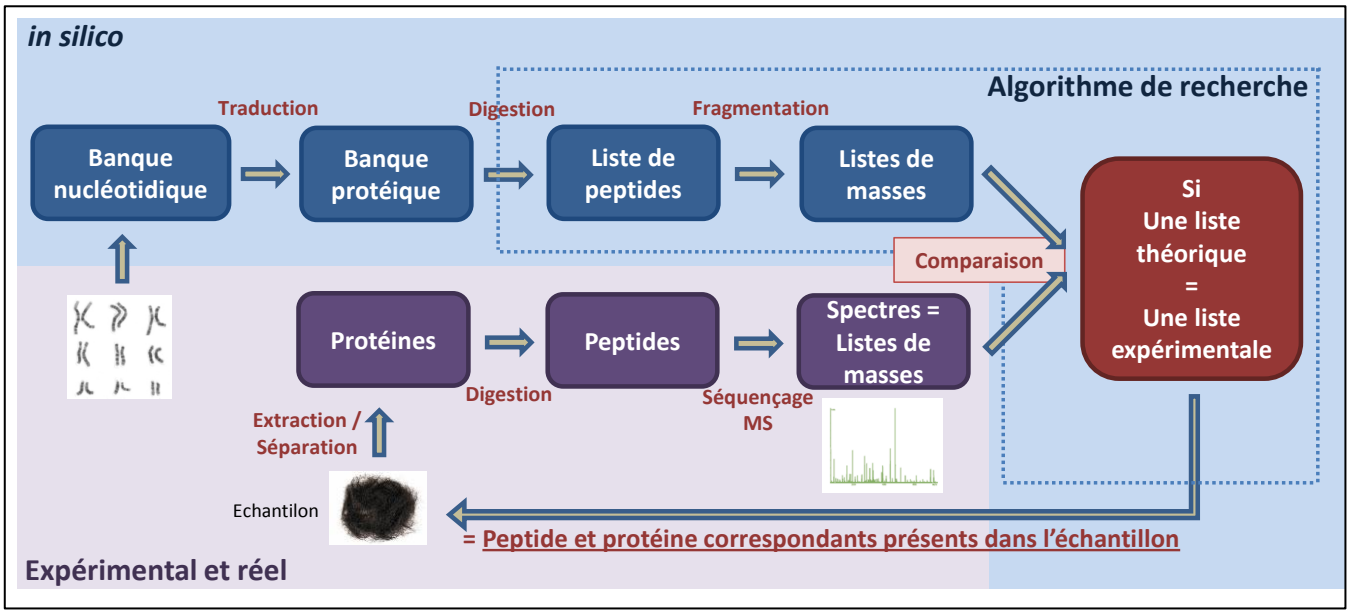


Figure 4 : *Principe de la confrontation des données expérimentales de séquençage MS-MS/MS aux données théoriques issues de la traduction et de l'annotation des banques nucléotidiques.*

b) La définition des critères de recherche

Il est nécessaire pour la spécificité de la recherche d'encadrer la création des listes de masses théoriques de manière à traduire la réalité des traitements et des analyses effectuées physiquement sur l'échantillon. Des critères de recherche sont ainsi fixés à l'algorithme afin de comparer les données expérimentales à des données théoriques cohérentes. Ces critères vont directement impacter sur le nombre de listes de masses théoriques comparées avec chaque liste de masse expérimentale. Chaque étape réelle de manipulation de l'échantillon protéique définit un équivalent de traitement *in silico* définissant des paramètres de recherche :

La condition expérimentale	définit	le paramètre de recherche
organisme / groupe d'organismes étudiés	->	banque protéique
tissus, protéines cibles	->	sous banque protéique
modifications post-traductionnelles	->	modifications variables
modifications chimiques spécifiques	->	modifications fixes
enzyme, spécificité de digestion	->	digestion *in silico* = masses des peptides à considérer
analyseur 1	->	tolérance de masse sur le parent
mécanisme de collision	->	masse des fragments à considérer pour chaque parent
analyseur 2	->	tolérance de masse sur les fragments

Tableau 1 : *Parallèle proposé entre les processus réels subis par l'échantillon et les paramètres de recherche utilisé pour l'identification en protéomique.*

Cette définition de recherche entraine une limite puisque par cette approche, les identifications sont restreintes à ce qui est recherché. Par exemple, il n'est pas possible d'identifier la modification d'un résidu si celle-ci n'a pas été considérée. De la même façon, une protéine absente de la banque utilisée pour la recherche ne sera pas identifiée.

4) La décomplexification de l'échantillon pour accroitre les capacités de séquençage en protéomique.

L'approche bottom-up de digestion des protéines en peptides entraine une augmentation de la complexité de l'échantillon [189]. Plusieurs obstacles s'opposent à l'introduction directe des mélanges peptidiques dans le spectromètre de masse pour leur séquençage :

- La présence d'un nombre trop important de peptides dans la source d'ionisation entraine un effet de suppression ionique [190]. Il existe dans la source un effet de compétition à l'ionisation et au transfert des composés vers l'analyseur. Si un nombre trop important de peptides est introduit, les analytes dont l'ionisation est la plus efficace seront favorisés. Parallèlement, si des composés sont très majoritaires en concentration par rapport à d'autres, ils peuvent également être favorisés.
- L'isolement spécifique d'ions ne peut être réalisée sur des espèces isobariques ou de rapports m/z voisins [191]. La fragmentation simultanée de peptides ionisés non séparés conduit à un mélange dans le spectre de fragmentation des différentes espèces. L'interprétation du spectre par comparaison avec les données théoriques résultant est alors impossible.
- Les étapes d'analyse de l'ion parent, de son isolement, de sa fragmentation et de l'analyse des masses de ses fragments nécessitent un certain temps de fonctionnement du spectromètre. Pendant ce temps, il n'est pas possible d'analyser d'autres ions. L'introduction de l'échantillon pendant un temps insuffisant pour séquencer l'ensemble des peptides entraine donc de la perte d'information [192].

Pour l'ensemble de ces raisons, la stratégie protéomique a toujours recours à des techniques de décomplexification de l'échantillon. Ces techniques peuvent être menées par séparation en amont des protéines (électrophorèses, chromatographies, précipitations…), séparation des peptides (chromatographies) ou combinaison des deux. Elles permettent de diviser par la suite l'échantillon en une multitude d'échantillons dans lesquels les peptides sont séparés physiquement. Sur ce principe, l'injection des peptides dans le spectromètre va être réalisée séquentiellement avec des complexités suffisamment faibles pour permettre le séquençage par le spectromètre de masse.

L'analyse de digest est ainsi essentiellement réalisée par couplage du spectromètre de masse avec un système de chromatographie liquide. Les sources d'ionisation électrospray sont idéalement adaptées pour réaliser ce couplage en direct. Elles permettent d'utiliser le spectromètre de masse comme un analyseur des composés élués mais également comme détecteur des courants d'ions générés en fonction du temps de chromatographie. Cette mesure des courants d'ions en fonction du temps de chromatographie est à la base de la majorité des applications quantitatives de la spectrométrie de masse en protéomique.

L'utilisation de la source MALDI pour le couplage est également possible mais nécessite une étape intermédiaire de cristallisation des analytes [193]. Le couplage entre chromatographie et spectrométrie de masse est donc indirect et passe par des étapes de collecte des éluats, de préparation de ces collectes avant analyse séquentielle par MALDI-MS/MS. La nécessité de collecter rend plus difficile l'utilisation de la LC-MALDI-MS pour réaliser de la quantification.

5) L'acquisition automatisée des données spectrales en mode dépendant des données (DDA)

L'obtention d'informations spectrales à haut débit requise en protéomique bottom-up rend peu efficace une intervention en direct de l'utilisateur [181]. Les étapes de sélection, d'isolement et de fragmentation nécessaires sont donc programmées. L'automatisation est rendue possible par l'utilisation d'outils d'analyses « à la volée » permettant l'examen rapide des spectres MS réalisés pendant l'acquisition. Ces spectres MS contiennent un certain nombre d'informations qui vont permettre de choisir les composés à fragmenter mais également de fixer les conditions de fragmentation et d'obtention des spectres des fragments. Cette acquisition est dynamique dans le cas du couplage en ligne LC-ESI-MS et statique en couplage indirect LC-MALDI-MS.

Lors de l'analyse dynamique des peptides élués par la chromatographie, l'instant auquel est réalisée une acquisition est plus ou moins aléatoire. Un spectre MS contient plusieurs composés dont l'intensité est mesurée. L'analyse à la volée permet de trier les composés en fonction de leur intensité puis de déterminer leur état de charge. En fonction de ces données et de paramètres définis avant l'analyse par l'expérimentateur (seuil d'intensité de sélection, nombre de composés à sélectionner, règles de fragmentation…), des spectres de

fragmentation des composés sélectionnés sont obtenus pour les acquisitions suivantes. Lorsque ces acquisitions sont achevées, un nouveau spectre MS est réalisé et le cycle d'analyse reprend. Il est possible de paramétrer une exclusion temporaire des masses des composés ayant précédemment été analysés pour éviter une redondance de sélection au détriment d'autres composés non analysés.

L'inconvénient du mode d'acquisition dynamique est qu'il existe une probabilité non négligeable de ne pas sélectionner l'ensemble des composés. Si l'analyse n'est pas suffisamment rapide, le nombre de composés sélectionnés est insuffisant pour analyser l'ensemble des composés de l'échantillon. Le choix des paramètres d'acquisition, de réglage du spectromètre et de la chromatographie ainsi que les stratégies de décomplexification peuvent s'avérer critiques pour la génération de données complètes et exhaustives. Nous décrirons plus en détail ces considérations dans la suite de cette partie.

L'acquisition statique de données permise par la stratégie MALDI-MS s'affranchit de ces difficultés puisque le temps imparti pour l'analyse de l'échantillon n'est pas limité.

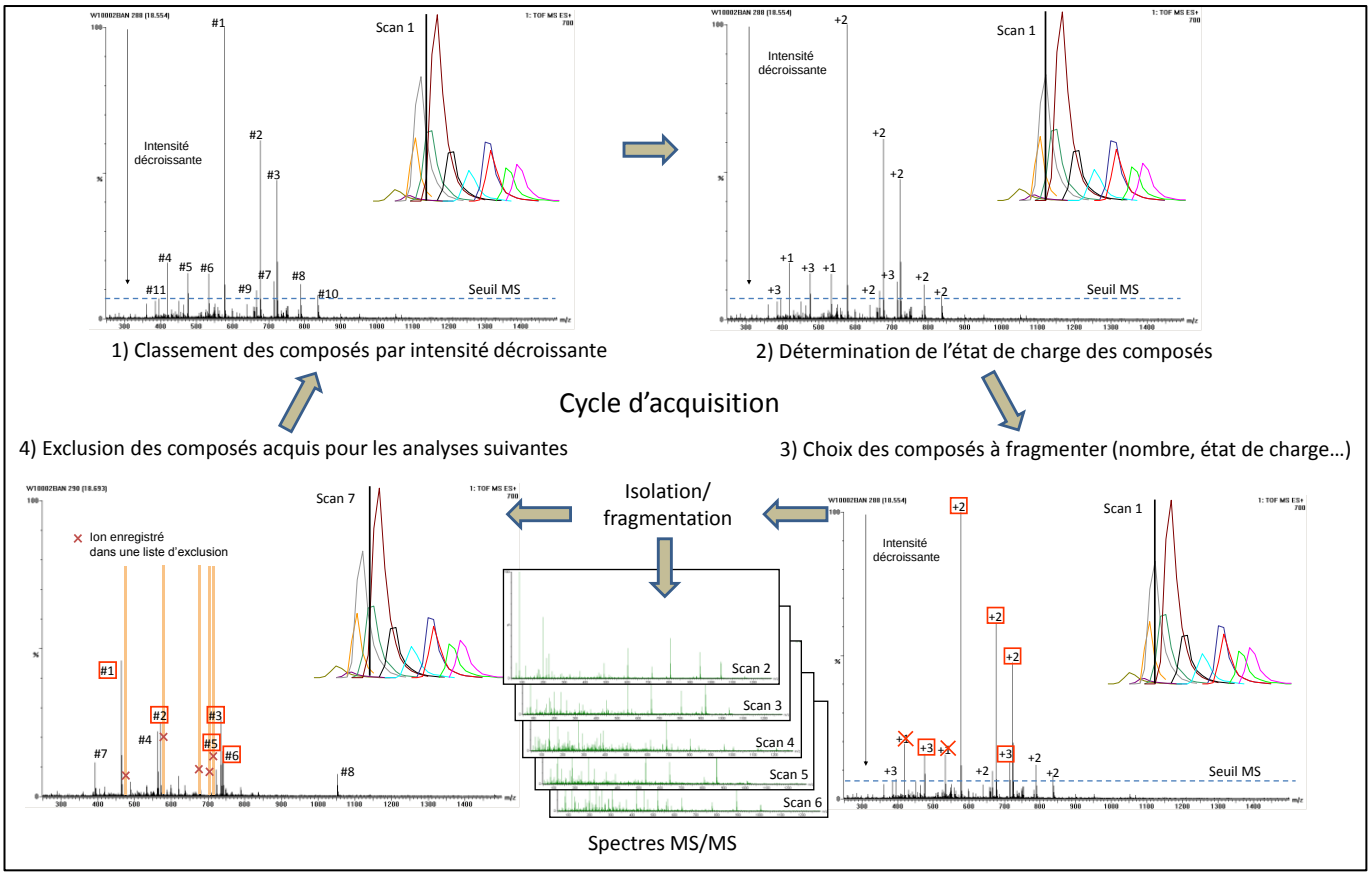


Figure 5 : *Schématisation du cycle d'acquisition en mode dépendant des données en LC-MS/MS.*

6) La notion d'erreur d'identification en protéomique et son contrôle

a) Les faux positifs et les vrais négatifs

L'approche protéomique par comparaison de listes de masses permet d'apporter un haut débit d'identification. Elle pose néanmoins le problème de la certitude de l'identification ainsi réalisée [194].

La donnée expérimentale (le spectre MS/MS) peut être de qualité insuffisante (peu de fragments ou fragments peu intenses par rapport au bruit de fond de la détection). La liste de masses expérimentales est alors peu spécifique et peut correspondre, après comparaison par l'algorithme de recherche, à plusieurs séquences théoriques dont la qualité de corrélation est faible. Il est alors probable que l'algorithme assigne une

identification fausse au spectre correspondant. Cette identification fausse completera la liste des protéines identifiées dans l'échantillon : c'est un faux positif.

Les critères de recherche adoptés pour générer les listes de masses théoriques peuvent être peu spécifiques. Dans ce cas, le nombre de séquences proposées à la comparaison d'une liste de masse expérimentale est très important. La probabilité d'y associer une séquence théorique fausse augmente.

Si au contraire les critères de recherche sont trop restrictifs, alors la probabilité de ne pas générer une liste de masses correspondant à une liste de masses expérimentales augmente et des identifications peuvent être manquées : ce sont des vrais négatifs.

La spécification des critères de recherche est un compromis entre une recherche large qui ouvre un espace large d'identification augmentant le risque de faux positifs et une recherche restreinte qui expose au risque de vrais négatifs.

Pour tenir compte de la spécificité de la recherche liée au choix des critères choisis pour la recherche, la plupart des algorithmes de recherche pondèrent le score de corrélation d'une identification avec la probabilité que l'identification soit due au hasard. L'identification de la liste de masses a alors un score dit probabiliste fonction de la qualité de la corrélation entre la donnée expérimentale et la donnée théorique et fonction du nombre de données théoriques utilisées pour la comparaison ([195] pour plus de détails sur les principes de calcul de ces scores).

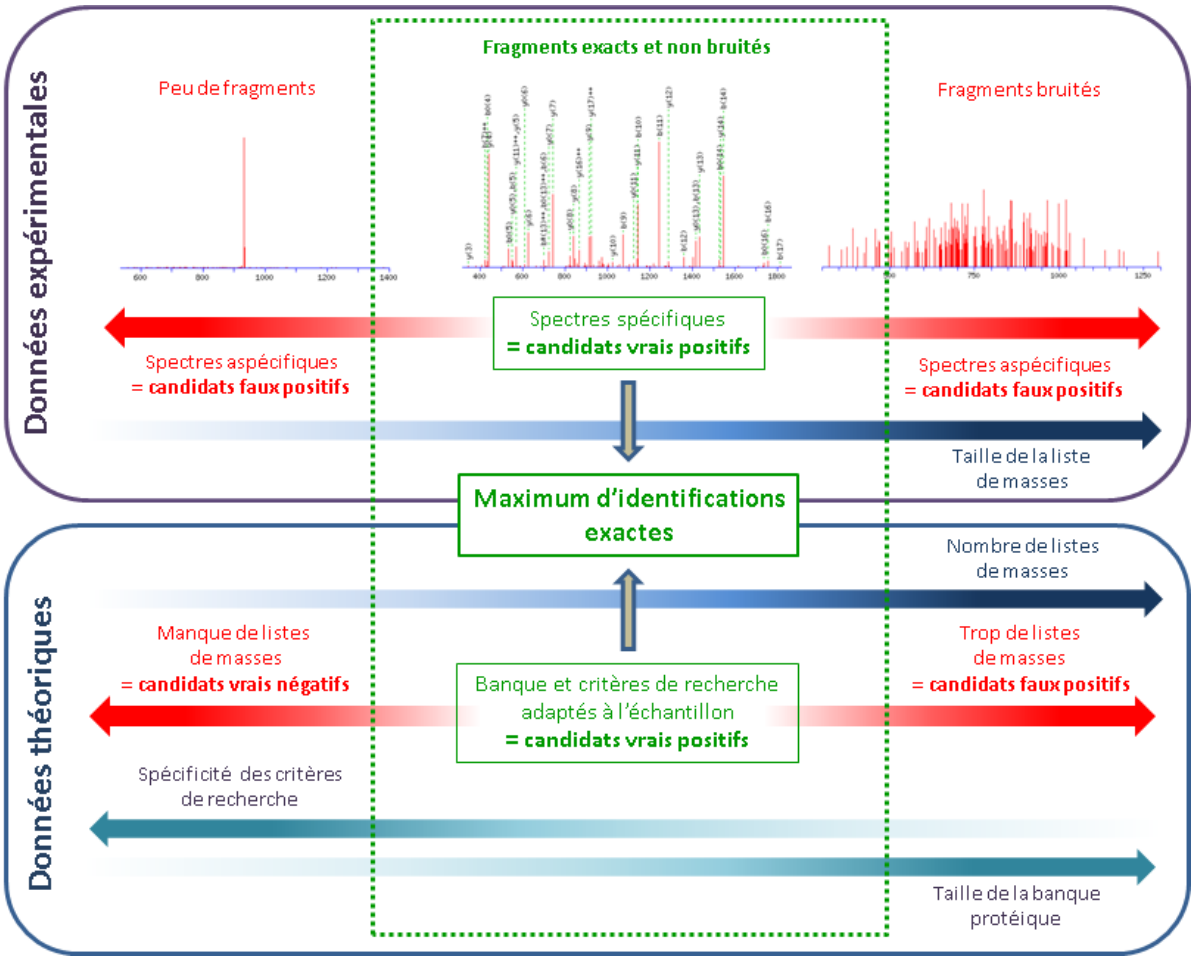


Figure 6 : *Illustration des causes d'introduction d'erreur d'identification (faux positifs et vrais négatifs) dans le cadre d'une approche d'identification protéomique utilisant les spectres de fragmentation peptidique.*

b) La validation des identifications

Le problème de la validation des spectres identifiés par l'approche décrite précédemment est de distinguer les vrais positifs des faux positifs et d'assurer un filtrage de ces derniers. Il est évidemment impossible d'effectuer une vérification manuelle de chaque spectre compte tenu du nombre de données et de la subjectivité de cette méthode [196].
Le fitrage peut être réalisé grâce à l'utilisation des scores de probabilité calculés par l'algorithme de recherche pour l'identification de chaque spectre assigné. La difficulté est d'évaluer objectivement le seuil de score à utiliser pour filtrer les identifications de faux positifs tout en conservant les identifications de vrais positifs.
La première approche consiste à joindre à la banque protéique une banque protéique leurre contenant des séquences incohérentes vis-à-vis de l'échantillon étudié. La banque « leurre » est choisie afin que les listes de masses expérimentales soient comparées à un nombre identique de listes de masses théoriques et de listes de masses « leurres » [197]. Classiquement, ces banques leurres sont obtenues en inversant ou en mélangeant l'ordre des acides aminés de chaque séquence protéique de la banque protéique utilisée pour l'analyse.

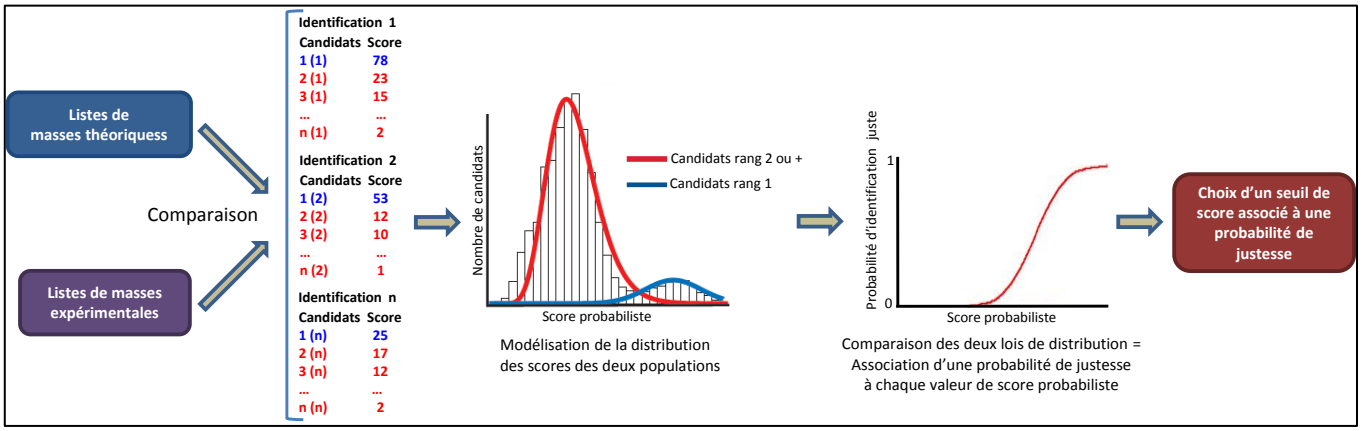


Figure 7 : *Illustration du principe de fixation de seuil par l'utilisation des approches empiriques de Bayes [180].*

Les résultats d'identifications réalisées dans la banque théorique/leurre contiennent alors des vrais et des faux positifs mélangés à des identifications de la banque leurre. Le choix d'un filtre de score probabiliste peut alors se faire en contrôlant le nombre d'identification de faux positif de la banque leurre par rapport au nombre d'identifications issues de la banque théorique. On parle alors de contrôle du taux de faux positifs.

La seconde approche, utilisant les approches empiriques de Bayes, consiste à utiliser les scores probabilistes obtenus pour l'ensemble des listes de masses candidates à une identification [180]. Ces scores ont été calculés par l'algorithme lors de la comparaison des listes de masses théoriques à chacune des listes de masses expérimentales soumises. Deux populations de scores sont alors considérées : les scores des listes de masses théoriques ayant conduit à une identification à un rang 1 et les scores des listes de masse ayant conduit à une identification à un rang 2 ou supérieur. Les distributions des scores des deux populations sont modélisées par deux lois binomiales. Par comparaison de ces deux lois binomiales, il est alors possible d'attribuer à chaque score probabiliste une probalité d'appartenir à l'une ou l'autre des populations (Figure 7).

Chapitre II Evaluation et développement d'une nouvelle stratégie expérimentale protéomique pour l'identification des isoformes du cheveu

L'analyse de la bibliographie et des données actuelles concernant la biologie du cheveu a montré qu'il existait des inconnues concernant l'expression des protéines, particulièrement des KAP, dans les types cellulaires constituant la fibre capillaire. La démonstration de leur expression avec les outils de l'analyse protéomique suppose d'obtenir des informations de séquences peptidiques spécifiques à chaque protéine issue des gènes correspondants.

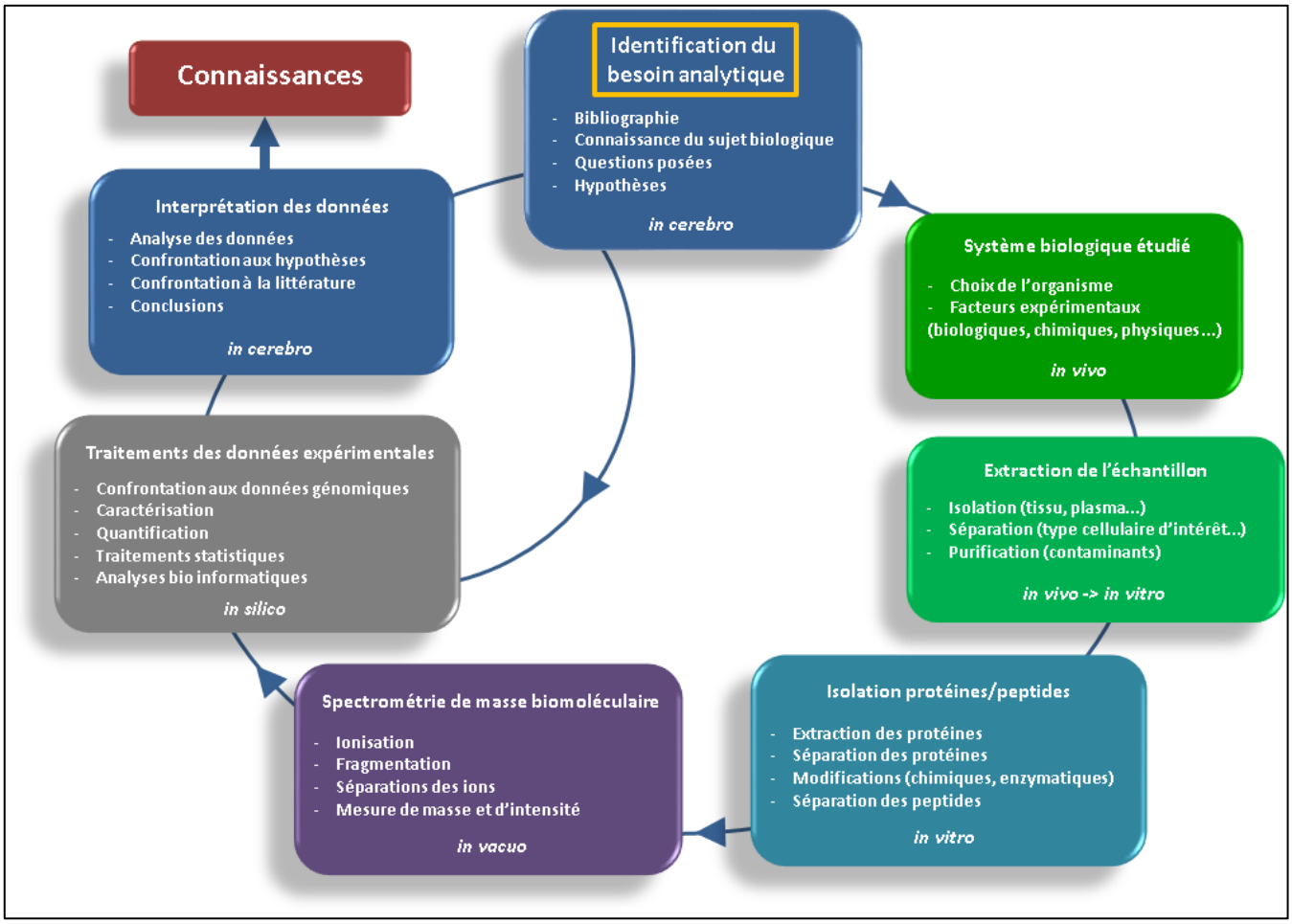


Figure 1 *: Les différentes étapes conduisant à l'acquisition de nouvelles connaissances des systèmes biologiques grâce aux stratégies de l'analyse protéomique combinées à la bioinformatique. Pour chaque étape, les événements clés qui y sont associés sont cités.*

L'obtention de ces séquences nécessite de s'inscrire dans une succession d'étapes expérimentales (Figure 1) :

- l'isolation spécifique des cellules ou des extraits cellulaires contenant le protéome d'intérêt (cortex, cuticule…),
- la décomplexification et la purification des extraits obtenus en passant par des étapes de séparation des protéines et/ou des peptides obtenus après digestion,
- le séquençage par spectrométrie de masse de l'ensemble des peptides issus de la digestion des protéines,
- le traitement des données de séquençage afin de mettre en évidence les protéines présentes dans les protéomes étudiés tout en y associant des informations quantitatives d'expression,

- l'interprétation des données d'expression et leur confrontation aux données de la littérature afin d'identifier de nouvelles informations permettant d'envisager de nouvelles hypothèses d'organisation des protéines dans ces cellules.

Précédemment dans la partie bibliographique, nous avons montré la difficulté d'accéder à la preuve de l'expression en protéine d'un ensemble de gènes de KAP provenant de familles multigéniques. La protéomique peut être utilisée pour résoudre cette problématique insolvable avec les stratégies d'immunohistochimie. Néanmoins, l'analyse des isoformes en protéomique n'est pas triviale et il va être nécessaire d'adapter les différentes étapes expérimentales dans le cadre de cette problématique.

Dans ce chapitre, nous commenterons les difficultés pouvant être rencontrées à chaque étape du processus expérimental et discuterons de la pertinence des stratégies analytiques pouvant être adoptées pour répondre à l'analyse des isoformes issues de familles multigéniques. L'ensemble de ces considérations nous permettra de conclure quant à la stratégie la plus adaptée pour répondre aux difficultés de l'étude du protéome du cheveu.

1) L'homologie de séquence : un rempart à la discrimination des isoformes

La difficulté d'obtenir des peptides antigéniques pour identifier spécifiquement par immunohistochimie une isoforme par rapport à une autre est la conséquence de leur homologie de séquence en acides aminés. Si l'homologie de séquence est trop importante, il n'est pas possible d'obtenir un anticorps spécifique à une isoforme particulière.
L'analyse de ces protéines par spectrométrie de masse peut permettre de discriminer des isoformes s'il existe entre leurs séquences quelques substitutions d'acides aminés. S'il est possible après digestion de la protéine d'obtenir des peptides portant le ou les acides aminés substitués, le séquençage de ces peptides permet l'identification sans ambiguïté de ces isoformes.
Plusieurs conditions doivent donc être réunies pour identifier spécifiquement une isoforme :

- La séquence doit posséder au moins une substitution (ou des insertions/délétions) d'acides aminés unique à l'isoforme. La substitution ne doit pas être isobare d'une autre isoforme.
- La séquence doit posséder des sites de coupures enzymatiques encadrant les sites substitués et permettant d'obtenir des peptides dits discriminants ou protéotypiques.
- Le peptide discriminant doit être transféré dans le spectromètre de masse, sélectionné et fragmenté de manière à obtenir une liste de masses spécifique et non ambiguë.
- Le gène correspondant a été décrit et la protéine théorique traduite se trouve dans la banque avec une séquence identique à celle de la protéine dans l'échantillon.
- L'algorithme de recherche parvient à attribuer la liste de masses expérimentales avec la liste de masses théoriques.

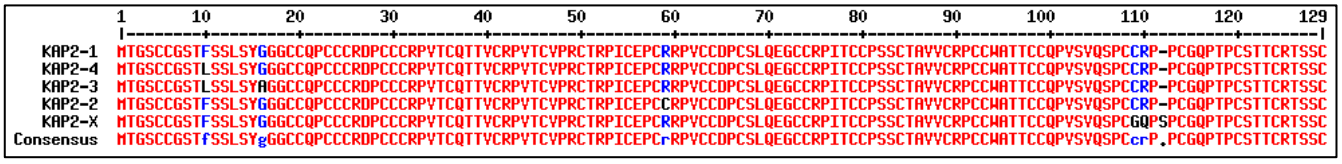


Figure 2 : *Illustration de l'homologie de séquence des séquences prédites des KAP de la famille 2. Séquences extraites de SwissProt.*

Cette difficulté justifie dans notre cas de concevoir une stratégie analytique permettant de satisfaire chacune de ces étapes. Dans le cas contraire, l'identification de l'isoforme sera compromise.
Dans le cas des isoformes de kératines, les séquences têtes et queues sont suffisamment dégénérées pour apporter des zones spécifiques d'identification. Ce sont ces séquences qui ont été précédemment utilisées pour

l'obtention de peptides antigéniques. Pour les KAPs, la majorité des familles présente de l'homologie tout le long de leur séquence. L'homologie peut être plus ou moins prononcée. Le cas extrême de la famille des KAP 2 peut être cité : moins de quatre sites de substitution peuvent être exploités pour étudier l'expression de 5 gènes potentiels (Figure 2).

2) Une approche multienzymatique pour augmenter le nombre de peptides protéotypiques

a) Les enzymes pour le séquençage par spectrométrie de masse

Les enzymes utilisées pour générer des peptides en protéomique sont des endoprotéases spécifiques issues d'extractions d'organismes. L'endoprotéase utilisée se fixe spécifiquement à certains résidus des séquences protéiques et catalyse l'hydrolyse de la liaison peptidique soit du côté N-terminal (N-ter) soit du côté C-terminal (C-ter).

Endoprotéase	Spécificité	Endoprotéase	Spécificité
Trypsine	**K**, **R** (C-ter)	Arg-C	**R**, K (C-ter)
Chymotrypsine	**F**, **W**, **Y**, L (C-ter)	**Asp-N**	**E**, **C**, D (N-ter)
Glu-C (V8)	**E**, **D** (C-ter)	**Lys-C**	**K** (C-ter)

Tableau 1 *: Tableau de différentes endoprotéases pouvant être utilisées pour le séquençage. L'activité et la spécificité de ces enzymes sont fonctions de la nature du tampon et de son pH ainsi que du ratio entre l'enzyme et le substrat. La nature du résidu situé de l'autre côté de la liaison peptidique hydrolysée peut influer sur cette activité.*

Le choix de ces enzymes se fait en fonction de l'abondance et de la distribution des résidus ciblés dans la séquence protéique. Les peptides obtenus doivent être suffisamment longs (>5-6 résidus) pour obtenir une séquence spécifique à la protéine. Le peptide ne doit cependant pas être trop long (<20-40 résidus selon les spectromètres) car il est dans ce cas plus difficile d'obtenir des informations complètes de leur fragmentation.
Compte tenu de l'abondance et de bonne répartition des résidus basiques (lysine et arginine) dans les protéomes et les propriétés d'ionisation et de fragmentation des peptides trypsiques (localisations des charges en positions N-terminales et en C-terminales du peptide ionisé en mode positif), la trypsine est couramment utilisée en protéomique.
Si les enzymes utilisées sont choisies pour leur spécificité, il convient tout de même de considérer que cette spécificité n'est pas absolue. Des coupures non spécifiques peuvent se produire minoritairement et des peptides dits semi-spécifiques existent alors dans le digest. A l'inverse, certaines coupures attendues sont cinétiquement moins favorables que d'autres. Il convient ainsi d'envisager dans le digest des peptides contenant ces sites de coupures manquantes.

b) Une stratégie adaptée pour l'étude des kératines et des KAPs

Les kératines ne sont pas des protéines dont l'identification représente une difficulté analytique particulière. La présence d'un nombre et d'une répartition équilibrée de sites de coupure enzymatique permet d'envisager un recouvrement correct de leur séquence avec la trypsine. Les peptides trypsiques de kératines issues de la dégradation de la couche cornée de l'épiderme des manipulateurs sont connus pour être une source de contamination des échantillons analysés en protéomique. Néanmoins, la caractérisation d'une protéine, c'est-à-dire l'étude de l'ensemble de sa séquence suppose d'obtenir des peptides permettant de recouvrir l'ensemble de cette séquence. Dans ce cas, l'utilisation d'une seule enzyme comme la trypsine ne permet pas d'envisager une telle caractérisation. Le recours à des analyses des protéines avec l'utilisation d'autres enzymes complémentaires doit être alors envisagé [198].
Dans le cas d'une partie des familles de KAPs, la composition en acides aminés n'est pas usuelle et certaines portions de leurs séquences sont presque dépourvues de sites de clivage trypsique (KAP 1, 3, …). Pour d'autres de

ces familles, la présence régulière de prolines du côté C-terminal du site de coupure trypsique (KAP 4, 2 et 1) laisse envisager une cinétique très défavorable pour la génération des peptides correspondants. Pour ces raisons, l'utilisation d'autres enzymes a été envisagée. L'analyse des séquences des protéines ciblées montre la présence de résidus acides (KAP 1, 3, 4) ou de résidus hydrophobes (KAP 1, 3, 11, 13, 7, 8, 19) pouvant être ciblés pour des digestions à la Glu-C ou à la chymotrypsine.

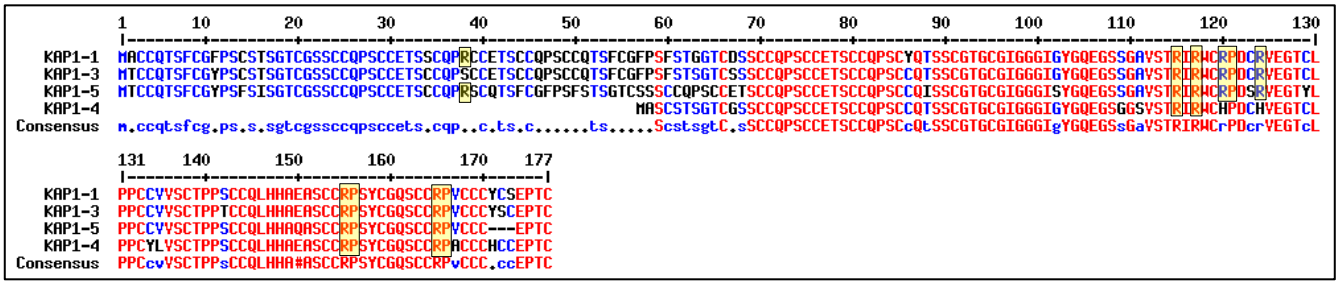


Figure 3 : *Alignement des séquences des KAPs décrites pour la famille 1 illustrant la faible abondance et la mauvaise répartition des sites de coupures trypsiques dans ces séquences. Nous noterons la présence de site impliquant la proline du côté C-terminal.*

Il parait ainsi fondamental pour tenter de contourner la difficulté d'obtention de peptides de ne pas se restreindre à l'étude des digests trypsiques de cheveu. Ainsi, nous utiliserons par la suite la complémentarité des enzymes pour réaliser nos caractérisations de mélanges d'isoformes.

3) L'extraction des protéines des structures kératinisées

a) Principe d'extraction des protéines du cortex

La structure corticale est maintenue dans la fibre par un vaste réseau de ponts disulfures. La solubilisation des protéines du cortex nécessite de détruire ce réseau de liaisons inter et intra chaines. La réduction de ces liaisons formées par oxydation des cystéines est couramment réalisée par l'utilisation d'agents réducteurs tels que le dithiothreitol (DTT), le mercaptoéthanol ou l'acide thioglycolique (le thioglycolate d'ammonium est utilisé pour la réalisation des permanentes). La réduction est réalisée en présence d'agents chaotropes permettant de déstructurer les protéines des édifices macrofibrillaires et de faciliter leur extraction (Figure 4).

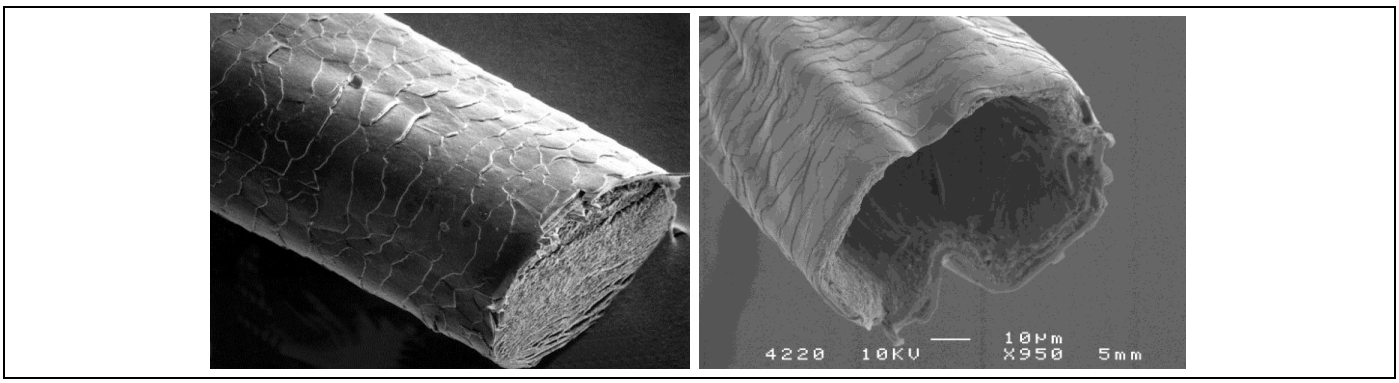


Figure 4 *: Clichés de microscopie électronique à balayage d'une fibre de cheveu avant et après extraction réductrice avec des agents chaotropes. La fibre est vidée de ses cellules corticales, les cellules cuticulaires restent intactes. (Crédits photographiques : L'Oréal recherche).*

Afin d'éviter la réoxydation des ponts disulfures lors de l'exposition prolongée de l'échantillon à l'air, l'alkylation des cystéines libres est réalisée en présence d'un agent alkylant. Ces réactifs, couramment utilisés en protéomique peuvent être l'iodoacétamide ou l'acide iodoacétique. L'ajout de ces groupements aux résidus des

protéines modifie leurs propriétés physico-chimiques. Par exemple, la solubilité des KAPs riches en soufre peut être sensiblement augmentée par réaction de l'acide iodoacétique compte tenu de leur abondance en cystéines.

b) Principe d'extraction des protéines de la cuticule

L'extraction des protéines de la cuticule est beaucoup plus compliquée à réaliser. Les techniques de réduction/extraction des protéines des cellules corticales n'affectent pas la cuticule. Cette insolubilité est liée au rôle biologique des cellules cuticulaires qui permettent de maintenir la structure de la fibre même en présence d'agents chimiques. Elle peut être attribuée à la l'imperméabilité, à la structure des cellules cuticulaires et à la densité des liaisons inter protéines au sein de ces cellules.
Une des techniques d'isolement des cellules cuticulaires est d'utiliser cette différence de solubilité. Par extraction répétée du cortex, l'insoluble restant peut être considéré comme de la cuticule. L'insoluble restant peut alors être digéré afin d'obtenir des peptides permettant d'étudier les protéines cuticulaires. Elle exclut l'utilisation par la suite de techniques de manipulation et de séparation au niveau protéique.
Les techniques pouvant être développées pour accéder au protéome cuticulaire seront décrites en troisième partie de ce manuscrit dans le cadre de l'étude de ce protéome.

4) Evaluation de l'efficacité de l'étude d'isoformes par séparation au niveau protéique

Nous avons précédemment discuté de la nécessité de décomplexifier les échantillons afin de compenser certaines limites de l'analyse directe des digests peptidiques par spectrométrie de masse. Différentes techniques sont utilisées dans ce but en protéomique. Nous avons souhaité évaluer ces techniques et nous discuterons des avantages et des inconvénients de chacune ainsi que de la cohérence de leur utilisation dans le cadre de notre problématique de séparation des isoformes des cheveux. Cette évaluation sera réalisée sur la base de simulations réalisées grâce à la connaissance des séquences protéiques attendues. L'évaluation sera également complétée d'expériences de séparation des protéines avec différentes techniques usuelles. Nous présenterons et commenterons ces expériences qui ont été en partie réalisées dans le cadre du travail de thèse d'Audrey Bednarczyck [199].

a) Principe des stratégies de séparation des protéines

Les techniques de séparation des protéines basées sur leurs propriétés physico-chimiques sont couramment utilisées pour la décomplexification de l'échantillon avant l'analyse par spectrométrie de masse. La composition différente en acides aminés des protéines permet d'envisager différences stratégies de séparation :

- Les différences de poids moléculaire des protéines impliquent une grande diversité de tailles. Les techniques d'électrophorèse sur gel monodimensionnel avec Sodium Dodecyl Sulfate (SDS) comme détergent [200] et de chromatographie d'exclusion stérique (chromatographie d'exclusion stérique, SEC, ou chromatographie de perméation de gel, GPC) utilisent les différences d'encombrement stérique pour réaliser la séparation.
- Les différences de composition en acides aminés basiques et acides impliquent des états de protonation des protéines en solution différents qui varient selon les conditions de tampon. Pour chaque protéine, il existe un pH pour lequel la somme des acides aminés cationiques (basiques protonés) est égale à la somme des acides aminés acides anioniques (acides déprotonés). La charge de la protéine à ce pH est globalement neutre : il correspond à son point isoélectrique (pI). Cette propriété est utilisée pour les techniques d'isoélectrofocalisation qui consistent à faire migrer par électrophorèse les protéines au travers d'un système composé de gradients de pH. Lorsque la protéine atteint la zone de pH

correspondant à son pI, sa neutralité la rend insensible à la différence de potentiel électrique appliqué ce qui interrompt sa migration.

- Les différences en composition en acides aminés sont à l'origine de différence d'hydrophobicité/hydrophilicité entre les protéines. Les techniques de chromatographie de partage peuvent être ainsi utilisées pour un préfractionnement des protéines. Différents modes chromatographiques peuvent être envisagés comme les chromatographies en phase inverse et d'échange d'ions.
- La solubilité des protéines dépend des interactions qui existent entre les résidus de la protéine et le solvant. La modification des compositions de solvant peut permettre de précipiter différentiellement des groupes de protéines.

b) Evaluation théorique du potentiel des techniques de séparation protéiques par modélisation de la physico-chimie des protéines étudiées

Avant d'évaluer expérimentalement le potentiel de l'électro focalisation des protéines étudiées, nous avons souhaité prédire théoriquement les profils électrophorétiques pouvant être attendus notamment pour les isoformes recherchées.

Les propriétés physico-chimiques des protéines peuvent être modélisées grâce à la connaissance de leur structure primaire prédite. Les masses moléculaires et les points isoélectriques attendus sont calculés grâce à des outils prédictifs utilisant les séquences en acides aminés (« compute pI/MW » disponible en ligne dans les outils du serveur protéomique Expasy). Pour tenir compte de la modification de ces propriétés qu'entraine l'alkylation des cystéines des protéines étudiées, nous avons substitué pour chacune de leur séquence les cystéines en acide glutamique pour simuler la carboxyméthylation ou en glutamine pour simuler la carbamidométhylation. La compilation de ces résultats peut être illustrée par des profils électrophorétiques théoriques (Figure 5).

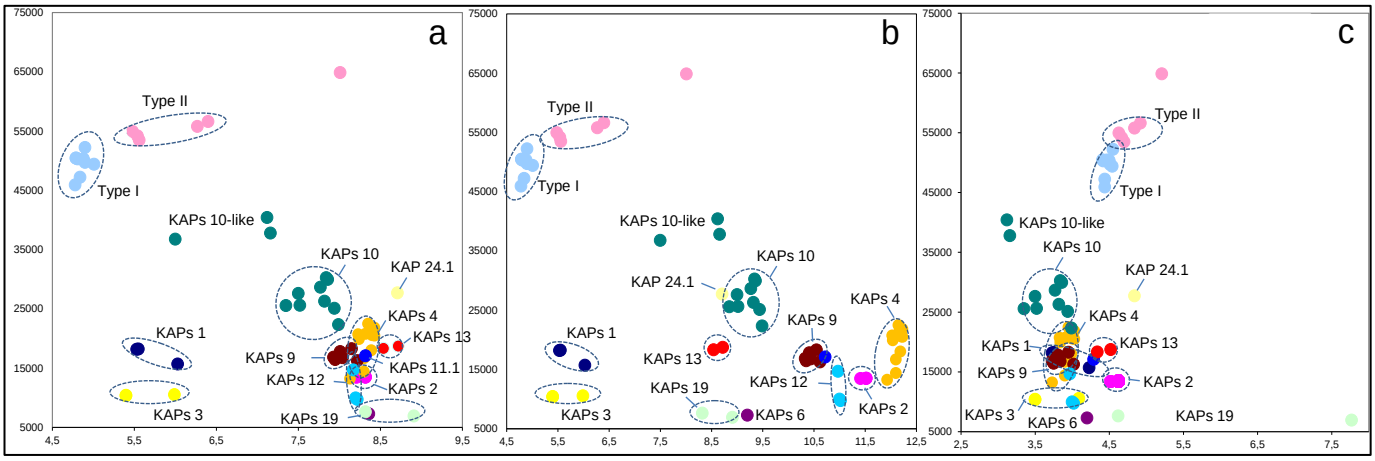


Figure 5 *: Profils électrophorétiques théoriques des kératines et des KAP pouvant être attendues dans le cheveu. En fonction de la dérivation des cystéines, les points isoélectriques des protéines (en abscisse) et notamment de certaines familles de KAPs sont significativement modifiés. a) Sans alkylation b) Alkylation avec l'iodoacétamide c) Alkylation avec l'acide iodoacétique.*

Cette modélisation montre comment le choix de l'agent alkylant peut théoriquement impacter sur les propriétés des familles de KAP. Sans modifications, les KAP riches en soufre des familles 2, 4, 9 sont relativement basiques (pI ≈ 8). Cette basicité augmente significativement avec l'ajout important de fonctions amides (plus de 65 résidus modifiés en moyenne pour les KAP 4) et ces mêmes protéines ont alors un pI théorique supérieur à 10. Ce résultat de modélisation est plutôt inattendu compte tenu de l'absence de propriétés acido-basiques du résidu ajouté. Cette modification des points isoélectriques semblerait plutôt attribuable à la perte de la fonction thiol qui elle est protonable. A l'inverse, l'ajout de fonctions acides aux résidus cystéines entraine pour l'ensemble des KAP

modélisées une diminution cruciale du pI vers des pH acides (pI < 4). Les kératines ont moins de cystéines et sont plus grandes en taille. Leurs pI sont ainsi moins affectés par la modification de quelques résidus dans la séquence. Quelle que soit la stratégie adoptée pour marquer ou non les cystéines, il apparait difficile mais néanmoins envisageable de séparer par différence de points isoélectriques et de masse les différentes familles de protéines. Il semble par contre impossible de séparer les différentes isoformes prédites au sein de chaque famille. Cette difficulté est en partie à l'origine de la méconnaissance du nombre de gènes exprimés pour chaque famille de KAP dans le protéome du cheveu.

c) Evaluation expérimentale de l'isoélectrofocalisation des protéines pour le préfractionnement par approche off-gel.

Après avoir évalué théoriquement les différences de physico-chimie des protéines ciblées, nous les avons confronté à des profils électrophorétiques pouvant être obtenus expérimentalement sur des extraits de cortex.
Le fractionnement off-gel permet la séparation des protéines en fonction de leur point isoélectrique. Les protéines soumises à un champ électrique migrent en solution en passant à travers des puits contenant une solution d'ampholyte [201]. Le passage de puits en puits est réalisé par l'intermédiaire d'une bande de gel à gradient de pH immobilisé (IPG). En fin de migration, les protéines sont isolées dans un puits correspondant à leur point isoélectrique.
L'extrait de cortex a ainsi été fractionné en douze fractions. Les masses moléculaires des protéines fractionnées dans chaque puits ont été analysées par SDS PAGE 12% (Figure 6).
Les résultats de cette expérience montrent qu'il est possible de séparer les kératines de type I des kératines de type II. Cette séparation est rendue possible par les différences de masses de ces deux groupes, conséquence des différences dans les longueurs de chaines entre les segments têtes et queues respectifs. Les différences de points isoélectriques s'expliquent par le fait que les kératines de type II possèdent dans leur séquence de tige plus de résidus basiques que les kératines de type I, les proportions de résidus acides entre les deux familles étant environ les mêmes.

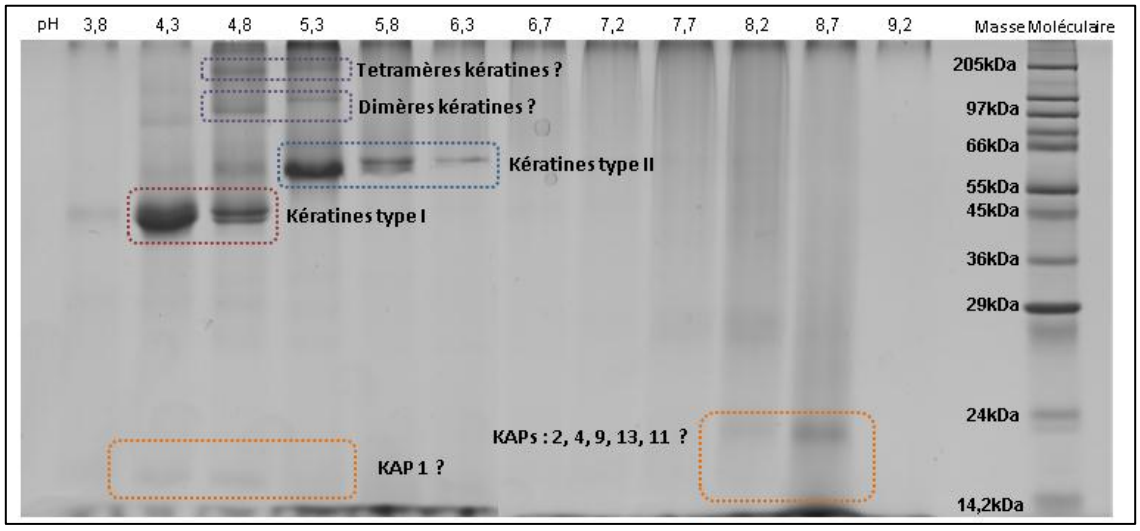


Figure 6 *: Gel SDS-PAGE 12% de 12 fractions off-gel d'un extrait cortical réduit/carbamidométhylé. Ampholytes pH 3-10. Coloration au bleu de Coomassie. D'après les travaux d'Audrey Bednarczyck.*

Au sein de ces familles, nous pouvons distinguer des dédoublements en termes de masses moléculaires. Ces dédoublements peuvent s'expliquer par des différences de tailles des segments de queue pour les kératines de type I (par exemple plus long pour K31 que pour K33a/K33b) ou des différences de taille des segments de tête pour les kératines de type II (plus long pour K85 que pour K86, K83 et K81). Nous pouvons noter, dans des zones de masses moléculaires correspondant respectivement à deux fois et à quatre fois la masse d'une kératine, la

présence de bandes de coloration. Nous noterons également qu'au moins deux espèces de masse moléculaire et de point isoélectrique différents existent dans la région correspondant à 100 kDa environ. L'analyse de ces bandes a montré la présence simultanée de kératines de type I et de type II (résultats non montrés). La présence de ces protéines dans ces zones du gel peut être expliquée par des dimères et des tétramères constituants initialement les filaments intermédiaires extraits. Le fait qu'une minorité de ces unités structurales soit toujours constituée même après réduction et alkylation pourrait être la conséquence d'une à plusieurs liaisons covalentes non réductibles décrites en première partie.

Si l'examen dans les zones de migration correspondant aux kératines apporte des informations, il apparait en revanche beaucoup moins évident d'étudier les KAP avec cette technique. En effet, ces dernières ne présentent que de faibles colorations dans les zones de migration attendues.

d) Evaluation expérimentale de l'analyse des protéines par électrophorèse bidimensionnelle.

Pour compléter les observations réalisées lors de l'évaluation de la technique de préfractionnement off-gel, nous avons évalué le potentiel de la séparation des extraits par électrophorèse bidimensionnelle. Les résultats obtenus montrent que les signaux correspondant aux kératines tendent à prédominer sur le gel (Figure 7). Il est néanmoins possible de distinguer dans des zones correspondant à de plus faibles masses moléculaires quelques spots dont la localisation peut correspondre aux KAP 1 et 3. L'identité de ces spots a été confirmée par analyse protéomique des digests de ces spots découpés. Nous noterons que l'identification de ces protéines recherchées pour notre étude n'est réalisée qu'avec un nombre restreint de peptides ce qui rend difficile l'assignation des spots à une isoforme particulière de la famille plutôt qu'à une autre. Les couvertures de séquences obtenues sont ainsi insuffisantes pour réaliser une caractérisation de ces protéines en passant par cette technique.

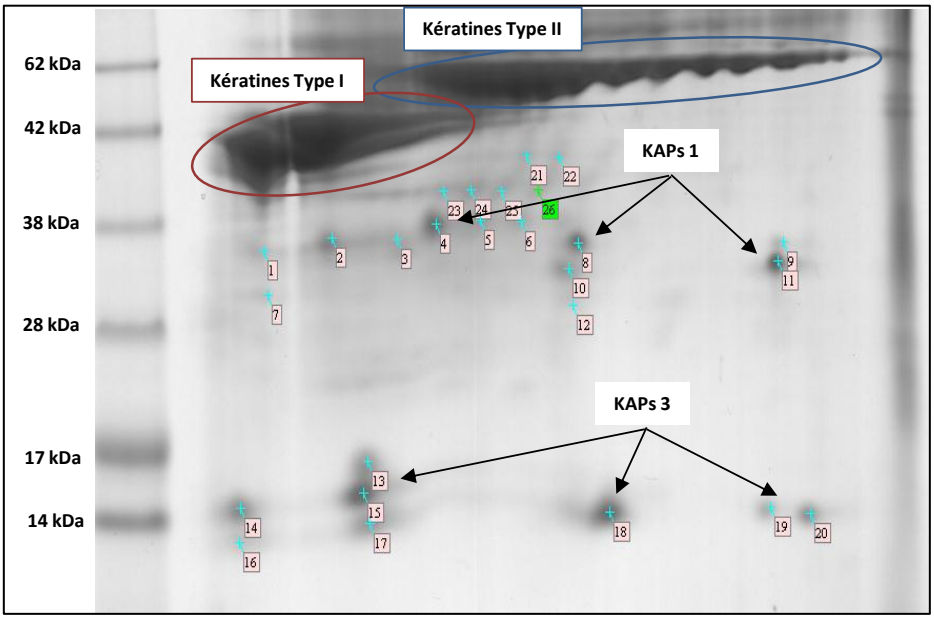


Figure 7 *: Profil d'électrophorèse bidimensionnelle d'un extrait de cortex réduit et alkylés à l'iodoacétamide. D'après les travaux d'Audrey Bednarczyck.*

Le nombre limité de peptides obtenus lors de l'analyse de ces spots peut s'expliquer par la faible quantité de matériel initialement présent dans chaque spot. Il peut également être attribué à des rendements de digestions insuffisants pour obtenir une identification par spectrométrie de masse.

Récemment, Plowman et *al*. ont montré qu'il était possible d'optimiser les conditions expérimentales pour privilégier l'extraction de KAP de fibre de laine par rapport aux kératines. Les extraits, ainsi déplétés en partie des

kératines, peuvent alors être analysé par électrophorèse bidimensionnelle. Ces analyses montre qu'il est possible de détecter des KAP minoritaires [94].
S'il semble que des développements soient possibles pour améliorer l'analyse des extraits corticaux par séparations électrophorétiques, les différents essais d'analyse suggèrent que leur emploi limite l'identification des KAP notamment celles très riches en soufre et basiques.

e) Evaluation expérimentale de l'analyse des protéines par chromatographie d'exclusion

La séparation des protéines peut être réalisée sans avoir recours à des techniques d'électrophorèse. L'utilisation de la chromatographie d'exclusion a ainsi été évaluée. En solubilisant les extraits corticaux réduits et alkylés, la chromatographie de l'extrait peut être effectuée, les protéines pouvant être détectées par absorption dans l'ultraviolet.
La chromatographie d'exclusion implique un ordre d'élution des protéines les plus grandes en tailles suivi des protéines les plus petites. Ainsi les kératines, plus grandes, sont éluées les premières et donnent un pic intense. La résolution du système n'est cependant pas suffisante pour séparer les kératines de type I des kératines de type II. La suite du profil chromatographique montre que des protéines plus petites, les KAPs, sont éluées mais la résolution est insuffisante pour permettre de séparer les différentes isoformes. Nous noterons qu'un pic intense est observé entre 8 et 9 minutes qui pourrait correspondre à une famille de KAPs dont les isoformes possèdent des masses communes.

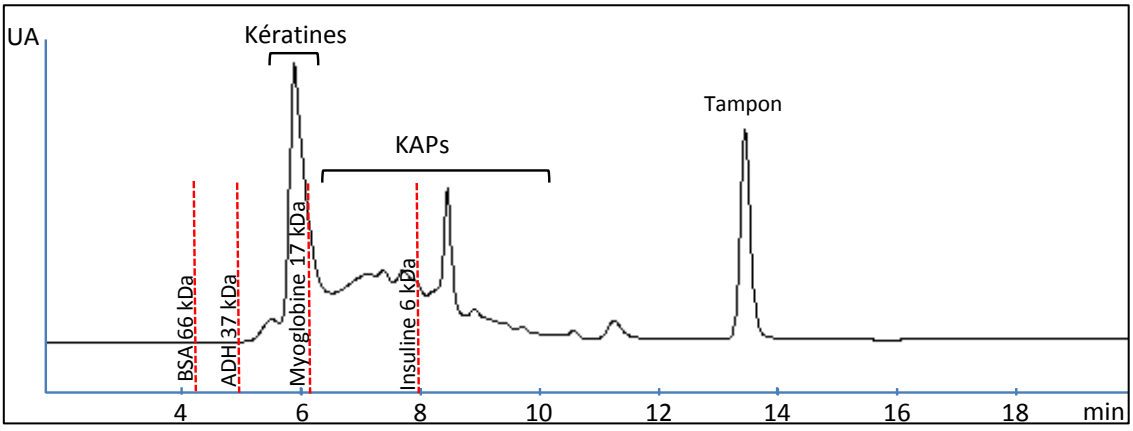


***Figure 8** : Chromatogramme SEC-UV d'un extrait cortical réduit/carbamidométhylé dialysé puis resolubilisé. Les temps d'élution des standards protéiques utilisés pour calibration sont indiqués. Colonne Phenomenex Biosep Sec-S2000, 7.8 x 300mm. Phase mobile Tris-HCl 20mM, pH 7.5. Débit 1mL /min. Détection UV à 280 nm. D'après les travaux d'Audrey Bednarczyck.*

Cette technique, même si elle parait peu résolutive, peut néanmoins être envisagée pour séparer les kératines des KAPs tout en gardant les protéines en solution.

f) Les techniques de précipitations sélectives

Bien que nous n'ayons pas évalué cette technique, nous pouvons tout de même noter l'existence d'un protocole de précipitations sélectives basé sur les différences de propriétés physico-chimiques des familles de protéines constituant la laine [108] (Figure 9). Cette technique utilise notamment les différences de proportions en cystéines des KAPs riches en soufre par rapport aux kératines et aux KAPs riches en glycine et tyrosine. En réalisant la carboxyméthylation des cystéines de l'échantillon, les KAPs riches en soufre possèdent alors un nombre important de fonction acides augmentant considérablement leur solubilité. En augmentant la force ionique dans l'extrait, les protéines les moins hydrophiles précipitent. Ainsi les KAPs riches en soufre restent en solution tandis que les autres protéines de l'extrait sont précipitées.

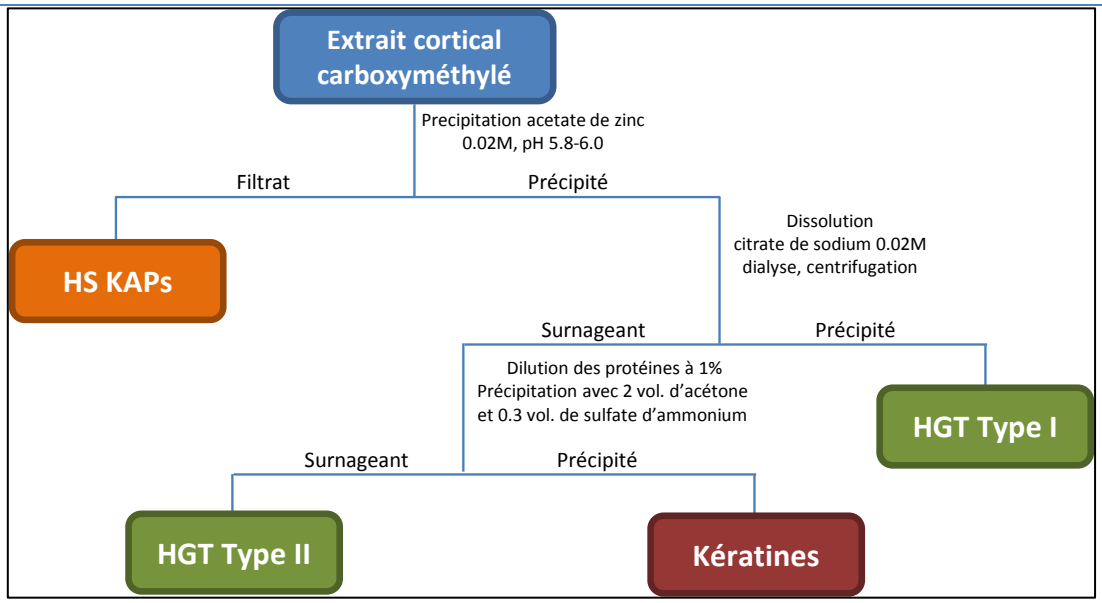


***Figure 9** : Principe du protocole de séparation des différentes familles de protéines de l'extrait de cortex par utilisation de leur propriétés physico-chimiques [108].*

Il est par la suite possible de resolubiliser une partie du précipité protéique. L'insoluble restant est décrit comme riche en KAPs riches en glycine et tyrosine. Les kératines peuvent alors être précipitées en ajoutant un solvant organique.

g) Conclusion : limites des approches de séparation des protéines pour l'analyse des isoformes du cheveu

L'ensemble de ces stratégies permet de discriminer une protéine par rapport à une autre à condition que la masse moléculaire et la composition en résidus soient suffisamment différentes. Dans le cas d'isoformes issues de familles multigéniques, ces différences, parfois de seulement quelques acides aminés, ne suffisent généralement pas à apporter une séparation suffisante pour répondre à nos problématiques d'identification et de caractérisation. D'autres part, il semble que la détection des KAP après passage de l'échantillon sur gel d'électrophorèse soit limitée.
Il semble néanmoins possible d'envisager la séparation de certaines familles d'isoformes les unes des autres et d'utiliser ces techniques à des fins de décomplexification de l'échantillon.
Ces stratégies supposent également qu'il soit possible d'extraire les protéines intactes des tissus étudiés. Dans le cas des protéines de la cuticule densément réticulées, il parait difficile d'envisager leur extraction et seule la digestion *in situ* peut alors apporter des informations de séquence.
Nous avons ainsi fait le choix de ne pas utiliser dans le cadre de la problématique de séparation des protéines pour nous affranchir des problèmes de leur manipulation. Suite à ces évaluations, il nous a semblé plus adapté de travailler à la séparation et à l'analyse des peptides obtenus après digestion directe de l'échantillon.

5) Evaluation de l'efficacité de décomplexification au niveau peptidique

a) L'analyse des peptides comme alternative à l'isolement des protéines

Dans le cadre de ce travail de thèse, nous avons choisi de nous focaliser sur la recherche de peptides protéotypiques sans passer par une étape de séparation des protéines. Cette stratégie est baptisée « shotgun » . L'échantillon est digéré enzymatiquement ce qui permet d'obtenir un mélange de peptides contenant notamment les peptides protéotypiques que nous cherchons à mettre en évidence. L'inconvénient du shotgun est

la perte du lien entre protéine et peptides dès lors que ces derniers sont partagés potentiellement par plusieurs protéines [202].
L'absence de décomplexification de l'échantillon à cette étape introduit une complexité importante de l'échantillon qui ne peut être simplement compensé par l'unique dimension de séparation des peptides lors de l'analyse LC-MS. Les peptides ont une meilleure solubilité que les protéines et sont plus faciles à séparer. Pour décomposer l'échantillon en plusieurs échantillons de peptides décomplexifiés, il est possible d'utiliser en amont des techniques de préfractionnement [203]. Pour qu'il soit efficace, le préfractionnement doit être complémentaire à la dimension de chromatographie utilisée en couplage avec la spectrométrie de masse.

b) Les techniques multidimensionnelles de séparation des peptides

La technique de séparation des peptides la plus courante est la chromatographie. Le fractionnement utilisant l'électrophorèse off-gel peut également être cité. L'ensemble de ces techniques s'appuie, comme pour les protéines, sur les différences de composition des résidus des séquences peptidiques.
Les différences de charges des peptides en solution peuvent être utilisées pour réaliser leur séparation grâce aux modes chromatographiques par échanges d'ions (chromatographies d'échange de cations ou d'anions) ou par électrophorèse. Les différences hydrophobicité et d'hydrophilicité des peptides peuvent également être utilisées (respectivement, chromatographie en phase inverse et chromatographie d'interaction hydrophile). Chacune de ces techniques étant complémentaires aux autres, il est donc possible de combiner successivement plusieurs séparations pour la décomplexification [204]. Les différences d'efficacité de séparation inhérentes aux modes chromatographiques et l'orthogonalité des modes chromatographiques les uns par rapport aux autres impliquent que certaines associations soient plus performantes que d'autres [205]. Le terme de chromatographie bidimensionnelle est utilisé lorsqu'une étape de fractionnement chromatographique précède l'analyse par LC-MS. Trois stratégies peuvent être employées pour la réaliser [206, 207] :

- La stratégie dite « online » consiste à réaliser la chromatographie en première dimension et à transférer séquentiellement les éluats vers la seconde dimension de séparation. La collecte des fractions de première dimension est réalisée au moyen d'une interface composée de deux espaces de stockage réalisant alternativement la collecte et le transfert vers la seconde dimension. L'interface peut être constituée par deux boucles d'échantillonnage ou par deux précolonnes. Ce système nécessite de réaliser la séparation sur la seconde dimension pendant le temps correspondant au temps de stockage de l'éluat de la première dimension. Cette configuration limite les possibilités d'augmentation du temps de séparation sur la deuxième dimension qui, nous le verrons, peut être déterminant pour améliorer la qualité de la séparation.

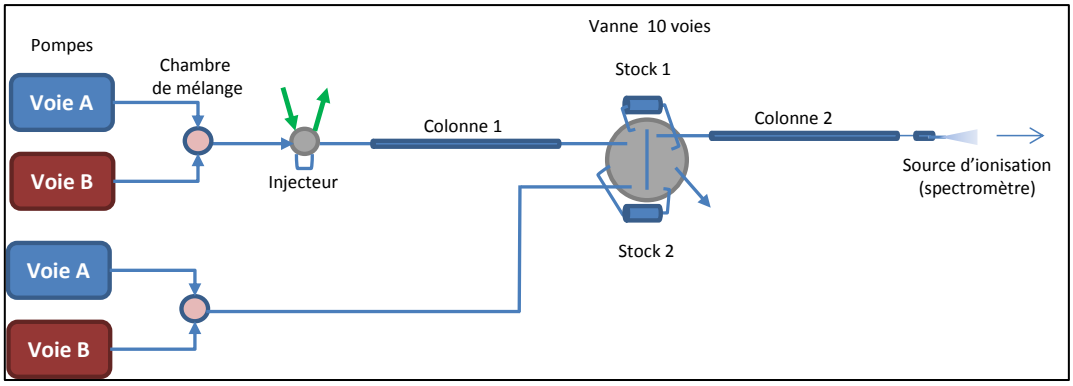


Figure 10 : *Montage utilisé pour la chromatographie bidimensionnelle « online ».*

- La stratégie dite « stop-and-go » consiste à réaliser séquentiellement la première dimension de séparation et à transférer l'éluat vers la deuxième dimension de séparation. Pendant qu'est réalisée la deuxième dimension de chromatographie, la première dimension est à l'arrêt et les peptides n'ayant pas été élués sont piégés sur

la première colonne (peak parking). Lorsque l'analyse en seconde dimension est terminée, une séquence d'élution sur la première dimension est à nouveau réalisée. Cette technique présente l'inconvénient de perdre en capacité de séparation sur la première dimension mais permet d'atteindre des niveaux de décomplexification supérieurs à la précédente.

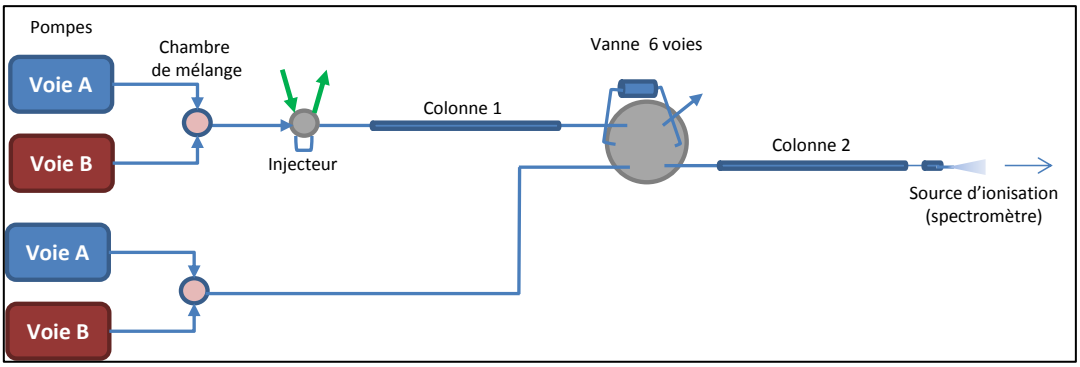


Figure 11 : *Montage utilisé pour la chromatographie bidimensionnelle en « stop-and-go »*

- La stratégie dite « offline » consiste simplement à collecter à l'aide d'un collecteur des fractions de la première dimension pour les analyser individuellement en deuxième dimension. Cette stratégie s'affranchit des contraintes de temps ce qui offre la possibilité d'optimiser indépendamment les deux systèmes chromatographiques et d'adapter le nombre de collectes de la première dimension.

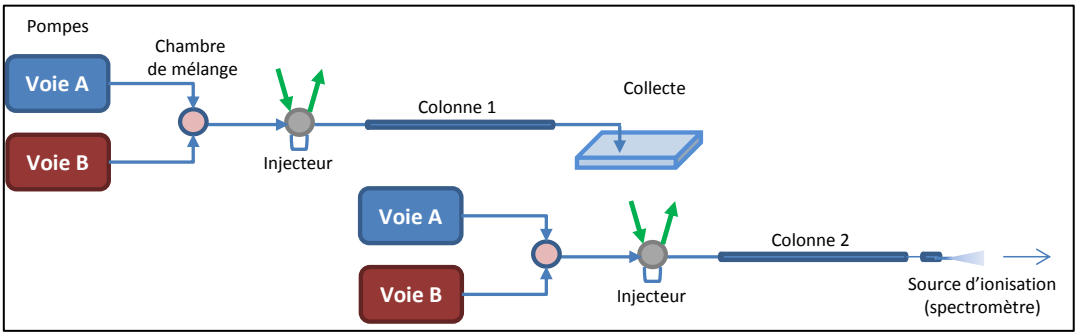


Figure 12 : *Montage utilisé pour la chromatographie bidimensionnelle « offline ».*

Chacun de ces systèmes peut être optimisé pour obtenir des capacités de séparation accrues. Celles-ci seront toujours obtenues en augmentant le temps global de l'analyse mais l'optimisation peut permettre de conserver la même efficacité de la séparation en réduisant le temps d'analyse [208-210].

c) Principe et évaluation de techniques de chromatographie bidimensionnelles

Nous utiliserons pour la suite des travaux une stratégie de préfractionnement avec, en première dimension, une séparation par chromatographie en phase inverse à pH basique [211]. Ce mode chromatographique apparait comme une alternative à l'utilisation plus courante de la chromatographie d'échange de cation [212]. La stratégie de chromatographie bidimensionnelle utilisée sera « offline ». L'orthogonalité entre la phase inverse à pH basique en première dimension et la phase inverse à pH acide est basée sur la différence des états de protonation des peptides en fonction du pH de la phase mobile utilisée. La première dimension est réalisée à pH 10 en présence d'un agent de paire d'ion cationique (ammonium, triéthylammonium). En seconde dimension, la modification de l'état de charge sur les fonctions amine, phénol et acide carboxylique et l'utilisation d'un agent de paire d'ion anionique (formate, trifluoroacétate) modifie suffisamment l'hydrophobicité des peptides pour que des composés coélués en première dimension soient séparés en seconde dimension.

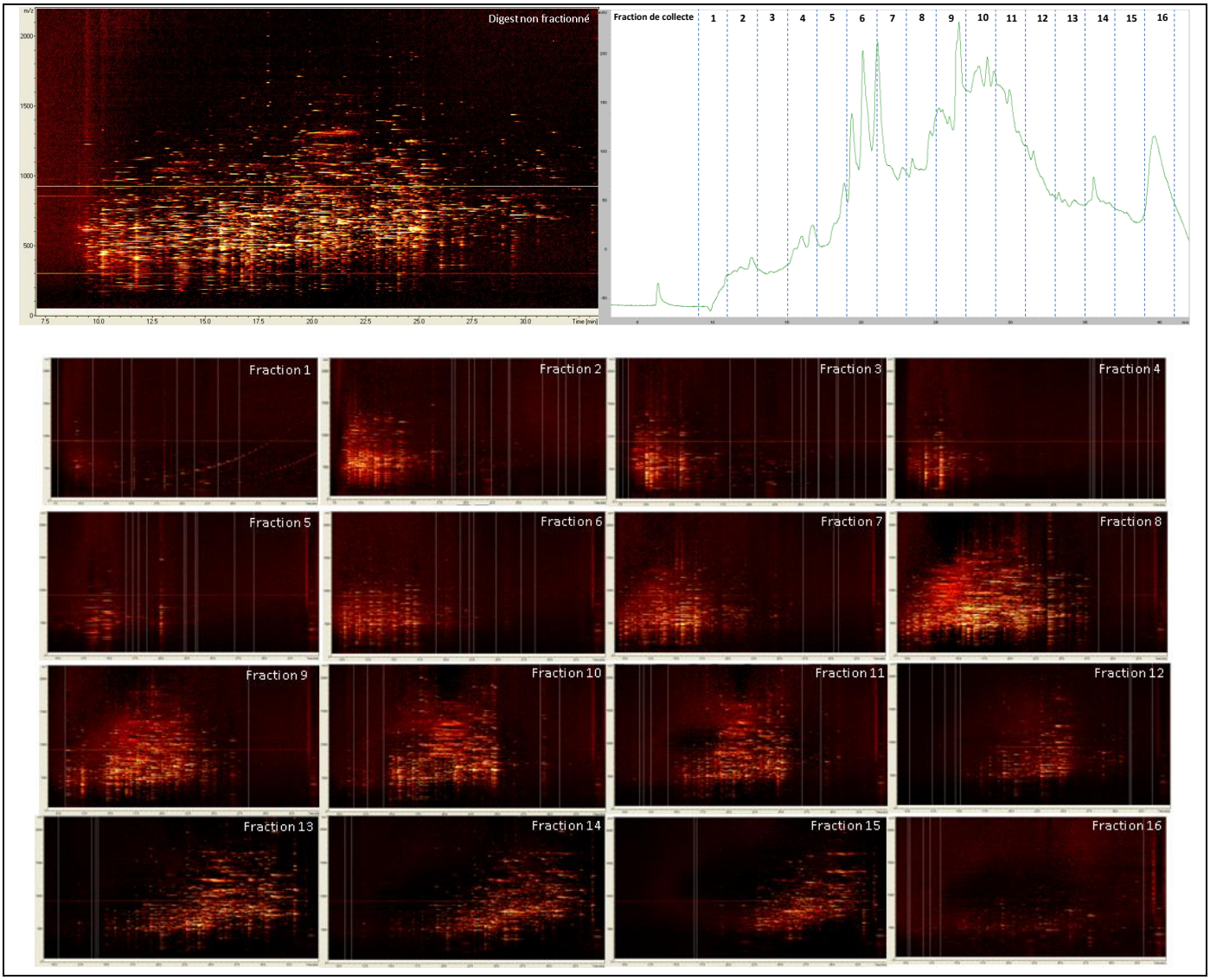


Figure 13 *: En haut à gauche, profil LC-MS obtenu en phase inverse à pH acide du digest trypsique d'un extrait de cortex. En haut à droite, chromatogramme en détection UV du digest séparé et collecté en première dimension à pH basique. En bas, profils LC-MS des différentes fractions séparées en deuxième dimension à pH acide.*

Si l'orthogonalité entre les deux modes chromatographique est suffisante pour éliminer la plupart des coélutions, nous noterons que les fractions obtenues en première dimension restent intrinsèquement d'hydrophobicité voisine en seconde dimension. L'analyse des différentes collectes en utilisant le même gradient chromatographique montre que les peptides ne sont pas toujours répartis sur l'ensemble du chromatogramme (Figure 13). Une solution permettant de limiter le temps pendant lequel aucun peptide n'est élué consiste à rassembler deux à deux des collectes complémentaires [213].

L'intérêt de la décomplexification apportée par le préfractionnement a été évalué. Les résultats d'identification obtenus par l'analyse d'un digest de cortex préfractionné en douze collectes ont été comparés avec les résultats obtenus par l'analyse répétée du même échantillon mais sans préfractionnement. Dans ce second cas, nous avons choisi d'améliorer l'acquisition de données en utilisant une stratégie de fractionnement en phase gazeuse (GPF). Le GPF consiste à ne sélectionner les peptides à fragmenter que sur une gamme restreinte de masses d'ions parents ce qui permet de diminuer les phénomènes de sous échantillonnage lors de l'acquisition [214, 215]. Les

répétitions de l'analyse de cet échantillon ont ainsi été réalisées sur quatre domaines de masse (définis selon [216]) chacun répété trois fois.

La comparaison des deux séries d'analyses montre qu'avec le même temps d'analyse et une méthode d'acquisition similaire, un gain significatif est obtenu grâce au préfractionnement par chromatographie de l'échantillon (tableau 2). Le nombre de spectres acquis est tout d'abord plus important bien que cette différence puisse être une conséquence des différences de quantités injectées entre les deux séries. Néanmoins, en proportion de la population de spectres examinée, ceux obtenus sur les fractions de chromatographies permettent d'obtenir plus de spectres et de peptides assignés. Nous montrons que le nombre de peptides uniques est considérablement plus important en décomplexifiant l'échantillon avant l'analyse. Cette différence très significative peut en partie être attribuée à la diminution du nombre de coélution.

	nanoLC-MS/MS : Fractionnement en phase gazeuse (GPF)	LC-nanoLC-MS/MS Fractionnement par chromatographie liquide	Gain LC-2D/GPF
Temps d'analyse MS/MS	12 x 50min	12 x 50min	=
Spectres acquis	6609	9299	+40%
Spectres assignés	985	1555	+58%
Peptides assignés	169	284	+68%
Peptides uniques	94	189	+101%

***Tableau 2** : Illustration de l'intérêt du préfractionnement des échantillons peptidiques avant analyse en couplage nanoLC-MS/MS par comparaison de l'analyse du même échantillon avec une stratégie de fractionnement en phase gazeuse ou une stratégie de fractionnement par chromatographie bidimensionnelle. Critères de filtre des faux positifs adopté : Ion score-Identity score = 7. 70 protéines sont identifiées dans la totalité des analyses pour une identification de séquence leurre.*

Nous montrons ainsi la nécessité d'analyser nos digests en ayant recourt au préfractionnement préalable de l'échantillon par chromatographie pour augmenter la probabilité d'identification de peptides uniques. Nous utiliserons par la suite cette stratégie de décomplexification systématiquement pour l'analyse des digests des extraits de cheveu.

6) Choix de la stratégie de traitements des données

Le traitement des données générées par l'analyse par spectrométrie de masse des peptides est une étape critique pour l'obtention de résultats. La stratégie de traitement utilisée peut jouer un rôle essentiel dans l'obtention d'informations supplémentaires. Elle passe par un choix cohérent de la banque de données théoriques utilisée pour la comparaison aux données expérimentales. Le mode opératoire choisis pour comparer ces données peut également impacter sur le nombre d'identifications obtenu ainsi que sur la qualité de la validation des résultats. Nous commentons ici la stratégie de traitement que nous utiliserons par la suite pour effectuer les identifications de nos échantillons.

a) Le choix des banques protéiques utilisées pour la recherche

Différentes banques de séquences protéiques sont à la disposition de la communauté scientifique et peuvent donc être envisagées pour réaliser des études protéomiques sur l'humain.

La banque UniprotKB/SwissProt Homo Sapiens comme banque de référence

La banque de référence pour l'humain est actuellement la banque UniProtKB/Swiss-Prot restreinte à la taxonomie homo sapiens [217, 218]. Cette banque, dont chaque entrée a été minutieusement annotée à partir des informations de la littérature, tient compte de la plupart des séquences des KAPs ayant pu être décrites suite aux travaux précédemment évoqués dans la première partie de ce manuscrit. Nous l'utiliserons dans la suite de nos

études dans une optique où nos résultats puissent être utilisés pour compléter les informations de cette banque pour les protéines que nous pourrons identifier dans le cheveu.

La version Varsplic de UniprotKB/SwissProt Homo Sapiens comme banque pour la recherche de variants dans les séquences

Le génome humain a été particulièrement étudié ces dix dernières années. Grâce à des projets visant à comprendre l'origine des phénotypes humains en comparant un nombre significatif de génomes, différents sites de variants naturels sont désormais décrits pour les protéines humaines [219-222]. Ces variants peuvent être recherchés dans une problématique de caractérisation mais il convient alors de choisir une méthodologie appropriée pour soumettre ces données *in silico* aux données expérimentales de nos échantillons.

L'annotation des séquences dans UniProt dispose en partie de ces informations en plus de référencement d'isoformes issues de sites d'épissages alternatifs. Certains variants observés au niveau du transcriptome pour certaines familles de KAP (KAP 1 et 4) suggèrent l'existence de différents allèles pour une partie des gènes correspondants [74]. Les variants décrits dans ces études, dont certains pourraient être à l'origine de polymorphismes en taille des protéines si elles étaient exprimées, sont répertoriés dans la banque UniprotKB. L'outil Varsplic [223], peut être utilisé pour extraire ces informations et constituer pour chaque variant décrit une entrée protéique particulière. Le nombre de séquence ainsi générées est très important (plus de 700 000 par rapport aux 20 300 initiales). Nous évaluerons l'utilisation de ces séquences supplémentaires pour la caractérisation des protéines du cheveu.

La banque NCBI nr Homo Sapiens

Si la banque UniProt est particulièrement bien annotée, nous pouvons considérer qu'il puisse exister dans d'autres banques protéiques, des séquences différentes de celles répertoriées et qui pourraient correspondre à certaines se trouvant dans nos échantillons. Il peut ainsi être intéressant d'utiliser d'autres banques protéiques en complément d'une recherche dans UniProt. Dans ce cadre, nous avons choisi d'utiliser une banque complémentaire, « NCBI non redondante » contenant plus d'entrées protéiques qu'UniProt. Les séquences de cette banque sont issues du rassemblement de différentes banques (dont UniProt). Lorsque des séquences protéiques de ces sources différentes sont identiques, une unique entrée est constituée. Les séquences portant des variants ou des conflits de séquences constituent des entrées uniques qui peuvent être utilisées dans une optique de caractérisation. Le principal inconvénient de cette banque est la moindre qualité de l'annotation et son volume d'entrée qui respectivement peuvent compliquer l'attribution des séquences peptidiques identifiées, allonger la durée de la recherche et augmenter la probabilité d'obtenir des faux positifs lors de l'identification.

b) L'utilisation de la complémentarité des moteurs de recherche pour renforcer la validation des résultats

Pour accroitre la confiance qui peut être accordée aux résultats d'identification, plusieurs algorithmes de recherche peuvent être utilisés parallèlement [224, 225]. Les principes de comparaison des données expérimentales aux données théoriques et les systèmes de calcul de score de corrélation sont différents entre les algorithmes [226]. Ces différences peuvent être utilisées pour bénéficier d'une complémentarité des performances des algorithmes pour l'identification. En validant indépendamment les résultats pouvant être obtenus puis en combinant les résultats de validation, il est possible d'obtenir des identifications supplémentaires à celles obtenues avec un seul algorithme (Figure 14). Par ailleurs, les séquences identifiées avec plusieurs algorithmes sont validées de manière croisée ce qui renforce leur identification. Notre laboratoire a développé une procédure permettant de travailler avec deux algorithmes de recherche : Mascot et OMSSA [227]. Nous utiliserons conjointement ces deux moteurs lors de nos recherches dans la suite de nos études.

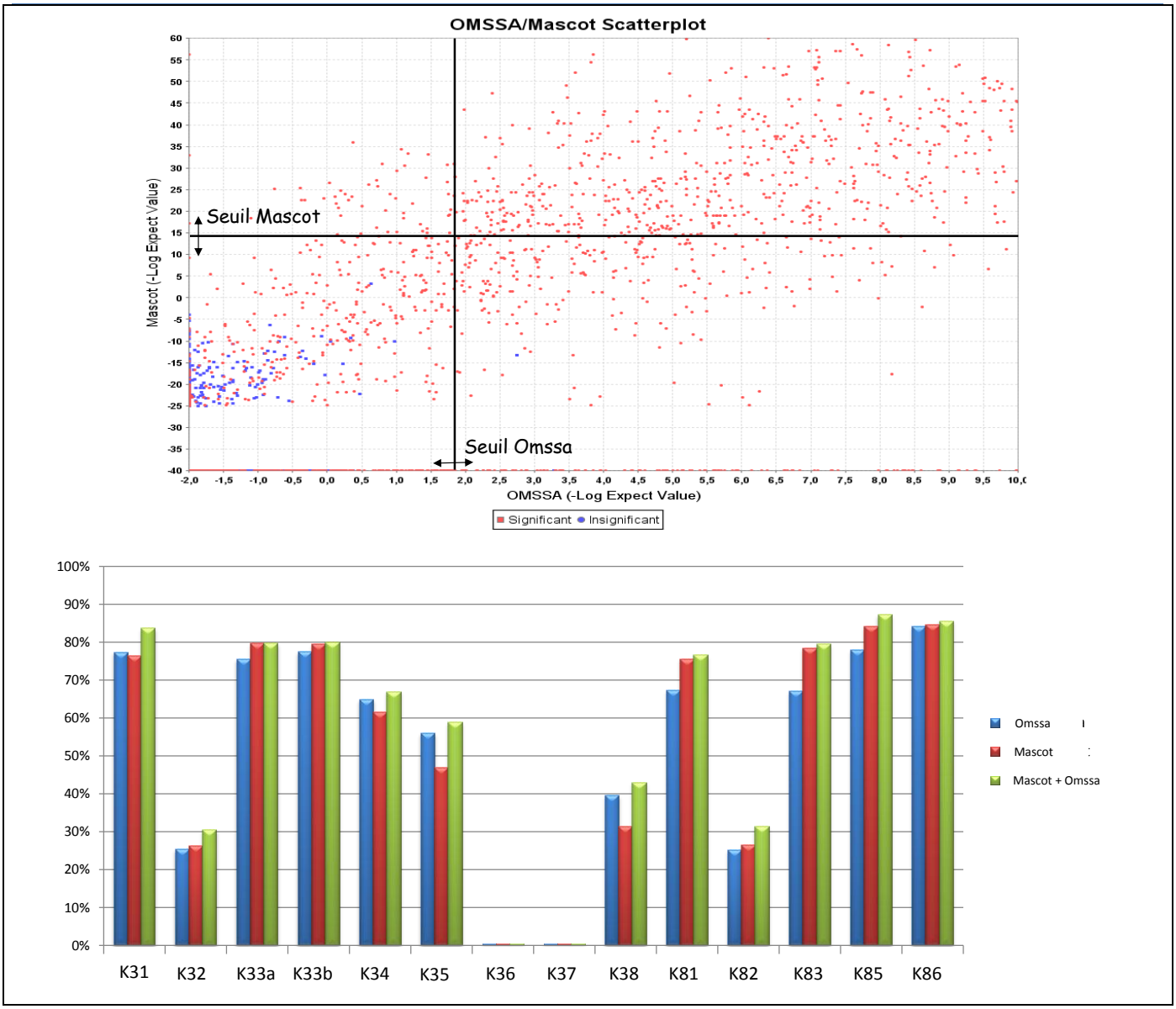


***Figure 14** : Illustration de la complémentarité des moteurs de recherche. En haut, chaque point représente un spectre de fragmentation auquel chaque moteur de recherche attribue un score probabiliste. La population de spectres peut être triée avec deux seuils indépendants. En bas, illustration de l'augmentation des couvertures de séquences des kératines grâce à la combinaison des deux moteurs de recherche pour l'identification de résultats de séquençage d'un digest trypsique de cortex.*

Le processus de traitement et de validation des résultats obtenus avec plusieurs moteurs de recherche nécessite de les traiter avec un outil bioinformatique adapté. Nous utiliserons le logiciel Scaffold (Proteome Software) pour le rassemblement et la validation par contrôle du taux de faux positif des résultats des deux moteurs [228]. La validation sera effectuée par contrôle du taux de faux positif des résultats obtenus pour chaque moteur par approche target-decoy. Le logiciel permet, à partir des données d'identifications peptidiques obtenues par les moteurs de recherche dans les banques, de déterminer les identifications des protéines. Il associe également des probabilités de justesse aux peptides et aux protéines en utilisant les approches empiriques de Bayes.

Si l'utilisation de ces outils permet de considérablement faciliter la tâche de traitement des données, les résultats apportés nécessitent tout de même un examen critique dans certain cas. Certaines ambiguïtés peuvent demeurer et un examen manuel global des résultats peut permettre d'identifier des erreurs. Dans le cas de l'identification de protéines identifiées avec un peptide unique et tout particulièrement celles provenant de familles d'isoformes, nous contrôlerons dans une dernière étape de validation les spectres correspondants. La figure 15 illustre un cas

où la vérification manuelle des identifications permet d'identifier des faux positifs. Ce contrôle peut également s'avérer nécessaire pour valider en dernière étape des spectres issus de peptides modifiés.

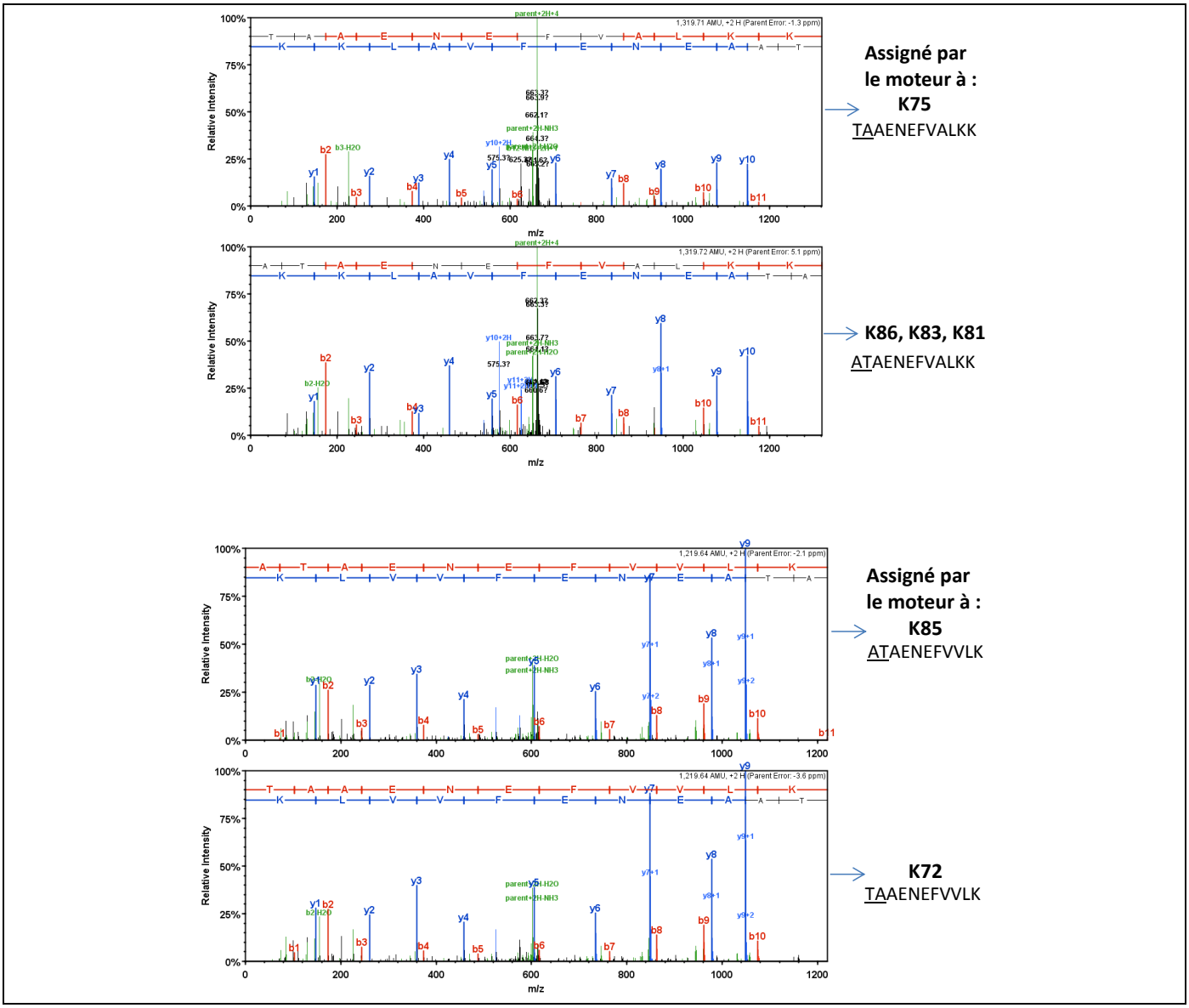


Figure 15 *: Illustration de la possibilité d'erreur d'assignation par l'algorithme de recherche et nécessitant le contrôle manuel des protéines identifiées à un peptide unique. Dans ces deux cas de figure, les identifications de K72 et K75 dans le protéome du cortex alors analysé ne sont pas attendues. Ces identifications sont manuellement écartées grâce à nos connaissances de la bibliographie et de l'existence de séquences isobariques parmi les kératines du cortex. Nous noterons également qu'aucun autre peptide unique à ces protéines n'est retrouvé dans l'échantillon contrairement aux autres isoformes K85 et K86, K83 et K81.*

c) L'analyse séquentielle des données pour réguler la probabilité d'entrées de faux positifs dans les listes d'identification

Nous avons précédemment commenté l'importance du choix des critères de recherche les plus précis possibles pour améliorer la confiance accordée aux résultats d'identification en diminuant le risque d'introduction de faux positifs. Des critères de recherche trop précis peuvent pourtant être source de perte d'informations. Par exemple, des informations supplémentaires peuvent être obtenues par l'identification des peptides obtenus par une digestion semi spécifique [229]. L'identification de ces peptides parait d'autant plus importante dans le cas où les protéines analysées possèdent peu de sites de coupures enzymatiques spécifiques. Néanmoins, la recherche de

ces peptides simultanément avec celle de peptides issus de coupures enzymatiques spécifiques diminue intrinsèquement la spécificité de la recherche. Il en va de même lorsque l'on souhaite rechercher d'éventuels peptides porteurs de modifications. Il est alors nécessaire d'adopter des critères de validation plus stricts au détriment du nombre d'identification.

Pour pallier à ce problème, nous réaliserons nos recherches en utilisant une approche séquentielle. Cette approche consiste à réaliser successivement différentes recherches pour lesquelles les paramètres sont choisis pour cibler spécifiquement une population de peptides particulière à chaque étape. La validation par contrôle du taux de faux positif peut alors être réalisée spécifiquement sur cette population. Nous pouvons envisager, dans un premier temps, la recherche de l'ensemble des peptides issus d'une digestion spécifiques avec quelques modifications supplémentaires spécifiques (carbamylation des cystéines, oxydation des méthionines, acétylation N-terminale des protéines). Dans un second temps, l'ensemble des spectres n'ayant pas donné d'identification à la première étape est resoumis au cours d'une seconde étape pour laquelle sont recherchés les peptides issus d'une digestion semi spécifique. Il est par la suite possible de poursuivre d'autres étapes pour lesquelles sont recherchées d'autres modifications ou d'effectuer des recherches dans d'autres banques. A chacune de ces étapes, les résultats sont spécifiquement validés pour la population de peptides identifiés. Ce contrôle adapté des résultats d'identification en fait le principal avantage de cette stratégie.

L'inconvénient de la recherche séquentielle réside dans la difficulté de fusionner les résultats obtenus à chaque étape en une seule liste d'identification. En effet, les logiciels de rassemblement des données à notre disposition ne permettent actuellement pas de rassembler les données générées de cette manière. Ce problème se pose également lors du rassemblement de données issues de différentes digestions enzymatiques. Dans le cas d'analyse d'isoformes, il est nécessaire de considérer manuellement l'ensemble des identifications obtenues lors des différentes recherches afin d'extraire les peptides spécifiques à chaque protéines. Des données, comme le recouvrement ou le nombre de peptides d'une protéine, nécessitent alors d'être calculées.

d) Le choix de l'analyseur : l'utilisation de l'apport de la précision de la mesure de masse

Parmi les spectromètres de masse du laboratoire, deux types d'analyseurs peuvent être envisagés pour la problématique : les analyseurs trappes d'ions et les Q-TOF. Les trappes ont la réputation d'être des appareils plus sensibles et plus rapides pour l'acquisition de données MS et MS/MS que les Q-TOF mais ont le principal inconvénient de ne pas apporter une très grande précision de mesure de masse [230]. La précision de mesure de masse sur les ions parents permet pourtant d'adopter des critères de masses strictes au cours des recherches dans les banques et diminue la probabilité d'identifier des faux positifs [231]. La précision de mesure sur les ions fragments parait également très appropriée pour l'identification sans ambiguïté de peptides modifiés et peut être un critère essentiel pour appuyer l'identification de nouvelles protéines [230]. Dans le contexte de notre étude, nous avons ainsi choisi de travailler exclusivement avec des instruments de type Q-TOF afin de bénéficier de cette précision de mesure.

Dans l'intégralité du chapitre suivant, nous consacrerons toute une étude destinée à l'amélioration de l'acquisition des données sur ce spectromètre de masse pour combiner la précision de mesure tout en garantissant un haut débit de séquençage des peptides.

7) Estimation de l'abondance des protéines par mesure des digests peptidiques

L'analyse protéomique se confronte actuellement à la nécessité de fournir, en complément des données de caractérisation des protéines présentes dans un mélange, des données relatives à leur quantité [232]. L'ensemble des méthodes de quantification précises développées ces dernières années répond à une problématique de quantification des différences d'expression des protéines lors de l'évolution d'un protéome soumis à différentes conditions expérimentales [178].

Ces méthodes sont en revanche moins adaptées pour quantifier de manière absolue des protéines différentes les unes par rapport aux autres dans un même protéome. Cette information est néanmoins nécessaire à notre problématique pour apporter une idée de la stœchiométrie des protéines constituant l'édifice moléculaire. Il parait indispensable d'évaluer les constituants majoritaires de la structure finale qui, nous le postulons, doivent apporter la majeure partie de la contribution aux propriétés mécaniques de la structure.
L'intensité des signaux des peptides ionisés détectée par spectrométrie de masse est fonction de la quantité de protéine correspondante et initialement présente. Cependant, l'obtention de ces peptides, leur ionisation et la méthode utilisée pour leur détection impactent sur la valeur quantitative mesurée finale (Figure 16).

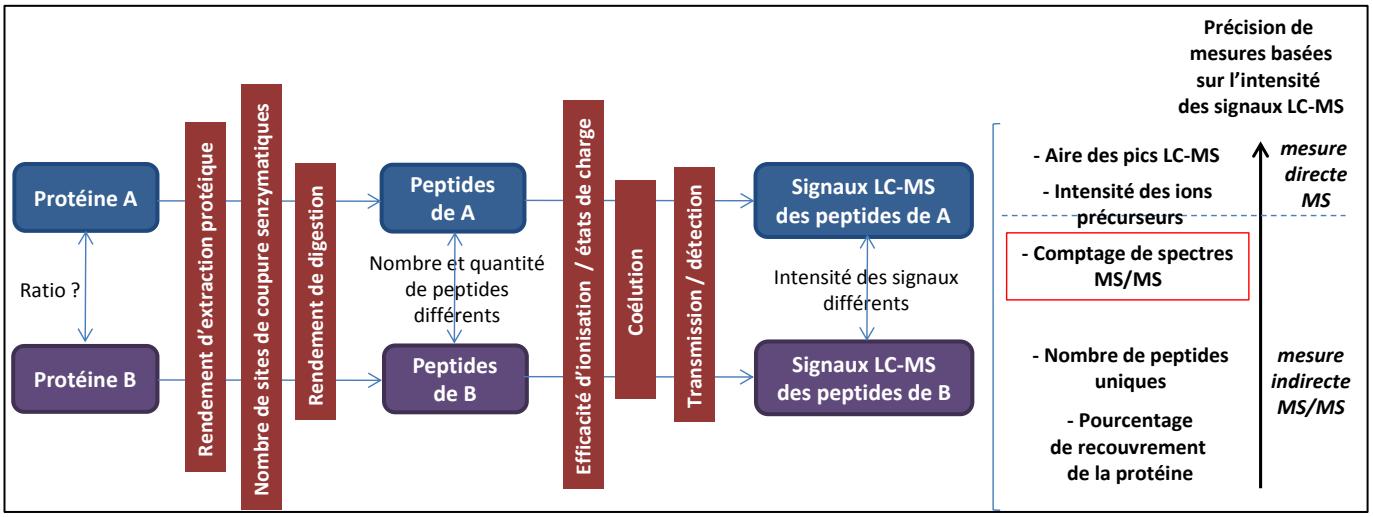


Figure 16 *: Les différents biais impactant sur le lien entre protéine et signal des ions des peptides mesurés par spectrométrie de masse.*

Afin d'apporter une notion semi quantitative à nos résultats d'identification, nous avons choisi d'utiliser les valeurs de nombres de spectres enregistrés au cours des futures acquisitions LC-MS/MS pour estimer l'abondance des protéines majoritaires. Conscients des limites de la précision d'une telle technique, nous resterons prudents quant aux résultats ainsi apportés. Nous nous contenterons d'utiliser ces données pour évaluer une expression importante ou non de la protéine identifiée dans le protéome analysé. Compte tenu de l'abondance de peptides partagés par plusieurs protéines dans le protéome analysé, nous n'utiliserons que la mesure des signaux obtenus sur les peptides uniques de chaque protéine.

Chapitre III Optimisations instrumentales du couplage nanoLC-ESI-Q-TOF : de la compréhension du système à son optimisation pour l'analyse protéomique

Les systèmes analytiques disponibles au laboratoire utilisent exclusivement, pour l'analyse de mélanges peptidiques, le couplage direct entre la chromatographie liquide et la spectrométrie de masse par l'intermédiaire d'une source électrospray. Les systèmes de chromatographie employés sont des systèmes fonctionnant en nanodébit.

Dans le cadre de notre problématique d'identification des isoformes, nous avons précédemment décrit notre choix de travailler à l'optimisation des instruments de type Q-TOF du laboratoire (SYNAPT G1 de Waters et MaXis de Bruker) permettant les meilleures précisions de mesure de masses. Nous avons souhaité optimiser ce type de couplage dans le but d'obtenir un système précis permettant des interprétations des données limitant le risque de faux positifs. L'identification des isoformes se basant sur l'analyse d'un nombre très limité de peptides uniques, l'optimisation doit également permettre d'obtenir une analyse complète des échantillons en minimisant tout risque de sous échantillonnage de ces peptides uniques.

Une optimisation ne peut être réalisée sans une connaissance globale des paramètres influant sur les résultats recherchés. Après avoir détaillé les principes de fonctionnement de cette instrumentation, nous présenterons une étude détaillée décrivant et évaluant l'impact des différentes étapes de l'analyse des peptides pour par la suite les optimiser dans le but d'obtenir un système performant.

1) Architectures et principe de fonctionnement d'un ESI-Q-TOF

a) La source électrospray

Le principe de l'électrospray consiste, à partir d'analytes initialement en solution, à générer des ions et à les transférer en phase gazeuse. Cette source fonctionne à pression atmosphérique et le mécanisme est doux ce qui permet de préserver l'intégrité des molécules.

Le liquide contenant les analytes passe au travers d'un capillaire sur lequel est appliquée une forte différence de potentiel électrique par rapport à une contre électrode située au niveau de l'interface du spectromètre. Il se forme ainsi un cône qui éclate en un spray de gouttelettes chargées. La formation de ce spray peut être assistée à l'aide d'un gaz de nébulisation injecté au travers d'une gaine entourant le capillaire.

Les gouttelettes chargées sont désolvatées sous l'effet de la température et/ou d'un gaz de séchage ce qui entraine une augmentation de la densité de charges à la périphérie des gouttelettes. Lorsque cette densité devient trop importante, la gouttelette explose en gouttelettes filles qui peuvent à leur tour subir le même mécanisme.

Les analytes présents dans les gouttelettes sont initialement sous forme ionisée en solution (résidus protonés, adduits avec des cations ou des anions). Au cours de ce processus, ils vont être transférés par désolvatation en phase gazeuse. Les différences de potentiel électrique et de pression appliquées au sein de la source vont permettre de les transférer vers l'interface du spectromètre.

En protéomique, cette source est principalement utilisée en mode d'ionisation positif du fait de la sensibilité apportée à l'analyse des peptides trypsiques et de la possibilité de mécanismes de fragmentation favorables à l'identification des séquences peptidiques. Les peptides élués de la colonne chromatographique sont dessalés et présents dans un tampon acide favorisant la protonation des résidus basiques.

L'efficacité d'ionisation des peptides est dépendante de leur séquence : de cette composition dépendent l'affinité protonique et l'enthalpie de solvatation du peptide. Un peptide possédant des résidus basiques aura plus de probabilité d'être associé à un proton et un peptide hydrophobe aura plus de probabilité d'être désolvaté de la gouttelette aqueuse et chargée. Par ailleurs, les peptides ont une conformation en solution : cette conformation induite par les interactions entre les résidus peut favoriser l'ajout de protons.

Cette compétition à l'ionisation fonction de la physico-chimie des analytes est à l'origine de phénomènes de suppression ionique pouvant exister dans la source électrospray.
L'ionisation est néanmoins favorisée par la concentration des analytes initialement en solution. Il existe donc un lien entre cette concentration et la quantité d'ions générée.

b) De la source à l'interface

Dans l'architecture Q-TOF utilisée, les ions issus de la source électrospray sont séparés des composés neutres grâce à un premier niveau d'interface situé après le cône d'échantillonnage. Plusieurs géométries sont utilisées par les constructeurs pour permettre cette séparation et le transfert des ions de la pression atmosphérique vers un premier vide primaire (typiquement de l'ordre de quelques millibars).
L'interface Z-spray de la source Waters contraint les ions à une trajectoire en Z lors du premier palier de vide primaire afin de les transférer vers le second étage de vide.
L'interface Bruker associant un capillaire de transmission en verre disposé excentré à un guide d'ions composé d'une série de lentilles annulaires de diamètre interne décroissant (ion funnels). L'ensemble permet de collecter, transporter et focaliser les ions vers le second étage de vide (Figure 1).

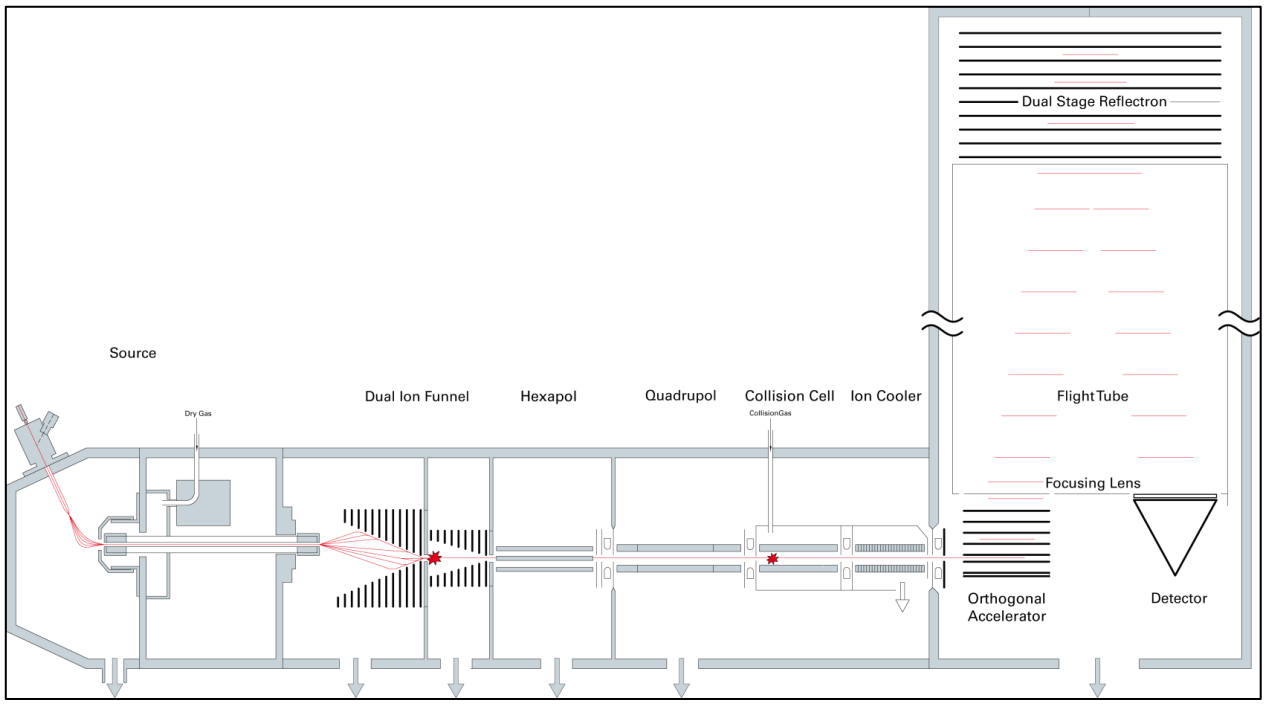


Figure 1 *: Schématisation de l'architecture générale du spectromètre de masse ESI-Q-TOF MaXis.*

c) La transmission

Par la suite les ions nécessitent d'être transmis et focalisés jusqu'à l'analyseur. L'objectif au cours de ces étapes est de transmettre le maximum d'ions tout en obtenant un faisceau focalisé, nécessaire à la mesure dans le temps de vol. La transmission est réalisée par l'intermédiaire de multipôles (Bruker) ou de lentilles annulaires de diamètre interne constant (T-Wave, Waters) associés à des lentilles de faibles diamètres permettant la focalisation.

d) Les quadripôles pour la sélection et la fragmentation

Les ions traversent sur ce parcours un quadripôle permettant de travailler soit en transmission des ions vers la suite de l'interface en mode MS, soit en filtre de masse en mode MS/MS. Le quadripôle est suivi d'une cellule de

collision qui peut être un hexapôle (Bruker) ou un T-Wave (Waters), partiellement isolé dans une enceinte où est maintenue une pression de gaz. En mode MS, la différence de potentiel aux bornes de ce quadripôle n'apporte pas une énergie cinétique suffisante pour entrainer la fragmentation des ions. Cette différence de potentiel est augmentée en mode MS/MS et permet par collision avec les molécules de gaz d'obtenir de la fragmentation (fragmentation à dissociation induite par collision, CID).
Le faisceau d'ions, fragmentés ou non à la suite de cette traversée, est focalisé avant l'introduction dans l'analyseur à temps de vol.

e) L'analyseur à temps de vol

L'analyseur à temps de vol consiste à accélérer les ions sous l'effet d'une différence de potentiel. En sortie de cette zone d'accélération, les ions ont une vitesse fonction de leur charge, de leur masse et de la tension qui leur a été communément appliquée. Les ions traversent alors une zone libre de champ où règne un vide leur permettant un libre parcours moyen suffisant atteindre le détecteur : c'est le tube de vol. Les ions atteignent alors le détecteur à des temps différents fonction de leur rapport m/z. Ceci permet d'enregistrer une intensité en fonction du temps. En étalonnant le système, il est alors possible de remonter à la correspondance m/z en fonction du temps et de reconstituer un spectre de masse I=f (m/z). La résolution de l'instrument (mesurée m/Δm ; m, la masse de l'ion ; Δm, la largeur à mi hauteur du signal de l'ion) est d'autant plus importante que le tube de vol est long, mais cette longueur expose à une perte d'ions plus importante lors de leur parcours dans le tube de vol, ce qui se traduit par une perte de sensibilité.

f) L'injection orthogonale

Le couplage de ce type d'analyseur nécessite de passer d'un faisceau d'ion continu issu de la première partie du spectromètre à un paquet d'ions focalisés spatialement et ne possédant pas de vitesse initiale dans l'axe du tube de vol.
Pour cela le faisceau d'ions continu est échantillonné et envoyé dans le tube de vol au moyen d'un système d'injection orthogonale (pusher). Il permet l'accélération des ions vers le tube de vol avec une vitesse qui doit être strictement supérieure à la composante de vitesse perpendiculaire du faisceau d'ion. L'accélération est réalisée par pulse sur le faisceau d'ions qui initie le top départ du temps de vol des ions.
Le faisceau d'ions peut être ralenti à la sortie de l'interface par un système de thermalisation des ions qui permet de regrouper les ions en un paquet dont la composante de vitesse perpendiculaire au pusher est réduite. Cette étape de thermalisation permet un gain de résolution mais aussi de sensibilité puisqu'un faisceau d'ions plus dense peut être amené au moment de l'injection orthogonale.

g) Le réflecteur électrostatique

Les ions de même masse peuvent, après accélération, posséder des vitesses légèrement différentes. Pour compenser ce phénomène, un réflecteur électrostatique est utilisé dans le tube de vol. Il est constitué d'une série d'anneaux ou de grilles permettant de réfléchir les ions vers le détecteur. Pour des ions de même rapport m/z, le trajet parcouru par les ions les plus rapides est plus important que celui des ions les plus lents. Ces ions de même masse peuvent alors être refocalisés au niveau du détecteur, ce qui permet un gain substantiel de résolution qui s'ajoute à celui apporté par l'augmentation de la longueur de la distance de vol.

h) La détection

Deux types de détecteurs sont utilisés avec les analyseurs à temps de vol. Le SYNAPT G1 est équipé d'un détecteur MCP tandis que le MaXis possède un détecteur MagneTOF. Ces détecteurs convertissent les ions en électrons secondaires qui sont amplifiés. La conversion ion/électron doit être réalisée sur une surface pour

permettre l'enregistrement précis du temps de vol. La géométrie de ces détecteurs est nécessairement sous forme de plaque.

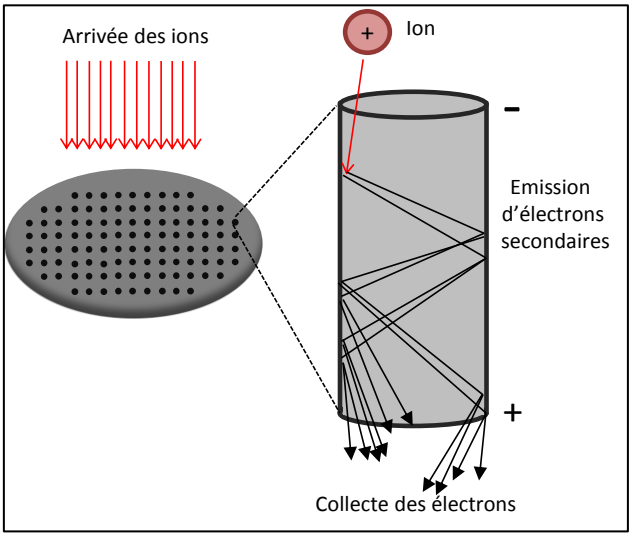


***Figure 2** : Schéma de fonctionnement d'un détecteur MCP.*

Le détecteur MCP (micro channel plate), consiste en un disque dans lequel sont perforés de petits canaux dont la surface interne est une couche semi conductrice [233, 234]. Lorsqu'un ion pénètre dans un canal, l'impact avec la surface génère un électron secondaire. La tension appliquée sur le détecteur entraine l'accélération de l'électron émis qui, par collision en cascade dans le canal, va conduire à l'émission d'autres électrons et à l'amplification du signal : on parle de dynode de conversion. Les électrons issus de l'ensemble des canaux ayant reçu un ion sont alors collectés par une anode puis le signal est digitalisé (Figure 2). Après l'impact d'un ion dans un des canaux, un certain temps est nécessaire avant que le canal soit à nouveau capable d'émettre un signal. Ce temps est du même ordre de grandeur que le temps correspondant à l'intervalle pendant lequel les ions de même m/z parviennent au détecteur.
La saturation du détecteur est constatée lorsque l'ensemble des canaux du MCP collecte simultanément un ion. Le détecteur est alors aveugle le temps que les canaux soient à nouveau en mesure de transmettre le signal induit d'un autre ion.

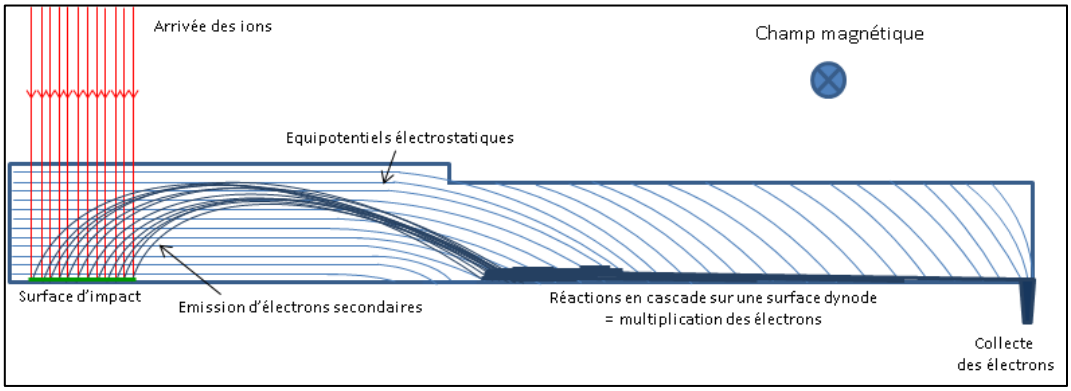


***Figure 3** : Schéma de principe du fonctionnement d'un détecteur magnétique pour temps de vol. Adapté de l'article technique de D. Stresau et al., SGE Analytical Science.*

Le détecteur magnétique TOF (magneTOF), consiste en une surface plane émettant également des électrons secondaires suite à l'impact des ions. Les électrons secondaires sont collectés hors du système grâce à une géométrie particulière de potentiels électrostatiques déformés par un champ magnétique. Leur trajectoire est

alors contrainte par le champ magnétique qui entraine leur focalisation sur une surface dynode et la réaction en cascade permettant l'amplification du signal électronique qui est par la suite collecté et digitalisé (Figure 3).

Le détecteur magnétique bénéficie d'un temps de réponse plus court que le MCP ce qui peut être mis à profit pour obtenir une meilleure résolution temporelle du signal.

i) La digitalisation

La digitalisation sur les analyseurs à temps de vol peut être réalisée par deux types de digitaliseurs. Elle est assurée à une fréquence de digitalisation qui définit le nombre de points enregistrés par unité de temps. Il est nécessaire d'avoir une fréquence de digitalisation cohérente par rapport à la résolution instrumentale. Un nombre de points inadapté se traduira par des pertes de précision sur la mesure de masse, la résolution digitale et l'intensité enregistrée pour le signal. Un nombre de 10 points environ est nécessaire pour digitaliser fidèlement un pic. L'augmentation de la résolution instrumentale des analyseurs à temps de vol va donc de pair avec l'augmentation des fréquences de digitalisation des digitaliseurs employés. Les dernières générations d'appareils permettant des résolutions instrumentales de l'ordre de 50000 doivent être utilisées avec des digitaliseurs échantillonnant à des fréquences de l'ordre de 1 à 4 GHz.

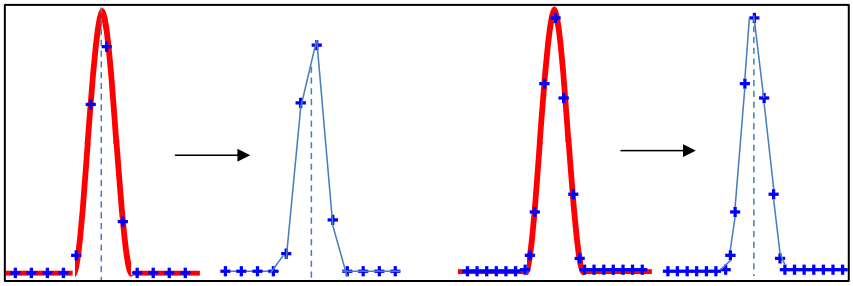


Figure 4 *: Illustration de l'impact de la fréquence de digitalisation sur la reconstitution d'un signal expérimental. A gauche, le signal est digitalisé avec 4 points, à droite avec 7. Dans le second cas, le signal reconstitué est plus fidèle au signal initial. Cela se traduit par de meilleures mesures de la masse, de la hauteur et de la largeur à mi hauteur pour le pic analysé.*

Le digitaliseur TDC (Time to Digital Converter) permet de mesurer des signaux en fonction du temps. Les temps des signaux sont digitalisés au moment où l'intensité issue du détecteur est supérieure à un seuil fixe de digitalisation (Figure 5). La mesure est réalisée sur une période et se répète à une fréquence fixe de digitalisation. Une seule mesure de temps peut être réalisée par période. Le signal ainsi enregistré à une période donnée est indépendant de l'intensité issue du détecteur dès lors qu'elle dépasse le seuil de digitalisation.

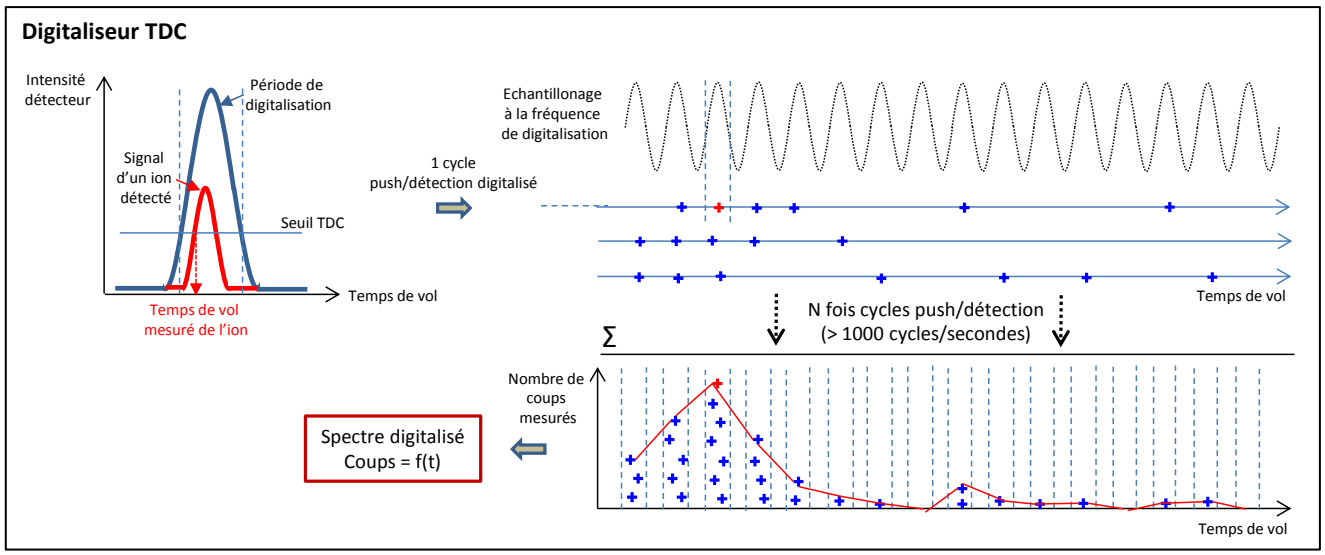


Figure 5 *: Principe de digitalisation du signal et de la constitution d'un spectre de masse avec le digitaliseur TDC.*

Pour chaque cycle injection (push)/détection au niveau de l'analyseur, une liste de temps associés à la détection de signaux est digitalisée. La combinaison de plusieurs milliers de ces cycles pendant un scan (de l'ordre de la seconde) permet, par sommation de l'ensemble des données, de remonter à un nombre de signaux digitalisés (coups) en fonction du temps de vol.

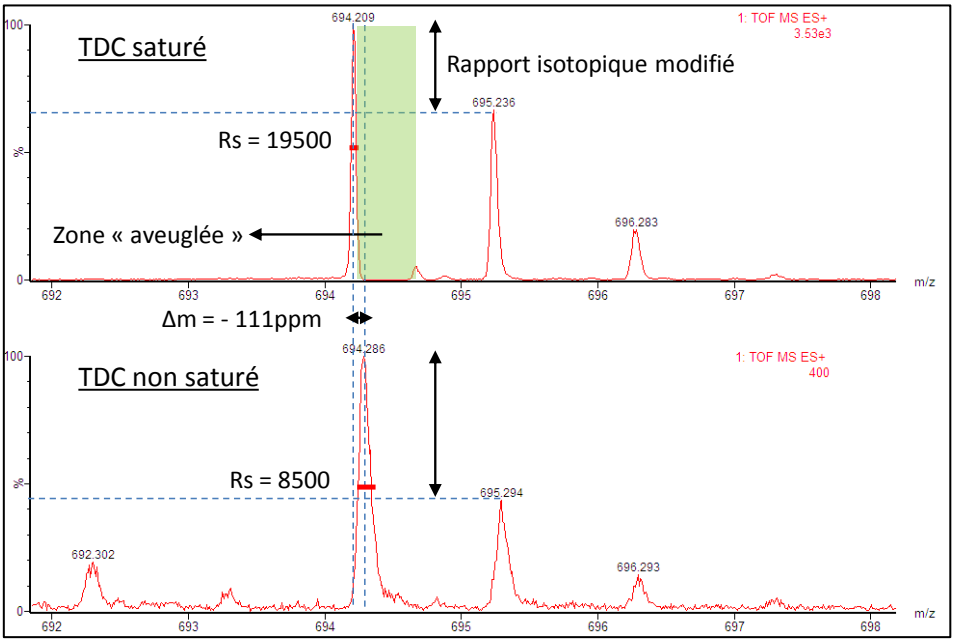


Figure 6 *: Illustration de la modification d'un signal suite à la saturation du digitaliseur TDC. Lorsque le signal est saturé, seuls les ions émettant des électrons aux temps de vol les plus courts sont enregistrés. Les ions arrivant à des temps de vol plus longs impactent la dynode au moment de l'aveuglement du digitaliseur et ne sont ainsi pas digitalisés. Le signal est artificiellement plus fin et décalé vers une masse plus faible que la réelle. L'intensité réelle du signal est sous estimée et peut s'observer par la modification des rapports isotopique de l'ion considéré.*

Le digitaliseur ADC (Analogic to Digital Converter) permet d'associer une valeur digitale à l'intensité issue du détecteur. Cette association est réalisée par intégration du signal sur une période à la fréquence fixe de digitalisation. La résolution du digitaliseur définit un nombre de niveaux qui est associé à une fenêtre d'intensité. Pour chaque cycle push/détection, une liste d'intensités associées à une fenêtre de temps de vol est constituée. La sommation des cycles au cours d'un scan permet de sommer l'ensemble des intensités intégrées dans chaque fenêtre de temps et d'obtenir une intensité totale en fonction du temps de vol (Figure 7).

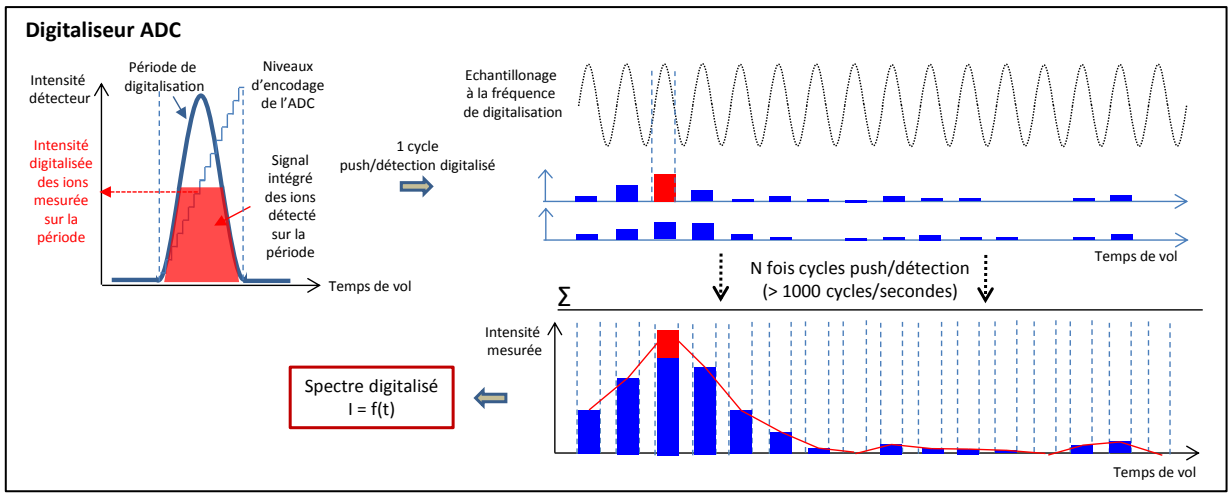


Figure 7 *: Principe de digitalisation du signal et de la constitution d'un spectre de masse avec le digitaliseur ADC.*

Lorsque l'intensité issue du détecteur est supérieure à la fenêtre d'intensités correspondant au niveau le plus haut pouvant être digitalisé, le digitaliseur est à saturation et associera par défaut sa valeur d'intensité limite. Le signal aura « la tête coupée ».

Nous avons vu que dans chaque configuration de digitalisation, le spectre de masse est le résultat de la sommation de l'ensemble des données intégrées sur une plage de temps regroupant des milliers de cycles de mesure de temps de vol. Cette accumulation permet l'augmentation du rapport signal sur bruit. Le bruit étant un phénomène aléatoire contrairement aux signaux des ions accumulés, le gain en rapport signal sur bruit est proportionnel à la racine carrée du nombre de sommations.

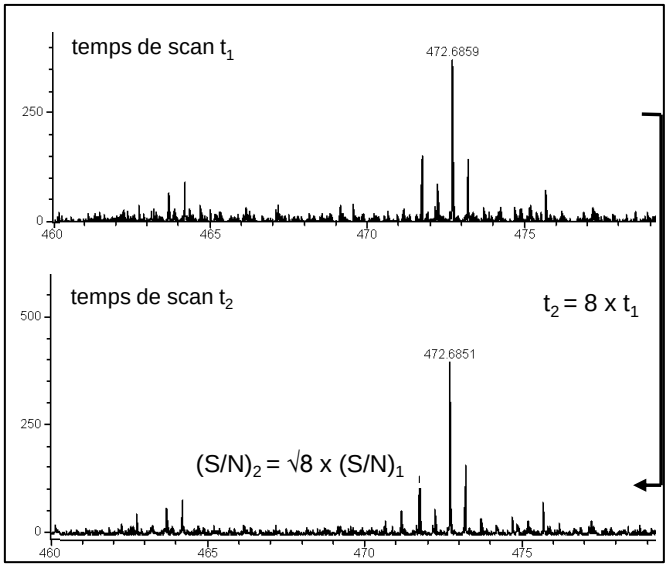


***Figure 8** : Illustration du gain de signal apporté sur un spectre par l'augmentation du temps d'accumulation sur les analyseurs TOF. S/N, rapport signal sur bruit.*

Cette propriété rend la sensibilité des analyseurs à temps de vol dépendante du temps passé à accumuler (Figure 8). Elle explique l'intérêt d'avoir en amont une interface permettant de transmettre les ions le plus efficacement possible jusqu'au détecteur afin de diminuer significativement le temps d'acquisition pour l'obtention d'un spectre de rapport signal sur bruit satisfaisant. Nous verrons par la suite l'importance que revêt le temps d'acquisition dans le cadre du séquençage des peptides par utilisation du couplage direct entre Q-TOF et chromatographie liquide.

j) La mesure de masse et son importance en protéomique

Les analyseurs TOF sont des spectromètres permettant d'atteindre des précisions de mesure importantes grâce à leur bonne résolution spectrale (de 10 000 à 60 000 pour les dernières générations d'instruments). La mesure de la masse de l'ion étant fonction de la distance parcourue pendant le temps de vol, il est nécessaire que cette distance soit constante dès lors que le spectromètre a été étalonné. Dans les faits, les tubes de vol sont soumis à la dilatation fonction de la température ambiante ce qui entraine des dérives de l'étalonnage en fonction du temps. La précision des systèmes électronique peut également être affectée par la température.

Une première solution pour éviter les dérives de mesure est la thermostatation de la pièce voire de l'instrument : les éventuelles variations de la régulation de la température entrainent néanmoins des variations pouvant se répercuter en variations sur la mesure de masse.

La seconde solution est de mesurer très précisément les variations de température au niveau du tube de vol et d'impacter la variation mesurée au coefficient de la fonction de calibration utilisée.

Une autre solution est d'introduire une ou plusieurs références internes : la mesure de la variation de la masse expérimentale de chaque composé de référence peut être utilisée pour ré-étalonner la fonction de calibration associée au spectre (Figure 9). Différentes façons peuvent être envisagées pour la calibration interne. Il est

possible d'utiliser des ions contaminants présents dans les solvants utilisés (par exemple les polydiméthylcyclosiloxanes), d'utiliser un second sprayeur pour infuser séquentiellement des références (système Waters LockSpray[tm]) ou de diffuser en permanence dans la source d'ionisation des composés volatils déposés sur un adsorbant (système Bruker).

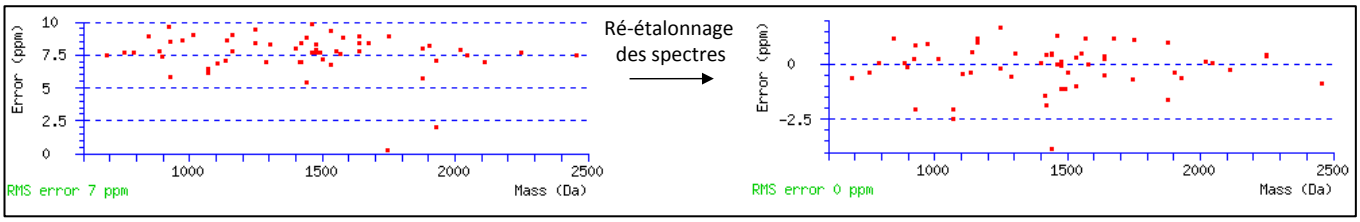


Figure 9 *: Illustration du bénéfice du réétalonnage des spectres sur la mesure des ions parents grâce à l'utilisation d'une calibrant interne.*

L'intérêt de la précision de mesure en protéomique réside dans le fait qu'elle permet la restriction sur des bases expérimentales du nombre de listes de masses théoriques comparés. Les spectromètres de masse hybrides Q-TOF font parti, avec l'IT-TOF et l'Orbitrap des spectromètres utilisés en protéomique pouvant apporter de la précision de mesure de masse de la même manière sur le parent et sur ses fragments en réalisant un haut débit de cycles d'acquisitions. La précision de masse sur les fragments permet de diminuer le risque d'identifier un faux positif lors de la comparaison avec la liste de masses de fragments expérimentale.

2) Optimisation du système ESI-Q-TOF pour l'amélioration des acquisitions de données MS et MS/MS

a) Optimisation des paramètres de source, de transmission et d'isolement

L'optimisation de la source d'ionisation est réalisée à la suite de chaque intervention sur cette dernière, par exemple après son nettoyage. Elle consiste à infuser en continu une solution contenant des analytes et de suivre en direct leur signal. Les différents réglages de tension appliquée entre le capillaire et le cône, de position du spray, du débit de gaz appliqué au niveau du sprayeur peuvent ainsi être réalisés pour optimiser l'intensité du signal.

De la même façon, les paramètres de tensions appliquées sur les différents étages des éléments de l'interface influent sur l'intensité des ions mesurée au niveau de l'analyseur. La transmission des ions en fonction de leur rapport m/z peut ainsi être modifiée de manière critique. Ces paramètres sont optimisés à l'installation de l'instrument par le constructeur qui établit des méthodes standards en fonction des applications envisagées. Pour l'utilisation de l'instrument dans une application d'analyse protéomique, les réglages de l'instrument doivent être adaptés pour une transmission des ions optimale sur la gamme de m/z correspondant à la population d'ions issus d'un digest protéique (typiquement 300-1500 m/z) lors de la mesure de masse des ions parents. La transmission après la cellule de collision doit permettre de transmettre des fragments sur une gamme de masse s'étendant de 50 à 2000 m/z.

Le choix des paramètres d'isolement au niveau du quadripôle de sélection est un compromis entre la transmission de l'ion parent et la largeur de la fenêtre d'isolement. La fenêtre d'isolement est choisie pour permettre la transmission du massif isotopique. Cet isolement permet la conservation du massif isotopique des fragments nécessaire à la détermination de leur état de charge et est suffisamment large pour permettre un rendement suffisant de transmission du parent. Les paramètres d'isolement pour l'analyse d'ions de peptides sont définis par le constructeur et nous ne les avons pas modifiés.

b) Optimisation de la fragmentation sur le SYNAPT G1

Principe de fragmentation

La qualité de la fragmentation des peptides dans la cellule de collision est une étape cruciale lors du séquençage des peptides par spectrométrie de masse en tandem. Le maximum d'informations de séquences doit pouvoir être obtenu à cette étape. L'utilisation du mode de collision en CID des peptides protonés permet préférentiellement de réaliser des fragmentations au niveau de la liaison peptidique ce qui, dans la nomenclature de fragmentation des peptides de Biemann, permet d'obtenir des fragments nommés y et b (Figure 10).

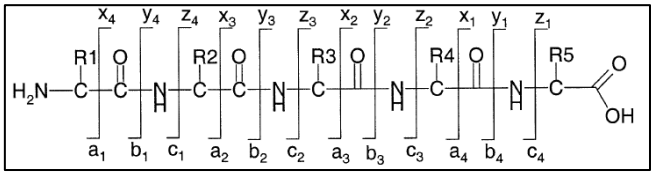


***Figure 10** : Nomenclature de fragmentation des peptides de Biemann.*

Les énergies de collision influent sur la génération de fragments informatifs. Choisir ces énergies consiste à déterminer les tensions à appliquer aux ions parents dans la cellule de collision afin d'obtenir l'activation permettant d'obtenir le maximum de fragments de l'ensemble des liaisons peptidiques. Une énergie trop importante induit une dissociation des fragments et donc une perte d'information. La radiofréquence et son amplitude appliquées au niveau de la cellule de collision peuvent également impacter sur la transmission des fragments générés. Ce choix est donc un compromis à réaliser en fonction de données expérimentales. Couramment les peptides issus de la source électrospray peuvent présenter des états de charge allant de 1 à 5 en fonction de la séquence. Les états de charge 2 et 3 sont les plus courants pour les peptides trypsiques et permettent en CID l'obtention des spectres de fragmentation les plus informatifs.

Expérience d'optimisation

Etablir la tension idéale pour la fragmentation de chaque peptide ionisé s'avère difficile à envisager dans le cadre du séquençage haut débit d'un ensemble de peptide. Afin d'établir sur une population significative de peptides les tendances de fragmentation, nous avons mis en place une expérience. Notre problématique envisage d'obtenir des peptides protéotypiques par analyse de digests obtenus avec différentes enzymes, nous avons souhaité étudier ces tendances sur les digests trypsique, chymotrypsique et GluC (Table 1).

L'analyse par nanoLC-MS/MS du même digest a été réalisée à une énergie de fragmentation constante mais différente pour chaque répétition. Les scores d'ions d'identification des peptides obtenus par Mascot pour chaque répétition sont collectés : ils sont fonctions du nombre de fragments identifiés par le moteur de recherche. Pour chaque peptide, la valeur de score mesurée est normalisée par rapport au meilleur score obtenu pour ce peptide dans la totalité des analyses. L'état de charge de l'ion parent correspondant est pris en compte pour différencier les peptides chargés deux ou trois fois.

Enzyme/Etat de charge	2+	3+	(4+)
Trypsine	188	73	14
Chymotrypsine	150	44	7
GluC	136	59	19

***Table 1** : Nombre de peptides analysés dans l'ensemble des expériences. Les peptides chargés quatre fois sont insuffisants pour réaliser une étude significative de leurs propriétés de fragmentation.*

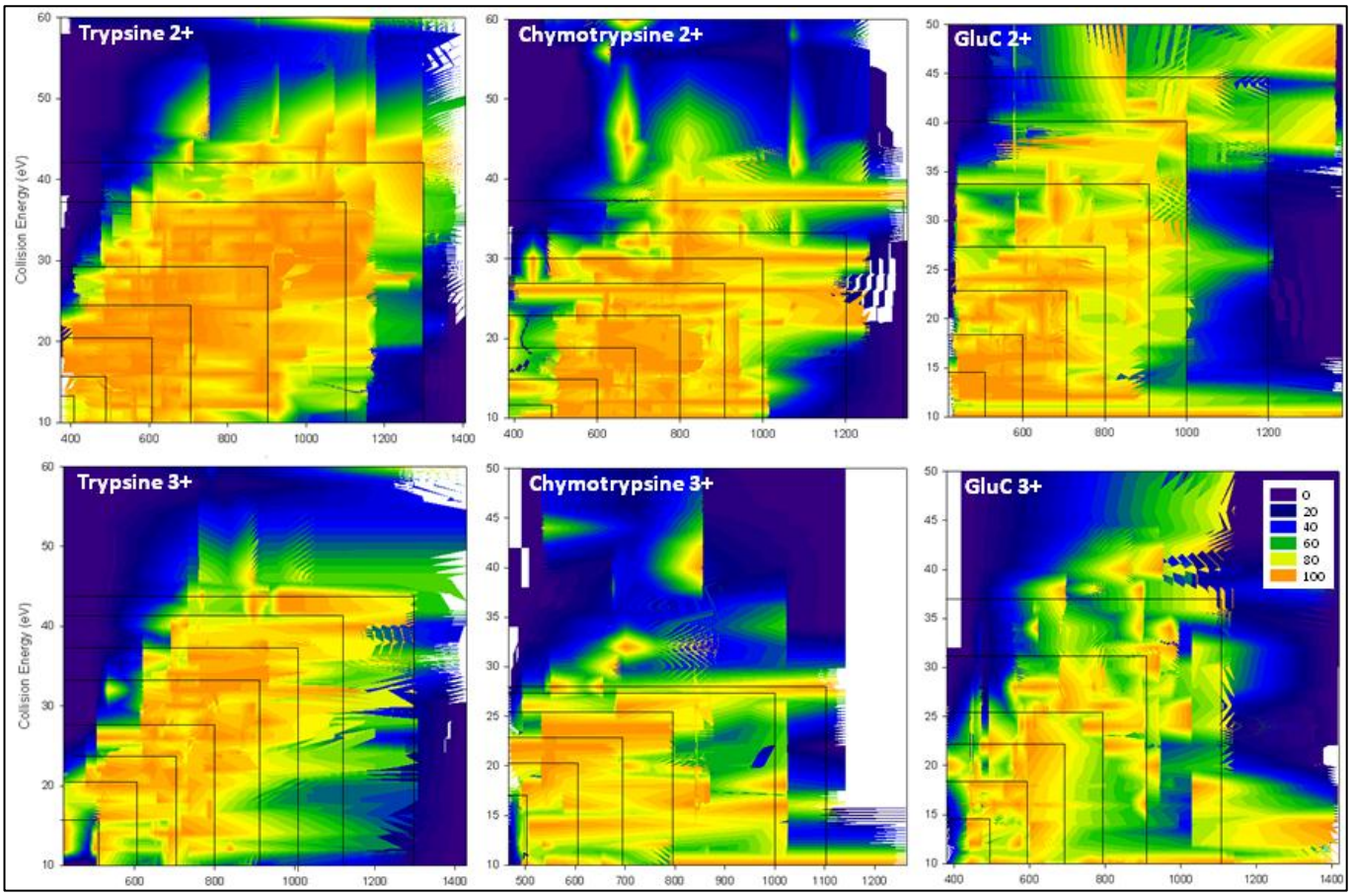


Figure 11 : *Résultats graphiques Score d'ion = f (CE ; m/z). CE est l'énergie de collision appliquée (eV) d'après le constructeur à chaque expérience. Les couleurs, des plus froides aux plus chaudes, rapportent la valeur normalisée du score d'ion d'identification Mascot normalisé à 100 pour le meilleur score obtenu pour l'ion correspondant. Energies étudiées de 10 à 60 eV par pas de 2 eV. Instrument SYNAPT G1. Gaz de collision Argon, pression 1.10^{-2} mbar.*

Nous obtenons ainsi pour chaque digest à un état de charge donné un tableau score d'identification = f (énergie de collision ; m/z). Lorsqu'un peptide n'a pas été sélectionné pour une énergie donnée alors qu'il a donné lieu à une identification pour les deux énergies adjacentes, une valeur moyenne des scores des deux résultats est calculée pour remplacer la case laissée vide dans le tableau. Cette étape permet de réaliser un lissage des valeurs afin de faciliter la lecture des résultats.

Le tableau est par la suite converti en un graphique grâce au logiciel d'analyse de données SigmaPlot® (v11 Systat Software Inc). Les graphiques correspondant aux trois digests pour les deux états de charge étudiés sont présentés (Figure 11).

Résultats de l'optimisation

L'analyse des données montre que davantage d'identifications sont obtenues avec les états de charge 2+ quel que soit les digests utilisés. Ces identifications préférentielles de cet état de charge tiennent plus du fait que ces ions sont majoritaires en sortie de source ESI que du fait d'une fragmentation plus efficace. La répartition des zones de meilleurs scores d'identification est relativement large et suggère qu'une gamme d'énergies de collision peut être envisagée pour une bonne fragmentation d'un même ion.

Il est possible de distinguer au sein de ces graphes un lien entre l'énergie de collision et le rapport m/z. En reportant les énergies considérées centrales de la zone « chaude » pour un rapport m/z donné, les courbes Energie de collision = f (m/z) peuvent être obtenues (Figure 12).

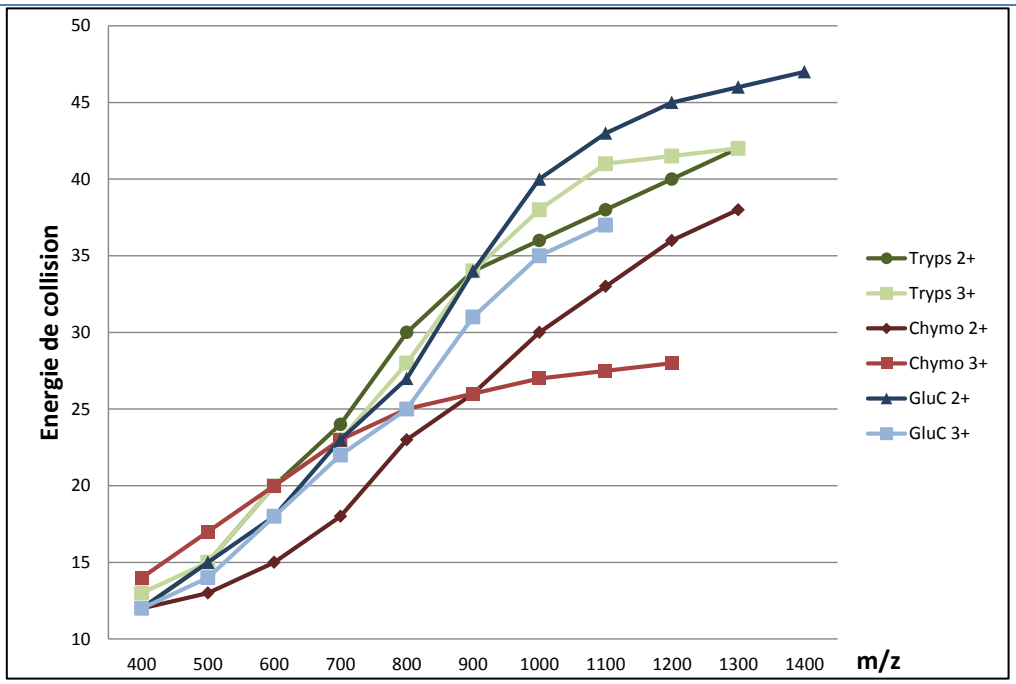


***Figure 12** : Courbes obtenues aux valeurs centrales de densité observées pour les différentes expériences réalisées.*

Ces courbes montrent que les valeurs centrales sont similaires sur l'ensemble des populations d'ions recherchés dans la gamme située entre 400 et 800 m/z. des divergences sont observés entre les différents digests pour des m/z plus importants. Cette dernière observation est à pondérer avec le fait que moins de données ont été obtenues pour des peptides de haut rapport m/z.

Ainsi, nous pouvons suggérer que des courbes d'énergie communes CE = f (m/z) peuvent être utilisées quel que soit le type de digest analysé sur l'instrument étudié. La dispersion des meilleurs résultats autour des valeurs centrales définies suggère la nécessité de ne pas adopter une énergie pour un m/z donné mais plutôt de réaliser des balayages d'énergies de collision entre des valeurs relativement larges et dépendantes du m/z. Une alternative au balayage sur une gamme est de diviser chaque scan de fragmentation en trois scans où une énergie intermédiaire, une énergie basse et une énergie haute sont appliquées séquentiellement.

c) La détection

L'optimisation du détecteur est à réaliser à intervalles réguliers. Elle consiste à adapter la tension appliquée aux bornes du détecteur de manière à avoir un rapport signal sur bruit le plus élevé possible. Les impacts réguliers d'ions sur le détecteur ont pour conséquence une diminution du rendement d'émission d'électrons secondaires de la surface. Pour compenser cette diminution, l'augmentation de la tension sera progressivement réalisée tout au long de la durée de vie du détecteur. Il est à noter qu'une tension appropriée permet de limiter les différences de rendement d'émission secondaires pouvant exister entre des ions de masse différentes dont les vitesses à l'entrée du détecteur diffèrent.

d) Optimisation de l'acquisition des données et de leur traitement

Différentes optimisations de paramétrage dans l'acquisition, l'enregistrement et le traitement des données générées doivent être envisagées pour contrôler la taille et la qualité des données d'acquisition obtenues en protéomique.

Enregistrement des données d'acquisition MS

La nécessité de cette réflexion résulte du constat suivant : les analyses protéomiques en haut débit passent souvent par la décomplexification de l'échantillon avant la réalisation d'analyse en couplage nanoLC-MS/MS.

Cette décomplexification entraine la multiplication du nombre d'analyses. Pour chaque analyse, la quantité de données générées consiste en des listes de données numériques permettant de remonter aux informations de type « valeur de signal = f (temps ; m/z) » pour les analyses MS et les analyses MS/MS. Ces listes sont enregistrées et nécessitent d'être stockées pour la conservation de l'information. Les espaces de stockage dédiés sont donc fonction du nombre d'analyses, des temps d'acquisition et du nombre de points choisis pour la description des spectres signal = f (m/z).

Nous avons vu que l'augmentation de la résolution instrumentale des analyseurs à temps de vol s'accompagnait nécessairement d'une augmentation des capacités de digitalisation et donc du nombre de points enregistrés par unité de temps de vol donc de m/z. La forte fréquence de digitalisation n'a pour but que d'obtenir les informations précises d'intensité, de rapport m/z et de résolution pour l'ensemble des signaux sur la gamme de m/z. La conservation de l'ensemble des points permettant d'y accéder est donc superflue dès lors que ces informations ont été obtenues. Des algorithmes de calcul permettent d'obtenir ces informations directement pendant l'acquisition par traitement des données issues de la digitalisation, stockées temporairement puis effacées après calcul. Les données peuvent alors être compressées sous la forme de signaux à un m/z donné pour lesquels les données d'intensité, de résolution voire d'état de charge associé sont enregistrées (Figure 13). Il est ainsi possible en fonction des logiciels d'acquisition d'avoir des données centroïde (I = f (m/z) pour Waters) ou line ((I ; Rs) = f (m/z) pour Bruker).

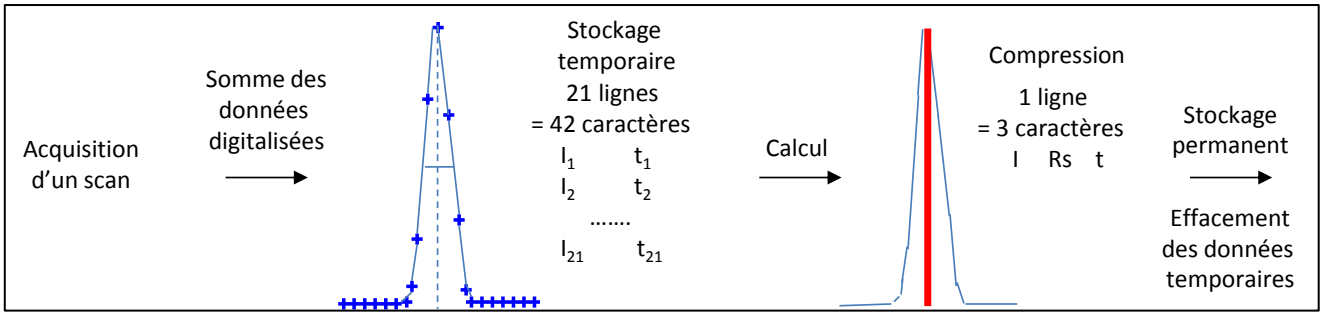


***Figure 13** : Principe de compression des données générées par conversion des signaux digitalisés en points en signaux « bâtons ».*

L'utilisation de ces modes d'acquisition permet de diminuer significativement la taille des données brutes générées et ainsi d'économiser des investissements sur l'espace de stockage dédié à l'instrument.

Enregistrement des données d'acquisition LC-MS

Le nombre de scans acquis par analyse est également un paramètre qui influe sur la taille des données brutes associées. Lors d'un couplage en LC-MS, un choix cohérent dans la fenêtre d'acquisition utilisée peut permettre de diminuer la taille des fichiers en supprimant des données correspondant à des temps où aucune information ne peut être disponible (par exemple hors de la fenêtre des temps d'élution des peptides analysés). Le choix du temps de scan utilisé pour définir un chromatogramme peut également influer sur la taille du fichier, des temps de scan courts entrainant une multiplication du nombre de spectres enregistrés (Figure 14).

La définition d'un pic chromatographique suit la même règle de digitalisation que celle décrite précédemment pour le signal de masse. Lors d'analyses LC-MS pour lesquelles la mesure de l'aire du pic chromatographique est utilisée pour la quantification, un nombre de 10 points minimum sera nécessaire. Le temps de scan à choisir sera fonction de la largeur estimée du pic fonction des conditions de séparation. Ce paramètre peut donc être optimisé pour la description du signal chromatographique lors de problématiques de quantification en LC-MS. L'optimisation du nombre de points nécessaire à l'analyse permet de réduire la taille des données brutes. Cette diminution est cependant moins significative sur des données contenant des spectres déjà compressés.

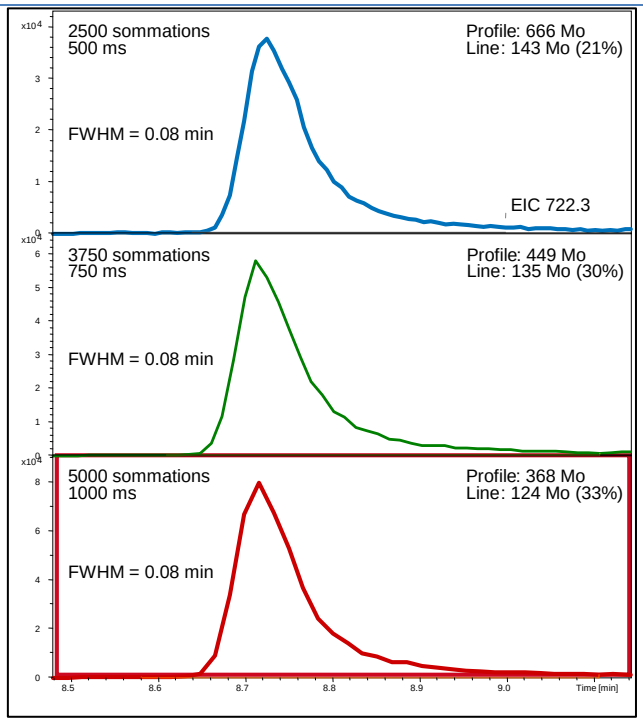


Figure 14 *: Illustration de l'influence du mode d'enregistrement d'une analyse LC-MS sur la taille des données brutes générées. Comparaison de l'acquisition en mode profile (acquisition et enregistrement sans compression), en mode line (acquisition et enregistrement de donnés compressées) en fonction du nombre de sommations de scans utilisé. Parallèlement, cette comparaison montre l'impact du nombre de sommations sur la digitalisation d'un pic chromatographique.*

La génération des listes de masses expérimentales

Après l'analyse d'un échantillon en LC-MS et MS/MS, les données expérimentales obtenues doivent être converties en listes de masses pouvant être comparées aux données théoriques des banques protéiques (Figure 15). La conversion est réalisée par un logiciel de traitement qui effectue différentes étapes :

- La création de « composés » qui consiste à regrouper les couples parent/fragments sous une même identité. Les informations du rapport m/z, de l'intensité et de l'instant où s'est réalisée l'acquisition (temps de rétention et/ou numéro de scan) du parent sont associées aux informations signal = f (m/z) du spectre de fragmentation correspondant. A cette étape, des spectres MS/MS correspondant à des ions parents équivalents peuvent être rassemblés. Si une référence interne pour l'évaluation de la dérive de mesure de masse a été utilisée, le réétalonnage du spectre peut être effectué à ce moment. L'ensemble de ces calculs est effectué sur des données compressées : si l'acquisition a été réalisée en mode point, une étape préliminaire de compression des données en centroïde sera nécessaire.
- La recherche de l'état de charge z du parent et des fragments. Cette étape se base sur la reconnaissance des massifs isotopiques dont les signaux sont séparés par 1/z dalton. La connaissance de l'état de charge permet de connaitre la masse expérimentale de l'ion considéré. Cette masse étant par la suite comparée aux listes de masses théoriques, une erreur dans l'attribution de l'état de charge peut être à l'origine d'une perte d'information. L'état de charge d'un ion n'est pas toujours déterminé. Cette absence d'information peut par la suite être compensée lors de la recherche : il est possible de définir au moteur de recherche des hypothèses d'état de charge pour chaque signal considéré (par exemple 2 ou 3 ou 4).
- L'export des listes de masses des fragments à considérer pour la comparaison avec les données des listes théoriques. L'étape consiste à créer un fichier contenant pour chaque composé les informations qui seront soumises lors de la recherche dans un format lisible par l'algorithme.

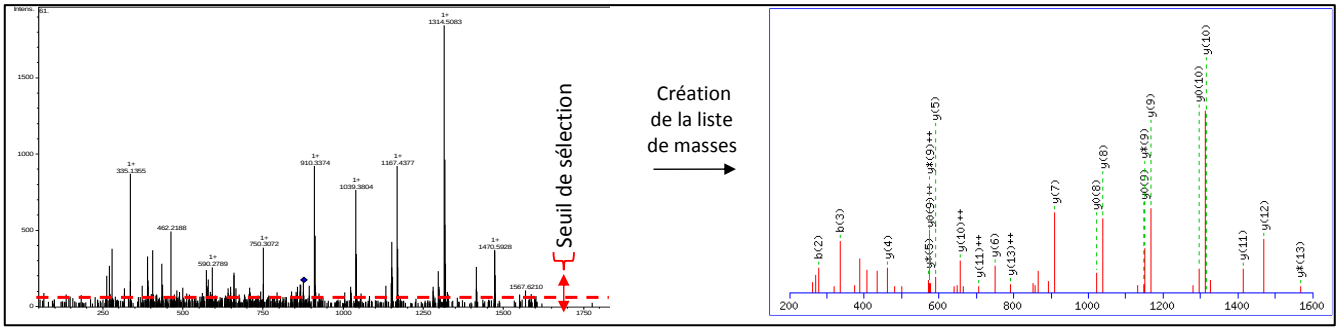


Figure 15 *: illustration de la conversion d'un spectre de fragmentation en listes de masses utilisables pour la recherche protéomique. Le seuil de sélection détermine principalement les données conservées à l'export.*

En fonction des constructeurs, il est possible de modifier un certain nombre de critères influant sur la nature des données exportées. La connaissance de l'instrument et de ses capacités d'acquisition permet de régler précisément un certain nombre de ces paramètres. La connaissance du niveau absolu de bruit présent sur les spectres MS/MS obtenus l'est particulièrement. Elle permet de fixer les seuils d'intensité pour n'effectuer les traitements et les exports que pour des valeurs pouvant être associées avec une grande probabilité à un signal réel.

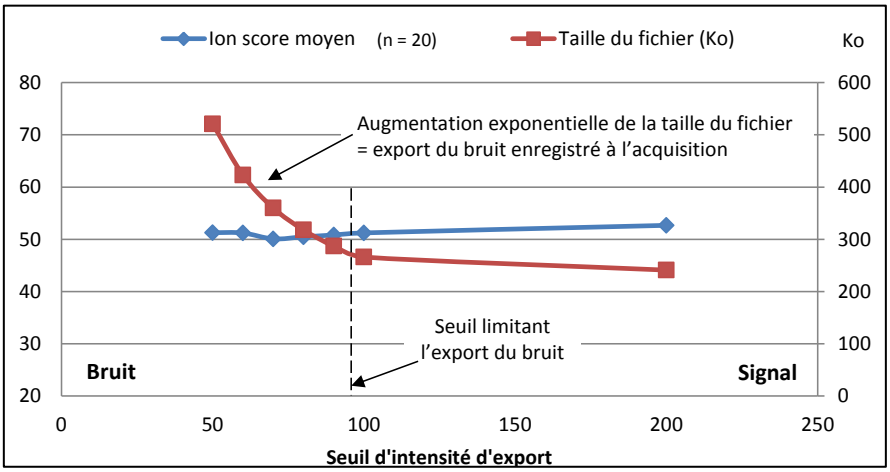


Figure 16 *: Illustration de l'impact du seuil d'intensité d'export des signaux des listes MS/MS sur le résultat d'identification moyen (n = 20 peptides) et sur la taille du fichier d'export. L'ion score moyen des peptides ne varie pas significativement mais devrait diminuer pour des valeurs de seuil plus élevées supprimant les signaux spécifiques aux peptides.*

Nous avons réalisé l'étude de l'évolution de la taille des données d'export en fonction du seuil d'intensité fixé pour illustrer la problématique de la gestion du bruit enregistré à l'acquisition. Le seuil d'intensité est le seuil au dessous duquel les points correspondants ne sont pas enregistrés dans les fichiers d'export. Nous démontrons que les données permettant l'identification des composés sont toujours présentes dès lors que le seuil choisi est suffisamment bas. Jusqu'à un certain seuil d'intensité, la taille du fichier d'export n'augmente que très peu. En dessous de ce seuil, l'introduction de données correspondant aux signaux aléatoires de bruit fait augmenter exponentiellement la taille du fichier de listes de masses enregistrées. Sur un TOF, le niveau de bruit d'un spectre est directement proportionnel au temps de sommation utilisé. En connaissant le niveau de bruit enregistré pour les acquisitions les plus longues de la séquence d'analyse, le seuil d'intensité d'export peut être choisi précisément. L'utilisation d'un seuil relatif (id : seuil fixé par rapport à un pourcentage du signal le plus intense du spectre, utilisé par Waters) plutôt que d'un seuil absolu ne permet pas de tenir compte de cette donnée instrumentale.

Les résultats de cette étude ont permis d'optimiser les seuils d'export utilisés sur le MaXis. Cette optimisation a permis de réduire considérablement la taille des listes de masses et a pour conséquence directe une diminution significative de la durée de traitement lors de l'étape d'identification avec le moteur de recherche (temps de chargement et de traitement réduits).

Nous noterons qu'aucun constructeur ne permet de tenir compte de l'évolution du niveau de bruit en fonction de la gamme de m/z. Cette propriété instrumentale pourrait être utilisée pour améliorer l'export des données de haut m/z, d'intensités souvent faibles mais moins bruitées.

3) Optimisation des paramètres chromatographiques du couplage nanoLC-ESI-MS

a) L'optimisation de l'architecture des systèmes nanoLC

La quantité d'échantillon est souvent limitée lors de l'analyse d'échantillons en protéomique. Les systèmes chromatographiques utilisés pour la séparation de ces faibles quantités de matériel tendent à se miniaturiser en diminuant les diamètres des colonnes utilisées pour la séparation. Les débits d'utilisation sont alors très faibles et permettent des séparations sans dilution de l'échantillon. Les diamètres de colonne couramment utilisés dans ce cadre varient de 75 µm (nanoLC avec des débits de 200-800 nL /min) à 100-300 µm (micro LC avec des débits de 1 à 20 µL/min). Les sources électrospray classiques permettent de travailler avec des microdébits. En revanche, l'utilisation de débits inférieurs nécessite de travailler avec des sprayeurs dont le diamètre du capillaire et la géométrie sont adaptés et désignés sous l'appellation nanosprayeurs. L'utilisation de ces nanosprayeurs permet l'obtention de gouttelettes initialement plus fines qui favorisent le mécanisme d'ionisation électrospray, permettant des gains en sensibilité.
Les chaines chromatographiques à notre disposition en couplage avec les spectromètres de masse utilisés sont des systèmes de nanoLC.

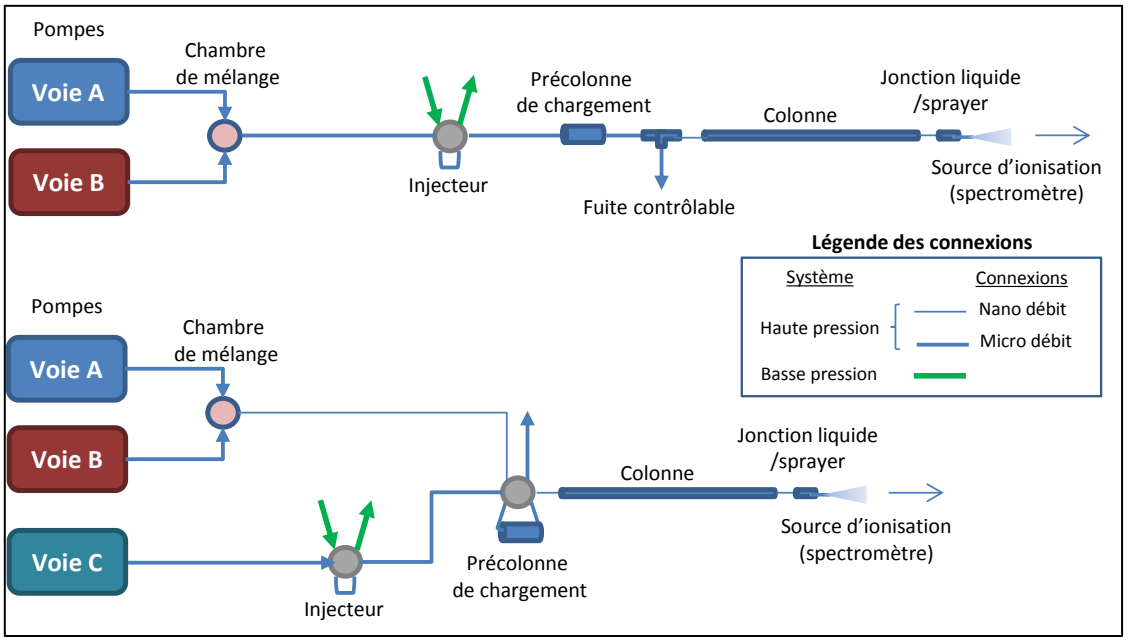


***Figure 17** : Schéma des deux types de montage chromatographiques utilisés pour travailler à nano débits (en haut, système à une pompe ; en bas, système à deux pompes).*

Une difficulté avec de tels systèmes chromatographiques est d'introduire l'échantillon dont le volume, couramment de quelques µL, excède le volume total de la colonne. L'injection se passe toujours en deux étapes :

- la première consiste à introduire l'échantillon dans un système microfluidique par l'intermédiaire d'un injecteur (vanne 6 voies avec une boucle). L'échantillon est alors transféré et piégé sur une précolonne de chargement dont le diamètre est compatible avec le débit de chargement.
- La seconde consiste à d'introduire la précolonne de chargement dans un système nanofluidique dans lequel se trouve la colonne de séparation. Par l'application d'un gradient d'élution en nanodébit, les composés immobilisés sur la précolonne sont tour à tour élués et chromatographiés sur la colonne analytique.

Deux types de montages peuvent être envisagés pour réaliser ce transfert de l'échantillon (Figure 17). Chacun nécessite de disposer d'une pompe binaire permettant d'assurer le gradient d'élution à un nanodébit.
Dans le premier montage, la voie A permet le chargement de l'échantillon à un microdébit. Une fuite entre la précolonne et la colonne est réalisée afin de permettre au système de fonctionner à ce débit incompatible avec le diamètre et la longueur de la colonne analytique. Cette fuite permet également l'élimination des sels présents dans l'échantillon lors de la phase de chargement. Lorsque l'échantillon a été transféré sur la précolonne, la fuite est fermée et les pompes A et B délivrent alors un gradient en nanodébit qui passe par la colonne analytique.

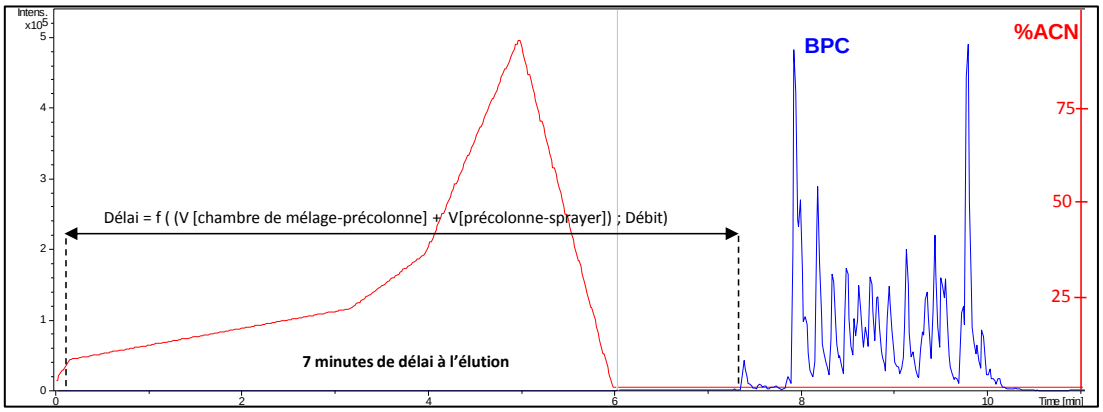


Figure 18 *: Evaluation de l'impact de l'inertie de l'application du gradient sur l'utilisation du spectromètre avec un système nanoLC 1 pompe. L'analyse effective de l'échantillon avec un gradient court est de 2,5 min. Environ 7 minutes sont nécessaires pour que l'élution soit réalisée à compter du basculement du système en nanodébit. Près de 70% de ce temps est imputable au temps mis par le gradient pour arriver sur la précolonne. Pendant environ 60% du temps total de l'analyse, le spectromètre de masse est inutilisé (sans prendre en compte le temps nécessaire à l'injection de l'échantillon dans le système). Débit 450 nl /min, colonne 75 µm x 200 mm, précolonne 180 µm*2 mm. Volume mesuré du système 2,9 µL, volume liquide mesuré de la colonne 0,52 µL, volume liquide calculé de la précolonne 0,15 µL, volume de connexion entre précolonne et colonne 0,25 µL.*

Le second montage nécessite une deuxième pompe (voie C). Dans ce système, la voie C est utilisée uniquement pour le chargement de l'échantillon à un microdébit sur la précolonne. La précolonne est isolée du système nanofluidique par l'intermédiaire d'une vanne 6 voies. Après le transfert de l'échantillon sur la précolonne, cette dernière est introduite dans le système nanofluidique par basculement de la vanne. Dans ce système, les voies A et B délivrent toujours un nano débit et la colonne analytique est toujours parcourue par du liquide.
L'intérêt de l'ajout d'une pompe supplémentaire dédiée au chargement de l'échantillon réside principalement dans le gain de temps apporté à l'analyse de l'échantillon.
En effet, l'utilisation d'une pompe commune au chargement et à l'analyse de l'échantillon présente l'inconvénient d'une inertie importante à l'application du gradient d'élution lors de basculement en nano débit (Figure 18). Cette inertie est la conséquence du volume mort plus important qui existe entre la chambre de mélange et la précolonne dans le premier système dont les connections sont obligatoirement dimensionnées pour des microdébits.

L'ajout de la pompe supplémentaire pour le chargement permet également d'anticiper le gradient d'élution de façon à l'appliquer directement sur la précolonne au moment de son basculement. Le temps entre l'arrivée du gradient sur la précolonne et l'arrivée de l'échantillon au niveau de la source est alors réduit au temps nécessaire au passage de l'échantillon à travers la colonne et la précolonne.

b) Séparation des peptides en phase inverse

La chromatographie des peptides couramment utilisée en couplage avec la spectrométrie de masse est un mode de chromatographie de partage en phase inverse sur une phase stationnaire de type C18.

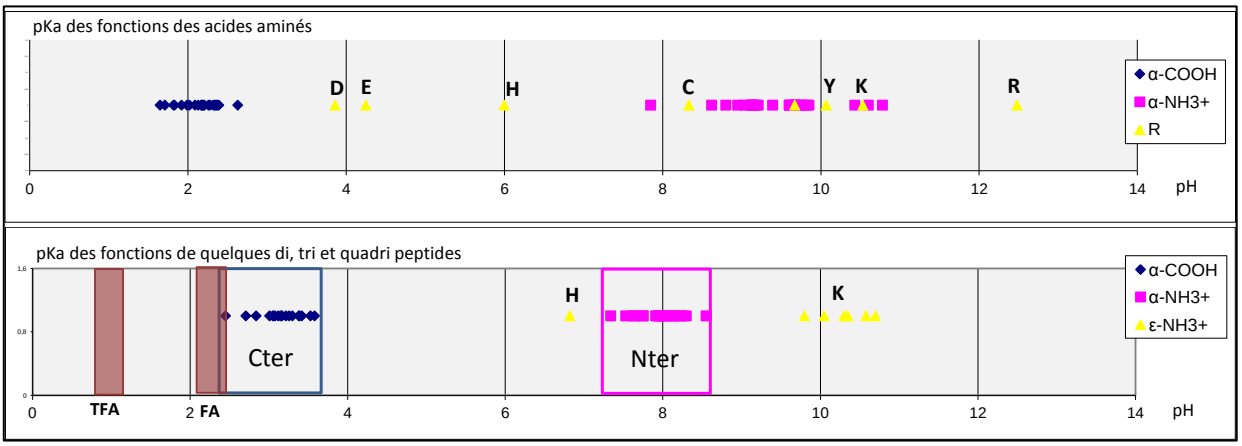


***Figure 19** : pKa des fonctions acido-basiques des acides aminés et de quelques peptides (valeurs extraites de Dawson, RMC et al., Data for Biochemical Research, Oxford, Clarendon Press, 1959). L'utilisation d'agent de paires d'ions acides comme l'acide trifluoroacétique (TFA) ou l'acide formique (FA) montre les gammes de pH dans lesquelles les peptides sont chromatographiés (C < 1 M). Seules les fonctions C-terminales des peptides peuvent être en équilibre entre la forme acide et la forme basique à ces valeurs de pH. Cette possibilité est considérablement diminuée avec l'utilisation du TFA.*

La présence au sein des séquences peptidiques de fonctions très polaires possédant des propriétés acido-basiques nécessite de réaliser un système de paires d'ions pour chromatographier ces composés. L'agent de paires d'ions permet de tamponner la phase mobile à un pH favorisant la prédominance des fonctions acido-basiques sous une forme majoritaire de protonation ou de déprotonation. Les fonctions restant ionisées et donc très polaires sont écrantées par une association avec l'agent de paire d'ion, ce qui augmente l'hydrophobicité de l'ensemble et favorise sa rétention. Les agents de paires d'ions couramment utilisés pour les couplages nanoLC-MS sont l'acide formique ou l'acide trifluoroacétique. Ces agents sont compatibles avec la spectrométrie de masse et permettent de travailler à un pH acide où la majorité des fonctions carboxylate et la totalité des fonctions amine sont protonées. Les amines protonées s'associent en paires d'ions avec la base conjuguée de l'agent de paire d'ion utilisé.
L'élution est réalisée par augmentation de la proportion de phase organique (couramment l'acétonitrile ou le méthanol) dans la phase mobile. Pour la suite de ces travaux en couplage, l'agent de paire d'ion utilisé est l'acide formique (moins polluant et plus favorable à l'ionisation que le TFA), le solvant organique est l'acétonitrile (plus efficace en terme de séparation et de force élutive supérieure au méthanol).

c) Optimisation du gradient d'élution

Le choix du gradient chromatographique est un des principaux paramètres qui influe sur la séparation des composés. Afin d'améliorer l'efficacité de la séparation, nous avons souhaité étudier l'impact de ce paramètre sur le profil d'élution des digests protéiques. Le choix du gradient peut se faire en plusieurs étapes.

Pourcentages de solvant organique initial et final.

Il est indispensable de connaitre les pourcentages de solvant organique permettant l'élution des peptides les plus hydrophiles et les plus hydrophobes. Cette connaissance permet d'appliquer directement le gradient nécessaire à l'élution des peptides les plus polaires et de commencer le rinçage de la colonne dès que le dernier peptide a été décroché de la précolonne. Il en résulte un gain de temps d'analyse significatif puisque le gradient est spécifiquement adapté à l'échantillon chromatographié.

Pour cela, la mesure du temps de rétention du premier et du dernier peptide retenu lors de l'analyse d'un digest avec un gradient linéaire de 0 à 100% de solvant organique est réalisée. En connaissant le délai d'élution du système chromatographique utilisé, il est possible de remonter à une estimation du pourcentage de solvant organique correspondant à chacun de ces peptides.

En pratique, avec l'utilisation de l'acétonitrile, nous avons mesuré qu'un pourcentage de 6% permet l'élution des premiers peptides. En ce qui concerne l'élution des peptides les plus hydrophobes, le pourcentage est échantillon dépendant. Nous avons retenu que les peptides les plus hydrophobes des digests de protéines sur gel étaient correctement élués à 35% mais qu'un extrait de digest direct d'un échantillon complexe (digest d'un protéome de levure) générait des peptides hydrophobes élués à 40%. Le gradient se termine couramment par une montée vers de hautes valeurs de solvant organiques afin de rincer le système d'éventuelles protéines non digérées ou de contaminants organiques.

Evaluation de l'efficacité d'un gradient d'élution linéaire et méthode d'optimisation

Les gradients utilisés en protéomique sont couramment des gradients simples linéaires. La pertinence de ce choix sur la qualité de la séparation a été évaluée en examinant deux critères : la largeur des pics chromatographiques et la dispersion sur le domaine de séparation des composés chromatographiés. L'analyse de ces critères est réalisée grâce à la détection automatique de composés *via* la fonction Disect de DataAnalysis 4.0 (Bruker).

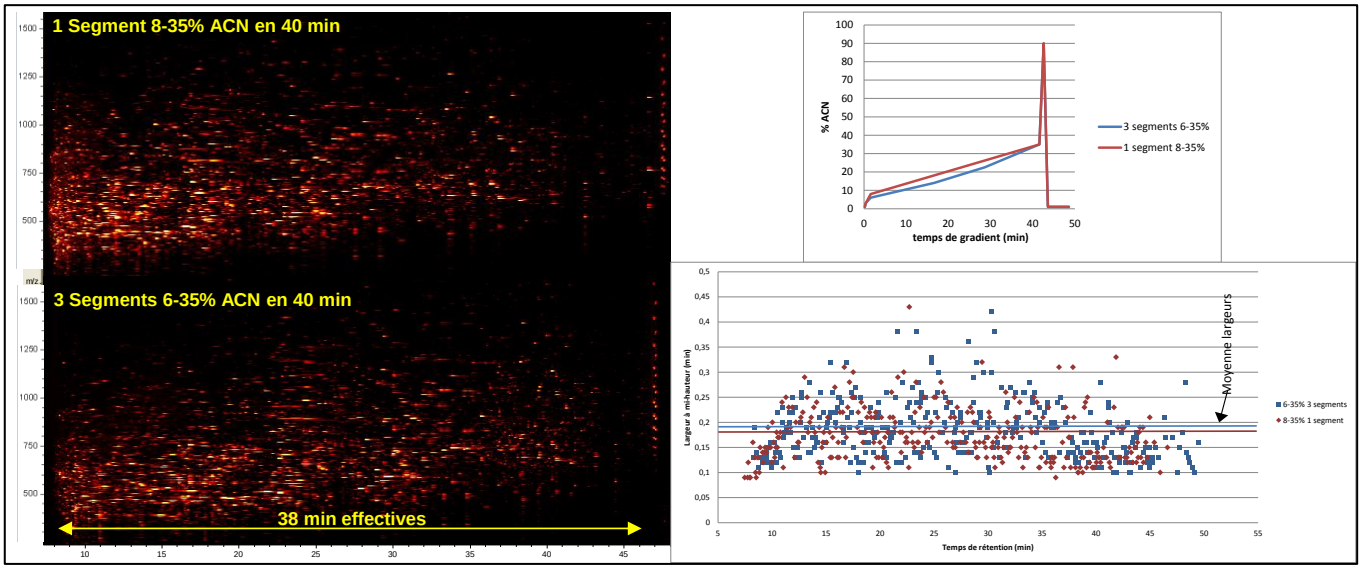


Figure 20 *: Comparaison de deux gradients d'élution : le premier (en bas à gauche), optimisé à trois segments possède trois ruptures de pentes, le second est un gradient linéaire utilisé avant optimisation (en haut à gauche). La comparaison graphique des deux profils d'élution montre une répartition entre la première et la deuxième partie du chromatogramme plus équilibrée dans le cas de l'élution avec le gradient à trois segments que dans le cas du gradient linéaire. La représentation du gradient linéaire commençant à 8% d'acétonitrile montre l'importance de choisir correctement le pourcentage initial de solvant organique : dans ce cas, la densité de composés en début de chromatogramme est trop importante par rapport au reste du chromatogramme. La représentation des largeurs à mi-hauteur en fonction du temps de rétention (en bas à droite) montre que les changements de pente pendant le gradient n'affectent pas l'homogénéité des largeurs de pics chromatographiques. Colonne 200 mm x 75 µm. Débit 450 nL /min.*

Nous avons constaté que les gradients linéaires entrainaient une distribution irrégulière des composés pendant le temps d'élution. La tendance est à une plus grande densité de composés sur la première moitié du chromatogramme par rapport à la seconde moitié. Pour compenser ce phénomène, nous avons développé un gradient comprenant trois segments de pentes permettant de mieux disperser les composés sortant dans la première partie de l'élution et au contraire de concentrer les composés sortant en deuxième partie.
Afin de pouvoir utiliser cette méthode à différents temps de gradients chromatographiques, les valeurs de pente déterminées peuvent être augmentées ou diminuées en appliquant un coefficient multiplicateur commun à chacun des segments.

Coeff	1		1.5	
%ACN	RT (min)	Pente (%/min)	RT (min)	Pente (%/min)
1	0.2		0.2	
3	0.4		0.4	
6	1.6	2.6	2.2	1.7
14	16.6	0.5	24.7	0.4
22.5	28.6	0.7	42.7	0.5
35	41.6	1.0	62.2	0.6
90	42.6		63.2	
1	43.6		64.2	
1	48.6		69.2	

***Tableau 2** : Séquence de gradient à 3 segments optimisée au cours de cette étude et exemple de transposition à un gradient plus long par application d'un coefficient commun aux pentes de chaque segment.*

d) Optimisation des temps de gradient en fonction de la capacité de pic

Le temps pendant lequel est appliqué le gradient d'élution définit une pente de gradient. Avec l'augmentation de ce paramètre, les temps de rétention et la largeur des pics des composés vont augmenter (Figure 21). La capacité du système à séparer est évaluée en tenant compte de ces deux phénomènes. Un système efficace sera capable d'étaler les composés sur une gamme d'élution importante tout en élargissant le moins possible les pics des composés.

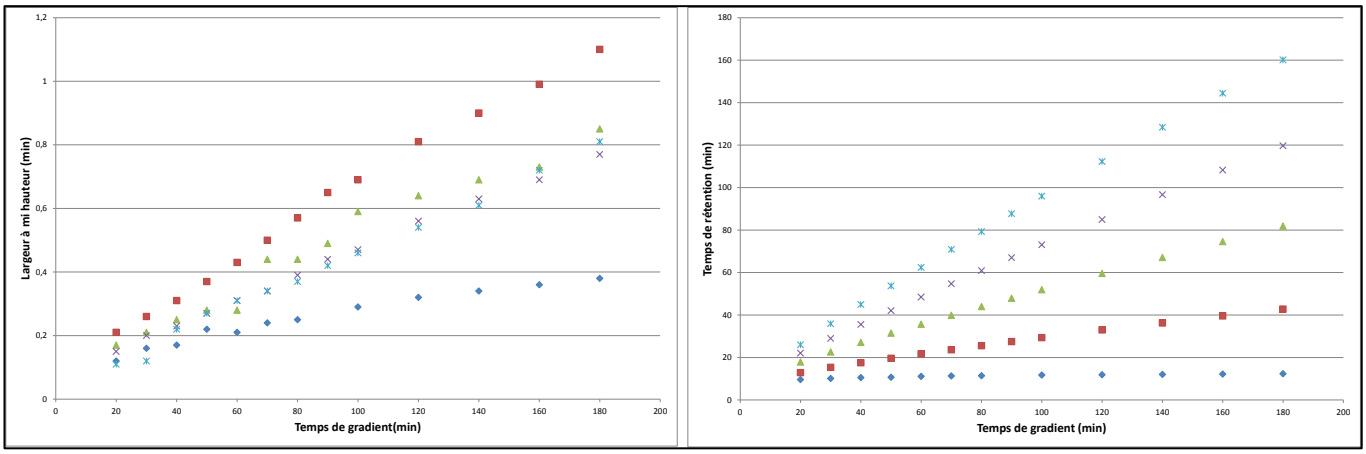


***Figure 21** : Graphiques illustrant pour 5 composés l'impact de l'augmentation du temps de gradient sur la largeur des pics (à gauche) et sur les temps de rétention des composés (à droite). Il apparait que les composés sont de plus en plus séparés temporellement avec l'augmentation du temps de gradient. Néanmoins, cette augmentation de séparation temporelle se traduit par une augmentation de la largeur des pics chromatographiques qui va à l'encontre de la séparation. Ces deux courbes ne permettent pas d'établir à quel temps de gradient le système est le plus efficace.*

La mesure de l'aptitude du système à la séparation utilisant des gradients d'élution est évaluée par la mesure de la capacité de pic. La capacité de pic (C_p) est calculée en mesurant la largeur à mi-hauteur $\omega_{1/2}$ moyenne d'un ensemble significatif de composés et la différence de temps de rétention entre le dernier et le premier peptide élués sur le chromatogramme [206, 210, 235, 236]. Soit :

$$C_p = \frac{\left(t_{rfin} - tr_{début}\right)}{\omega_{1/2}} \quad (1)$$

La valeur de la capacité de pic peut se traduire par le nombre de composés qu'un système est capable de résoudre sur l'ensemble de la gamme d'élution avec des pics séparés les uns des autres à mi-hauteur.
Théoriquement, la capacité de pic dépend de différents paramètres opératoires :

$$C_p = 1 + \frac{\sqrt{N}}{4} \frac{B\Delta C}{B\Delta C\left(\frac{t_0}{tg}\right) + 1} \quad (2)$$

qui peut s'écrire, lorsque le temps de gradient t_g est significativement supérieur au temps mort de la colonne t_0 :

$$C_p = \frac{\sqrt{N}}{4} B\Delta C = a\sqrt{\frac{L}{d_p}} \quad (3)$$

où N est le nombre de plateaux théoriques du système, B la pente de la droite décrite par la relation reliant ln k à la composition de phase mobile (k est le facteur de rétention) et ΔC la variation de composition de solvant organique entre le début et la fin du gradient.
Le nombre de plateaux théoriques N augmente proportionnellement avec l'augmentation de la longueur de la colonne L et la diminution du diamètre des particules de remplissage de la colonne d_p.

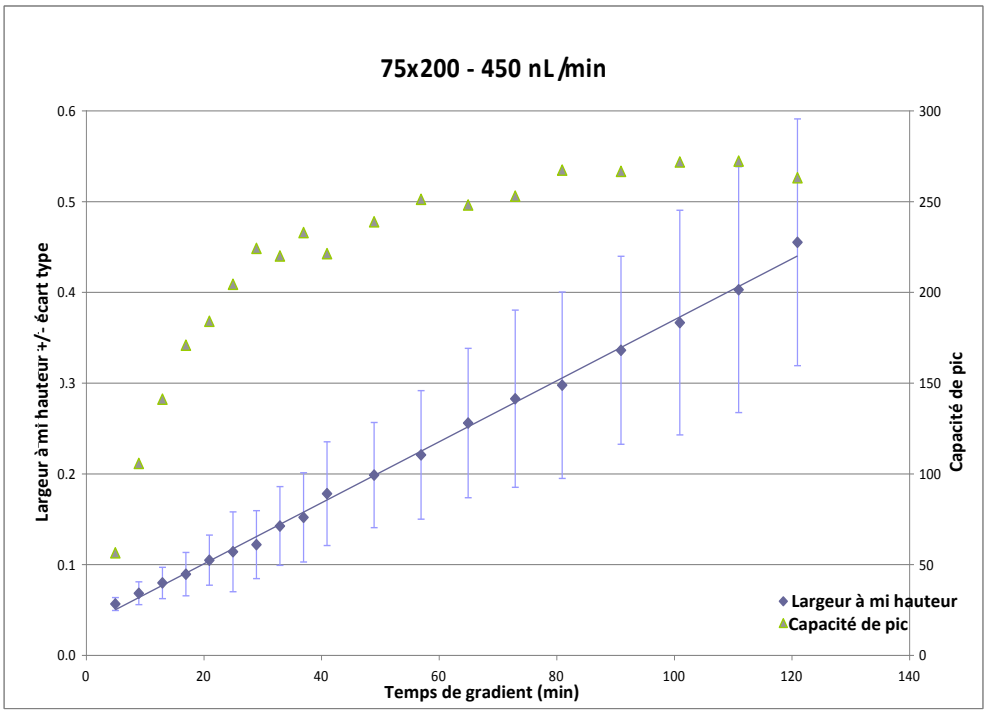


Figure 22 *: Evolution de la capacité de pic en fonction du temps de gradient (analyse d'un digest trypsique de cortex, nombre de composés intégrés allant de 82 à 295). Sur le système utilisé, l'efficacité de séparation augmente de manière critique pour un temps de gradient compris entre 3 et 30 minutes. Cette augmentation est beaucoup moins prononcée pour des gradients plus élevés et un pallier est atteint vers 90 minutes. Parallèlement, la largeur des pics et la dispersion des valeurs mesurées augmentent proportionnellement avec le temps de gradient. Si l'efficacité du système est la plus importante à 90 minutes, elle n'est cependant supérieure que de 15 % à celle mesurée à 30 minutes. Le meilleur compromis entre le temps passé à l'élution et la qualité de la séparation se situe donc vers 30 minutes de temps de gradient.*

Mesurée expérimentalement, la capacité de pic est plus une estimation qu'une valeur absolue. La mesure des largeurs de pics à mi-hauteur moyennes étant évaluée sur la population de composés chromatographiés, la valeur de la capacité de pic peut être impactée par le type et la quantité d'échantillon utilisé. La façon dont les largeurs à mi-hauteur sont estimées peut également impacter sur la mesure. Les résultats de mesure peuvent être différents par exemple lors d'une mesure sur une population de composés ou sur un échantillon de quelques composés chromatographiés. La valeur peut également être affectée par le choix du composé hydrophobe utilisé pour la mesure de la fin de l'élution.

L'étude de l'évolution de la capacité en fonction du temps de gradient utilisé montre l'impact de ce paramètre sur la décomplexification de l'échantillon (figure 22). La connaissance de cette évolution pour le système considéré a permis de choisir le temps de gradient à utiliser de manière à réaliser la meilleure séparation en un minimum de temps d'élution.

e) Influence du débit et de la longueur de la colonne sur la séparation

Au cours de cette étude, la géométrie des particules utilisées pour le remplissage des colonnes est de 1,7 µm. La diminution du diamètre des particules de remplissage de la colonne implique une augmentation de la pression qui, associée à l'augmentation du débit, nécessite des pompes et des systèmes permettant de la supporter. Les constructeurs proposent ainsi des systèmes permettant de supporter des pressions de près de 700 bars. Cette capacité à résister à la pression peut être mise à contribution pour augmenter la longueur des colonnes chromatographiques afin de gagner en nombre de plateaux théoriques.
L'impact de ces deux paramètres chromatographiques sur la capacité de séparation a été évalué expérimentalement. Compte tenu de l'influence du temps de gradient sur la séparation, nous avons souhaité confronter cette influence à celle des paramètres évalués.

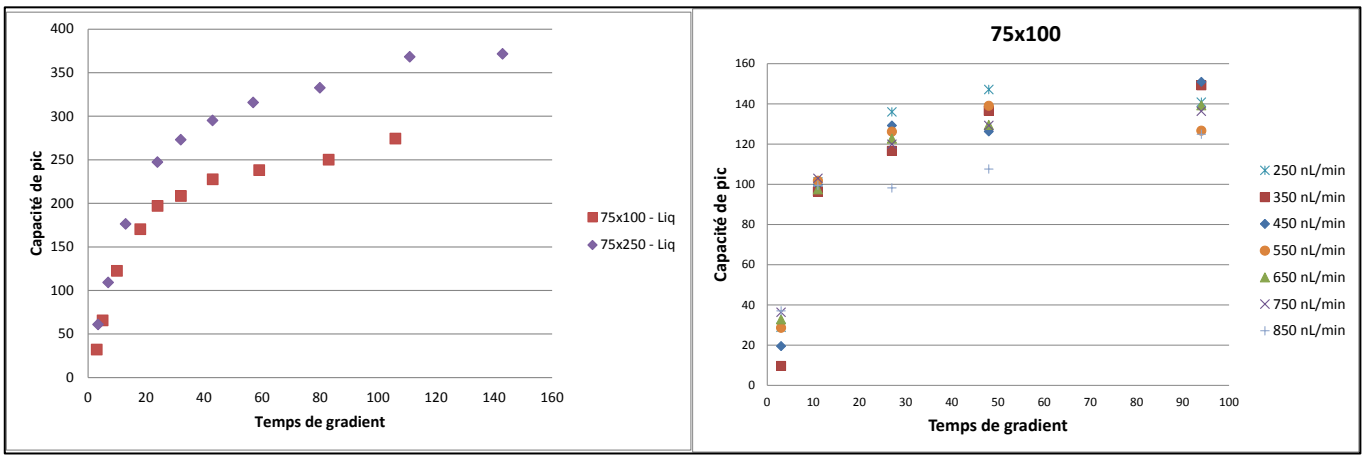


***Figure 23** : Evolution de la capacité de pic en fonction du temps de gradient pour des longueurs de colonne différentes (à gauche, débit 450 nL /min) et pour des débits de phase mobile différents (à droite).*
Nous montrons que l'augmentation de la longueur de la colonne se traduit par un gain de séparation qui peut être combiné au bénéfice de l'augmentation du temps de gradient. Théoriquement, le gain de capacité de pic doit être proportionnel à la racine carrée de la longueur soit dans notre cas un gain de 1,58 entre 100 mm et 250 mm (Equation 3). Expérimentalement ce gain est inférieur (environ 1,3 pour des temps de gradient suffisamment longs).
La modification du débit ne se traduit pas par une modification significative de la courbe d'évolution de la capacité de pic suggérant que ce paramètre n'influe pas sur la largeur des pics chromatographiques. Pour indication, la vitesse linéaire à 500 nL/min est de 1,9 mm/s.

Cette étude montre l'intérêt d'augmenter les longueurs de colonne afin de gagner en capacité de séparation et cela quelque soit le temps de gradient adopté. Le bénéfice de l'augmentation du temps de gradient en vue d'améliorer les capacités de séparation est d'autant plus important que la colonne est longue.
La modification du débit de phase mobile adopté pour l'élution n'impacte pas significativement sur la capacité de séparation du système. Théoriquement, la capacité de pic n'est pas affectée par le débit tant que le temps mort t_0 est significativement inférieur au temps de gradient (équation 2). Il ne peut donc pas être conclu que le faible diamètre des particules est responsable de cette robustesse de capacité de séparation en fonction du débit. La variation du débit est non significative par rapport à l'effet du gradient d'élution sur la séparation quelque soit le diamètre de particule utilisé. La même étude avec des diamètres de particules plus importants mériterait d'être réalisée afin de confirmer cette observation.
Nous montrons qu'une grande gamme de débit peut être utilisée sur notre système sans poser de problème de perte de séparation. Cette propriété semble pouvoir être utilisée pour gagner en capacité de séparation par augmentation de la longueur des colonnes chromatographiques. La diminution du débit peut permettre de compenser la surpression conséquente à l'augmentation de la perte de charge de la colonne plus longue.

f) Mise en évidence de l'influence des paramètres chromatographiques sur la sensibilité nanoESI-MS

L'influence des principaux paramètres chromatographiques sur la qualité de la séparation a été évaluée. Il convient d'évaluer désormais comment ces mêmes paramètres doivent être adaptés dans le cadre du couplage avec le système ESI-MS.
Pour mesurer l'impact de ces paramètres sur le couplage global, nous avons choisi d'étudier leur influence sur le signal mesuré en MS. Ce signal est directement lié à l'aire et à l'intensité des pics chromatographiques observés. L'aire est en théorie indépendante des paramètres chromatographiques et est seulement fonction de la réponse du détecteur. L'intensité du pic est une composante entre la réponse du détecteur et la largeur du pic chromatographique qui dépend du gradient d'élution (partie d).

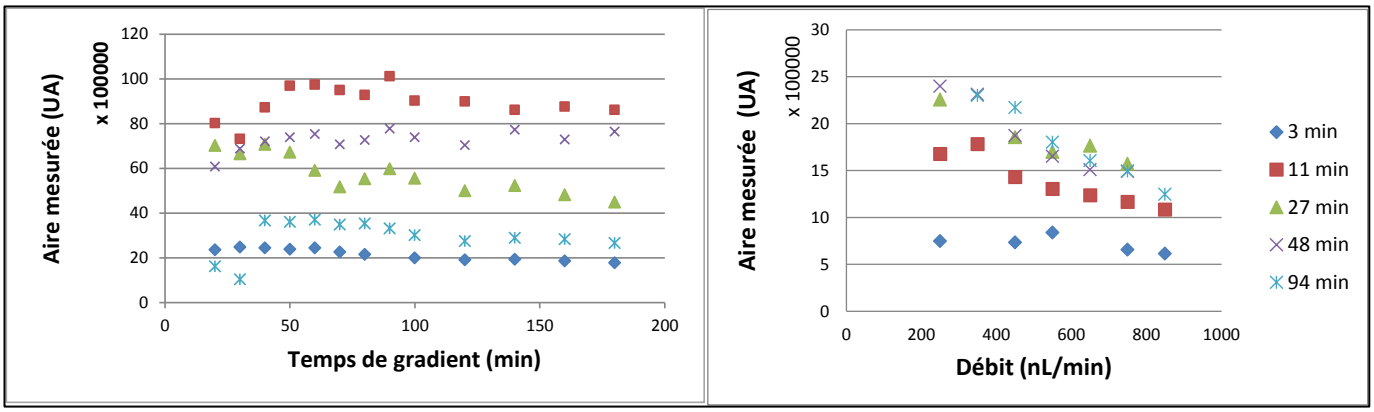


Figure 24 *: A gauche : Etude de l'évolution de l'aire mesurée sur 5 ions en fonction du temps de gradient. L'évolution de ces ions suggère une légère diminution de l'aire mesurée entre 50 et 200 minutes pour la majorité des ions examinés. Des décrochements d'intensité sont observés pour certains ions en dessous de 40 minutes.*
A droite : Etude de l'évolution de l'aire mesurée pour un ion en fonction du débit et de 5 temps de gradients. L'évolution de l'aire en fonction du débit montre une diminution significative pour tous les gradients au-dessus de 11 minutes. Les aires mesurées pour des gradients courts de 3 et 11 minutes sont significativement plus basses que pour les autres gradients dont les valeurs d'aires sont voisines à un débit donné.

L'étude montre que le temps de gradient n'a pas d'influence particulière sur les aires mesurées pour des temps de gradient relativement longs. Une légère tendance à la diminution peut néanmoins être notée pour certains composés. Elle suggère également que l'aire à tendance à diminuer pour des temps de gradients correspondants à une plus faible capacité de séparation du système. Cette diminution du signal avec la diminution de la

séparation ne peut s'expliquer que par un effet de suppression ionique au niveau de la source d'ionisation. L'arrivée simultanée de plus de composés peut effectivement être à l'origine de ce phénomène, diminuant pour des gradients plus longs avec l'augmentation de la séparation.

La diminution du signal avec l'augmentation du débit est tout à fait cohérente avec le fait que la source électrospray génère un courant d'ion fonction de la concentration en analyte. L'utilisation d'un débit de phase mobile plus important lors de l'élution entraine une dilution de l'analyte. La concentration de l'analyte au moment de son introduction en source est plus faible à fort débit ce qui explique la diminution du courant d'ion observée pour un composé donné.

En tenant compte de l'augmentation de la largeur des pics chromatographiques et de la légère décroissance de leur aire avec le temps de gradient, l'intensité des pics chromatographiques décroit donc avec l'augmentation de ce même temps de gradient.

g) Evaluation de l'impact de la quantité injectée sur la dynamique du système LC-MS

La chromatographie permet une concentration de la quantité de chaque type d'analyte au sein de quelques plateaux théoriques. En sortie de colonne la concentration de l'analyte est donc fonction de sa quantité initiale dans l'échantillon. La source d'ionisation étant concentration dépendante, la quantité d'ions obtenue est donc indirectement liée à la quantité injectée. Si cette quantité injectée est trop importante par rapport à la capacité de la source, le phénomène de suppression ionique peut être constaté : les analytes minoritaires disparaissent au profit des majoritaires.

Un second phénomène pouvant expliquer cette perte de la détection des composés minoritaires est la saturation chromatographique. A saturation, les composés majoritaires s'étalent sur plusieurs plateaux théoriques. Les composés minoritaires coélués avec ces derniers pourraient se répartir de la même façon entre les plateaux saturés ce qui aurait pour conséquence leur dilution.

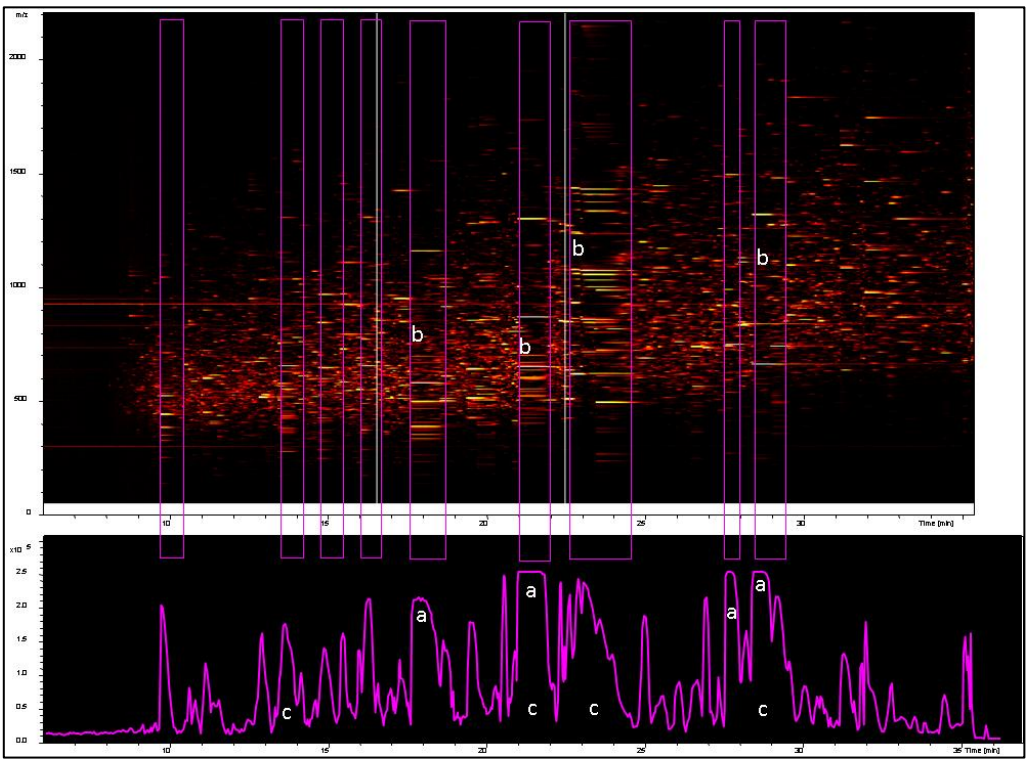


Figure 25 *: Effet de la saturation en quantité de matériel sur un système nanoLC-QTOF. (a) La saturation du digitaliseur ADC (détecteur MagneTOF) est illustrée par des pics chromatographiques « créneaux ». (b) La densité de signaux minoritaires diminue au profit des composés à saturation. (c) La saturation chromatographique est illustrée par des élargissements anormaux des pics chromatographiques à saturation. Colonne 200 mm x 75 µm 1,7 µm.*

Nous avons précédemment illustré dans la partie « digitaliseur » le phénomène de saturation du digitaliseur TDC, se traduisant par des erreurs de mesure de masse, de résolution et d'intensité. Ce phénomène de saturation impacte sur la gamme dynamique d'utilisation des systèmes nanoLC-QTOF avec détection MCP. Cette dynamique a pu être évaluée sur le système nanoLC SYNAPT G1 par le suivi de l'évolution des ratios isotopiques. Les effets de saturation du détecteur (diminution de la précision de mesure de masse, augmentation artificielle de la résolution, erreur de mesure d'intensité), commencent à être observés au-delà d'une dynamique de détection d'un ordre de 10 au-dessus du seuil de détection d'un composé [195].

Nous avons évalué la gamme dynamique du système nanoLC-QTOF MaXis équipé d'une détecteur MagneTOF à digitalisation ADC. Par injection de quantités croissantes d'un digest, nous avons observé les conséquences sur l'intensité, l'aire et la largeur de signaux chromatographiques et sur la mesure de masse et la résolution des ions considérés. Dans un premier temps et contrairement à un système MCP/TDC (résultats non présentés), la résolution des composés analysés sur la gamme étudiée n'est pas modifiée. Lorsque le système commence à saturer, une diminution de la résolution mesurée est observée : cette perte de résolution correspond au dépassement du seuil d'intensité de saturation du digitaliseur sur une partie des cycles injection/détection. La mesure de la largeur du pic n'est plus enregistrée à mi-hauteur mais en dessous : cela se traduit par sa surestimation qui conduit à une diminution de la résolution mesurée. Nous avons attribué l'absence d'erreur sur la mesure de masse à saturation au fait que le signal décrit reste symétrique. Le barycentre des points décrivant le signal restant le même, la mesure du temps de vol n'est pas significativement modifiée.

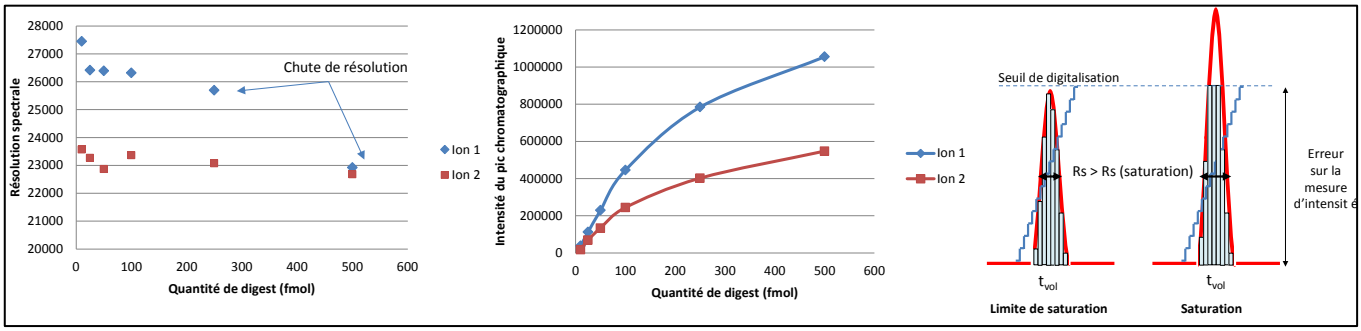


Figure 26 *: Evaluation de l'impact de l'augmentation de la quantité injectée sur les paramètres spectraux enregistrés par un détecteur MagneTOF à digitalisation ADC. A gauche : l'augmentation de l'intensité du signal se traduit à partir d'un certain seuil par une chute de la résolution mesurée. Cette chute s'explique par l'atteinte du seuil de digitalisation (schéma à droite) : la hauteur du signal est sous-estimée, le signal mesuré est artificiellement plus large. Au centre, l'évolution de l'intensité du pic chromatographique en fonction de la quantité injectée est une fonction quadratique sur la gamme étudiée.*

Du point de vue de la chromatographie, une quantité injectée trop importante se traduit par une saturation de la capacité des plateaux théoriques. L'analyte « déborde » de ces plateaux pour occuper les plateaux adjacents, ce qui entraine une augmentation de la largeur du pic chromatographique correspondant.

La combinaison de ces phénomènes peut expliquer l'évolution quadratique de la mesure de l'intensité avec la quantité de matériel injectée sur le système. La saturation chromatographique semble être la principale composante de ce phénomène puisque l'aire des pics chromatographiques n'a pas une évolution quadratique si prononcée. Ceci semble indiquer que dans un couplage nanoLC Q-TOF muni d'un digitaliseur ADC, la saturation chromatographique intervient avant la saturation spectrale. Ceci n'est pas le cas avec un Q-TOF muni d'un digitaliseur TDC où la saturation spectrale intervient bien avant la saturation chromatographique.

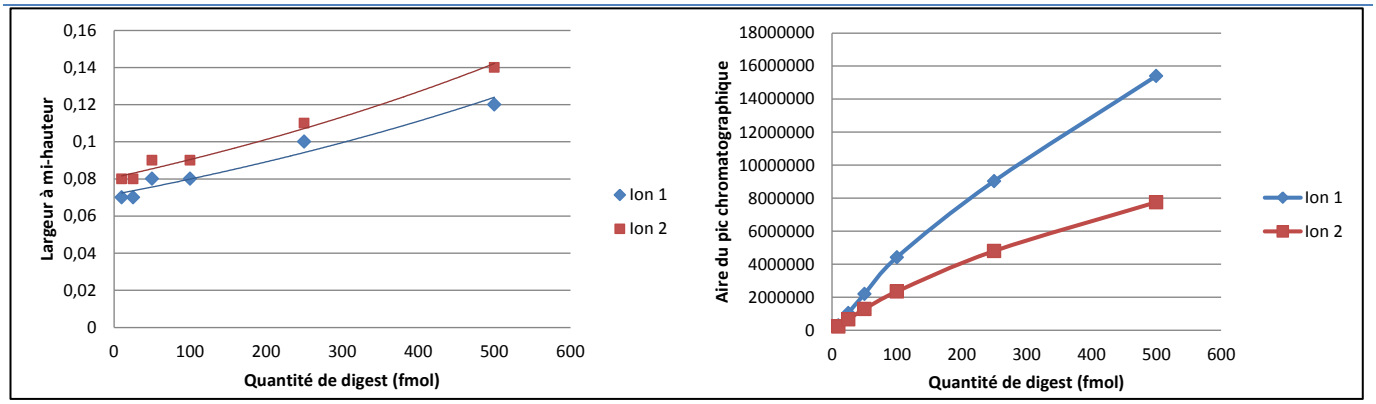


Figure 27 *: Evaluation de l'impact de la quantité injectée sur les paramètres chromatographiques de largeur à mi-hauteur (à gauche) et de l'aire des pics chromatographiques (à droite). La quantité impacte progressivement sur l'augmentation de la largeur à mi-hauteur du pic chromatographique illustrant la saturation progressive des plateaux théoriques. L'aire des pics chromatographiques contrairement à l'intensité présente un profil moins quadratique : la dynamique de mesure du signal est donc meilleure par mesure d'aire que d'intensité. Colonne 100 µm x 100 mm 1,7 µm.*

4) Etude des paramètres influant sur les identifications en protéomique par l'utilisation d'un plan d'expériences

a) La problématique de l'optimisation des paramètres d'acquisition pour la génération des données de séquençage par spectrométrie de masse.

Nous avons décrit dans le premier chapitre de cette partie le rôle de l'automatisation dans l'acquisition des données spectrales utilisées pour le séquençage peptidique en couplage LC-MS/MS. Bien qu'automatisée, cette acquisition nécessite le choix de différents paramètres permettant la programmation.

La grande question qui se pose à tous les protéomistes avant l'analyse est généralement : « *comment tirer le meilleur parti de l'instrument afin d'avoir un maximum de résultat d'identification pour mes échantillons* ? ». Cette problématique s'accompagne souvent de contraintes : la quantité de matériel et le temps d'utilisation du spectromètre ne sont généralement pas illimités. En effet, des séries de plusieurs dizaines voire centaines de fractions d'un protéome décomplexifié limite les temps d'analyse d'un échantillon individuel. Dans ces conditions, les méthodes d'acquisition utilisée sont standards et une incertitude demeure quant à leur adéquation avec l'analyse de l'échantillon étudié. La problématique n'est pas d'éviter d'analyser incorrectement un échantillon, l'analyse protéomique donnant toujours des résultats, mais de tendre vers une identification complète de l'ensemble des composés pouvant passer au-dessus des seuils de détection et d'identification.

Le problème se pose de manière similaire lors de l'identification d'isoformes. Le sous échantillonnage des extraits peptidiques analysés confronte à une probabilité non négligeable de manquer l'identification de peptides uniques et de perdre potentiellement l'identification d'une ou de plusieurs protéines.
Un réglage adapté de l'instrument pour ce type d'acquisition nécessite de connaitre les effets des paramètres de programmation pouvant influer sur l'efficacité et la qualité du séquençage. Les expériences à mettre en œuvre sont difficiles à planifier dès lors qu'un nombre important de paramètres doit être évalué. Pour pallier à ce problème, nous avons choisi d'utiliser la méthodologie des plans d'expérience.

La méthodologie des plans d'expériences est une approche mathématique et statistique d'organisation des expériences à mener pour obtenir une étude précise et fiable d'un système en un minimum d'expériences [237-242].

b) Terminologie du plan d'expérience

Avant d'introduire le principe et la réalisation d'un plan d'expériences, le vocabulaire utilisé nécessite d'être défini. Nous allons étudier un système dans lequel des **facteurs** variables ou paramètres x_i influent sur un phénomène pour lequel il est possible de mesurer des **réponses** y_i. L'étude du système passe par la connaissance des **effets** $y_i = f(x_i)$ qui existent entre les facteurs et les réponses. Afin de connaitre ces liens, il est nécessaire d'expérimenter et de réaliser des essais en faisant varier les facteurs x_i et en mesurant leur impact sur la ou les réponses y_i. Des interactions peuvent exister entre les facteurs pour une réponse données : cela signifie que l'effet d'un facteur x_1 sur une réponse est différent en fonction de la valeur utilisée pour un autre facteur x_2. Le système à étudier définit un domaine expérimental. L'étude du système peut être restreinte à un domaine d'étude, fraction de ce domaine expérimental.

Il est possible de réaliser une optimisation classique en faisant varier un seul paramètre sur le domaine d'étude pendant que les autres demeurent constants. Les différentes réponses peuvent alors être étudiées en fonction de cet unique paramètre. Au contraire, l'approche par plans d'expériences consiste à faire varier simultanément plusieurs paramètres à chaque essai. Cette variation non linéaire est déterminée par un plan préétabli et optimisé pour obtenir le maximum d'information en un minimum d'essais. Cette démarche va permettre de quantifier les effets des facteurs sur la réponse et d'établir d'éventuelles interactions entre les facteurs. Il est alors possible de hiérarchiser ces effets afin de connaitre les paramètres les plus influents pour une réponse donnée.

Le plan d'expérience définit les variations des paramètres à fixer pour étudier la réponse. Les variations sont réalisées sur un domaine d'étude codé pour chaque paramètre entre des valeurs définies par un **niveau bas** : **-1** et un **niveau haut : +1** pour chaque paramètre. Une fois le plan choisi, il est nécessaire de traduire les valeurs codées en valeurs réelles. Le choix du domaine d'étude doit être défini par l'expérimentateur qui détermine les valeurs réelles du paramètre étudié entre lesquelles il souhaite observer l'effet sur la ou les réponses. Une matrice d'expérience est alors constituée : elle définit l'ensemble des valeurs réelles des paramètres à fixer pour chaque expérience.

c) Les catégories de plan d'expérience

Différents plans d'expériences ont été développés et peuvent être choisis en fonctions des contraintes expérimentales et du degré de précision avec lequel on souhaite étudier un système. Les paramètres peuvent définir des valeurs discrètes (par exemple le nombre d'ion sélectionnés) ou continues (par exemple le temps de gradient). Le nombre d'expérience peut être limité par des coûts expérimentaux (en temps ou en difficulté de réalisation par exemple). Dans le cas le plus simple, les effets des facteurs sur les réponses peuvent définir un lien linéaire (fonction du premier degré). Dans d'autre cas, l'effet ne peut être correctement modélisé que par une fonction de degré supérieur permettant d'exprimer une relation quadratique (fonction du second degré).

Les plans d'expériences les plus simples sont réalisés en faisant varier les paramètres pour deux modalités : le niveau haut et le niveau bas. Ces plans sont appelés plans factoriels complets à deux niveaux et sont adaptés pour un nombre limité de facteurs à étudier. Il est nécessaire d'effectuer 2^n expériences avec n le nombre de facteurs. Ces plans permettent de quantifier les effets des facteurs et de leurs interactions mais n'apportent pas de connaissance sur la nature de l'effet reliant facteur et réponse (premier ou second degré). Il est possible d'avoir une estimation de la nature de cet effet en réalisant une expérience au centre pour laquelle tous les facteurs sont mis à un **niveau central** : **0**. Si cet essai est répété, il est également possible d'avoir accès à l'erreur expérimentale et d'établir une éventuelle déviation de la réponse au cours de l'expérience.
Lorsqu'un nombre important de paramètres doit être étudié, d'autres plans peuvent être utilisés comme alternative. Les plans factoriels fractionnés permettent de diminuer le nombre des essais d'un plan factoriel complet.

Lorsqu'il est envisagé de modéliser les effets des variables avec un modèle du second degré, les plans factoriels sont inadéquats. Les plans du second degré peuvent alors être utilisés. Ces plans consistent à réaliser des mesures à plus de deux niveaux par facteur afin d'accéder à l'information de la composante quadratique éventuelle de chaque effet.

d) Construction du plan

La vocation de notre plan d'expérience n'est pas de déterminer quelles sont les valeurs optimales des paramètres à fixer pour avoir le maximum d'identification : il sera utilisé afin d'étudier et hiérarchiser les effets des facteurs sélectionnés sur la réponse choisie et de mettre en évidence ou non d'éventuelles interactions entre les facteurs. Il convient de définir dans un premier temps les paramètres à étudier puis de choisir la matrice d'expériences la plus appropriée à l'étude. Par la suite nous définirons les bornes du domaine d'étude de chaque paramètre. Le système étudié sera un échantillon standard composé d'un digest de 8 protéines injecté sur un couplage nanoAcquity-SYNAPT G1.

Choix des paramètres étudiés

Nous souhaitons connaitre l'effet des paramètres d'acquisition des spectres sur les résultats de séquençage de l'échantillonnage. La programmation de l'acquisition en mode dépendant des données nécessite le choix de valeurs pour différents paramètres.

Parmi ces paramètres nous avons identifié cinq variables d'étude. Nous avons choisi d'étudier les temps de sommation lors des acquisitions MS et MS/MS, le nombre d'ions analysés en MS/MS après chaque analyse MS (ou nombre de précurseurs), le seuil d'intensité au-delà duquel un ion peut être sélectionné en vue d'être fragmenté et le temps pendant lequel un précurseur est exclu de la sélection. D'autres paramètres pouvant être fixés à la programmation peuvent être considérés comme optimisés indépendamment (gammes de m/z acquises en MS et MS/MS, état de charge considéré pour la sélection, énergie appliquée pour la fragmentation, critère de retour à la MS).

Nous avons précédemment établi l'effet de paramètres instrumentaux au niveau du chromatographe et du spectromètre de masse sur les réponses du type intensité détectée, géométrie des pics chromatographiques et séparation des composés. La connaissance du système nous permet d'ores et déjà d'écarter un certain nombre de paramètres d'acquisition qui peuvent être choisis indépendamment des autres.

Nous avons choisi de travailler à géométrie de colonne fixe, température et débit constants au niveau du chromatographe. Aucun paramètre n'est modifié au niveau de la source, de la transmission, de l'isolement, de la fragmentation, de la détection et du traitement des données. Nous pouvons en effet envisager que chacune de ces composantes peut être optimisée pas à pas indépendamment de toutes les autres.
Les seuls paramètres instrumentaux qui nous ont parus pertinents d'étudier en combinaison avec les paramètres d'acquisition automatique des spectres MS/MS sont le temps de gradient et la quantité injectée. Nous avons précédemment déterminé que ces paramètres étaient les plus influents respectivement sur la séparation des composés et leur temps d'élution, et sur l'aire des pics chromatographiques.

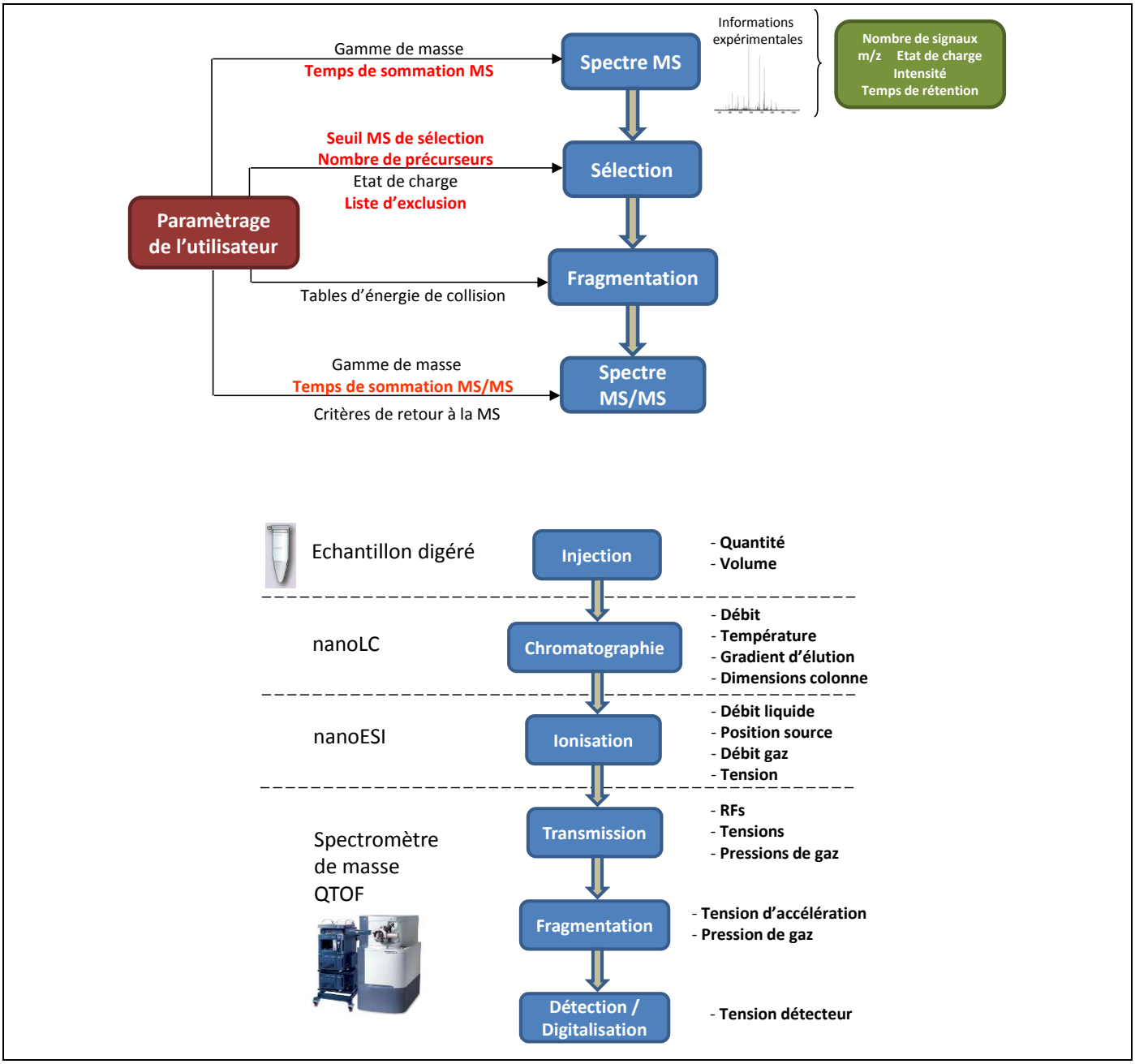


Figure 28 *: Représentation des différents paramètres pouvant influer sur le séquençage des peptides par MS/MS. En haut, représentation des paramètres d'acquisition en mode dépendant des données ; en rouge les paramètres investigués. En bas, les paramètres du système nanoLC-MS-MS/MS.*

Définition du domaine d'étude

Une fois les facteurs choisis, il est nécessaire de définir les bornes des domaines d'étude de chaque paramètre. Ce choix est réalisé sur la base de nos connaissances préliminaires.

La quantité d'échantillon injectée a été choisie entre une valeur basse estimée comme suffisante pour obtenir des résultats corrects d'identification sur les protéines majoritaires et une valeur haute estimée comme proche de la saturation du couplage du point de vue de la chromatographie et de la détection.

Les temps de sommation MS et MS/MS ont été choisis entre une valeur basse estimée comme courte mais suffisante pour obtenir un signal et une valeur plus longue permettant d'augmenter le rapport signal sur bruit (augmentation théorique du rapport signal sur bruit d'un facteur 2,2 entre le niveau bas et le niveau haut en MS et d'un facteur 1,7 en MS/MS).

Le temps de gradient a été choisi entre une valeur basse et une valeur haute correspondant à une augmentation du temps efficace de séparation d'un facteur 2.
Le temps d'exclusion a été fixé entre une valeur basse correspondant à une absence d'exclusion et une valeur haute correspondant à l'exclusion pendant la durée moyenne d'un pic chromatographique (entre 25 et 45 secondes à la base).

Paramètre	Désignation	Valeur basse -1	Valeur centrale 0	Valeur haute +1
X1	Quantité (fmol)	12,5/65	38/195	62,5/325
X2	Temps de sommation MS (s)	0,25	0,75	1,25
X3	Temps de gradient (min)	43,5	68	92,5
X4	Temps d'exclusion (s)	3	17	31
X5	Temps de sommation MS/MS (s)	0,5	1	1,5
X6	Seuil MS (Cps/s)	10	30	50
X7	Nombre d'ion	2	4	6

***Tableau 3** : Domaine d'étude défini pour les 7 facteurs étudiés au cours du plan d'expérience. Les protéines huit protéines ne sont pas injectées en quantités équimolaires ; leurs quantités sont comprises entre deux valeurs indiquées dans le tableau.*

Le seuil d'intensité MS a été choisi entre une valeur basse proche du bruit de fond de l'instrument et une valeur haute pour laquelle les signaux sont significativement distingués du bruit.
Enfin le nombre d'ions choisi pour la fragmentation a été fixé entre un niveau bas de 2 et un niveau haut de 6. Parmi les paramètres étudiés, il s'agit de la seule variable discrète.

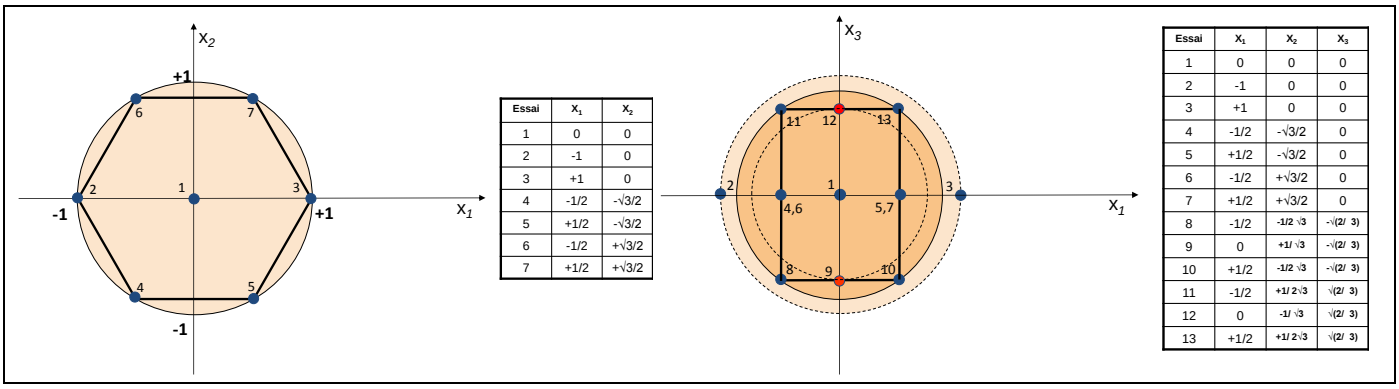
Choix des réponses étudiées

Il est possible d'étudier une multitude de réponses en utilisant la même matrice d'expériences. Dans le cas d'une étude protéomique, différentes réponses peuvent être quantifiées. Suite à chaque essai, le **nombre de spectres réalisés** en MS/MS est mesuré.
La recherche d'identification dans les banques de données apporte différents résultats :
Il est possible de considérer le **nombre de spectres assignés** c'est-à-dire ayant donné lieu à une identification. Ce nombre ne tient pas compte du fait que plusieurs spectres peuvent donner la même identification (sélection du même peptide à différents états de charge, redondance de sélection).
La mesure du **nombre de spectres uniques** est réalisée. Les spectres uniques excluent du comptage l'ensemble des spectres redondants.
Afin d'obtenir une notion de la qualité des spectres obtenus, les scores d'identification des ions identifiés ont été extraits. En pondérant la somme des scores d'ions par le nombre de spectres assignés, un **score d'ion moyen** est obtenu pour chaque essai.
Le nombre de spectres assignés a également été pondéré par le nombre de spectres réalisés. La mesure du **pourcentage de spectres assignés** est ainsi obtenue.
La différence entre le nombre de spectres assignés et le nombre de spectres uniques permet le calcul du nombre de spectres redondants. En pondérant le nombre de spectres redondants par le nombre de spectres réalisés, le **pourcentage de spectres redondants** est calculé.

e) Choix et principe de construction de la matrice de Doehlert

Nous avons fait le choix d'étudier l'ensemble des facteurs en partant de l'hypothèse qu'un certain nombre d'entre eux pouvaient avoir une influence quadratique sur les réponses. Ce choix nous a orientés vers un plan

d'expériences du second degré. Le nombre d'expériences pouvant être réalisé n'étant dans notre cas pas limité (il est possible de programmer une séquence d'analyses sur plusieurs jours), nous nous sommes tournés vers une matrice permettant de décrire précisément l'espace expérimental : une matrice de Doehlert. Cette matrice va permettre l'étude des effets du premier et du second degré des facteurs et d'étudier les interactions du premier ordre pouvant exister entre les 7 facteurs.



Essai	X_1	X_2
1	0	0
2	-1	0
3	+1	0
4	-1/2	-√3/2
5	+1/2	-√3/2
6	-1/2	+√3/2
7	+1/2	+√3/2

Essai	X_1	X_2	X_3
1	0	0	0
2	-1	0	0
3	+1	0	0
4	-1/2	-√3/2	0
5	+1/2	-√3/2	0
6	-1/2	+√3/2	0
7	+1/2	+√3/2	0
8	-1/2	-1/2 √3	-√(2/ 3)
9	0	+1/ √3	-√(2/ 3)
10	+1/2	-1/2 √3	-√(2/ 3)
11	-1/2	+1/ 2√3	√(2/ 3)
12	0	-1/ √3	√(2/ 3)
13	+1/2	+1/ 2√3	√(2/ 3)

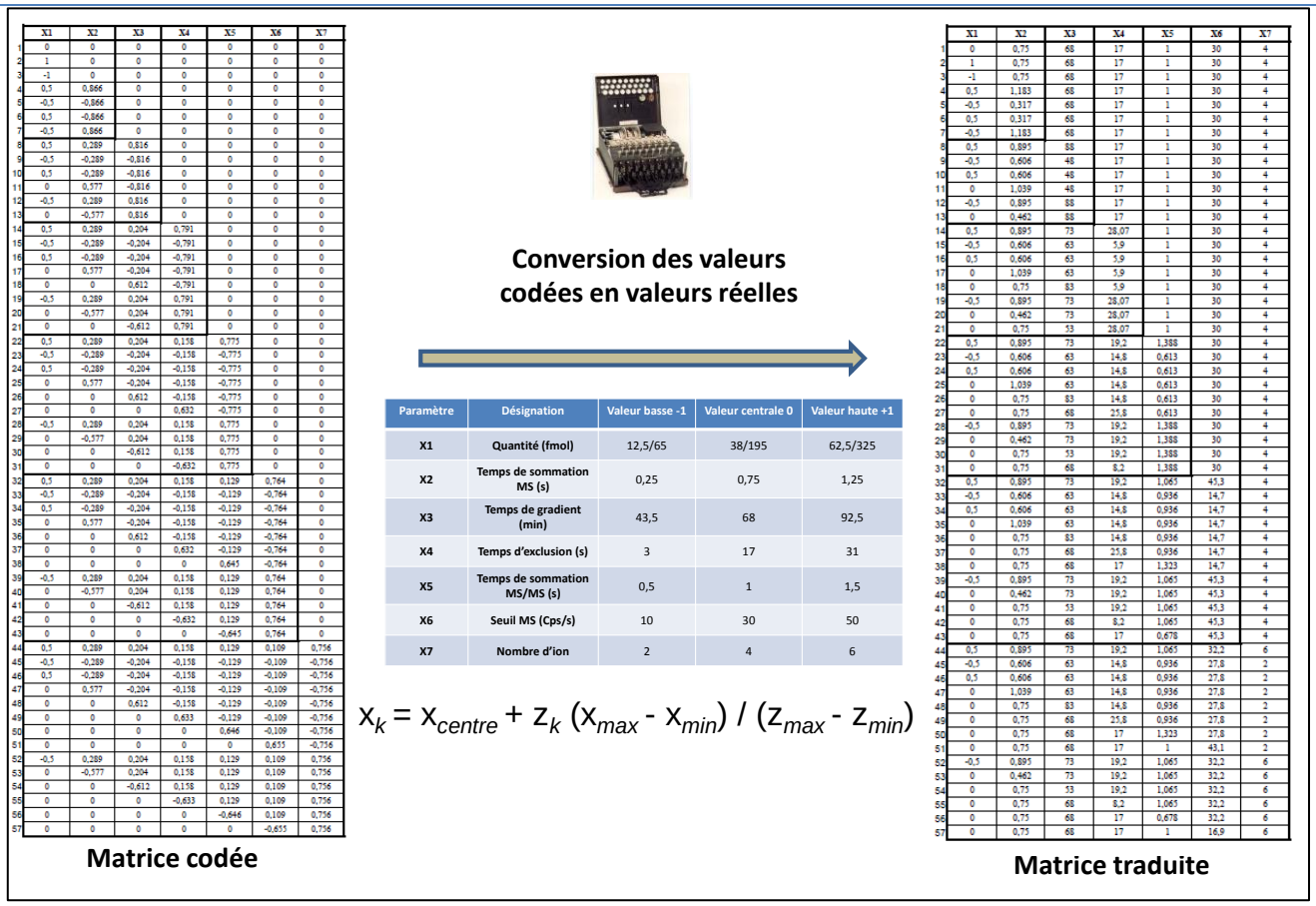
***Figure 29** : Illustration de la répartition des essais décrits par la matrice de Doehlert pour l'étude de deux facteurs x_1 et x_2 (à gauche) et pour l'extension à un facteur supplémentaire x_3 (à droite). L'extension à un facteur supplémentaire introduit 4 essais (8, 10, 11 et 13) permettant de construire un cercle sur la dimension (x_1 ; x_3) et 6 essais (de 8 à 13) permettant de construire un cercle sur la dimension (x_2 ; x_3). La distribution des essais établie permet d'avoir un nombre de points important pour décrire chaque domaine expérimental bidimensionnel. Ce nombre de points augmente avec l'augmentation du nombre de facteurs étudiés.*

La matrice de Doehlert [243] est une matrice construite pour une exploration circulaire du domaine expérimental. Pour chaque étude de couple de facteurs, des combinaisons de valeurs permettent d'inscrire les essais dans un cercle autour du niveau central 0 de ces paramètres. Le nombre de modalités pour chaque paramètre est toujours supérieur à 3 afin de quantifier les effets quadratiques. L'interaction du premier ordre est également quantifiée pour chaque couple de paramètre. Pour l'étude d'un facteur supplémentaire des essais sont ajoutés aux précédents afin de pouvoir décrire deux nouveaux cercles étudiant les deux nouveaux couples de facteur. Le principe est le même à chaque ajout d'un nouveau paramètre, la matrice peut croître sans limitations en nombre. Le domaine expérimental est ainsi étudié par des essais inscrits dans des cercles sur n-1 dimensions (n, nombre de facteurs). La répartition des essais sur le cercle offre la possibilité de réaliser ultérieurement la transposition du domaine d'étude par ajout d'essais supplémentaires.
Les avantages de cette matrice d'expérience résident dans sa flexibilité (l'étude d'un nouveau paramètre utilise les résultats des précédents essais) et sa précision (le nombre de points utilisé pour la quantification des effets est important).

f) Construction de la matrice expérimentale

La matrice de Doehlert à 7 facteurs a été utilisée. L'attribution des facteurs étudiés est réalisée de manière à faciliter la mise en place de l'expérience. Le facteur x_1 de la matrice ne varie que sur 5 modalités : il a ainsi été associé à la quantité injecté afin de permettre de ne préparer physiquement que 5 niveaux de concentration d'échantillon. Le facteur x_7 varie sur trois modalités : il a été associé au nombre d'ions sélectionnés, le seul facteur représentant des variables discrètes. La conversion en valeurs réelles est ainsi facilitée.

La conversion des valeurs codées de la matrice d'expérience en valeur réelles est réalisée par une simple formule en fonction des valeurs déterminées pour les niveaux bas, haut et au centre du facteur et de la valeur codée associée à l'essai. Cette conversion permet de construire la matrice d'expériences.



Paramètre	Désignation	Valeur basse -1	Valeur centrale 0	Valeur haute +1
X1	Quantité (fmol)	12,5/65	38/195	62,5/325
X2	Temps de sommation MS (s)	0,25	0,75	1,25
X3	Temps de gradient (min)	43,5	68	92,5
X4	Temps d'exclusion (s)	3	17	31
X5	Temps de sommation MS/MS (s)	0,5	1	1,5
X6	Seuil MS (Cps/s)	10	30	50
X7	Nombre d'ion	2	4	6

$$x_k = x_{centre} + z_k\,(x_{max} - x_{min}) \,/\, (z_{max} - z_{min})$$

Figure 30 *: Principe de conversion de la matrice de Doehlert en une matrice d'expériences contenant des valeurs réelles.*

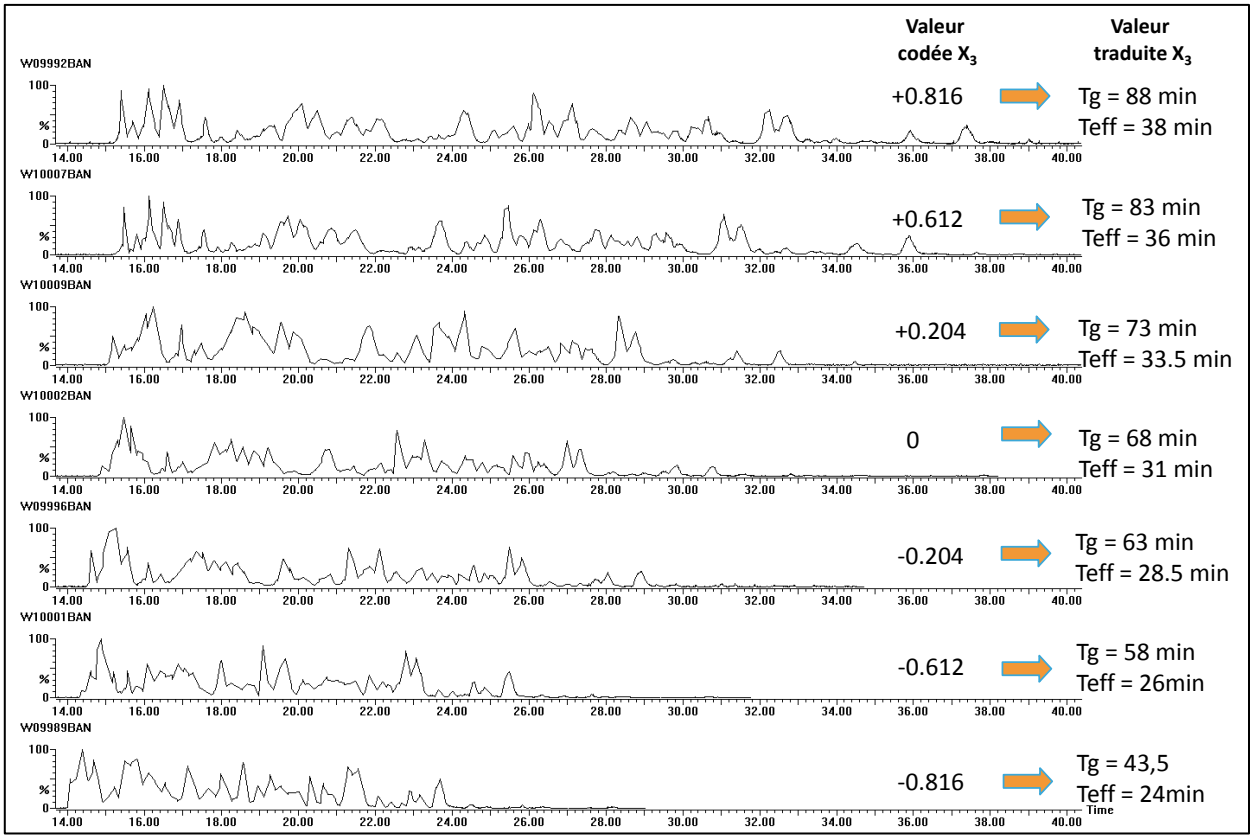


Figure 31 *: Illustration de l'évolution du facteur temps de gradient sur les 7 modalités définies par la matrice de Doehlert (facteur x_3).*

g) Réalisation de la matrice réponse et principe de la mesure des effets

La programmation de la séquence d'acquisition a été réalisée. Pour chaque essai, une méthode d'acquisition MS est créée. Un nouvel échantillon, à la concentration déterminée par la modalité du paramètre « quantité injectée » est utilisé lors de chaque essai. Le volume d'injection est ainsi gardé constant. Toutes les huit analyses, un essai au centre est enregistré afin de mesurer les variations de la réponse au cours de l'expérience et ainsi connaitre l'erreur expérimentale sur les mesures.

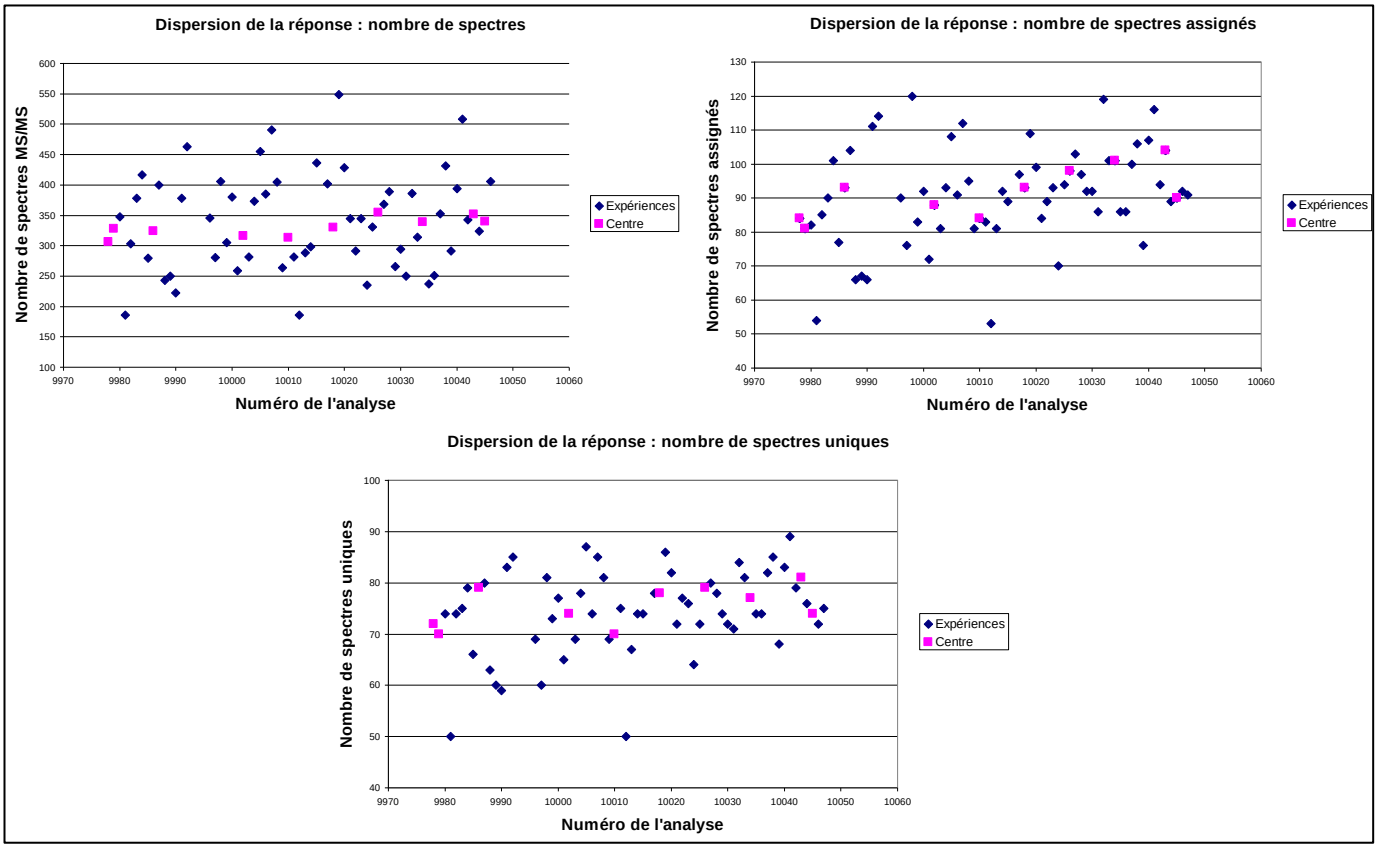


Figure 32 *: Illustration de la dispersion de différentes réponses calculées en fonctions des essais réalisés au cours de l'expérience par rapport aux essais effectués au centre.*

Pour chaque essai, les réponses précédemment décrites ont été calculées. La visualisation de la valeur des réponses en fonction des essais réalisés montre une dispersion significative traduisant la présence d'effets des facteurs. L'analyse de ces effets et des facteurs responsables est réalisée par traitement statistique. Le principe consiste à résoudre un système d'équations, un pour chaque couple de facteurs, où les inconnues sont les 6 coefficients des équations du second degré modélisant les effets. Cette résolution est obtenue en utilisant la méthode des moindres carrés *via* le logiciel STATISTICA (StatSoft) [244].

Les coefficients déterminés, les effets linéaires, quadratiques et les interactions du premier ordre sont mesurés. La significativité de ces effets est alors évaluée par analyse de la variance (ANOVA). L'ANOVA peut être réalisée car la matrice d'expérience comporte plus d'essais que de facteurs étudiés. Elle est affinée par la répétition des essais au centre.

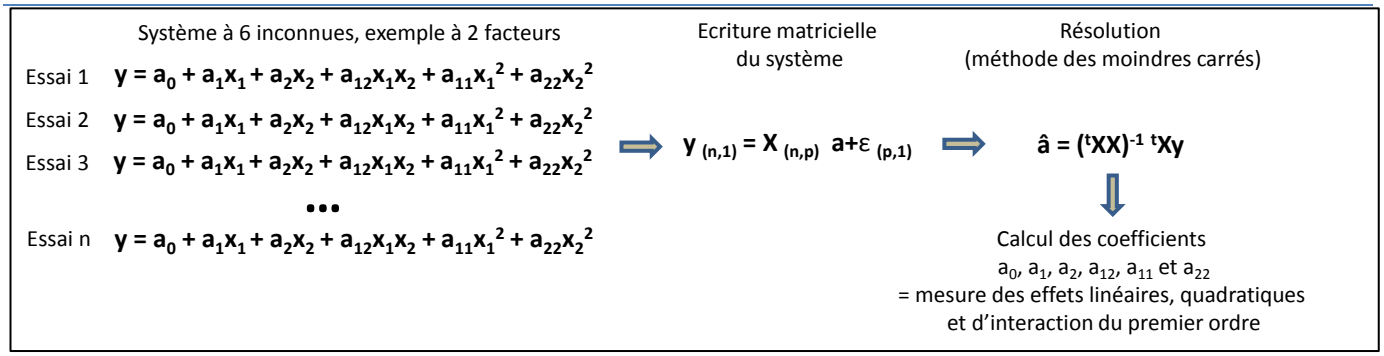


Figure 33 *: principe de modélisation du système expérimental à partir des données issues des essais effectués lors du plan d'expérience.*

h) Identification des paramètres influents sur les réponses

Les résultats de l'analyse de la variance pour chaque effet permettent de les classer par ordre de significativité.
Les effets sont jugés significatifs au-dessus d'une valeur p correspondant à 95% de confiance. Pour chaque réponse les estimations des effets obtenus sont présentés sous forme de diagramme de Pareto. Le signe de l'effet ainsi que sa valeur permettent de décrire l'influence du facteur correspondant. Lorsqu'un facteur possède des effets linéaire et quadratique significatifs, l'effet peut être interprété comme la combinaison d'un effet croissant ou décroissant dont la courbe est plus ou moins « bombée » en fonction de l'importance de l'effet quadratique. Un effet quadratique positif entraine une courbe bombée vers le bas et vers le haut s'il est négatif.
La lecture des diagrammes de Pareto (Figure 34) permet d'apporter les informations suivantes sur l'influence des différents paramètres :

Influence du temps de gradient

Le temps de gradient est le paramètre le plus influent sur le nombre de spectres MS/MS réalisés. Logiquement l'augmentation du temps disponible pour l'acquisition permet d'augmenter le nombre de spectres assignés et de la même façon, le nombre de spectres uniques acquis. Ces effets peuvent être associés à la décomplexification apportée par l'augmentation de la capacité résolutive du système chromatographique mais également à la diminution du sous-échantillonnage permise par l'augmentation du nombre de spectres réalisés.
Si l'augmentation du temps de gradient apporte des effets positifs, nous constatons qu'il impacte sur la diminution du score d'ion moyen des spectres assignés. La première explication de ce phénomène peut être que les spectres assignés en plus grâce à ce facteur peuvent provenir d'ions parents ayant moins de probabilité d'être sélectionnés du fait de leur plus faible intensité. Leur identification a plus de chance de donner un résultat dont le score associé sera plus faible, entrainant une diminution de la moyenne globalement mesurée.
L'augmentation du temps de gradient a également un impact sur la redondance. Nous savons que des gradients plus longs impliquent l'élargissement des pics chromatographiques. Cet élargissement accroit la probabilité de sélectionner deux fois le même composé.

Influence du temps d'exclusion

Ce facteur a principalement un effet sur la redondance, contraire à celui du temps de gradient. Cette diminution d'effet sur la redondance est plafonnée comme le traduit la significativité de la composante quadratique pour ce facteur.
Le temps d'exclusion a un effet défavorable à la limite de la significativité sur le nombre de spectres enregistrés et assignés. L'absence d'effet significatif sur le nombre de spectres uniques montre que les spectres non réalisés sous l'influence du temps d'exclusion sont principalement des spectres redondants.

Il est ainsi possible d'utiliser ce paramètre de manière concertée avec le temps de gradient choisi pour compenser l'effet de redondance. La valeur du temps d'exclusion doit donc être choisie sur des valeurs hautes du domaine d'étude qui correspondent environ à la largeur du pic chromatographique à la base.

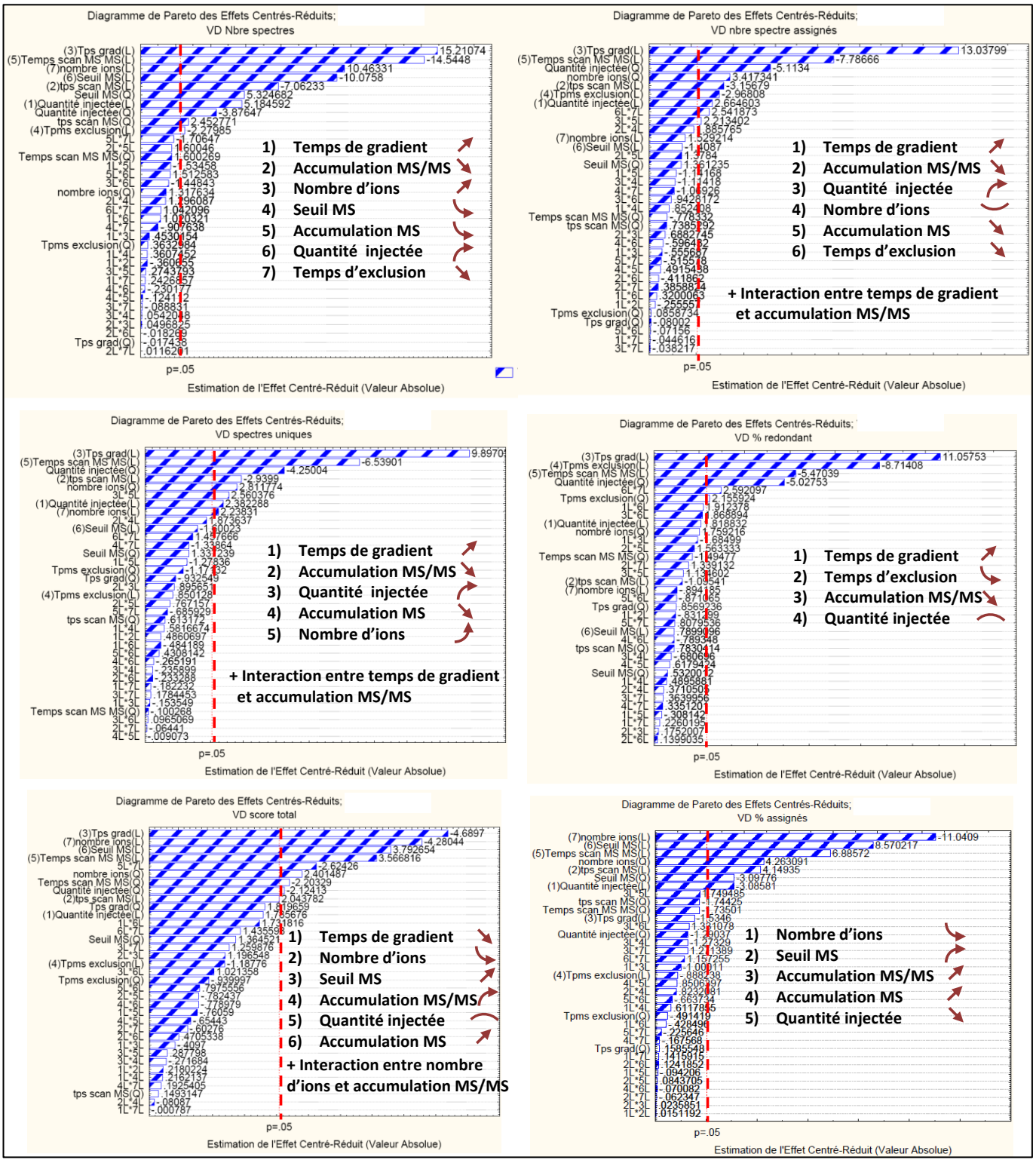


Figure 34 *: Diagramme de Pareto des effets centrés réduits calculés pour les 6 réponses investiguées au cours de cette étude. Les effets des facteurs sont jugés significatifs au dessus d'une valeur p correspondant à 95 % de confiance. Pour chaque réponse les effets des facteurs sont classés en tenant compte de la combinaison des effets linéaires et quadratiques.*

Le temps de sommation MS

Ce facteur impacte défavorablement sur le nombre de spectre réalisés et dans une moindre mesure sur le nombre de spectres assignés et uniques. Ce résultat suggère qu'il peut y avoir un bénéfice à diminuer le temps de sommation MS pour laisser davantage de temps disponible à l'acquisition de spectres MS/MS supplémentaires. Le temps minimum utilisé sur le domaine d'étude (0,3 seconde) peut donc être envisagé pour améliorer légèrement les résultats.

Le temps de sommation MS/MS

Après le temps de gradient, le temps de sommation MS/MS est le paramètre qui influe le plus sur le nombre de spectres réalisés. La diminution de ce temps permet logiquement d'augmenter le nombre de spectre MS/MS réalisés. L'effet de la diminution de ce paramètre se traduit également par une augmentation du nombre de spectres assignés et uniques. Cet effet est néanmoins moins influent que celui du temps de gradient sur les résultats d'identification. Ce résultat peut être interprété par le fait que la diminution du temps d'acquisition se traduit par une diminution du rapport signal sur bruit, préjudiciable à certaines identifications. D'autre part, l'augmentation du temps de gradient permet une décomplexification de l'échantillon et une diminution de la suppression ionique, ce que ne permet pas la modification du temps de sommation.
Nous noterons qu'un effet d'interaction significatif est observé entre le temps de gradient et le temps d'accumulation MS/MS pour les réponses nombre de spectres assignés et nombre de spectres uniques. Cette interaction se traduit par un effet bénéfique de l'augmentation du temps de gradient sur le nombre d'identifications moins important lorsque le temps de sommation est le plus court. Ce résultat suggère qu'il est possible de diminuer le temps de gradient donc d'analyse si le temps d'accumulation MS/MS est raccourci sans perte d'identification.

Ce facteur influe également sur la redondance. La réalisation d'un nombre de spectres plus important par diminution du temps d'acquisition favorise la sélection de composés déjà identifiés. De la même manière que pour le temps de gradient, ce phénomène peut être corrigé avec le temps d'exclusion.

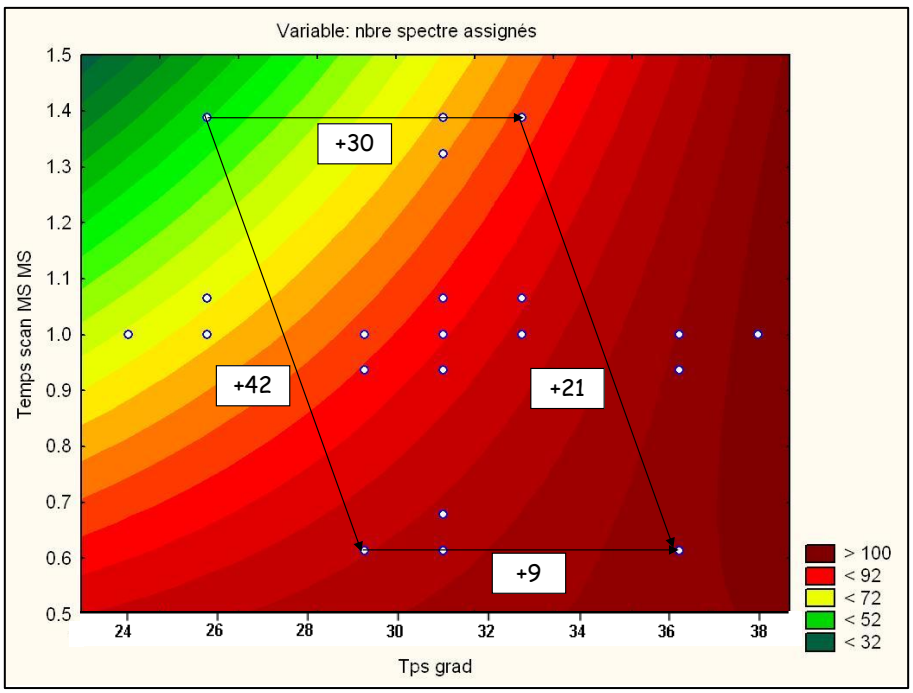


***Figure 35** : Courbes d'isoréponse « nombre de spectres assignés » en fonction des facteurs les plus influents « temps de gradient » et « temps de sommation MS/MS ». L'existence d'une interaction significative entre ces facteurs est illustrée par la différence de l'effet d'un des facteurs en fonction du niveau haut ou du niveau bas de l'autre facteur.*

Le bénéfice de l'augmentation du rapport signal sur bruit par augmentation du temps de sommation est illustré par son effet sur le score d'ion moyen des peptides identifiés. Pour autant, ce résultat doit être pondéré par le fait que l'augmentation du temps des sommations diminue le nombre d'acquisitions et favorise donc la sélection de composés dont l'intensité du parent est plus élevée. Le même raisonnement explique que l'augmentation du temps MS/MS augmente le pourcentage de spectres assignés.

Ces différents résultats suggèrent que le temps d'accumulation MS/MS doit être fixé au plus court possible afin de permettre un gain sur le nombre d'identifications. Ce raccourcissement expose cependant à une diminution de certaines identifications nécessitant un nombre de sommation plus élevé, permettant d'améliorer le rapport signal sur bruit.

La quantité injectée

La quantité injectée est le troisième facteur dont l'effet est le plus significatif sur le nombre de spectres assignés et le nombre de spectres uniques. Ce facteur est moins influent sur le nombre total de spectres mais montre que plus d'ions passent au-dessus du seuil de sélection MS avec l'augmentation de la quantité. L'augmentation de la quantité injectée influe ainsi directement sur l'augmentation de l'assignation des spectres sélectionnés. Nous noterons cependant une composante d'effet quadratique important de ce facteur pour les réponses nombre de spectres assignés et uniques : l'augmentation du nombre de composés identifiés « sature » pour les quantités les plus importantes utilisées sur le domaine d'étude. Le phénomène de saturation observé précédemment lors de l'étude du couplage LC-MS se traduit donc également par une saturation en termes de composés identifiés. Le phénomène de suppression ionique associée à l'augmentation de la largeur des pics chromatographiques semblent donc avoir un impact sur les identifications.
L'effet quadratique du facteur quantité injectée est également retrouvé pour les réponses mesurant la redondance et le score d'ion moyen.

Ces résultats suggèrent qu'une augmentation de la quantité injectée est bénéfique jusqu'aux limites de capacité du système LC-MS.

Le seuil de sélection MS

Ce facteur a un effet principalement sur le nombre de spectres MS/MS réalisés auquel s'additionne un léger effet quadratique. La diminution du nombre de spectre acquis avec l'augmentation du seuil de sélection est facilement appréhendable. En revanche, les effets de ce facteur sont peu significatifs sur les réponses nombre de spectres assignés et uniques. Cette information suggère que les composés non sélectionnés suite à l'augmentation du seuil ne sont pas des composés qui apportent significativement des identifications supplémentaires.

Nous pouvons également noter que l'augmentation du seuil MS permet l'augmentation du score d'ion moyen des spectres assignés. Sachant que ce facteur impacte peu sur le nombre de spectres assignés, nous pouvons en conclure que ce facteur peut être utilisé afin d'améliorer la qualité des spectres MS/MS acquis. Ce phénomène peut s'expliquer par le fait qu'une sélection prématurée (au pied du pic chromatographique) se traduit par une intensité plus faible lors de la fragmentation du composé sélectionné en montée de pic chromatographique. Le rapport signal sur bruit sur les fragments peut donc dépendre du moment où est réalisée la sélection du parent.

Ce résultat suggère que le seuil MS peut être utilisé pour contrôler la qualité des spectres d'identification sans compromis avec le nombre d'identifications.

Le nombre d'ions sélectionnés pour la fragmentation

L'augmentation du nombre d'ions sélectionnés se traduit par une augmentation du nombre de spectres MS/MS réalisés. Cette tendance s'explique par le fait qu'une augmentation de ce nombre implique une diminution du nombre d'acquisition MS au profil du nombre d'acquisition MS/MS. Les spectres supplémentaires ainsi acquis ne semblent pas donner lieu à une augmentation significative du nombre de spectre assignés. La combinaison de ces deux résultats explique l'effet négatif important du facteur sur le pourcentage de spectres assignés. Les spectres MS/MS supplémentaires ont plus de probabilité de provenir de parents dont l'intensité est faible et ne donnent pas de spectres de fragmentation présentant un signal sur bruit correct. Ce phénomène s'illustre par l'effet négatif de l'augmentation du nombre d'ion sur le score d'ion moyen par analyse.

Le choix de la valeur de ce facteur est ainsi plus ambigu. Si un nombre important d'ions est sélectionné au-dessus du seuil de sélection MS et que ce dernier n'est pas adapté, les spectres MS/MS réalisés pour des ions parents de faible intensité ont une forte probabilité de ne pas donner d'identification.

i) Développement d'une stratégie d'optimisation

La hiérarchisation des effets des différents facteurs étudiés sur les réponses traduisant l'efficacité et la qualité du séquençage de peptide par MS/MS a ainsi été réalisée. Cette hiérarchisation permet d'envisager de manière réfléchie une procédure d'optimisation pas à pas de l'ensemble de ces facteurs.

Nous avons démontré que la plupart des facteurs sont indépendants les uns des autres, ce qui permet d'optimiser prioritairement les paramètres les plus influents sur les réponses avant d'envisager l'optimisation des autres.
La quantité injectée permet un bénéfice quant au nombre de spectre identifié. Si la quantité d'échantillon le permet, elle doit être fixée proche du niveau de saturation du spectromètre avant l'apparition des premiers effets de suppression ionique et l'élargissement prématuré des pics chromatographiques.

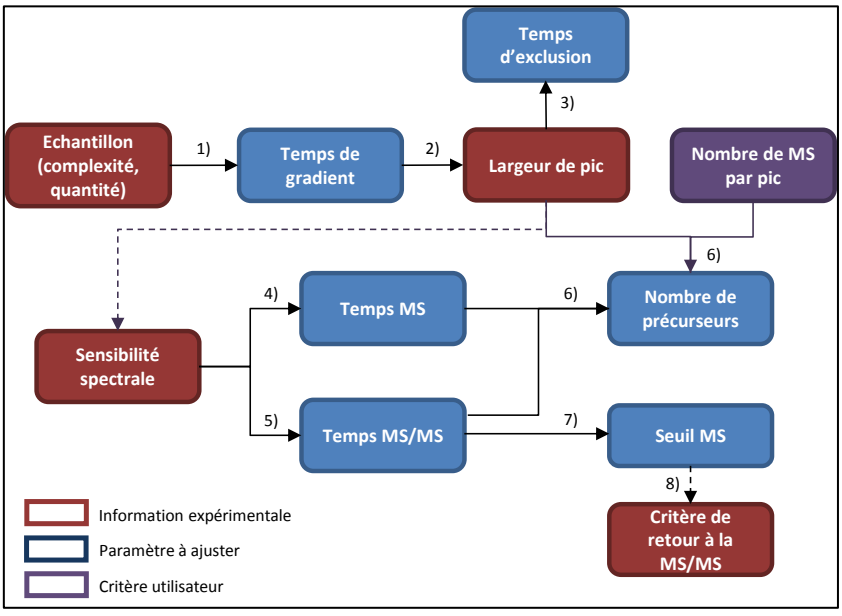


***Figure 36** : Principe de l'optimisation pas à pas du SYNAPT G1 proposée à la suite de l'étude.*

Dans un second temps, la complexité de l'échantillon peut être compensée par l'augmentation du temps de chromatographie en jouant sur le temps de gradient afin de bénéficier de la diminution de la suppression ionique, de l'augmentation de la capacité de pic du système et de l'augmentation du temps disponible pour le séquençage. Cette augmentation de temps ne peut être infinie et est limitée par le plafonnement de

l'augmentation de la capacité de pic du système et la diminution de l'intensité des pics chromatographiques induite par leur élargissement. La redondance dans l'acquisition, consécutive à l'élargissement des pics chromatographiques peut être compensée par le choix d'un temps d'exclusion adapté. La diminution de l'intensité peut être compensée dans une certaine mesure par l'augmentation du temps de sommation MS/MS.

Le temps de sommation MS doit être simplement suffisant pour permettre l'obtention des informations m/z, état de charge et intensité permettant d'initier le processus de sélection. Une augmentation de ce temps en vue de gagner en rapport signal sur bruit n'est pas justifiée. Lorsqu'un temps de sommation MS/MS fixe doit être choisi, il convient de la raccourcir suffisamment afin d'augmenter le nombre de spectres MS/MS réalisés. Le problème d'un temps de sommation fixé est qu'il n'est pas possible de prendre en compte les différences d'abondance des ions afin d'améliorer le rapport signal sur bruit en fonction du besoin. Le compromis entre temps court pour gagner en nombre d'identification et temps long pour compenser un signal faible ne peut être obtenu sans régulation de ce paramètre.

Lorsque les temps de sommation ont été déterminés, le nombre de précurseurs peut être fixé. Il doit être choisi de façon à ne pas réaliser de spectres MS inutiles, c'est-à-dire apportant des informations similaires au spectre MS acquis précédemment. En tenant compte du temps de sommation MS/MS et de la largeur moyenne d'un pic chromatographique au temps de gradient utilisé, il est possible de choisir un nombre d'ions de manière à obtenir des spectres MS à une fréquence cohérente. Il peut être choisi, par exemple, un spectre MS à des intervalles de temps correspondants à une largeur à mi-hauteur du pic chromatographique moyen.

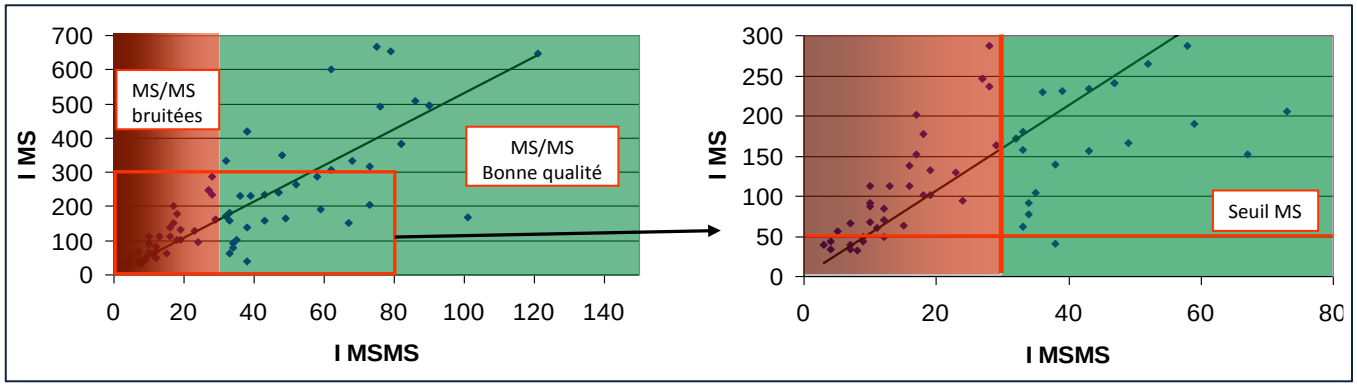


Figure 37 *: Evaluation sur une centaine d'ions de la correspondance entre l'intensité des ions parents et l'intensité des pics de base des spectres MS/MS résultants. En rouge la zone correspondant à des MS/MS bruitées, en vert la zone correspondant à des MS/MS de rapport S/N satisfaisants. Temps de sommation : 0,5 s en MS et 0,7 s en MS/MS. Instrument SYNAPT G1.*

Le seuil MS peut être déterminé après une étude de la correspondance entre l'intensité MS des ions parents au temps de sommation MS choisi et l'intensité des fragments des spectres MS/MS au temps de sommation MS/MS choisi. Cette étude montre qu'une tendance lie expérimentalement les deux intensités. Il est alors possible de fixer un seuil MS correspondant à l'intensité minimale de l'ion parent qui a permis d'avoir expérimentalement un spectre MS/MS avec un rapport signal sur bruit suffisant.

La fixation de ce seuil ne permet cependant pas de contrôler que l'ensemble des composés sélectionnés permettra d'obtenir systématiquement des spectres de bonne qualité. Pour une certaine population de composés, l'augmentation du temps d'accumulation MS/MS est tout de même nécessaire.

j) Bilan

Nous noterons que l'obtention de ces informations et résultats sans plan d'expérience et traitement logiciel aurait nécessité de réaliser manuellement 42 courbes (7 facteurs x 6 réponses). L'étude de l'effet de chaque facteur avec 5 points pour décrire correctement d'éventuels profils de courbes du second degré aurait nécessité 35 essais (7 facteurs x 5 points). Pour connaitre la significativité des courbes obtenues par rapport à l'erreur expérimentale, ces essais auraient dû être répétés, augmentant sensiblement le nombre total d'analyses. Ces résultats auraient été obtenus pour un facteur variable pendant que tous les autres restaient à valeurs constantes. La comparaison des effets des facteurs les uns par rapport aux autres aurait ainsi été beaucoup moins bien appréhendée. La quantification et la comparaison des effets auraient nécessité l'examen manuel des profils des 42 courbes « réponses en fonction des facteurs ». Une inconnue serait demeurée quant à l'existence ou non d'interactions entre les facteurs. La gestion d'une éventuelle dérive de la mesure des réponses en fonction du temps aurait été impossible.

Par comparaison, l'utilisation d'un plan d'expérience avec traitement logiciel des résultats n'a été couteuse en temps qu'au niveau du nombre d'essais réalisés (57 + 9 essais au total soit moins de 2 jours d'acquisition instrumentale). Ce nombre important d'essais est inhérent à la précision de la modélisation de la réponse apportée par la matrice de Doehlert. L'obtention des résultats de mesure des effets est entièrement gérée par logiciel. L'interprétation des résultats est facilitée grâce à la combinaison de l'utilisation des résultats d'analyse de variance dont la visualisation est facilitée par la lecture des diagrammes de Pareto. Chaque réponse peut être visualisée pour chaque couple possible parmi les 7 facteurs.

5) Application à l'optimisation des acquisitions dépendantes des données des Q-TOF SYNAPT G1 et MaXis

a) Constitution d'échantillons standards

L'évaluation des bénéfices d'une optimisation des paramètres d'acquisition sur des échantillons de protéomique nécessite de constituer des échantillons adéquats. Les échantillons ne doivent pas être une source de variation pouvant impacter sur la mesure et doivent donc être strictement identiques pour l'ensemble des méthodes comparées. La comparaison des méthodes doit être appliquée à un nombre d'échantillons suffisant afin de démontrer un bénéfice significatif des conditions opératoires choisies.

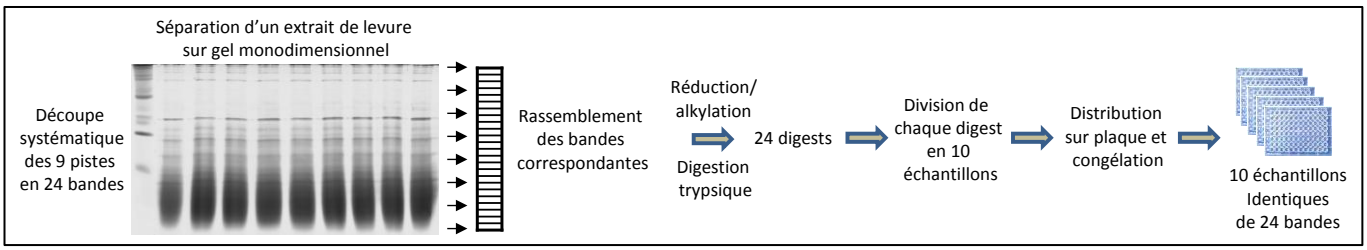


Figure 38 *: Schématisation du protocole employé pour la réalisation des échantillons standards utilisés lors de l'évaluation de l'efficacité des méthodes d'acquisition.*

Nous avons dans ce cadre préparé des séries identiques de 24 échantillons issus du rassemblement des pistes de gel monodimensionnel d'un même extrait de levure. Les échantillons ainsi constitués seront par la suite utilisés pour des comparaisons de méthodes et d'instruments.

b) Comparaison de méthodes d'acquisition sur le SYNAPT G1

Les enseignements tirés de l'étude précédente ont été mis à profil pour le développement de méthodes protéomiques sur le SYNAPT G1. Deux méthodes ont été créées afin d'évaluer l'impact des paramètres d'acquisition sur les résultats d'identification apportés à l'analyse des 24 échantillons standards. Les méthodes ont été comparées à temps de gradient et quantité injectée identiques.

	Méthode d'acquisition 1	Méthode d'acquisition 2
Temps de gradient	30 min	30 min
Nombre d'ions	3	5
Temps d'accumulation MS	1 s	0,5s
Temps d'accumulation MS/MS	2 x 0,7 s	0,7 s / 3x 0,7 s
Seuil MS	30	20
Temps d'exclusion	10 s	30 s

Tableau 4 *: Paramètres d'acquisition utilisés pour les deux méthodes d'acquisition comparées sur le SYNAPT G1.*

Pour la création de ces méthodes, nous avons choisi d'utiliser une fonctionnalité supplémentaire du logiciel d'acquisition. Il est possible de fixer un critère de retour à la MS, qui consiste à partir de la donnée d'intensité mesurée sur le spectre MS/MS d'un composé de choisir de réaliser une acquisition ou plusieurs acquisitions MS/MS supplémentaires. Nous avons évalué sur la base de l'examen d'une centaine de spectres, qu'un spectre MS/MS avait un rapport signal sur bruit correct lorsque l'intensité du pic majoritaire était de l'ordre de 4000 coups/sec. Deux populations de spectres MS/MS peuvent ainsi être considérées en fonction du dépassement ou non de ce seuil. Si ce seuil n'est pas dépassé, nous avons choisi de réaliser deux acquisitions MS/MS supplémentaires pour la méthode 2.

Le fait de réaliser plusieurs acquisitions MS/MS d'un même composé a été utilisé pour réaliser des fragmentations à différentes énergies de collision autour de la valeur d'énergie moyenne déterminée en partie III.3.b. .

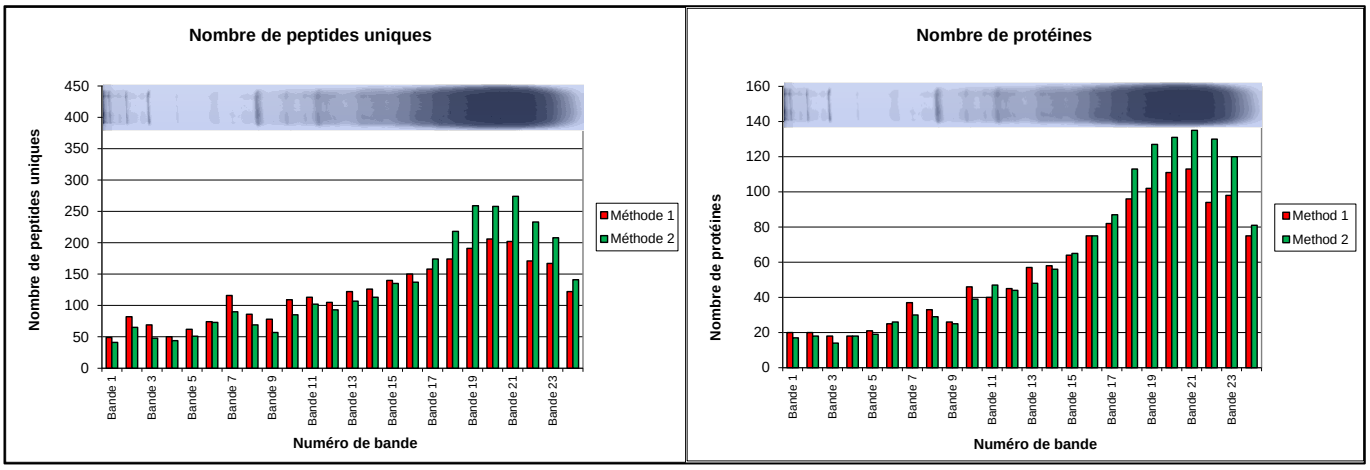


Figure 39 *: Résultats d'identification obtenus pour le nombre de peptides uniques et le nombre de protéines identifiées sur les mêmes échantillons avec les deux méthodes d'acquisition différentes.*

La comparaison des deux méthodes a été effectuée en mesurant le nombre de peptides uniques et le nombre de protéines identifiés dans chaque bande. Les résultats montrent que les différentes bandes du gel ne présentent pas la même complexité. Les bandes (1-9) correspondant aux plus hauts poids moléculaires contiennent entre 20 et 40 protéines identifiées par rapport aux autres bandes (10-24) qui contiennent de 50 à 140 protéines. L'efficacité des deux méthodes n'est donc pas la même en fonction de la complexité de l'échantillon. Il apparait que la méthode 1 identifie davantage de peptides uniques pour les échantillons contenant moins de 60 protéines alors que la seconde méthode permet d'augmenter le nombre d'identification dès lors que le nombre de

protéines est supérieur. Ces différences d'efficacité entre les deux méthodes s'expliquent principalement par la différence dans le temps de sommation MS/MS, facteur vu comme étant le plus influent sur les résultats d'identification. Pour la méthode 2, les échantillons les plus complexes contiennent un nombre important de peptides dont l'intensité des ions générés est suffisante pour déclencher des temps d'accumulation de 0,7 s permettant d'augmenter le nombre de composés identifiés. Dans le cas de la méthode 1, l'acquisition est de 1,4 s dans tous les cas. A partir d'un certain niveau de complexité, ce temps devient trop long, ce qui se traduit par un sous-échantillonnage. Dans le cas des échantillons les moins complexes, le sous-échantillonnage de la méthode 2 par rapport à la méthode 1 est imputable à l'analyse d'un certain nombre de composés avec un temps de sommation de 2,1 s.

Ces résultats montrent que des différences de complexité entre échantillons nécessitent une adaptation de la méthode d'acquisition afin de diminuer les phénomènes de sous-échantillonnage. Ils soulignent la nécessité d'adapter les temps de sommation en fonction de ce paramètre. En pratique, il n'est pas possible sans une analyse préalable de déterminer finement le niveau de complexité d'un échantillon.

Cette problématique montre de nouveau la nécessité de pouvoir adapter finement le temps de sommation MS/MS pour utiliser efficacement le temps d'analyse disponible en évitant au maximum les phénomènes de sous-échantillonnage. Cette possibilité est limitée par les logiciels d'acquisition en place sur le SYNAPT G1.

c) Mise en place d'une stratégie de développement de séquences d'acquisitions dépendantes des données sur le MaXis

Le logiciel d'acquisition utilisé par le MaXis permet d'adapter le temps de sommation MS/MS en fonction de l'intensité MS des précurseurs. Les paramètres de temps de gradient et de temps d'exclusion peuvent être choisis en fonction de la complexité de l'échantillon. Le temps de MS est choisi suffisamment court tout en permettant de pouvoir réaliser correctement la sélection des précurseurs. Le temps de sommation MS/MS n'a plus besoins d'être déterminé. La régulation permet de s'adapter à la quantité injectée et de compenser la diminution de l'intensité des précurseurs avec l'augmentation du temps de gradient. Il reste alors à choisir le seuil de sélection MS au-dessus duquel peut être réalisée la sélection. Le seuil doit être suffisamment haut pour ne pas déclencher la sélection immédiatement au pied du pic chromatographique.

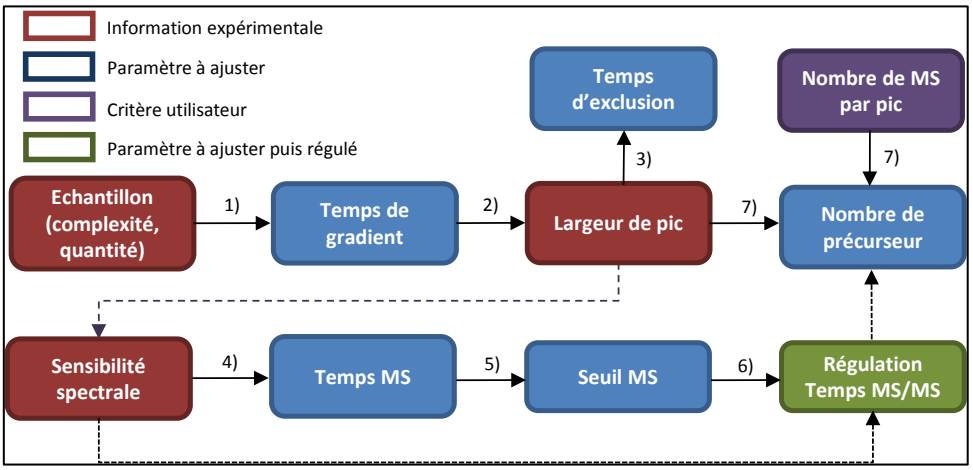


Figure 40 *: Principe de l'optimisation pas à pas du MaXis utilisée dans le cadre de cette étude.*

Le nombre de précurseurs doit toujours être choisi en fonction du nombre de MS devant être réalisées par unité de temps. L'introduction de la régulation du temps MS/MS entraine une variabilité de cet intervalle de temps qui dépend de l'intensité des ions sélectionnés. Plus que le choix d'un nombre de précurseurs, un temps maximum de cycle avant le retour à la MS serait plus approprié. En considérant le temps nécessaire à la fragmentation des ions

d'intensité données vus en MS, il serait possible de calculer le nombre maximum d'ions à sélectionner pour ne pas dépasser un temps de cycle d'acquisition donné et ainsi réguler le nombre de précurseurs.

d) Optimisation d'une option de régulation du temps de sommation MS/MS sur le Q-TOF MaXis

Le logiciel d'acquisition en mode dépendant des données en place sur le MaXis reprend en partie la même logique que celle utilisée sur le SYNAPT G1. La principale différence réside dans l'introduction d'une nouvelle fonctionnalité permettant d'envisager la régulation du temps de sommation MS/MS en fonction de l'intensité du parent sélectionné pour la fragmentation. Nous avons précédemment vu qu'il existait un lien de pseudo proportionnalité entre l'intensité du parent et intensité enregistrée des fragments. En utilisant cette propriété, il est possible d'envisager un temps d'accumulation spécifique à chaque ion sélectionné en fonction de son intensité vue en MS. Bien qu'une méthode « tps MS/MS = f (I MS) » soit fournie à l'installation de l'instrument par le constructeur, nous avons souhaité développer une procédure permettant d'évaluer la corrélation existant entre ces deux paramètres à partir de données expérimentales.

Mise en place d'une procédure d'optimisation de la régulation du temps MS/MS

Nous avons utilisé un digest d'un standard de 4 protéines qui a été injecté en différentes quantités pour permettre de modifier l'intensité du signal des ions parents. Chaque analyse est réalisée à 8 temps de sommation MS/MS différents (de 200 à 1600 ms). Pour chaque résultat d'identification, la mesure du score d'ion est réalisée et l'information de l'intensité de l'ion parent est extraite. Nous obtenons ainsi pour chaque quantité un graphe du type « score d'ion = f (Intensité MS ; temps de sommation MS/MS). L'interprétation des résultats a été réalisée pour la série correspondant à 50 fmol de digest injecté.

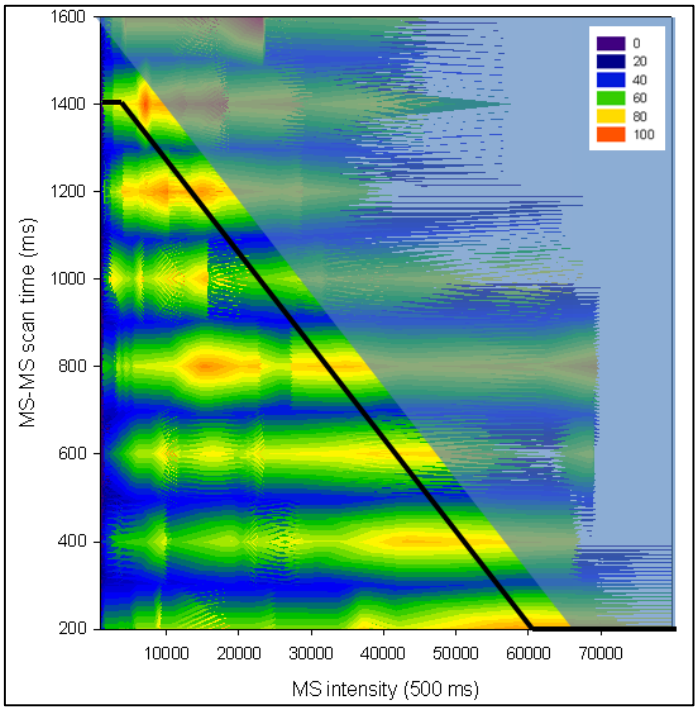


Figure 41 *: Graphe obtenu pour l'analyse d'un digest de 4 protéines (MassPREP Digestion Standard, Waters) à 8 temps de sommation MS/MS différents. Temps de sommation MS 500ms. Paramètres d'acquisition constants. Score d'ion en fonction de l'intensité du parent mesuré en MS et du temps de sommation MS/MS.*

La courbe étudiée montre, pour le temps de sommation le plus court, qu'une population d'ions parents autour de 60000 d'intensité obtient des résultats corrects d'identification. Nous sommes donc partis de ce point pour considérer que l'ensemble des ions parents d'intensités supérieures pouvaient être analysés à ce temps de sommation. Pour les ions d'intensités inférieures, nous avons considéré à chaque temps de sommation utilisé les

zones permettant d'obtenir les meilleurs scores d'ions. Ainsi pour 400 ms de temps MS/MS, les ions d'intensité 50 000 apportent de bons résultats d'identification. A 1200 ms, les ions d'intensité 20 000 apportent de bons scores d'ions. La combinaison de ces résultats nous permet de tracer une droite reliant le temps MS/MS à l'intensité de l'ion parent de manière à obtenir de bons scores d'ions en un minimum de temps.

Le temps de sommation fixé sera ainsi compris entre 200 et 1400 ms pour des intensités MS respectivement de 60 000 et 400 (seuil de sélection). Les temps de sommation entre ces valeurs sont fixés proportionnellement à l'intensité du parent. Les ions sélectionnés au-dessus de 60 000 seront analysés par défaut à 200 ms. Ces valeurs sont plus courtes que celles fixées de manière standard par le constructeur dont les temps considérés sont compris entre 600 et 2000 ms.

Nous nous sommes interrogés sur le fait que l'utilisation de temps de sommation MS/MS longs avait pour conséquence d'avoir plus d'ions considérés d'intensité faible malgré des échantillons analysés identiques. Ce phénomène peut être la conséquence de la diminution du nombre de MS réalisées avec l'augmentation du temps de sommation MS/MS. Nous suggérons que cette diminution peut diminuer la probabilité de sélectionner un composé à son maximum d'intensité chromatographique, les données acquises ayant plus de probabilité d'être sélectionnées en fin de montée ou en descente voire sur la trainée du pic chromatographique.

Evaluation de la méthode de régulation pour l'analyse de bandes de gel 1D

Afin d'évaluer le bénéfice de l'optimisation de la régulation des temps de sommation par rapport à la méthode constructeur, nous avons choisi d'utiliser notre protocole de comparaison de méthodes par utilisation des 24 échantillons standards. Trois méthodes ont été évaluées : les deux premières sont celles proposées par défaut par le constructeur respectivement pour l'analyse dite de mélanges simples et l'autre pour l'analyse de mélanges complexes. Cette dernière utilise la fonction de régulation du temps de sommation MS/MS. La dernière méthode utilise la régulation du temps MS/MS développée d'après notre étude.

	Méthode simple	Méthode complexe	Méthode optimisée
Temps de gradient	29 min	29 min	29 min
Nombre d'ions	3	5	5
Temps de sommation MS	800 ms	500 ms	200 ms
Temps de sommation MS/MS	1600 ms	si I=1000 -> 2000 ms si I=10000 -> 600 ms	si I =400 -> 1400 ms si I=60000 -> 200 ms
Seuil MS	3000	1000	400
Temps d'exclusion	0,8 min	0,33 min	0,8 min

***Tableau 5** : Paramètres d'acquisition utilisés pour les deux méthodes d'acquisition constructeurs comparées à celle optimisée dans le cadre de l'étude sur le MaXis.*

La comparaison montre tout d'abord que la méthode « constructeur complexe » apporte systématiquement des résultats inférieurs en termes de peptides uniques à ceux des deux autres méthodes. La différence significative entre les deux méthodes utilisant la régulation des temps de MS/MS montre la nécessité de régler expérimentalement les paramètres de régulation du temps de sommation. La méthode optimisée apporte 25% de peptides uniques supplémentaires par rapport à la méthode « constructeur complexe ».

La méthode « constructeur simple » donne dans l'ensemble des résultats voisins voire un peu meilleur pour les bandes présentant moins de 50 protéines par bande. Au-delà, la méthode optimisée est systématiquement meilleure. Le fait que la méthode optimisée ne s'adapte pas aussi bien que la méthode simple à l'analyse d'échantillons moins complexes est imputable au nombre de précurseurs trop important par rapport à la

complexité de l'échantillon. Cet inconvénient pourrait probablement être compensé par une régulation du nombre de précurseurs.

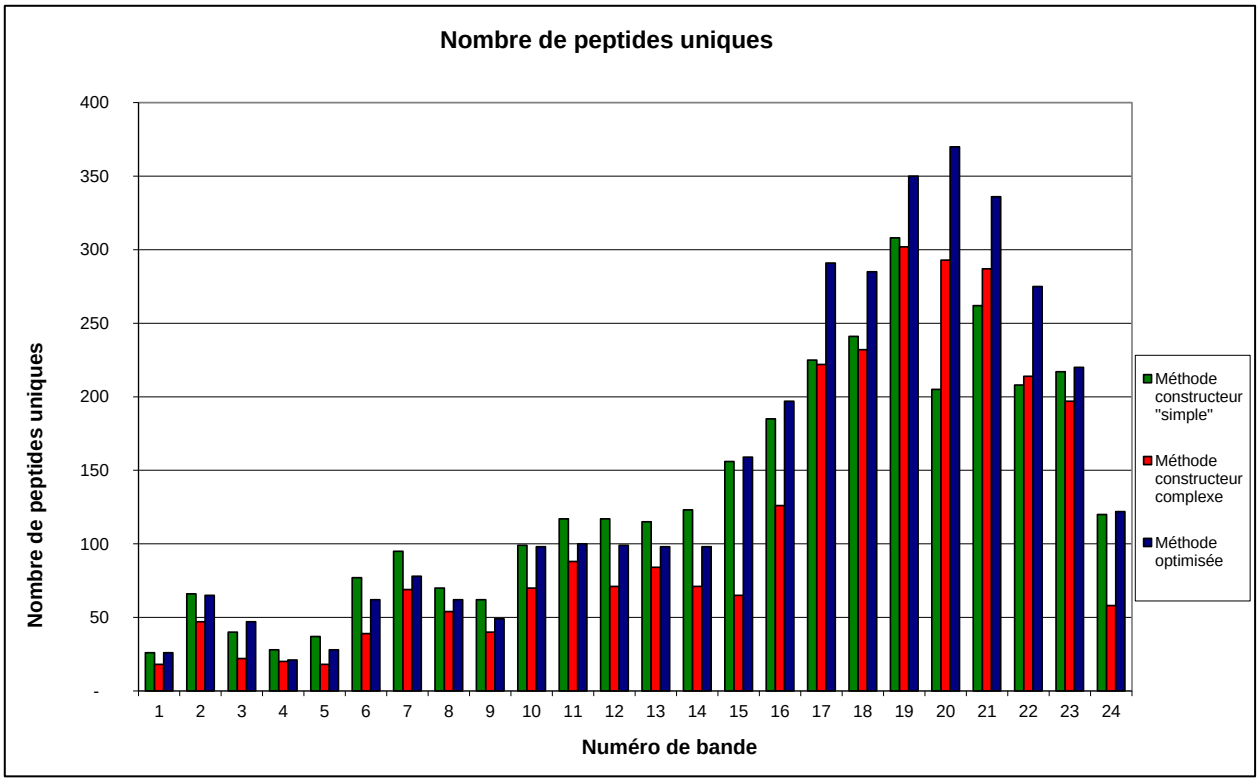


***Figure 42** : Comparaison des méthodes d'acquisition fournies par le constructeur et celle optimisée dans le cadre de notre étude. Les paramètres d'acquisition correspondent à ceux indiqués dans le tableau 5. Extrait de levure décomplexifié en 24 bandes de gel monodimensionnel. Le nombre de peptides uniques indiqué est le nombre de peptides uniques pour une analyse de la bande considérée.*

Nous noterons que par comparaison aux résultats obtenus pour les mêmes échantillons sur le SYNAPT G1, ce dernier apporte plus d'identification sur les échantillons de faible complexité que le MaXis. Cette comparaison suggère que le SYNAPT permettrait d'identifier des peptides moins abondants. En revanche les résultats du MaXis montrent la capacité de l'instrument à réaliser plus d'identifications dès lors que la complexité augmente et cela quelque soit la méthode d'acquisition choisie.

Evaluation de la méthode de régulation pour l'analyse de digests totaux

La méthode optimisée a montré son efficacité pour l'analyse des échantillons les plus complexes. Nous avons souhaité évaluer les performances des méthodes pour l'analyse du même échantillon d'extrait de levure mais cette fois ci sans utiliser la décomplexification par gel 1D. Les extraits totaux de levures ont ainsi été digérés puis analysés avec les trois méthodes précédemment décrites. Les analyses ont été réalisées à trois temps de gradient différents pour comparer l'impact de la méthode utilisée à celui de la décomplexification par chromatographie.

Les résultats de cette étude montrent l'impact du choix de la méthode. La méthode optimisée permet quel que soit le temps de gradient utilisé d'obtenir entre 90 et 130 % d'identifications de peptides supplémentaires par rapport à la méthode « constructeur simple » et entre 60 et 80 % par rapport à la méthode « constructeur complexe ». Si l'augmentation du temps de gradient permet de diminuer l'impact des différences d'échantillonnage entre les méthodes d'acquisition, l'écart n'en reste pas moins très significatif.

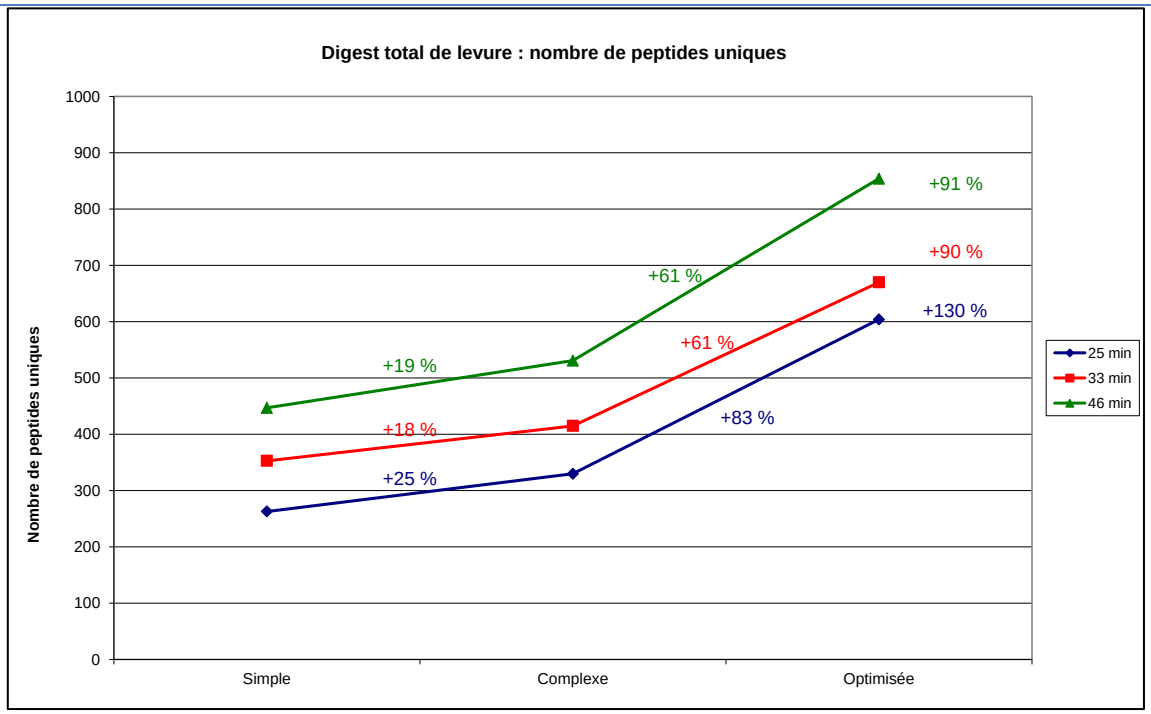


Figure 43 *: Comparaison des méthodes d'acquisition fournies par le constructeur et de celle optimisée dans le cadre de notre étude. Extrait total de levure analysé pour trois temps de gradient différents.*

Nous montrons ici que l'optimisation de la méthode d'acquisition peut compenser dans une certaine mesure l'absence de décomplexification en amont d'un échantillon complexe. Nous noterons néanmoins que les résultats d'identification obtenus avec la méthode optimisée en une analyse restent inférieurs en termes de peptides uniques et de protéines identifiés à ceux obtenus après décomplexification par gel monodimensionnel. L'analyse du digest total avec le temps de gradient de 46 minutes et la méthode optimisée a permis l'identification de 245 protéines contre 320 par analyse des bandes de gel 1D avec la méthode optimisée et un temps de gradient de 29 minutes.

e) La répétition pour l'augmentation des résultats d'identification

Principe de la répétition des analyses pour accroitre le nombre d'identification

La comparaison des résultats d'identification obtenus entre le digest d'extrait de levure total et l'analyse des bandes de gel 1D du même échantillon montrent qu'un certain nombre de protéines ne sont pas identifiées dans l'analyse directe de l'échantillon. L'amélioration apportée par l'augmentation du temps de gradient suggère qu'il existe toujours un phénomène de sous échantillonnage des ions lors de l'analyse de ce mélange.

Le sous échantillonnage observé lors de l'analyse d'échantillon de complexité importante peut être compensé par la répétition de l'analyse du même échantillon dans les mêmes conditions. Compte tenu de l'aspect pseudo aléatoire de la sélection en mode dépendant des données, il existe une probabilité de sélectionner des composés dont la sélection n'a pas été réalisée lors de la première acquisition.

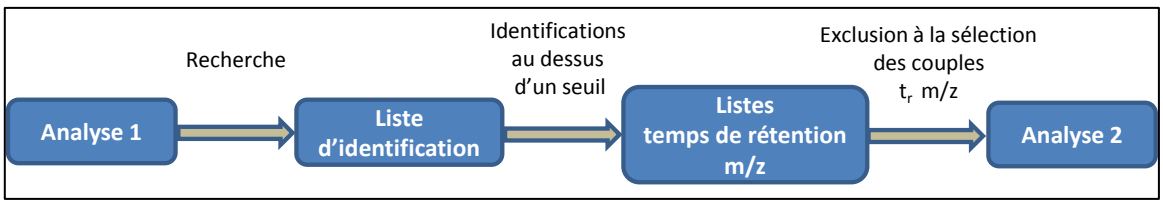


Figure 44 *: Principe de répétition d'analyse utilisant une liste d'exclusion.*

Le logiciel d'acquisition Bruker offre la possibilité à partir des composés identifiés lors d'une première analyse de réaliser une liste permettant d'exclure de la sélection les composés déjà identifiés (liste d'exclusion ou Schedule Precursor List : SPL). La seconde acquisition peut ainsi être réalisée en tenant compte des identifications déjà réalisées et cibler les ions n'ayant pas fait l'objet d'une sélection ou d'une identification.

Evaluation du bénéfice de la répétition de l'analyse avec et sans liste d'exclusion sur un mélange complexe.

Nous avons réalisé la comparaison des résultats pouvant être obtenus lors de la répétition de l'extrait total de levure. Six répétitions ont été réalisées à trois temps de gradient différents afin d'évaluer le bénéfice de la répétition par rapport à ce facteur. Chaque répétition a été réalisée avec et sans liste d'exclusion.

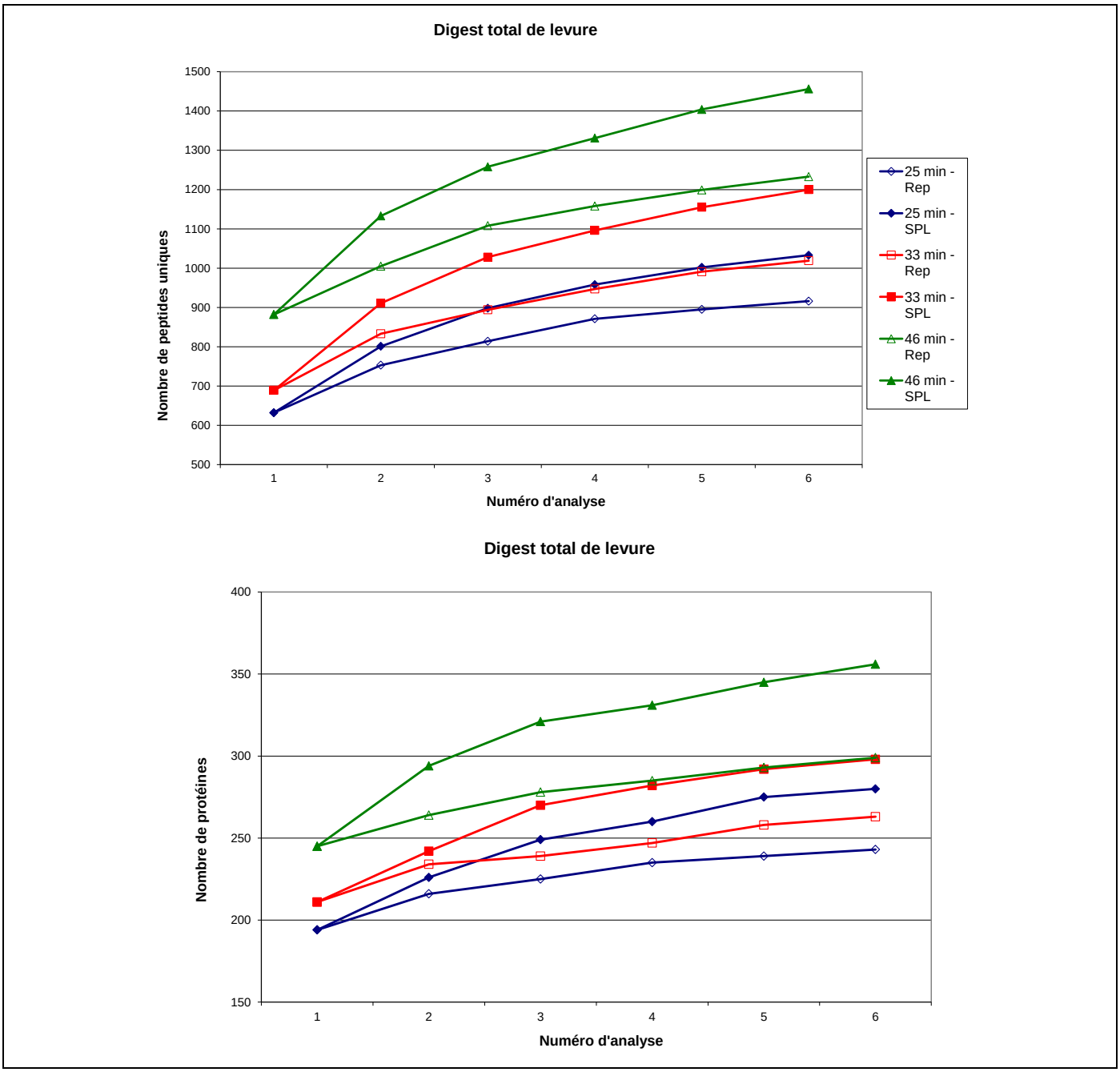


Figure 45 *: Courbes d'accumulation des résultats d'identification en nombre de peptides uniques (en haut) et en protéines (en bas) obtenus par répétition avec (SPL) et sans liste d'exclusion (Rep). Extrait total de levure analysé pour trois temps de gradient différents. Pour les analyses réalisées avec liste d'exclusion, cette dernière est constituée de l'ensemble des identifications obtenues avec un score d'ion supérieur à 30 lors des analyses précédentes de la même série.*

La comparaison montre que l'utilisation d'une liste d'exclusion permet d'apporter un bénéfice en termes d'identifications de peptides uniques et de protéines par rapport à une simple répétition quelque soit le temps de gradient utilisé. La différence entre l'utilisation ou non d'une liste d'exclusion est d'autant plus importante que le temps de gradient est élevé. De l'ordre de 10 % de peptides uniques et de protéines supplémentaires sont ainsi obtenus pour un nombre de répétitions équivalent. Pour une répétition avec liste d'exclusion, le nombre d'identifications est légèrement supérieur à ce qui peut être obtenu avec deux répétitions sans liste d'exclusion. Nous notons également qu'une répétition permet d'obtenir à peu près autant d'identification qu'un allongement du gradient d'une dizaine de minutes. Le gain d'identification est le plus important à la suite de la première répétition. Ce gain diminue avec le nombre de répétitions mais est toujours plus important avec l'utilisation d'une liste d'exclusion.

Le fait que les courbes d'accumulation de peptides uniques continuent à augmenter après cinq répétitions montre que le phénomène de sous-échantillonnage est toujours présent. L'utilisation d'une liste d'exclusion favorise donc l'efficacité de la sélection principalement lors des premières répétitions d'analyse mais ne permet pas de conduire à une saturation en identifications. Les courbes d'accumulation des protéines identifiées avec le nombre de répétitions continuant d'augmenter, nous pouvons suggérer que les nouveaux peptides identifiés dans cet échantillon sont pour la plupart issus de protéines non identifiées dans les précédentes répétitions. L'hypothèse que les signaux de ces peptides soient peu abondants peut être émise. Il y aurait donc un intérêt à réaliser des répétitions afin d'augmenter la probabilité d'identification de peptides correspondants à des protéines minoritaires.

Nous noterons que les résultats d'identification en termes de protéines identifiées sont plus importants avec une analyse répétée 5 fois avec liste d'exclusion et temps de gradient de 46 minutes que ceux obtenus lors de l'analyse de l'ensemble des bandes de gel 1D avec un gradient de 29 minutes (356 contre 320). Cette différence doit être pondérée par le fait qu'expérimentalement la quantité injectée de digest total est relativement plus importante que celle des extraits peptidiques de gel 1D obtenus (données non présentées).

Evaluation du bénéfice de la répétition de l'analyse avec et sans liste d'exclusion pour l'analyse de bandes de gel 1D.

La répétition des analyses a été réalisée sur nos 24 échantillons standards de gel monodimensionnel d'extrait de levure. Ces expériences ont été réalisées afin d'évaluer l'effet de la répétition pour l'analyse d'échantillons moins complexes.

Comme attendu, les résultats de cette étude montrent que la répétition apporte un bénéfice systématique. Le bénéfice en termes de nombre de peptides uniques identifiés supplémentaires augmente avec la complexité de l'échantillon. L'apport de la liste d'exclusion par rapport à une simple répétition est moins important que ce qui a pu être décrit dans le paragraphe précédent. Pour certaines bandes les moins complexes, l'analyse répétée apporte même un peu plus d'identification que la répétition avec liste d'exclusion. Au total 360 protéines ont été identifiées par les analyses répétées.

En conclusion, l'utilisation de la répétition peut être envisagée sur l'instrument dès qu'une complexité importante est attendue. L'augmentation de la quantité de matériel devrait favoriser les bénéfices de la répétition. L'utilisation des listes d'exclusion présente un intérêt pour les échantillons très complexes et est d'autant plus efficace que le gradient utilisé est long et que le nombre de répétitions est important.

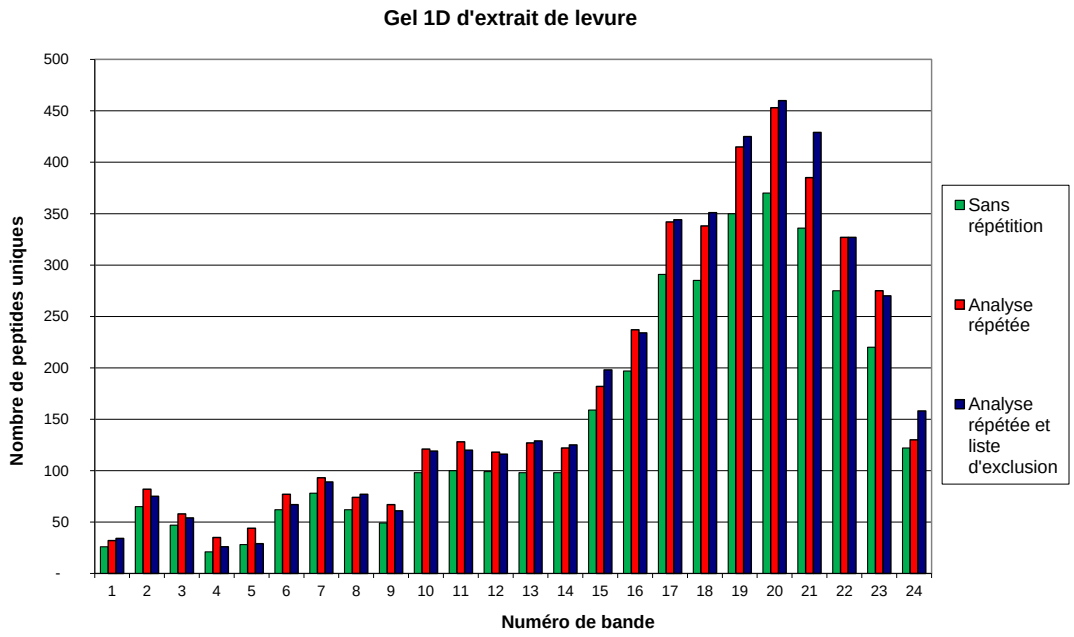


Figure 46 *: Evaluation de l'impact de la répétition de l'analyse avec et sans liste d'exclusion. Résultats d'identification obtenus en termes de nombre de peptides uniques sur les mêmes échantillons avec la même méthode d'acquisition.*

f) Constitution de séquences d'acquisition dépendantes des données du MaXis

En tenant compte de l'ensemble des études et optimisations réalisées, nous avons développé sur le MaXis différentes méthodes d'acquisition standard disponibles pour les utilisateurs en fonction du type d'échantillon analysé. Les échantillons considérés peuvent être :

- des spots de gel bidimensionnels de faible complexité ne nécessitant pas une décomplexification poussée mais provenant la plupart du temps de grandes séries de découpe. Ces échantillons peuvent être analysés avec des gradients rapides permettant d'augmenter le nombre d'analyse par heure. Le raccourcissement du gradient a pour effet la diminution des largeurs de pic et une augmentation de l'intensité qui doivent être considérés pour le temps d'exclusion et le nombre de précurseurs.
- des spots de gel monodimensionnels complexes nécessitant au contraire une décomplexification par augmentation des capacités de séparation chromatographiques et de temps disponible au séquençage.
- des spots de gel monodimensionnels peu complexes.

En considérant les niveaux de complexité attendus, les temps de gradient ont été adaptés. La connaissance des largeurs moyennes des pics chromatographiques permet de fixer le temps d'exclusion pour chaque méthode à environ six fois la largeur à mi-hauteur des pics les plus larges. Le nombre d'ions sélectionnés est choisi de façon à avoir au moins un cycle d'acquisition par largeur à mi-hauteur de pic chromatographique bien que la possibilité de fixer une période maximum de cycle d'acquisition aurait été plus appropriée.

Echantillon	Gradient	Temps total	Largeur à mi hauteur	Temps d'exclusion	Nombre d'ions
Gel 2D	4 min	15 min	0,06 ± 0.015 min	0,3 min	2
Gel 1D simple	9 min	20 min	0,07 ± 0.025 min	0,4 min	3
Gel 1D	21 min	32 min	0,10 ± 0.055 min	0,6 min	4
Gel 1D complexe	29 min	40 min	0,12 ± 0.075 min	0,8 min	5

Tableau 6 *: Récapitulatif des principaux paramètres d'acquisition de méthodes nanoLC-MS adaptés en fonction du type et de la complexité de l'échantillon analysé.*

Ces séquences d'acquisition sont désormais appliquées à l'ensemble des analyses protéomiques réalisées sur le MaXis. L'utilisation de gradients plus longs combinée à des listes d'exclusion est également proposée lorsque la

décomplexification n'est pas suffisante ou dans le cadre de projets nécessitant la caractérisation d'échantillons complexes.

Conclusion

La stratégie protéomique définit un processus analytique dont les étapes sont bien définies. Néanmoins, la problématique de l'analyse des isoformes du cheveu nécessite d'adapter certaines de ces étapes et de s'affranchir de l'étape de séparation des protéines notamment par électrophorèse sur gel. L'analyse des digests des extraits de cheveu apparait dans ce contexte comme une bonne alternative. Les possibilités de décomplexification des mélanges peptidiques par plusieurs dimensions de chromatographie ainsi que l'optimisation des couplages nanoLC-MS pour l'efficacité du séquençage permettent d'envisager une identification exhaustive des peptides présents dans l'extrait biologique. L'analyse des digests d'extraits corticaux et cuticulaires va pouvoir sur les bases de l'utilisation de cette stratégie être réalisée.

Partie III Applications des technologies protéomiques à l'étude du protéome du cheveu

Le but principal de cette étude réside dans l'établissement d'une catalogue complet des protéines exprimées dans les cellules du cortex et de la cuticule de la fibre capillaire tout en cherchant à connaitre leurs éventuelles modifications.

Dans cette partie nous exposerons les résultats qui peuvent être obtenus en utilisant principalement la méthodologie d'analyse protéomique développée précédemment. Les protéomes du cortex et de la cuticule seront investigués afin d'établir la nature des protéines de structures qui les composent. L'évidence de l'expression protéique d'un grand nombre d'isoformes de KAP sera démontrée. Nous compléterons ces résultats d'étude de caractérisation des protéines identifiées par la recherche de leurs modifications chimiques et post traductionnelles. Nous montrerons comment des informations semi quantitatives et quantitatives peuvent être utilisées dans le cadre de recherche d'expression ou de modifications différentielles des protéines.

Nous compléterons l'étude des protéines de structure du cheveu par une comparaison avec un protéome de composition proche, le protéome de la plaque unguéale.

Chapitre I Le Protéome des cellules corticales humaines

Avant propos

La stratégie d'analyse développée et exposée en deuxième partie a été appliquée à l'analyse des extraits corticaux de cheveu humain dans le but de répondre à la problématique de caractérisation de ce protéome. Ces travaux ont conduit à la rédaction d'un article soumis au journal Molecular and Cellular Proteomics que nous incorporons à ce manuscrit comme élément de ce premier chapitre consacré à l'étude du protéome du cortex humain. Il présente l'ensemble des identifications protéiques obtenues grâce à la stratégie de combinaison de l'utilisation de plusieurs enzymes, de la chromatographie bidimensionnelle en phase inverse et de traitement séquentiel des données de spectrométrie de masse obtenues sur le Q-TOF SYNAPT G1.

La répétition des analyses effectuées sur le digest trypsique a été utilisée, d'une part pour apporter plus d'identification, mais également pour disposer d'un nombre significatif de données spectrales pour obtenir une notion semi quantitative par comptage de spectre de l'abondance des protéines identifiées. Ces travaux succèdent à ceux de Lee et al. publiés en 2006 qui exposaient principalement l'analyse protéomique d'un extrait insoluble obtenu après extraction successive du cheveu [89]. Cet extrait a été décrit par les auteurs comme un complexe insoluble contenant les protéines substrats de la transglutaminase. Cette méthode d'extraction, qui nous semble loin d'être spécifique, pose la question de l'origine exacte de ces extraits protéiques qui pourraient correspondre simultanément à de la cuticule, de la médulla et des éléments insolubles du cortex.

Dans ce cadre, nous avons souhaité confronter nos résultats d'identification obtenus grâce à une technique d'extraction plus spécifique du cortex. Les données d'identification issues de l'étude comparée ont été mises à disposition par les auteurs à la communauté. En partant de ces données, les informations relatives aux nombres de spectres obtenus pour les différents peptides uniques identifiés dans cette étude ont été extraites. La comparaison des abondances relatives des kératines et des KAP mesurées par comptage de spectre dans les deux études a ainsi été réalisée.

En utilisant les données d'expression protéiques et transcriptomiques réalisées sur le cortex et la cuticule pour les kératines et les KAP et décrites en première partie de ce manuscrit, nous avons établi la composition en cuticule, en cortex et en médulla des deux échantillons. Ces résultats montrent que les protéines identifiées dans notre échantillon sont très majoritairement issues du cortex.

Nous montrons dans cette étude comment notre stratégie analytique a permis de caractériser les protéines de structures du cortex. Un des résultats clé de ce travail est la démonstration de l'expression d'un nombre important de gène de KAP. La difficulté de la multigénicité de ces protéines pour leur identification a été contournée grâce à l'obtention de couverture de séquence quasiment totale pour un certain nombre des individus de ces familles.

Un second résultat réside dans la suggestion, sur la base des abondances spectrales enregistrées par comptage, de l'expression importante des KAP des familles 4, 9, 3, 2 et 1 dans la structure intermicrofibrillaire.

L'abondance de données spectrales nous a également permis de réaliser une étude de caractérisation des protéines majoritairement identifiées.

A Designed Proteomic Strategy to Characterize Isoforms from Multigenic Families in Human Hair Cortical Cells

Nicolas R. Barthélemy[1], Christine Schaeffer-Reiss[1], Christine Carapito[1], Alexandre Burel[1], Patrick Guterl[1], Dominique Jullien[2], Alain Van Dorsselaer[1] and Nükhet Cavusoglu[2]

1: Université de Strasbourg, IPHC, CNRS, UMR7178, 25 rue Becquerel 67087 Strasbourg, France
2: L'Oréal Research and Innovation, Aulnay-sous-Bois, France.

The human hair cortex proteome was investigated to focus on keratin intermediate filament and keratin associated protein isoform identification. The strategy has required (i) multiple enzymatic digestion, (ii) high resolution two dimensional chromatography with high-pH/low-pH peptide separation, (iii) accurate MS and MS/MS measurements and (iv) sequential searches with two search algorithms. This approach has led to the identification of 123 proteins among which 40 KAP isoform proteins with high sequence coverage and allowed sequence refinements and studies on their modifications and natural variants. Spectral abundance of over 60 hair proteins was used to estimate the hair shaft protein composition. Intermediate filament and matrix protein distribution was in a good agreement with previous cortical macrofibril visualization with microscopic techniques. Abundance suggests higher expression of some KAP families which might have an important contribution on hair matrix structure. Results were compared to previous transcriptomic and proteomic works in human hair to establish expression specificity between cortex and cuticle for identified proteins.

1. Introduction

Hair fibre is a common structure in Mammalia species which has contributed to their colonization even in the harsh places of Earth biotopes. Hair have multiple functions according to differentiation such as coat to protect organism against coldness and wetness in air or water, eyelashes to avoid dust deposition in eyes, spines to dissuade predation, vibrissae as sensitive detector to interact with environment.

In human, epidermal hairiness is characterized by short length which helps heat regulation in sweating process and allows better resistance to endurance effort in warm environment than other animals [5, 6]. This characteristic has probably been a critical advantage during Homo sapiens evolution in East Africa before using animal hair and coat to conquer colder place in the world. Nevertheless, longer head hair have subsisted with many phenotypes and today human head of hair is a symbol of beauty and a subject of fashion [117].

Hair fibre is divided into three components composed of dead cells generated after several phases of differentiation into the follicle: (i) an outer protective cuticle layer made of a stack of plate-shaped cuticle cells, (ii) an inner cortex and (iii) a central hair medulla, a column of large and randomly staggered cells which may be discontinuous and is not found in all hair types [101]. The cortex represents the large part of hair fibre and consists of spindle-shaped cells separated by the cell membrane complex between outer cellular membranes where melanin granules are inserted. Each cortical cell contains a dozen of macrofibrils separated by a minor intermacrofibrillar matrix. Macrofibrils are formed of hundreds of keratin intermediate filaments (KIFs) called microfibrils embedded in intermicrofibrillar matrix. Shaft size, number of cuticle cells in the cuticle layer, medulla cell abundance, ratio between microfibrils and intermicrofibrillar matrix vary according to hair functions and species [108]. Along these variations, cortical cells can be divided into ortho- and para-cortical cells whose microfibrillar packing is different [107, 114]. Abundance and localization of these cells into the fibre can also vary.

Human proteins of keratin intermediate filaments and intermicrofibrillar matrix are the main constituents of hair cortex. Keratin intermediate filament proteins (IFPs) are divided into two families: type I with 11 members (K31-

K40) and type II with 6 members (K81-K86) [3, 82, 83, 142]. Although these proteins are described as hair keratins, not all are cortical KIFs. Several are described as specifically expressed in cuticle (K82, K32 and K40) and even in cytoskeletal of tongue cells (K84) [82, 83, 142]. Type I and type II keratins are associated as heterodimers whose lateral arrangement of 16 units forms the microfibril width which lengthways grow [85, 86, 133]. Other proteins are described as keratin associated proteins (KAP), divided in three groups based on main amino acid composition [73] : (i) the ultra high sulphur (UHS) KAP (>30% cysteine content) including 5 families (4, 5, 9, 17 and 28), (ii) the high sulphur (HS) KAP (<30% cysteine content) consisting of 12 families (1, 2, 3, 10, 11, 12, 13, 24, 25, 26, 27, 29) and (iii) the high glycine-tyrosine (HGT) KAP including 6 families (6, 7, 8, 19, 20 and 21). These proteins are small in size (typically from 6 to 25 kDa) and are described as putative actors in the linkage between KIFs. KAP probably make the main part of intermicrofibrillar matrix but their molecular structures and interactions including disulphide linkages are unknown although certainly essential to hair rigidity and flexibility.

Most of these hair keratins and KAP families belong to multigenic families leading to expression of isoform proteins with high sequence similarities. These isoforms are the expression of paralog genes derived from the same, several times duplicated, ancestral gene whose copies have independently evolved and kept similar function [122].

Many studies have been performed on KAP at both genomic and transcriptional levels. To date, about 100 human KAP genes have been described and *in situ* hybridization in hair follicle has allowed mRNA localization in cortex and cuticle of the majority of KAP families [73, 77, 78, 80, 81, 139, 140]. Nevertheless, few studies have been performed at the protein level and evidence of KAP translation into protein remain challenging owing to difficulties to generate KAP-specific antibodies due to their small size and amino acid composition [75]. In this context, analysis with proteomic tools to establish the existence of multigenic KAP proteins in mature hair may appear to be specifically appropriate to this issue [245]. However, classical proteomic analysis is confronted with the extensive sequence homology within as well as between KAP families, paucity of tryptic cleavage sites and low solubility of these proteins [92].

To address these technical shortcomings in the proteomic characterization of hair proteins, we have developed a specific proteomic approach. Using a refined LC-MS/MS-based Shotgun analysis, we established an exhaustive keratin and KAP composition of human hair extracts and compared with predicted sequences from genomic data.

Characterization with high sequence coverage could also detect potential database error or sequence ambiguity to get reliable specific peptide sequences for further quantification on human hair proteome. This step is essential for future investigation on human hair to understand potential correlation between protein expression and hair characteristics in dermatologic and cosmetic contexts.

2. Experimental Procedures

a) Cortical protein extraction

Melted hair (100mg) from three Caucasian individuals was delipidated by soaking fibers in ethanol followed by cyclohexane. Cortical proteins were extracted in a solution containing 7 M urea, 2 M thiourea, 50 mM Tris-HCl, 50mM DL-dithiothreitol (DTT) and 0.1% Triton X100 for 18 hours at 37°C. The protein extract was separated from the insoluble cuticle, collected *via* filtration and then alkylated with a solution of 1 M iodoacetamide and 3 M Tris-HCl at pH 8.4 for 10 minutes in the dark at room temperature. The solution was dialyzed in 3,500 MWCO dialysis cassettes (Pierce, Rockford, USA) against water over a period of 48 hours. The solution was then lyophilized and the freeze-dried sample was stored in a -80°C freezer.

b) Digestions

For each experiment, 1mg of extract was reduced for 1 hour at 60°C in 1 mL of 20 M D,L-dithiothreitol (DTT) and 25 mM ammonium hydrogen carbonate (NH_4HCO_3) then alkylated by adding solid iodoacetamide mixed by gentle vortex to 40mM and incubated at room

temperature for 1 hour in the dark. Proteins were then precipitated by adding 2.5 volumes of ethanol and rinsed twice with 70% ethanol. Finally, proteins were resuspended in digestion buffer in 100 mM NH_4HCO_3 and 2 M urea for trypsin or chymotrypsin digestion and in 25 mM potassium hydrogen phosphate (K_2HPO_4) and 1 M urea for GluC digestion. Modified porcine trypsin (Promega, Madison, WI, USA), bovine chymotrypsin and staphylococcus aureus V8 endoproteinase GluC (Roche, Mannheim, Germany) were added to the corresponding buffer to give a protein/enzyme ratio of 1/100, 1/80 and 1/40 by mass, respectively. Digests were incubated for 15 h at 37°C for trypsin, 9h at room temperature for chymotrypsin and 15 h at 25°C for GluC. Trypsin digestion was performed in triplicate. Reactions were quenched by adding formic acid to pH 3. Digests were loaded on 1 mL (100mg) tC18 SepPak cartridges (Waters, Mildford, MA, USA), desalted and lipid and melanin content removed before eluting with 60% acetonitrile 0.1% formic acid then concentrating by vacuum centrifugation.

c) First dimension: High-pH reverse phase HPLC

Digested samples were injected on a Waters 625 LC System (Waters) using a XTerra C18 column (150 mm x 4.6 mm, 5 µm i.d., 125 Å pore size) (Waters). Sample loading was performed at 1 mL/min with (A) 72 mmol /L triethylamine titrated to pH 10.0 with acetic acid. Elution was performed with (B) 72 mmol/L triethylamine, 65 mmol/L acetic acid. The gradient was 2-55% B in 45 min followed by isocratic conditions at 100% B for 2 min. 28 fractions were collected every two minutes, concentrated by vacuum centrifugation and acidified to pH 3 with formic acid.

d) Second dimension: Low-pH reversed phase nanoLC-MS/MS analysis

Fractions of each digest were analyzed by nanoLC-ESI-MS/MS using a nanoACQUITY UPLC (Waters, Milford, MA). The samples were trapped on a 20 x 0.18 mm, 5 µm Symmetry C18 precolumn (Waters, Milford, MA), and the peptides were separated on ACQUITY UPLC® BEH130 C18 column (Waters, Milford, MA), 75 µm x 200 mm, 1.7 µm particle size. The solvent system consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile. Trapping was performed for 3 min at 5 µL/min with 99% A and 1% B. Elution was performed at a flow rate of 400 nL/min, using 1-40% B over 35 min at 45°C followed by 65% B over 5 min. The MS and MS/MS analyzes were performed on a SYNAPT hybrid quadrupole orthogonal acceleration time-of-flight tandem mass spectrometer (Waters, Milford, MA), equipped with a Z-spray ion source and a lock mass system in the positive ion mode. The capillary voltage was set at 3.5 kV and the cone voltage at 35 V. Mass calibration of the TOF was achieved using phosphoric acid (H_3PO_4) on the [50-2000] m/z range. Online correction of this calibration was performed with a lock-mass. For tandem MS experiments, the system was operated with automatic switching between MS and MS/MS modes (MS 0.5 s/scan on m/z range [250;1500] and MS/MS 0.7 s/scan on m/z range [50;2000]). The three most abundant ions (intensity threshold 60 counts/s), preferably doubly and triply charged ions, were selected on each MS spectrum for further isolation and CID fragmentation with two energies set using collision energy profile, then were excluded for 10 seconds. Fragmentation of each ion was performed during two scans or was stopped when intensity fell below 15 counts/s. Fragmentation was performed using argon as the collision gas. The complete system was fully controlled by MassLynx 4.1 (SCN 566, Waters, Milford, MA). Raw data collected during nanoLC-MS/MS analyses were processed and converted with ProteinLynx Browser 2.3 (Waters, Milford, MA) into .pkl peak list format. Normal background subtraction type was used for both MS and MS/MS with 5% threshold and polynomial correction of order 5, and deisotoping were performed.

e) Data analysis

Mass data collected during LC-MS/MS analyses were processed using the software tool ProteinLynx Global Server (version 2.3, Waters, Milford, MA) converted into pkl files. All files corresponding to one digest were combined into a single mgf file using homemade software adapted from merge.pl (www.matrixscience.com). The three trypsin digests results were merged together. Mass data collected during nanoLC-MS/MS analyses were searched using MASCOT 2.2.0 (Matrix Science, London, UK) and OMSSA (NCBI) algorithms. Several steps of searches were performed for each digest to first extract fully specific peptides then semi-specific peptides in second step and modified peptides for the 2 last steps. All spectra were searched against a target-decoy version of Swissprot database restricted to Homo sapiens (v 57.13, January 28, 2010, 40556 entries) with a mass tolerance of 15 ppm for MS and 0.05 Da for MS/MS data. For all steps, Carbamidomethylation on Cys, N-acetylation at protein N-termini and oxidation on Met were specified as variable modifications. Enzyme specificities rules were set to C-terminal of the Lys and Arg allowing two missed cleavages (without Pro rule) for trypsin, C-terminal of Phe, Trp, Tyr and Leu allowing three missed cleavages for chymotrypsin and C-terminal of Glu and Asp allowing three missed cleavages for GluC. Results from the two search algorithms were combined using Scaffold software (Proteome Software, Portland, OR, USA). Criteria used for validation of peptide identifications are described (supplementary data I) and were chosen according the estimation of the false discovery rate on protein identifications. All spectra that did not satisfy criteria in the first step were exported in a new .mgf file using Scaffold and submitted to the next search step as previously described. Proteins were assigned when at least one unique peptide was detected above search thresholds and sequence coverages were calculated using all validated peptides from the specific and semi specific searches. After the two first search steps with current modifications and specific and semispecific enzyme, other modifications were investigated. First, unmatched spectra from previous search steps were searched adding expected modifications such as deamidation on glutamine and asparagine and cysteine trioxidation described as common alterations in hair sample. Second, a Mascot error tolerant search [246] was performed to evaluate unsuspected modifications on previously identified proteins on unmatched spectra including biological and

chemical modifications and variants with amino acid substitutions. No enzyme was specified during these searches. Accurate mass measurement on MS and MS/MS performed on Q-TOF allowed to cut down candidates among the large number of alterations tested. During the fourth step, other chemical modifications described in the results part were searched. Unmatched spectra after the fourth step were searched with Mascot algorithm against a Varsplic version of Human Swissprot database (v57.6, August 5, 2010, 719464 sequences) from which isoforms, variants and sequence conflicts were extracted [223]. At the same time, same unmatched spectra were searched against Human NCBI nonredundant database (March 20, 2010, 438240 sequences) to assign potential peptide sequences unreferenced in Swissprot database. All peptides identified with variant were blasted to assign unambiguously variant to protein. Substituted peptides shared with several sequence were assigned according to Swissprot annotation for variant and sequence conflict.

	Mascot Ion - Identity score	Omssa -log E	assigned spectra	identified proteins	Mascot reverse	Omssa reverse
Trypsin (43085 spectra)						
specific	13,55	7,15	5515	124	2	2
semi specific	7,9	7	1402	53	0	0
specific with modifications 1	12,8	7,25	659	48	0	0
specific with modifications 2	13,4	8,1	583	26	0	0
Chymotrypsin (15123 spectra)						
specific	5	8,6	282	26	0	0
semi specific	-5	8,6	269	31	0	0
specific with modifications 1	3,8	8,6	121	23	0	0
specific with modifications 2	13	8,1	96	15	0	0
GluC (9224 spectra)						
specific	4,7	5,8	294	32	0	0
semi specific	-5	8,8	43	12	0	0
specific with modifications 1	8	5,6	37	11	0	0
specific with modifications 2	2,35	7,2	26	5	0	0

***Supplementary data I** . Search criteria used for the validation of MS/MS spectra in each search step.*

3. Results

a) Experimental strategy and identification results

Identification of isoform families involved in keratinous constituents extracted from human hair cortex was performed after multiple digestions with three complementary endoproteases, namely trypsin, chymotrypsin and GluC (Figure 1). Multiple enzymatic digestions of the same sample led to increased sequence coverages, higher confidence in proteins assignments [198, 247, 248] and identification of more proteotypic peptides. In addition, trypsin digestion was performed in triplicate to increase MS/MS data set and allow a better dynamic range for subsequent spectral counting analysis. High-pH/low-pH 2D-LC-MS/MS allowed high separation efficiency, homogeneous distribution and easy experimental handling without desalting [211]. Improvement of peptides separation led to reduced isobaric co-elution and MS/MS undersampling. The high mass accuracy on MS and MS/MS measurements (<15ppm) conducted on a Q-TOF mass spectrometer permitted discriminating near isobaric compounds and restricting the space search. Both Mascot and OMSSA were employed to increase the number of MS/MS hits [249] and protein validation was performed using a target-decoy strategy. An iterative peptide identification strategy was employed. After a conventional search with enzymatic specific cleavage in the Human Swissprot curated database, semi specific peptides were searched to increase the number and sequence coverage of identified proteins.

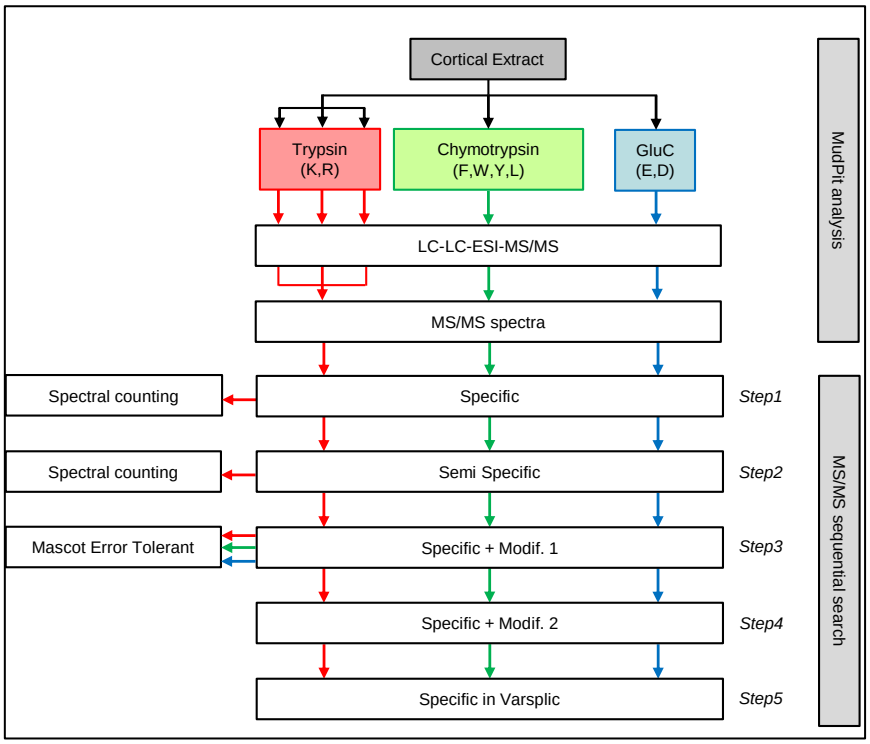


***Figure 1**: Experimental strategy for the identification of proteins from hair cortex.*

After the combination of specific and semi specific peptide identifications from the three digests, proteins identified with at least one unique peptide were retained and global sequence coverage was calculated. Table I shows the 63 keratins and KAP components identified after these two steps of iterative searches. In addition, 59 non-keratin proteins were identified (supplementary data 1). These proteins could be mainly attributed to nuclear and cytoplasmic remnants of keratinous cells. During the last stage of cortex keratinization process, nucleus and organelles are degraded and incorporated into intermacrofibrillar matrix. Finally, this compartment constitutes a minor fraction of the cortical cell.

Swissprot accession number	Description	MW (kDa)	Group[a]	Location[b]	unique peptides[c]						Search algorithm[d]	SC[e] (%)	TSC[f]	Protein existence[g]
					T	ST	C	SC	V8	Total				
Q15323	Hair keratin Type I K31	47,2	IFP	cortex	15	17	8	10	9	**59**	m+o	**99**	**12**	p
O76009	Hair keratin Type I K33a	45,9	IFP	cortex	13	8	8	3	7	**39**	m+o	**96**	**14**	p
Q14525	Hair keratin Type I K33b	46,2	IFP	cortex	10	11	2	4	5	**32**	m+o	**97**	**15**	p
O76011	Hair keratin Type I K34	49,4	IFP	cortex	19	21	15	10	10	**75**	m+o	**99****	**13**	p
Q92764	Hair keratin Type I K35	50,3	IFP	cort., cut.	14	4	2	3	4	**27**	m+o	**81**	**4**	p
O76013	Hair keratin Type I K36	52,2	IFP	cortex	2	1	1	0	0	**4**	m+o	**27**	**4**	p
O76015	Hair keratin Type I K38	50,5	IFP	cortex	4	2	0	0	4	**10**	m+o	**23**	**4**	p
Q6A163	Hair keratin Type I K39	55,6	IFP	cort., cut., med.	10	0	1	0	2	**13**	m+o	**35**	**3**	p
Q14533	Hair keratin Type II K81	54,9	IFP	cortex	2	4	2	1	0	**9**	m+o	**95**	**24**	p
P78385	Hair keratin Type II K83	54,2	IFP	cortex	9	7	1	3	3	**23**	m+o	**97**	**10**	p
P78386	Hair keratin Type II K85	55,8	IFP	cort., cut.	22	38	13	16	7	**96**	m+o	**98**	**15**	p
O43790	Hair keratin Type II K86	53,5	IFP	cortex	6	4	0	3	2	**15**	m+o	**94**	**14**	p
Q07627	KAP 1-1	18,2	HS	cortex	3	1	3	0	4	**11**	m+o	**89**	**2**	t
Q8IUG1	KAP 1-3	18,2	HS	cortex	1	0	0	0	4	**5**	m+o	**83**	**1**	t
Q9BYS1	KAP 1-5	18,0	HS	cortex	2	8	1	2	2	**15**	m+o	**91**	**3**	t
Q9BYU5	KAP 2-1	13,5	HS	cortex	12*	1	0	5*	0	**18***	m+o	**97**	**7**	t
A8MTN3	Putative KAP 2-X	13,5	HS	-	0	1	0	0	0	**1**	o	**89**	**1*****	ibh
Q9BYR8	KAP 3-1	10,5	HS	cortex	3	9	8	2	2	**24**	m+o	**93**	**11**	p
Q9BYR7	KAP 3-2	10,4	HS	cortex	1	0	1	0	0	**2**	m+o	**100**	**8**	p
Q9BYR6	KAP 3-3	10,4	HS	cortex	1	0	1	0	0	**2**	m+o	**100**	**6**	p
Q9BYQ7	KAP 4-1	13,2	UHS	cortex	7	1	0	0	0	**8**	m+o	**64**	**5**	t
Q9BYR5	KAP 4-2	14,4	UHS	cortex	4	0	0	0	0	**4**	m+o	**71**	**4**	t
Q9BYR4	KAP 4-3	20,5	UHS	cortex	8	1	0	0	1	**10**	m+o	**90**	**3**	t
Q9BYR3	KAP 4-4	18,0	UHS	cortex	6	1	0	0	0	**7**	m+o	**79**	**3**	t
Q9BYQ5	KAP 4-6	21,8	UHS	cortex	4	0	0	0	1	**5**	m+o	**97**	**3**	t
Q9BYR0	KAP 4-7	22,5	UHS	cortex	2	0	1	0	1	**4**	m+o	**81**	**4**	t

Q9BYQ9	KAP 4-8	20,7	UHS	cortex	2	0	1	0	0	**3**	m+o	**55**	**2**	t
Q9BYQ8	KAP 4-9 fragment	20,6	UHS	cortex	1	0	0	0	0	**1**	m+o	**78**	**9**	t
A8MTL4	*Putative KAP 4-X* [h]	16,7	UHS	-	1	0	0	0	0	**1**	m+o	**77**	**1**	-
Q9BYQ6	KAP 4-11	20,8	UHS	cortex	0	1	0	0	0	**1**	o	**55**	**1*** **	t
Q9BQ66	*KAP 4-12*	21,4	UHS	cortex	0	0	0	0	0	**0**		**83**	**-**	t
Q9BYR2	*KAP 4-5* [i]	19,9	UHS	cortex	-	-	-	-	-	**-**	-	**-**	**-**	t
Q9BYQ4	KAP 9-2	18,3	UHS	cortex	1	0	0	0	0	**1**	m	**66**	**1**	p
Q9BYQ3	KAP 9-3	16,8	UHS	cortex	2	2	0	3	0	**7**	m+o	**80**	**3**	p
Q9BYQ2	KAP 9-4	16,4	UHS	cortex	0	1	0	0	1	**2**	m+o	**72**	**1*** **	t
A8MVA2	KAP 9-6	16,8	UHS	-	1	0	0	0	0	**1**	m	**26**	**1**	ibh
A8MTY7	KAP 9-7	17,8	UHS	-	1	0	0	0	0	**1**	m+o	**37**	**2**	ibh
Q9BYQ0	KAP 9-8	16,8	UHS	cortex	2	0	0	0	0	**2**	m+o	**76**	**6**	t
Q9BYP9	KAP 9-9	16,5	UHS	cortex	2*	0	0	1	0	**3**	m+o	**61**	**3**	t
Q8IUC1	KAP 11-1	17,1	HS	cortex	4	5	3	3	0	**15**	m+o	**76**	**8**	p
Q8IUC0	KAP 13-1	18,3	HS	cortex	6	0	2	0	0	**8**	m+o	**89**	**2**	t
Q52LG2	KAP 13-2	18,7	HS	cortex	10	0	1	1	0	**12**	m+o	**75**	**2**	p
Q3LI64	KAP 6-1	7,3	HGT	cortex	1	0	0	0	0	**1**	m+o	**14**	**2**	t
Q8IUC3	KAP 7-1	9,3	HGT	cortex	2	8	2	4	0	**16**	m+o	**91**	**4**	t
Q8IUC2	KAP 8-1	6,8	HGT	cortex	0	2	2	0	0	**4**	m+o	**49**	**1*** **	t
Q8IUB9	KAP 19-1	9,0	HGT	cort./cut.	1	1	0	1	0	**3**	m+o	**44**	**1**	t
Q7Z4W3	KAP 19-3	8,2	HGT	cort./cut.	1	0	0	0	0	**1**	m	**14**	**1**	t
Q3LI72	KAP 19-5	7,6	HGT	cort./cut.	3	2	0	2	0	**7**	m+o	**86**	**3**	p
Q3SYF9	KAP 19-7	6,6	HGT	cort./cut.	1	0	0	0	0	**1**	m	**18**	**1**	t
Q3LI61	*KAP 20-2* [j]	6,9	HGT	cortex	-	-	-	-	-	**-**	-	**-**	**-**	t
P02533	Keratin Type I epithelial K14	51,5	IFP	medulla	2	0	0	0	0	**2**	m	**12**	**1**	p
P13647	Keratin Type II epithelial K5	62,4	IFP	medulla	2	0	0	0	0	**2**	m+o	**7**	**1**	p
P04259	Keratin Type II epithelial K6B	60,1	IFP	medulla	1	0	0	0	0	**1**	m	**7**	**2**	p
Q6KB66	Keratin Type II epithelial K80	50,5	IFP	medulla	1	0	0	0	0	**1**	m	**3**	**1**	p
Q14532	Hair keratin Type I K32	50,3	IFP	cuticle	4	0	0	0	2	**6**	m+o	**29**	**4**	p
Q9NSB4	Hair keratin Type II K82	56,6	IFP	cuticle	1	0	0	0	0	**1**	m+o	**10**	**3**	p
Q6A162	Hair keratin Type I K40	48,1	IFP	cuticle	1	0	0	0	0	**1**	m+o	**10**	**1**	p
Q701N4	KAP 5-2	16,3	UHS	cuticle	1	0	0	0	0	**1**	m	**11**	**1**	t
P60331	KAP 10-1	28,6	UHS	cuticle	3	0	0	0	0	**3**	m+o	**11**	**3**	t
P60014	KAP 10-10	25,6	UHS	cuticle	1	0	0	0	0	**1**	m	**6**	**1**	p
A8MUX0	Putative KAP 10-like	53,9	UHS	-	2	0	0	0	0	**2**	m+o	**3**	**1**	ibh
P59991	KAP 12-2	14,7	UHS	cuticle	0	0	0	2	0	**2**	m+o	**18**	-	p
P13645	Keratin Type I epithelial K10	58,8	IFP	epiderm	7	1	0	0	0	**8**	m+o	**25**	**2**	p
P04264	Keratin Type II epithelial K1	66,0	IFP	epiderm	11	0	0	0	0	**11**	m+o	**27**	**3**	p
P35908	Keratin Type II epithelial K2	65,4	IFP	epiderm	4	0	0	0	0	**4**	m+o	**13**	**3**	p
P35527	Keratin Type I epithelial K9	62,1	IFP	epiderm	8	0	0	0	0	**8**	m+o	**17**	**2**	p

***Table I** : Keratin and KAP protein identification.*

[a] According to Swissprot annotation.

[b] According to Rogers et al. and Langbein et al. [73, 101, 139, 140, 142].

[c] Unique peptides identified for trypsin (T), semi trypsin (ST), chymotrypsin (C), semi chymotrypsin (SC) and GluC (V8) digests. Shared peptides with other proteins of the table were not counted. Example of unique peptide spectrum for each protein is reported in supplementary data.

[d] Search algorithm used for the identification of unique peptides, Mascot (m), Omssa (o).

[e] SC, sequence coverage calculated with unique and shared peptides identified in all sequential search results.

[f] TSC, trypsin spectral counts based on spectral counts of unique peptides identified in T or ST for proteins marked with (**).

[g] According to Swissprot annotation. p, evidence at protein level, t, evidence at transcript level, ibh, inferred by homology.

[h] Unique peptide from KAP 4-X can be assigned to a polymorphic variant of KAP 4-7.

[i] KAP 4-5 was identified after search in NCBI database with tryptic peptides (table IV).

[j] KAP 20-2 can be identified from trypsin experiment with low restrictive search criteria (fig 8).

* Peptides of KAP 2-1 shared with KAP 2-X were considered as unique.

** After N-terminal correction.

*** According to semi trypsin experiment.

The results have shown that complementary digests allowed increase in sequence coverage rather than in the number of identified proteins (data not shown). The trypsin replicates have accounted for most of the identifications. The illustration of the benefit of our strategy for KAP 3 and 1 families is shown (Figure 2) with 93 to 100% and 83 to 91% sequence coverage, respectively. Proteins were classified according to locations obtained by in-situ hybridization and immunohistochemistry on keratins and KAP from previous works [3, 73, 77, 78, 80-83, 101, 139, 140, 142]. Several KAP identifications were based on a low number of unique peptides but with high sequence coverage due to high sequence homology among KAP families (inducing high numbers of shared peptides) and too short lengths of KAP sequences. For the first time, these results showed expression at the cortical proteome level of a large number of human KAP genes whose expression had only been proven at the transcript level or even only inferred by homology.

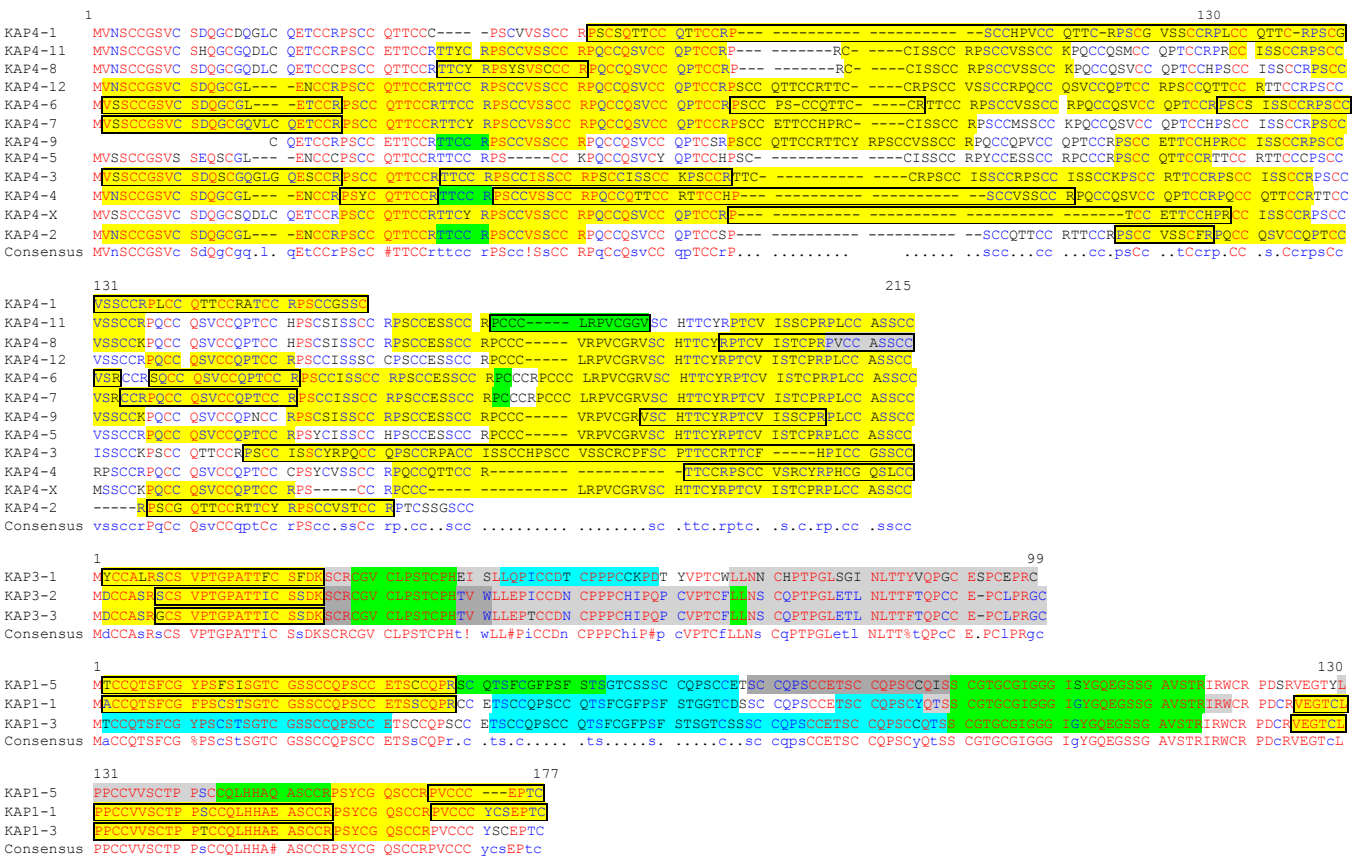


***Figure 2** : **Sequence coverage examples obtained for KAP 4, 3, 1 families after the two first steps of sequential search**. Sequence alignments were performed using MultAlin [250]. Framed, discriminant zone for trypsin digest. Yellow, trypsin ; green, semi trypsin ; light grey, chymotrypsin ; grey, semi chymotrypsin ; blue, gluC.*

b) Keratin and KAP Abundances and Compartmentalization

In addition to qualitative information on protein isoforms, spectral abundances obtained from trypsin digests were evaluated. In cortex analysis, high number of shared peptide among isoforms could hinder true abundance estimations of these proteins. To take into account sequence homology and size dissimilarities for protein abundance measurements, spectral abundance of unique peptides identified for each protein was chosen. Spectral counting (SC) was based on a large spectra set from trypsin replicates (43085 MS/MS spectra) considered as sufficient to obtain a good dynamic range count for the majority of keratins and KAP. Peptides shared by multiple proteins were not used. The SC value for a protein was calculated as follows:

$$SC = \frac{\sum ups}{up}$$

ups = number of spectra associated with unique peptides for the protein

up = number of unique peptides for the protein

Figure 3 shows cortical type I and type II keratins as the most abundant proteins with 28% and 25% of spectral counts associated to unique peptides of keratins and KAP proteins, respectively. UHS KAP 4 and 9 represent 21% of estimated abundance; important contribution of this group was probably related to the high isoform number in these families. Spectral abundance obtained must be balanced by the tryptic peptide nature for these protein families, particularly for KAP 4, 1 and 2. The major part of peptides used for abundance measurement was obtained with unfavorable tryptic cleavage between arginine and proline or cysteine which could lead to underestimating this protein abundance [251]. The results suggest that KAP 4 family could be the first constituent of matrix proteins. The three KAP 3 proteins are the major HS KAP expressed in the cortex, and individual spectral count of KAP 3-1 is close to some major cortical keratins. The HS KAP 11-1 and 2 have a non negligible individual contribution whereas KAP 1 and 13 are detected with lower abundance. The HGT KAP are the lowest represented

class with KAP 7-1, 6-1, 8-1 and 19. For this last protein family, KAP 19-5 seems to be more abundant than the other KAP 19 detected. Spectral count for KAP 8-1 was only obtained from semi trypsin results and was probably under-estimated. In fact, spectrum abundance observed in chymotrypsin experiments suggested more important representation of this protein among the HGT class (data not shown).

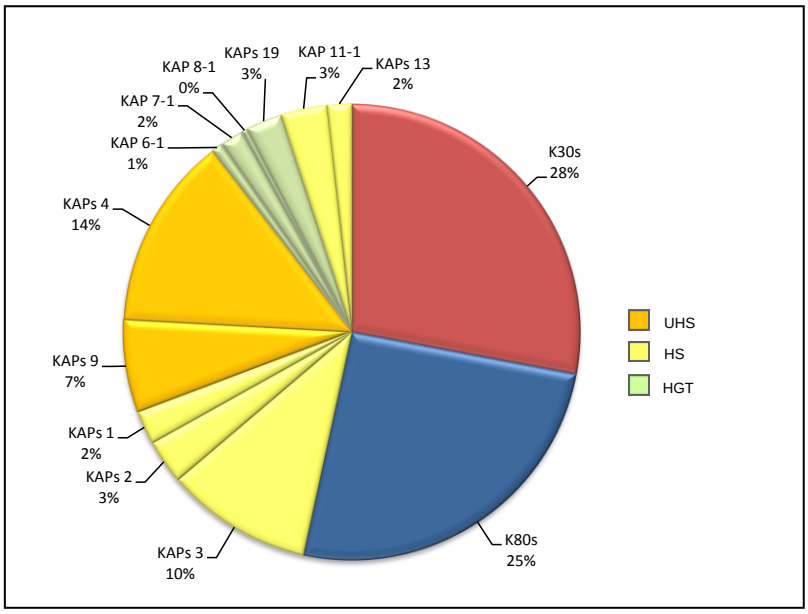


Figure 3 *: Spectral count distribution of cortical hair keratins and KAP according to Table 1 results.*

We have compared our abundance results to previous proteomic works on proteins from the insoluble extract of human hair shaft described by Lee at al. [89]. Conversion of IPI from Lee experiment to Uniprot accession numbers were performed with a home made software. Spectral counts for keratins and KAP were normalized for the two experiments on total spectral counts obtained for type I keratin in each trypsin digest Shotgun analysis. Then, protein abundances between the two experiments were compared (figure 4A and 4B). In addition, spectral counts from proteins classified as cortical, cuticular and medullar were dissociated to establish the origin of sample based on keratin and KAP composition (Figure 4A' and 4B'). Preliminary examination has suggested the cuticular origin of Lee's sample with the presence of a large number of cuticular KAP 10 and 12. Consequently, proteins described as co-expressed in cuticle and cortex (K85, K35, K39) were attributed to cuticle contribution for Lee's data and to cortex for our experiments. Classification and abundance analyses clearly indicate cortical origin of our sample with near to 90% of spectral purity. The presence of medulla proteins expected in the sample as this minor hair compartment in human might be extracted during the cortex solubilization. In addition, we have detected minor contamination of cuticle (K82, K32 and KAP 10-1) and epiderm. In Lee's experiment, approximately 60% of keratin and KAP unique spectra is from cuticular origin. The sample could be described as insoluble material formed by insoluble cuticle with unextracted cortex contamination and minor medulla. New protein name assignments have revealed detection of KAP 24-1 and 26-1 in this sample. These KAP have been described after Lee's work [139, 140]. Differential analysis of the spectral counts obtained from the two experiments has allowed us to confirm specific expression of cortical sulfur KAP 4, 9, 3, 2, 1 and 13 as described by previous in-situ hybridization experiments. HGT KAP 7-1, 6-1, 8-1 and 19 were also specifically detected in our cortical sample. In the same way, sulfur KAP 10, 12 and KAP 24-1 and 26-1 were specific to the insoluble sample despite negligible detection of some KAP 10 in the cortex sample. Only one ultra high sulfur KAP 5-2 was detected in both experiments. This member of KAP family attributed to cuticle by in situ hybridization was only found as minor in cortical sample. KAP 5-2 was found with only one peptide and the lack of the corresponding sequence in the IPI database (added in 3.06 version) used by Lee could likely explain lack of detection in this experiment. Nevertheless, absence of KAP 5 in cuticle sample remains unexplained.

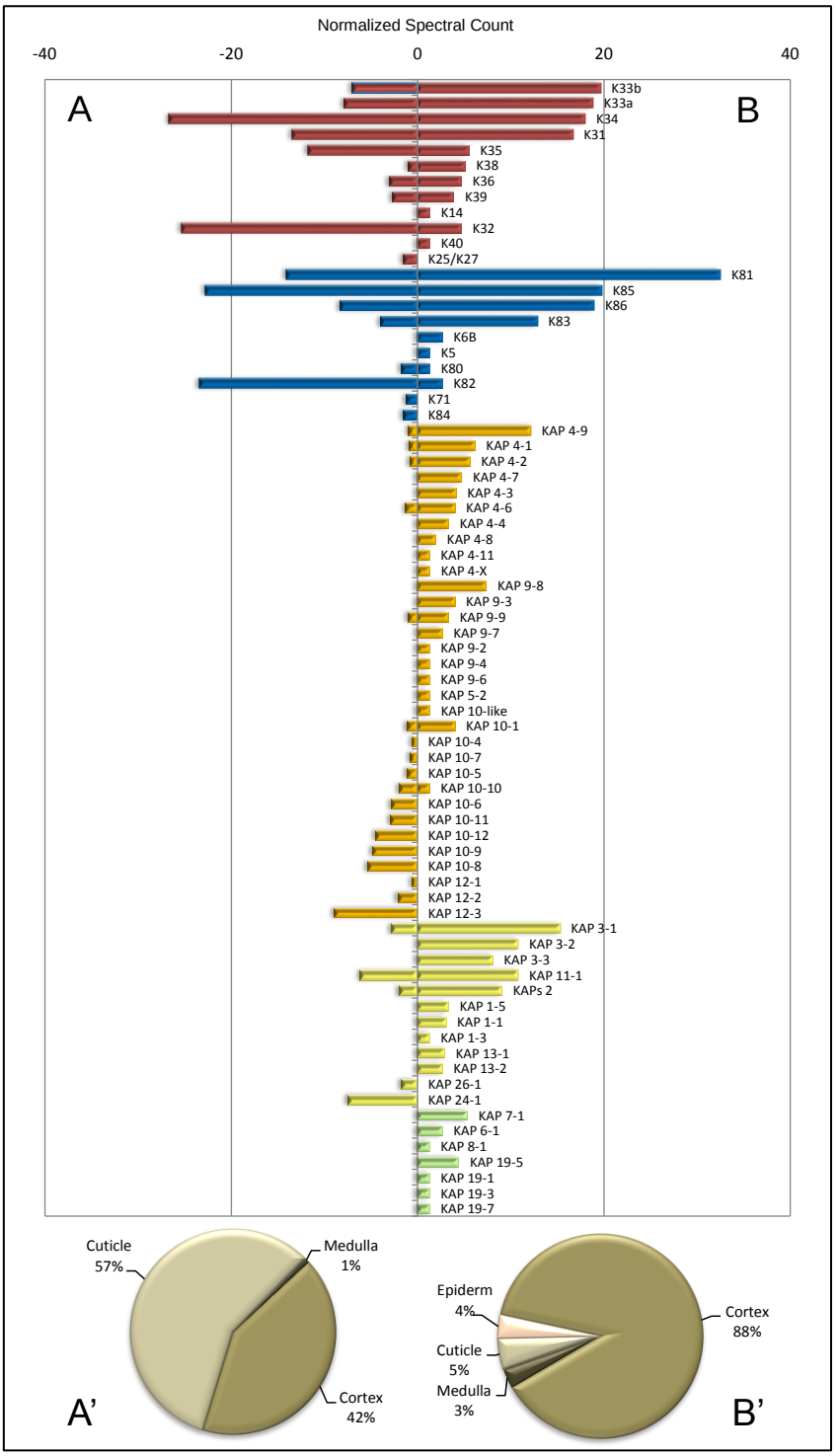


Figure 4 *: **Keratin and KAP spectral abundance differential analysis. Results from Lee et al. (A, A') and from this study (B, B')**. A and B spectral counts are normalized on the total spectral count of Type I keratins detected in each Shotgun experiment. Red, Type I keratins. Blue, Type II keratins. Orange, UHS KAP. Yellow, HS KAP. Green, HGT KAP. (*) based on semi trypsin experiments.*

A', B', compartment distribution of spectral counts. Protein locations were assigned according to results from Rogers, Langbein et al.. [73, 74, 101, 139, 140]. Proteins shared in several compartments are considered to belong in priority to cuticle in A' and to cortex then medulla in B'. For example K85, K35 and K39 were assigned to cuticle in A' and to cortex in B'.

Keratin's abundance in cortical sample suggests the major expression of expected type I K33b, K33a, K34, K31 and type II K81, K85, K86 and K83. Higher estimated expression of K81 protein must be balanced by the lowest number of unique peptides (2) compared to other type II keratins (respectively 22, 6 and 9 unique peptides for K85, K86 and K83). Other type I cortical keratins K35, K38, K36 and K39 were less abundant. Part of cortical

contamination in the cuticular sample has impacted on unambiguous location attribution of major cortical components. Thus, HS KAP 11-1 previously described as cortical protein was detected in the two experiments but seemed to be more abundant in the cortical sample. Type I and type II keratins previously described in the cortical sample were also found in the cuticular sample. But K81 contains only two possible unique peptides which were not detected by Lee even though they could be considered as initially present in the sample. Thus, spectral count obtained for K86 was divided according to the ratio obtained for these two proteins in the cortex sample. Nevertheless, spectral abundance on cuticle sample suggested expected overexpression of K32 and K82 as K85 and K35. Surprisingly, cortical K34 showed an unexpected overexpression in cuticle sample.

c) Modifications, Mutation Detection and Database Annotation Refinement

A very relevant illustration of the potential of our search strategy was the refinement of K34 sequence (Figure 5). After the iterative search step, K34 was covered except for 50 amino acids at the start of the sequence. An adapted search was performed to localize the correct N-terminal of this protein and brought attribution of acetylated peptide corresponding to the start of protein after the 43th amino acid of the sequence.

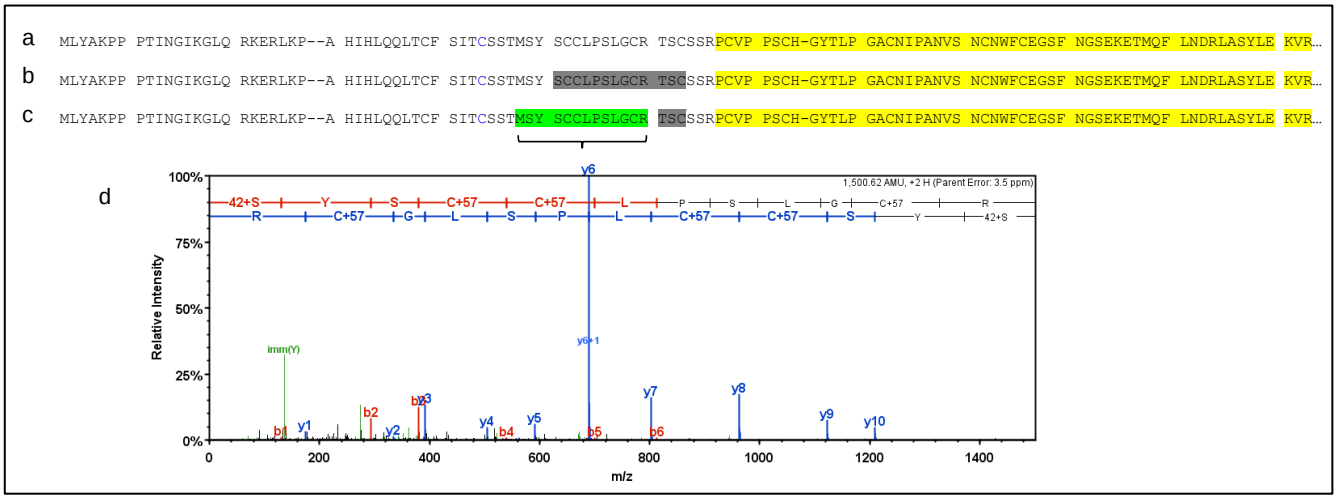


***Figure 5** : **Strategy for N-terminal assignment of K34**. After the two steps of iterative search, no semi specific spectra at the end of the covered sequence was identified (a, trypsin and semi trypsin ; b, semi chymotrypsin). We have suggested another position of the initiator with a systematic N-acetylation of the protein. To locate real N-terminal of K34, another step in the sequential search was performed. Unmatched spectra after the second step were searched with semi specific cleavage and N-acetylation on peptide N-terminal as variable modification. We have unambiguously identified 8 MS/MS spectra for a N-acetylated tryptic peptide localizing the start of the sequence after the second methionine of K34 sequence (c). Considering this sequence, the coverage for K34 was 99%. MS/MS spectrum of K34 N-terminal peptide $S^{acetyl}YSC^{am}C^{am}LPSLGC^{am}R$ after Met-43 of the Swissprot sequence O76011. Spectrum was identified after trypsin semi specific search with N-acetylation on peptide N-terminus as variable modification (d).*

In general, our dataset allowed 27 N-terminal sequence identifications corresponding to 38 proteins (31%) (Table II). Consequently, we have also used our results to study N-acetylation of the identified proteins, as this modification had already been described as occurring in isoform families previously identified. N-terminal acetylation is mediated by N-terminal acetyltransferase (NAT) activity and three main NATs are described to modify substrates with different specificity [252, 253]. Our results showed that N-terminal acetylation occurred differently according to isoform family and suggested sequence dependent action of NATs as described elsewhere and consistent with previous studies in humans (see Table II for details). After the modifications search step, trioxidations on KIF and KAP cysteines were located (Table III). Because cysteines were protected during extraction, the alterations can be unambiguously assigned to the initial sample. They seem to occur in the same way for K80s and K30s keratin proteins and for major sulfur KAP 4, 9, 3 and 2 but also for High Glycin and Tyrosine KAP 7. Cysteic acid, negligible in infant hair, is naturally present in virgin hair as a result of weathering and

probably photolytic reaction [254]. It is produced by the hemolytic cleavage of disulfide bonds and oxidation to sulfonic acid by air oxygen and water [255]. Type I and type II keratins are affected on head, coil and tail part of their protein sequence. Our results suggest that these chemical alterations indistinctively affect microfibrils and matrix proteins. Evidence of site specificity was not brought out as otherwise shown for the modified peptide detected for KAP 9 (R.PACCETTCCR.T) for which 4 distinct MS/MS spectra showed possible location of the modification on the four cysteines. On the contrary, only one site was found for KAP 2 with two different peptides.

Proteins	Peptide sequence	Digest	Number of spectra	
			free NH_2	N-acetyl
K31/K33b	PYNFC^{am}LPSLSC^{am}R	tryps	41	3
K33a	SYSC^{am}GLPSLSC^{am}R	tryps	0	6
K34	SYSC^{am}C^{am}LPSLGC^{am}R	tryps	0	8
K81	TC^{am}GSGFGGR	tryps	6	0
K83	TC^{am}GFNSIGC^{am}GFR	tryps	7	1
	TC^{am}GFNSIGC^{am}GFRPGNFSC^{am}VSAC^{am}GPR	tryps	1	1
	TC^{am}GFNSIGC^{am}GFRPGNF	semitryps	5	0
K86	TC^{am}GSYC^{am}GGR	tryps	9	1
KAP 1-1	AC^{am}C^{am}QTSFC^{am}GFPSC^{am}STSGTC^{am}GSSC^{am}C^{am}QPSC^{am}C^{am}ETSSC^{am}QPR	tryps	0	5
	AC^{am}C^{am}QTSFC^{am}GFPSC^{am}STSGTC^{am}GSSC^{am}C^{am}QPSC^{am}C^{am}E	GluC	0	1
KAP 1-3	TC^{am}C^{am}QTSFC^{am}GYPSC^{am}STSGTC^{am}GSSC^{am}C^{am}QPSC^{am}C^{am}E	GluC	1	1
KAP 1-5	TC^{am}C^{am}QTSFC^{am}GYPSFSISGTC^{am}GSSC^{am}C^{am}QPSC^{am}C^{am}ETSC^{am}C^{am}QPR	tryps	2	2
	TC^{am}C^{am}QTSFC^{am}GYPSFSISGTC^{am}GSSC^{am}C^{am}QPSC^{am}C^{am}E	GluC	0	1
KAP 2-1/2-X	TGSC^{am}C^{am}GSTFSSLSYGGGC^{am}C^{am}QPC^{am}C^{am}C^{am}R	tryps	4	1
	TGSC^{am}C^{am}GSTFSSL	semitryps	1	0
KAP 3-1	M^{ox}YC^{am}C^{am}ALR	tryps	15	0
	M^{ox}YC^{am}C^{am}ALRSC^{am}SVPTGPATTF	chymo	1	0
KAP 3-2/3-3	MDC^{am}C^{am}ASR	tryps	0	1
KAP 4-3	VSSC^{am}C^{am}GSVC^{am}SDQSC^{am}GQGLGQESC^{am}C^{am}R	tryps	1	0
	VSSC^{am}C^{am}GSVC^{am}SDQSC^{am}GQGLGQE	GluC	1	0
KAP 4-6	VSSC^{am}C^{am}GSVC^{am}SDQGC^{am}GLETC^{am}C^{am}R	tryps	2	0
	VSSC^{am}C^{am}GSVC^{am}SDQGC^{am}GLE	GluC	3	0
KAP 4-7	VSSC^{am}C^{am}GSVC^{am}SDQGC^{am}GQVLC^{am}QETC^{am}C^{am}R	tryps	4	0
	VSSC^{am}C^{am}GSVC^{am}SDQGC^{am}GQVL	chymo	1	0
	VSSC^{am}C^{am}GSVC^{am}SDQGC^{am}GQVLC^{am}QE	GluC	3	0
KAP 4-2/4-4	VNSC^{am}C^{am}GSVC^{am}SDQGC^{am}GLENC^{am}C^{am}R	tryps	2	0
	VNSC^{am}C^{am}GSVC^{am}SDQGC^{am}GLE	GluC	4	0
KAP 7-1	TRYFC^{am}C^{am}GSYFPGYPIYGTNFHGTFR	tryps	4	0
	TRYFC^{am}C^{am}GSYFPGYPIYGTNFHGTF	semichymo	4	0
KAP 8-1	M^{ox}LC^{am}DNFPGAVFPGC^{am}Y	chymo	8	0
	M^{ox}LC^{am}DNFPGAVFPGC^{am}YW	chymo	1	0
KAP 9-2/9-7/9-3/9-8/9-4/9-1/9-9	THC^{am}C^{am}SPC^{am}C^{am}QPTC^{am}C^{am}R	tryps	10	6
KAP 9-6	THC^{am}C^{am}SPGC^{am}QPTC^{am}C^{am}R	tryps	0	1
KAP 13-1/13-2	SYNC^{am}C^{am}SGNFSSR	tryps	0	4
KAP 19-1	SHYGSYYGGLGY	semitryps	2	0
Calmodulin-like protein 3	ADQLTEEQVTEFK	tryps	0	2
Cell division protein kinase 7	ALDVKSR	tryps	0	2
Macrophage migration inhibitory factor	PM^{ox}FIVNTNVPR	tryps	4	0
Glutathione S-transferase P	PPYTVVYFPVR	tryps	1	0
Protein S100-A3	ARPLEQAVAAIVC^{am}	semitryps	0	2
	ARPLEQAVAAIVC^{am}TFQEY	chymo	0	3

***Table II**. Proteins N-terminal characterization with peptides identified from first steps of the three enzymatic searches. Proteins starting with proline are not NAT substrates and coherently K31 and K33b were found to be free at their N-termini. Nevertheless, minor N-acetylation was also detected on proline of these proteins despite no previous description of this modification on the site. K30s proteins could be free (K31, K33b) or acetylated (K33a, K34) according to their starting amino acids. KAP 3-1 with rare N-terminal methionine-tyrosine showed free N-terminal. KAP 4 family members are also free as predicted for valine starting sequences, proteins starting with alanine (KAP 1-1, calmodulin-like 3, cell division kinase 7) are systematically acetylated and proteins starting with threonine (K80s, KAP 1-3 and 1-5, KAP 9) are partially acetylated. Sequences starting with serine are typically acetylated but unexpected free serine on HGT KAP 19-1 was detected and suggested systematic free N-terminal amine on HGT KAP identified (8-1, 7-1 and 19-1).*

Error tolerant searches have suggested amino acid substitution on keratin and KAP peptides and several chemical alterations such as carbamylation on amines, modifications on peptides N-terminal eg. glutamine cyclization, acylation with glutamic acid or acetamided cysteine. Such chemical modifications attributable to sample handling during extraction and separation were specifically looked for in the fourth step (Figure 1). Indeed, non negligible deamidation, and production of pyro-glu on peptides with glutamine, acetamided cysteine or glutamic acid on N-terminus were observed after combined analysis of the third and fourth search steps (Figure. 6). These amino acid lateral functions modifications involved in high-pH conditions suggest impact of the first stage of separation on amide hydrolysis. High-pH reverse phase chromatographic mode seems to be less appropriate to perform labile modification measurements and sample deamidation estimation. Carbamylation of N-terminal amine and epsilon-lysine subsequent to urea exposure during extraction and digestions was also observed. No methylations were found, although previously detected on hair keratins [89]. It suggests the low level of these modifications usually described on nucleus histones and exclude a significant regulation mechanism on keratins.

Proteins	Peptide sequence	Number of spectra	
		with modification	without modification
K31/K33b	M.PYNFC^{am}LPSLSC^{tox}R.T	6	44
K31/K33b/K34	R.SQQQEPLLC^{tox}PSYQSYFK.T	3	18
K33a	R.SQQQEPLVC^{tox}ASYQSYFK.T	1	7
K31/K34	R.ILDELTLC^{tox}K.S	2	85
K33a/K33b	R.ILDELTLC^{tox}R.S	2	70
K34	K.SDLESQVESLREELIC^{tox}LK.K	2	6
K33a/K31/K35/K34	R.ARLEC^{tox}EINTYR.S	3	11
K33a/K31/K33b/K35/K34	R.LEC^{tox}EINTYR.S	4	50
K33a	K.STGPC^{am}ISNPC^{tox}GLR.A	1	11
K31	R.C^{am}GPC^{tox}NSFVR-	1	24
K86	M.TC^{tox}GSYC^{am}GGR.A	2	10
K86/K81	R.AFSC^{am}ISAC^{tox}GPRPGR.C	3	2
K86/K81	R.AFSC^{tox}ISAC^{am}GPRPGR.C	1	2
K86/K81/K83	R.GLTGGFGSHSVC^{tox}GGFR.A	6	41
K86/K81	R.EC^{am}C^{tox}QSNLEPLFEGYIETLR.R	4	33
K83	R.EC^{am}C^{tox}QSNLEPLFAGYIETLR.R	3	17
K85	R.C^{am}C^{tox}ESNLEPLFSGYIETLR.R	2	26
K85/K81/K83/K86	R.REAEC^{tox}VEADSGR.L	1	2
K85/K81/K83	R.EAEC^{tox}VEADSGR.L	3	88
K85/K81/K86	R.DLNMDC^{tox}IIAEIK.A	1	175
K85/K81/K86	R.DLNM^{ox}DC^{tox}IIAEIK.A	1	175
K86/K81/K83	R.LTAEVENAKC^{tox}QNSK.L	1	0
K86/K81/K83	K.AKQDMAC^{tox}LIR.E	2	35
K85/K81/K83/K86	R.C^{tox}KLAELEGALQK.A	7	2
K86	R.GGVVC^{am}GDLC^{tox}ASTTAPVVSTR.V	5	35
K86	R.GGVVC^{tox}GDLC^{am}ASTTAPVVSTR.V	3	35
K85	R.GGVSC^{tox}GGLSYSTTPGR.Q	6	30
K85	R.QITSGPSAIGGSITVVAPDSC^{tox}APC^{am}QPR.S	1	23
K85	R.SSSFSC^{tox}GSSR.S	2	10
KAP 2-1	R.DPC^{am}C^{am}C^{am}RPVTC^{tox}QTTVC^{am}R.P	2	12
KAP 2-1	R.PVTC^{tox}QTTVC^{am}RPVTC^{am}VPR.C	2	18
KAP 3-1	R.SC^{am}SVPTGPATTFC^{tox}SFDK.S	1	16
KAP 3-1	R.SC^{tox}SVPTGPATTFC^{am}SFDK.S	1	16
KAP 3-2	R.SC^{tox}SVPTGPATTIC^{am}SSDK.S	2	8
KAP 4-4	R.PQCC^{tox}Q^{dam}TTC^{am}C^{am}R.T	1	2
KAP 4-7/4-11/4-8/4-6/4-9	R.PSC^{tox}C^{am}ESSC^{am}C^{am}R.P	1	6
KAP 4-7/4-6/4-X	R.VSC^{am}HTTC^{am}YRPTC^{tox}VISTC^{am}PR.P	1	17
KAP 4-7/4-6/4-X	R.VSC^{tox}HTTC^{am}YRPTC^{am}VISTC^{am}PR.P	1	17
KAP 4-7/4-9	R.PSC^{tox}C^{am}ETTC^{am}C^{am}HPR.C	1	17
KAP 7-1	R.YFC^{am}C^{tox}GSYFPGYPIYGTNFHGTFR.A	1	4
KAP 9-2/9-7/9-3/9-8/9-9	R.PAC^{am}C^{am}ETTC^{am}C^{tox}R.T	1	37
KAP 9-2/9-7/9-3/9-8/9-9	R.PAC^{am}C^{am}ETTC^{tox}C^{am}R.T	1	37
KAP 9-2/9-7/9-3/9-8/9-9	R.PAC^{tox}C^{am}ETTC^{am}C^{am}R.T	2	37
KAP 9-2/9-7/9-3/9-8/9-9	R.PAC^{am}C^{tox}ETTC^{am}C^{am}R.T	2	37

Table III *: Trioxidized peptides identified from third step of trypsin search. Numbers of spectra are counted from the third step search for trioxidized peptides and from first step search for unmodified peptides.*

Finally, we have focused our searches on variant identifications. The results obtained from Mascot error tolerant, Varsplic and NCBI searches are shown in Table IV. According to Varsplic searches, we have identified 2 annotated variants for K30s, 4 variants and 2 sequence conflicts annotated for K80s. Varsplic searches also identified a long peptide corresponding to isoform 2 of KAP 9-9. This identification on a previously uncovered part of KAP 9-9 sequence suggested unique expression of this isoform in the cortical sample.

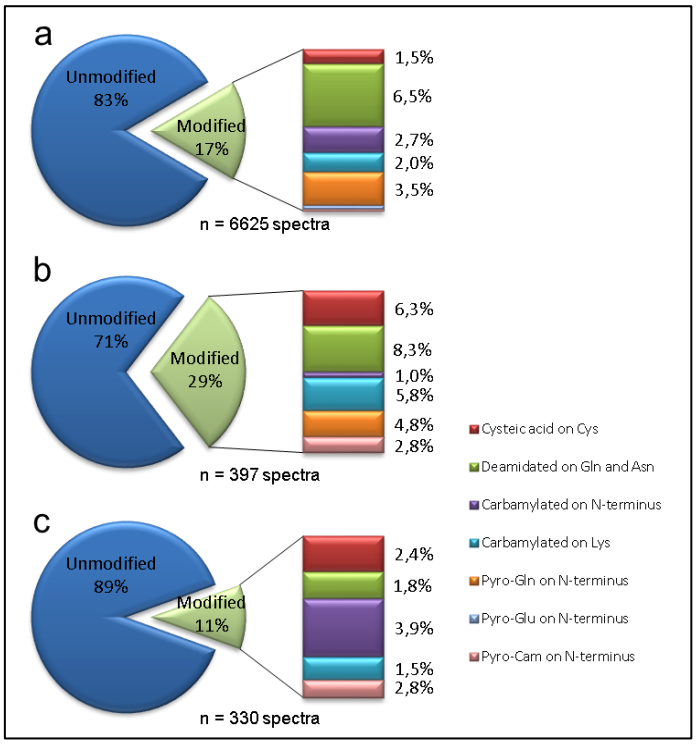


Figure 6 *: Spectral abundance of unmodified and modified peptides. (a) trypsin digest ; (b) chymotrypsin digest ; (c) GluC digest. Spectra of unmodified peptides were extracted from the first step search. Trioxidized and deamidated peptides were extracted from the third step search. Other modified peptides were extracted from the fourth step search. Validation for these spectra were performed according to criteria described in Supplementary data I.*

NCBI searches revealed several sequence errors in Swissprot related to previously unidentified KAP 4-5 which was detected with two unique peptides in our experiment. One of them was the N-terminal peptide. This sequence correction thus led to identify new additional protein from our dataset. KAP 4-5 referenced as Q9BYR2 contains two erroneous amino acid substitutions compared to the NCBI sequence. Another error was detected at the N-terminal part of KAP 4-11 with a substitution of one amino acid. The N-terminal sequence of KAP 4-9, incomplete in Swissprot, was successfully matched for the KAP 4-9 version in NCBI and could be corrected. We also detected one annotated variant for KAP 4-1 previously described as belonging to polymorphic form of KRTAP 4-1 gene found in 97% of Japanese corresponding allele population [256]. In parallel, substitution on KAP 4-8 was identified and corresponded to a polymorphic form of this gene described as occurring in 52% of Japanese allelic population [256]. A sequence conflict on KAP 9-8 was detected with a complementary sequence of previous TCYHPTTVCLPGCLNQSCGSSCCQPCCR peptide with one amino acid substitution. Simultaneous detection of this two peptides in our sample suggests two polymorphic expressions of KAP 9-8 gene.

Moreover Mascot error tolerant search results suggested four unreferenced substitutions on K30s or K80s. An unreferenced substitution was also observed with glutamine to leucine on shared KAP 4 peptide PSCCQTTCCR. Due to sequence homology these five substitutions could not be assigned to distinct genes.

Proteins	Peptide sequence	Digest	Search strategy			Ref. Uniprot
			Mascot error tolerant	Varsplic	NCBI	
K31	R.DN(**A->V**)ELENLIR.E	Tryps	x	x		VAR_046990
K31/K34/K36	R.(**R->K**)ILDELTLCamK.S	Tryps	x			
K33a/K31/K33b	R.SQYE(**A->V**)LVETNR.R	Tryps	x			
K33a/K33b	K.QVVSSSEQLQSYQ(**A->V**)EIIELR.R	Tryps	x	x		VAR_054432
K81	K.LLETKL(**Q->P**)FYQNR.E	Tryps		x		Sequence Conflict
K81	K.L(**Q->P**)FYQNR.E	Tryps		x		Sequence Conflict
K81	F.LEQQNKLLETKL(**Q->P**)F.Y	Chymo	x	x		Sequence Conflict
K81	E.TKL(**Q->P**)FYQNRE.C	GluC	x	x		Sequence Conflict
K81	E.I(**R->L**)ILQSHISD.T	GluC		x		VAR_018114
K81	R.GLTGGFGSHSVC(**C->R**).G	Tryps		x		VAR_018113
K83	R.DLNMDCam(**I->M**)VAEIK.A	Tryps	x	x		VAR_018120
K83	E.TKLQFYQN(**R->C**)E.C	GluC		x		VAR_018119
K83	L.NTTCamGGGSCamGQGR(**H->Y**)-	Chymo		x		Sequence Conflict
K85[#]	R.AGSCamG(**R->H**)SFGYR.S	Tryps	x	x		VAR_029657
K81/K83/K85/K86	R.EAECam(**V->A**)EADSGR.L	Tryps	x			
K81/K83/K85/K86	K.L(**A->P**)ELEGALQK.A	Tryps	x			
KAP 4-5 [a]	M.VSSCamCamGSVSSEQSCamGLENCamCam**R**.P	Tryps			x	
KAP 4-5 [a]	R.PS**Cam**CamISSCamCamHPSCamCamESSCamCamR.P	Tryps			x	
KAP 4-11 [b]	M.VNSCamCamGSVCamSHQGCamG**R**.D	Tryps			x	
KAP 4-11 [b]	R.DLCamQETCamCamR.P	Tryps			x	
KAP 4-9 [c]	M.**VSSCamCamGSVCamSDQGCamGQDL**CamQETCam CamR.P	Tryps			x	
KAP 4-1 [d]	R.PSCamCam(**H->R**)PVCamCamQTTCamR.P	Tryps			x	VAR_047044
KAP 4-8 [e]	P(**T->A**)CamVISTCamPR	Tryps	x		x	
KAP 4-2/4-6/4-8/4-X/4-3/ 4-9/4-5/4-12/4-7	PSCamCam(**Q->I/L**)TTCamCamR	Tryps	x			
KAP 9-9 isoform2	R.TTCamCamQPTCamL**TSCamCamQPSCamCamSTTCam CamQPICamCamG**SSCamCamGQTSCamGSSCamGQSSSCamAPVYCam R.R	Tryps		x	x	VSP_028981
KAP 9-8 [f]	R.TCamYHPTTVCamLPGCamLNQSCamGS(**S>N**)CamCamQPCamCamR.	Tryps			x	Sequence Conflict

***Table IV** : Variant and sequence conflict detection and database annotation refinement identified with the 3 search strategies. See supplementary data for spectrum details.*

[a] Q9BYR2 corrected.
[b] Q9BYQ6 corrected, corresponding to polymorphic variant 4-14 described [256].
[c] Q9BYQ8 corrected
[d] Polymorphic variant 4-10 [256].
[e] Polymorphic variant 4-8 v1 [256].
[f] Our results showed the two substituted sequences with Ser et Asn. This position could be described as a variant for KAP 9-8.
This variant on K85 is associated with ectodermal dysplasia disease according to Human Intermediate Filament Database.

4. Discussion

Human hair cortical proteome mainly consists of specific IFPs and KAP from macrofibrils of cortical cells. Characterization of these proteins has requested an adapted strategy which allowed successful identification of 10 out of 11 type I and 5 out of 6 type II hair keratins [127] and close to half of predicted KAP proteins [138]. Sequence of proteins identified as the major constituents thanks to spectral abundance analysis were next totally covered in experiments. These results also allowed protein modification, protein sequence variation and database error for a large part of these proteins to be studied. Among unidentified hair KIF proteins, K37 and K84 were not previously described as expressed in hair but in vellus hair and tongue papilla, respectively [82]. Four keratin sequences from medulla were also partially covered according to the low content of these cells in our sample.

a) Abundance results

Abundance results on mature hair could be compared to keratin expression and KAP in situ hybridization in the follicle. Among proteins whose expression starting is detected in the mitotic zone of the bulb called matrix and the pre-elongation zone [73, 75, 78, 79, 81-83, 108, 131, 139, 140, 257], K31 and K85 showed a strong expression in the final structure whereas K35 was detected as a minor keratin. It suggests K85 and K31 as potential heterodimers association partners during intermediate filament formation start. Regarding matrix KAP, KAP 11-1 is individually more expressed, KAP 13-1 and 13-2 are slightly expressed whereas HGT KAP 8-1 abundance is ambiguous but probably important. For other proteins detected as expressed in the upper cortex corresponding to elongation and keratogenous zones [82, 83, 108], K81, K86, K83, K33a, K33b and K34 were found as the major proteins. The high expression of these IFP isoforms illustrates their contribution to the high level of protein synthesis during keratinization stage. These observations are valid for KAP expressed at the last stage of keratinization process and incorporated into growing macrofibrils [73]. Indeed, abundance results suggest that approximately 35% of keratinous proteins are HS KAP from 4, 3, 9, 2 and 1 families, from the 17q21.2 chromosomal locus. HGT KAP 19, 7-1 and 6-1 described as expressed approximately at the same time are largely less represented. K36, K38 and K40 type I keratins are also less expressed. Surprisingly, spectral abundance ratio between keratins from microfibrils and KAP expected from the matrix obtained in our study was 53/47 and was very similar to previous results from X-ray scattering on human hair (54/46) [121]. Additionally, the near equivalent spectral abundance between type I and type II hair KIFs (53/47) is consistent with equimolar ratio expected due to heterodimer association [86]. It suggests consistency of the measurement despite expected low accuracy of spectral count used for quantification.

b) KAP gene expression evidence

Although the whole set of expected cortical KIFs was detected, identified KAP in the cortex compared to the high number of expected genes might be discussed according to gene annotation and description. For KAP 1 family, 3 over 4 genes were found expressed at the protein level. Interestingly, the unidentified KAP 1-4 exhibits the highest sequence dissimilarity compared to the 3 identified proteins with a different N-terminal and was firstly not identified at the transcript level [81]. These two results suggest non coding nature for the corresponding gene despite in situ hybridization result described elsewhere [258]. KAP 2 family is described with 5 genes and 1 pseudo gene [73]. These genes present the highest sequence homology among KAP families with only one or two amino acid substitution to differentiate protein sequences. Analyses have allowed only unambiguous KAP 2-1 identification with 97% of sequence coverage corresponding to simultaneous identification of KRTAP 2-1A and KRTAP2-1B gene expression. Unique peptides corresponding to 2-2, 2-3 and 2-4 were not found despite potential chymotryptic and tryptic specific peptides. Regarding KAP 2-2, only one unique peptide differs from 2-1 with only one amino acid substitution. Nevertheless, database annotation describes this position as sequence conflict and suggests putative total similarity between the two sequences which could explain unidentification. A surprising specific peptide was found for the pseudo gene described in the protein database. Several KAP pseudo genes are described for KAP families or hair keratins. Pseudogene character is based on the presence of premature stop codons or frameshifts in the coding regions of the genes [81]. Sequence described for this pseudo gene differs on the C-terminal sequence compared to 2-1 and was covered with a semi tryptic peptide with unambiguous MS/MS spectra. Identification of this peptide suggests unexpected expression of the corresponding locus. The fact that KAP 2-X was identified with semi tryptic peptide R.PccWATTccQPVSVQSPcG.Q corresponding to an aspecific cleavage at the C-terminal questioned whether glycine was the C-terminal of the protein. Protein expression of 3 over 4 of the KAP 3 genes was observed, the undetected protein was the pseudogene not described in database used. Regarding KAP 4 family, compared to the 11 genes [73] and 3 pseudogenes described [138], 10 were identified, including KAP 4-5 after search in NCBI protein database. Our search also identified 3 new proteotypic peptides for KAP 4-11, 4-5 and 4-9 suggesting 3 N-terminal corrections in the database. KAP 4-12 could not be

unambiguously assigned with Shotgun strategy due to the lack of specific peptide compared to other isoforms but it was virtually covered at 83% of sequence. Regarding pseudo gene expression of KAP 4-X, unique peptide PTccETTccHPR identified for this sequence is shared by 2 polymorphic variants for the KAP 4-7 described elsewhere [256] whose expression is more likely.

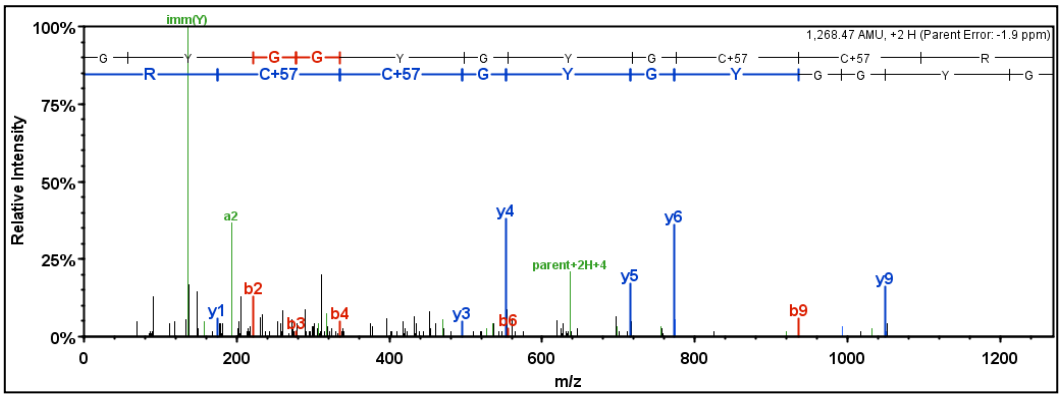


***Figure 7**. MS/MS spectrum of HGT KAP 20-2 peptide R.GYGGYGYGC^{am}C^{am}R.P. Search criteria used for the validation were too restrictive to validate this spectrum.*

For KAP 9, 7 proteins were identified. The number of expected KRTAP9 genes was 8. The only unidentified expected protein arises from KRTAP 9-1 gene, the only one located away from the family cluster and close to KRTAP 4 gene's cluster [73]. KAP 11-1, 7-1 and 8-1 subfamilies described with e one gene each were identified. Only one of the three expected gene products of KRTAP 6, KAP 6-1, was observed. As regards KAP 13 family, recently extended to 6 genes with grouping of KRTAP 15 family [138], only KAP 13-1 and 13-2 were identified. Finally HGT KAP 19 were found with 4 members compared to the 8 sequences present in the database and the 7 to 9 predicted genes depending on literature sources and grouping of KRTAP 22 into the subfamily [73, 138].

Undetected, but expected cortical KAP families were HGT KAP 20 and 21 and HS KAP 23-1. Among undetected KAP corresponding to predictive genes expressed or not at transcript level, either protein could poorly be expressed or maybe not expressed at all. To take into account the probability of non validated MS/MS spectra from these putative proteins due to restrictive parameters in our search criteria, we have attempted to search results with less stringent parameters. Indeed, this approach led to identify a spectrum corresponding to a peptide from KAP 20-2 (Figure 7) which has completed our list of identified cortical KAP summarized on table V.

KAP family	Category	Chromosomal location	Gene Identified	Identified	Missing	Note
KAP 1	HS	17q21.2	3/4	1.1 ,1.3, 1.5	1.4	1.4 pseudogene?
KAP 2	HS	17q21.2	3/5	2.1a/2.1b, 2.X	2.2, 2.3, 2.4	2.1a and 2.1b have 100% of sequence homology KAP 2.X described as 2.2 in ref [138]
KAP 3	HS	17q21.2	3/3	3.1, 3.2, 3.3		
KAP 4	UHS	17q21.2	10/11	4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.11	4.12	4.12 sequence was virtually covered with 83% of sequence coverage
KAP 5	UHS	11p15.5 / 11q13.5	1/12	5.2	5.1, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11	expected in hair cuticle
KAP 6	HGT	21q22.1	1/3	6.1	6.2, 6.3	low abundance compound
KAP 7	HGT	21q22.1	1/1	7.1		
KAP 8	HGT	21q22.1	1/1	8.1		
KAP 9	UHS	17q21.2	7/8	9.2, 9.3, 9.4, 9.6, 9.7, 9.8, 9.9	9.1	9.1 pseudogène?
KAP 10	UHS	21q22.3	2/12	10.1, 10.10	10.2, 10.3, 10.4, 10.5, 10.6, 10.7, 10.8, 10.9, 10.11, 10.12	expected in hair cuticle

KAP 11	HS	21q22.1	1/1	11.1		
KAP 12	UHS	21q22.3	1/4	12.2	12.1, 12.3, 12.4	expected in hair cuticle
KAP 13	HS	21q22.1	2/4	13.1, 13.2	13.3, 13.4	
KAP 15	HS	21q22.1	0/1		15.1	
KAP 16	HS	17q21.2	1/1	16.1		KAP 10-like 1 described as 16.1 [138]]
KAP 17	UHS	17q21.2	0/1		17.1	expected in hair cuticle
KAP 19	HGT	21q22.1	4/8	19.1, 19.3, 19.5, 19.7	19.2, 19.4, 19.6, 19.8	low abundance compounds
KAP 20	HGT	21q22.1	1/4	20.2	20.1, 20.3, 20.4	low abundance compounds
KAP 21	HGT	21q22.1	0/3		21.1, 21.2, 21.3	
KAP 22	HGT	21q22.1	0/2		22.1, 22.2	
KAP 23	HS	21q22.1	0/1		23.1	
KAP 24	HS	21q22.1	0/1		24.1	expected in hair cuticle
KAP 25	HS	21q22.1	0/1		25.1	inferred by homology
KAP 26	HS	21q22.1	0/1		26.1	expected in hair cuticle
KAP 27	HS	21q22.1	0/1		27.1	inferred by homology
KAP 28	UHS	2q36.3	0/8		28.1, 28.2, 28.3, 28.4, 28.5, 28.6, 28.7, 28.8	inferred by homology
KAP 29	HS	17q21.2	0/1			inferred by homology
Total			42/103			

***Table V** : Summary of the KAP gene expression identified at the protein level in hair cortex.*

Regarding other human KAP from cuticle, data from Lee et al. suggested protein existence of the 12 HS KAP 10 with low sequence coverage and of 3/4 HS KAP 12. We have also shown protein identification of KAP 24-1 and 26-1 in their samples. In these samples, none of the 11 or 12 UHS KAP 5 were found in spite of several sites of potential trypsin cleavage on lysine and a high expression of KAP 5 shown at the transcript level in the upper cuticle [77].

5. Perspectives

The present study shows agreement between protein expression of the majority of cortical KAP previously described on follicle transcriptome and it provides new information on their abundance in mature hair macrofibrils. Suggestion of highest expression for several cortical KAP could lead to focus attention on these families to understand KAP structural functions and nature of matrix/microfibril interactions. Using the list of proteotypic peptides obtained in this study could be a perspective for absolute quantitation of KIF and KAP components to determine their stoichiometry in the structure and bring new hypothesis about arrangement mechanisms during keratinization process.

Variant analysis on keratins and KAP shows the polymorphic character of these proteins and the potential to correlate several genomic data with our proteomic strategy. Substitution variant could easily be described on keratins and the ability to detect described variant implicating monilethrix diseases [99] could be imagined by direct hair protein sample analysis.
Previous studies on cortical KAP 4 and 1 families at the transcript level have shown polyallelism for a majority of these genes with substitutions and insertion/deletion which lead to size polymorphism [74, 75, 256, 258]. Influence of these variations, particularly size polymorphisms, could help to presume matrix and macrofibril mechanical properties. Nevertheless, evidence of this polymorphism type is compromised by our strategy without appropriate protein variant databases describing individual genomes used in this study. Getting this information stays as an improvement perspective into MS-based proteomics pipelines [177]. Using data from recent studies on human multi genome sequencing [221] seems to be the better way to detect and study the impact of these variations on the polymorphic nature of human hair.

We have shown the importance of using previous transcriptional works to describe purity of hair extract according to protein expression specificity and abundance measurement. Using this information allows cuticle, cortex, medulla and epidermal content in hair extract to be analyzed which is necessary to describe protein expression in these different cell types.

Comparison with extract analysis reported by Lee et al. showed that the related extract contained 60% cuticle keratins substantially contaminated with major cortical keratins and KAP and a very low content of medulla proteins. This observation could temper Lee et al. conclusion and restricts potential crosslinked proteins to cuticular specific proteins. High proportion of cortex in hair compared to cuticle and medulla could easily explain difficulties to isolate pure cuticle or medulla with current extraction procedures. This phenomenon seems also critical for protein extraction from wool whose recent proteome analysis could suggest co-existence of cortical and cuticular IFPs and KAP in wool cuticular and soluble extracts [96, 98, 150]. Future proteomic studies on hair and particularly on cuticle and medulla might require specific technique extraction procedures to unambiguously characterize keratin and refine human KAP catalogue in these differentiated cells.

6. Résultats d'identifications supplémentaires réalisées en analyse « Shotgun » du protéome cortical

En complément de l'étude décrite précédemment, nous présentons succinctement des résultats de spectres obtenus en supplément et en parallèle sur d'autres analyses d'échantillons corticaux.

a) Evidence de l'expression de la KAP 2.4

Des analyses, ultérieures à l'étude présentée dans ce chapitre, ont été réalisées sur des extraits fractionnés de cortex analysés sur le couplage nanoAcquity-MaXis. Une identification notable a été réalisée en supplément des listes de protéines établies précédemment. L'identification d'un peptide trypsique et protéotypique de la KAP 2.4 (référence SwissProt Q9BYR9) a été réalisée. Elle permet de compléter le catalogue des KAP identifiées avec un gène exprimé dans le protéome du cortex de la famille des KAP 2 (Figure 8).

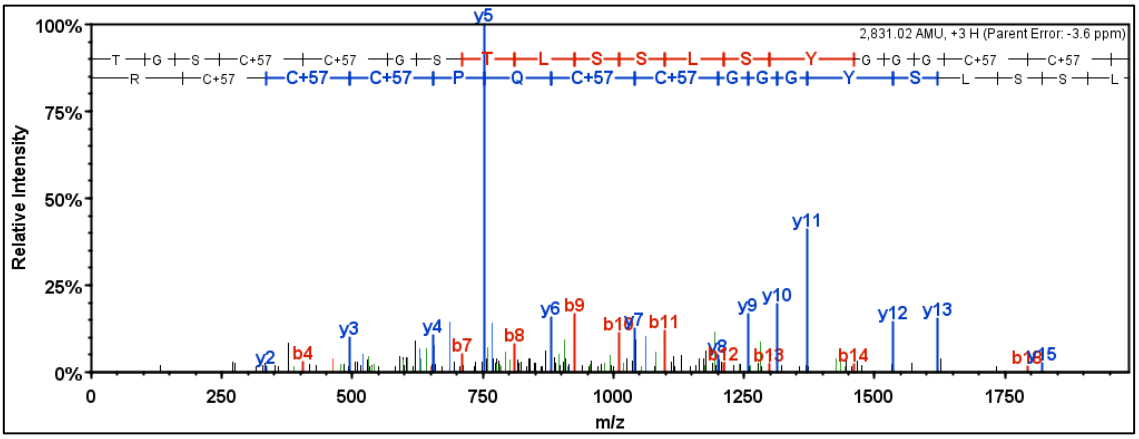


***Figure 8** : Spectre de fragmentation permettant de démontrer l'expression de la KAP 2.4 par identification du peptide T(acétyl)GSCamCamGSTLSSLSYGGGCamCamQPCamCamCamR.*

b) Détection de sites de phosphorylation sur les kératines de type II

Un examen des données issues des recherches en mode tolérant aux erreurs a suggéré la présence de peptides modifiés portant des phosphosérines sur les kératines du cortex. Les phosphorylations sont des modifications post traductionnelles très bien décrites pour les kératines du cytosquelette qui se retrouvent localisées

exclusivement dans les domaines tête et queue [132, 259]. Une série de spectres de fragmentation, caractérisés par la présence de fragments présentant des pertes de masse de -98 Da (perte d'acide phosphorique), sont retrouvés et correspondent à des peptides phosphorylés des kératines de type II, K85, K86, K83 et K81.

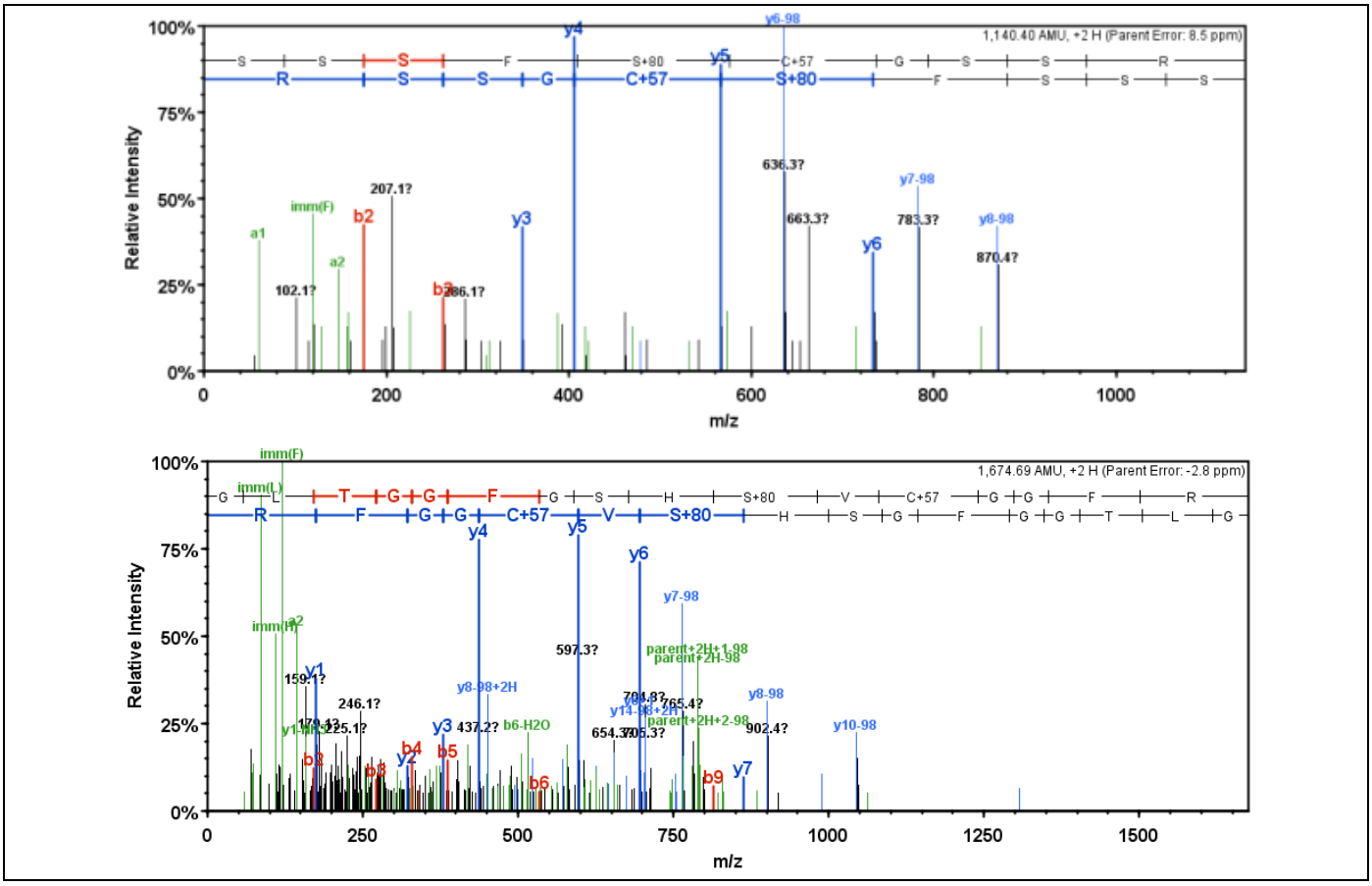


Figure 9 *: Spectres de fragmentation de peptides phosphorylés SSSFSCamGSSR (situé sur le segment queue de K85) et GLTGGFGSHSVCamGGFR (situé sur le segment tête communément de K81, K83, K86) et retrouvés dans le protéome cortical.*

Des précédentes études, menées sur les kératines de la laine et analysées par électrophorèse bidimensionnelle différentielle avec et sans traitement à la phosphatase alcaline, avaient déjà permis de mettre en évidence de telles modifications sur les kératines de type II [260, 261]. Nos résultats protéomiques montrent que la phosphorylation est retrouvée sur les kératines de type II du cortex du cheveu humain sur sept sites portant partiellement des phosphosérines et précisent leur localisation sur différentes sérines des domaines tête et queue de ces protéines (Figure 9 et Annexe 2). Les kératines de type I ne semblent pas être affectées par cette modification puisqu'aucune phosphorylation n'y est retrouvée.

Cette présence de phosphorylations, spécifique aux kératines de type II et exclusivement localisée sur les domaines tête et queue, peut laisser supposer des possibilités de régulation des mécanismes d'interactions entre les kératines des cheveux. L'introduction d'un groupement phosphate anionique dans ces domaines majoritairement basiques et hydrophobes, pourrait permettre des associations contribuant à favoriser la croissance des filaments.

c) Détection de modifications chimiques supplémentaires

Dans la première partie de ce manuscrit, nous avons décrit les modifications chimiques des protéines pouvant être attendues suite à l'exposition de la fibre à différentes conditions environnementales ou à des traitements (Chapitre III 6.e). Les analyses du protéome du cortex précédemment décrites ont permis l'identification de trioxydations des cystéines pouvant être attribuées à des modifications des protéines de la fibre avant leur

extraction. La déamidation des kératines et des KAP a également été détectée, bien qu'il ne soit en revanche pas exclu que ces modifications puissent être en partie une conséquence de la manipulation de l'échantillon pendant son analyse.

L'examen de résultats obtenus en mode tolérant aux erreurs montre que d'autres types de modifications chimiques des peptides peuvent être envisagés sur les kératines de type I et II. Parmi les spectres de fragmentation obtenus, nous retrouvons des résidus issus de l'oxydation de résidus aromatiques comme le tryptophane (hydroxytryptophane, kynurenine, formylkynurenine), la tyrosine (dihydroxyphenylalanine) et de l'histidine (hydroxyhistidine). Nous n'avons pas détecté de modification sur la phénylalanine. La méthionine, dont l'oxydation est une modification communément observée en protéomique, est également retrouvée minoritairement modifiée sous forme de sulfone (dioxydation) et d'homosérine. La modification de la cystéine en déhydroalanine, intermédiaire avant la formation de ponts interprotéines comme la lanthionine, a été observée (Annexe 3).

La présence de ces modifications pose la question de leur impact sur les propriétés mécaniques du cheveu, bien que la majorité de ces résidus soient relativement minoritaires dans le cortex. Le suivi quantitatif de ces modifications pourrait, par la suite, constituer une donnée supplémentaire à celle obtenue par la quantification de l'acide cystéique (couramment dosés sur les hydrolysats de cortex) pour le suivi de l'impact des traitements cosmétiques sur les protéines de la fibre.

Nous pouvons également noter que la partie N-terminale des kératines K31 et K33b, commençant par une proline que nous avons décrite comme très majoritairement non acétylée, présente des modifications de type formylation et succinylation.
Une modification d'une arginine de kératine de type I par le méthylglyoxal est également détectée.

7. Détection de sites de ruptures au sein des segments tiges des kératines par électrophorèse bidimensionnelle

L'approche Shotgun pour l'analyse du cortex nous a permis d'apporter un nombre exhaustif d'informations des protéines présentes, de leurs polymorphismes et de leurs modifications. Néanmoins, l'utilisation de l'électrophorèse bidimensionnelle sur gel (2D-GE) nous a également permis d'accéder à des informations supplémentaires relatives aux propriétés des kératines vis-à-vis du stress mécanique.

Dans une étude s'inscrivant dans la continuité des travaux d'Audrey Bednarczyck [199], nous décrivons l'analyse d'un extrait de cortex par 2D-GE. La combinaison de l'analyse protéomique des spots obtenus, complétées d'un examen des signaux LC-MS correspondants, nous a conduits à mettre en évidence des zones de ruptures préférentielles au sein des segments tige des kératines de type II. Cette étude, associée à certains résultats obtenus en analyse Shotgun et présentés dans le manuscrit d'Audrey Bednarczyck, a fait l'objet de la rédaction d'un article soumis dans le journal Analytical Biochemistry. L'article correspondant est présenté en Annexe 4 de ce manuscrit.

8. Les approches « Label free » pour le suivi de l'impact des traitements cosmétiques sur les protéines du cortex

La détection de modifications chimiques réalisée lors des analyses Shotgun nous a conduits à nous intéresser de plus près aux possibilités du suivi des signaux MS correspondants. Une fois que les peptides modifiés ont été identifiés sur un couplage LC-MS/MS, il est possible, sur la base des informations de leurs temps de rétention et de leurs rapports m/z, d'extraire leurs courants d'ions respectifs. La comparaison des intensités de ces courants

d'ions obtenus à partir d'échantillons différents permet alors de comparer ces derniers. Les différences d'intensité de signal peuvent alors être attribuées aux différences existant entre les échantillons analysés. Cette stratégie suppose l'utilisation d'un système chromatographique robuste et résolutif et d'un spectromètre de masse permettant de bénéficier d'une très bonne précision de mesure.

Nous avons évalué cette approche pour la comparaison d'un échantillon ayant subi différents traitement cosmétiques. L'instrumentation utilisée est un couplage nanoLC-QTOF (nanoAcquity-MaXis) dont la capacité de séparation chromatographique a été optimisée en utilisant la colonne la plus résolutive du laboratoire (250 mm x 75 µm 1,7 µm) et en travaillant à un temps de gradient permettant le meilleur compromis entre capacité de pic et temps de chromatographie. Les échantillons traités et le témoin non traité ont préalablement été analysés sur le même système en nanoLC-MS/MS, chaque échantillon étant préalablement fractionné comme décrit précédemment. Des listes de composés ont ainsi été obtenues avec, parmi eux, des peptides portant des modifications chimiques. Les échantillons sont par la suite analysés en LC-MS sans préfractionnement préalable. Les signaux des peptides modifiés et non modifiés peuvent simultanément être extraits et comparés entre les différents échantillons.

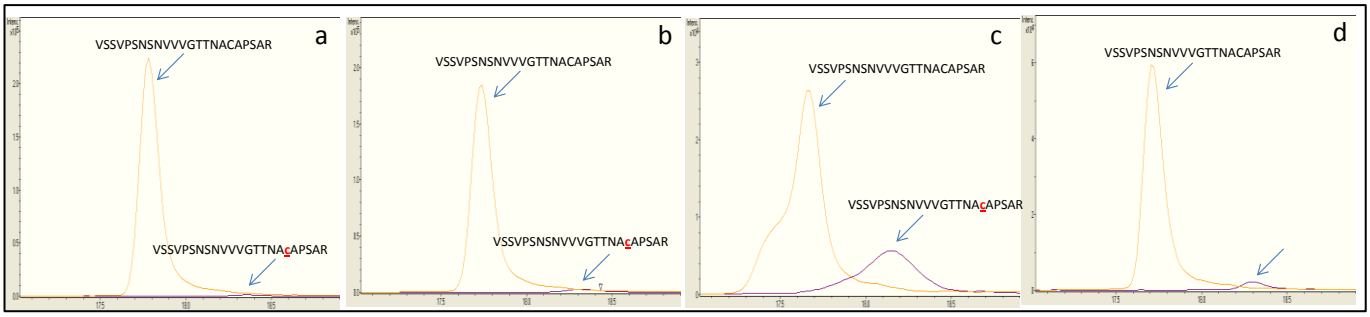


Figure 10 : *Courants d'ions extraits pour le peptide de K86 VSSVPSNSNVVVGTTNACAPSAR avec et sans trioxydation de la cystéine. A : fibre non traitée ; b : fibre permanentée ; c : fibre décolorée ; d : fibre colorée.*

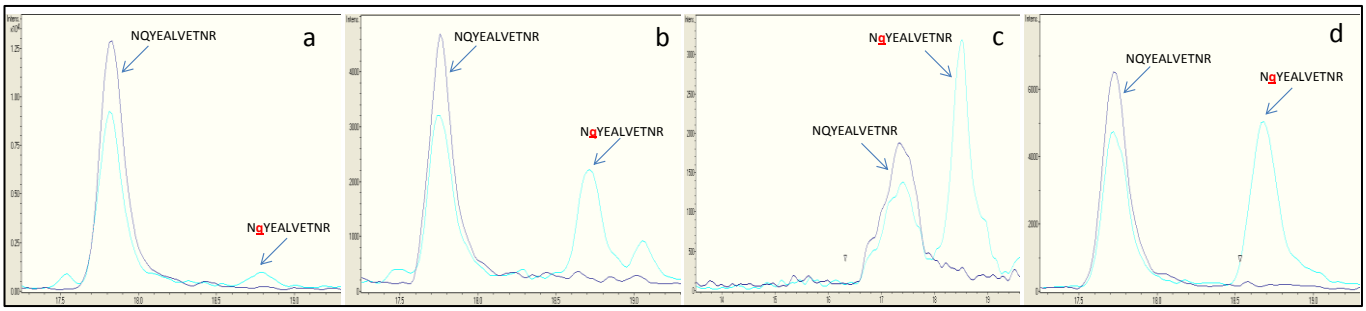


Figure 11 : *Courants d'ions extraits pour le peptide de K33b NQYEALVETNR avec et sans déamidation de la glutamine ou des asparagines. A : fibre non traitée ; b : fibre permanentée ; c : fibre décolorée ; d : fibre colorée.*

Les résultats obtenus (Figure 10 et 11) montrent par exemple que le traitement de permanente (réduction au thioglycolate d'ammonium) n'entraine pas d'oxydation irréversible et prononcée des cystéines de la fibre mais conduit par contre à de la déamidation. La décoloration (oxydation) en revanche présente une importante trioxydation des cystéines et un important effet de déamidation. La coloration a un léger effet d'oxydation et un effet important de déamidation. Nous pourrons envisager par la suite une évaluation plus précise des modifications par intégration des aires des pics chromatographiques correspondants.

A notre connaissance, c'est la première fois qu'est évalué l'impact de traitements cosmétiques sur la déamidation des kératines. Les suivis d'impact de traitements sont habituellement réalisés sur les hydrolysats de cheveux, la technique ne permettant pas le suivi de ces modifications sur des fonctions labiles.

Nous pouvons noter que des études de cinétique de la dégradation des résidus en fonction du temps de traitement pourraient peut être apporter des informations structurales. En effet, les résidus des protéines

pourraient ne pas être exposés de la même façon aux réactifs en fonction de leur localisation au sein de la structure. Nous pouvons par exemple supposer que les résidus se trouvant au sein de la matrice interfilamentaire hydrophobe ont moins de chance d'être modifiés que des résidus se trouvant à la surface des microfibrilles hydrophiles.

Chapitre II Etude du protéome des cellules cuticulaires

Après avoir étudié et développé des stratégies d'analyse adaptées à l'analyse du cortex, nous nous sommes intéressés à l'analyse protéomique de la cuticule. Cette problématique s'avère plus ardue que l'étude du cortex comptes tenus de la difficulté à séparer spécifiquement les cellules cuticulaires des cellules corticales et de la grande insolubilité des structures qui la compose. Cette insolubilité est due à la structure de ces cellules mais également à la présence au sein des protéines qui la constitue d'un nombre important de modifications induites par l'activité des transglutaminases. Dans ce contexte, l'étude de la cuticule nécessite de disposer d'outils analytiques supplémentaires en compléments des stratégies de protéomiques. Nous présentons ici ces développements puis l'étude protéomique de la cuticule.

1. Mise en place d'une stratégie d'identification de la modification GGEL

Une modification importante présente dans les cheveux est le pont GGEL établi entre lysine et glutamine suite à l'activité de la transglutaminase. Dans ce contexte, il nous a paru indispensable de disposer d'une technique analytique permettant d'évaluer l'importance de cette modification dans nos échantillons. Elle apparait d'autant plus importante dans le cadre de l'étude de la cuticule où le GGEL est décrit comme particulièrement abondant [22, 53-55].

a) Génération du dipeptide GGEL

La liaison formant le GGEL est une liaison amide qui s'hydrolyse en milieu acide comme les liaisons peptidiques. Les analyses classiques d'acides aminés utilisant l'hydrolyse acide à chaud pour dégrader les protéines et étudier les résidus détruisent donc également la liaison GGEL. Il est ainsi nécessaire d'envisager une méthode permettant de couper la chaine peptidique en conservant l'intégrité du pont.

Dans ce cas, la digestion enzymatique peut dans ce cas être utilisée. Il existe parmi la palette des enzymes à disposition du biochimiste des enzymes aspécifiques permettant de dégrader les chaines peptidiques. Ces enzymes ne permettent pas d'hydrolyser la liaison GGEL du fait de sa conformation. Il est à noter que les destabilases, enzymes pouvant être extraites des organismes comme les sangsues [262-266], permettent spécifiquement de réaliser cette coupure mais ne sont pas disponibles commercialement.

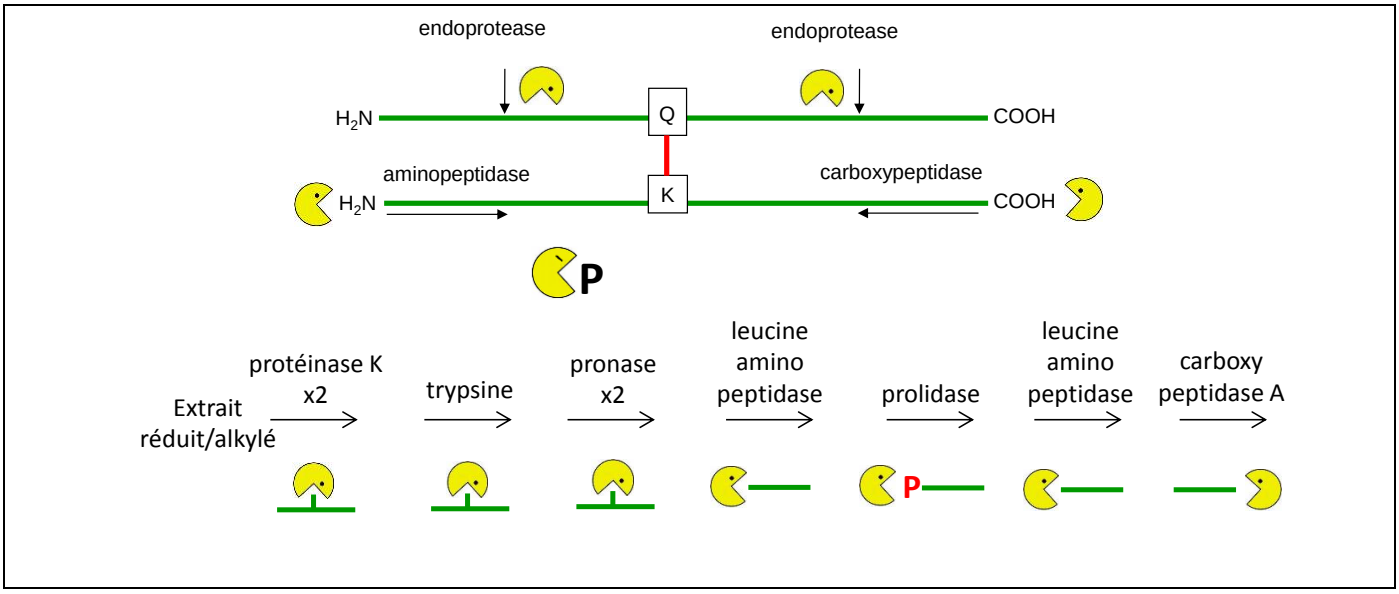


***Figure 1** : Principe d'hydrolyse de la chaine peptidique pour la génération de dipeptides GGEL par l'utilisation d'un « cocktail » enzymatique et combinaison des enzymes utilisée dans le cadre de notre étude.*

Des méthodes combinant différentes digestions enzymatiques sont décrites dans la littérature et peuvent être envisagées afin de générer le GGEL à partir d'extraits protéiques pontés [267]. L'utilisation d'endoprotéases aspécifiques (pronase, protéinase K, papaïne…) combinée à l'utilisation d'aminopeptidases et de carboxypeptidases peut permettre une digestion quasi complète de l'ensemble des liaisons peptidiques de l'extrait protéique. Compte tenu de l'inactivité des exopeptidases sur la proline, la prolidase peut être utilisée pour hydrolyser la liaison peptidique de ce résidu entre deux digestions utilisant une exopeptidase.

***Figure 2** : Illustration de l'activité de digestion de la protéinase K sur des fibres de cheveu dont le cortex a été préalablement extrait.*

La protéinase K a été choisie comme première enzyme de digestion. Extraite d'un champignon, *Tritirachium album*, elle doit son nom à sa capacité à digérer la kératine native et semble appropriée à la problématique [268]. L'activité de cette enzyme a été évaluée, le résultat montrant effectivement sa capacité à digérer la fibre.

Le principal inconvénient de cette méthode est le temps de digestion totale et le nombre de manipulations de l'échantillon nécessaires pour obtenir le protéolysat. La digestion s'étale sur une dizaine de jours. Des travaux sont actuellement en cours pour optimiser le temps de digestion en augmentant son efficacité par assistance micro onde de la digestion [269].

b) Analyses du dipeptide GGEL dans les digests enzymatiques

L'analyse des extraits ainsi générés peut être réalisée à l'aide de systèmes d'analyses classiques d'acides aminés. Nous disposons d'un analyseur d'acides aminés fonctionnant après dérivation en échange de cation et détection par fluorescence. Nous avons néanmoins souhaité développer un couplage permettant l'analyse des acides aminés par spectrométrie de masse en évaluant l'utilisation d'un mode chromatographique compatible avec la spectrométrie de masse et ne nécessitant pas de réaction de dérivation des acides aminés. Dans ce cadre, nous avons choisi d'évaluer la chromatographie d'interactions hydrophiles (HILIC).

Développement d'une approche d'analyse LC-MS d'analyse d'acides aminés par chromatographie d'interaction hydrophile

Les techniques de chromatographie employées pour l'analyse des acides aminés sont nombreuses mais l'utilisation de la spectrométrie de masse pour réaliser leur analyse se limite à quelques techniques. Nous pouvons citer les stratégies de dérivation suivies de l'analyse en chromatographie en phase gazeuse ou liquide en phase inverse.
Le mode chromatographique HILIC qui consiste à retenir les analytes sur des phases stationnaires fonctionnalisées de groupes polaires s'est considérablement développé en moins d'une décennie pour la séparation et l'analyse par spectrométrie de masse de composés polaires. Les constructeurs proposent désormais une palette de phases stationnaires qui peuvent être envisagées entre autres pour l'analyse des acides aminés.

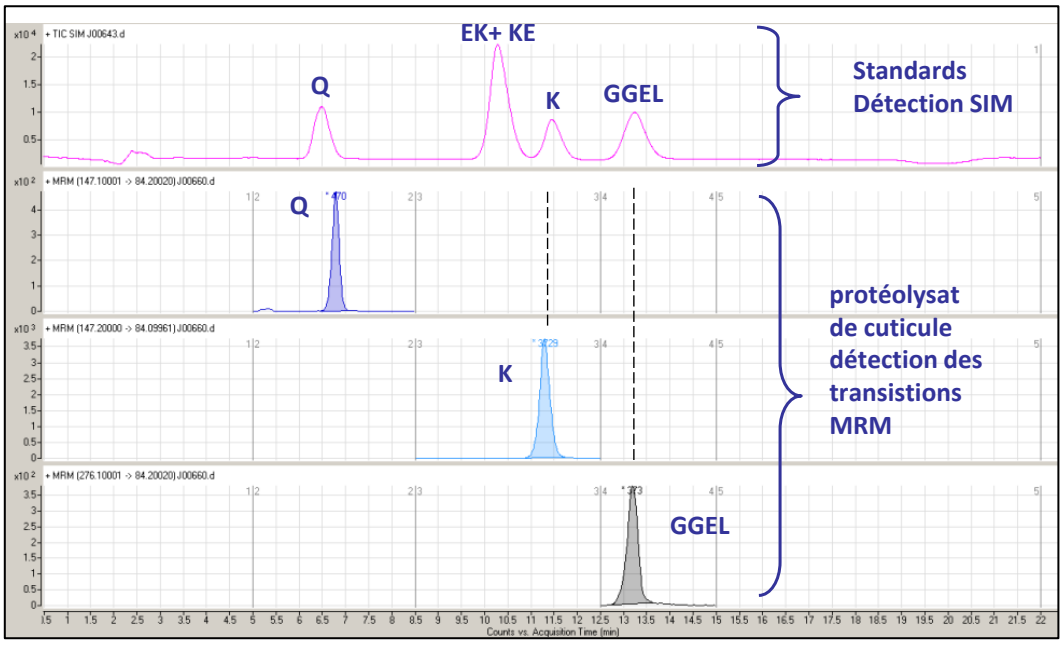


Figure 3 *: Séparation d'un mélange standard dans une problématique de suivi de la modification GGEL et détection de la glutamine, de la lysine et du GGEL dans un protéolysat de cuticule. Colonne TSKgel Amide 80, 25 cm x 2,0 mm, 5 µm. Voie A H2O, voie B ACN. Gradient de 10 à 60 % de A en 15 min. Débit 200 µL/min.*

Nous avons dans ce cadre évalué la possibilité de chromatographier les hydrolysats enzymatiques obtenus avec une colonne remplie d'une phase greffée avec un alkyle fonctionnalisé d'un groupement polaire amide. L'élution des composés retenus sur ce type de phase est réalisée par augmentation de la proportion d'eau par rapport à celle de solvant organique (ici l'acétonitrile). Les premiers essais se sont révélés concluants et nous ont permis d'envisager une méthode de séparation permettant de séparer les deux résidus substrats de la transglutaminase (glutamine et lysine) et le dipeptide GGEL tout en réalisant leur détection en couplage avec un spectromètre de masse de type triple quadripôle. Afin de s'assurer que le signal du dipeptide ne puisse pas être interféré par des composés isobares, nous avons contrôlé que d'autres dipeptides de même masse (EK et KE, MH^+=276.1) et possédant des fragments similaires au dipeptide GGEL étaient élués à des temps différents (Figure 3). Pour chaque composé analysé, une méthode de fragmentation MS/MS a été développée afin de détecter les composés grâce à leurs transitions parents / fragments spécifiques.

Cette méthode a été utilisée pour évaluer la présence du GGEL dans un protéolysat de cuticule et montre également la présence de glutamine. Cette méthode rend possible l'évaluation dans l'échantillon du pourcentage de glutamine ou de lysine restant libres par rapport aux résidus pontés.

Au cours de l'évaluation de cette méthode, la colonne a été soumise à une phase mobile contenant de l'acide formique. Les conséquences de ce traitement ont été une modification de la sélectivité de la séparation avec une augmentation générale de la rétention des composés. Ce phénomène nous a suggéré que la robustesse de la méthode devait être améliorée en tamponnant le pH de la phase mobile.

L'utilisation du couplage LC-MS n'a pas été actuellement poursuivie dans la suite de nos études mais les perspectives que pourrait apporter une telle méthode pour la recherche et la quantification de modifications particulières dans les protéolysats de cuticule mais également de cortex doivent être envisagées. Cette technique pourrait être particulièrement adaptée à la recherche de modifications labiles susceptibles d'être perdues suite à l'hydrolyse chimique. Elle permettrait également de compléter les données quantitatives d'acides aminés sur la teneur en tryptophane, en glutamine et en asparagine dans le cheveu, ces informations étant perdues par hydrolyse chimique.

Evaluation différentielle de l'abondance du GGEL par chromatographie d'échange d'ions et détection UV

La poursuite de l'étude de la modification GGEL dans les hydrolysats a été réalisée avec un analyseur d'acides aminés à détection UV. L'identification des différents acides aminés avec cette technique est réalisée sur la base des comparaisons des temps de rétention des composés mesurés avec des standards. Ces standards sont par la suite utilisés pour la quantification.
L'identification du GGEL, pour qu'elle soit réalisée sans ambiguïté dans les protéolysats, a été réalisée par la technique dite de dopage de pic. Les échantillons sont analysés deux fois : la première fois, l'échantillon seul est analysé ; la seconde, le même échantillon est analysé en présence d'un ajout d'une quantité de standard GGEL. La comparaison des deux chromatogrammes permet de mettre en évidence un pic différentiel correspondant au GGEL ajouté. Si le pic correspondant au GGEL est initialement sur le chromatogramme de l'échantillon seul, alors le composé peut être considéré présent dans l'échantillon aux limites de détection près.

Cette technique nous a ainsi permis de montrer que le protéolysat de cortex ne contenait pas de GGEL contrairement à celui de cuticule dans lequel le dipeptide est relativement abondant.

La méthode d'analyse a été par la suite utilisée pour évaluer d'éventuelles différences entre deux échantillons de cheveux décortexés provenant d'individus différents. Le premier échantillon est issu d'un mélange de cheveux d'individus caucasiens et l'autre issu d'un mélange de cheveux d'individus chinois. Les deux échantillons de cheveux ont été extraits de leur cortex dans les mêmes conditions puis digérés en parallèles. Pour chaque échantillon, une répétition du protocole de digestion a été réalisée.

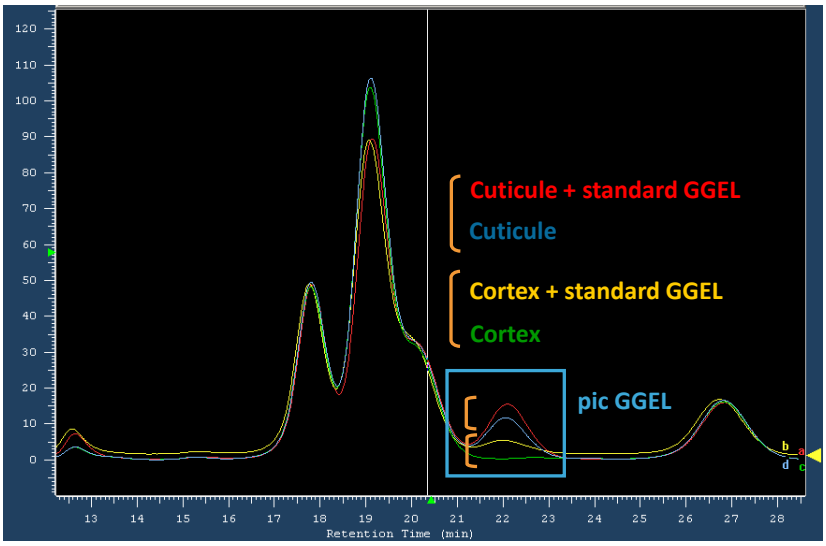


Figure 4 *: Résultats d'identification du GGEL dans les protéolysats de cuticule et de cortex par méthode de dopage de pic avec un standard (le terme cuticule correspond ici à l'insoluble restant après extraction extensive du cortex)*

Les différents protéolysats obtenus ont par la suite été analysés. Les résultats de cette étude montrent une différence de proportion de GGEL entre les deux protéolysats. Ils montrent également que d'autres acides aminés ont une proportion différentielle dans les deux protéolysats suggérant de légères différences de compositions protéiques. La différence pourrait s'expliquer par une expression du GGEL plus importante dans l'échantillon chinois. Elle pourrait également être expliquée par une différence de contamination de la cuticule par le cortex plus important dans l'échantillon caucasien, favorisant la diminution de la proportion de GGEL dans l'échantillon. Bien que les échantillons aient été extraits en parallèles, il n'est pas exclu que des différences de géométrie de fibre entre les deux populations soient à l'origine de différences d'extraction. Par comparaison des quantités dosées de lysine et de GGEL, nous pouvons déduire qu'il y a entre 2,5 et 3,8 fois plus de lysine que de GGEL dans ces échantillons ce qui permet d'estimer qu'environ une lysine sur quatre est pontée.

Le contrôle de l'extraction et l'estimation du niveau de contamination n'ayant pas été évalué sur les deux échantillons, nous resterons donc relativement prudents quant à l'interprétation qui peut être faite de ces résultats. Une confrontation de ces résultats avec des observations microscopiques de coupes de cheveux représentatives des deux échantillons et notamment des structures cuticulaires mériterait d'être réalisée.

	zelaca cuticule Z1 LSMBO **(1)**	zelaca cuticule Z2 LSMBO **(2)**	Moyenne (1) et (2) **(5)**	Ecart (1) et (2) **(6)**	chinois cuticule C1 LSMBO **(3)**	chinois cuticule C2 LSMBO **(4)**	Moyenne (3) et (4) **(7)**	Ecart (3) et (4) **(8)**	Ecart à la moyenne (5)-(7) **(9)**	Moyenne des écarts (6) et (8) **(10)**	F= (9)/(10)
Glutathion réduit	1,13	1,17	**1,15**	**0,04**	1,24	1,15	**1,20**	**0,09**	**-0,05**	**0,06**	**-0,78**
Ac. aspartique	3,31	3,51	**3,41**	**0,20**	3,48	3,35	**3,41**	**0,13**	**0,00**	**0,16**	**-0,01**
Thréonine	8,30	8,58	**8,44**	**0,28**	7,64	7,73	**7,69**	**0,09**	**0,75**	**0,19**	4,03
Sérine	12,64	12,94	**12,79**	**0,29**	14,20	13,82	**14,01**	**0,38**	**-1,22**	**0,34**	-3,62
Asparagine	3,97	4,03	**4,00**	**0,06**	3,69	3,73	**3,71**	**0,04**	**0,29**	**0,05**	5,74
Ac. glutamique	7,25	7,47	**7,36**	**0,22**	6,76	7,26	**7,01**	**0,50**	**0,35**	**0,36**	**0,97**
Glutamine	0,65	0,61	**0,63**	**0,04**	0,66	0,66	**0,66**	**0,00**	**-0,03**	**0,02**	**-1,44**
Proline	1,31	1,41	**1,36**	**0,10**	1,58	1,26	**1,42**	**0,31**	**-0,06**	**0,21**	**-0,27**
Glycine	4,63	4,72	**4,67**	**0,09**	5,39	5,04	**5,21**	**0,34**	**-0,54**	**0,22**	**-2,52**
Alanine	4,78	4,95	**4,87**	**0,17**	5,40	5,29	**5,35**	**0,11**	**-0,48**	**0,14**	-3,49
Valine	6,66	6,80	**6,73**	**0,14**	7,04	6,97	**7,01**	**0,07**	**-0,27**	**0,11**	**-2,52**
Cystine	0,85	0,67	**0,76**	**0,17**	0,34	0,31	**0,33**	**0,03**	**0,43**	**0,10**	4,27
Méthionine	0,57	0,59	**0,58**	**0,03**	0,42	0,28	**0,35**	**0,14**	**0,23**	**0,09**	**2,70**
Isoleucine	4,04	4,10	**4,07**	**0,06**	4,06	4,14	**4,10**	**0,08**	**-0,03**	**0,07**	**-0,40**
Leucine	8,74	8,91	**8,82**	**0,17**	8,85	9,21	**9,03**	**0,36**	**-0,21**	**0,27**	**-0,79**
Tyrosine	5,79	5,70	**5,74**	**0,10**	6,33	6,16	**6,25**	**0,17**	**-0,50**	**0,13**	-3,79
Phénylalanine	3,13	3,07	**3,10**	**0,06**	3,10	3,16	**3,13**	**0,06**	**-0,03**	**0,06**	**-0,53**
Tryptophane	8,46	6,88	**7,67**	**1,58**	6,01	6,42	**6,22**	**0,41**	**1,45**	**1,00**	**1,46**
Ethanolamine	0,09	0,08	**0,08**	**0,01**	0,05	0,07	**0,06**	**0,02**	**0,02**	**0,01**	**1,70**
Lysine	3,21	3,24	**3,23**	**0,03**	3,24	3,28	**3,26**	**0,04**	**-0,03**	**0,04**	**-0,89**
Histidine	1,16	1,11	**1,14**	**0,05**	1,15	1,12	**1,13**	**0,03**	**0,00**	**0,04**	**0,06**
Arginine	7,73	8,00	**7,87**	**-0,27**	6,94	7,20	**7,07**	**-0,25**	**0,80**	**-0,26**	-3,04
GGEL	1,61	1,47	**1,54**	**0,14**	2,42	2,37	**2,39**	**0,05**	**-0,85**	**0,09**	-9,16

***Tableau 1** : Résultats de l'analyse d'acides aminés, exprimée en g /100 g de protéolysat, incluant la mesure GGEL des protéolysats de deux échantillons de cheveux enrichis en cuticule et provenant de deux populations distinctes (zelaca = cheveu de référence ; chinois = échantillon constitué du mélange de cheveux de 30 individus chinois). Les différences pour chaque acide aminé ont été jugées significatives lorsque les écarts entre les moyennes mesurées étaient 3 fois supérieurs à la moyenne des valeurs absolues des écarts des répétitions (facteur F).*

Perspectives de l'étude de la modification GGEL

L'étude du réseau GGEL par la quantification et la localisation de son expression dans la cuticule entre des échantillons de cheveux de natures différentes peut être envisagée aux vues des stratégies analytiques développées. Ces outils et notamment l'étape de digestion sont actuellement en cours d'optimisation. Si la cuticule ne représente qu'une faible contribution dans la structure de la fibre, sa fonction de protection mérite tout de même d'être étudiée si d'éventuelles différences de résistances mécaniques existent entre les populations. La résistance de ce pont aux traitements cosmétiques mérite également d'être évaluée.

D'un point de vue de la compréhension de la biologie du cheveu, la présence de ce réseau de liaisons non réductibles permet d'envisager la recherche des substrats protéiques ainsi pontés comme cela a été réalisé dans le *stratum corneum* de l'épiderme [153]. La connaissance de ces substrats et la localisation des modifications dans leur séquence pourrait être particulièrement informative puisqu'elle apporterait des éléments relatifs à l'organisation de l'édifice moléculaire. Un modèle d'organisation de ce type cellulaire pourrait ainsi être établi dans la cuticule comme il l'a été pour le *stratum corneum*.

La recherche de ce type de modifications directement dans le milieu biologique nécessite d'isoler un extrait suffisamment enrichi des protéines pontées (généralement insolubles en conditions réductrices). L'abondance de la modification dans l'extrait doit idéalement être confirmée dans l'échantillon par une technique permettant sa

détection et sa quantification. Cette étape peut être réalisée par analyse du pont dans le protéolysat ou par test Elisa bien que de l'aspécificité ait été décrite pour l'anticorps monoclonal utilisé contre le GGEL [270].

L'identification des peptides porteurs de ces modifications est difficile et nécessite de trouver une stratégie de digestion permettant de les générer et suppose des sites de coupures adéquats. La caractérisation des peptides pontés dans un digest nécessite également de les séparer des peptides non pontés afin de décomplexifier l'échantillon et de gagner en sensibilité. La physico-chimie particulière de ces peptides pontés leur confère deux fonctions N-terminales et deux fonctions C-terminales pouvant être utilisées pour les retenir de manière plus importante sur des résines échangeuses d'ions. Ces peptides sont potentiellement plus gros que les peptides non pontés ce qui peut leur conférer une plus grande hydrophobicité utilisable en chromatographie en phase inverse [152, 153]. Par ailleurs, nous noterons que l'immunocapture consistant à complexer les peptides pontés avec un anticorps puis à séparer les des peptides non complexés par filtration sur membrane a été utilisée avec succès pour isoler des peptides pontés [271].

Le séquençage des peptides pontés a pu être réalisé par séquençage d'Edman des peptides isolés [152, 153]. L'inconvénient de cette technique réside dans la nécessité de reconstituer les séquences en envisageant les différentes combinaisons de séquences possibles, deux acides aminés étant identifiés à chaque étape de séquençage. La localisation de la position du pont est néanmoins facilitée par la détection d'un signal spécifique. Ce type de séquençage est réalisé à bas débit et l'identification des séquences est difficile voire ambiguë [272].

La spectrométrie de masse en tandem pour le séquençage de ces peptides pontés apparait comme une technique alternative, leurs spectres de fragmentation pouvant être utilisés pour les caractériser. La dérivation différentielle des extrémités terminales des dipeptides puis l'analyse différentielles par spectrométrie de masse des échantillons dérivés et non dérivés peut permettre de distinguer les espèces porteuses de deux motifs de dérivation. Ces peptides potentiellement pontés peuvent alors être sélectionnés pour la fragmentation. La caractérisation des spectres de fragmentation n'est pas possible directement avec les outils classiques d'identification protéomique. Ces peptides étant constitués de deux chaines peptidiques liées, leurs spectres de fragmentation ne peut pas correspondre avec des spectres théoriques classiques calculés *in silico* par les algorithmes de recherche couramment utilisés. Des moteurs adaptés à cette problématique sont développés dans le cadre des recherches d'informations structurales sur des protéines pontés chimiquement *in vitro.* Mais une caractérisation automatique des peptides pontés implique généralement de connaitre préalablement la ou les protéines substrats [273, 274]. L'interprétation des spectres de fragmentation est elle-même compliquée, les séries d'ions fragments de chaque chaine se mélangeant simultanément sur le même spectre.

La recherche et la caractérisation des substrats du GGEL dans la cuticule nécessite ainsi de connaitre les protéines potentiellement pontées. L'identification des protéines majoritaires dans les extraits insolubles est donc une étape nécessaire pour permettre de poursuivre cette recherche. Par ailleurs, il semble indispensable de disposer d'un protocole permettant d'obtenir des extraits cuticulaires suffisamment purs.

2. Développement d'une stratégie d'extraction et d'analyse des cellules cuticulaires

L'analyse du protéome de l'insoluble d'extrait de cheveu décrit par Lee et *al*. a montré qu'il était possible d'accéder au protéome de la cuticule humaine. Néanmoins, la technique d'extraction répétée du cheveu semblerait isoler des extraits cuticulaires presque à moitié constitués de cortex. La contamination corticale doit ainsi être diminuée afin de permettre de mieux étudier la cuticule minoritaire par rapport au cortex.

a) Extraction physique de la cuticule

Afin d'améliorer la pureté de l'extrait cuticulaire, nous nous sommes inspirés d'une technique d'isolement de la cuticule développée par Swift et *al*.[50-52]. Elle consiste à extraire physiquement la cuticule sans avoir recours à des agents réducteurs et chaotropiques pouvant parallèlement solubiliser le cortex. L'intégrité de l'empilement des cellules cuticulaires est sensible à la friction. En traitant la fibre pour retirer les lipides liés à sa surface, la susceptibilité de la fibre à la friction est accentuée. En favorisant le frottement des fibres les unes par rapport aux autres, des fragments des tuiles de cuticules sont ainsi décrochées (Figure 5). La friction peut être réalisée en soumettant les fibres à l'agitation dans une émulsion air/eau. Les ratios entre la quantité de fibres, les volumes d'eau et d'air ainsi que le temps et la fréquence d'agitation à employer ont été optimisés par Swift et *al*. dans l'optique d'obtenir un bon rendement d'extraction cuticulaire tout en diminuant le risque de contamination corticale. Le temps ne doit pas être trop long pour ne pas abraser les fibres de cortex situées aux extrémités de chaque capillaire.

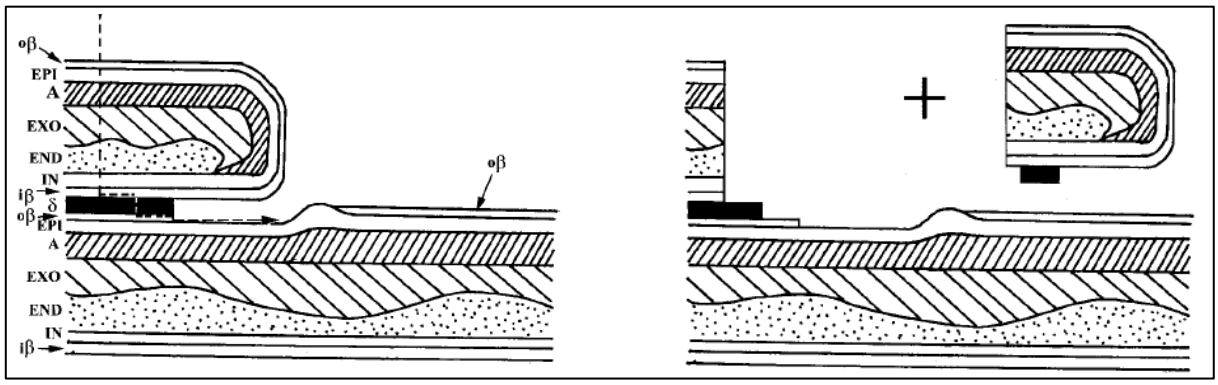


Figure 5 *: Principe de décrochement de fragments de cellules cuticulaires par frictions proposé par Swift [109].*

Le protocole physique de décutilation a ainsi été adapté en reprennant les étapes de délipidation des cheveux, de découpe, de mise en solution et d'agitation à l'aide d'un bras articulé. Après quelques heures d'agitation, une suspension blanche est obtenue en solution. La solution est ensuite séparée des fibres restées intactes puis la suspension est récupérée par centrifugation. L'absence de couleur de cet extrait insoluble suggère l'absence de mélanines pouvant provenir du cortex. Cet extrait a par la suite été utilisé pour les analyses protéomiques de la cuticule.

b) Digestion de la cuticule

Les premiers essais de digestion trypsique sur ces extraits réduits et alkylés d'insoluble ont montré que la trypsine ne parvenait pas visuellement à solubiliser l'extrait cuticulaire. Cette observation est rationelle avec des observations de la digestion de la cuticule suivie par microscopie et montrant que seule l'endocuticule était affectée par l'activité de cette enzyme [50-52]. Nous nous attendons donc à obtenir un protéome principalement enrichi en protéines provenant de cette partie de la cuticule.

Dans une optique d'analyse de l'exocuticule constituant probablement la majorité de la fraction non digérée de cuticule, nous avons choisi en complément des endoprotéases précédemment utilisées comme la trypsine, la chymotrypsine et la GluC d'utiliser une digestion chimique permettant une réaction sur les cystéines et générant des sites des coupures spécifiques [96, 150, 275]. Cette réaction est réalisée en milieu réducteur sans agent d'alkylation des cystéines. Le réactif, l'acide 2-nitro-5-thiocyanobenzoique (NTCB), permet d'effectuer la cyanylation du thiol de la cystéine. En milieu basique, une réaction d'attaque nucléophile d'un ion hydroxyde sur le carbonyle de la liaison peptidique conduit à l'hydrolyse de la liaison peptidique. L'amine générée réagit sur la fonction cyanile qui conduit à une réaction de cyclisation.

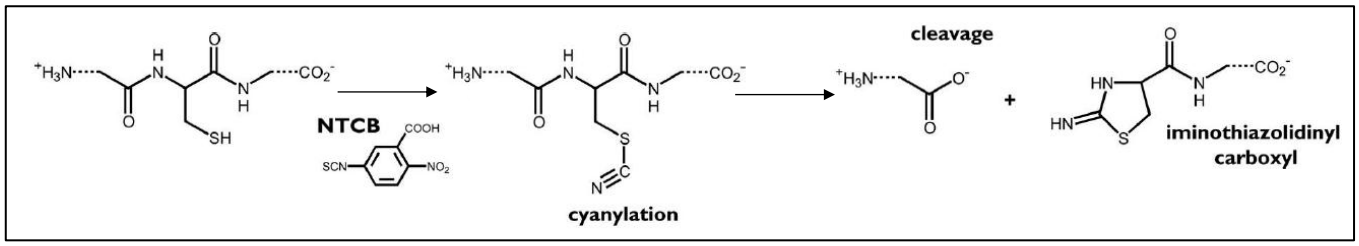


***Figure 6** : Principe de la digestion chimique au NTCB employée comme alternative à la digestion de la cuticule [150]*

Cette digestion a été réalisée en parallèle sur des extraits cuticulaires puis complétée d'une digestion avec une endoprotéase. Lors de cette digestion chimique, nous avons observé que l'extrait insoluble gonflait en comparaison d'un extrait digéré simplement à la trypsine. Ce gonflement peut s'interpréter par une modification du réseau protéique ponté qui, ainsi modifié, est capable d'absober des molécules d'eau. Néanmoins et suite à cette étape, l'insoluble conserve une intégrité apparente probablement grâce au maintien du réseau de liaisons covalentes GGEL. Dans ces conditions, nous pouvons supposer que l'exocuticule puisse être partiellement échantillonée par digestion.

3. Analyses protéomiques des digests de cuticule

a) Méthode

Les digests d'échantillons de cuticule obtenus après extraction physique et digestion ont été analysés en protéomique. La stratégie de multi digestion utilisée pour l'analyse du cortex a de nouveau été employée. 6 digests associant ou non une digestion NTCB à la trypsine, la chymotrypsine ou la GluC ont ainsi été réalisés. Pour cette dernière enzyme, les digestions se sont révélées insatisfaisantes et les résultats n'ont pas par la suite été exploités. Les surnageants obtenus après chaque digest ont été préalablement préfractionnés par LC bidimensionnelle en phase inverse alternant pH basique en première dimension et pH acide en seconde dimension de séparation. La quantité de matériel et le rendement de digestion de la cuticule étant bien moins important que pour le cortex, la totalité du digest à été préfractionnée en première dimension sur un système de µLC (colonne Zorbax extend C18, 300 µm x150 mm 3,5 µm) pour éviter toute dilution superflue de l'échantillon. La totalité de chaque digest a ainsi été injectée en première dimension et il a été possible de réaliser des répétitions des analyses des fractions en seconde dimension couplée avec la spectrométrie de masse. Les conditions d'acquisition sont similaires à celles employées pour l'analyse du cortex. Compte tenu de la quantité de matériel à disposition pour chaque digest, le digest trypsique a été analysé en triplicata, le digest chymotrypsique sans répétition, les digests « NTCB + trypsine » et « NTCB + chymotrypsine » en duplicata.

Les spectres obtenus ont été soumis à la recherche dans une banque Swissprot en version « target/decoy » (téléchargée le 31/01/2011) et complétée de séquences des KAP absentes dans la banque mais décrites dans les travaux de Wu et *al* . [138] et dont les erreurs de séquences concernant les kératines et les KAP du cortex détectées précédemement ont été manuellement corrigées. La recherche a été effectuée uniquement avec le moteur de recherche Mascot avec des ciitères de recherche similaires à ceux utilisés pour l'étude du cortex.

L'utilisation du NTCB a nécessité la configuration de cette coupure *in silico*, complétée de la spécification d'une modification impliquant un écart de masse de + 24,995 Da sur toute cystéine en N-ter du peptide. Une enzyme étant systématiquement utilisée en complément de cette digestion chimique, nous avons ajouté pour chacune de ces conditions les sites de coupures permettant de considérer les coupures NTCB + trypsine (C, R et K sans règle de la proline) et NTCB + chymotrypsine (C, Y, W, F et L). Il convient par ailleurs de rajouter la subsitution variable de la cystéine en dehydroalanine : la réaction de coupure en milieu basique peut conduire également à la désulfuration des cystéines. Dans notre cas, cette réaction secondaire est plutôt favorable puisqu'elle permet d'envisager la génération de peptides qui ne seront pas trop courts et donc plus informatifs.

b) Résultats d'identification

Traitements des résultats d'acquisition et caractérisation de l'extrait

Après validation des résultats d'identification par contrôle du taux de faux positif (fixé à moins de 1 séquence decoy pour 100 protéines pour l'analyse de chaque digest), rassemblement des données d'identification obtenues globalement et élimination des protéines identifiées à un peptide unique avec des spectres de fragmentation jugés ambigus, plus de 340 protéines ont été identifiées dans l'ensemble de ces digests (Annexe 5).

Parmi ces protéines, 20 kératines des filaments intermédiaires et 19 KAP ont été identifiées. Les autres protéines sont des protéines pouvant constituer des organelles du cytoplasme, du noyau mais aussi des membranes, ce qui est tout à fait rationnel avec ce qui était attendu pour un digest d'endocuticule. Nous noterons parmi ces protéines, la présence de deux types de transglutaminase, TGM1 et TGM3.

Parmi les KAP, l'expression de gènes non encore étudiés ou non démontrés au niveau protéique peut être mise en évidence comme nous le détaillerons par la suite.

Pour chaque protéine identifiée, les peptides uniques ont été extraits. Une attention particulière a été accordée à l'assignation des peptides uniques permettant d'identifier les individus des KAP 10 présentant une très forte homologie dans leur séquence. Une valeur semi quantitative basée sur le comptage de spectres provenant de ces peptides uniques a été extraite de la même façon que lors de l'étude du cortex. Les valeurs obtenues ne sont donc pas impactées par l'assignation des peptides partagés par plusieurs protéines isoformes.

Pour chaque protéine identifiée dans plusieurs digests, la valeur semi quantitative de comptage de spectre est calculée en additionnant les valeurs respectives de chacun des digests. Les valeurs correspondant aux kératines et aux KAP ont été extraites, ces dernières pouvant être regroupées en fonction de leur description dans la cuticule mais aussi dans le cortex et l'épiderme. Lorsque la description d'une protéine est attendue simultanément dans la cuticule et le cortex, celle-ci est considérée comme cuticulaire.

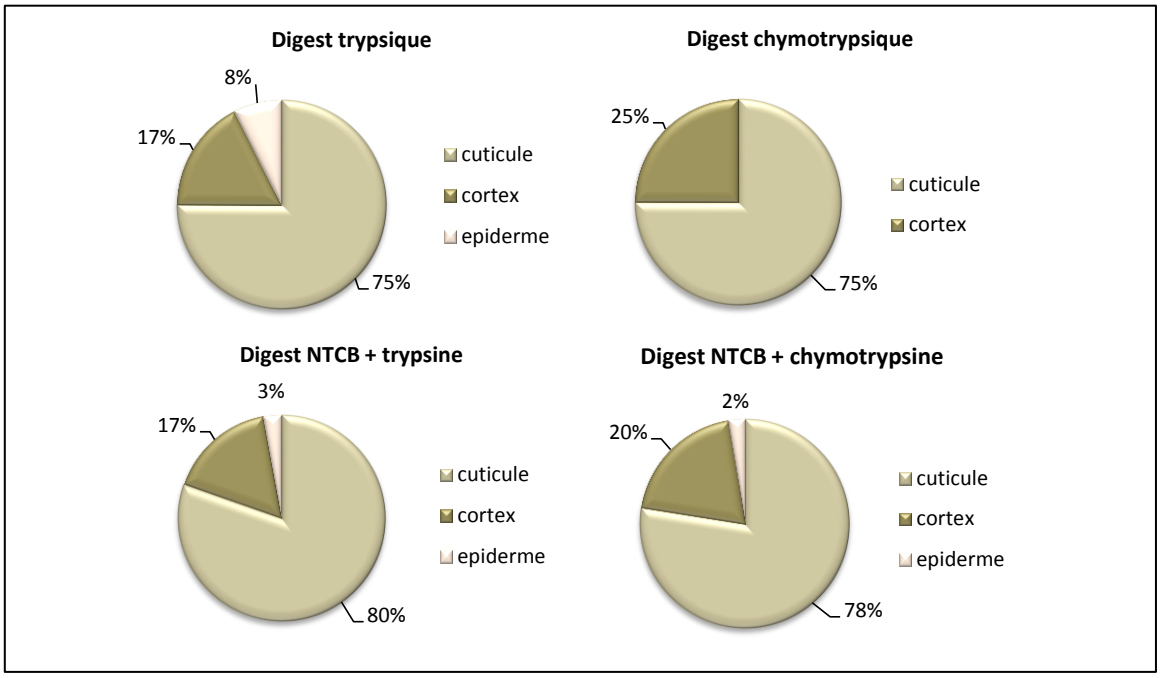


***Figure 7** : Répartition des valeurs de comptage de spectres des kératines et des KAP dans les différents digests en fonction de la prédiction de leur expression dans la cuticule le cortex ou l'épiderme.*

L'estimation du niveau de contamination de l'échantillon par les protéines corticales et épidermales afin d'évaluer l'efficacité de la méthode d'extraction des cuticules a été réalisée pour chaque digest. Les résultats montrent que plus des trois quarts des valeurs de comptage des spectres uniques pour ces protéines proviennent de

composants cuticulaires et cela quelque soit le digest étudié. Nous noterons que les digests réalisés avec le NTCB apportent plus de données cuticulaires notamment grâce à l'obtention de peptides uniques provenant de la famille des KAP 10. Nous pouvons également constater que le digest trypsique semble contenir une proportion plus importante de contamination épidermale. En comparaison des résultats obtenus par Lee et *al*. et présentés précédemment, la contamination corticale est atténuée signe qu'un bénéfice a été obtenu avec l'utilisation de la méthode d'extraction physique cuticulaire.

Identification des kératines et des KAP spécifiques de la cuticule et estimation de leur abondance

Les kératines cuticulaires identifiées dans les extraits protéiques n'ont pas fait l'objet d'une caractérisation comme nous l'avons précédemment exposée dans le cortex. Les kératines cuticulaires sont celles décrites par les travaux de Rogers, Schweizer, Langbein et *al*. (Figure 8). Leurs abondances relatives suggèrent une prévalence des kératines de type I et particulièrement de K32, exprimée dans la matrice et la zone de pré-élongation pendant la croissance folliculaire. La proportion de K35 exprimée dans la matrice semble beaucoup plus importante que ce qui a pu être observé dans le cortex. Les kératines de type I, K39 et K40, exprimés dans la zone de kératinisation sont retrouvées mais ne semblent pas prévaloir devant les kératines exprimées plus tôt au cours de la différenciation folliculaire comme c'est le cas dans le cortex. Les kératines de type II identifiées sont K85 retrouvée précédemment dans le cortex et K82 décrite spécifiquement dans la cuticule. K80, dont l'expression au niveau de la zone d'élongation a été décrite très récemment au niveau de la zone d'élongation dans la cuticule est également retrouvée [131]. Nous noterons une faible expression de K84 précédemment observée dans l'analyse de Lee.

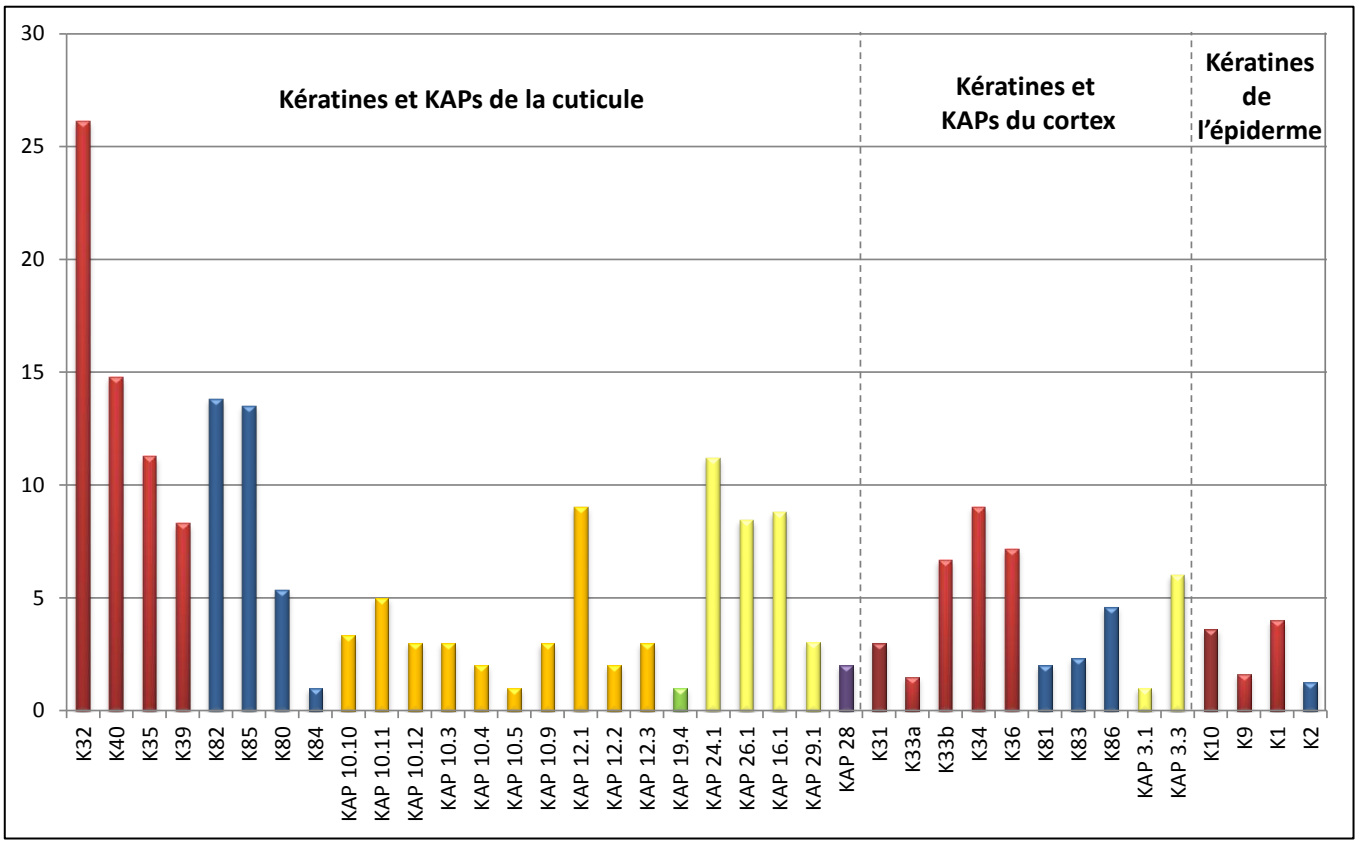


Figure 8 *: Mesure d'expression basée sur le comptage de spectres uniques des kératines et des KAP identifiées dans les digests de l'extrait cuticulaire.*

L'identification de ces kératines est à mettre en perspective de la compréhension des mécanismes conduisant à l'absence de structures fibrillaires organisées dans la cuticule en fin de différentiation. Le mécanisme conduisant à cette structure implique pourtant initialement K35 et K85 exprimées au même niveau de la croissance folliculaire dans les cellules corticales. Un rôle éventuel des kératines K32 et K82 voire de certaines KAP dans l'inhibition de la

croissance des microfibrilles mériterait d'être étudié dans une optique de compréhension des mécanismes d'initiation de formation de ces structures dans le cortex.

Les principaux résultats de ces analyses sont relatifs aux KAP identifiées.
Pour la première fois à notre connaissance, l'expression de la famille des KAP 28 est démontrée avec l'identification d'un peptide commun à l'ensemble des séquences prédites et décrites par Wu et *al*. [138]. Parmi les autres KAP riches en cystéine et en glycine, nous notons l'absence dans ces échantillons des KAP 5 et de la KAP 17.1 dont l'activité a pourtant été démontrée dans la cuticule au niveau du transcriptome.

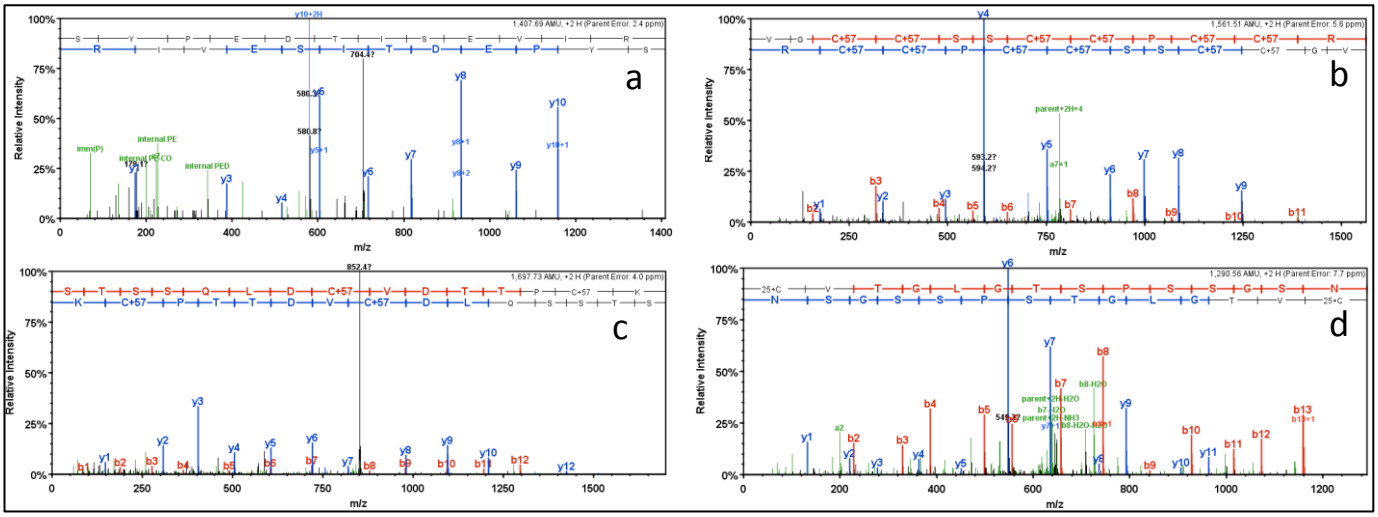


Figure 9 *: Spectres uniques des peptides issus de protéine dont l'expression n'avait auparavant pas été démontrée dans la cuticule et globalement dans le protéome humain. a) peptide SYPEDTISEVIR de la KAP 19.4. b) peptide VGCamCamSSCamCamPCamCamR de la famille des KAP 28. c) STSSQLDCamVDTTPCamK de la KAP 16.1. d) peptide CitaVTGLGTSPSSGSN de la KAP 29.1.*

La KAP 16.1, référencée dans Swissprot comme une séquence putative « 10-like » et détectée minoritairement dans le cortex est exprimée comme une protéine abondante dans la cuticule. Par ailleurs, une autre séquence « 10-like » correspondant à la séquence décrite par Wu et *al*. comme la KAP 29.1 est identifiée sans ambiguïté (Figure 9).
Nous notons également l'identification de la KAP 19.4 paralogue des KAP 19 identifiées dans le cortex. Le transcrit du gène correspondant avait été auparavant détecté spécifiquement dans la cuticule. Notre identification semble donc confirmer ce résultat qui suggèrerait l'expression spécifique de gènes paralogues de KAP dans des compartiments cellulaires différents. La KAP 19.4 est ainsi la seule KAP HGT à avoir été identifiée dans la cuticule et son expression semble être spécifique.
Les KAP 24.1 et 26.1 sont trouvées abondamment comme il avait été décrit dans les échantillons de Lee et par immunohistochimie [139, 140].

Concernant les KAP très riches en soufre, trois des 4 gènes de la famille des KAP 12 sont retrouvés comme protéines abondantes dans l'extrait. L'absence de la KAP 12.4 dans nos échantillons comme dans ceux de Lee et al. suggère un statut de pseudogène du gène correspondant. 7 des 12 individus de la famille des KAP 10 sont identifiés avec des peptides uniques parmi l'ensemble des isoformes (cf. paragraphe 3.c.).

Concernant l'identification des kératines et de KAP du cortex dans ces échantillons, l'hypothèse d'une contamination de l'échantillon cuticulaire n'écarte pas une éventuelle expression de ces protéines minoritairement dans la cuticule. Si l'expression des protéines spécifiques à la cuticule (K32, K82, K40, K80, KAP 10, 12, 5, 26, 24) peut facilement être détectée par les techniques d'immunohistochimie ou d'hybridation *in situ* des transcrits du follicule, il semble plus difficile de discerner des expressions de protéines pouvant être

coexprimées dans les deux types cellulaires. Cette difficulté peut être suggérée compte tenu de la résolution de ces techniques et de la faible taille de la couche de cellule cuticulaire. Par exemple, l'expression de K36, trouvée minoritairement dans nos analyses protéomiques du cortex et plutôt abondamment dans l'échantillon de cuticule pose la question de son expression dans la cuticule. Néanmoins, les valeurs semi quantitatives obtenues par comptage de spectres nécessitent d'être considérées avec précautions.

Ces résultats permettent de mettre en évidence l'expression, au niveau du protéome de la cuticule, de kératines et de KAP spécifiques. Ils suggèrent parmi les protéines de structure impliquées dans les mécanismes de kératinisation de ce type cellulaire une expression spécifique de certaines familles de KAP.

c) Identification des KAP de la famille 10

La famille des KAP 10 semble être particulièrement représentée dans les échantillons cuticulaires. Cette observation pourrait expliquer l'identification de certains de ces peptides comme contaminants dans le cortex.
Nous noterons que ces identifications sont principalement issues de peptides obtenus après digestion NTCB et présentant des spectres de fragmentation tout à fait informatifs. Ce type de digestion s'avère donc tout à fait adapté à la problématique de l'analyse protéomique des KAP riches en soufre et semble pourrait être utilisé pour l'analyse dans le cortex de protéines comme celles des familles 4 et 9.

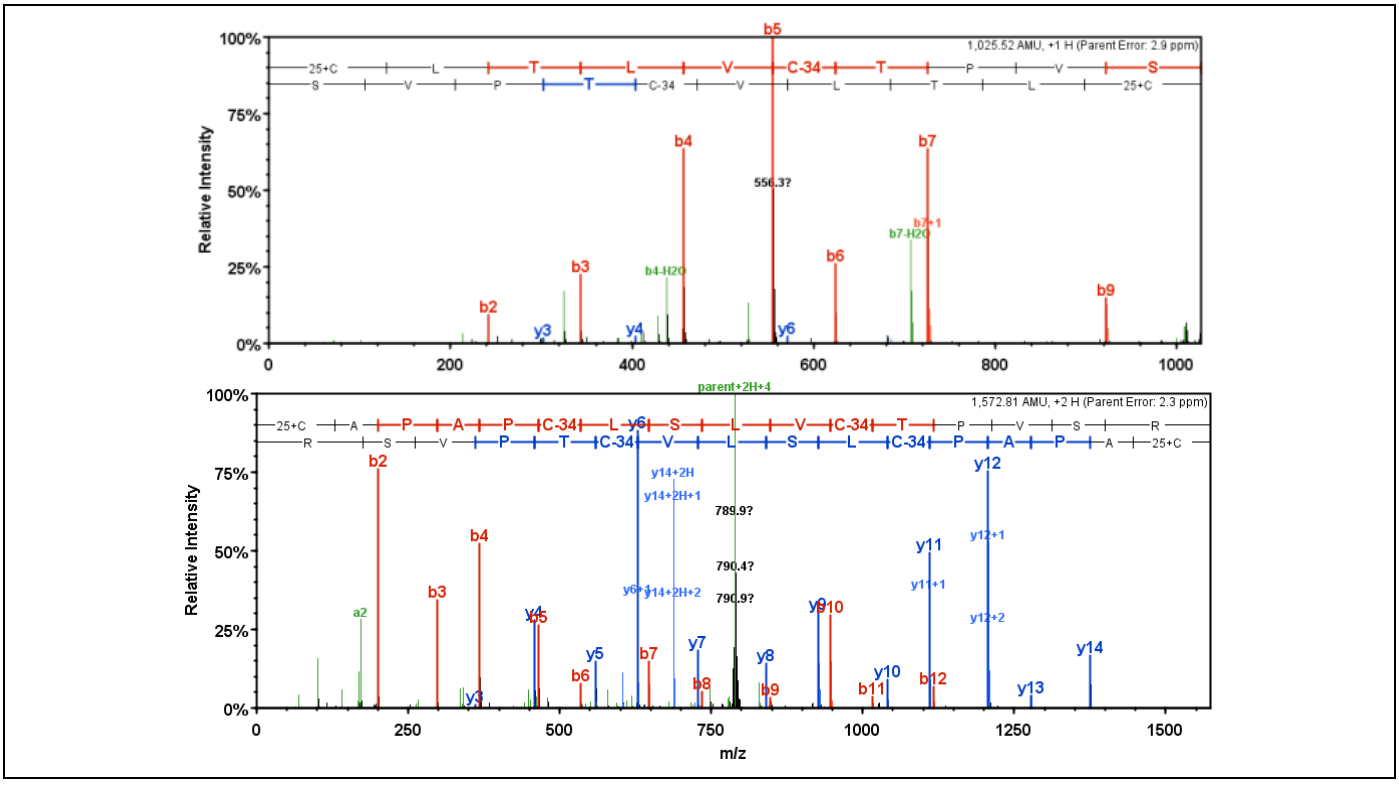


Figure 10 *: Exemple de spectres de fragmentation obtenus pour des peptides issus d'une digestion au NTCB. A gauche, peptide CitaLTLVCdhaTPVS de la famille des KAP 10 ne portant pas de site basique : la succession de fragments b indique que la charge est principalement localisée du côté de la cystéine modifiée en iminothiazolidine (ita). A droite, peptide CitaAPAPCdhaLSLVCdhaTPVSR de la famille des KAP 10 obtenu par digestion au NTCB et digestion trypsique : le spectre de fragmentation contient un mélange de fragments b et de fragments y. Ces observations, représentatives d'un ensemble de spectres étudiés, suggèrent la capacité de la fonction iminothiazolidine à accepter une charge lors de l'ionisation en électrospray.*

La forte homologie des KAP 10 nécessite, lors du rassemblement des données de séquençage, un examen minutieux de l'unicité des peptides identifiés. Lors de l'identification dans les différents digests des peptides par l'algorithme puis de leur assignation à une protéine, nous avons constaté que les résultats obtenus

automatiquement pouvaient s'avérer erronés. Dans certains cas, la combinaison de plusieurs peptides partagés par plusieurs isoformes est traduite en une identification d'une isoforme particulière par le moteur sans que cette dernière possède pour autant un peptide unique identifié dans l'analyse. Il est donc nécessaire d'examiner les résultats en recouvrant, avec l'ensemble des peptides identifiés, les séquences alignées des différentes isoformes puis de contrôler les peptides qui sont effectivement uniques à une isoforme. Cet examen nous permet de démontrer l'expression des KAP 10.3, 10.4, 10.5, 10.9, 10.10, 10.11 et 10.12. Les KAP 10.6 et 10.7 ne sont pas discriminées l'une de l'autre mais un peptide situé en N-terminal est unique à ces deux protéines. Ainsi, l'un ou l'autre ou les deux gènes pourraient être exprimés. La comparaison des séquences souligne une ambiguïté sur la position du N-terminal. L'identification de deux peptides N-terminaux suggère que la séquence la plus longue pourrait être considérée. Dans chacun des cas, les peptides sont N-acétylés.

d) Les KAP 10, des substrats potentiels de la transglutaminase et des candidats à la composition de l'exocuticule

Les recouvrements de séquences obtenus pour les KAP 10 suite à la combinaison de l'ensemble des résultats des digests sont plutôt faibles (typiquement de l'ordre de 25%). Malgré la multiplicité des sites de coupures potentiels notamment avec la combinaison de digestion (NTCB + trypsine), des zones entières des protéines ne sont pas recouvertes (Figure 11).
Le recouvrement systématique de certaines zones homologues pour ces différentes isoformes, nous a amené à envisager la possibilité que les KAP 10 soient des substrats de la transglutaminase et, de ce fait, potentiellement pontées au niveau de leurs glutamines ou de leurs lysines. La recherche dans les séquences de ces résidus (6 à 8 % de glutamine et 1 à 2% de lysine) montre que les zones riches en glutamine sont systématiquement non recouvertes. En considérant qu'une proportion importante de GGEL est décrite dans l'endocuticule (une glutamine sur sept pontée) et l'exocuticule (une glutamine sur quatre) [53], nous pouvons attribuer ces absences de couverture de séquence à une conséquence de ce type de modification.

Ces résultats nous amènent à suggérer que la famille des KAP 10 est un substrat des transglutaminases dans la cuticule et que les zones riches en glutamine non recouvertes en analyse protéomique sont potentiellement porteuses de ponts non réductibles. Compte tenu de l'absence d'autres KAP riches en glutamine décrites comme exprimées dans ce type cellulaire, nous pouvons également faire l'hypothèse que ces KAP riches en soufre, et potentiellement en GGEL, sont une des composantes de l'exocuticule qui contient abondamment de ces deux éléments [50, 51, 53].

Dans ce cadre, une redigestion de l'insoluble obtenu suite à la digestion (NTCB+trypsine) et préalablement rincé a été réalisée en utilisant de nouveau la trypsine dans l'optique d'observer les protéines qui pouvaient alors être détectées. Les résultats de cette analyse (non montrés) indiquent que le nombre de protéines identifiées est considérablement plus faible que lors de la première digestion (15 contre 169). Parmi ces 15 protéines, nous retrouvons les KAP 26.1, 10, 16.1, 12.1, 12.2, 12.3 et 29.1 ainsi que les kératines K32, K82 et K85 qui pourraient, entre autres constituer l'exocuticule insoluble. Néanmoins, cette analyse préliminaire ne nous a pas permis d'identifier ces protéines avec des couvertures de séquences aussi importantes que lors des analyses de l'endocuticule. Une optimisation du protocole, par exemple en combinant une digestion trypsique permettant d'éliminer l'endocuticule puis une digestion de l'insoluble au NTCB puis à la trypsine pourrait être par la suite envisagée pour une étude ciblée de l'exocuticule.

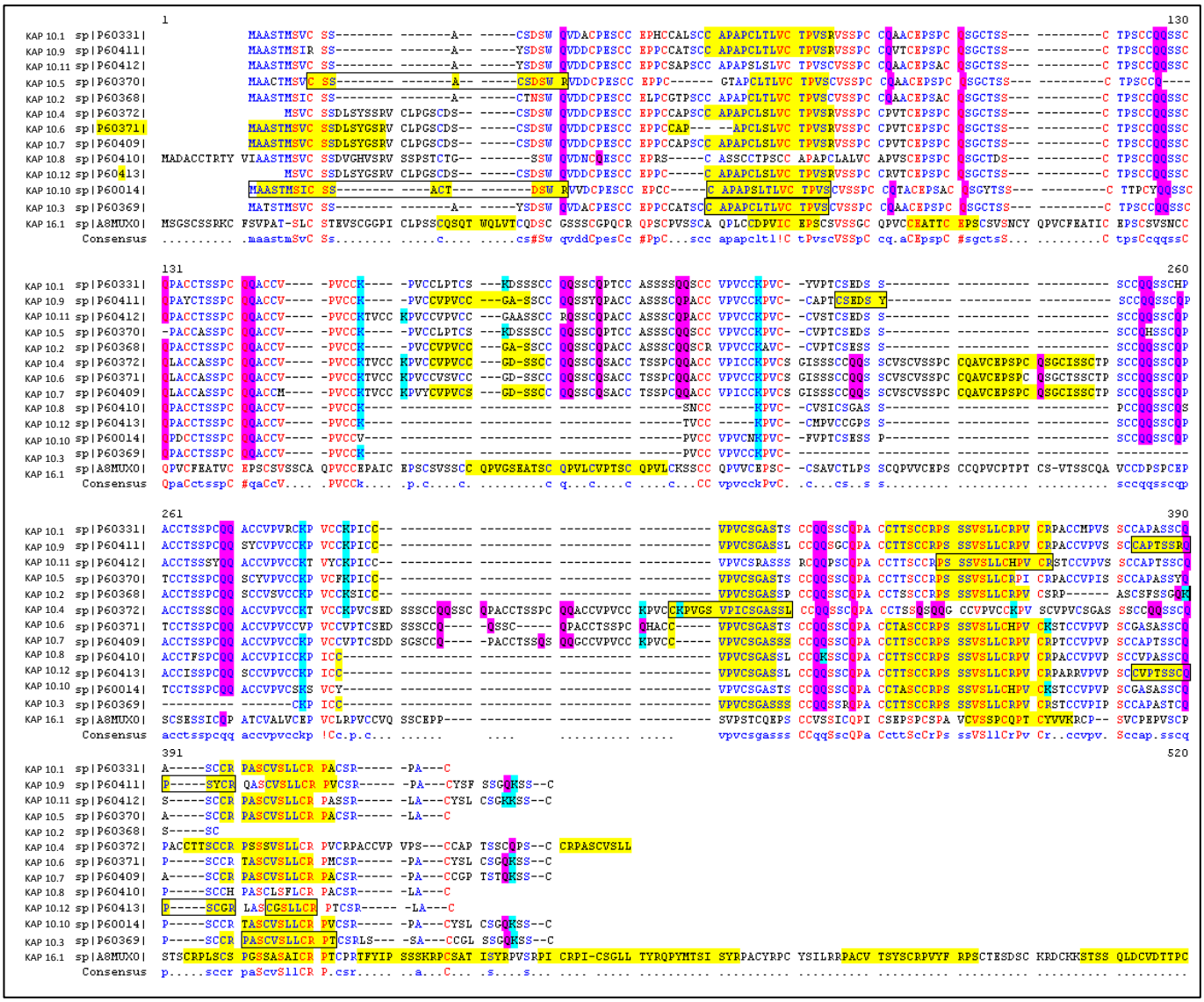


Figure 11 *: Alignement des séquences des KAP 10 décrites dans la banque Swissprot et recouvrements de séquences (en jaune) obtenus suite à l'analyse protéomique de la cuticule développée au cours de cette étude. Les peptides identifiés comme protéotypiques et permettant de démontrer l'expression d'un gène particulier pour les KAP 10 sont encadrés. Pour cette même famille, les glutamines et les lysines, potentiellement substrats de la transglutaminase sont représentées. Nous montrons ici que les peptides identifiés pour la famille 10 sont exclusivement localisés dans les zones pauvres en glutamines des séquences.*

La nouvelle KAP 16.1 identifiée et dénommé « 10-like » dans Swissprot montre des homologies de séquence avec la famille 10 mais est suffisamment différente pour ne pas être considérée comme une isoforme de cette famille.

4. Perspectives

a) L'analyse de l'exocuticule et de la couche A

La recherche des substrats de la transglutaminase suggère d'envisager les isoformes de la famille des KAP 10 comme candidats engageant potentiellement leurs glutamines. La question reste cependant posée quant à l'identification de protéines potentiellement donneuses de lysines. L'absence de résultats d'identification pour la famille des KAP 5, par ailleurs vues par hybridation *in situ* dans la cuticule au niveau du follicule et dont les séquences contiennent une certaine abondance en lysine (Figure 12), pourrait être expliquée par cette fonction de substrat pour ces protéines. En effet, de nombreux sites de coupures sont accessibles à la coupure trypsique et NTCB et devraient, en théorie, permettre d'obtenir des peptides de KAP 5 pouvant être séquencés par

spectrométrie de masse. Néanmoins, les zones N-terminales de ces protéines, dépourvues de lysine et de glutamine auraient tout de même pu être séquencées et potentiellement vues dans les digests d'endocuticule. Les protéomes de l'exocuticule et de la couche A n'ayant pas été particulièrement caractérisés dans notre étude, nous pouvons envisager d'y rechercher cette famille de protéines.

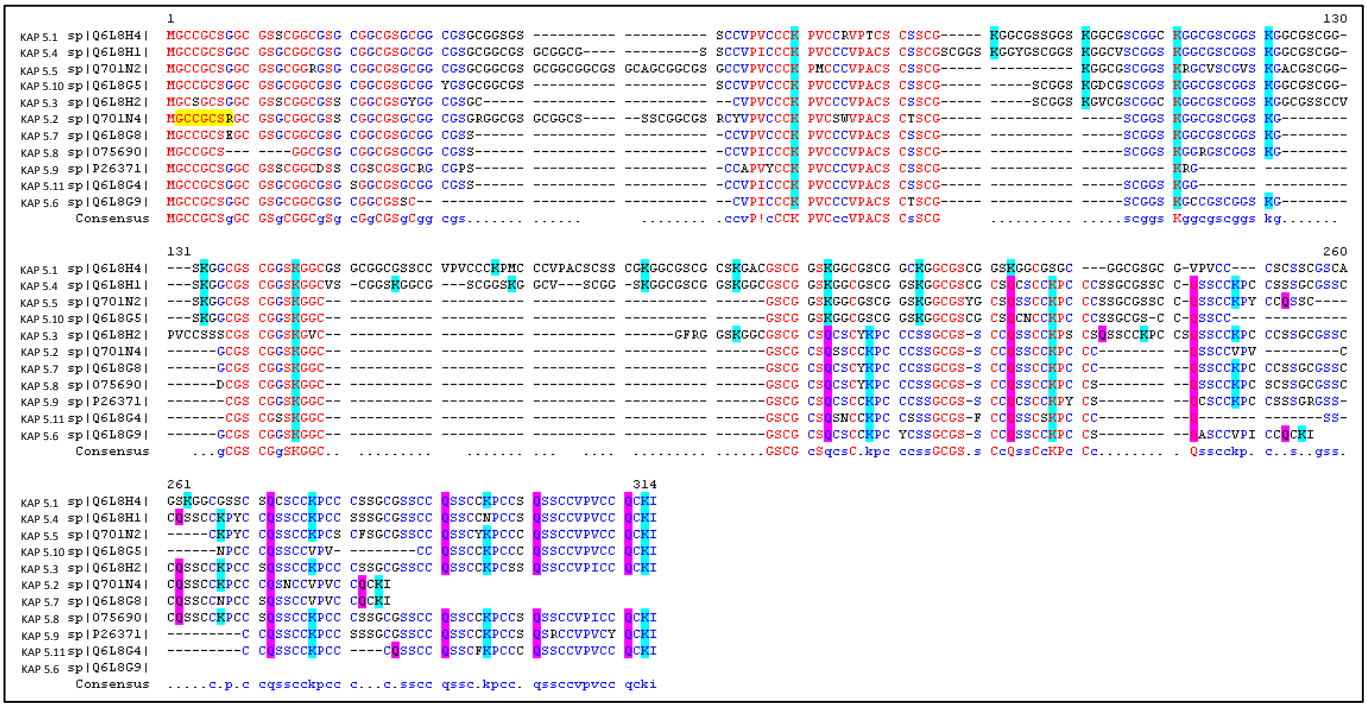


Figure 12 *: Alignement des séquences des KAP 5 décrites dans la banque Swissprot. Seul un peptide (recouvert en jaune) a été séquencé en tant que composé minoritaire dans le cortex. Nous présentons ici la localisation dans les séquences des lysines et des glutamines pouvant être considérés comme des sites de substrats potentiels de la transglutaminase.*

Envisager ces protéines comme donneuses de lysine et de glutamines pour la formation de liaisons GGEL ne doit cependant pas écarter la possibilité de telles liaisons avec les kératines retrouvées dans la cuticule. Les kératines contiennent elles-mêmes une certaine proportion de ces acides aminés notamment dans leurs domaines hélicoïdaux (respectivement 7% de lysine et 5% de glutamine pour les kératines de type II et 3-4% et 10% pour les kératines de type I). Cette hypothèse doit être d'autant plus considérée que l'expression de la TGM3 a été détectée par immunohistochimie dans la cuticule dès la zone de pré-élongation où seules les kératines K85, K32 et K82 sont observées comme exprimées [82, 83, 114].

Pour autant, ces hypothèses ne pourront être confirmées que par l'identification et la caractérisation des peptides pontés correspondants. Les études protéomiques réalisées ici sur la cuticule permettent de restreindre cette recherche à certaines familles de protéines et montrent le potentiel de la digestion chimique NTCB pour envisager la génération de ces peptides pouvant être attendus dans l'exocuticule voire dans la couche A. Néanmoins, il parait dans un premier temps nécessaire de parvenir à isoler l'exocuticule et la couche A de l'endocuticule où la liaison GGEL devrait être représentée. Les compositions particulières de l'exocuticule et de la couche A (la densité en résidus cystéines semble plus importante dans la couche A) posent la question d'une différence de composition en protéines qui les constituent. Dans un second temps, une stratégie de détection et de caractérisation des peptides pontés devra être développée.

b) L'utilisation de la stratégie de traitement des données séquentielles pour affiner les connaissances du protéome de la cuticule.

La stratégie de traitements des données développée pour le cortex par traitement séquentiel des résultats n'a, à l'heure actuelle, pas été adaptée à la caractérisation des protéines identifiées dans l'endocuticule. Le nombre important de données générées permet pourtant d'y envisager la recherche de modifications chimiques et de modifications post traductionnelles, qui pourraient être communes à celles retrouvées dans le cortex. Des modifications spécifiques à la cuticule et accessibles aux stratégies d'identifications protéomiques pourraient également être recherchées. Zahn et *al*. [53] décrivent dans leurs extraits cuticulaires la présence de citrulline qui pourrait ainsi être recherchée sur les protéines identifiées. Une recherche sur les données n'ayant pas conduit à des résultats en utilisant le mode d'identification tolérant aux erreurs ainsi que la recherche dans des banques complémentaires comme SwissProt en version Varsplic et NCBI nr pourraient également apporter d'autres informations dans une problématique de correction de certaines séquences et d'identification de variants.

Chapitre III Etude du protéome des onychocytes et des cellules corticales

Une étude protéomique très récente de Rice et *al.* [276] a focalisé notre attention sur le protéome des ongles. Les résultats de ces travaux montrent, du point de vue des protéines de structure constituant la majeure partie des extraits analysés, que les mêmes kératines dures que celles trouvées dans le cortex y sont globalement identifiées. Concernant les KAP, seules des KAP des familles 2, 13 et 11 sont décrites comme retrouvées dans ces échantillons. L'absence dans ces résultats d'identification de KAP riches en soufre telles que celles retrouvées abondamment dans le cortex comme les KAP 4, 9, 3 et 1, nous a conduit à réaliser nos propres analyses afin d'évaluer si ces protéines étaient effectivement absentes de ce type cellulaire. Nous présentons dans ce chapitre les premiers éléments de résultats de l'analyse protéomique obtenus en utilisant la même stratégie décrite appliquée précédemment pour le cortex et la cuticule.

1. Présentation de l'appareil unguéal

Les ongles constituent des plaques cornées en croissance sur la face dorsale des phalanges ayant pour fonctions de protéger les extrémités des doigts et d'être utilisées comme outil chez les primates. La plaque constituée par l'ongle est issue de la croissance continue de cellules au niveau de la matrice puis d'un processus de kératinisation qui peut être mis en parallèle avec la différenciation observées pour les cellules du cortex. Contrairement au follicule, il n'existe pas de cycle de renouvellement de ce dérivé épidermique. Nous ne présenterons pas plus en détail le fonctionnement de ce système biologique mais décrirons la structure kératinisée et morte constituée par la plaque unguéale après la kératinisation.

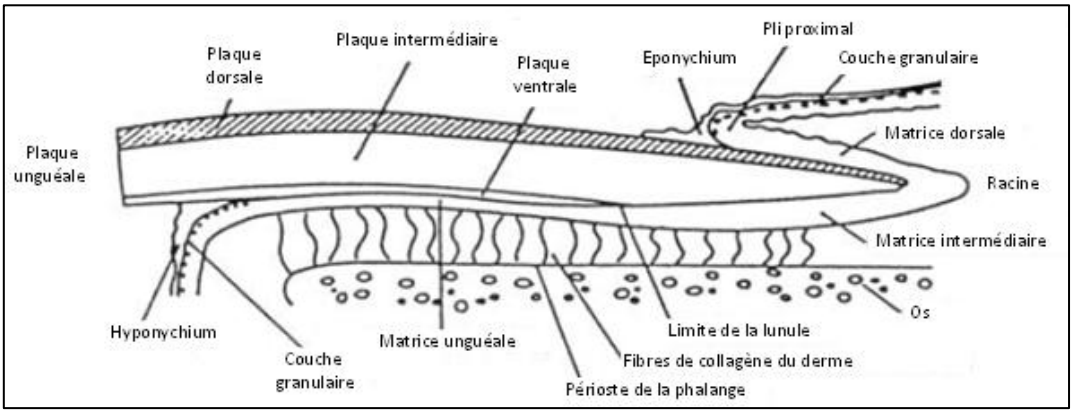


***Figure 1** : schéma représentant une section de l'appareil unguéal.*

La plaque unguéale se divise en trois plaques distinctes : la plaque dorsale, la plaque intermédiaire et la plaque ventrale (Figure 1). L'ensemble glisse le long de la matrice unguéale. La plaque intermédiaire constitue la majorité de la structure et se compose de cellules polyédriques allongées, les onychocytes, remplies de structures très proches des macrofibrilles observées dans le cortex (Figure 2). Les onychocytes, situés à la base de la plaque intermédiaire, ont la particularité de posséder des structures membranaires formant de larges villosités permettant un enchevêtrement de la membrane avec la membrane de la cellule adjacente. Cette structure, différente du complexe de membrane cellulaire du cortex permet sans aucun doute d'ancrer les cellules les unes aux autres et de conférer une intégrité à la plaque unguéale [64, 277-281]. La différence structurale entre les différentes plaques réside principalement dans une orientation différente des macrofibrilles qui les constituent. Les macrofibrilles des cellules de la plaque dorsale et intermédiaires sont perpendiculaires les unes des autres, chacune étant parallèle à la surface. Il semble également que leurs structures membranaires soient légèrement différentes.

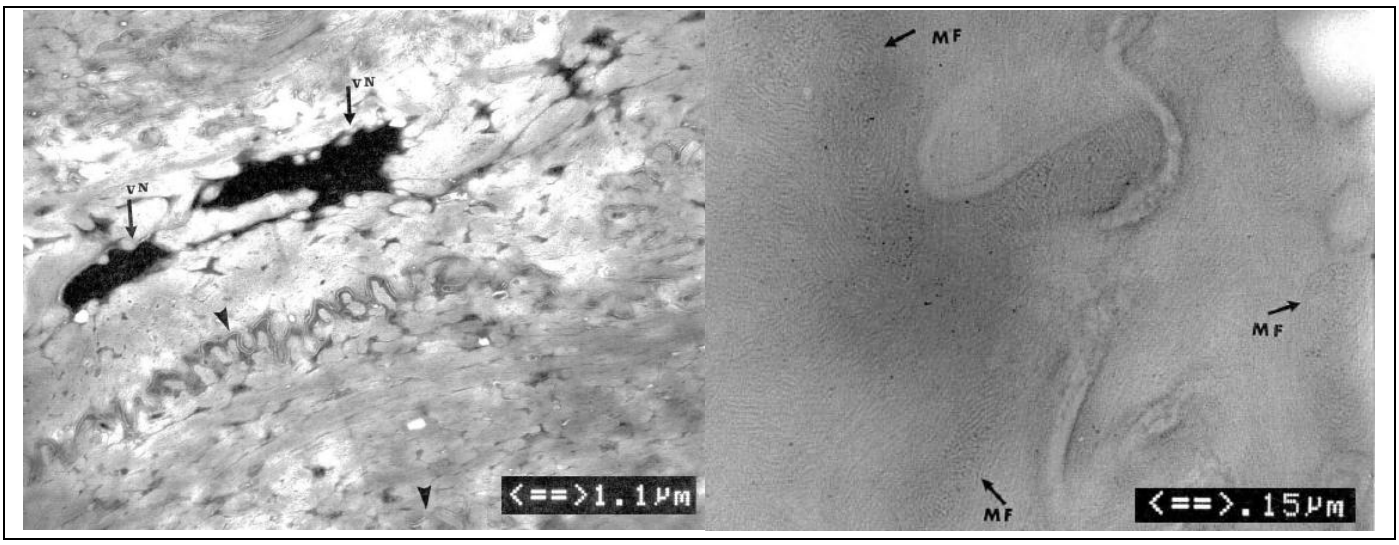


Figure 2 *: Visualisation en microscopie électronique à transmission des structures des onychocytes. A gauche, observation des structures macrofibrillaires de la partie intermédiaire de la plaque unguéale ; les membranes cellulaires (têtes de flèche) et les vestiges nucléaires (flèches entières) sont visibles. A droite, structures des microfibrilles composant les macrofibrilles. Données internes L'Oréal.*

2. Extraction des protéines

Des rognures d'ongles ont été préalablement nettoyées à l'aide de détergent puis rincées abondamment. Les ongles sont placés pendant 12 heures à 45°C dans la même solution que celle utilisée pour l'extraction du cortex. Les ongles réduits sont par la suite agités vigoureusement pour réaliser leur pulvérisation à 45°C pendant 12 heures à l'aide d'un agitateur magnétique. L'extrait hétérogène obtenu est précipité avec l'ajout d'éthanol. Le précipité est rincé, puis resolubilisé dans un tampon de réduction et alkylé avec l'iodoacétamide. L'extrait obtenu est à nouveau précipité puis divisé en différents échantillons placés dans un tampon de digestion puis digérés à la trypsine, à la chymotrypsine ou à la GluC. Les digests sont par la suite préparés et analysés de la même manière que ce qui a été précédemment décrit. Des triplicatas d'analyses des collectes des digests trypsique et chymotrypsique ont été réalisées. Les analyses des collectes du digest GluC n'ont pas été répétées.

3. Analyse protéomique

Comme pour les précédentes études, nous avons tout d'abord évalué, sur la base du comptage de spectres des individus identifiés, la répartition des différents types de kératines retrouvées dans l'échantillon. Les kératines et les KAP attendues dans le cheveu ont été regroupées tout comme les kératines épidermales. Un groupe supplémentaire à ceux constitués lors des précédentes études a été considéré : différentes kératines communes à celles exprimées dans la gaine externe du follicule ont en effet été retrouvées.

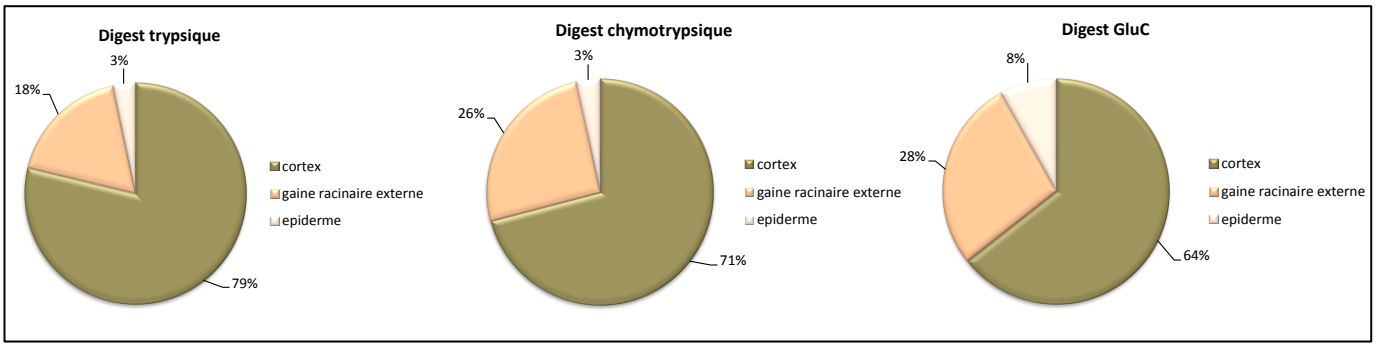


Figure 3 : *Proportion des kératines et des KAP des différents groupes envisagés dans les digests de la plaque unguéale.*

La majorité des kératines et des KAP identifiées sont des kératines dures et des KAP retrouvées également dans le cheveu (Figure 3). Nous retrouvons également une légère contamination épidermale. Une proportion non négligeable de kératines dites de la gaine externe est retrouvée. Certaines de ces kératines avaient été précédemment attribuées dans le cortex à une éventuelle présence de médulla dans les échantillons corticaux (K14, K5 et K6b). Compte tenu du peu de données histologiques relatives aux ongles, il est ici difficile de déterminer l'origine cellulaire précise de ces protéines. Ces protéines pourraient être issues de la plaque ventrale dont la structure filamentaire parait relativement différente de celle observée dans la plaque intermédiaire. Néanmoins, rien ne peut exclure une éventuelle présence de ces kératines dans les autres compartiments de la plaque unguéale.

Parmi les kératines et les KAP pouvant être également décrites dans le cheveu, nous retrouvons la majorité de protéines auparavant identifiées dans le cortex (Figure 4).

Les kératines corticales de type I, K31, K33a, K33b, K34, K36 et K39 sont identifiées. K35, exprimée en début de différenciation dans le cortex et la cuticule est retrouvée minoritairement. La kératine cuticulaire de type I, K32, est retrouvée très minoritairement. Nous noterons l'expression spécifique de K37 qui n'avait pas été identifiée dans le cheveu. La kératine corticale K38, dont la séquence est fortement homologue à K37, n'est pas détectée.

Parmi les kératines de type II, celles retrouvées dans le cortex comme K85, K81, K83 et K86 sont également exprimées abondamment dans la plaque unguéale. La kératine cuticulaire K82 est retrouvée très minoritairement. K84, qui n'avait été détectée que minoritairement dans la cuticule, semble être exprimée spécifiquement dans l'ongle.

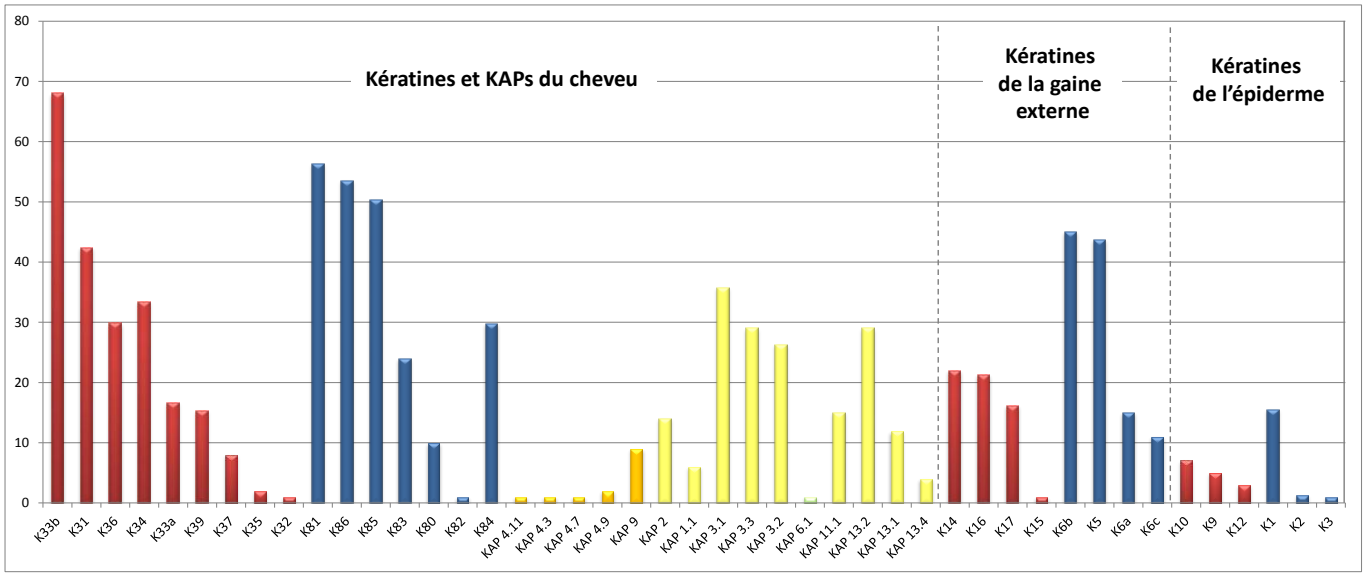


Figure 4 : *Mesure d'expression basée sur le comptage de spectres uniques des kératines et des KAP identifiées dans les digests de l'extrait unguéal.*

Le nombre d'identification d'individus de KAP s'avère moins complet que ce qui avait été obtenu pour le cortex. Cependant, nous montrons en supplément de l'étude de Rice et *al.* que les différentes familles de KAP riches en soufre 4, 9, 1, 2 et 3 sont, comme dans le cortex, représentées. Les KAP 11.1 et 13 vues dans le cortex sont également retrouvées. Nous noterons que la KAP 13.4 est identifiée, ce qui ajoute une nouvelle KAP à la liste des KAP exprimées dans le protéome humain. La KAP 6.1, du groupe des KAP HGT, est minoritairement exprimée.

Les similarités entre le protéome du cortex et le protéome de la plaque unguéale appuient de données moléculaires les observations des structures voisines effectuées en microscopie électronique. Néanmoins, l'expression spécifique de certaines protéines suggère des différences qui pourraient traduire des spécificités structurales propres à la différenciation de chacun de ces types cellulaires. Nous retiendrons principalement des premiers résultats de cette étude que l'observation de structures microfibrillaires dans le cortex et dans la plaque

unguéale semble étroitement lié à l'expression simultanée des kératines dures et de KAP riches en soufre spécifiques.

En plus des protéines précédemment citées, plus de 110 autres protéines ont été identifiées. Ce nombre est significativement supérieur à ce qui a été obtenu dans le cortex (de l'ordre d'une cinquantaine de protéines). Nous n'avons pour l'instant pas étudié en détail ces protéines (Annexe 5). La comparaison avec celles identifiées dans la cuticule et le cortex permet toutefois d'envisager la mise en évidence de protéines pouvant avoir des fonctions spécifiques à la différenciation de ces cellules. Une comparaison de ce type a déjà été en partie réalisée dans l'étude de Rice et *al*. , ces derniers ayant notamment comparé les protéines retrouvées dans les ongles et celles identifiées dans l'extrait insoluble de cheveux. Les résultats supplémentaires que nous avons obtenus grâce à l'utilisation d'autres enzymes de digestion, associé à la plus grande spécificité de la technique d'extraction cuticulaire employée, pourraient cependant être utilisés par la suite pour affiner ces comparaisons.

Conclusions et perspectives à l'exploration des protéomes des phanères

Affiner la connaissance des protéomes du cortex, de la cuticule et des ongles

Les études des protéomes décrits dans ce manuscrit ont permis l'identification sans ambiguïté d'un minimum de 56 KAP. Le nombre total de gènes de KAP exprimés dans le protéome humain doit cependant être supérieur si l'on considère l'expression de la famille multigénique 28 retrouvée pour la première fois avec un peptide commun aux huit isoformes décrites dans le génome. Certaines familles, potentiellement substrats de la transglutaminase dans la cuticule, paraissent difficiles à caractériser (onze individus de la famille des KAP 5). Sur la centaine de gènes décrits chez l'humain, nos études tendent à montrer que certains d'entre eux seraient des pseudogènes (KAP 1.4, KAP 9.1, KAP 12.4). L'absence d'identification de certains individus (KAP 22, 23, 25 et 27) pose la question de leur expression effective. Nous pouvons cependant considérer que certains développements peuvent encore être réalisés afin de caractériser les protéines minoritaires de ces protéomes et dont pourraient faire partie les KAP encore non identifiées. Les nouvelles techniques de déplétions des protéines majoritaires pourraient ainsi être employées [282, 283] tout comme les techniques de séparation protéiques présentées en deuxième partie de ce manuscrit.

Une analyse plus approfondie de la cuticule pourra probablement à l'avenir permettre de démontrer l'expression de gènes supplémentaires attendus mais actuellement manquants (individus des KAP 10, 5, 28 voire 17.1).

Concernant les KAP du cortex, nous disposons désormais de suffisamment de données expérimentales pour considérer que les protéines qui participent majoritairement à la structure macrofibrillaire sont désormais caractérisées.

La cartographie du génome humain est encore susceptible d'évoluer et la proposition de nouvelles séquences candidates à une expression protéique dans le cheveu pourrait encore à l'avenir être à considérer. L'exemple de la description du cluster de gènes des KAP 28, il y a seulement 3 ans, illustre bien cette problématique. Les données expérimentales obtenues suite à ces travaux sur ces protéomes pourront être ultérieurement réutilisées pour les confronter à de futurs gènes candidats.

Vers une exploration des « protéomes cousins » ?

Confirmer ou infirmer l'expression des gènes de KAP passe également par leur recherche dans des protéomes voisins de ceux précédemment étudiés. Les autres fibres capillaires que possède l'humain méritent dans ce cadre d'être étudiées. Les poils de barbe contiennent une proportion importante de médulla et pourraient être utilisés pour accéder au protéome de ce type cellulaire qui n'a pas encore été étudié par analyse protéomique. Les KAP riches en glycine et en tyrosine n'ont pas été vus comme très abondantes dans le cheveu. Une partie des gènes de KAP, toujours peu ou pas identifiés (familles 6, 20, 21, 22), pourrait être recherchée dans les différentes structures capillaires de l'humain.

Les comparaisons de l'expression de protéines communes à différents protéomes ont fait l'objet d'un certain nombre de nos travaux. Ces comparaisons, effectuées sur la base d'estimation d'abondance par comptage de spectres, pourraient par la suite être affinées par l'utilisation de stratégies de quantification plus précises. Les comparaisons des aires des pics chromatographiques, ou comparaison « label free », que nous avons commencé à développer dans le cadre de notre problématique, pourraient ainsi être utilisées. D'éventuelles différences d'expression entre ces structures pourraient être recherchées afin d'apporter des éléments pouvant expliquer ces polymorphismes de fibres observés au sein d'un même génome.

Une compréhension future des polymorphismes des cheveux grâce à l'analyse protéomique ?

Si les origines des polymorphismes des fibres capillaires observés au sein des populations humaines semblent devoir être expliquées en partie par des différences de régulation de la croissance folliculaire, des différences d'expression des protéines des structures entre les individus ne sont pas à exclure. Les comparaisons des

protéomes entre individus différents peuvent ainsi être envisagées pour les mettre en évidence. Cependant, la comparaison de protéomes dont les génomes sont différents expose à une difficulté analytique supplémentaire. Il sera en effet indispensable de considérer l'introduction des variations issues du poly allélisme, existant pour une partie des gènes des kératines et des KAP exprimés dans le cheveu, avant de pouvoir considérer les variations induites par des différences de l'expression des protéines elles-mêmes.

L'analyse protéomique pour qualifier précisément l'impact des traitements cosmétiques

Les perspectives du suivi des modifications induites par les traitements cosmétiques afin d'affiner l'évaluation de leur impact et de trouver des conditions opératoires minimisant leur effets négatifs sur la fibre semblent pouvoir être envisagées par l'utilisation des stratégies « label free ». Ces techniques pourraient se révéler à l'avenir des compléments indispensables aux analyses d'acides aminés des fibres capillaires traitées compte tenus des informations supplémentaires apportées. La localisation au sein des protéines des résidus affectés par les traitements ainsi que le suivi des modifications sur les résidus labiles vont pouvoir être considérés.

Partie IV Perspectives à l'analyse des protéomes des kératinocytes : la compréhension du rôle et de l'origine des protéines dans la formation des structures du cheveu.

Les analyses protéomiques réalisées au cours de ce travail de thèse ont permis de montrer l'expression des différentes familles de KAP dans trois types cellulaires exprimant les kératines dures. Ces résultats suggèrent l'importante contribution des familles de kératines et de KAP, retrouvées majoritairement dans le cortex, dans la cuticule et les ongles. Comme nous l'avons décrit en première partie, les connaissances relatives à la structure des kératines dans les microfibrilles sont de plus en plus détaillées. La contribution et le rôle des KAP sont en revanche bien moins appréhendés. A la lumière de la démonstration de l'expression des KAP et des informations semi quantitatives obtenues dans les différents protéomes analysés, nous nous sommes intéressés à la recherche d'informations pouvant nous orienter sur l'origine et les fonctions de ces protéines.
L'objectif de cette partie est de trouver des éléments supplémentaires afin de proposer des hypothèses d'arrangements des KAP entre elles et avec les filaments de kératines.

Dans un premier chapitre, nous présentons une étude réalisée pour évaluer si les données de composition en acides aminés obtenues dès les années 70 pour la structure corticale peuvent être utilisées en complément des données d'identifications protéomique. Nous montrerons comment nous avons pu estimer une composition quantitative des protéines majoritaires du cortex qui nous permet de considérer les familles multigéniques vraisemblablement majoritaires dans l'espace intermicrofibrillaire.

Dans un second chapitre, nous montrerons qu'une partie des KAP considérées comme majoritaires dans le cortex possèdent des motifs de répétitions pentapeptidiques caractéristiques et communs dans différentes familles. Nous évaluerons, parmi un ensemble de séquences orthologues obtenues chez d'autres mammifères, la conservation de ces motifs au cours de l'évolution. La comparaison des séquences des mammifères nous permettra également de montrer des structures de séquences primaires de certaines familles de KAP communément retrouvées parmi les espèces considérées. Les motifs de répétitions mis en évidence seront étudiés par modélisation moléculaire en considérant des hypothèses d'arrangements s'appuyant sur des événements de maturation des protéines survenant pendant les phases de kératinisation. Des propriétés structurales de ces motifs seront alors envisagées et nous proposerons un modèle décrivant l'association des filaments et des KAP de la matrice interfilamentaire. Ce modèle s'appuiera sur nos connaissances des séquences primaires des protéines et sur le principe que les associations inter protéines sont gouvernées par de simples interactions physico-chimiques.
Sur la considération du rôle fondamental joué par les motifs pentapeptidiques pour l'établissement de la structure finale du cortex, nous tenterons d'établir l'origine de ces gènes codant pour ces séquences uniques aux mammifères. Nos résultats suggéreront l'introduction chez l'ancêtre des mammifères, il y a 200 à 300 millions d'années, d'une séquence exogène dont une séquence homologue peut être retrouvée parmi une bactérie tout à fait particulière peuplant les océans.

Chapitre I Développement d'une stratégie originale de quantification des protéomes des kératinocytes

Les mesures semi quantitatives obtenues précédemment avec les approches par comptage de spectres peuvent s'avérer insuffisantes pour précisément évaluer la composition quantitative des structures kératinisées. La comparaison de l'expression de différentes protéines les unes par rapport aux autres sur la base de l'expression des signaux des ions des peptides obtenus après digestion comporte différents biais. Il est en effet nécessaire de considérer les différences de cinétique de digestion qui peuvent exister pour générer ces peptides.
La quantité de peptide obtenue après digestion enzymatique dépend sensiblement des sites de coupures enzymatiques spécifiques mais également aspécifiques pouvant conduire à la coupure du peptide mesuré. La cinétique de digestion est ainsi dépendante des résidus adjacents aux sites de coupure ce qui peut entrainer des biais significatifs sur l'estimation de l'expression de protéines avec des peptides. Dans notre cas, les mesures réalisées sur l'expression des KAP possédant des sites de coupures de type R.P (quasiment systématiques pour les KAP 4, 1, 2 et 9) sont sans aucun doute sous estimées.
Les signaux mesurés par LC-ESI-MS dépendent de la quantité de peptide analysé mais sont également entre autre fonction du rendement d'ionisation et des éventuels effets de suppressions ioniques. Il est possible d'utiliser des peptides standards pour réaliser la quantification absolue des peptides dans le digest. Cette stratégie nécessite de disposer de tels standards et ne s'affranchit pas du biais pouvant exister entre la quantité de peptide mesurée et le rendement de digestion de la protéine qui reste inconnu.
Dans le cadre de notre problématique, il semble donc nécessaire d'envisager une alternative permettant d'accéder à des données de quantification en s'affranchissant de ce biais introduit par la digestion.

1. L'analyse de la composition en acides aminés des extraits protéiques : une données à exploiter

Depuis le développement de techniques d'analyses par acides aminés, un grand nombre de données a pu être obtenu pour décrire la composition des hydrolysats de cheveu. Ces analyses n'ont jamais pu constituer des données suffisantes pour connaitre quelles étaient les familles de protéines pouvant être retrouvées dans les cheveux. Elles ont pour autant pu être utilisées pour évaluer des différences de composition des fibres en montrant par exemple une proportion plus importante de glycine et de tyrosine dans la laine que dans le cheveu suggérant une différence significative de l'expression des KAP HGT. Cette technique a également permis de montrer l'impact de l'apport nutritionnel sur la proportion de cystéines retrouvées dans les fibres de laine suggérant que l'expression de certaines protéines riches en soufre était dépendante de l'alimentation. La composition en acides aminés mesurée sur les hydrolysats contient donc potentiellement des informations quantitatives sur l'expression des protéines qui s'y trouvent. La proportion de certains acides aminés comme la cystéine, la sérine ou la proline peut suggérer que certaines protéines riches en ces résidus sont majoritaires dans la structure.

Nos études protéomiques du cortex ont montré que les extraits corticaux étaient majoritairement constitués de kératines et de KAP dont nous avons réalisé l'identification. En considérant ces résultats et le fait que la composition en acides aminés au sein des familles multigéniques reste relativement constante nous pouvons simplifier la composition du cortex à un mélange de kératines dures de type I et II et des différentes familles de KAP corticales auxquelles peuvent s'ajouter des rémanents cytoplasmiques et du noyau ainsi que des membranes.

Nous avons alors envisagé d'utiliser les données de composition en acides aminés de ces protéines identifiées par protéomique pour tenter de modéliser la composition en acides aminés mesurée expérimentalement dans les hydrolysats. Cette modélisation implique de trouver la contribution de chacune des protéines identifiées pour expliquer la composition totale en acides aminés. La comparaison des contributions de chaque protéine dans le modèle peut donc donner une information quantitative sur la composition du cortex.

2. Stratégie de modélisation de la composition en acides aminés du cortex

Nous avons choisi de constituer une matrice de données théorique contenant les proportions en acides aminés connus à partir des séquences des différentes protéines choisies pour modéliser la composition du système. Nous avons pondéré chaque colonne par un facteur variable pouvant être traduit comme une proportion de chaque protéine par rapport aux autres. En tenant compte de ces facteurs de pondération, une proportion de chaque acide aminé avec les facteurs fixés peut être calculée dans les conditions fixées. Cette proportion peut alors être comparée aux proportions mesurées expérimentalement. Les hydrolysats ne permettent pas d'obtenir de valeur pour la glutamine, l'asparagine et le tryptophane. Les fonctions amides étant converties pendant l'hydrolyse en acides carboxyliques, nous pouvons regrouper les valeurs théoriques de glutamine et d'acide glutamique pour les comparer à l'acide glutamique mesuré. Nous faisons de même pour l'asparagine et l'acide aspartique comparés à la valeur mesurée acide aspartique. La valeur du tryptophane n'est pas considérée.
Afin d'évaluer l'écart entre les proportions d'acides aminés calculés et la proportion mesurée expérimentalement, une mesure des écarts est réalisée. La somme des carrés de ces écarts doit ainsi être minimale pour que les proportions des protéines fixées dans le modèle permettent d'obtenir des proportions d'acides aminés calculés les plus proches de la donnée expérimentale.

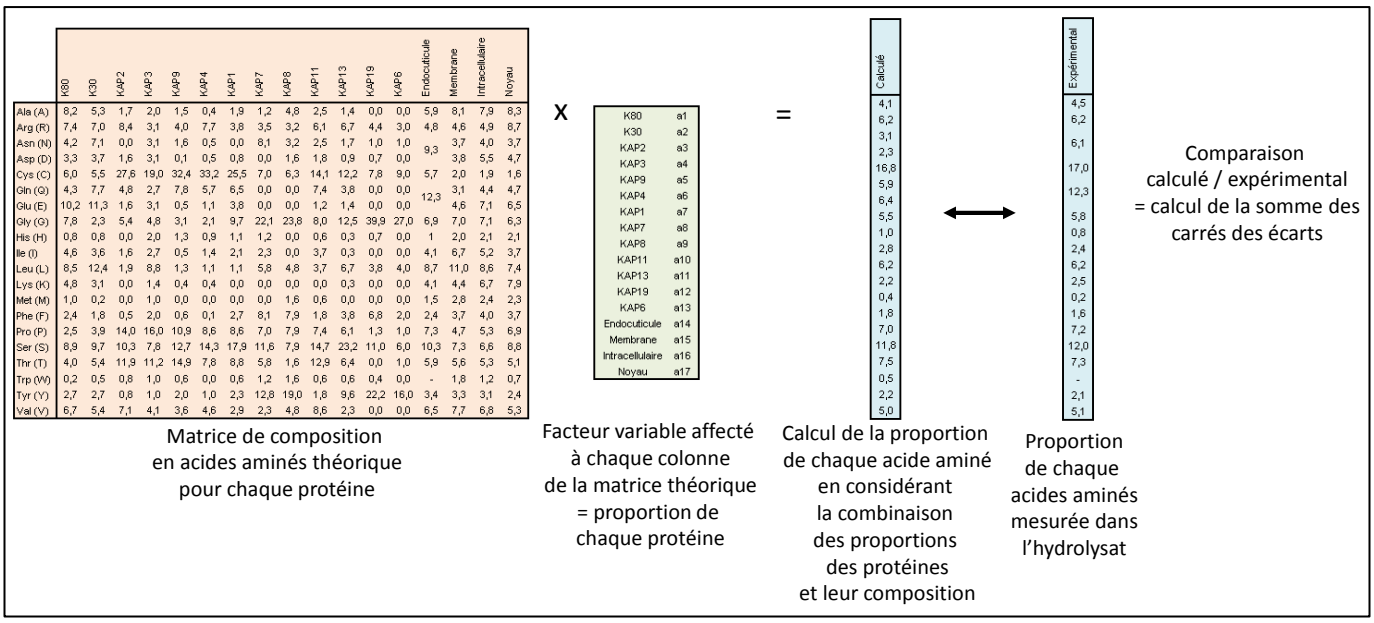


Figure 1 *: Principe de calcul de la composition en acides aminés d'un mélange de protéines présentes en différentes proportions à partir de leurs compositions respectives en acides aminés connues sur la base de leurs séquences et de la pondération pour chacune d'un facteur de proportionnalité.*

Il convient alors de trouver la combinaison des facteurs telle que la somme des carrés des écarts soit la plus faible possible. Pour cela, nous avons eu recours à un algorithme évolutionniste (intégré au complément Solver, Microsoft Excel 2010) pour réaliser cette minimisation qui peut s'assimiler à une optimisation. L'algorithme fonctionne en faisant varier les valeurs des facteurs fixés initialement. Nous fixons des conditions définissant un domaine de variation pour chacun des facteurs (ici le même pour chaque facteur) et une condition fixant un critère d'arrêt lorsque l'algorithme a atteint une valeur (ici la plus faible possible, pour la valeur correspondant à

la somme des carrés des écarts). Une boucle est ainsi constituée. L'algorithme construit au cours d'un premier essai une population d'individus chacun constitué de combinaisons des facteurs différentes et obtenues aléatoirement. L'algorithme sélectionne parmi l'ensemble des individus obtenus une sous population satisfaisant le mieux les critères définis (dans notre cas une diminution de la somme des carrés des écarts). On parle de reproduction des individus sélectionnés. Les nouveaux individus subissent ensuite des modifications, dites mutations, de leur facteurs variables ce qui constitue une nouvelle population d'individus dans laquelle une sous population est à nouveau sélectionnée comme précédemment. Les cycles se poursuivent jusqu'à ce que l'algorithme ne parvienne plus, au sein d'une nouvelle population, à obtenir d'individus satisfaisant mieux le critère défini initialement ou si le critère d'arrêt est atteint pour un individu. L'algorithme propose alors l'individu ayant le mieux satisfait les critères définis.

3. Quantification des familles multigéniques dans le protéome cortical humain

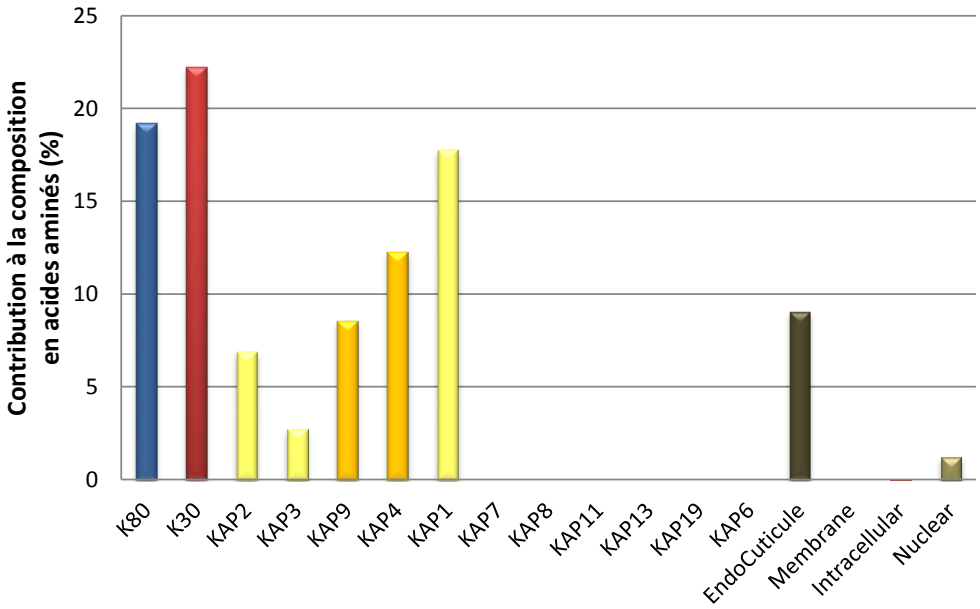
Nous avons ainsi employé cette stratégie pour évaluer la composition du cortex. Nous avons utilisé une composition d'acides aminés du cortex publiée [50, 51] et jugée représentative d'autres sources de la littérature [162]. Afin de tenir compte d'une éventuelle contamination cuticulaire de l'échantillon, nous avons intégré à la matrice une colonne correspondant aux valeurs expérimentales de la composition en acides aminés de l'endocuticule déterminé dans la même série d'études. Nous avons par ailleurs ajouté trois colonnes supplémentaires contenant les valeurs moyennes de composition en acides aminés pouvant être attendues pour les protéines membranaires, intracellulaires et nucléaires [284].

Le résultat proposé par l'algorithme permet d'obtenir un modèle se rapprochant de très près de la composition expérimentale (Figure 2). La contribution des protéines établie dans ce modèle montre des proportions proches entre les kératines de type I et de type II et donc tout à fait cohérentes avec ce qui est attendu compte tenu de la composition en hétérodimères des microfibrilles.

Nous noterons que, comme pour les résultats obtenus lors du comptage de spectres, la mesure obtenue pour les kératines de type I est légèrement supérieure à celle obtenue pour les kératines de type II. Ces résultats montrent également que les KAP des familles 1, 4, 9 et 2 pourraient apporter la plus grande contribution à l'espace intermicrofibrillaire. Contrairement aux résultats par comptage de spectres qui suggéraient une contribution importante des KAP 3, cette famille est ici peu représentée. Ce résultat est néanmoins à relativiser avec le fait que les KAP 3 sont plus courtes et qu'à quantité de protéine équivalente elles apportent moins d'acides aminés que les protéines contenant plus de résidus comme les KAP 1 ou 4. La contribution des familles de KAP dont les sites de coupures trypsiques dans leurs séquences sont défavorables compte tenu de la présence de proline est ici montrée comme importante et particulièrement pour les KAP 1. Cela confirmerait la sous évaluation de leur présence dans les analyses protéomiques. L'abondance en cystéine, en proline et en sérine retrouvées dans les hydrolysats peut ainsi être vraisemblablement attribuée à la forte expression de ces KAP très riches en soufre.
Les KAP HGT ne semblent pas contribuer significativement à la composition en acides aminés tout comme les KAP 11 et 13. La détection de ces protéines précédemment en protéomique montre cependant qu'elles peuvent être détectées même parmi les kératines et les autres KAP plus abondantes mais aussi plus difficiles à digérer. Nous les considèrerons par la suite comme négligeables dans la structure finale du cortex.

Une contribution significative d'éléments de composition voisine de ce qui est retrouvé dans l'endocuticule est retrouvée ce qui pourrait correspondre à une contamination cuticulaire mais qui pourrait également être une contribution des rémanents cellulaires du cortex de composition probablement voisine de celle de l'endocuticule. Une contribution de composition nucléaire est également détectée.

	Cortex		
	Expérimental	Calculé	Ecart
Ala (A)	4,5	4,1	0,4
Arg (R)	6,2	6,2	0,0
Asn (N)	6,1	2,7	0,6
Asp (D)		2,8	
Cys (C)	17,0	16,7	0,3
Gln (Q)	12,3	5,5	0,1
Glu (E)		6,7	
Gly (G)	5,8	5,5	0,3
His (H)	0,8	0,9	-0,1
Ile (I)	2,4	2,9	-0,5
Leu (L)	6,2	6,1	0,1
Lys (K)	2,5	2,2	0,3
Met (M)	0,2	0,5	-0,3
Phe (F)	1,6	1,8	-0,1
Pro (P)	7,2	7,0	0,2
Ser (S)	12,0	11,9	0,1
Thr (T)	7,3	7,5	-0,2
Trp (W)	-	0,4	-
Tyr (Y)	2,1	2,3	-0,2
Val (V)	5,1	5,1	0,0
		$\Sigma E^2 =$	1,1



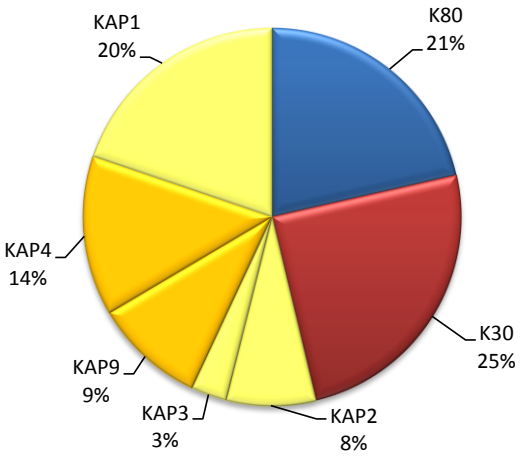


Figure 2 *: Résultats de la quantification obtenus par modélisation de la composition en acides aminés du cortex.*

4. Conclusion

Une évaluation de la stratégie mériterait d'être réalisée sur des mélanges quantifiés standards afin d'évaluer la précision et la justesse de la méthode de quantification proposée. Par ailleurs, l'utilisation de plus de données expérimentales issus d'hydrolysats d'extraits pourrait être envisagée afin d'appuyer ou non les résultats décrits dans ce chapitre. Néanmoins, la composition en acides aminés des hydrolysats parait très bien modélisée en considérant des proportions importantes des familles riches en soufre identifiées préalablement en protéomique. Il apparait ainsi raisonnable de considérer les familles 1, 4, 9 et 2 comme composants principaux de la matrice interfilamentaire au sein des macrofibrilles.

Chapitre II Etude des séquences particulières des KAP : une clé pour la compréhension de leur origine et de leur fonction

Afin de comprendre le rôle des KAP aux vues de leur importante expression dans l'espace interfilamentaire, nous avons recherché des informations pouvant permettre d'appréhender la structure de ces protéines.
Les structures des protéines sont classiquement déterminées grâce à des techniques analytiques comme la cristallographie ou la spectroscopie par résonance magnétique nucléaire. Ces techniques nécessitent généralement de disposer des protéines purifiées voire cristallisées. L'expression de ces gènes dans des bactéries s'avère très compliquée, principalement à cause de la forte abondance en cystéines de ces protéines. Nous avons précédemment vu, en seconde partie de ce manuscrit, les difficultés que pose l'isolement de ces protéines à partir des extraits corticaux.
Une alternative pourrait être utilisée en recherchant des protéines présentant des homologies de séquence avec les différentes familles de KAP et dont la structure tridimensionnelle a été par ailleurs étudiée. L'unicité des séquences de KAP conjuguées à l'abondance de cystéines dont l'arrangement doit influer notablement sur la structure, ne permet pas d'envisager de trouver des homologues suffisamment représentatifs pour réaliser cette recherche.
Dans tous les cas, il semble peut significatif d'envisager l'étude des structures de ces protéines seules sans envisager leur conformation adoptée en interaction avec les microfibrilles comme c'est le cas *in vivo*.
Une autre solution pourrait être la synthèse de segments caractéristiques de ces protéines et d'étudier l'interaction de ces fragments avec des filaments intermédiaires reconstitués *in vitro.* Cette méthode nécessiterait un important arsenal technique dont nous ne disposons pas. Les méthodes de pontages chimiques directement sur la structure pourraient être envisagées. Les principaux réactifs employés dans ce but utilisent comme substrats les fonctions des résidus des lysines qui bien qu'abondantes sur les tiges des kératines sont quasiment absentes des séquences des KAP attendues.
Nous avons donc choisi de nous intéresser à l'étude des séquences de ces protéines pour tenter de dégager certaines propriétés pouvant nous conduire à des indices de fonctions structurales. Cette étude des séquences a déjà été en partie réalisée dans les travaux de Parry et al. [84], mais nous avons choisi de les approfondir à la lumière de nouvelles données de séquences notamment obtenus chez d'autres mammifères dont les génomes séquencés et annotés ont pu être obtenus ces 5 dernières années.

1. Des structures pentapeptidiques spécifiques et conservées

Les premières analyses des séquences de KAP identifiées à la suite de la découverte des gènes ont montré des structures répétitives particulières pour une partie de ces familles présentes dans le cortex et principalement réparties sur les familles respectivement 4, 9, 1 et minoritairement sur les KAP 2 et 3. Les KAP 4 sont les isoformes qui contiennent chez l'humain le plus de ces motifs.
L'analyse particulière des séquences en acides aminés des isoformes de la famille 4 révèle un enchainement de motifs pentapeptidiques comprenant une cystéine en début et en fin de motif. Les motifs s'enchainent sur la majorité de la séquence exceptée des côtés N-terminaux et C-terminaux et l'enchainement pouvant s'interrompre sporadiquement dans la partie centrale de quelques isoformes (Figure 1).
Ces motifs pentapeptidiques peuvent se découper en deux catégories en fonction du résidu central : le résidu se retrouvant au centre du motif peut être une proline (motif CXPXC) ou un résidu portant une fonction alcool, une serine ou une thréonine (motif CX(S/T)XC).

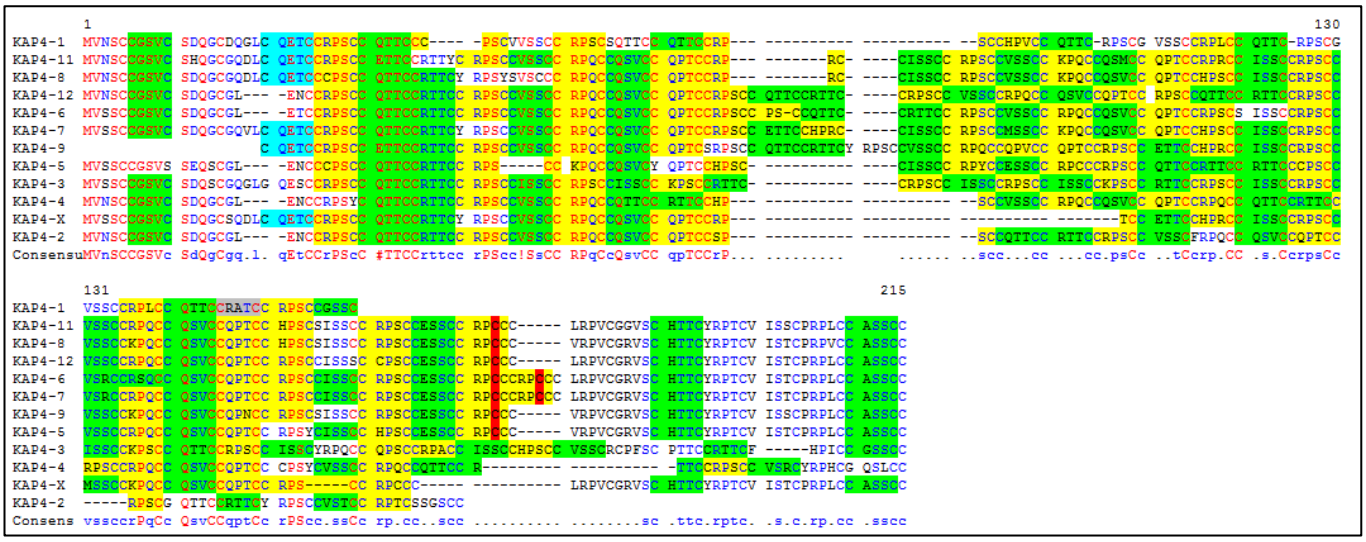


Figure 1 *: Séquences protéiques des KAP 4 humaines et position des motifs penta peptidiques CXPXC (en jaune) et CX(S/T)XC (en vert).*

Dans le cas du motif CXPXC, le second résidu est un acide aminé basique (le plus souvent arginine puis lysine et histidine) ou une glutamine. Le résidu en quatrième position est alors le plus souvent une serine, une thréonine ou une glutamine.

Dans le cas du motif CXSXC, un des résidus autour de la serine est dans la quasi-totalité des cas un acide aminé apolaire de type valine ou isoleucine. L'autre acide aminé restant peut être une glutamine en deuxième position (motif CQSVC) ou une sérine en quatrième position (motif CVSSC). Lorsque l'acide aminé central est une thréonine, le second acide aminé est une glutamine ou une arginine, le quatrième acide aminé étant alors une thréonine (motifs CQTTC ou CRTTC).

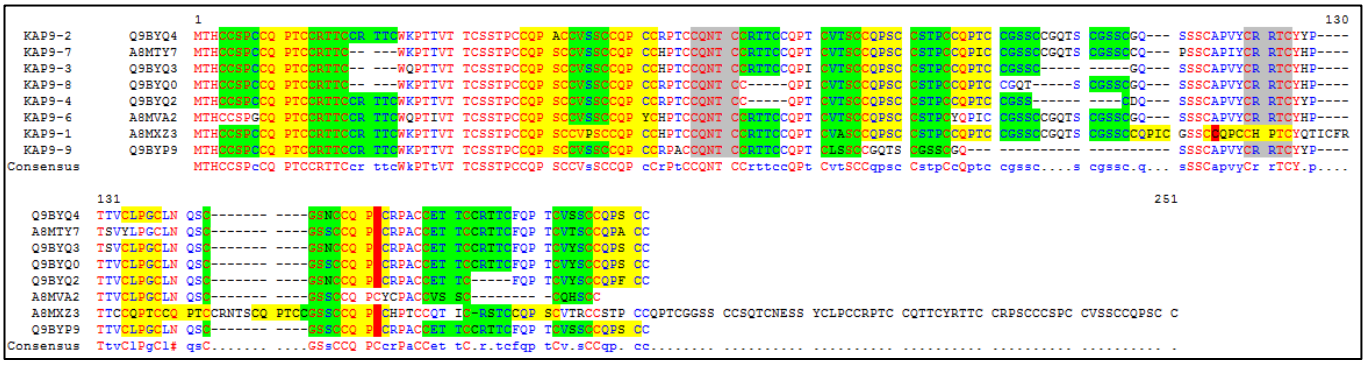


Figure 2 *: Séquences protéiques des KAP 9 humaines et position des motifs penta peptidiques.*

Pour les isoformes de la famille 9, les mêmes motifs peuvent être retrouvés mais leur répartition au sein de la séquence est différente de celle de la famille 4. La régularité de l'enchainement est moins parfaite avec des insertions entre les motifs et une structure des segments N-terminaux et C-terminaux différente de celle observée pour la famille 4. Les motifs d'insertion peuvent être plus ou moins longs (typiquement de 4 à 14 résidus).

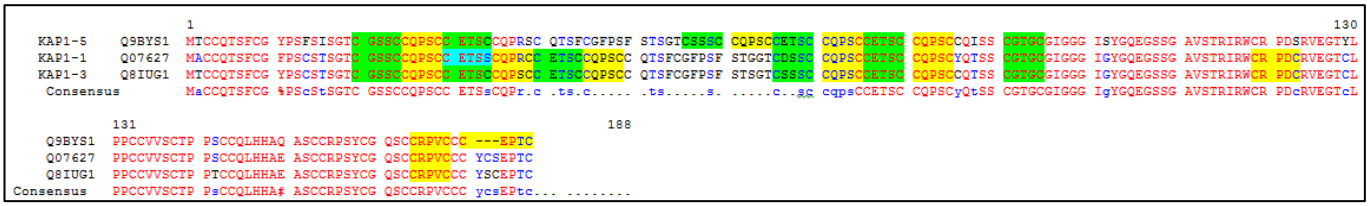


Figure 3 *: Séquences protéiques des KAP 1 humaines et position des motifs penta peptidiques.*

Les isoformes de la famille 1 présentent moins de ces motifs qui sont essentiellement présents sur un segment après le début de la séquence. La spécificité de ce segment est de comporter des motifs de type (CETSS) dans lequel un résidu acide glutamique est impliqué en position 2. Un à deux motifs sont également présents en fin de séquence.
La famille très homologue des KAP 2 ne présente que quelques motifs dispersés sur la séquence bien que des structures imparfaites de ces motifs puissent être décelées dans le reste de la séquence (CRPIT, CRPPC, RTSSC…), signes probables d'une perte de ces motifs au cours de l'évolution. Concernant les isoformes de la famille 3, seul un motif est retrouvé au sein de ces séquences.

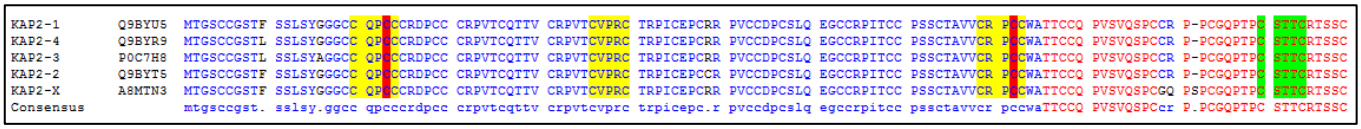


***Figure 4a** : Séquences protéiques des KAP 2 humaines et position des motifs penta peptidiques.*

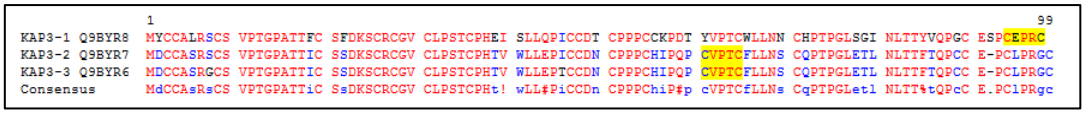


***Figure 4b** : Séquences protéiques des KAP 3 humaines et position des motifs penta peptidiques.*

Il est ainsi possible par examen des séquences des familles de KAP du cortex de les classer en fonction de l'occurrence de ces motifs caractéristiques.

2. La conservation des séquences pendant l'évolution des mammifères : un indice de la conservation d'une fonction des protéines

a) Les KAP à travers l'évolution des mammifères

Afin d'évaluer l'importance des KAP du cortex et comprendre l'importance que pourraient avoir certains éléments de leurs séquences, nous nous sommes intéressés à la comparaison des séquences humaines avec celles d'autres mammifères. En 2008, Wu et *al*. ont étudié, parmi différents mammifères dont le génome a été séquencé, la présence des différents gènes des KAP. Cette étude a permis, entre autre, de montrer l'ensemble des gènes des familles de KAP communs à l'ensemble des mammifères étudiés et de montrer l'évolution de la multigénicité parmi ces espèces. La présence parmi ces mammifères de l'ornithorynque permet de connaitre quels étaient les gènes déjà présents au moment de la divergence des différents groupes de mammifères existant encore aujourd'hui (Figure 5). L'estimation de la période à laquelle a évolué l'ancêtre commun de tous les mammifères est d'environ 170 millions d'années. Il existe également des preuves paléontologiques de l'existence de poils chez les mammifères antérieures à cet événement. Nous pouvons en déduire que les gènes que portaient cet ancêtre commun sont donc nécessaires et suffisants à la formation de la structure capillaire conservée depuis chez la quasi-totalité des mammifères.
Parmi ces gènes sont retrouvés, pour ceux exprimés dans le cortex, ceux correspondants aux KAP riches en soufre 4, 1, 2 et 3 pour les protéines majoritaires et ceux correspondants aux KAP 11 et 13. Les gènes correspondants aux KAP riches en glycine et en tyrosine 8, 20 et 21. Il apparait ainsi à la vue de cette comparaison que les KAP 9, 7, 6, 19 sont apparues plus tard et peuvent donc être considérées comme des gènes supplémentaires apportés au cours de l'évolution. Nous pouvons également considérer le fait que les KAP 13 et 21 ont été perdues chez l'espèce marsupiale considérée indiquant que ces gènes ne sont pas critiques pour la conservation de la fibre. Néanmoins, une perte de ces gènes n'exclut pas que des gènes d'autres KAP présentant certaines relations d'homologie ne compensent pas cette absence.

En ce qui concerne les KAP cuticulaires, nous retrouvons les KAP 5, 10, 16, 17, 26, 28 et 29 comme gènes initiaux. Les gènes correspondant aux KAP 12, 24, 25 et 27 sont apparus plus tardivement au cours de l'évolution. Nous noterons que la conservation de la famille 28 chez tous les mammifères avait ainsi indiqué qu'une forte suspicion pouvait exister quant à son expression dans le protéome humain. Certaines similitudes dans leur séquence avec les KAP 5 et 17 dont les transcrits avaient été décrits dans la cuticule nous avait encouragé à chercher avec succès cette famille dans ce type cellulaire.

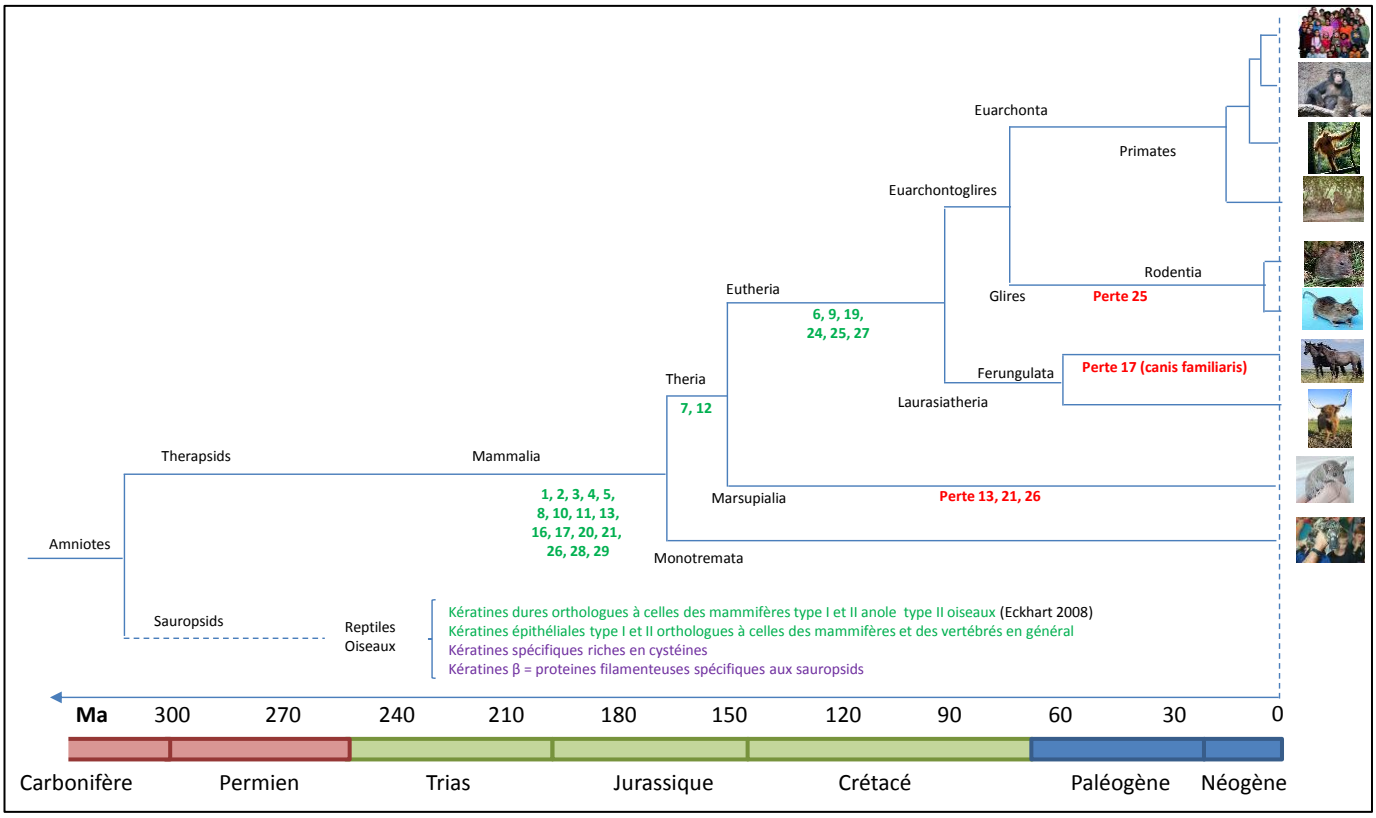


Figure 5 *: Reconstitution de l'arbre phylogénique des différentes lignées de mammifères dont les gènes des KAP ont été étudiés. Les résultats obtenus par les études de Wu et al. sur l'apparition ou de disparition des gènes des KAP sont présentés. Nous noterons que les espèces dont le génome a actuellement été séquencé et issus de la branche comprenant les reptiles, les oiseaux et les tortues ne semblent pas posséder de gènes similaires aux KAP dans les génomes récemment étudiés [138, 285].*

b) Comparaison inter mammifères des séquences protéiques des KAP majoritaires du cortex

Parmi les gènes des KAP conservés depuis 170 millions d'années, nous pouvons supposer que le mécanisme d'évolution a opéré sur les séquences de chaque espèce. La comparaison de ces séquences devrait pouvoir révéler des zones ou des structures de séquences susceptibles d'avoir été conservées indépendamment les unes des autres ou au contraire ayant fait l'objet de plus de mutations. Les zones conservées peuvent être assimilées à des morceaux de séquences nécessaires pour la fonction de la protéine.

Nous avons dans ce but extrait de la banque protéique NCBI les séquences des KAP du cortex correspondant à différentes taxonomies de mammifères dont les génomes ont été séquencés et annotés. La recherche de ces séquences a été réalisée en utilisant, pour chaque famille de KAP, une séquence humaine pour laquelle nous avons recherché les homologues dans chaque taxonomie avec l'outil « blast ». Les séquences homologues ainsi retrouvées sont alors enregistrées pour être par la suite étudiées.

Comparaison des orthologues de la famille des KAP 4

Pour la famille des KAP 4, 10 taxonomies (*Homo sapiens*, *Pan troglodytes*, *Pongo abelii*, *Macaca mulatta*, *Rattus norvegicus*, *Mus musculus*, *Equus caballus*, *Bos taurus*, *Monodelphis domestica* et *Ornithorhynchus anatinus*) ont

été étudiées. Pour chacune, nous avons recherché la présence des motifs pentapeptidiques afin de dégager une structure secondaire des isoformes dans chaque taxonomie. Les séquences ainsi étudiées pour chacune des taxonomies sont présentées en Annexe 6.

L'examen de ces motifs penta peptidiques montre que ces motifs décrits chez l'humain sont retrouvés sans exception pour l'ensemble des mammifères étudiés. Parmi les primates (*Pan troglodytes*, *Macaca mulatta* et *Pongo abelii*), les rongeurs (*Mus musculus* et *Rattus norvegicus*) et les Laurasiathériens (*Bos taurus* et *Equus cabalus*), les règles de disposition des résidus dans les motifs sont les mêmes. Pour des mammifères plus éloignés tels que les marsupiaux (*Monodelphis domestica*) et les monotrèmes (*Ornithorhynchus anatinus*), ces règles sont également retrouvées bien que d'autres motifs minoritaires puissent être observés comme l'association d'acides aminés apolaires dans les motifs CXPXC et CXTXC.

Concernant l'organisation des motifs au sein de la séquence, il est à noter que certaines espèces montrent des singularités dans l'enchainement de ces motifs. Chez *Monodelphis domestica*, les motifs sont retrouvés séparés alternativement de motifs tetrapeptidiques de type CRPC et CQPC. La présence de cette organisation est retrouvée également pour certains gènes d'*Ornithorhynchus anatinus*, ce dernier possédant également de propres variantes de ces insertions tetrapeptidiques avec des motifs PQPC et PRPC.

Pour les primates et les rongeurs, la succession de répétitions pentapeptidiques est quasiment parfaite. En revanche, les séquences des Laurasiathériens étudiés montrent des interruptions comme c'est le cas d'*Equus caballus* pour lequel il semble que des mutations d'une des deux cystéines d'une partie des motifs aient été réalisées au cours de l'évolution. Cette propriété est également observée chez *Bos taurus* pour lequel des insertions de tri et tétra peptides complètent également l'organisation de la séquence. Au sein de chaque espèce, le nombre de motifs de répétition peut présenter des hétérogénéités plus ou moins prononcées (de 15 à 65 répétitions de motifs chez les isoformes de *Pongo abelli*, de 25 à 35 chez *Rattus norvegicus*).

```
  NN        I   E
MVSSCCGSVCSDXSCG…
          T   G G
```

***Figure 6** : Motif N-terminal consensus des KAP 4 retrouvé pour l'ensemble des mammifères étudiés contenant une extrémité relativement hydrophobe, acide et des résidus cystéines conservés.*

Nous noterons que si les séquences centrales présentent quelques disparités, la séquence N-terminale présente une structure particulièrement bien conservée chez l'ensemble des mammifères étudiés.

Côté C-terminal, il est plus difficile de discerner des motifs consensuels mais nous pouvons noter que pour l'ensemble des mammifères étudiés, ces portions de séquences contiennent toujours des cystéines qui ne sont pas impliquées dans un motif pentapeptidique contenant 2 cystéines en position 1 et 5. Pour le groupe des thériens et le marsupial une cystéine isolée est toujours retrouvée en dernier résidu de la séquence. La partie C-terminale des séquences du monotrème semble plus différente des autres mammifères mais une ambiguïté sur l'assignation de l'interruption de la fin de séquence pourrait exister.

Comparaison des orthologues de la famille des KAP 1

11 taxonomies ont été examinées. Les séquences d'Ovis *aries* ont été trouvées en supplément. La comparaison des séquences entre les primates montre une très forte homologie des séquences suffisante pour considérer celles de l'humain comme représentative des autres séquences. Nous noterons que la séquence tronquée en N-ter de la KAP 1.4, qui nous avait amené précédemment à l'envisager comme pseudo gène, n'est retrouvée que pour Pan troglodyte.

L'alignement des séquences de 8 taxonomies a ainsi été réalisé. Cet alignement permet de distinguer différents segments dans ces séquences protéiques, déjà envisagés auparavant [84] mais pour lesquels la comparaison avec les autres mammifères permet d'affiner les propriétés des séquences primaires. Il est à noter que la recherche de structures « classiques » de type hélice α ou coude β ne donne pas de résultats sur ces séquences primaires.

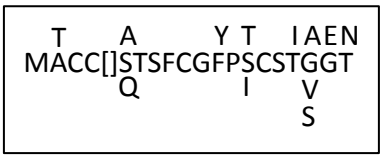


Figure 7 *: Motif N-terminal consensus des KAP 1 retrouvé pour l'ensemble des mammifères étudiés contenant une extrémité hydrophobe et des résidus cystéines conservés.*

La comparaison permet de montrer un segment N-terminal très conservé exception faite des séquences d'*Ornithorhynchus anatinus* qui semblent tronquées. Ce segment comporte quelques acides aminés hydrophobes et 2 positions de résidus portant systématiquement un résidu aromatique.

Après ce segment, nous retrouvons un segment de longueur variable entre les différents groupes et composé de motifs pentapeptidiques parfaits ou imparfaits (mutation de la cystéine en position 1 ou 5). Ce segment de longueur variable correspond à celui où sont retrouvées chez l'humain des insertions de motifs pentapeptidiques responsables du polymorphisme en taille de certains gènes de la famille [258]. Il apparait donc à la lumière de cette comparaison que ce segment est pour l'ensemble des mammifères soumis à ce phénomène d'insertion.

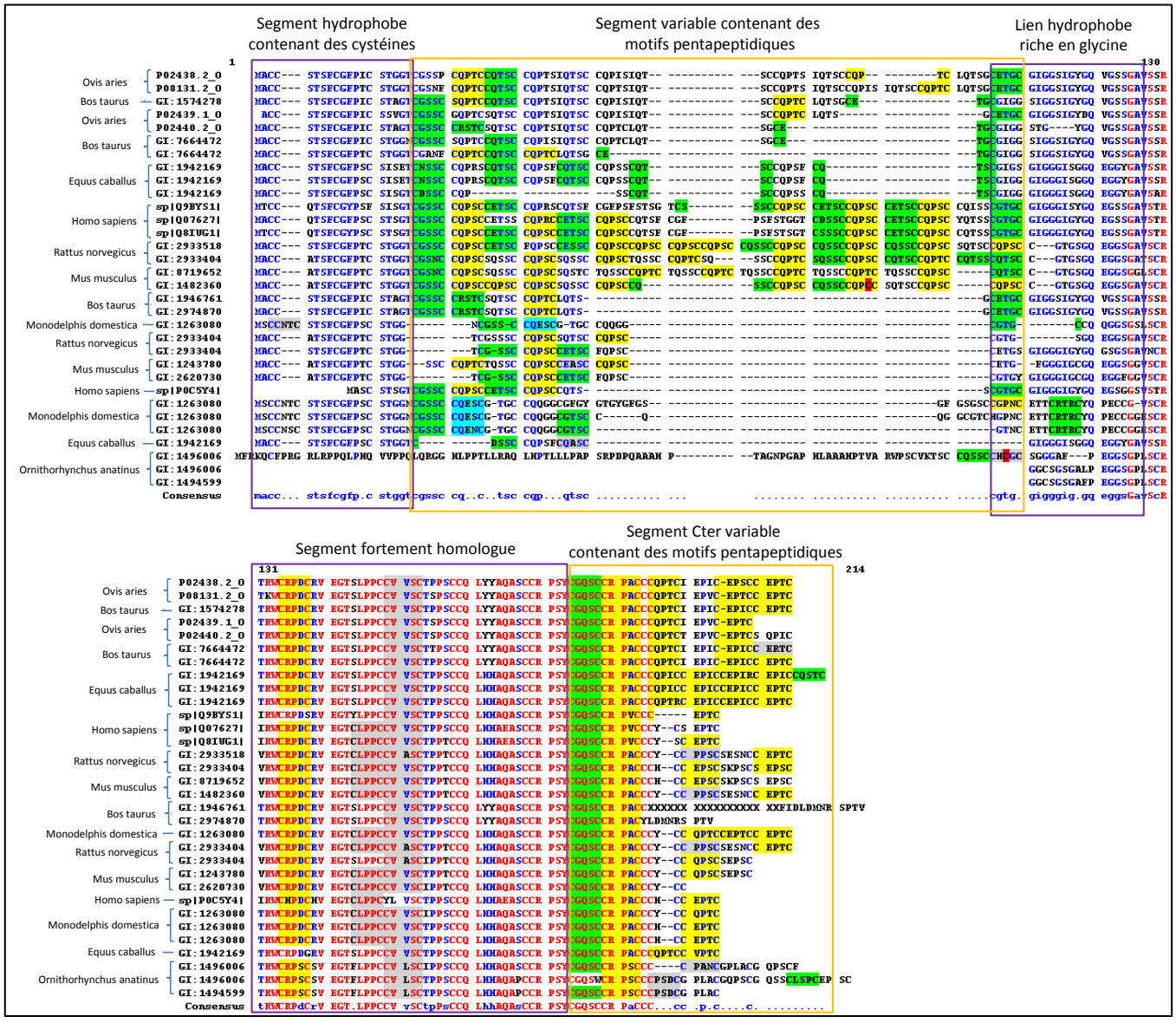


Figure 8 *: Alignement des séquences des KAP 1 de différents mammifères et identification des segments.*

Après ce segment est systématiquement retrouvé un segment caractéristique riche en glycines. Compte tenu du faible encombrement apporté par ces résidus, il est possible d'envisager ce segment comme un lien permettant une certaine flexibilité entre les deux parties de la protéine.

Le segment suivant montre une région très conservée chez l'ensemble des mammifères. Cette forte homologie de séquence associée à l'absence d'événements de types insertion/délétion suggère une zone structurée de la protéine.

Le segment C-terminal est de longueur variable et comporte de nouveau des motifs pentapeptidiques. L'enchainement de ces motifs est imparfait et toujours séparé par 1 à 4 résidus contenant au minimum une cystéine.

En comparant les motifs pentapeptidiques de la famille 1 et de la famille 4, il apparait une différence notable. Si les protéines de la famille 4 contiennent principalement en position 2 des arginines et des glutamines, la famille 1 en plus de ces résidus, est caractérisée par la présence de résidus acides glutamiques. Ces motifs pentapeptidiques « acides » sont retrouvés sur la partie C-terminale mais également sur le second segment des orthologues des primates et d'une des isoformes de *Rattus norvegicus*.

Comparaison des orthologues de la famille des KAP 3

La comparaison des orthologues des KAP 3 a également été réalisée sur 9 taxonomies. La comparaison des alignements montre que pour l'ensemble des séquences, les motifs pentapeptidiques sont largement sous représentés. Contrairement aux familles examinées précédemment, une segmentation des séquences ne peut être établie. Vers le début de séquence un segment de très forte homologie contenant plus d'une vingtaine de résidus peut être observé. Nous notons également dans la suite de la séquence certaines positions pour lesquelles les résidus ont été conservés pendant plus de 170 millions d'années.

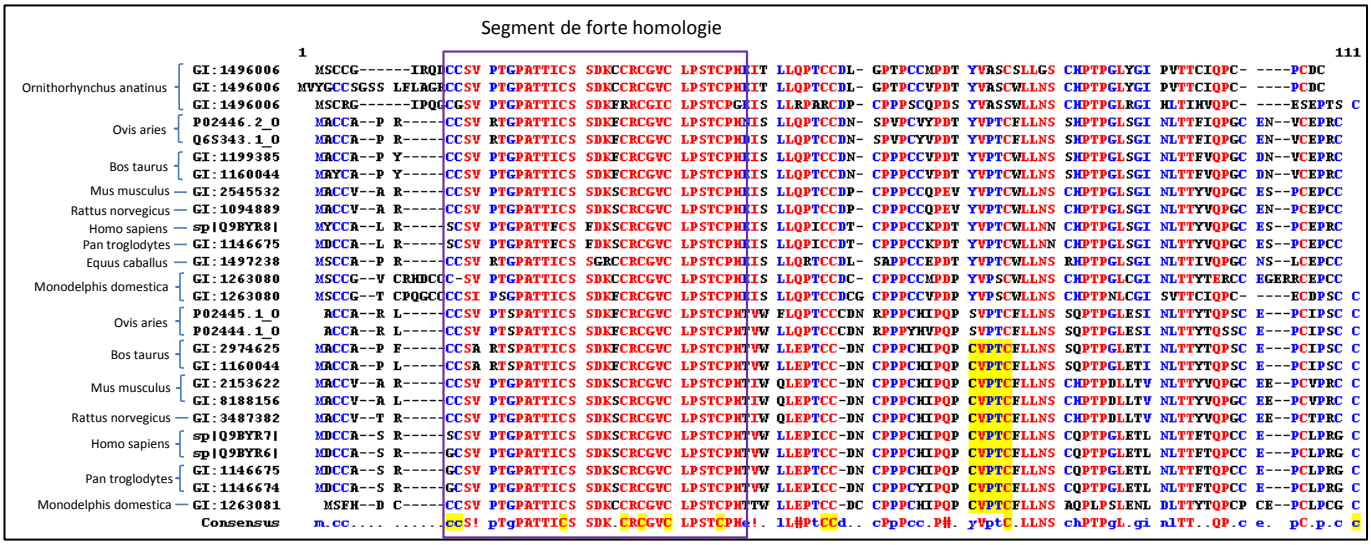


Figure 9 *: Alignement de séquence des KAP3 de différents mammifères. L'alignement montre une zone extrêmement conservée depuis plus de 170 millions d'années (d'après la comparaison avec les séquences de l'ornithorynque) qui pourrait correspondre à une zone fonctionnelle de la protéine.*

Ces homologies peuvent de nouveau suggérer la présence d'un motif structural dont la composition et la distribution des résidus sont essentielles pour la fonction de la famille de protéine.

Conclusion

La conservation du motif pentapeptidique chez les mammifères au cours de leurs évolutions indépendantes suggère que ces motifs ont un lien direct avec le rôle sans aucun doute structural d'une partie des familles de KAP riches en soufre et particulièrement les familles 4, 9 et 1. La connaissance de l'arrangement spatial de ces motifs

pourrait donc être une clé pour comprendre la structure de ces protéines et leurs interactions avec les filaments de kératine dont la structure et l'arrangement au sein de la cellule corticale sont mieux appréhendés.
Ces comparaisons pourraient être mises en parallèle d'études des propriétés microstructurales et mécaniques des fibres des différents mammifères ici étudiés. Les différences de structures primaires des protéines acquises au cours de l'évolution pourraient peut être expliqué des différences de souplesse, d'élasticité ou de capacité à obtenir des poils plus longs chez des groupes d'espèces en particulier.

3. Modélisation structurale des motifs consensuels composants les familles majoritaires des KAP du cortex

L'étude de Parry et *al*. a montré la possibilité, au sein d'un motif pentapeptidique CXPXC, de constituer un pont disulfure probablement favorisé par la présence de la proline contraignant le rapprochement des deux cystéines. Le motif ainsi formé est une boucle de 5 acides aminés. La conclusion des auteurs concernant les autres motifs (de type CX(S/T)XC) était que l'absence de proline rendait le rapprochement moins probable. Ces données ont été obtenues par modélisation moléculaire utilisant des données cristallographiques et de RMN obtenues sur des protéines possédant des motifs voisins de type CXPXC et CX(S/T)(S/T)C.
Néanmoins, la possibilité d'expliquer le grand nombre de ponts disulfures intra protéiques au sein de la cellule corticale par la formation systématique de boucles penta peptidiques au sein des KAP 4, 9 et 1 n'est pas à écarter. Cette hypothèse permettrait d'expliquer la conservation systématique de ces motifs parmi certaines familles de KAP des mammifères et de relier la position de ces cystéines à un rôle de structuration secondaire de ces protéines. Elle permettrait également d'apporter une explication à la relative facilité d'extraction en milieu réducteur des protéines de la matrice compte tenu du très grand nombre de cystéines comprises dans leur séquence [149].

a) Modélisation

Afin d'éprouver l'hypothèse d'une formation systématique des boucles penta peptidiques pendant l'oxydation se déroulant au cours du processus de kératinisation, nous avons choisi d'utiliser une stratégie de modélisation moléculaire.
Considérant que la boucle doit être formée lorsque les cystéines sont à une distance permettant la formation de la liaison pendant la réaction d'oxydation, nous allons étudier les liaisons non covalentes pouvant s'établir et les interactions stériques pouvant stabiliser ou déstabiliser cette boucle. Nous sommes partis de l'hypothèse que les KAP, synthétisées après la formation des microfibrilles, sont incorporées en solution au sein des espaces inter micro fibrillaires puis immobilisées. Les observations microscopiques des zones de transition entre la zone de kératinisation et la zone de consolidation suggèrent une déshydratation de la structure suivie d'une disparition des résidus cystéines libres pendant l'oxydation. Ces données permettent d'imaginer, à ce moment, des motifs dans lesquels sont insérés des molécules d'eau. Suite à la déshydratation, des liaisons hydrogènes doivent pouvoir s'établir au sein de ce motif. Nous allons donc étudier les mécanismes d'arrangement pouvant conduire au rapprochement des cystéines en formant *in silico* sur des modèles ces liaisons hydrogènes.

L'étude a été réalisée sur différentes séquences de motifs en utilisant le logiciel ChemBio3D Ultra (CambridgeSoft). Les motifs tridimensionnels ont été dessinés avec la conformation correcte des acides aminés. Afin de tenir compte d'éventuelles interactions provoquées par les fonctions des liaisons peptidiques à chaque extrémité du penta peptide, les cystéines adjacentes au motif ont été ajoutées. Leurs extrémités N-terminale et C-terminale ont été remplacées par des groupes encombrants tert-butyl afin de modéliser l'encombrement des chaines peptidiques adjacentes sans ajouter de fonctions polaires supplémentaires.

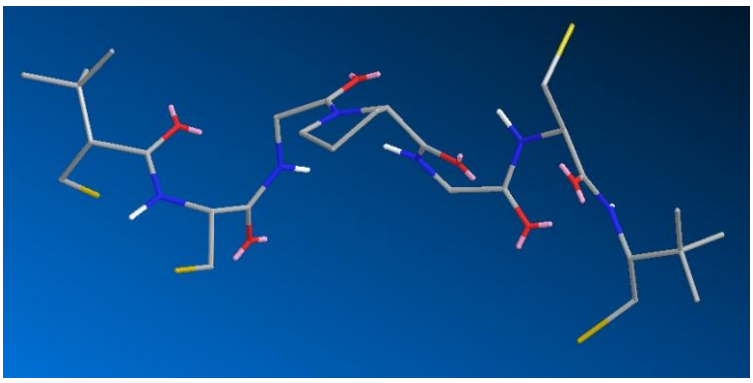


Figure 10 *: Motif tBuC(CGPGC)CtBu déstabilisé aléatoirement. Le système ne comporte aucune liaison hydrogène intra moléculaire comme cela pourrait être le cas en solution aqueuse.*

L'édifice moléculaire ainsi constitué est déstabilisé en lui conférant de l'énergie cinétique permettant d'établir une conformation aléatoire dans laquelle ne s'établissent pas de liaisons hydrogènes. La structure obtenue après déstabilisation est placée volontairement dans une conformation linéaire dans laquelle aucune liaison hydrogène n'est établie. Le modèle initial est ainsi libre de toute interaction intramoléculaire. Nous considérons ici une structure totalement solvatée telle qu'elle pourrait exister en solution aqueuse. Un calcul « MM2 dynamic » est ensuite appliqué à la molécule pendant lequel sa conformation est simulée à 300 K avec des pas de 10 K. La structure peut commencer à former des liaisons hydrogènes mais l'énergie appliquée au système est suffisamment importante pour que ces liaisons formées ne soient pas permanentes et puissent se défaire si une autre conformation de structure permettait de minimiser l'énergie de l'édifice. L'énergie de la structure finalement obtenue est alors minimisée par un calcul MM2 et des liaisons hydrogènes peuvent s'établir définitivement. Les conditions en termes d'interactions entre fonctions permettant la formation du coude sont au fur à mesure examinées.

b) L'établissement du réseau de liaisons hydrogènes dans la boucle penta peptidique

Différents essais ont suggéré que le rapprochement des deux cystéines était possible dès lors que se formaient des liaisons hydrogènes notamment entre les carbonyles et les protons portés par les azotes de la chaine peptidique. Ce rapprochement n'est pas systématique et peut être contrarié par les interactions des fonctions des chaines latérales avec notamment les fonctions amides présentes à l'extérieur du motif penta peptidique. Un événement est commun à tous les motifs expérimentés ayant permis le rapprochement : l'établissement d'une interaction entre le carbonyle de la liaison amide de la cystéine 1 avec l'hydrogène de la liaison amide de la cystéine 4 ou éventuellement avec l'hydrogène de la fonction alcool pouvant être porté par la sérine ou la thréonine en position 3.

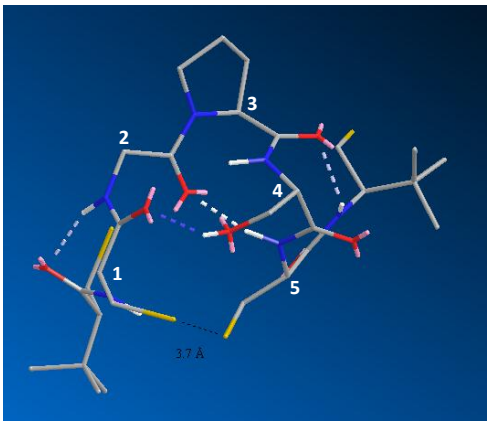


Figure 11 *: Motif tBuC(CGPSC)CtBu stabilisé. Dans cette conformation les cystéines sont voisines, la boucle formée est stabilisée par deux liaisons hydrogène (C=O de la cystéine en 1 avec OH de la serine en 4 ; C=O de la position 2 avec NH de la cystéine en 5).*

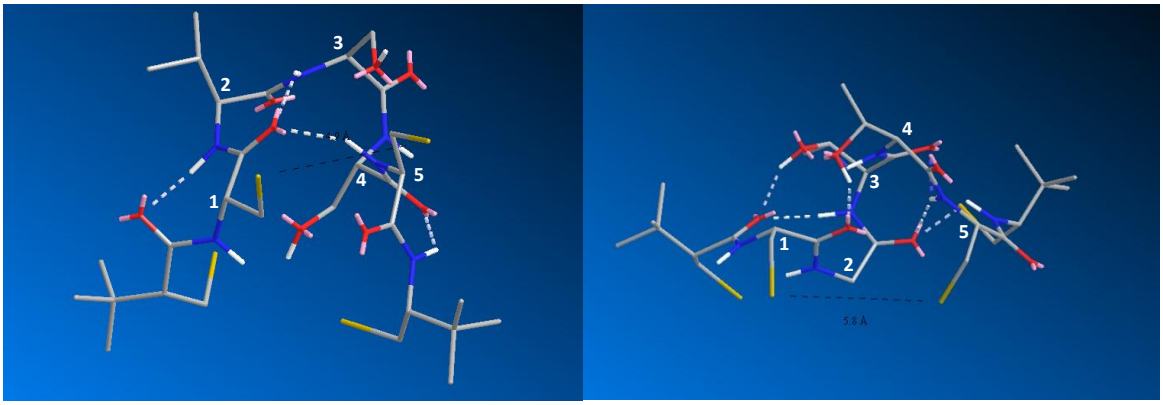


Figure 12 *: A gauche motif tBuC(CVSSC)CtBu stabilisé par deux liaisons hydrogène (C=O de la cystéine en 1 avec NH de la serine en 3 et de la serine en 4). A droite motif tBuC(CGTTC)CtBu stabilisé (C=O de la cystéine en 1 avec OH de la thréonine en 4 et C=O de la position 2 avec NH de la cystéine en 5).*

En résumé, la présence de la proline en position 3 dans le penta peptide semble favoriser l'interaction des fonctions de la chaine peptidique pour la formation du coude. Lorsqu'il s'agit d'une sérine ou d'une thréonine, la possibilité de formation d'une liaison hydrogène avec les carbonyles de la chaine peptidique doit également permettre le rapprochement. La présence en position 4 d'une sérine ou d'une thréonine permet l'interaction avec le carbonyle du côté C-ter de la cystéine 1. Cette liaison non covalente permet de verrouiller la boucle et peut être complétée ou substituée par la liaison avec l'amine de la chaine peptidique entre les positions 3 et 4.

c) Détection d'une interférence possible des résidus longs et polaires limitant la formation de la boucle

L'acide aminé en seconde position est communément une glutamine ou une arginine. La modélisation suggère, notamment dans le cas où l'arginine est à cette position, que la fonction interfère avec la formation de la boucle. En effet, le caractère polaire de l'arginine et la longueur de la chaine latérale permet la formation de liaisons polaires entre l'arginine et les fonctions consolidant la boucle, entrainant le déplacement des cystéines en opposition et défavorisant considérablement la formation du pont. Les motifs de type CRTTC ou CRPSC peuvent donc plus difficilement se cycliser.

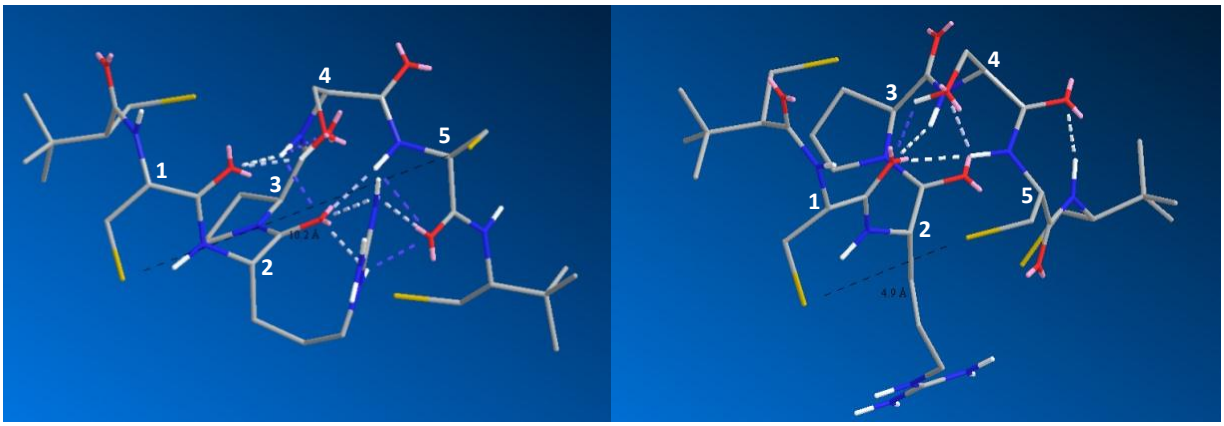


Figure 13 *: Motif tBuC(CRPSC)CtBu. A gauche, la chaine de l'arginine entre en interaction avec les fonctions de la chaine peptidique, repoussant les cystéines en opposition (distance supérieure à 10 Å). A droite, l'arginine est volontairement maintenue à l'extérieur du motif, les liaisons hydrogènes s'établissent dans la chaine et les cystéines se rapprochent (distance inférieure à 5 Å).*

En revanche, si l'arginine est en interaction avec une fonction complémentaire extérieure au motif penta peptidique, la cyclisation pourrait être réalisée. L'expérience de modélisation de ce motif montre en effet que si

l'arginine est maintenue hors de la boucle, l'établissement des liaisons hydrogènes favorables au rapprochement des cystéines peut être réalisé. La fonction ayant le plus d'affinité avec l'ammonium peut être un carboxylate, qui dans le cadre de la formation d'une liaison ionique pourrait exister initialement lorsque l'édifice est encore en solution. La faible abondance d'acides glutamique et aspartique dans les séquences des KAP 4 mais également dans les autres séquences des autres KAP corticales (proportions inférieures à 5%) ne permet pas d'envisager ce type d'interaction.

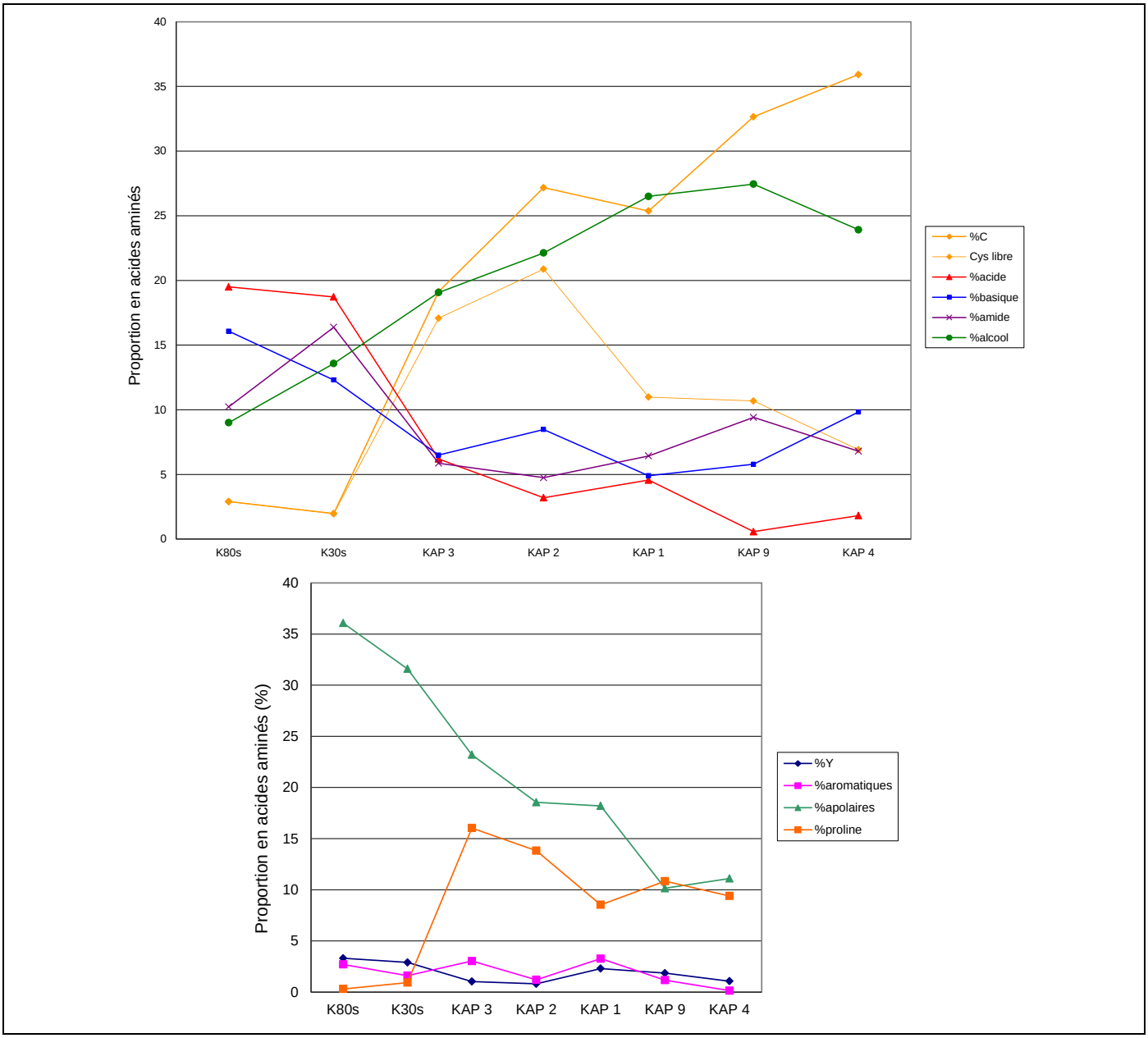


Figure 14 *: Proportion des fonctions des résidus présents dans les séquences des protéines majoritaires du cortex. L'examen des proportions de fonctions acides et basiques des kératines de type I et II montre qu'une proportion de résidus acides est en excès par rapport au nombre de résidus basiques. En supposant un appariement de ces résidus pour la formation de liaisons ioniques permettant les associations entre hétérodimères et tétramères, des fonctions acides doivent rester non appariés au sein des microfibrilles. La présence de certaines de ces fonctions libres à la surface des microfibrilles peut être envisagée. En ce qui concerne les KAP, les familles 2, 9 et 4 présentent un excès de résidus basiques (essentiellement des arginines et quelques histidines) tandis que les KAP 1 et 3 possèdent la même proportion de résidus acides et basiques.*

En revanche, les kératines des filaments intermédiaires possèdent une forte proportion de ces acides aminés (respectivement 19 et 20% pour les kératines de type I et de type II). L'excès en acides aminés acides par rapport

aux acides aminés basiques dans les séquences et particulièrement pour les kératines de type I indique qu'un certain nombre de ces fonctions pourraient être laissées libres et donc disponibles à la surface des microfibrilles après association des hétérodimères puis des tétramères. Dans ces conditions, l'interaction ionique entre les fonctions carboxylates libres des microfibrilles et les ammoniums des arginines comme ceux des KAP 4 et 9 peut être envisagée pour former systématiquement les boucles sans interférence du résidu en position 2.

4. Projection des résultats de modélisation à des propositions de structures tertiaires des KAP

a) Proposition d'une structure latérale de l'espace interfilamentaire

Suite aux résultats des expériences de modélisation, nous pouvons faire l'hypothèse de l'existence initiale d'interactions ioniques en solution dans la zone d'élongation. Ces interactions peuvent être réalisées entre les motifs pentapeptidiques des KAP et les fonctions à la surface des microfibrilles. Lors de l'assèchement du milieu, ces interactions se retrouvent sous formes de ponts salins et des liaisons hydrogènes peuvent s'établir au sein des motifs penta peptidiques, rapprochant alors les cystéines qui, pendant l'oxydation, peuvent former un pont disulfure. Après oxydation, l'enchainement de ces motifs comme c'est le cas dans les séquences des KAP 4, constituerait alors un maillage le long des microfibrilles. La structure ainsi constituée peut être assimilée à des maillons de chaines dont un côté est décoré de résidus polaires (arginine, glutamine voire acide glutamique pour les KAP 1).

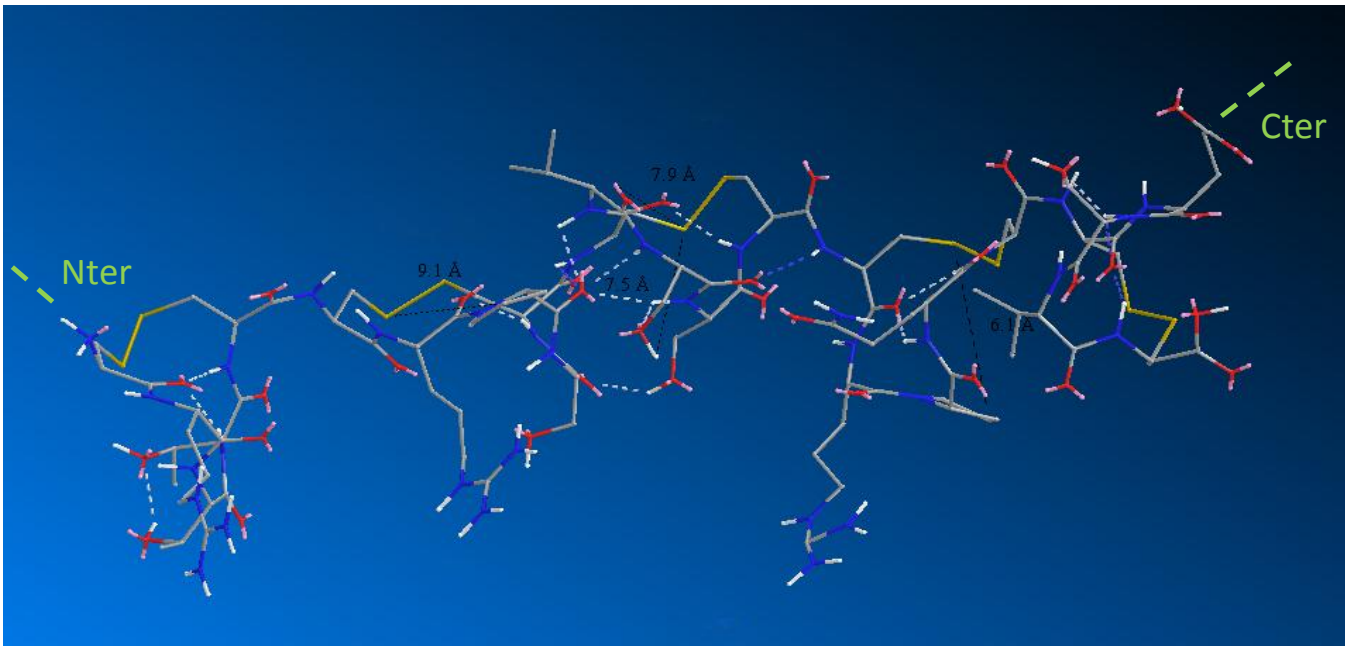


Figure 15 *: Modélisation d'un segment de 25 acides aminés extrait de la séquence de la KAP 4.9 et contenant 5 motifs pentapeptidiques. Les distances mesurées correspondent à des estimations de la largeur des boucles entre un atome de soufre de la cystéine 1 ou 5 et un atome de la chaine latérale de l'acide aminé 3.*

Au sein de chaque maillon, un réseau de liaisons hydrogènes est constitué et il est également possible d'envisager des liaisons hydrogènes avec les maillons adjacents. Le côté en opposition des résidus polaires est constitué par une tige hydrophobe où se trouvent les ponts disulfures. En construisant un modèle moléculaire constitué de l'enchainement de 5 motifs pentapeptidiques, il est possible d'estimer les dimensions de cette chaine. La largeur d'une boucle pentapeptidique mesurée sur le modèle constitué est d'environ 8 à 11 Å. En supposant que chaque segment polaire des motifs est dirigé vers la surface des microfibrilles et que le segment hydrophobe de ponts disulfures est repoussé vers l'extérieur, nous pouvons considérer qu'au minimum deux épaisseurs de ces boucles peuvent se retrouver insérées entre deux microfibrilles.

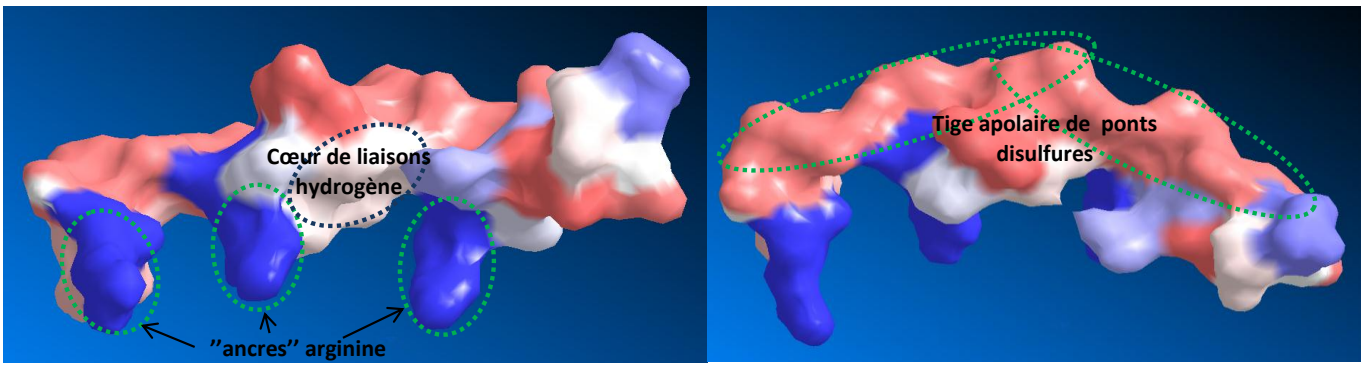


Figure 16 *: Représentation des surfaces de la structure du segment 4.9 modélisé. Les centres respectivement d'hydrophiles à hydrophobes sont représentés du bleu vers le rouge en passant par le blanc.*

Les données de diffraction X ont précédemment établi que la largeur d'une micro fibrille était de 7,5 nm et que l'espace inter micro fibrillaire chez l'homme était de 2 nm [141]. L'espace entre deux microfibrilles correspondrait environ à l'épaisseur de deux couches de tige constituée de boucles pentapeptidiques soit une épaisseur par microfibrille.

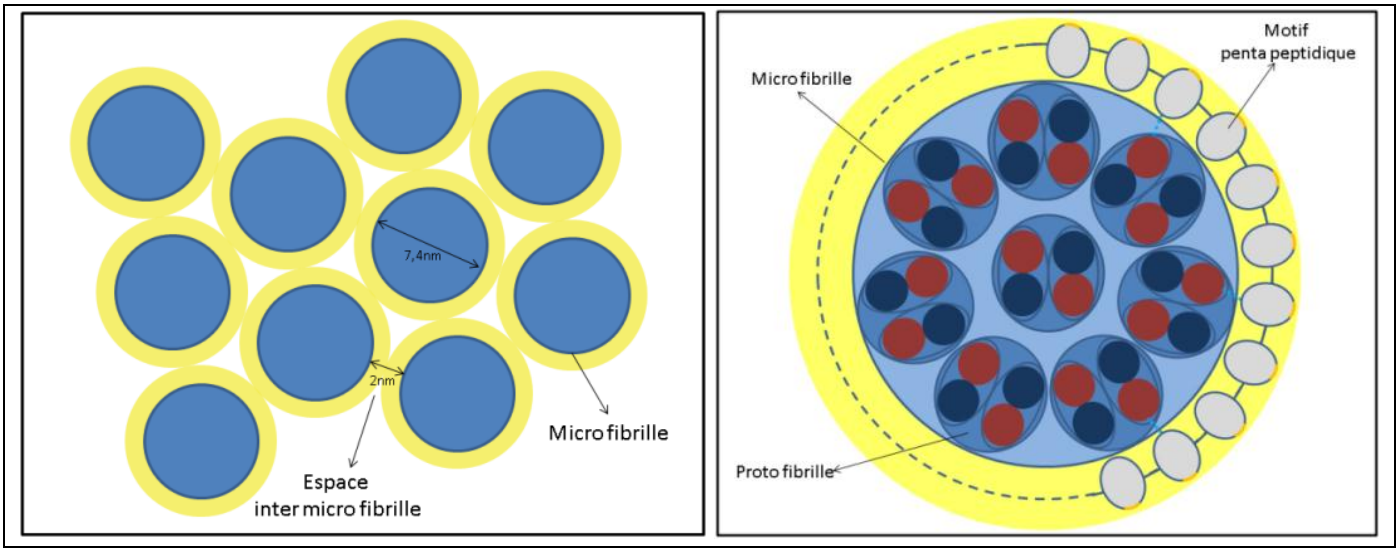


Figure 17 *: A gauche, modèle d'empilement des microfibrilles oxydées dans un système hexagonal compact formant les macrofibrilles. Les distances correspondent aux paramètres d'empilement établis par diffraction X. A droite, modèle proposé d'arrangement des segments protéiques riches en motifs pentapeptidiques oxydés sur la surface des microfibrilles.*

L'hypothèse de cet arrangement des KAP riches en motifs autour des microfibrilles permettrait donc de proposer un mode d'empilement des microfibrilles cohérent avec les observations expérimentales. Les microfibrilles présentent un excès de résidus acides au niveau des parties hélicoïdales des kératines assemblées. Sans l'insertion de protéines basiques, cet excès pourrait se traduire par une répulsion électrostatique des charges entre les microfibrilles laissant la possibilité d'insertion de protéines dans les espaces inter micro fibrillaires. Des KAP basiques exprimées après l'élongation, comme les KAP 4, 9 et 2 ont probablement suffisamment d'affinité pour assurer ce rôle et venir interagir avec les résidus acides potentiellement présents en surface des microfibrilles.

Parmi ces protéines insérées, une majorité peut former des boucles pentapeptidiques suite à la déshydratation et l'oxydation de la cellule ce qui contraint les fibrilles à être séparées d'un espace minimum correspondant à deux largeurs de ces boucles.

Cette architecture expliquerait l'empilement régulier des microfibrilles dans certaines zones du cortex. Elle remettrait en question la vision d'une matrice sans organisation, celle-ci étant alors orientée autour des filaments. La question de l'organisation de cette couche par rapport aux microfibrilles peut être posée. Dans le

cas des KAP 4 et 9, l'enchainement plutôt aléatoire des motifs contenant des arginines laisse envisager une disposition également aléatoire de ces segments sur les résidus acides libres des microfibrilles.
Une remarque peut à ce stade être soulevée : il est établi que les microfibrilles passent d'une organisation en cylindres creux à une organisation cylindrique compacte pendant l'oxydation [135]. Une contrainte doit être exercée pendant l'oxydation sur les microfibrilles pour permettre cette compaction. Nous proposons que la formation des boucles pentapeptidiques puisse se traduire par une augmentation de l'espace occupé par la matrice interfilamentaire. La structuration des motifs et leur répulsion stérique les uns par rapport aux autres pourraient ainsi permettre d'exercer une pression latérale suffisante pour assurer la compression progressive des filaments.

b) Proposition de mécanismes d'association des protéines dans l'espace intermicrofibrillaire

Des interactions hydrophobes et électrostatiques entre microfibrilles et KAP en solution

Nous postulons que la construction de la structure avant déshydratation et oxydation ne doit reposer que sur une combinaison d'effets hydrophobes et d'attractions/répulsions électrostatiques. Ces interactions permettent d'expliquer la possibilité d'insertion des KAP entre les microfibrilles elles mêmes repoussées les unes des autres par les répulsions électrostatiques des résidus acides en excès à leur surface. L'insertion des KAP dans l'espace intermicrofibrillaire puis la formation des boucles pentapeptidiques permettent d'expliquer les distances intermicrofibrillaires mesurées par diffraction X.
L'existence de différentes familles d'isoformes doit dans ce cadre être considérée. Nous avons précédemment étudié la structure primaire des différentes familles de KAP déterminées comme majoritaires dans le cortex, ce qui nous a permis de découper certaines séquences en segments dont la physico-chimie des résidus peut déterminer si ces segments sont hydrophobes et basiques ou acides.

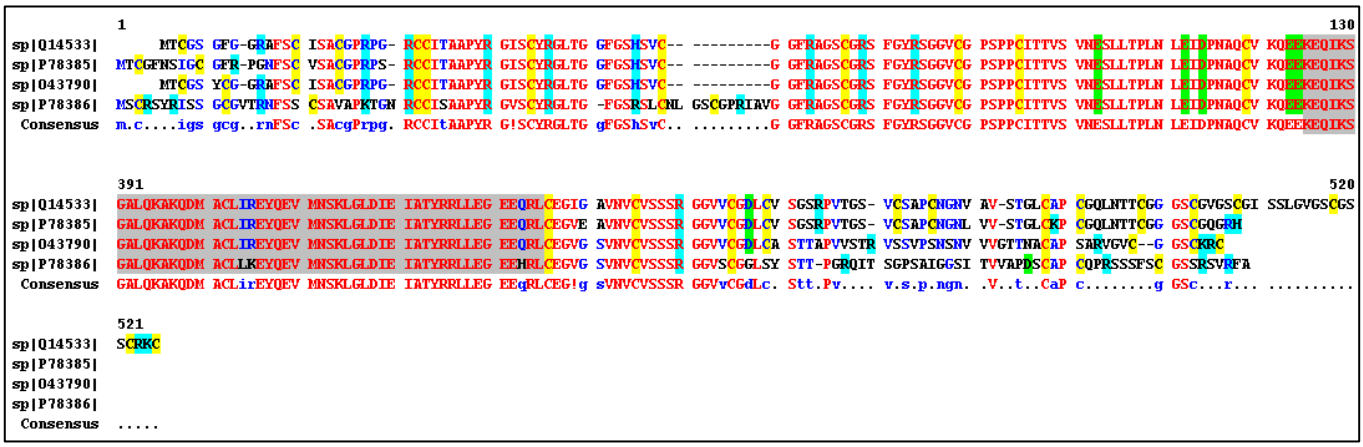


Figure 18 *: Alignement des segments N-terminaux et C-terminaux des kératines de type II majoritaires du cortex (respectivement K81, K83, K86 et K85). Nous montrons l'abondance de résidus basiques et de cystéines parmi des résidus apolaires au sein de ces séquences. En gris, les segments correspondants aux tiges en hélice α..*

Concernant les microfibrilles, nous pouvons supposer qu'une partie des domaines têtes et queues des kératines sont présents en surface des microfibrilles comme il l'est suggéré dans le modèle proposé par Parry et *al*. . L'examen des segments têtes et queues des kératines dures montre une composition très différente des segments tiges (Figure 19). Ces domaines contiennent principalement des résidus chargés exclusivement basiques (arginine et histidine), des résidus hydrophobes et également environ une cystéine pour 10 résidus. Dans ces conditions, des interactions hydrophobes et électrostatiques avec d'autres séquences protéiques portant des résidus acides et hydrophobes peuvent être envisagées. Les segments acides des KAP 1 peuvent être des candidats pour ces types d'interactions qui permettraient de les considérer en solution plutôt proches des domaines têtes et queues des kératines des microfibrilles.

Concernant les KAP 4, nous pouvons noter une polarisation entre les domaines N-terminaux et C-terminaux. En effet, le segment N-terminal contient les seuls résidus acides de la séquence protéique et cela pour l'ensemble des mammifères (généralement deux à quatre acides aspartique et glutamique). Le segment C-terminal peut être considéré comme relativement hydrophobe et contenant des résidus basiques. Nous pouvons alors envisager qu'il puisse exister une interaction du domaine N-terminal des KAP 4 avec les domaines N-terminaux et C-terminaux des kératines des filaments et/ou une interaction avec le domaine C-terminal d'une autre KAP 4.

Figure 19 *: Alignement des segments N-terminaux et C-terminaux des kératines de type I majoritaires du cortex (respectivement K31, K33a, K33b et K34). Ces segments sont plus courts que ceux des kératines de type II mais contiennent également des résidus basiques et des cystéines parmi des résidus apolaires.*

Les structures des autres KAP 9, 2 et 3 n'ont pas été examinées de la même manière. Pour les KAP 9, nous pouvons considérer l'abondance des résidus basiques comme voisin de ce qui a pu être décrit pour les KAP 4. Les différences de structures N-terminales et C-terminales ainsi que des insertions de segments non pentapeptidiques dans le domaine central constituent des spécificités qui les différencient cependant des KAP 4. Pour les familles des KAP 2 et 3, respectivement basiques et acides et de faible abondance en motifs pentapeptidiques, il semble nécessaire de considérer différemment leur arrangement structural et nous n'apporterons pas plus de discussions sur de potentielles fonctions de ces séquences.

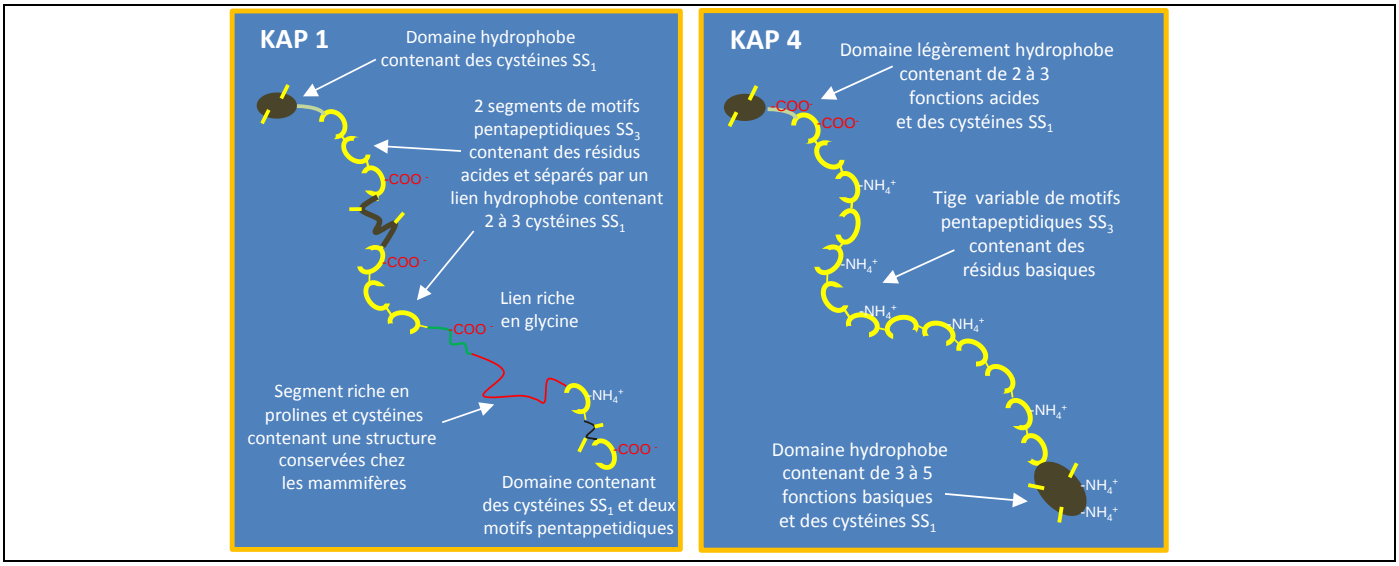


Figure 20 *: Proposition de structure secondaire des KAP 1 et 4 établie sur la base des analyses des séquences primaires de ces protéines.*

Hypothèses d'organisation des ponts disulfures dans l'espace interfilamentaire

La considération de ces arrangements structuraux en solution entre les microfibrilles et les KAP pourrait permettre de mieux appréhender le réseau de ponts disulfures formés par la suite pendant l'oxydation du système.

La formation des boucles pentapeptidiques permet d'ores et déjà d'expliquer la forte proportion de boucles dites SS_3 d'après la nomenclature de Naito (47% des ponts) [146]. Les cystéines restantes sont ainsi des candidates pour réaliser des ponts inter protéines de type SS_1 impliqués dans 35% des liaisons disulfures du cortex. La concentration relativement importante de cystéines dans les segments têtes et queues des kératines fait de ces domaines des donneurs potentiels de résidus cystéine pouvant ponter avec des KAP. Ces cystéines sont proches dans les séquences protéiques de résidus basiques : elles auraient ainsi une plus forte probabilité de ponter avec des cystéines libres d'autres protéines proches dans leur séquence de résidus acides comme ceux présents par exemple sur certains segments des KAP 1. Les cystines ainsi formées pourraient permettre de lier covalemment par l'intermédiaire d'une KAP 1 plusieurs domaines N-terminaux et C-terminaux des kératines et ainsi de solidariser longitudinalement et latéralement les protéines d'une microfibrille en un unique édifice. La possibilité de lier également deux microfibrilles l'une à l'autre par cette intermédiaire n'est pas à écarter. Les liens entre la matrice et les filaments peuvent donc être envisagés comme périodiques et au niveau des domaines terminaux présents à la surface des microfibrilles.

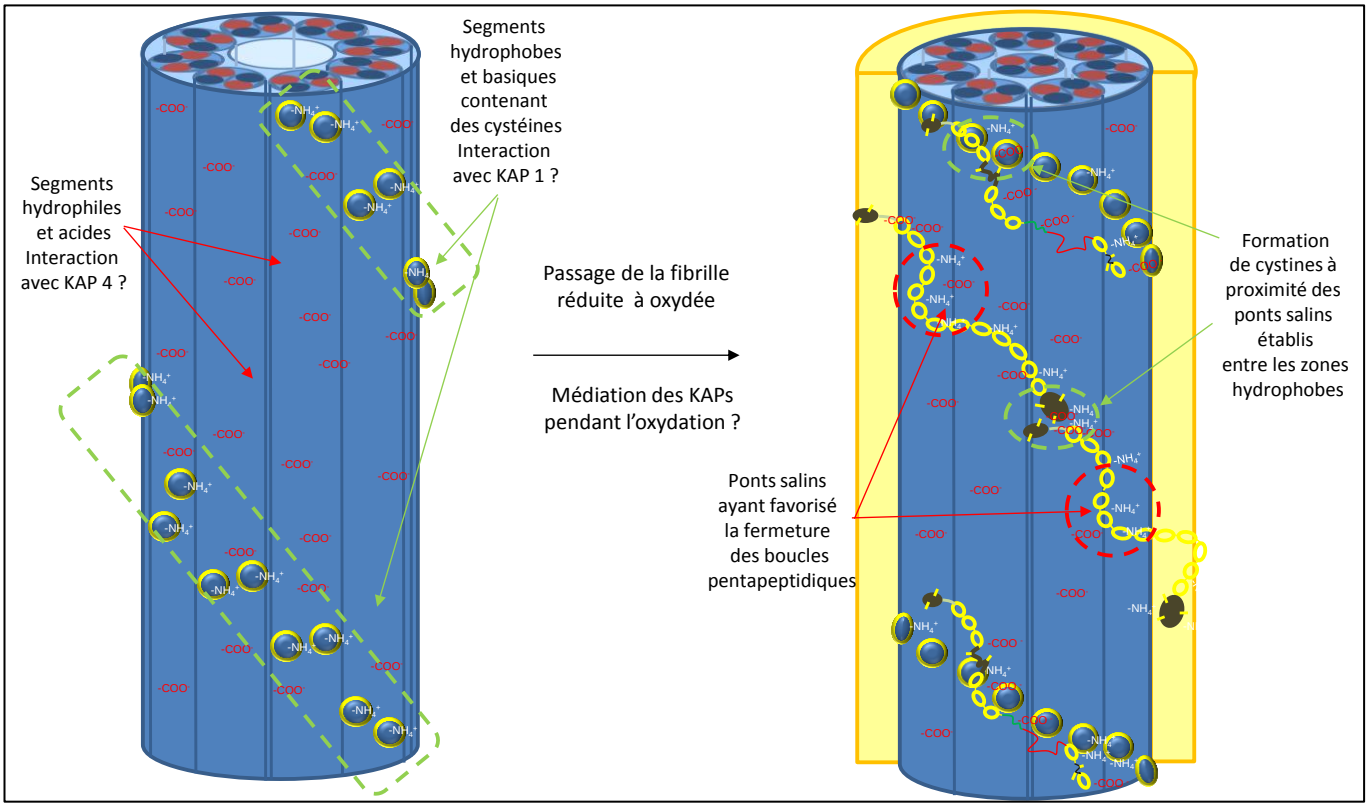


Figure 21 *: A gauche, modèle présentant les zones fonctionnelles pouvant être envisagées à la surface des microfibrilles avant l'expression des KAP dans la cellule. A droite, représentation du modèle de Parry de la microfibrille oxydée sur laquelle pourrait se répartir les KAP en fonction des interactions hydrophobes et électrostatiques établies initialement en solution.*

Concernant la structure des KAP 4, nous pouvons imaginer des liens cystines entre leurs domaines N-terminaux et C-terminaux initiés par des interactions hydrophobes et électrostatiques favorables. Une succession de plusieurs protéines liées les unes aux autres par ces enchainements têtes et queues peut être envisagée. Nous pouvons également envisager qu'il puisse exister des nœuds de réticulation où sont liées plusieurs extrémités de protéines

simultanément (par exemple deux domaines N-terminaux sur un domaine C-terminal). Une structure analogue avec une matière élastomérique serait alors obtenue.

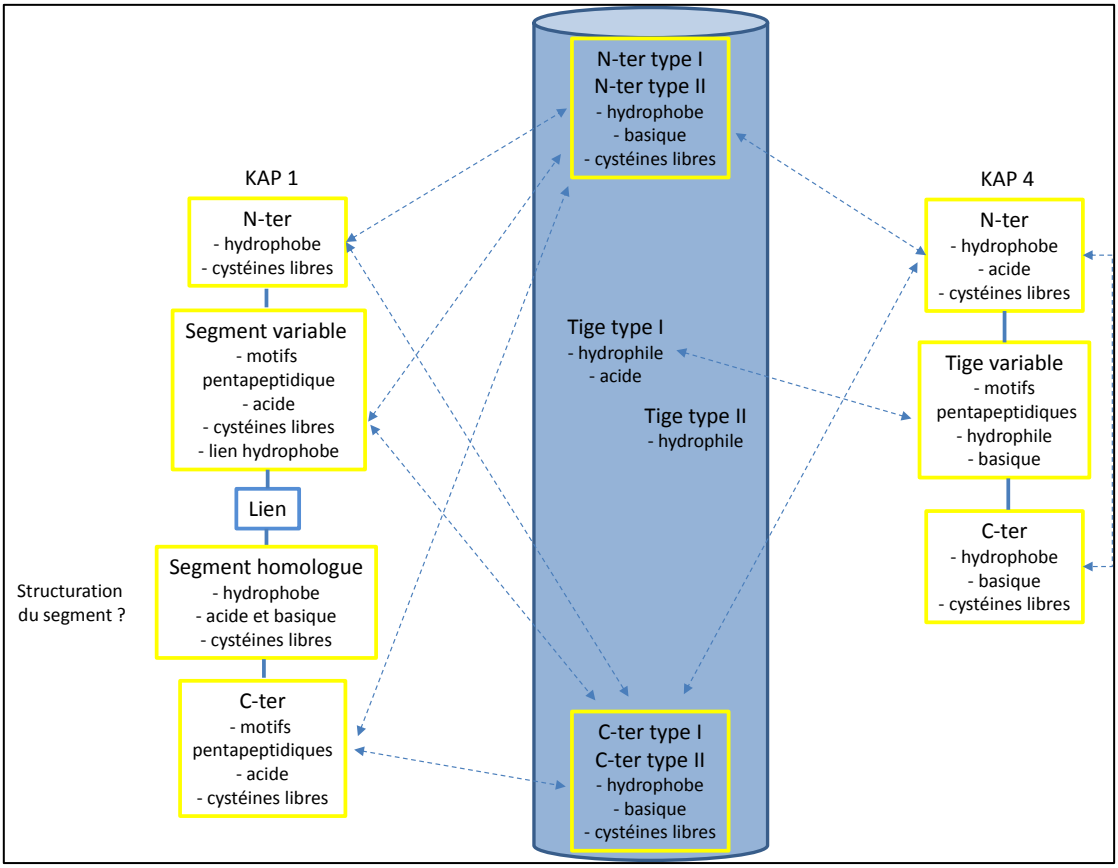


***Figure 22** : Proposition d'un réseau d'interactions entre les domaines des kératines en surface des microfibrilles et les domaines des KAP 1 et 4. La possibilité de ces interactions hydrophobe et électrostatique peut indiquer l'organisation préférentielle des structures en solution. Cette organisation doit être conservée et renforcée lorsque des cystéines libres (non impliquées dans des motifs pentapeptidiques) sont présentes respectivement sur deux segments en interaction.*

Ces hypothèses peuvent être confrontées aux différents modèles structuraux proposés dans la littérature pour expliquer les propriétés mécaniques de la fibre [149, 286]. La principale zone d'ombre suscitant le débat entre ces modèles reste l'arrangement de la matrice interfilamentaire.

Le modèle de Feughelman propose un arrangement des KAP en amas globulaires hydrophobes entourés d'une surface hydrophile. L'ensemble est décrit comme entouré d'eau isolant la matrice des structures constituées par les filaments de kératines. En comparaison avec notre modèle, la notion de globule peut sembler erronée. En revanche, le concept de surface hydrophile pourrait être cohérent avec le système de ponts salins proposé à l'interface entre les segments pentapeptidiques de KAP et les tiges acides de microfibrilles.

Le modèle proposé par Chapman et Hearle décrit la matrice comme un élastomère densément ponté de liaisons inter et intra moléculaires. Il propose également des liens périodiques entre les microfibrilles, au niveau des domaines terminaux, et les protéines de la matrice. Ce modèle est vraisemblablement en accord avec le modèle que nous proposons mais ne prend pas en compte les interactions de types ponts salins pouvant exister entre la matrice et les segments tiges des filaments de kératine.

Le modèle de Crewther décrit la matrice comme une chaine de perles liées les unes aux autres par un réseau peu dense de ponts disulfures et liés uniformément à la surface des microfibrilles. Ce concept de perles parait comparable aux boucles pentapeptidiques des KAP 4 qui sont également légèrement pontées les unes aux autres au niveau des segments où les motifs sont absents. En revanche, les liaisons entre ces éléments et les microfibrilles devraient être considérées plutôt périodiques qu'uniformes.

L'ensemble de ces modèles ne considèrent pas l'hétérogénéité apportée par la présence dans la matrice de plusieurs familles de protéines. Notre modèle permet de trouver des points d'accord avec les différents modèles précédemment proposés et pourrait affiner la compréhension de l'organisation de la matrice.

5. Recherche de l'origine des KAP chez les mammifères

Les oiseaux et les reptiles sont les plus proches parents des mammifères mais ne possèdent dans leur génome aucun signe de séquences cousines des KAP riches en soufre pouvant contenir les motifs pentapeptidiques précédemment décrits. En revanche, des kératines orthologues aux kératines dures des mammifères peuvent être identifiées. Le locus contenant les gènes des kératines de type I dans ces organismes est dépourvu de tout segment pouvant correspondre à une séquence orthologue à celles des mammifères [287-297].
Les protéines exprimées dans les fibres des plumes sont très différentes, les microstructures fibrillaires pouvant y être observées sont constituées par une protéine, la kératine béta, structurée en une chaine de coudes béta en interaction avec des domaines N-terminaux et C-terminaux. La matrice interfilamentaire est, dans ce système, constituée par les propres segments N-terminaux et C-terminaux de la protéine. Ces très grandes différences nous ont amené à nous interroger sur l'origine des KAP chez les mammifères. Nous avons souhaité étudier l'histoire de ces gènes en évaluant les liens d'homologie pouvant exister entre eux afin de tenter d'établir leur arbre phylogénique.

a) Phylogénique des KAP

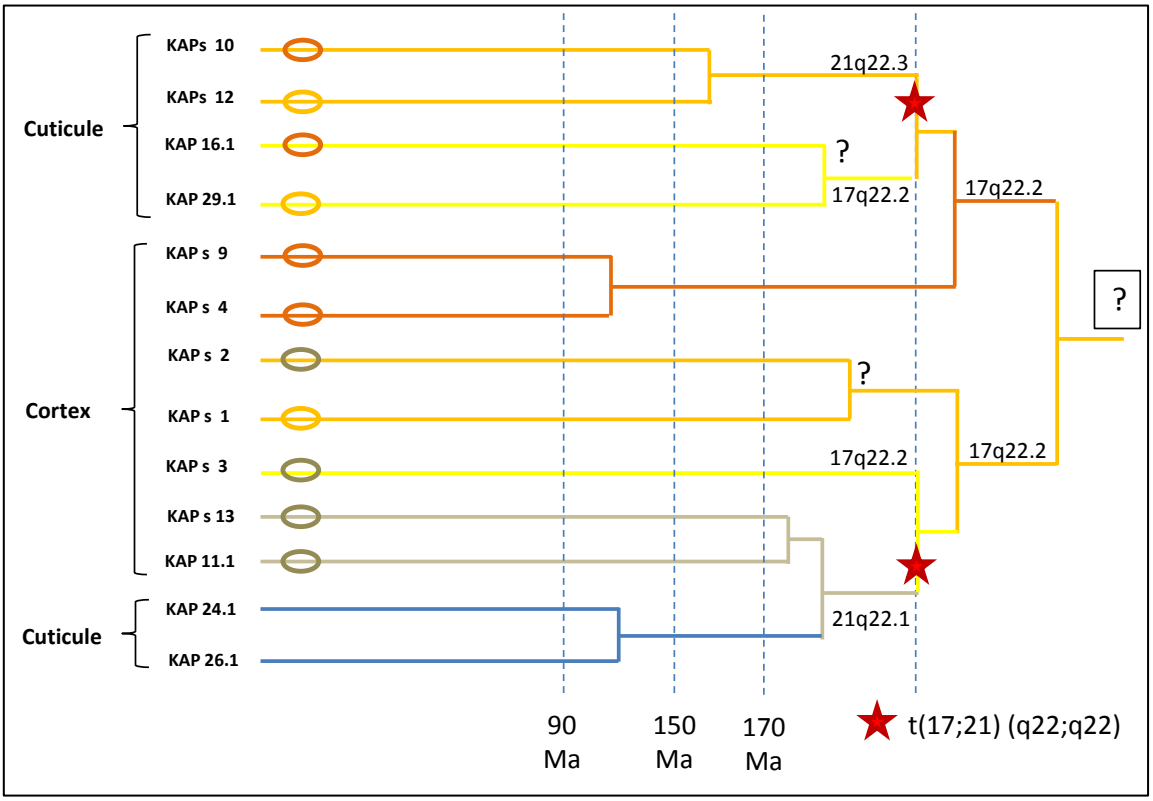


***Figure 23** : Proposition d'un arbre phylogénétique reliant la majeure partie des gènes actifs des familles de KAP contenant des cystéines présentes sur les locus chromosomiques 17q22.2, 21q22.1 et 21q22.3. La présence de boucles pentapeptidiques dans les séquences des familles est symbolisée par les boucles colorées (boucles résiduelles, en brun ; présence importante de boucles, en orange).*

En nous appuyant sur l'hypothèse d'un événement de translocation d'un fragment du chromosome 17 contenant le locus 17q22 vers le chromosome 21, nous pouvons expliquer des liens de parenté entre les gènes des différents locus. Cette hypothèse est appuyée par l'homologie existant entre les KAP 3, 11, 13, 24.1 et 26.1 (Figure 24) mais

également sur l'homologie existant entre les KAP 10, 16.1 et 29.1. Cet événement de translocation semblerait très antérieur à la divergence Monotrèmes/Thériens (environ 170 Ma) compte tenu des différences observées dans les séquences de ces gènes.
Il semble difficile d'établir si les événements de duplication antérieurs à cette divergence ont pu avoir lieu avant ou après l'événement de translocation. Il est également difficile sur la base des comparaisons de séquence d'établir l'ordre des duplications ayant données naissance aux différentes branches de familles de KAP comme les KAP 2, 1 et 4. Cependant pour ces dernières, l'ensemble des séquences présente des homologies qui peuvent nous conforter dans le fait qu'il ait existé un gène ancestral commun à l'ensemble des gènes de KAP riches en soufre existant aujourd'hui.

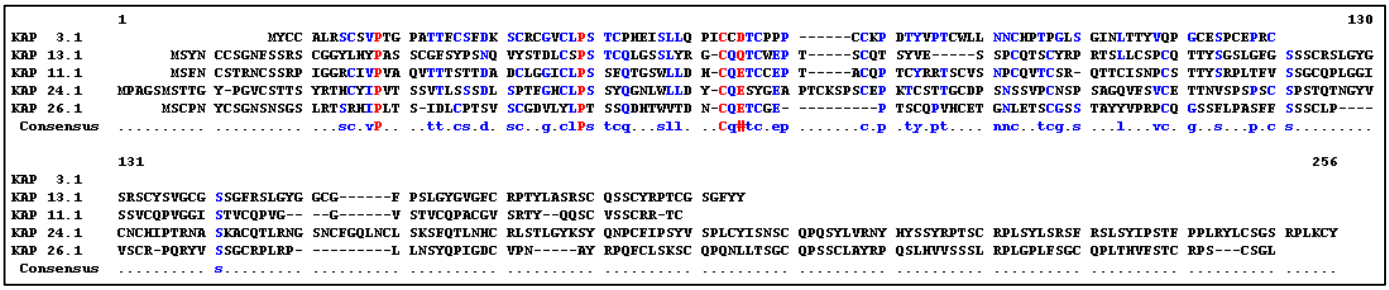


Figure 24 *: Alignement suggérant l'homologie existant en début de séquence entre les HS KAP du locus 21q22.1 et la famille des KAP 3 du locus 17q22.2 .*

Nous noterons que les motifs pentapeptidiques semblent conservés dans certaines branches de l'arbre (KAP 4, 9 et 10) et que d'autres branches ont perdu progressivement cette fonction (branche contenant les KAP 3, 11.1, 13, 24.1 et 26.1).

b) Recherche d'un gène ancestral commun à tous les gènes des KAP riches en soufre

L'arbre phylogénique des KAP suggère que l'ensemble des familles de KAP soufré des locus 17q21.2, 21q22.1 et 21q22.3 proviennent d'un ou deux gènes ancestraux initialement localisés sur le locus 17q22.2 et inséré entre des segments contenant les gènes des kératines du cheveu de type I. Dans les génomes des oiseaux et des reptiles, des gènes pouvant être parents de ces séquences sont absents dans le locus correspondant aux kératines de type I mais également dans l'ensemble du caryotype de ces espèces [285]. Cette absence nous a conduit à nous interroger sur l'origine de ces gènes ancestraux ayant par la suite été dupliqués et translocalisés vers un ou plusieurs chromosomes (en considérant une potentielle homologie des KAP 5, 28 avec la séquence 17 que nous n'avons pas étudiée en détail dans ce manuscrit).
Il est possible d'envisager une perte totale chez les sauropodes de l'expression de ce gène qui aurait pu exister chez l'ancêtre amniote et être conservé seulement par les théropodes. Néanmoins, des traces de ce gène perdu devraient pouvoir être retrouvées dans les génomes des descendants des sauropodes.
L'introduction d'un gène xénologue, c'est-à-dire d'un organisme à un autre, au sein du locus 17q21.2 dans le génome de l'ancêtre commun à tous les mammifères de la branche des théropodes pourrait être également une explication plausible permettant d'expliquer l'absence de ces gènes chez les descendants des sauropodes. Nous avons choisi d'explorer, par la suite, cette seconde hypothèse.

c) Recherche d'une origine des gènes des KAP hors du génome des mammifères

Les mécanismes de transmission d'un gène xénologue au sein d'un organisme multi cellulaire comme l'ancêtre des mammifères impliquerait qu'il ait été contaminé par un vecteur portant ce gène. Le vecteur doit alors infecter et intégrer son génome aux gamètes de l'individu contaminé. Dans cette hypothèse, des fragments de gènes du vecteur auraient été incorporés entre des gènes des kératines.

Parmi ces vecteurs, nous pouvons envisager des virus comme les rétrovirus capables d'intégrer leur génome dans une cellule hôte. 8% du génome humain est constitué de séquences rétrovirus endogènes, vestiges d'infections. La majorité de ces séquences ne sont pas codantes mais certaines pourraient être à l'origine de l'émergence de nouvelles espèces : des rétrovirus endogènes sont par exemple suspectés d'avoir contribué à l'apparition du placenta chez les mammifères placentaires [298].
Des études récentes suggèrent également la possibilité de mécanismes de transfert de gène horizontal impliquant des agents de transfert de gènes observés chez les bactéries océaniques *alpha-Proteobacteria* [299].
Nous avons choisi d'évaluer s'il existait, dans les génomes viraux ou bactériens déjà séquencés et disponibles à la communauté scientifique, des taxonomies possédant des protéines présentant des homologies de séquence avec celles des KAP riches en soufre des mammifères. Compte tenu de l'arbre phylogénique des KAP établi précédemment, nous avons considéré comme probable que le gène ancestral commun aux KAP riches en soufre possédait de nombreux motifs pentapeptidiques.
Les premiers essais de recherche dans la banque protéique NCBI nr rassemblant les protéines issues de la traduction des génomes viraux n'ont donné aucun résultat d'homologie avec les segments répétitifs des KAP 4.

Figure 25 *: En haut, alignement de la séquence trouvée chez Planctomyces brasiliensis comme homologue aux KAP riches en motifs pentapeptidiques. Cette séquence est alignée avec une KAP 4 humaine, une KAP 4 de Monodelphis domestica et une KAP 10 d'Ornithorhynchus anatinus. En bas, nous montrons que des motifs pentapeptidiques peuvent être retrouvés dans chacune de ces séquences.*

La tentative de recherche d'homologie parmi les protéines bactériennes a en revanche apporté un résultat présentant un bon score d'homologie. La séquence obtenue, provenant de la taxonomie *Planctomyces brasiliensis*, du groupe des bactéries aquatiques planctomycètes possède un segment complet au sein duquel l'enchainement des cystéines, des prolines et des sérines est analogue à celui retrouvé dans les séquences des KAP 4. La recherche des motifs pentapeptidiques dans la séquence montre une construction voisine de ce que nous avons pu décrire précédemment chez les mammifères. Dans la bactérie, les motifs penta peptidiques retrouvés sont de type (CAPSC) ce qui d'après nos observations de modélisation, pourrait constituer un motif très favorable à la formation du pont. Il est en effet possible d'envisager que l'alanine ne réalise pas d'interférence pour le rapprochement et que la sérine favorise la formation de liaisons hydrogènes rapprochant les deux

cystéines. Nous notons également la présence de quelques motifs (CAPEC) non retrouvés chez les séquences des mammifères.

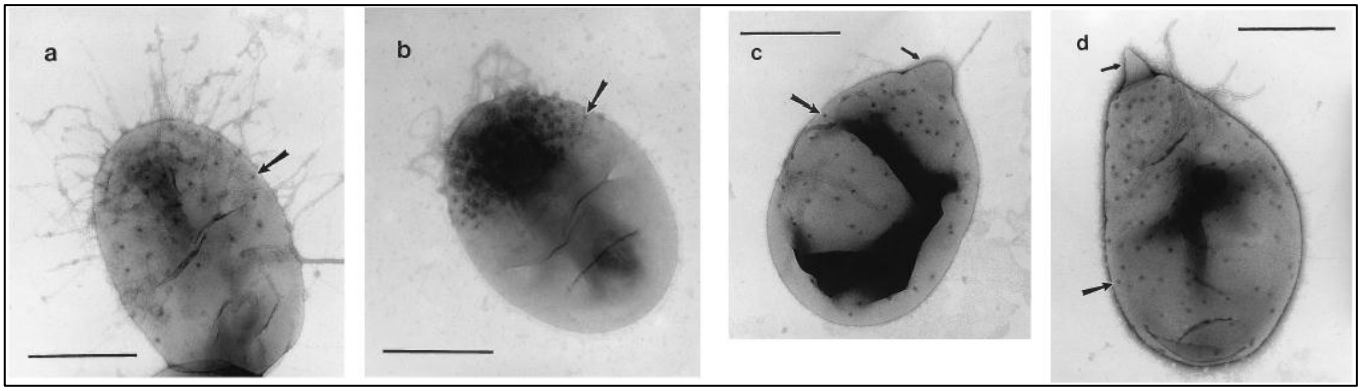


***Figure 26** : Observation en microscopie électronique de différentes espèces de Pirellulae isolées de larves de crustacé. a) AGA/M12 ; b) HGG/MP1 ; c) AGA/M18 ; d) Planctomyces brasiliensis ATCC 49424. Les deux dernières bactéries c) et d), décrites comme proches du point de vue phylogénique, possèdent un appendice visible sur la projection longitudinale. L'appendice observé pour Planctomyces brasiliensis a une forme de corne [300].*

Planctomyces brasiliensis est décrite dans une étude de Fuerst et *al*. ce qui nous a permis de comparer cette bactérie avec d'autres membres apparentés au même groupe des *Pirellulae*. La comparaison de ces quelques membres montre que *Planctomyces brasiliensis* a la particularité de posséder une corne à la différence des autres *Pirellulae* (Figure 26). Cette différence est troublante compte tenu du cheminement qui nous a conduit jusqu'à cette taxonomie. Elle nous amène à nous interroger pour savoir si la protéine identifiée comme homologue aux KAP des mammifères n'est pas en cause dans la structure particulière de cet appendice qui ressemble très fortement à une structure cornifiée.

La présence d'une telle séquence protéique dans un organisme si différent des mammifères pourrait permettre d'envisager qu'il a existé et qu'il existe encore parmi les microorganismes des espèces possédant des gènes homologues à certaines KAPs ailleurs que chez les mammifères. Cette observation rend probable un événement de contamination par transfert de gène horizontal il y a plus de 200 millions d'années du génome de l'ancêtre commun des mammifères actuels. Le mécanisme et le vecteur ayant permis ce transfert resterait alors à déterminer.

Un mécanisme d'évolution convergente conduisant à des séquences protéiques similaires entre les mammifères et cette bactérie pourrait également être envisagé. Néanmoins, il n'expliquerait pas l'absence de séquences ou de rémanents de séquences au sein du locus des kératines de type I chez les descendants des sauropodes. Afin d'écarter définitivement cette hypothèse, il parait nécessaire d'explorer plus en détail le génome non codant des oiseaux, tortues et reptiles à la recherche de rémanents de séquences de KAP.

Il pourrait également être intéressant d'étudier particulièrement *Planctomyces brasiliensis* afin d'estimer, sur la base de comparaisons des gènes de l'ARN ribosomique 16S, à quelle période cette espèce de bactérie aurait pu acquérir ce trait spécifique permettant de la distinguer des autres *Pirellulae*. Cette datation serait nécessaire pour déterminer si cette espèce a une probabilité d'avoir côtoyé sous cette forme les ancêtres des mammifères à une période où aurait pu avoir lieu une potentielle transmission horizontale du gène. Dans le cas contraire, elle permettrait d'envisager un autre vecteur de transmission du gène qui aurait pu communément contaminer l'ancêtre des mammifères et l'ancêtre de *Planctomyces brasiliensis*.

d) Remarque concernant la régulation des KAP dans le génome humain

Nous venons de voir que l'origine des KAP chez les mammifères pourrait être expliquée par l'insertion de gènes correspondants au sein du locus correspondant aux kératines dures de type I. Dans ce cas, nous pouvons imaginer que la position de cette insertion a joué un rôle vis-à-vis de leur expression simultanée avec les kératines dures.

La position des gènes précédemment étudiés au sein des locus devrait être examinée au regard de la connaissance de leur expression dans les cellules corticales, cuticulaires et unguéales. Cet examen pourrait permettre d'établir d'éventuels mécanismes de régulation des gènes conduisant aux différentes différenciations cellulaires. Nous pouvons noter que l'événement proposé de translocation de gènes de KAP riches en soufre vers le locus 21q22.1 pourrait avoir conduit à une translocation de séquences promotrices ayant permis l'activation de gènes de KAP riches en glycine et tyrosine situés initialement dans ce locus avant la translocation et exprimés dans le cortex. Il pourrait ainsi être intéressant par la suite d'étudier si des promoteurs similaires peuvent être retrouvés dans les différents locus où se situent les gènes des KAP pour mieux comprendre les mécanismes de régulation de ces gènes.

Conclusion générale

Nous avons présenté au cours de ce manuscrit comment la protéomique a permis d'apporter de nouvelles connaissances en démontrant l'expression protéique des composants principaux des structures du cheveu. Cette démonstration permet de compléter le catalogue des gènes exprimés chez l'être humain et de mieux comprendre les protéines impliquées dans les structures cellulaires composant la fibre capillaire comme le cortex et la cuticule. L'obtention de ces informations fait suite à une profonde réflexion quant à la stratégie à mettre en place pour répondre à la problématique de caractérisation de protéines isoformes ciblées et tout particulièrement des familles de protéines associées aux kératines.

L'instrumentation nano-LC-MS/MS a nécessité d'être finement optimisée pour maximiser les probabilités d'identification des peptides. Nous avons montré comment la compréhension de ces systèmes et l'identification des paramètres les plus influents sur l'acquisition des données pouvaient être mises à profit pour le séquençage des peptides en protéomique. Cette compréhension découle, d'une part de l'emploi d'une stratégie originale basée sur l'utilisation d'un plan d'expérience, et d'autre part d'une évaluation rigoureuse de l'impact sur les résultats d'identification des paramètres chromatographiques et des nouvelles fonctionnalités apportées par la dernière génération de spectromètres de masse.

L'obtention de nouvelles données apportées par nos études protéomiques nous a conduit à les exploiter dans un contexte de recherche d'informations quantitatives et structurales pour les familles de protéines majoritairement identifiées. L'étude des structures primaires de certaines KAP, associée à des expériences de modélisation moléculaire, nous a permis de suggérer un modèle définissant le rôle de ces protéines dans la structure de la fibre capillaire. Ce modèle s'établit sur la base de considérations de mécanismes d'interactions entre les KAP et les filaments de kératine pendant et après la phase de kératinisation.

L'étude des séquences des KAP nous permet enfin de suggérer que les gènes correspondant, de toute évidence fondamentaux à la structure corticale, pourraient constituer des éléments historiques de l'évolution des mammifères.

Références bibliographiques

1. Warren, W.C., et al., *Genome analysis of the platypus reveals unique signatures of evolution.* Nature, 2008. **453**(7192): p. 175-83.
2. Ji, Q., et al., *A swimming mammaliaform from the Middle Jurassic and ecomorphological diversification of early mammals.* Science, 2006. **311**(5764): p. 1123-7.
3. Langbein, L. and J. Schweizer, *Keratins of the human hair follicle.* Int Rev Cytol, 2005. **243**: p. 1-78.
4. Vullo, R., et al., *Mammalian hairs in Early Cretaceous amber.* Naturwissenschaften, 2010. **97**(7): p. 683-7.
5. Porter, A.M., *Why do we have apocrine and sebaceous glands?* J R Soc Med, 2001. **94**(5): p. 236-7.
6. Folk, G.E., Jr. and H.A. Semken, Jr., *The evolution of sweat glands.* Int J Biometeorol, 1991. **35**(3): p. 180-6.
7. Thurman, J., *Two cases in which the skin, hair and teeth were very imperfectly developed.* Medico-chirurgical transactions, 1848. **31**: p. 71-82.
8. Smilie, E.R., *Blanched Hair from Sudden Emotions.* Boston Med Surg J, 1851. **44**: p. 438-440.
9. Smith, W.G., *On a rare Nodose Condition of the Hair.* British Medical Journal, 1880.
10. Scott, W.B.a.J.A., *Moniliform Hairs.* The British Journal of Dermatology, 1892.
11. Unna, D., *Materia Medica and Therapeutics.* British Medical Journal, 1885.
12. Edison, T.A., *Filament for incandescent lamps*, in *US Patent 534206*. 1895.
13. Todd, R. and W. Bowman, *The Physiological Anatomy and Physiology of Man.* 1850.
14. Beigel, H., *The Human hair, its structure, growth, diseases and their treatment*, ed. H. Renshaw. 1869.
15. Hodgkinson and Sorby, *On the colouring matters found in Human Hair* The Journal of the Anthropological Institute of Great Britain and Ireland, 1878. **8**.
16. Riddle, O., *Our knowledge of melanin color formation and its bearing on the Mendelian description of heredity.* The Biological Bulletin, 1909.
17. Vickery, H. and C. Leavenworth, *The basic amino acid of wool.* Journal of Biological Chemistry, 1929. **86**: p. 107-111.
18. Leavenworth, H.V.a.C., *The separation of cystine from histidine : the basic amino acids of human hair.* The Journal of Biological Chemistry, 1929. **83**: p. 523-534.
19. Beveridge, C.L.J., *The analysis of hair keratin. A method for the quantitative removal of cystine fron keratin hydrolysates.* Biochemical Journal, 1940. **34**(10-11): p. 1356-1366.
20. Horn, M., D. Jones, and S. Ringel, *Isolation of a new sulfur-containing amino acid (lanthionine) from sodium carbonate-treated wool.* Journal of Biological Chemistry, 1941. **138**: p. 141-149.
21. Rogers, G.E. and D.H. Simmonds, *Content of citrulline and other amino-acids in a protein of hair follicles.* Nature, 1958. **182**(4629): p. 186-7.
22. Harding, H.W. and G.E. Rogers, *Epsilon-(gamma-glutamyl)lysine cross-linkage in citrulline-containing protein fractions from hair.* Biochemistry, 1971. **10**(4): p. 624-30.
23. Woods, W.A.a.H., *The X-Ray Interpretation of the Structure and Elastic Properties of Hair Keratin.* Nature, 1930. **126**: p. 913-914.
24. Street, W.A.a.A., *III X-Ray Studies of the Structure of Hair, Wool and Related Fibres.* Philosophical Transactions of the Royal Society of London, 1932. **230**: p. 75.
25. Woods, W.A.a.H., *X. X-Ray Studies of the Strucutre of Hair, Wool and Related Fibres. II The Molecular Structure and Elastic Properties of Hair Keratin.* Philosophical Transactions of the Royal Society of London, 1933. **232**: p. 333.
26. Fiala, G., *Preparation of hair for cross-section examination.* American Journal of Physical Anthropology, 1930. **14**(1): p. 73-74.
27. Trotter, M., *The form, size, and color of head hair in American whites.* American Journal of Physical Anthropology, 1930. **14**(3): p. 433-445.
28. Dawson, M.T.a.H., *The Hair of French Canadians.* American Journal of Physical Anthropology, 1934. **18**(3): p. 443-456.
29. Kneberg, M., *Improved Technique for Hair Examination.* American Journal of Physical Anthropology, 1935. **20**(1): p. 51-67.
30. Trotter, M., *Anthropometry A review of the classifications of hair.* American Journal of Physical Anthropology, 1938. **24**(1): p. 105-126.
31. Steggerda, M., *Cross Sections of Human Hair from Four Racial Groups.* Journal of Heredity, 1940: p. 474.
32. Kirk, P., *Human Hair Studies. 1. General considerations of hair individualization and its forensic importance.* JOurnal of Criminal Law and Criminology, 1940. **31**(4): p. 486-496.
33. Martin, F., *The microscopy of mammalian hair for anthropologists.* Proceedings of The American Philophical Society, 1942. **85**(3): p. 250-274.
34. Pauling, L., R.B. Corey, and H.R. Branson, *The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain.* Proc Natl Acad Sci U S A, 1951. **37**(4): p. 205-11.
35. Pauling, L. and R.B. Corey, *Compound helical configurations of polypeptide chains: structure of proteins of the alpha-keratin type.* Nature, 1953. **171**(4341): p. 59-61.
36. Birbeck, M.S. and E.H. Mercer, *The electron microscopy of the human hair follicle. III. The inner root sheath and trichohyaline.* J Biophys Biochem Cytol, 1957. **3**(2): p. 223-30.
37. Birbeck, M.S. and E.H. Mercer, *The electron microscopy of the human hair follicle. II. The hair cuticle.* J Biophys Biochem Cytol, 1957. **3**(2): p. 215-22.
38. Birbeck, M.S. and E.H. Mercer, *The electron microscopy of the human hair follicle. I. Introduction and the hair cortex.* J Biophys Biochem Cytol, 1957. **3**(2): p. 203-14.
39. Rogers, G.E., *Electron microscope observations on the glassy layer of the hair follicle.* Exp Cell Res, 1957. **13**(3): p. 521-8.
40. Rogers, G.E., *Some observations on the proteins of the inner root sheath cells of hair follicles.* Biochim Biophys Acta, 1958. **29**(1): p. 33-43.
41. Rogers, G.E., *Some aspects of the structure of the inner root sheath of hair follicles revealed by light and electron microscopy.* Exp Cell Res, 1958. **14**(2): p. 378-87.
42. Rogers, G.E., *Electron microscopy of wool.* J Ultrastruct Res, 1959. **2**(3): p. 309-30.
43. Rogers, G.E., *Electron microscope studies of hair and wool.* Ann N Y Acad Sci, 1959. **83**: p. 378-99.
44. Downes, A.M., L.F. Sharry, and G.E. Rogers, *Separate Synthesis of Fibrillar and Matrix Proteins in the Formation of Keratin.* Nature, 1963. **199**: p. 1059-61.
45. Chapman, R.E. and R.T. Gemmell, *An ultrastructural autoradiographic study of the incorporation of (35S)cystine in the wool fibre cortex.* J Cell Sci, 1973. **13**(3): p. 811-9.
46. Bern, H.A., D.R. Harkness, and S.M. Blair, *Radioautographic Studies of Keratin Formation.* Proc Natl Acad Sci U S A, 1955. **41**(1): p. 55-60.
47. Clarke, R.M. and G.E. Rogers, *Protein synthesis in the hair follicle. II. Polysomes and amino acid incorporation.* J Invest Dermatol, 1970. **55**(6): p. 425-32.
48. Rogers, G.E., *Isolation and Properties of Inner Sheath Cells of Hair Follicles.* Exp Cell Res, 1964. **33**: p. 264-76.
49. Steinert, P.M., H.W. Harding, and G.E. Rogers, *The characterisation of protein-bound citrulline.* Biochim Biophys Acta, 1969. **175**(1): p. 1-9.
50. Swift, J.A. and B. Bews, *The chemistry of human hair cuticle-II: The isolation and amino acid analysis of the cell membranes and A-layer.* Journal of the Society of Cosmetic Chemists, 1974. **25**(7): p. 355-366.

51. Swift, J.A. and B. Bews, *The chemistry of human hair cuticle•I: A new method for the physical isolation of cuticle.* Journal of the Society of Cosmetic Chemists, 1974. **25**: p. 13-22.
52. Swift, J.A. and B. Bews, *The chemistry of human hair cuticle•III: The isolation and amino acid analysis of various subfractions of the cuticle obtained by pronase and trypsin digestion.* Journal of the Society of Cosmetic Chemists, 1976. **27**(6): p. 289-300.
53. Zahn, H., et al., *Wool as a biological composite structure.* Industrial and Engineering Chemistry Product Research and Development, 1980. **19**(4): p. 496-501.
54. Harding, H.W. and G.E. Rogers, *Isolation of peptides containing citrulline and the cross-link, epsilon-(gamma-glutamyl)lysine, from hair medulla protein.* Biochim Biophys Acta, 1976. **427**(1): p. 315-24.
55. Harding, H.W. and G.E. Rogers, *The occurrence of the -(-glutamyl)lysine cross-link in the medulla of hair and quill.* Biochim Biophys Acta, 1972. **257**(1): p. 37-9.
56. Gillespie, J.M. and R.L. Darskus, *Relation between the tyrosine content of various wools and their content of a class of proteins rich in tyrosine and glycine.* Aust J Biol Sci, 1971. **24**(6): p. 1189-97.
57. Chapman, G.V. and J.H. Bradbury, *The chemical composition of wool. 7. Separation and analysis of orthocortex and paracortex.* Arch Biochem Biophys, 1968. **127**(1): p. 157-63.
58. Clarke, R.M. and G.E. Rogers, *Protein synthesis in the hair follicle. I. Extraction and partial characterization of follicle proteins.* J Invest Dermatol, 1970. **55**(6): p. 419-24.
59. Gillespie, J.M. and R.C. Marshall, *A comparison of the proteins of normal and trichothiodystrophic human hair.* J Invest Dermatol, 1983. **80**(3): p. 195-202.
60. Marshall, R.C. and J.M. Gillespie, *High-sulphur proteins from alpha-keratins. II. Isolation and partial characterization of purified components from mouse hair.* Aust J Biol Sci, 1976. **29**(1-2): p. 11-20.
61. Marshall, R.C. and J.M. Gillespie, *High-sulphur proteins from alpha-keratins. I. Heterogeneity of the proteins from mouse hair.* Aust J Biol Sci, 1976. **29**(1-2): p. 1-10.
62. Gillespie, J.M., *Proteins rich in glycine and tyrosine from keratins.* Comp Biochem Physiol B, 1972. **41**(4): p. 723-34.
63. Powell, B.C., et al., *Mammalian keratin gene families: organisation of genes coding for the B2 high-sulphur proteins of sheep wool.* Nucleic Acids Res, 1983. **11**(16): p. 5327-46.
64. Marshall, R.C., *Characterization of the proteins of human hair and nail by electrophoresis.* J Invest Dermatol, 1983. **80**(6): p. 519-24.
65. Carracedo, A., L. Concheiro, and I. Requena, *The isoelectric focusing of keratins in hair followed by silver staining.* Forensic Sci Int, 1985. **29**(1-2): p. 83-9.
66. Carracedo, A., et al., *Isoelectric focusing patterns of some mammalian keratins.* J Forensic Sci, 1987. **32**(1): p. 93-9.
67. Kuczek, E. and G.E. Rogers, *Sheep keratins: characterization of cDNA clones for the glycine + tyrosine-rich wool proteins using a synthetic probe.* Eur J Biochem, 1985. **146**(1): p. 89-93.
68. Frenkel, M.J., et al., *The keratin BIIIB gene family: isolation of cDNA clones and structure of a gene and a related pseudogene.* Genomics, 1989. **4**(2): p. 182-91.
69. Kuczek, E.S. and G.E. Rogers, *Sheep wool (glycine + tyrosine)-rich keratin genes. A family of low sequence homology.* Eur J Biochem, 1987. **166**(1): p. 79-85.
70. Rogers, G.E., *Genes for hair and avian keratins.* Ann N Y Acad Sci, 1985. **455**: p. 403-25.
71. Rogers, M.A., et al., *Sequence data and chromosomal localization of human type I and type II hair keratin genes.* Exp Cell Res, 1995. **220**(2): p. 357-62.
72. MacKinnon, P.J., et al., *An ultrahigh-sulphur keratin gene of the human hair cuticle is located at 11q13 and cross-hybridizes with sequences at 11p15.* Mamm Genome, 1991. **1**(1): p. 53-6.
73. Rogers, M.A., et al., *Human hair keratin-associated proteins (KAPs).* Int Rev Cytol, 2006. **251**: p. 209-63.
74. Shimomura, Y. and M. Ito, *Human hair keratin-associated proteins.* J Investig Dermatol Symp Proc, 2005. **10**(3): p. 230-3.
75. Rogers, M.A. and J. Schweizer, *Human KAP genes, only the half of it? Extensive size polymorphisms in hair keratin-associated protein genes.* J Invest Dermatol, 2005. **124**(6): p. vii-ix.
76. Rogers, M.A., et al., *Characterization of new members of the human type II keratin gene family and a general evaluation of the keratin gene domain on chromosome 12q13.13.* J Invest Dermatol, 2005. **124**(3): p. 536-44.
77. Yahagi, S., et al., *Identification of two novel clusters of ultrahigh-sulfur keratin-associated protein genes on human chromosome 11.* Biochem Biophys Res Commun, 2004. **318**(3): p. 655-64.
78. Rogers, M.A., et al., *Hair keratin associated proteins: characterization of a second high sulfur KAP gene domain on human chromosome 21.* J Invest Dermatol, 2004. **122**(1): p. 147-58.
79. Langbein, L., et al., *K6irs1, K6irs2, K6irs3, and K6irs4 represent the inner-root-sheath-specific type II epithelial keratins of the human hair follicle.* J Invest Dermatol, 2003. **120**(4): p. 512-22.
80. Rogers, M.A., et al., *Characterization of a first domain of human high glycine-tyrosine and high sulfur keratin-associated protein (KAP) genes on chromosome 21q22.1.* J Biol Chem, 2002. **277**(50): p. 48993-9002.
81. Rogers, M.A., et al., *Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein genes embedded in the type I keratin gene domain on chromosome 17q12-21.* J Biol Chem, 2001. **276**(22): p. 19440-51.
82. Langbein, L., et al., *The catalog of human hair keratins. II. Expression of the six type II members in the hair follicle and the combined catalog of human type I and II keratins.* J Biol Chem, 2001. **276**(37): p. 35123-32.
83. Langbein, L., et al., *The catalog of human hair keratins. I. Expression of the nine type I members in the hair follicle.* J Biol Chem, 1999. **274**(28): p. 19874-84.
84. Parry, D.A., et al., *Human hair keratin-associated proteins: sequence regularities and structural implications.* J Struct Biol, 2006. **155**(2): p. 361-9.
85. Smith, T., Parry DAD, *Three-dimensional modelling of interchain sequence similarities and differences in the coil-coil segments of keratin intermediate filament heterodimers highlight features important in assembly.* Journal of Structural Biology, 2008. **162**: p. 139-151.
86. Smith, T.A. and D.A. Parry, *Sequence analyses of Type I and Type II chains in human hair and epithelial keratin intermediate filaments: promiscuous obligate heterodimers, Type II template for molecule formation and a rationale for heterodimer formation.* J Struct Biol, 2007. **158**(3): p. 344-57.
87. Herrmann, H., et al., *Intermediate filaments: from cell architecture to nanomechanics.* Nat Rev Mol Cell Biol, 2007. **8**(7): p. 562-73.
88. Fraser, R.D.B. and D.A.D. Parry, *Structural changes in the trichocyte intermediate filaments accompanying the transition from the reduced to the oxidized form.* Journal of Structural Biology, 2007. **159**: p. 36-45.
89. Lee, Y.J., R.H. Rice, and Y.M. Lee, *Proteome analysis of human hair shaft: from protein identification to posttranslational modification.* Mol Cell Proteomics, 2006. **5**(5): p. 789-800.
90. Plowman, J.E., *Proteomic database of wool components.* J Chromatogr B Analyt Technol Biomed Life Sci, 2003. **787**(1): p. 63-76.
91. Plowman, J.E., *The proteomics of keratin proteins.* J Chromatogr B Analyt Technol Biomed Life Sci, 2007. **849**(1-2): p. 181-9.

92. Plowman, J.E., et al., *Problems associated with the identification of proteins in homologous families: the wool keratin family as a case study.* Anal Biochem, 2002. **300**(2): p. 221-9.
93. Plowman, J.E., W.G. Bryson, and T.W. Jordan, *Application of proteomics for determining protein markers for wool quality traits.* Electrophoresis, 2000. **21**(9): p. 1899-906.
94. Plowman, J.E., et al., *Characterisation of low abundance wool proteins through novel differential extraction techniques.* Electrophoresis, 2010. **31**(12): p. 1937-46.
95. Plowman, J.E., et al., *The effect of oxidation or alkylation on the separation of wool keratin proteins by two-dimensional gel electrophoresis.* Proteomics, 2003. **3**(6): p. 942-50.
96. Koehn, H., et al., *The proteome of the wool cuticle.* J Proteome Res, 2010. **9**(6): p. 2920-8.
97. Deb-Choudhury, S., et al., *Electrophoretic mapping of highly homologous keratins: a novel marker peptide approach.* Electrophoresis, 2010. **31**(17): p. 2894-902.
98. Clerens, S., et al., *Developing the wool proteome.* J Proteomics, 2010. **73**(9): p. 1722-31.
99. Shimomura, Y. and A.M. Christiano, *Biology and genetics of hair.* Annu Rev Genomics Hum Genet, 2010. **11**: p. 109-32.
100. Tobin, D.J., et al., *The Fate of Hair Follicle Melanocytes During the Hair Growth Cycle.* Journal of Investigative Dermatology Symposium Proceedings 2004. **4**: p. 323-332.
101. Langbein, L., et al., *The keratins of the human beard hair medulla: the riddle in the middle.* J Invest Dermatol, 2010. **130**(1): p. 55-73.
102. Wagner, R.d.C.C., et al., *Electron microscopic observations of human hair medulla.* J Microsc, 2007. **226**: p. 54-63.
103. Kreplak, L., et al., *Investigation of human hair cuticle structure by microdiffraction: direct observation of cell membrane complex swelling.* Biochim Biophys Acta, 2001. **1547**(2): p. 268-74.
104. Jones, L.N., *Hair structure anatomy and comparative anatomy.* Clin Dermatol, 2001. **19**(2): p. 95-103.
105. Wolfram, L.J., *Human hair: a unique physicochemical composite.* J Am Acad Dermatol, 2003. **48**(6 Suppl): p. S106-14.
106. Bryson, W.G., et al., *Cortical cell types and intermediate filament arrangements correlate with fiber curvature in Japanese human hair.* J Struct Biol, 2009. **166**(1): p. 46-58.
107. Harland, D.P., et al., *Arrangement of trichokeratin intermediate filaments and matrix in the cortex of Merino wool.* J Struct Biol, 2011. **173**(1): p. 29-37.
108. Marshall, R.C., D.F. Orwin, and J.M. Gillespie, *Structure and biochemistry of mammalian hard keratin.* Electron Microsc Rev, 1991. **4**(1): p. 47-83.
109. Swift, J.A. and J.R. Smith, *Microscopical investigations on the epicuticle of mammalian keratin fibres.* J Microsc, 2001. **204**(Pt 3): p. 203-11.
110. Rogers, G. and K. Koike, *Laser capture microscopy in a study of expression of structural proteins in the cuticle cells of human hair.* Exp Dermatol, 2009. **18**(6): p. 541-7.
111. Smith, J.R. and J.A. Swift, *Lamellar subcomponents of the cuticular cell membrane complex of mammalian keratin fibres show friction and hardness contrast by AFM.* J Microsc, 2002. **206**(Pt 3): p. 182-93.
112. Thibaut, S., et al., *Human hair keratin network and curvature.* Int J Dermatol, 2007. **46 Suppl 1**: p. 7-10.
113. Schlake, T., *Determination of hair structure and shape.* Semin Cell Dev Biol, 2007. **18**(2): p. 267-73.
114. Thibaut, S., et al., *Human hair shape is programmed from the bulb.* Br J Dermatol, 2005. **152**(4): p. 632-8.
115. Ortonne, J.P. and G. Prota, *Hair melanins and hair color: ultrastructural and biochemical aspects.* J Invest Dermatol, 1993. **101**(1 Suppl): p. 82S-89S.
116. Slominski, A., et al., *Melanin pigmentation in mammalian skin and its hormonal regulation.* Physiol Rev, 2004. **84**(4): p. 1155-228.
117. Popescu, C. and H. Hocker, *Hair--the most sophisticated biological composite material.* Chem Soc Rev, 2007. **36**(8): p. 1282-91.
118. Sick, S., et al., *WNT and DKK determine hair follicle spacing through a reaction-diffusion mechanism.* Science, 2006. **314**(5804): p. 1447-50.
119. Clement, J.L., et al., *Ultrastructural study of the medulla of mammalian hairs.* Scan Electron Microsc, 1981(Pt 3): p. 377-82.
120. Clement, J.L., A. Le Pareux, and P.F. Ceccaldi, *The specificity of the ultrastructure of human hair medulla.* J Forensic Sci Soc, 1982. **22**(4): p. 396-8.
121. Briki, F., B. Busson, and J. Doucet, *Organization of microfibrils in keratin fibers studied by X-ray scattering modelling using the paracrystal concept.* Biochim Biophys Acta, 1998. **1429**(1): p. 57-68.
122. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics.* Annu Rev Genet, 2005. **39**: p. 309-38.
123. Drummond, A.J., et al., *Relaxed phylogenetics and dating with confidence.* PLoS Biol, 2006. **4**(5): p. e88.
124. Atkinson, H.J., et al., *Using sequence similarity networks for visualization of relationships across diverse protein superfamilies.* PLoS One, 2009. **4**(2): p. e4345.
125. Strelkov, S.V., H. Herrmann, and U. Aebi, *Molecular architecture of intermediate filaments.* Bioessays, 2003. **25**(3): p. 243-51.
126. Elaine Fuchs, K.W., *INTERMEDIATE FILAMENTS: Structure, Dynamics, Function, and Disease.* Annual Review of Biochemistry, 1994. **63**: p. 345-382.
127. Schweizer, J., et al., *New consensus nomenclature for mammalian keratins.* J Cell Biol, 2006. **174**(2): p. 169-74.
128. Moll, R., M. Divo, and L. Langbein, *The human keratins: biology and pathology.* Histochem Cell Biol, 2008. **129**(6): p. 705-33.
129. Bragulla, H.H. and D.G. Homberger, *Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia.* J Anat, 2009. **214**(4): p. 516-59.
130. Aoki, N., et al., *A novel type II cytokeratin, mK6irs, is expressed in the Huxley and Henle layers of the mouse inner root sheath.* J Invest Dermatol, 2001. **116**(3): p. 359-65.
131. Langbein, L., et al., *Against the rules: human keratin K80: two functional alternative splice variants, K80 and K80.1, with special cellular localization in a wide range of epithelia.* J Biol Chem, 2010. **285**(47): p. 36909-21.
132. Coulombe, P. and M. Omary, *Hard and Soft principles defining the structure, function and regulation of keratin intermediate filaments.* Current Opinion in Cell Biology, 2002. **14**: p. 110-122.
133. Parry, D.A., et al., *Towards a molecular description of intermediate filament structure and assembly.* Exp Cell Res, 2007. **313**(10): p. 2204-16.
134. Godsel, L.M., R.P. Hobbs, and K.J. Green, *Intermediate filament assembly : dynamics to disease.* Trends in Cell Biology, 2007. **18**(1): p. 28-37.
135. Rafik, M.E., et al., *In vivo formation steps of the hard alpha-keratin intermediate filament along a hair follicle: evidence for structural polymorphism.* J Struct Biol, 2006. **154**(1): p. 79-88.
136. Watts, N.R., et al., *Cryo-electron microscopy of trichocyte (hard alpha-keratin) intermediate filaments reveals a low-density core.* J Struct Biol, 2002. **137**(1-2): p. 109-18.
137. Fraser, R.D., P.M. Steinert, and D.A. Parry, *Structural changes in trichocyte keratin intermediate filaments during keratinization.* J Struct Biol, 2003. **142**(2): p. 266-71.
138. Wu, D.D., D.M. Irwin, and Y.P. Zhang, *Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair.* BMC Evol Biol, 2008. **8**: p. 241.
139. Rogers, M.A., et al., *Characterization and expression analysis of the hair keratin associated protein KAP26.1.* Br J Dermatol, 2008. **159**(3): p. 725-9.
140. Rogers, M.A., et al., *Characterization of human KAP24.1, a cuticular hair keratin-associated protein with unusual amino-acid composition and repeat structure.* J Invest Dermatol, 2007. **127**(5): p. 1197-204.
141. Er Rafik, M., J. Doucet, and F. Briki, *The intermediate filament architecture as determined by X-ray diffraction modeling of hard alpha-keratin.* Biophys J, 2004. **86**(6): p. 3893-904.

142. Langbein, L., et al., *Novel type I hair keratins K39 and K40 are the last to be expressed in differentiation of the hair: completion of the human hair keratin catalog.* J Invest Dermatol, 2007. **127**(6): p. 1532-5.
143. Jones, L.N. and F.M. Pope, *Isolation of intermediate filament assemblies from human hair follicles.* J Cell Biol, 1985. **101**(4): p. 1569-77.
144. McKinnon, A.J., *The self-assembly of keratin intermediate filaments into macrofibrils: Is this process mediated by a mesophase?* Current Applied Physics, 2006. **6**: p. 375-378.
145. Rogers, G.E., *Hair follicle differentiation and regulation.* Int J Dev Biol, 2004. **48**(2-3): p. 163-70.
146. Naito, S. and K. Arai, *Type and location of SS Linkages in Human Hair and Their Relation to Fiber Properties in Water.* Journal of Applied Polymer Science, 1996. **61**(12): p. 2113-2118.
147. Fraser, R.D.B., et al., *Disulphide bonding in α-keratin* International Journal of Biological Macromolecules, 1988. **10**(2): p. 106-112.
148. Parry, D.A.D., R.D.B. Fraser, and T.P. MacRae, *Repeating patterns of amino acid residues in the sequences of some high sulphur proteins from α-keratin* Int. J. Biol. Macromol., 1979. **1**(1): p. 17-22.
149. Hearle, J.W., *A critical review of the structural mechanics of wool and hair fibres.* Int J Biol Macromol, 2000. **27**(2): p. 123-38.
150. Koehn, H., et al., *Higher sequence coverage and improved confidence in the identification of cysteine-rich proteins from the wool cuticle using combined chemical and enzymatic digestion.* J Proteomics, 2009. **73**(2): p. 323-30.
151. Kim, S.Y., T.M. Jeitner, and P.M. Steinert, *Transglutaminases in disease.* Neurochem Int, 2002. **40**(1): p. 85-103.
152. Steinert, P.M. and L.N. Marekov, *Direct evidence that involucrin is a major early isopeptide cross-linked component of the keratinocyte cornified cell envelope.* J Biol Chem, 1997. **272**(3): p. 2021-30.
153. Steinert, P.M. and L.N. Marekov, *The proteins elafin, filaggrin, keratin intermediate filaments, loricrin, and small proline-rich proteins 1 and 2 are isodipeptide cross-linked components of the human epidermal cornified cell envelope.* J Biol Chem, 1995. **270**(30): p. 17702-11.
154. Thibaut, S., et al., *Transglutaminase-3 enzyme: a putative actor in human hair shaft scaffolding?* J Invest Dermatol, 2009. **129**(2): p. 449-59.
155. Thibaut, S., et al., *Transglutaminase 5 expression in human hair follicle.* J Invest Dermatol, 2005. **125**(3): p. 581-5.
156. Kempson, I.M., W.M. Skinner, and P.K. Kirkbride, *Calcium distributions in human hair by ToF-SIMS.* Biochim Biophys Acta, 2003. **1624**(1-3): p. 1-5.
157. Gumbiner, B.M., *Cell adhesion: the molecular basis of tissue architecture and morphogenesis.* Cell, 1996. **84**(3): p. 345-57.
158. Bazzi, H., et al., *Desmoglein 4 is expressed in highly differentiated keratinocytes and trichocytes in human epidermisand hair follicle.* Differentiation, 2006. **74**: p. 129-140.
159. Jones, L.N. and D.E. Rivett, *The role of 18-methyleicosanoic acid in the structure and formation of mammalian hair fibres.* Micron, 1997. **28**(6): p. 469-85.
160. Breakspear, S., J.R. Smith, and G. Luengo, *Effect of the covalently linked fatty acid 18-MEA on the nanotribology of hair's outermost surface.* J Struct Biol, 2005. **149**(3): p. 235-42.
161. Kalinin, A.E., A.V. Kajava, and P.M. Steinert, *Epithelial barrier function: assembly and structural features of the cornified cell envelope.* Bioessays, 2002. **24**(9): p. 789-800.
162. Mitu, A.M., *Damage assessment of human hair by electrophoretical analysis of hair proteins*. 2004, Rheinisch-WestfälischenTechnischen Hochschule Aachen Aachen.
163. Dyer, J.M., et al., *Proteomic evaluation and location of UVB-induced photo-oxidation in wool.* Journal of Photochemistry and Photobiology B: Biology, 2010. **98**: p. 118-127.
164. Dawber, R., *Hair: its structure and response to cosmetic preparations.* Clin Dermatol, 1996. **14**(1): p. 105-12.
165. Stewart, K., et al., *Surface Layer of Wool. I. Dityrosine Synthesis and Characterization.* Journal of Applied Polymer Science, 1997. **66**: p. 2359-2363.
166. Zahn, H. and H.G. Gattner, *Hair sulfur amino acid analysis.* EXS, 1997. **78**: p. 239-58.
167. Hilterhaus-Bong, S. and H. Zahn, *Contributions to the chemistry of human hair. 1. Analyses of cystine, cysteine and cystine oxides in untreated human hair.* Int J Cosmet Sci, 1987. **9**(3): p. 101-10.
168. Robinson, N.E., *Protein deamidation.* Proc Natl Acad Sci U S A, 2002. **99**(8): p. 5283-8.
169. Kossiakoff, A.A., *Tertiary structure is a principal determinant to protein deamidation.* Science, 1988. **240**(4849): p. 191-4.
170. Shimomura, Y., et al., *Mutations in the keratin 85 (KRT85/hHb5) gene underlie pure hair and nail ectodermal dysplasia.* J Invest Dermatol, 2010. **130**(3): p. 892-5.
171. Korge, B.P., et al., *Identification of novel mutations in basic hair keratins hHb1 and hHb6 in monilethrix: implications for protein structure and clinical phenotype.* J Invest Dermatol, 1999. **113**(4): p. 607-12.
172. Shimomura, Y., et al., *Autosomal-dominant woolly hair resulting from disruption of keratin 74 (KRT74), a potential determinant of human hair texture.* Am J Hum Genet, 2010. **86**(4): p. 632-8.
173. Gillespie, J.M., R.C. Marshall, and M. Rogers, *Trichothiodystrophy--biochemical and clinical studies.* Australas J Dermatol, 1988. **29**(2): p. 85-93.
174. CL Gummer, R.D., VH Price, *Thichothiodystrophy: an electron-histochemical study of the hair shaft.* British Journal of Dermatology, 1984. **110**(4): p. 439-449.
175. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
176. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.
177. Nilsson, T., et al., *Mass spectrometry in high-throughput proteomics: ready for the big time.* Nat Methods, 2010. **7**(9): p. 681-5.
178. Domon, B. and R. Aebersold, *Options and considerations when selecting a quantitative proteomics strategy.* Nat Biotechnol, 2010. **28**(7): p. 710-21.
179. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
180. Nesvizhskii, A.I., O. Vitek, and R. Aebersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry.* Nat Methods, 2007. **4**(10): p. 787-97.
181. Domon, B. and R. Aebersold, *Mass spectrometry and protein analysis.* Science, 2006. **312**(5771): p. 212-7.
182. Yates, J.R., C.I. Ruse, and A. Nakorchevsky, *Proteomics by mass spectrometry: approaches, advances, and applications.* Annu Rev Biomed Eng, 2009. **11**: p. 49-79.
183. Cravatt, B.F., G.M. Simon, and J.R. Yates, 3rd, *The biological impact of mass-spectrometry-based proteomics.* Nature, 2007. **450**(7172): p. 991-1000.
184. Han, X., A. Aslanian, and J.R. Yates, 3rd, *Mass spectrometry for proteomics.* Curr Opin Chem Biol, 2008. **12**(5): p. 483-90.
185. Parks, B.A., et al., *Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers.* Anal Chem, 2007. **79**(21): p. 7984-91.
186. Dunn, M.J., *Proteomics reviews 2011.* Proteomics, 2011. **11**(4): p. 509-12.
187. Waridel, P., et al., *Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated de novo sequencing.* Proteomics, 2007. **7**(14): p. 2318-29.
188. Adamidi, C., et al., *De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics.* Genome Res, 2011.

189. Lee, H.J., et al., *Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics.* Curr Opin Chem Biol, 2006. **10**(1): p. 42-9.
190. Annesley, T.M., *Ion suppression in mass spectrometry.* Clin Chem, 2003. **49**(7): p. 1041-4.
191. Blackburn, K., et al., *Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation.* J Proteome Res, 2010. **9**(7): p. 3621-37.
192. Lu, B., et al., *Improving protein identification sensitivity by combining MS and MS/MS information for shotgun proteomics using LTQ-Orbitrap high mass accuracy data.* Anal Chem, 2008. **80**(6): p. 2018-25.
193. Pan, S., et al., *Application of targeted quantitative proteomics analysis in human cerebrospinal fluid using a liquid chromatography matrix-assisted laser desorption/ionization time-of-flight tandem mass spectrometer (LC MALDI TOF/TOF) platform.* J Proteome Res, 2008. **7**(2): p. 720-30.
194. Huttlin, E.L., et al., *Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy.* J Proteome Res, 2007. **6**(1): p. 392-8.
195. Gallien, S., *Nouvelles méthodologies protéomiques d'aide à l'annotation des génomes et à la validation des séquences protéiques*. 2009, Université de Strasbourg.
196. Nesvizhskii, A.I., *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.* J Proteomics, 2010. **73**(11): p. 2092-123.
197. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.
198. Swaney, D.L., C.D. Wenger, and J.J. Coon, *Value of using multiple proteases for large-scale mass spectrometry-based proteomics.* J Proteome Res, 2010. **9**(3): p. 1323-9.
199. Bednarczyk, A., *Nouvelles méthodologies en protéomique pour une caractérisation fine des protéines*. 2009, Université de Strasbourg.
200. Weber, K. and M. Osborn, *The reliability of molecular weight determinations by dodecyl sulfate-polyacrylamide gel electrophoresis.* J Biol Chem, 1969. **244**(16): p. 4406-12.
201. Heller, M., et al., *Two-stage Off-Gel isoelectric focusing: protein followed by peptide fractionation and application to proteome analysis of human plasma.* Electrophoresis, 2005. **26**(6): p. 1174-88.
202. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.
203. Motoyama, A. and J.R. Yates, 3rd, *Multidimensional LC separations in shotgun proteomics.* Anal Chem, 2008. **80**(19): p. 7187-93.
204. Issaq, H.J., et al., *Multidimensional separation of peptides for effective proteomic analysis.* J Chromatogr B Analyt Technol Biomed Life Sci, 2005. **817**(1): p. 35-47.
205. Gilar, M., et al., *Orthogonality of separation in two-dimensional liquid chromatography.* Anal Chem, 2005. **77**(19): p. 6426-34.
206. Fairchild, J.N., K. Horvath, and G. Guiochon, *Approaches to comprehensive multidimensional liquid chromatography systems.* J Chromatogr A, 2009. **1216**(9): p. 1363-71.
207. Fairchild, J.N., K. Horvath, and G. Guiochon, *Theoretical advantages and drawbacks of on-line, multidimensional liquid chromatography using multiple columns operated in parallel.* J Chromatogr A, 2009. **1216**(34): p. 6210-7.
208. Horvath, K., J.N. Fairchild, and G. Guiochon, *Generation and limitations of peak capacity in online two-dimensional liquid chromatography.* Anal Chem, 2009. **81**(10): p. 3879-88.
209. Horvath, K., J.N. Fairchild, and G. Guiochon, *Detection issues in two-dimensional on-line chromatography.* J Chromatogr A, 2009. **1216**(45): p. 7785-92.
210. Horvath, K., J. Fairchild, and G. Guiochon, *Optimization strategies for off-line two-dimensional liquid chromatography.* J Chromatogr A, 2009. **1216**(12): p. 2511-8.
211. Delmotte, N., et al., *Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis.* J Proteome Res, 2007. **6**(11): p. 4363-73.
212. Gilar, M., et al., *Comparison of 1-D and 2-D LC MS/MS methods for proteomic analysis of human serum.* Electrophoresis, 2009. **30**(7): p. 1157-67.
213. Dwivedi, R.C., et al., *Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics.* Anal Chem, 2008. **80**(18): p. 7036-42.
214. Breci, L., et al., *Comprehensive proteomics in yeast using chromatographic fractionation, gas phase fractionation, protein gel electrophoresis, and isoelectric focusing.* Proteomics, 2005. **5**(8): p. 2018-28.
215. Blonder, J., et al., *Proteomic investigation of natural killer cell microsomes using gas-phase fractionation by mass spectrometry.* Biochim Biophys Acta, 2004. **1698**(1): p. 87-95.
216. Scherl, A., et al., *Genome-specific gas-phase fractionation strategy for improved shotgun proteomic profiling of proteotypic peptides.* Anal Chem, 2008. **80**(4): p. 1182-91.
217. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database issue): p. D154-9.
218. O'Donovan, C. and R. Apweiler, *A guide to UniProt for protein scientists.* Methods Mol Biol, 2011. **694**: p. 25-35.
219. Kaiser, J., *DNA sequencing. A plan to capture human diversity in 1000 genomes.* Science, 2008. **319**(5862): p. 395.
220. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.
221. Durbin, R.M., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
222. Rothberg, J.M. and J.H. Leamon, *The development and impact of 454 sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1117-24.
223. Kersey, P., H. Hermjakob, and R. Apweiler, *VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL.* Bioinformatics, 2000. **16**(11): p. 1048-9.
224. Elias, J.E., et al., *Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations.* Nat Methods, 2005. **2**(9): p. 667-75.
225. Yu, W., et al., *Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines.* Proteomics, 2010. **10**(6): p. 1172-89.
226. Jones, A.R., et al., *Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines.* Proteomics, 2009. **9**(5): p. 1220-9.
227. Geer, L.Y., et al., *Open mass spectrometry search algorithm.* J Proteome Res, 2004. **3**(5): p. 958-64.
228. Searle, B.C., *Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies.* Proteomics, 2010. **10**(6): p. 1265-9.
229. Alves, P., et al., *Fast and accurate identification of semi-tryptic peptides in shotgun proteomics.* Bioinformatics, 2008. **24**(1): p. 102-9.
230. Mann, M. and N.L. Kelleher, *Precision proteomics: the case for high resolution and high mass accuracy.* Proc Natl Acad Sci U S A, 2008. **105**(47): p. 18132-8.
231. Zubarev, R. and M. Mann, *On the proper use of mass accuracy in proteomics.* Mol Cell Proteomics, 2007. **6**(3): p. 377-81.
232. Ong, S.E. and M. Mann, *Mass spectrometry-based proteomics turns quantitative.* Nat Chem Biol, 2005. **1**(5): p. 252-62.
233. H. Kume, K.K., K. Nakatsugawa, S. Suzuki, and David Fatlowitz, *Ultrafast microchannel plate photomultipliers.* Applied Optics, 1988. **27**(6): p. 1170.
234. Wiza, J., *MICROCHANNEL PLATE DETECTORS.* Nuclear Instruments and Methods, 1979. **162**: p. 587-601.

235. Neue, U.D., *Theory of peak capacity in gradient elution.* J Chromatogr A, 2005. **1079**(1-2): p. 153-61.
236. Li, X., D.R. Stoll, and P.W. Carr, *Equation for peak capacity estimation in two-dimensional liquid chromatography.* Anal Chem, 2009. **81**(2): p. 845-50.
237. Goupy, J., ed. *Plans d'Expériences pour Surfaces de Réponse*. 1999, Dunod: Paris.
238. Sado, G. and M.C. Sado, *Les Plans d'Expériences*. 1991, Paris: Afnor Technique.
239. Goupy, J., ed. *La méthode des Plans d'Expériences*. 1988, Dunod: Paris.
240. Goupy, J., *What kind of experimental design for finding and checking robustness of analytical methods? .* Analytica Chimica Acta, 2005. **544**(1-2): p. 184-190.
241. Goupy, J., *Introduction aux plans d'Expériences*. 2001, Paris: Dunod.
242. Goupy, J., ed. *Introduction aux plans d'expériences : avec applications*. 2009, Dunod: Paris.
243. Doehlert, D.H., Appl. Stat., 1970. **19**(231).
244. Statsoft, I., *STATISTICA, version 6, www.statsoft.com.* 2001.
245. Delalande, F., et al., *Multigenic families and proteomics: extended protein characterization as a tool for paralog gene identification.* Proteomics, 2005. **5**(2): p. 450-60.
246. Creasy, D.M. and J.S. Cottrell, *Error tolerant searching of uninterpreted tandem mass spectrometry data.* Proteomics, 2002. **2**(10): p. 1426-34.
247. Choudhary, G., et al., *Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS.* J Proteome Res, 2003. **2**(1): p. 59-67.
248. Biringer, R.G., et al., *Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS.* Brief Funct Genomic Proteomic, 2006. **5**(2): p. 144-53.
249. Balgley, B.M., et al., *Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy.* Mol Cell Proteomics, 2007. **6**(9): p. 1599-608.
250. Corpet, F., *Multiple sequence alignment with hierarchical clustering.* Nucleic Acids Res, 1988. **16**(22): p. 10881-90.
251. Rodriguez, J., et al., *Does trypsin cut before proline?* J Proteome Res, 2008. **7**(1): p. 300-5.
252. Polevoda, B. and F. Sherman, *N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins.* J Mol Biol, 2003. **325**(4): p. 595-622.
253. Arnesen, T., et al., *Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans.* Proc Natl Acad Sci U S A, 2009. **106**(20): p. 8157-62.
254. Robbins, C.R. and C.H. Kelly, *Amino acid composition of human hair.* Textile Res J, 1970. **40**: p. 891-896.
255. Millington, K.R. and J.S. Church, *The photodegradation of wool keratin II. Proposed mechanisms involving cystine.* Journal of Photochemistry and Photobiology B: Biology, 1997. **39**: p. 204-212.
256. Kariya, N., Y. Shimomura, and M. Ito, *Size polymorphisms in the human ultrahigh sulfur hair keratin-associated protein 4, KAP4, gene family.* J Invest Dermatol, 2005. **124**(6): p. 1111-8.
257. Shimomura, Y., et al., *Characterization of human keratin-associated protein 1 family members.* J Investig Dermatol Symp Proc, 2003. **8**(1): p. 96-9.
258. Shimomura, Y., et al., *Polymorphisms in the human high sulfur hair keratin-associated protein 1, KAP1, gene family.* J Biol Chem, 2002. **277**(47): p. 45493-501.
259. Ku, N.O., et al., *Mutation of a major keratin phosphorylation site predisposes to hepatotoxic injury in transgenic mice.* J Cell Biol, 1998. **143**(7): p. 2023-32.
260. Herbert, B.R., et al., *Characterisation of wool intermediate filament proteins separated by micropreparative two-dimensional electrophoresis.* Electrophoresis, 1997. **18**(3-4): p. 568-72.
261. Herbert, B.R., A.L. Chapman, and D.A. Rankin, *Investigation of wool protein heterogeneity using two-dimensional electrophoresis with immobilised pH gradients.* Electrophoresis, 1996. **17**(1): p. 239-43.
262. Baskova, I.P. and L.L. Zavalova, *[Polyfunctionality of destabilase, a lysozyme from a medicinal leech].* Bioorg Khim, 2008. **34**(3): p. 337-43.
263. Zavalova, L.L., et al., *Recombinant destabilase-lysozyme: synthesis de novo in E. coli and action mechanism of the enzyme expressed in Spodoptera frugiperda.* Biochemistry (Mosc), 2004. **69**(7): p. 776-81.
264. Zavalova, L.L., et al., *Multiple forms of medicinal leech destabilase-lysozyme.* Biochem Biophys Res Commun, 2003. **306**(1): p. 318-23.
265. Pilcher, H., *Medicinal leeches: stuck on you.* Nature, 2004. **432**(7013): p. 10-1.
266. Fradkov, A., et al., *Enzyme from the medicinal leech (Hirudo medicinalis) that specifically splits endo-epsilon(-gamma-Glu)-Lys isopeptide bonds: cDNA cloning and protein primary structure.* FEBS Lett, 1996. **390**(2): p. 145-8.
267. Schafer, C., et al., *Identification and quantification of epsilon-(gamma-glutamyl)lysine in digests of enzymatically cross-linked leguminous proteins by high-performance liquid chromatography-electrospray ionization mass spectrometry (HPLC-ESI-MS).* J Agric Food Chem, 2005. **53**(8): p. 2830-7.
268. Ebeling, W., et al., *Proteinase K from Tritirachium album Limber.* Eur J Biochem, 1974. **47**(1): p. 91-7.
269. Hahn, H.W., et al., *Ultrafast microwave-assisted in-tip digestion of proteins.* J Proteome Res, 2009. **8**(9): p. 4225-30.
270. Johnson, G.V. and R. LeShoure, Jr., *Immunoblot analysis reveals that isopeptide antibodies do not specifically recognize the epsilon-(gamma-glutamyl)lysine bonds formed by transglutaminase activity.* J Neurosci Methods, 2004. **134**(2): p. 151-8.
271. Nemes, Z., et al., *Cross-linking of ubiquitin, HSP27, parkin, and alpha-synuclein by gamma-glutamyl-epsilon-lysine bonds in Alzheimer's neurofibrillary tangles.* FASEB J, 2004. **18**(10): p. 1135-7.
272. Nemes, Z., G. Petrovski, and L. Fesus, *Tools for the detection and quantitation of protein transglutamination.* Anal Biochem, 2005. **342**(1): p. 1-10.
273. Tang, Y., et al., *CLPM: a cross-linked peptide mapping algorithm for mass spectrometric analysis.* BMC Bioinformatics, 2005. **6 Suppl 2**: p. S9.
274. Singh, P., et al., *Characterization of protein cross-links via mass spectrometry and an open-modification search strategy.* Anal Chem, 2008. **80**(22): p. 8799-806.
275. Bringans, S.D., et al., *Characterization of the exocuticle a-layer proteins of wool.* Exp Dermatol, 2007. **16**(11): p. 951-60.
276. Rice, R.H., et al., *Proteomic analysis of human nail plate.* J Proteome Res, 2010. **9**(12): p. 6752-8.
277. de Berker, D.A., J. Andre, and R. Baran, *Nail biology and nail science.* Int J Cosmet Sci, 2007. **29**(4): p. 241-75.
278. De Berker, D., et al., *Keratin expression in the normal nail unit: markers of regional differentiation.* Br J Dermatol, 2000. **142**(1): p. 89-96.
279. Kitahara, T. and H. Ogawa, *Variation of differentiation in nail and bovine hoof cells.* J Invest Dermatol, 1994. **102**(5): p. 725-9.
280. Conrads, A., et al., *In vitro reconstitution of nail intermediate filaments.* Naturwissenschaften, 1988. **75**(2): p. 100-1.
281. Baden, H.P., *The physical properties of nail.* J Invest Dermatol, 1970. **55**(2): p. 115-22.
282. Boschetti, E. and P.G. Righetti, *The ProteoMiner in the proteomic arena: a non-depleting tool for discovering low-abundance species.* J Proteomics, 2008. **71**(3): p. 255-64.
283. Righetti, P.G. and E. Boschetti, *The ProteoMiner and the FortyNiners: searching for gold nuggets in the proteomic arena.* Mass Spectrom Rev, 2008. **27**(6): p. 596-608.
284. Cedano, J., et al., *Relation between amino acid composition and cellular location of proteins.* J Mol Biol, 1997. **266**(3): p. 594-600.
285. Eckhart, L., et al., *Identification of reptilian genes encoding hair keratin-like proteins suggests a new scenario for the evolutionary origin of hair.* Proc Natl Acad Sci U S A, 2008. **105**(47): p. 18419-23.

286. Hearle, J.W., *Proteins fibers: structural mechanics and future opportunities.* J. Mater. Sci., 2007. **42**: p. 8010-8019.
287. Fraser, B.R.D. and D.A.D. Parry, *The structural basis of the filament-matrix texture in the avian/reptilian group of hard b-keratins.* Journal of Structural Biology, 2010.
288. Witmer, L.M., *Dinosaurs: Fuzzy origins for feathers.* Nature, 2009. **458**(7236): p. 293-5.
289. Witmer, L.M., *Palaeontology: Feathered dinosaurs in a tangle.* Nature, 2009. **461**(7264): p. 601-2.
290. Fraser, R.D. and D.A. Parry, *Molecular packing in the feather keratin filament.* J Struct Biol, 2008. **162**(1): p. 1-13.
291. Toni, M., L.D. Valle, and L. Alibardi, *Hard (Beta-)keratins in the epidermis of reptiles: composition, sequence, and molecular organization.* J Proteome Res, 2007. **6**(9): p. 3377-92.
292. Alibardi, L. and B.J. Gill, *Epidermal differentiation in embryos of the tuatara Sphenodon punctatus (Reptilia, Sphenodontidae) in comparison with the epidermis of other reptiles.* J Anat, 2007. **211**(1): p. 92-103.
293. Alibardi, L., *Structural and immunocytochemical characterization of keratinization in vertebrate epidermis and epidermal derivatives.* Int Rev Cytol, 2006. **253**: p. 177-259.
294. Prum, R.O. and A.H. Brush, *The evolutionary origin and diversification of feathers.* Q Rev Biol, 2002. **77**(3): p. 261-95.
295. Zhang, F. and Z. Zhou, *A primitive enantiornithine bird and the origin of feathers.* Science, 2000. **290**(5498): p. 1955-9.
296. Prum, R.O., *Development and evolutionary origin of feathers.* J Exp Zool, 1999. **285**(4): p. 291-306.
297. Spearman, R.I., *The keratinization of epidermal scales, feathers and hairs.* Biol Rev Camb Philos Soc, 1966. **41**(1): p. 59-96.
298. de Parseval, N., et al., *Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins.* J Virol, 2003. **77**(19): p. 10414-22.
299. McDaniel, L.D., et al., *High frequency of horizontal gene transfer in the oceans.* Science, 2010. **330**(6000): p. 50.
300. Fuerst, J.A., et al., *Isolation and molecular identification of planctomycete bacteria from postlarvae of the giant tiger prawn, Penaeus monodon.* Appl Environ Microbiol, 1997. **63**(1): p. 254-62.
301. Fraser, R.D. and D.A. Parry, *Macrofibril assembly in trichocyte (hard alpha-) keratins.* J Struct Biol, 2003. **142**(2): p. 319-25.
302. Fournier, M.L., et al., *Multidimensional separations-based shotgun proteomics.* Chem Rev, 2007. **107**(8): p. 3654-86.
303. Wolters, D.A., M.P. Washburn, and J.R. Yates, 3rd, *An automated multidimensional protein identification technology for shotgun proteomics.* Anal Chem, 2001. **73**(23): p. 5683-90.
304. Folk, J.E. and J.S. Finlayson, *The epsilon-(gamma-glutamyl)lysine crosslink and the catalytic role of transglutaminases.* Adv Protein Chem, 1977. **31**: p. 1-133.
305. Yoneda, K., et al., *Expression of transglutaminase 1 in human hair follicles, sebaceous glands and sweat glands.* Br J Dermatol, 1998. **138**(1): p. 37-44.
306. Zhang, Y. and D. Reinberg, *Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails.* Genes Dev, 2001. **15**(18): p. 2343-60.

Partie expérimentale

Analyse du protéome endo cuticulaire

Extraction physique de la cuticule

Inspiré du protocole décrit par Swift et Bews [51].

Des mèches d'environ 3-4 cm de cheveu caucasien non traité (référence BAN) sont utilisées. Une poignée de mèches est trempée dans un grand bécher avec une solution aqueuse de SDS à 1%. L'ensemble est placé dans un bac à ultrason pendant une demi heure. Les cheveux sont alors rincés abondamment avec de l'eau au moyen d'un chinois puis avec de l'eau déminéralisée. Les cheveux nettoyés sont séchés.

Les lipides libres sont extraits avec un mélange dichlorométhane/méthanol 50/50 (200 mL) sous ultrasons pendant une heure. Les mèches sont filtrées et les lipides fixés sont alors extraits avec une solution de KOH 0,1 M dans le méthanol pendant 30 min. Des précautions permettant d'éviter la contamination épidermale sont nécessaires à partir de cette étape. Les fibres sont récupérés et rincées abondamment avec de l'eau déminéralisée L'extraction des lipides libres est réalisée une nouvelle fois puis les cheveux sont séchés. Les fibres après séchage sont rêches et cassantes.

300 mg de ces cheveux sont pesés et coupés à 2 cm de long environ. Ils sont placés dans un flacon plastique cylindrique de 40 mL avec 15 mL d'eau disposé sur un bras articulé (Intelli-Mixer RM-2, ELMI Ltd., Riga, Latvia). Deux flacons peuvent être réalisés à la fois sur le dispositif. Le programme d'agitation choisi est u30 réalisé pendant quelques heures (typiquement 4 heures). La solution obtenue est trouble. Les cheveux sont retirés et la suspension est récupérée et placée dans un falcon de 15 mL. Les falcons sont centrifugés pour récupérer les suspensions de fragments cuticulaires pendant 10 min à 4000 trs/min. L'eau ne contenant pas de suspension est retirée ce qui permet de concentrer. Les éventuels résidus de cheveu pouvant encore être présents sont enlevés à l'aide d'une spatule pour éviter une contamination corticale. 12 suspensions récupérées et rassemblées (soit 12 x 300 mg de cheveu initial) conduisent après lyophilisation dans un eppendorf taré à environ 10 mg de matériel (NB : la lyophilisation est facultative et rend l'échantillon plus difficile à resuspendre en solution) . L'échantillon est conservé à -20°C.

Préparation des extraits cuticulaires pour la digestion enzymatique et la digestion chimique

L'échantillon est divisé en deux après avoir été suspendu dans de l'eau. La première partie est utilisée pour réaliser directement les digestions enzymatiques, la seconde pour la digestion chimique.

Digestions enzymatiques

L'échantillon utilisé pour la digestion enzymatique est réduit 16 h à 37°C dans 1 mL de tampon d'extraction (50 mM Tris HCl, 50 mM DTT, 7 M urée, 2 M thiourée, 2% SDS) puis soniqué trois heures. L'ensemble est précipité (lyophilisation à environ 300 µl puis ajout d'1 mL d'EtOH froid puis centrifugation). L'insoluble réduit est séparé du surnageant et alkylé pendant 1 heure à 60°C avec 100 mg d'iodoacétamide dans 1,5 mL de tampon carbonate d'ammonium à 25 mM. La solution est concentrée par speedvac (environ 500 µL) puis 1 ml d'EtOH froid est ajouté. Centrifugation et récupération de l'insoluble resuspendu dans 150 µL de tampon carbonate d'ammonium à 25 mM. L'échantillon resupsendu est divisé en trois respectivement pour les digestions à la trypsine, la chymotrypsine et la GluC.

Chaque échantillon est respectivement digéré dans 700 µL de tampon contenant 100 mM de NH_4HCO_3 et 2 M d'urée avec 10 µg de trypsine ou 17 µg de chymotrypsine. La digestion à la GluC est réalisée dans 700 µL de

tampon phosphate dibasique à 25 mM et 1 M d'urée avec 33 µg d'enzyme. La digestion est réalisée toute la nuit à 37°C pour la trypsine et à 25 °C pour la chymotrypsine et la GluC.

Digestion chimique NTCB

L'extrait cuticulaire est placé dans 1 mL de tampon de digestion composé de 5 M guanidine-HCl, DTT 20 mM, Tris pH 8.2 25 mM et glycine 4 M (d'après Koehn et al [150]).

Le NTCB est ajouté à 50 mM (solution jaune). Le pH est contrôlé pour que la réaction se fasse autour de pH 9 (solution orange à ce pH). Quelques gouttes de NH_4OH peuvent être ajoutées pour remonter le pH. Une sonication peut être réalisée pour solubiliser le NTCB. La réaction est réalisée une nuit à 37°C. Un gonflement de la cuticule peut être observé suite à cette étape par rapport aux fragments cellulaires de départ. L'ensemble est centrifugé et le surnageant enlevé. Le culot est rincé avec du tampon Tris-HCl puis divisé pour les différentes digestions. Chaque échantillon est centrifugé et le culot est placé dans 700 µL de tampon de digestion comme réalisé précédemment.

Purification des digests

Les différents surnageants obtenus à la suite des 6 digestions sont récupérés et purifiés sur des cartouches SepPack C18 de 50 mg selon le protocole déjà employé pour le cortex (lavage de la cartouche avec 2x1 mL de MeOH puis 2x1 mL d'ACN et 3x1mL H_2O 0,1% acide formique ; chargement de l'échantillon acidifié puis rinçage avec 3x1mL H_2O 0,1% acide formique et élution avec 600 µL d'ACN 50% 0,1% acide formique). Les digests sont alors speedvaqués à une dizaine de µL pour être injectés pour fractionnement sur microLC puis analyse des collectes en nanoLC-MS/MS.

Analyse du protéome des ongles

Extraction des onychocytes.

Adaptation du protocole de Rice et al [276] en utilisant le mélange d'extraction utilisé précédemment pour l'extraction du cortex.

350 mg d'extrémités d'ongles sont collectées. Environ 30 mg sont prélevés soit trois ou quatre morceaux. Les morceaux sont nettoyés dans un tube avec une solution de SDS à 2% sous agitation. Ils sont ensuite rincés abondamment à l'eau à l'aide d'une pince à thé.

Les ongles sont placés à 45°C dans 1 mL de tampon d'extraction pour réduction (50 mM Tris HCl, 50 mM DTT, 7 M urée, 2 M thiourée, 2% SDS) et laissés une nuit à 45°C sans agitation. Suite à cette étape les ongles gonflent. Les ongles sont alors déposés avec le tampon dans un petit erlen de 10 mL et 3 mL de tampon d'extraction sont ajoutés ainsi qu'un mini barreau magnétique. L'ensemble est mis au bain marie à 45°C sur un agitateur magnétique chauffant. L'agitation est vigoureuse, le barreau tape dans l'erlen pour obtenir un régime turbulent. La pulvérisation est réalisée pendant 12 heures. Suite à cette étape, les morceaux d'ongles semblent dissouts et l'extrait se présente comme une solution hétérogène. L'extrait est concentré au speedvac puis précipité en ajoutant de l'éthanol à froid (2 volumes d'éthanol pour 1 volume d'extrait concentré). Le précipité est repris dans le tampon de réduction puis l'alkylation est réalisée pendant une heure à l'abri de la lumière après avoir ajouté 100 mg d'iodoacétamide. L'extrait réduit et alkylé est précipité à l'éthanol. Un précipité gélatineux est obtenu et est divisé à la spatule pour réaliser les digestions à la trypsine, à la chymotrypsine et à la GluC.

Les digestions sont réalisées respectivement dans 700 µL de tampon NH_4HCO_3 100 mM 2 M urée avec 10 µg de trypsine ou 17 µg de chymotrypsine. La digestion GluC est réalisée dans 700 µL de tampon tampon phosphate dibasique 25 mM, 1 M urée avec 33 µg d'enzyme. Les quantités d'enzyme ont été choisies sur une base d'environ 1 mg de matériel digéré.

Annexe 1 Communications par affiches

1. 58th ASMS Conference on Mass Spectrometry and Allied Topics, Salt Lake City, Utah, Mai 2010. ***Comprehensive study of significant nano-LC-MS/MS parameters for proteomics analysis on Q-TOF mass spectrometers thanks to design of experiment***. Barthélemy N, Brennetot R, Carapito C, Schaeffer C, Van Dorsselaer A.

2. 27èmes Journées Françaises de la Spectrométrie de Masse, Clermont Ferrand, Septembre 2010. ***Data Dependent Acquisition optimization on a Q-TOF instrument for proteomic applications***. Colas C; Barthélemy N, Carapito C, Husser C, Schaeffer C, Van Dorsselaer A.

Comprehensive study of significant nanoLC-MS/MS parameters for proteomic analysis on Q-TOF mass spectrometers thanks to design of experiment

Nicolas Barthélemy[1]; René Brennetot[2]; Christine Carapito[1]; Christine Schaeffer[1]; Alain Van Dorsselaer[1]

1- *LSMBO, IPHC-DSA, UMR7178, F-67087 Strasbourg, France*
2- *CEA Saclay DEN/DANSDPC/SECR/LANIE, F-91191 Gif sur Yvette, France*

Overview

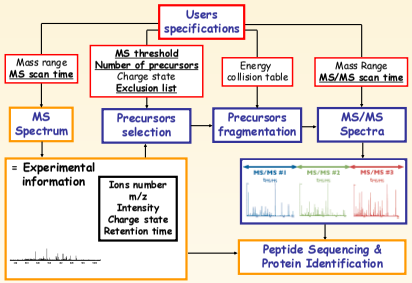
Aim :
Investigating the significance of several nanoLC-MS/MS parameters on protein identification using Q-TOF mass spectrometers.

Methods :
Using Doehlert design of experiment to establish existing links between 7 nanoLC-MS/MS variables and 6 bottom-up proteomic responses.

Results :
- Accurate system modeling found with a fast and optimized design.
- Significant effects obtained for each studied response and quantification of these effects.
- Well-thought-out optimization strategy used to find suitable acquisition methods.

Introduction

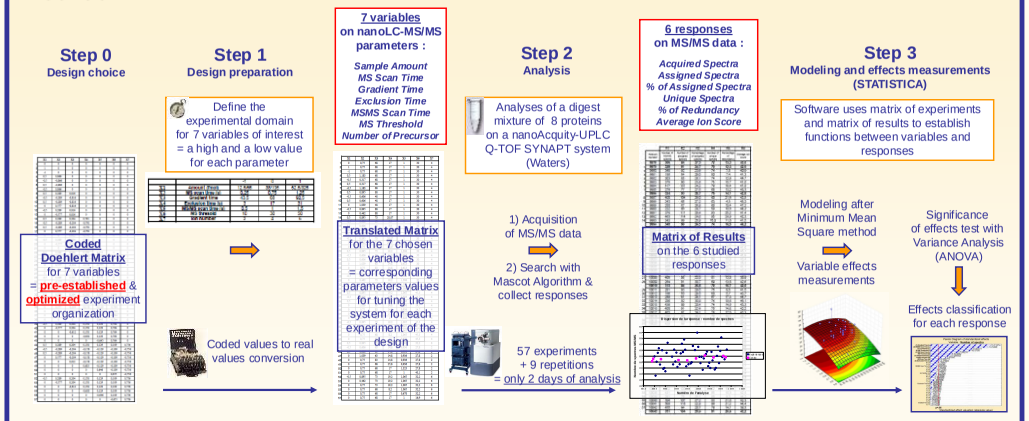
In bottom-up proteomic analysis, getting the most information in a single nanoLC-MS/MS run to save time and sample consuming requires the best tuning of the chromatograph and the mass spectrometer.
Data Dependent Acquisition (DDA) LC-MS/MS analyses depend on several factors that influence the numbers of peptides and proteins identified and thus the quality of the results.



Data Dependant Acquisition scheme

To understand which parameters have the most critical impact on the results, a statistical design of experiment was performed as an alternative to the conventional method consisting of varying the parameters one by one. In addition to 5 DDA parameters (underlined on the DDA scheme), we also focused on the chromatographic gradient time and the sample injected amount.

Method



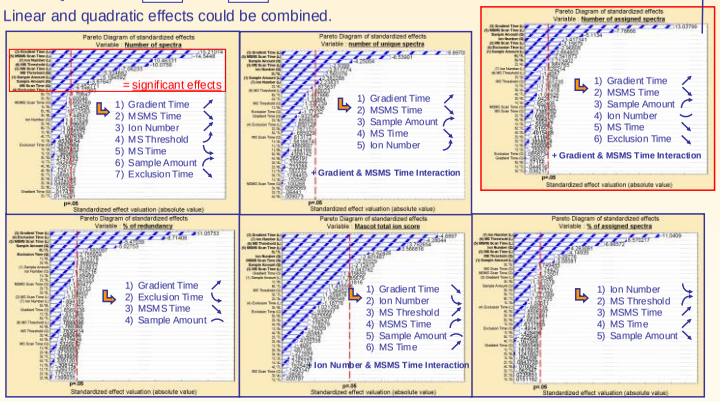
Results

For each response, parameters effects were measured. An effect could be :

- Linear = (+ ↗) or (- ↘)
- Quadratic = (+ ⌣) or (- ⌢)

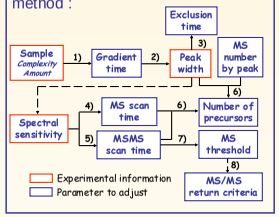
Linear and quadratic effects could be combined.

Quantitative and standardized values of effects was obtained and classified. Information were summarized in **Pareto diagrams**.
Only significant effects after ANOVA was kept as results.



For example, the number of identified peptides raises with the increasing of gradient time and the decreasing of MS/MS scan time.
= Scan times impact on identification suggests the need to control these values to obtain spectra in an accumulation time as short as possible.

Now this study allows us to propose a workflow to develop an optimized method :



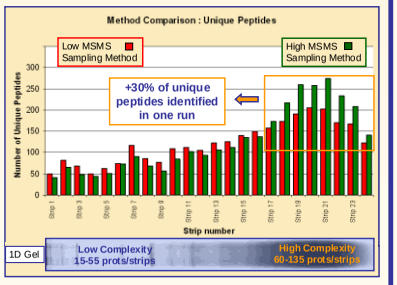
Application

According to our optimized workflow, a DDA method was developed to get a high MS/MS sampling for complex sample. This method was confronted with a non optimized one.

	Low MSMS Sampling Method	High MSMS Sampling Method
Gradient Time	30 min	30 min
Precursor Number	3	5
MS Scan Time	1s	0.5s
MS/MS Scan Time	2x0.7s	0.7s or 3x0.7s
MS Threshold	30	20
Exclusion time	10s	30s

Parameters values summarization

To compare method efficiency, a protein extract from Yeast was fractionated with 1D gel electrophoresis. Gel was cut on 24 strips analyzed with bottom up proteomic workflow. Each strip was consecutively analyzed with the two respective methods.



The high MS/MS sampling method is more efficient for complex sample.

Conclusion

The results obtained in this study allow a comprehensive adjustment of LC-MS/MS parameters to have an optimized method for proteomic studies. Complex sample (60 to 135 proteins by run) can be analyzed with an improvement of 30% in the number of unique peptides compare to a non optimized method.
Using design of experiments reveals to be an efficient tool to better understand the mechanisms of Data Dependant Acquisition on LC-MS/MS.
The resulting method improvement allows identifying more peptides for better protein sequence coverage in a single run.

Data Dependant Acquisition optimization on a Q-TOF instrument for proteomic applications

Cyril COLAS[1,2], Nicolas BARTHELEMY[1,2], Christine CARAPITO[1,2], Chrystel HUSSER[1,2], Christine SCHAEFFER[1,2] and Alain VAN DORSSELAER[1,2]
[1]Université de Strasbourg, IPHC, 25 rue Becquerel 67087 Strasbourg, France
[2]CNRS, UMR 7178, 67037 Strasbourg, France

Overview

Purpose: To improve peptide identification rates in nanoLC-MS/MS analyses.

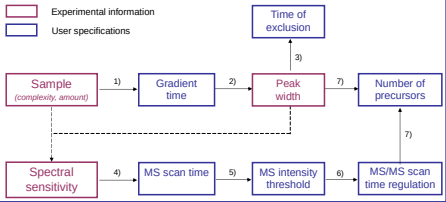
Method: Data Dependent Acquisition (DDA) parameters optimized.

Results: Increase of the number of identified proteins with a better sequence coverage.

Introduction

Because of the stochastic behaviour of the nanoLC-MS/MS experiments, the number of peptides usually identified by MS/MS is lower than that eluted from the chromatography. To increase the number of identified peptides, a Data Dependent Acquisition (DDA) mode for LC-MS/MS analyses has been introduced by manufacturers. This DDA mode depends on several parameters, which can be optimized to improve identification rates of peptides.

We determined the best optimization sequence of the different parameters (see diagram below) and explained how they were set to their optimal value, including the new and specific functionality of Bruker's Q-TOF: regulation of MS/MS scan time according to precursor MS intensities.
We present here a method which allows optimizing the DDA parameters on a nanoAcquity coupled to a maXis mass spectrometer.



nanoLC-MS/MS optimization diagram
This diagram shows the order to optimize the different parameters, including DDA ones.

Chromatography

Chromatography was performed using a nanoAcquity UPLC BEH130 C_{18} column (75 µm x 200 mm, 1.7 µm) heated at 50°C with water and acetonitrile, both with 0.1% formic acid at a flow rate of 450 nL/min. In these conditions the dead volume of the system was 6.6 min.

	Acetonitrile (%)	RT (min)	Slope (%/min)		RT (min)	Slope (%/min)		RT (min)	Slope (%/min)
	1	0.2			0.2			0.2	
	3	0.4			0.4			0.4	
4-S lope gradient	6	1.2	3.64		0.8	7.29		2.0	1.82
	14	11.7	0.76		6.1	1.52		23.0	0.38
	22.5	20.1	1.01		10.3	2.02		39.8	0.51
Gradient time	**35**	**29.2**	**1.37**		**14.8**	**2.75**		**58.0**	**0.69**
Rinsing	90	30.2	+ 1 min		15.8			59.0	
	1	31.2	+ 1 min		16.8			60.0	
Equilibration	1	36.2	+ 5 min		21.8			65.0	

x 2 (first → second column)
x 0.5 (second → third column)

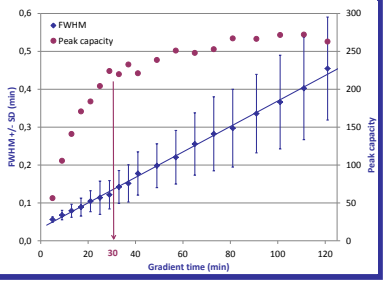
4-Slope gradient table
Slopes were adjusted in order to obtain constant peak width during the whole run, giving a 4-slope gradient.

→ Other gradient times with constant peak width can be computed by use of multiplicative coefficients on slopes

$$Peak\_capacity = \frac{Rt_{last} - Rt_{first}}{FWHM}$$

FWHM, Peak capacity = f (gradient time)
The Full Width at Half Maximum (FWHM) linearly increases with gradient time.

→ The optimum peak capacity is obtained with a gradient of 30 min



Mass spectrometry

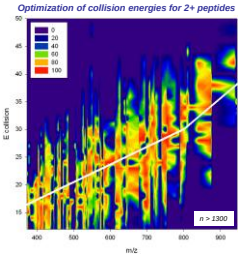
DDA Parameters

Collision energies
LC-MS/MS runs of a 4-protein mixture digest were performed at constant collision energies from 12 to 50 eV.

- *The quality of the fragmentation was evaluated by normalizing best Mascot MS/MS ion scores for each peptide to 100*

→ Peptides are well fragmented on a wide range of energies
→ Good MS/MS can be obtained even if some precursor is still present
→ Some peptides can't be fragmented whatever collision energy

Optimization of collision energies for 2+ peptides

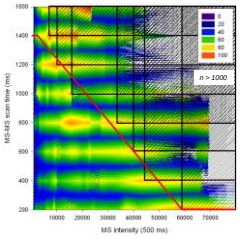


MS/MS scan time
A new and specific functionality of Bruker's Q-TOF is the regulation of MS/MS scan time according to precursor MS intensities.
8 LC-MS/MS runs of a 4-protein mixture digest were performed at constant MS/MS scan times from 200 to 1600 ms.

- *The quality of the fragmentation was evaluated by use of absolute Mascot MS/MS ion scores*
- *The correlation between MS/MS scan time and MS intensity was established for each MS/MS scan time (see Figure)*

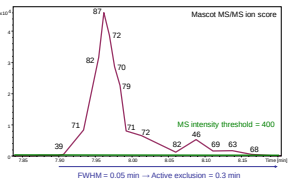
→ Adjustment of MS/MS scan time according to sample amounts is useless
→ More peptides identified per analysis

Optimization of MS/MS scan time according to MS intensities



Time of exclusion
Exclusion of an ion, which has already been selected for MS/MS from the list of precursors, reduces the number of redundant spectra and allows taking less abundant ions as precursors, thus leading to increase the number of identified peptides.
→ Exclusion after 1 spectrum
→ Exclusion time = whole width of the peak (~ 6 FWHM)

Lack of a "peptide exclusion" algorithm, taking charge states into account (~ 15% of redundant peptides).



Number of precursors
The number of precursor ions has been adjusted to have between 1 and 5 MS spectra per peak width at FWHM.

Lack of a retro-control to automatically adjust the number of precursor ions: the number of precursors has to be adjusted according to sample amounts.

Gradient time (min)	FWHM (s)	Exclusion time (s)	number of precursors	Cycle time (s) min - max
4	3.5	18	2	0.6 - 3.0
9	4.1	24	3	0.8 - 4.4
21	6.2	36	4	1.0 - 5.8
29	7.3	48	5	1.2 - 7.2

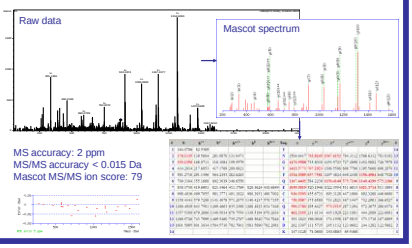
Export of raw data

Conversion of raw data to export files
This step of the process is crucial to maximize sequence information without false positive results.

How to improve Mascot MS/MS ion scores?
→ Limited noise export
- *Mascot ion scores are mainly dependent of the ratio between the number of identified fragments and the number of submitted ions*

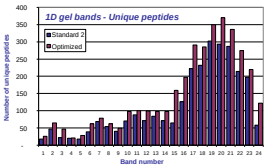
How to decrease the risk of false positive?
→ Use of high accuracy to reduce database searches:
- *UHRMS (Ultra High Resolution Mass Spectrometry)*
- *Lock Mass*

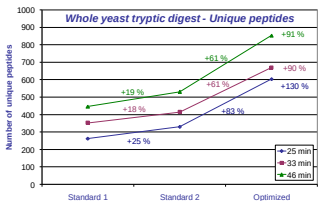
→ Error tolerance of 5 ppm for MS and 0.02 Da for MS/MS in search engines



Application to a yeast digest

1D gel bands
Our optimized set of parameters was first tested on 1mm 1D gel bands of a yeast tryptic digest. The number of identified peptides was increased by about 25% compared with the standard parameters. As a consequence, the number of identified proteins was increased by 10%.





Whole yeast tryptic digest
On a much complex sample, 3 methods with 3 gradient times were evaluated:
-Standard 1: 3 precursors with constant MS/MS scan time (1600 ms)
-Standard 2: 5 precursors with regulated MS/MS scan time (from 600 to 2000 ms)
-Optimized: 5 precursors with regulated MS/MS scan time (from 200 to 1400 ms)

The optimized parameters significantly increased the number of identified peptides and proteins.

Conclusions

Methodology
The methodology described here to optimize Data Dependent Acquisition parameters on a Q-TOF instrument for proteomic applications has been used with success on a nanoAcquity coupled to a maXis mass spectrometer. These methodology can be applied on any Q-TOF.

Results
After optimization of the Data Dependent Acquisition parameters, the number of identified peptides was increased by about 25% for low complexity samples (1D gel bands) and by more than 90% for more complex samples (whole digest).
Thanks to the MS/MS scan time regulation, a new and specific functionality of Bruker's Q-TOF, which needs to be correctly adjusted, these improvements were possible.

Software limitations
Some software improvements can still be performed:
- *Introduction of a retro-control loop to automatically adjust the number of precursor ions between two MS spectra performed at constant interval of time.*

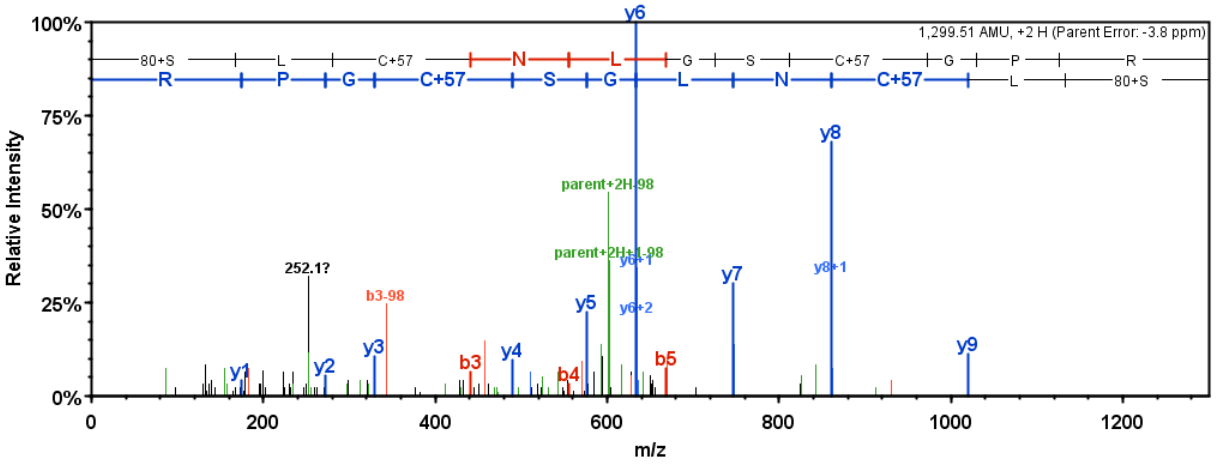
→ Only exclusion time and interval between two MS spectra will have to be adjusted according to gradient time (depending on sample complexity), independently of sample amounts

- *Introduction of a "peptide exclusion" algorithm, taking charge states into account.*

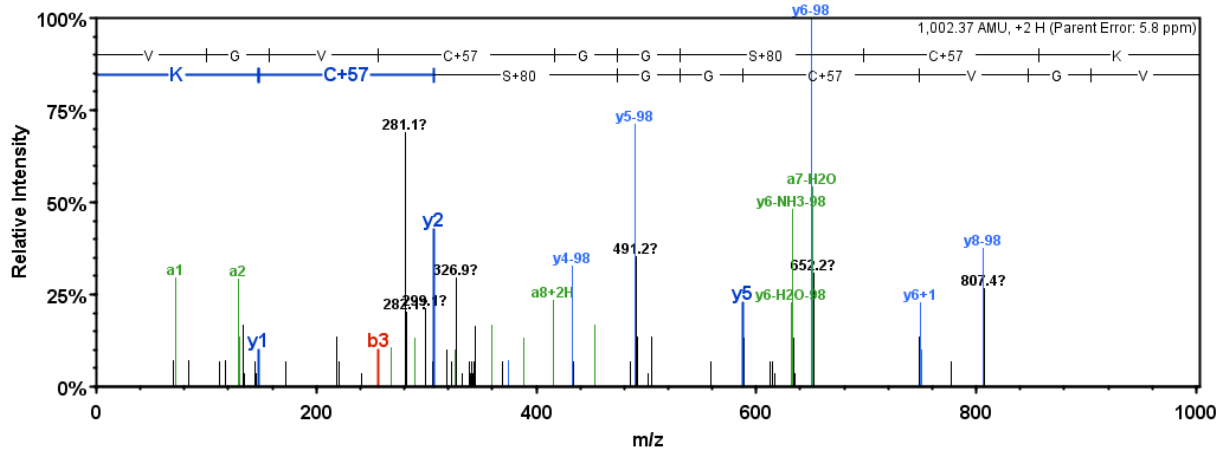
→ Identification of more compounds by increased exclusion of redundant peptides: about 15% of MS/MS spectra were attributed to redundant peptides with different charge states

Annexe 2 : Spectres de fragmentation de peptides phosphorylés des kératines de type II

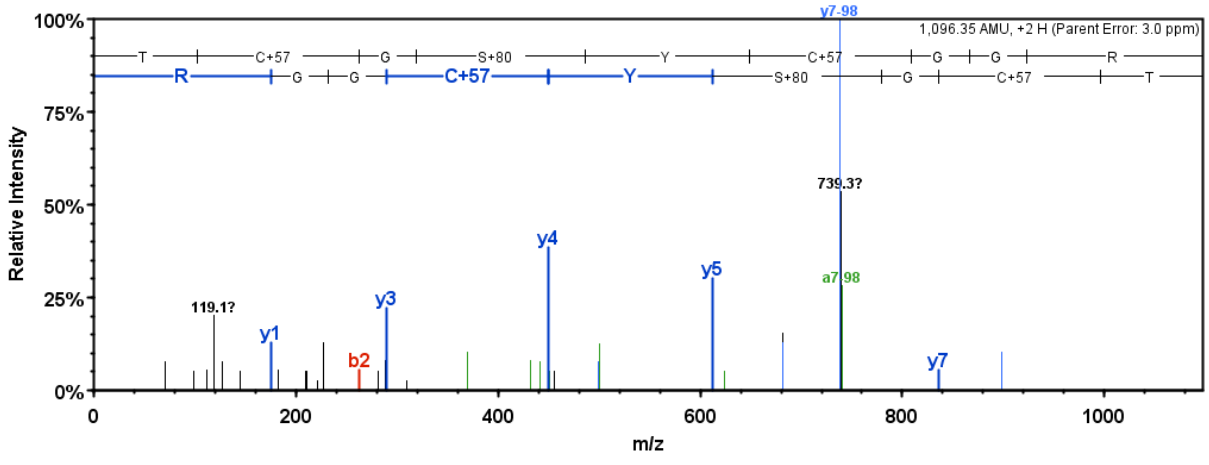
Spectres identifiés dans les données de séquençage des digests de cortex après la recherche en mode error tolerant.



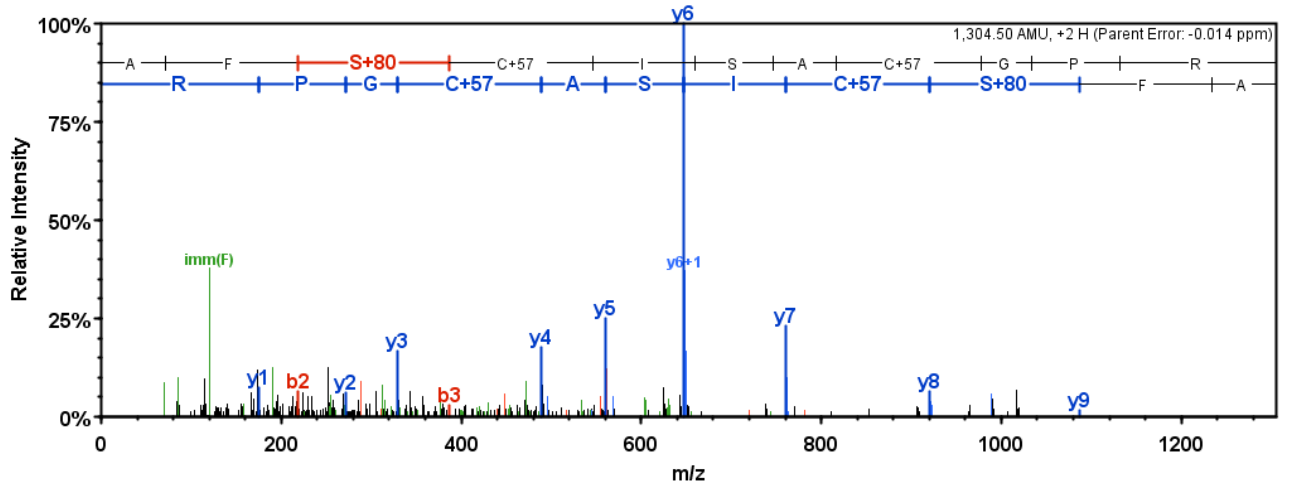
SLCamNLGSCamGPR (Tête K85)



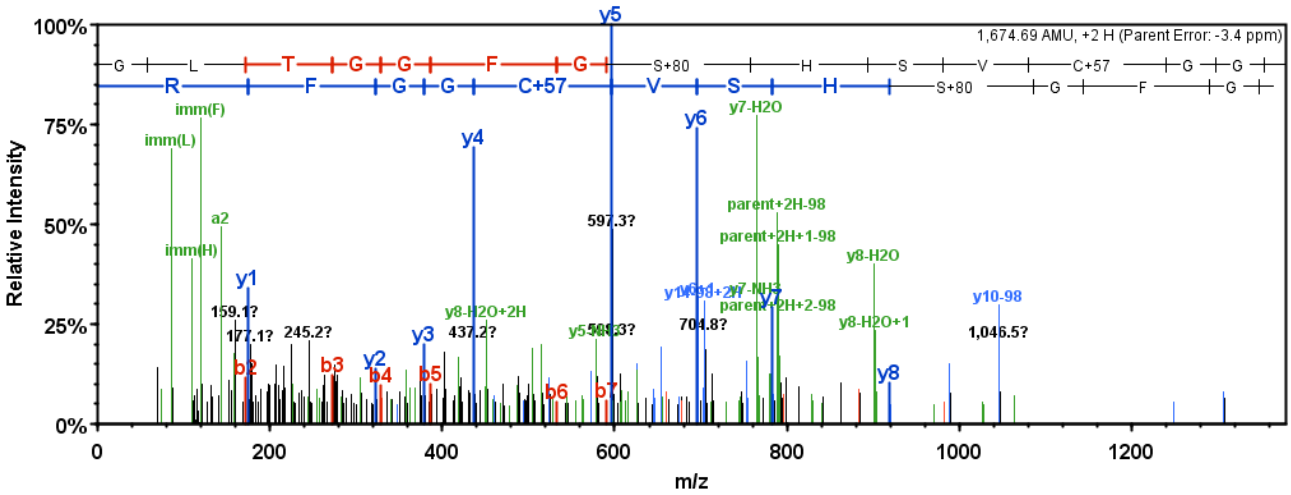
VGVCamGGSCamK (Queue K86)



TCamGSYCamGGR (tête K86)



AFSCamISACamGPR (Tête K81, K86)



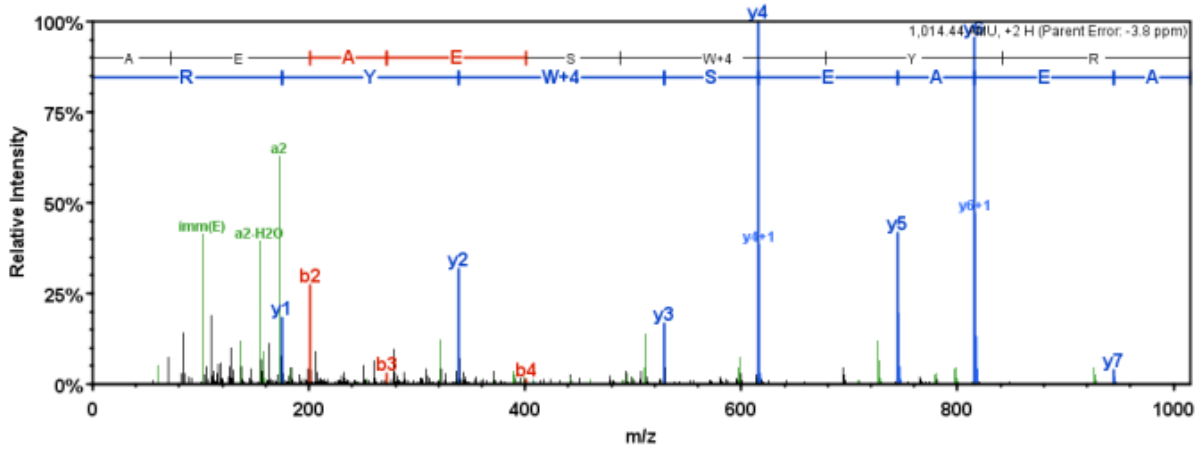
GLTGGFGSHSVCamGGFR (Tête K81, K83, K86)

Annexe 3 : Spectres de fragmentations correspondant à des modifications des résidus observées sur les kératines de type I et II

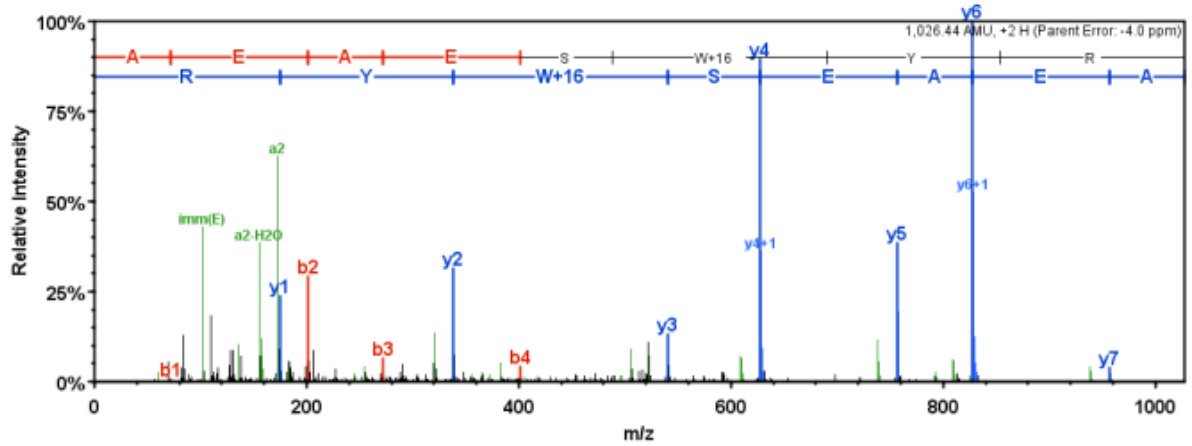
Spectres identifiés dans les données de séquençage des digests de cortex après la recherche en mode error tolerant.

K86

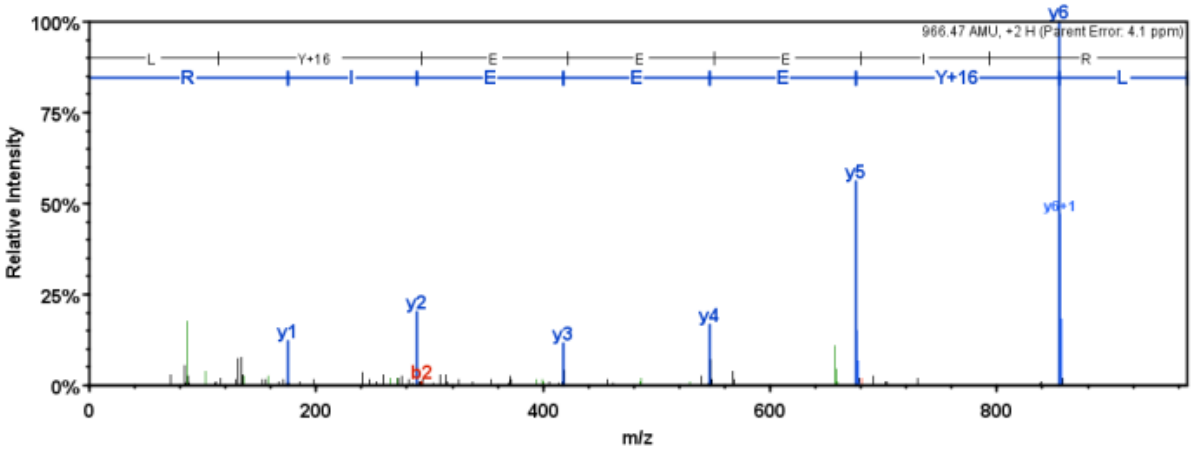
AEAESWYR kynurenine



AEAESWYR oxydation

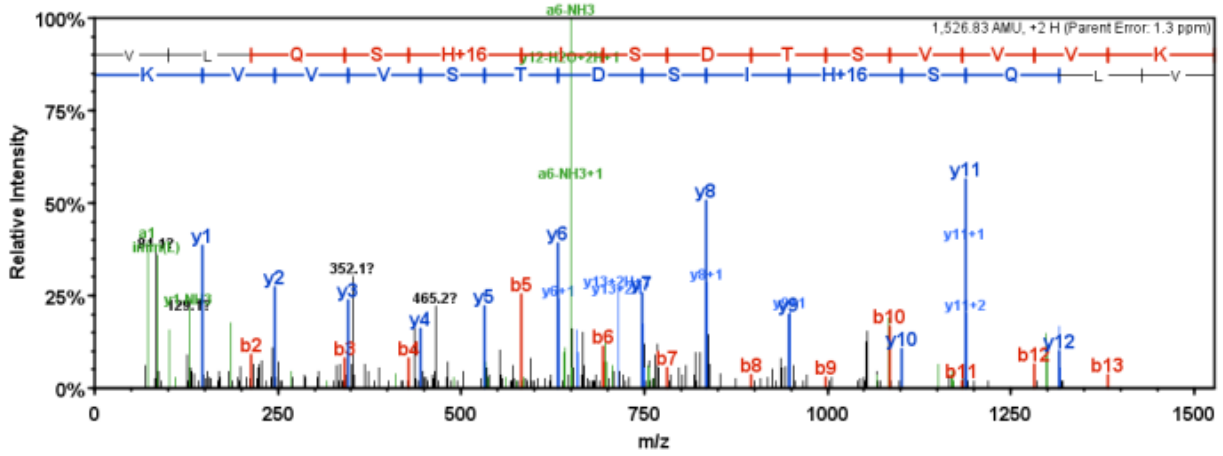


LYEEEIR hydroxylation

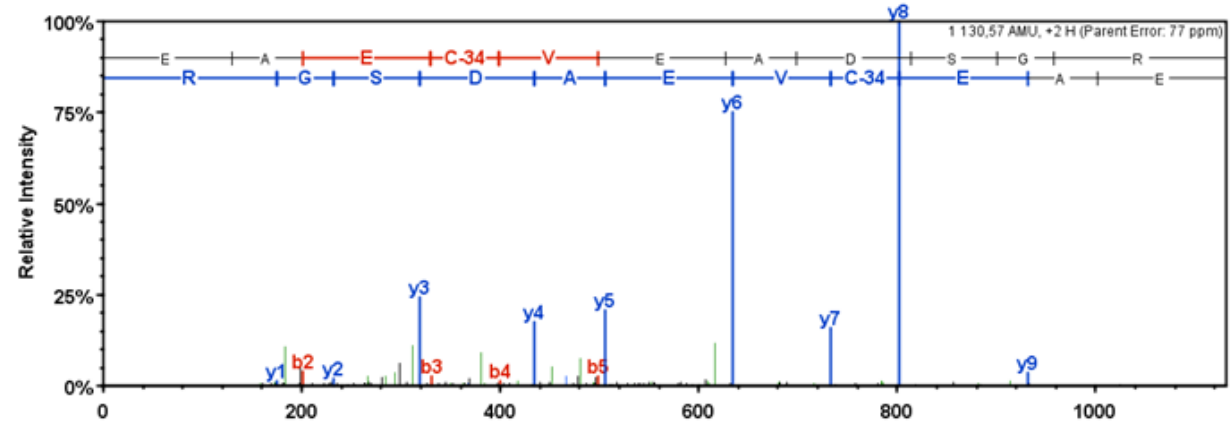


Annexe 3 : Spectres de fragmentations correspondant à des modifications des résidus observées sur les kératines de type I et II

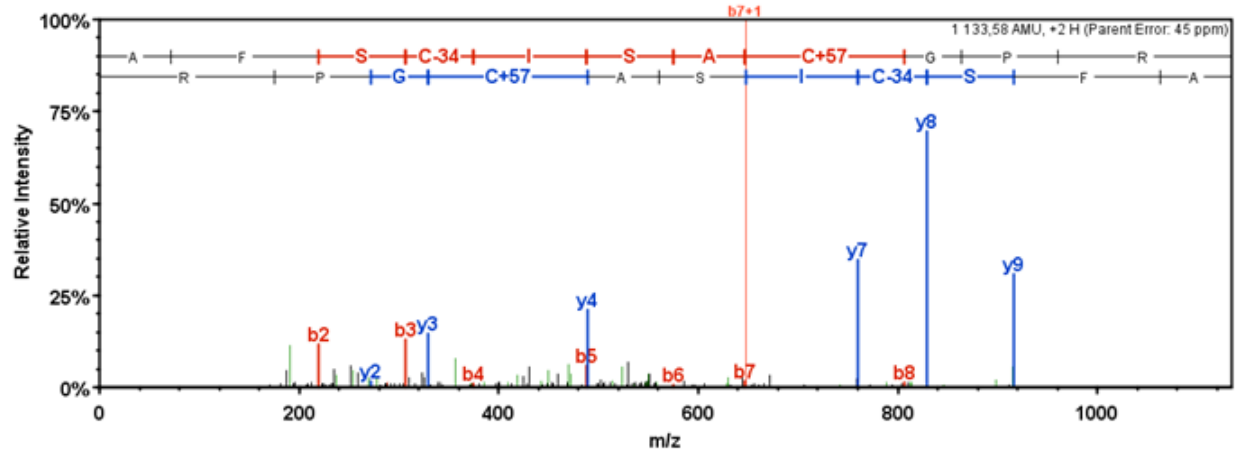
VLQSHISDTSVVVK oxydation



EAECVEADSGR dehydroalanine

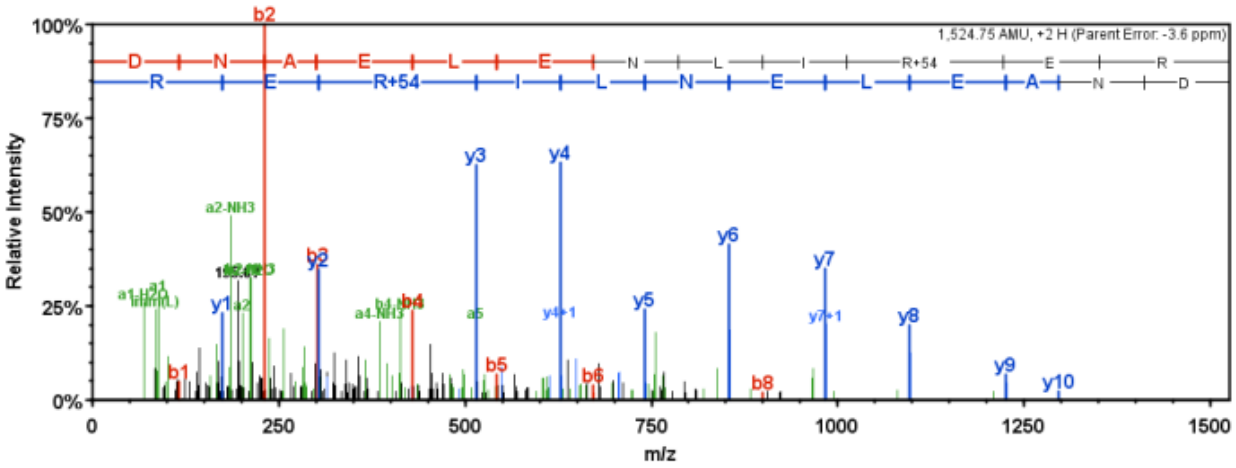


AFSCISACGPR dehydroalanine

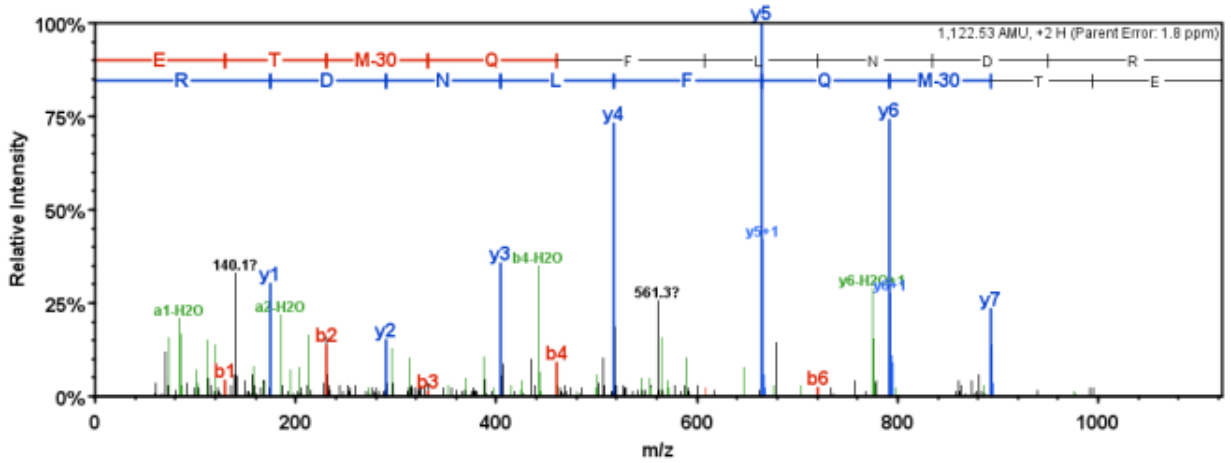


K31

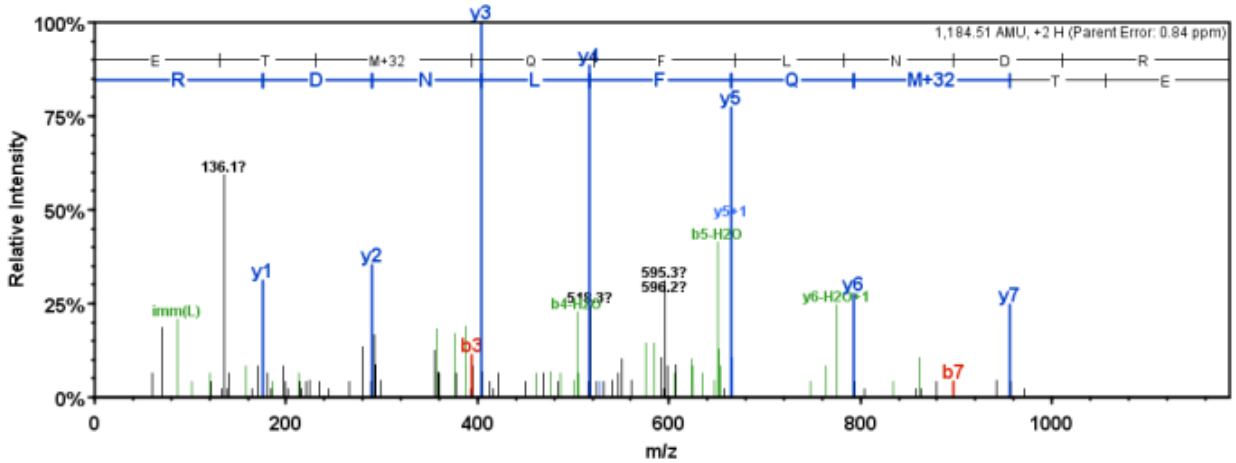
DNAELENLIRER methylglyoxal



ETMQFLNDR homoserine

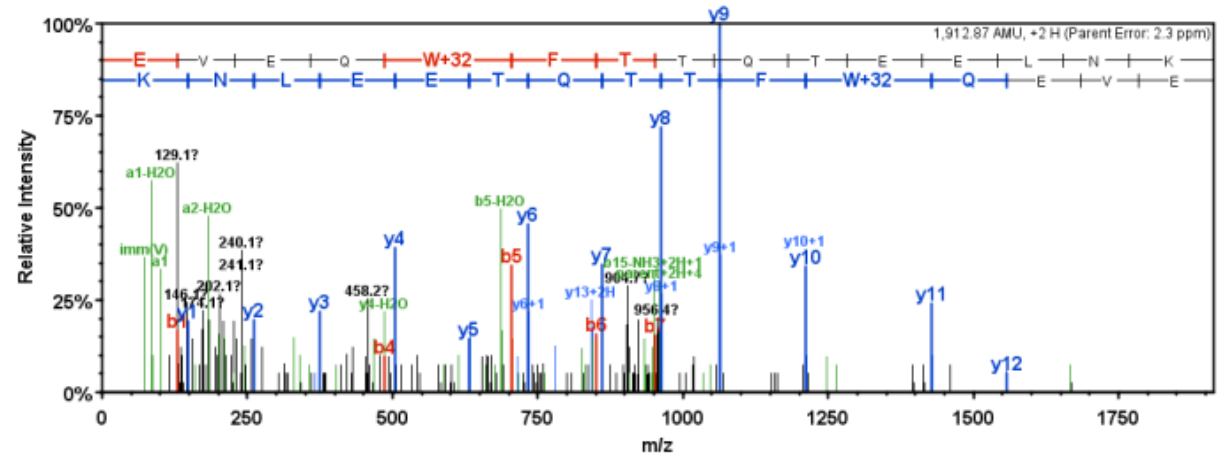


ETMQFLNDR sulphone

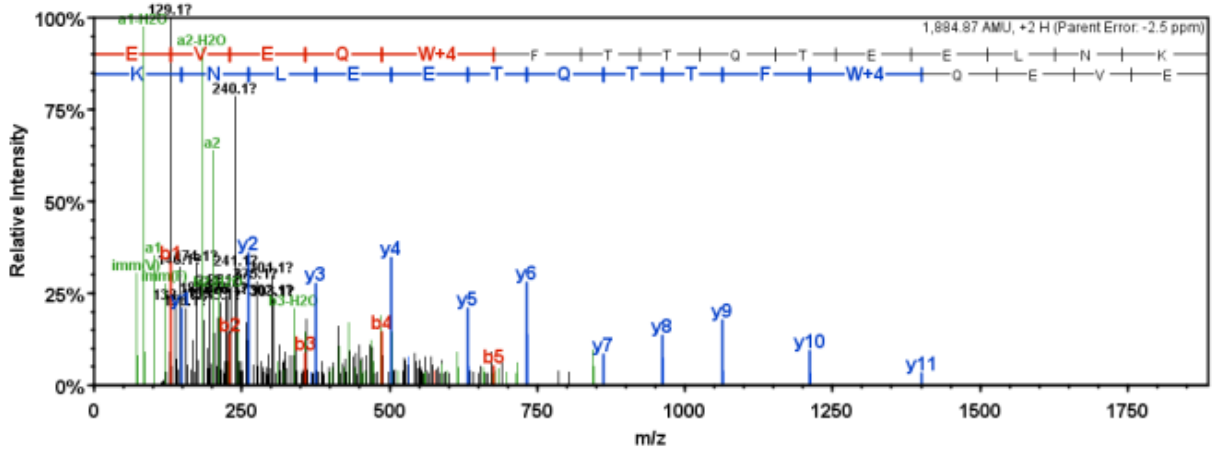


Annexe 3 : Spectres de fragmentations correspondant à des modifications des résidus observées sur les kératines de type I et II

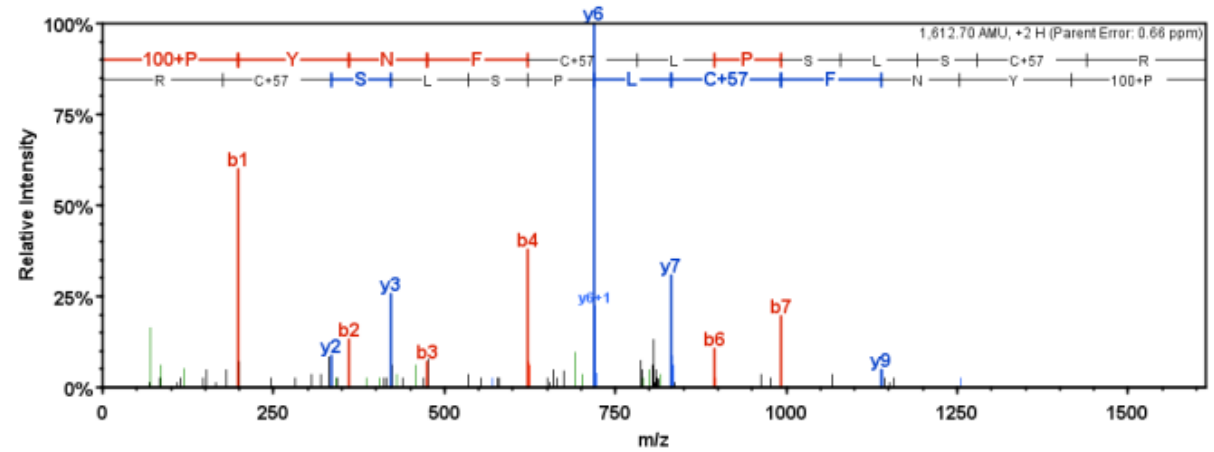
EVEQWFTTQTEELNK formylkynurenin



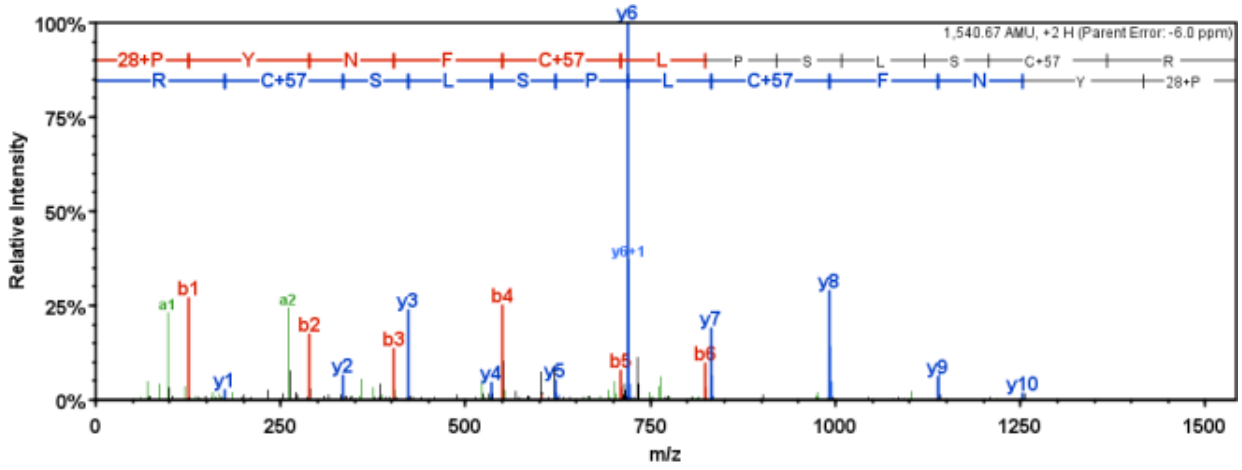
EVEQWFTTQTEELNK kynurenine



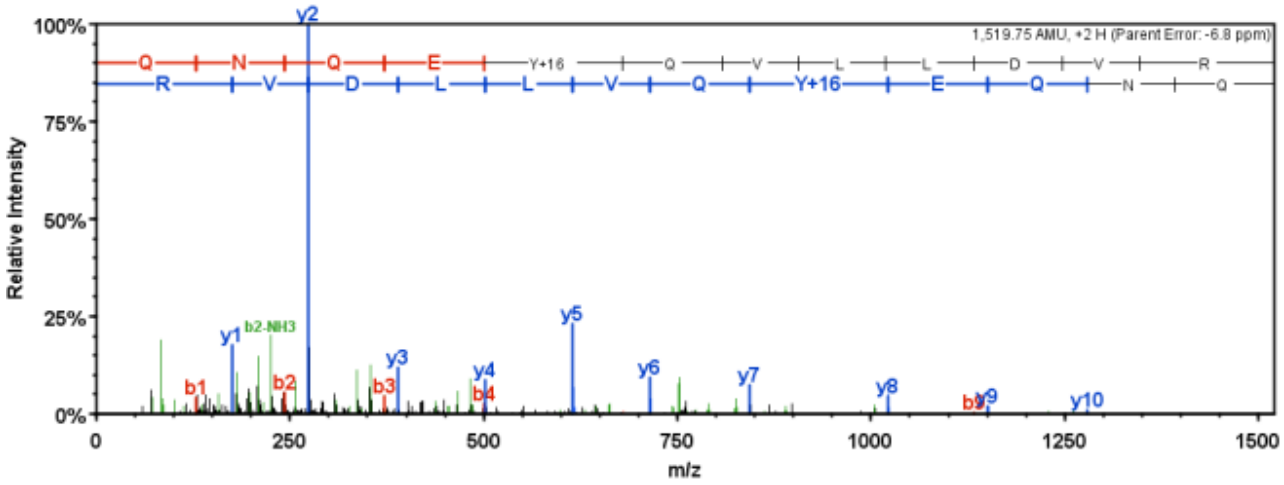
PYNFCLPSLSCR N-succinyl (Nter prot)



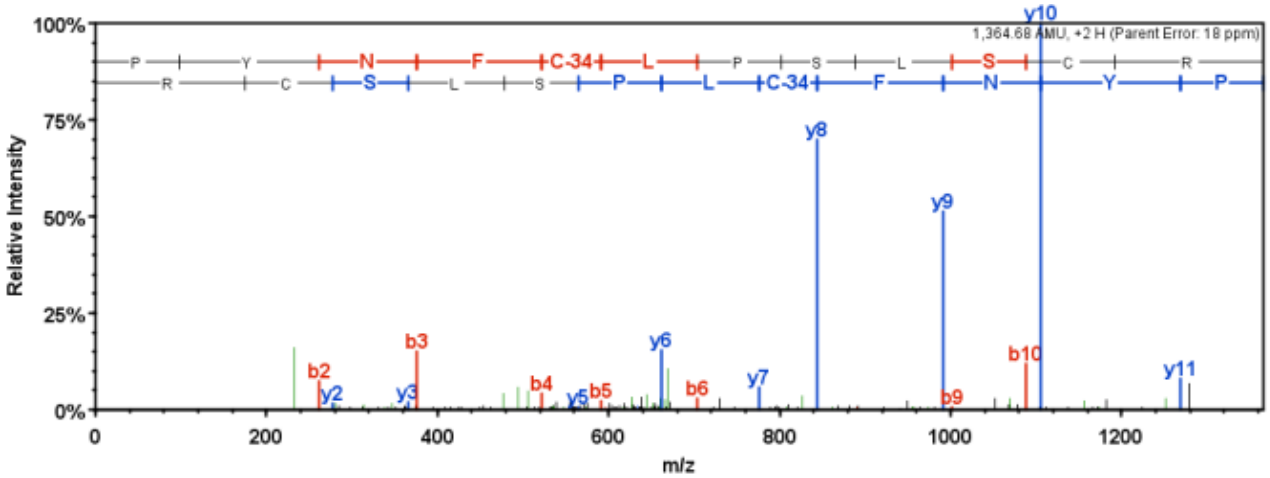
PYNFCLPSLSCR N-formyl (Nter prot)



QNQEYQVLLDVR hydroxylation



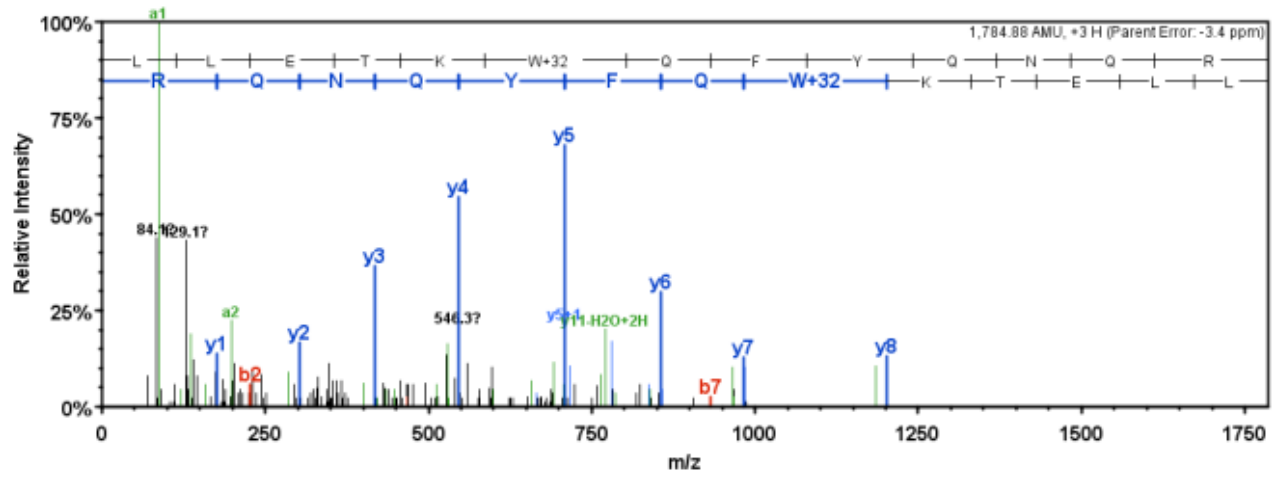
PYNFCPSLSCR dehydroalanine



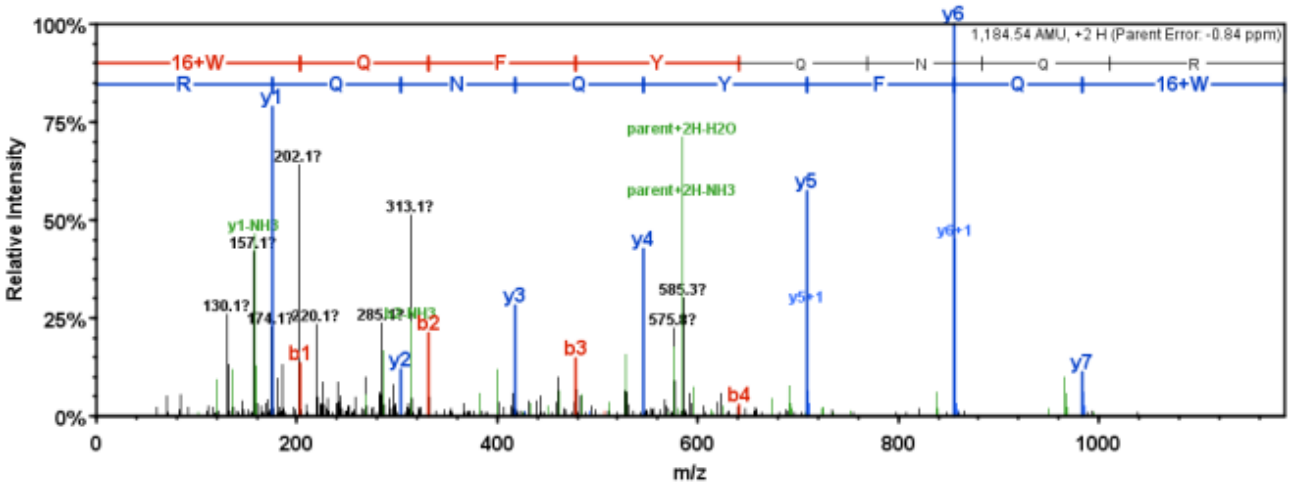
Annexe 3 : Spectres de fragmentations correspondant à des modifications des résidus observées sur les kératines de type I et II

K85

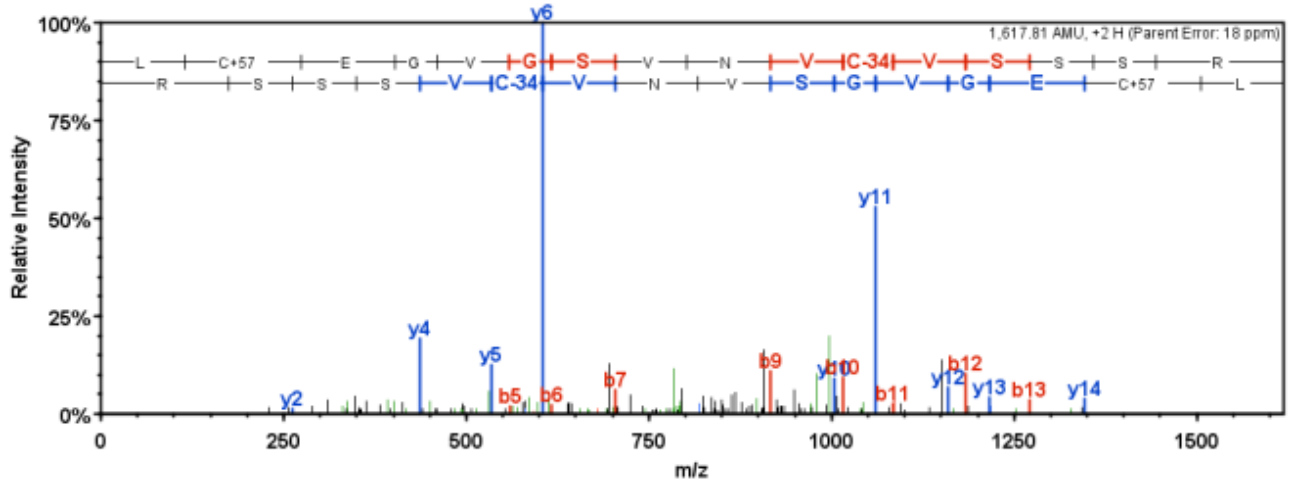
LLETKWQFYQNQR formylkynurenine



WQFYQNQR oxydation



LCEGVGSVNVCVSSSR dehydroalanine



Annexe 4 : Article accepté dans le journal Analytical Biochemistry

TITLE

Proteomic Tools for the Investigation of Human Hair Structural Proteins and Evidence of Weakness Sites on Hair Keratins Coil Segments

Nicolas R Barthélemy [1 §], Audrey Bednarczyk [1 §], Christine Schaeffer-Reiss [1], Dominique Jullien [2], Alain Van Dorsselaer [1] and Nükhet Cavusoglu [2*]

1: Université de Strasbourg, IPHC, CNRS, UMR7178, 25 rue Becquerel 67087 Strasbourg, France

2: L'Oréal Research and Innovation, Aulnay-sous-Bois, France.

* To whom correspondence should be addressed.
§ These authors contributed equally.

Running Title: Human hair proteomics

Address correspondence to: CAVUSOGLU Nükhet, L'Oréal R&I, laboratoire de protéomique et analyse des protéines, 1 rue Eugène Schueller, 93601 Aulnay-sous-Bois, France.

Tel : 01.48.68.96.42, Fax: 01.48.68.97.73. E-mail : NCAVUSOGLU@rd.loreal.com.

Abbreviations: KIF, keratin intermediate filament; KAP, keratin associated protein; 2D LC-MS/MS, offline two dimensional liquid chromatography coupled with tandem mass spectrometry; 2-DE gel MS, two dimensional gel mass spectrometry; HSP, high sulphur protein; UHSP, ultra high sulphur protein; HGTP, high glycine-tyrosine protein; PTM, post-translational modification .

ABSTRACT

Human hair is principally composed of hair keratins and keratin associated proteins (KAPs) which form a complex network that give to the hair its rigidity and mechanical properties.

However during their growth, hairs are subject to various treatments that can induce irreversible damage. For a better understanding of the human hair protein structures, proteomic MS-based strategies could assist to characterize numerous isoforms and post-translational modifications of human hair fibre proteins. However, due to their physico-chemical properties, characterization of human hair proteins using classical proteomic approaches is still a challenge.

To address this issue, we have used two complementary approaches to analyze proteins from the human hair cortex.

The MudPit approach allowed identifying all keratins and the major KAPs present in the hair as well as post-translational modifications in keratins such as cysteine tri-oxidation, lysine and histidine methylation.

Then, two dimensional gel electrophoresis coupled with mass spectrometry (2DE-gel MS) allowed us to obtain the most complete 2-DE gel pattern of human hair proteins which revealed an unexpected heterogeneity of keratin structures. Analyses of these structures by differential peptide mapping have brought evidence of cleaved species in hair keratins and suggest a preferential breaking zone in α-helical segments.

INTRODUCTION

Over the last decades most research regarding keratin proteins has generally been focused on the chemical make up of wool for textile and breeding purposes. However, during recent years, investigations on human hair proteins have been meeting with a growing interest, especially in the field of cosmetic and dermatological sciences. A better understanding of this biological material and the structural organization of human hair protein may assist in product development and other potentially disease-related issues pertaining to hair and hair follicle proteins [117, 164]

Outer hair is made of dead cells mainly composed of proteins which represent a range of 60 to95% of total chemical composition. The other constituents are lipids, water, and metals whose levels may vary depending on the hair [105].

Hair fibre can be divided into three general components [108]: (i) an outer cuticle cell layer (ii) an inner cortex and (iii) a central medulla in some cases. The cortex contains different cell types and each cell contains 500 to 800 keratin intermediate filaments (KIFs) that are the main proteins expressed in human hair. Human KIFs are composed of keratin proteins that can be sorted into two families: 1) the acidic Type I keratins (K31-K38) and 2) the neutral-basic Type II keratins (K81-K86) which contain respectively 9 and 6 members [82, 83]. The complex interaction of Type I and Type II keratins results in the formation of heteropolymers [104, 117, 133, 286].
These proteins are surrounded by an amorphous matrix of keratin-associated proteins (KAPs). The 23 KAPs families represent more than 100 KAPs with high sequence homologies in each family. Based on amino acid composition, KAPs are classified into three groups: 1) the High Sulphur Proteins (HSPs) (< 30% cysteine content) including 9 KAP families, 2) the Ultra High Sulphur Proteins (UHSPs) (> 30% cysteine content) consisting of 4 KAP families and 3) the High Glycine-Tyrosine Proteins (HGTPs) including 7 KAP families [73].
These components form a strong compact and complex network of proteins bound together by intra and intermolecular interactions, such as disulfide linkages, hydrogen bonds, electrostatic salt bonds and amide bonds [149, 301]. This variety of internal bonds are essential to impart rigidity to the structure of the hair and make it resistant to environmental factors such as UV, pollutants and weather, or chemical treatments [105].
Although the expression of Type I and Type II keratins and KAPs in the hair follicle is well established, their expression in mature hair, their post-translational modifications and the way they can be altered in the hair shaft remain elusive. Their role in the shape and quality of hair, as well as the mechanisms involved, is not well identified [73, 149].
Proteomics could assist to characterize numerous isoforms and post-translational modifications (PTMs) of human hair fibre proteins. Some challenges do remain as regards hair protein characterization using the general MS based strategies. Most of these challenges are due to the insolubility of hair proteins and the difficulty of extracting and solubilizing the proteins in solvents that are compatible with gel electrophoresis or liquid chromatography. Moreover, most of hair proteins are keratins, and the issue of separating and detecting minor proteins, for instance the KAPs, is still critical. Regarding this dynamic range issue, the high sequence homologies of human hair keratins (70-90%) and KAPs raises a major additional challenge which requires a high level of expertise to interpret generated mass data and identify the proteins [3, 84, 86, 92]. The small number of unique peptides called proteotypic peptides does not allow us to determine which particular members of these families are present in the sample.
So far, only a few proteomic approaches have been described in the way of analyzing human hair proteins. The first studies on keratin proteins were based on electrophoresis separation and were performed on wool keratins [95]. Similar studies were reported on human hair keratins, which were identified by two dimensional gel electrophoresis (2-DE gel) and Western Blot analysis using antibodies specific to each family member [3, 82, 83]. The 2-DE gel of human hair keratins showed a pattern similar to the pattern of wool keratins with a long train of proteins in the 62 kDa area between pI 5-7 which correspond to Type II keratins and a cluster of proteins at a lower isoelectric point (pI) and molecular weight (MW) which correspond to Type I KIFs (pI 4-5, MW 45 kDa). Beside this typical 2-DE gel pattern, few other spots appeared in the lower MW part of the gel.
However, most of the 2D gel studies on the human hair shaft have only demonstrated the strong expression of some Type I (K31, K33a, K33b, K34 and K35) and Type II keratins (K81, K82, K83, K85 and K86) without mass spectrometry identifications [94]. As an alternative method for analyzing human hair proteins, Lee *et al*. suggested a multi-dimensional protein identification technology (MudPit). This approach involved separating a peptide complex mixture resulting from the digestion of total hair protein extract by Strong Cation Exchange (SCX) chromatography followed by reversed phase nanoliquid chromatography coupled to tandem mass spectrometry (nanoLC-MS/MS) [302, 303]. This approach, which eliminated solubility problems and signal suppression due to the high number of generated peptides, allowed the authors to identify hair keratins, KAPs and many proteins involved in the formation and structural organization of the hair shaft. In addition, Lee *et al*. showed evidence of post-translational methylation, dimethylation and trimethylation on hair proteins [89].
In the present study, we investigate an approach based on a combination of 2D-gel and 2D-LC followed by nanoLC-MS/MS in order to improve identification of human hair proteins.

MATERIALS AND METHODS

1 Protein extraction

All hair samples used in this study are constituted in the blending of untreated scalp hair from three individuals. Different extraction procedures were applied in accordance with the analytical method used. Before extraction, hair samples were delipidated by soaking fibers in ethanol, then in cyclohexane.

1.1 Protein extraction for 2D-LC

Hair samples were extracted following the experimental procedure described by Lee *et al*. [89]. Short time after delipidation, proteins were extracted in a solution containing 2% sodium dodecylsulfate (SDS), 50 mM sodium phosphate (pH 7.8), 20 mM

DL-dithiothreitol (DTT) and incubated overnight at 65°C. After centrifugation, the insoluble material was submitted to repeated extraction procedure (6 times) as described, giving finally two samples, an insoluble material and a soluble material.

1.2 Protein extraction for 2D-gels

Proteins were extracted in a solution containing 7 M Urea, 2 M thiourea, 50 mM Tris-HCl, 50 mM DTT, 0.1% Triton X100 for 18 hours at 37°C. The protein extract was collected *via* filtration and then alkylated with a solution of 1 M iodoacetamide and 3 M Tris-HCl at pH 8.4 for 10 minutes in the dark at room temperature. The solution was dialyzed with 3,500 MWCO dialysis cassettes (Pierce, Rockford, USA) against water over a period of 48 hours. The solution was then lyophilized and the freeze-dried sample was stored in a -80°C freezer.

2 Separation and MS/MS analysis

2.1 Protein digestion and peptide separation by SCX chromatography

Both soluble and insoluble hair extracts were incubated during 1 hour at 57°C in 2% SDS, 20 mM DTT and 50 mM disodium hydrogen phosphate (Na_2HPO_4) to reduce all cystine groups. Reduced cysteines were then alkylated by adding 40 mM iodoacetamide during 1 hour in the dark. Proteins were precipitated by adding 2.5 volumes of ethanol and rinsed twice with 70% ethanol in order to eliminate SDS, DTT, Na_2HPO_4 and iodoacetamide excess. Finally, proteins were resuspended in 100 mM ammonium bicarbonate (NH_4HCO_3) and 2 M urea and digested overnight at 37°C by adding modified porcine trypsin in a ratio 1 part of enzyme's weight for 20 parts of protein's weight.

Both tryptic peptide mixtures were fractionated with a Waters 625 LC System (Waters, Milford, MA, USA) using a PolySULFOETHYLA™ column (100 mm x 2.1 mm, 5 µm i.d., 300 Å pore size) (PolyLC$_{INC}$, Columbia, MD) working at a flow rate of 200 µL/min. After samples were loaded, a 15 minutes isocratic run with 100% solvent A (5 mM KH_2PO_4, 25% acetonitrile, pH 3) was performed. Peptides were eluted using a two step gradient from: (1) 0% to 25% of solvent B (5 mM KH_2PO_4, 350 mM KCl, 25% acetonitrile, pH 3) in 30 minutes to (2) 25% to 100% of solvent B in 20 minutes. Two minutes interval fractions were collected, concentrated by vacuum centrifugation, and desalted using ZipTipC18 Pipette Tips (Millipore, Bedford, MA, USA). 35 fractions of each soluble and insoluble tryptic digestion material were collected for a second dimension of separation on reversed phase nanoLC coupled with mass spectrometry.

2.2 Protein separation by 2-DE gel and in gel tryptic digestion

400 µg of protein sample was suspended in the rehydration buffer consisting of 7 M urea, 2 M thiourea, 2% CHAPS, 0.5% ampholytes pH 3-11 and DeStreak rehydration solution (GE Healthcare, Uppsala, Sweden) and incubated with IPG strips pH 3-11 from GE Healthcare in a PROTEAN isoelectric focusing cell (BioRad) for 16 hours. Isoelectric focusing was performed by a stepwise voltage increase until reaching 14400 VH. Before running the second dimension, the strips were equilibrated in a 6 M urea, 2% SDS and 0.5 M DTT solution and loaded on 7 cm NuPAGE 10% Bis-Tris gels (Invitrogen). Electrophoresis was carried out at 200 V for 40 minutes on the Novex mini cell system from Invitrogen (Cergy Pontoise, France). All the gels were stained by Coomassie blue (SimplyBlue, Invitrogen).

Two-dimensional gel spots of interest were excised and digested. In-gel digestion was performed with an automated protein digestion system, MassPREP station (Waters, Milford Massachusetts, USA). The gel spots were washed twice with 100 µL of 25 mM ammonium bicarbonate (NH_4HCO_3) acetonitrile (ratio 1/1). The cystine groups were first reduced by 50 µl of 10 mM DTT at 57°C for one hour and then alkylated using 50 µl of 55 mM iodoacetamide at room temperature for 20 minutes. After dehydration of the gel bands with acetonitrile, proteins were digested overnight in gel with 15 µl of 12.5 ng/µl modified porcine trypsin (Promega, Madison, WI, USA) in 25 mM ammonium bicarbonate (NH_4HCO_3) at room temperature. The generated peptides were extracted with 60% acetonitrile in 5% formic acid (HCOOH) followed by removing excess acetonitrile.

3 NanoLC-Chip-MS/MS analysis of SCX fractions and 2D gel spots

The tryptic digests from each fraction from SCX chromatography or 2D spots were analyzed by nanoLC-MS/MS using an Agilent 1100 series HPLC-Chip/MS system (Agilent Technologies, Palo Alto, USA) coupled to an HCTultra ion trap (Bruker Daltonics, Bremen, Germany). For all experiments, water was purified using a Direct-Q™ from Millipore and acetonitrile HPLC grade was purchased from Carlo Erba Reactifs-SDS (Val de Reuil, France). The solvent system consisted of 2% v/v acetonitrile and 0.1% v/v formic acid in water (solvent A), and 2% v/v water and 0.1% v/v formic acid in acetonitrile (solvent B). Peptides were separated with a reversed phase C18 column (Zorbax 300SB-C18, 75 µm x 150 mm, 5 µm i.d.) using an acetonitrile gradient from 8 to 40% solvent B in 30 minutes at a flow rate of 300 nL/min. Mass spectrometer was operated in positive ion mode and the voltage applied to the capillary cap was optimized to -1850 V. Tandem MS experiments were performed by CID and the system was operated with automatic switching between MS and MS/MS modes. The three most abundant peptides

and preferentially doubly charged ions were selected on each MS spectrum for further isolation and fragmentation. MS/MS scanning was performed with the ultrascan resolution mode at a scan rate of 26,000 m/z per second. A total of 6 scans were averaged to obtain a MS/MS spectrum. The complete system was fully controlled by ChemStation (Agilent Technologies) and EsquireControl (Bruker Daltonics) softwares.

4 Protein identification and validation

4.1 2D-LC-MS/MS data

Mass data collected during nanoLC-MS/MS analysis were processed, converted into *.mgf files and analyzed using the MASCOT 2.2.0 algorithm (Matrix Science, London, UK). Spectra were searched with a mass tolerance of 0.3 Da for MS and MS/MS data, allowing a maximum of one missed cleavage, tryptic peptides with a proline as C-terminal amino acid of the cleavage site and with carbamidomethylation of cysteine, N-acetylation at the N-terminus of the protein and oxidation of methionines specified as variable modification. Spectra were first searched in a target-decoy version of UniprotKB database (v 14.8, February 17, 2009, 165400 entries). All files corresponding to one extract were combined using homemade software (concatene.exe) and were searched in the same target-decoy UniprotKB database described in the gel 2D protein identification part, using the Mascot search algorithm. Mass tolerance for the MS and MS/MS ions was also fixed at 0.3 Da. The database search parameters were the same as described before and the searching was performed in two steps. During the first step, protein identifications were considered correct when one peptide was detected (less than 10 points below Mascot's threshold ion score of identity at 95% confidence level and ion score more than 15, 61 proteins, 0 reverse for total extract and 45 proteins, 0 reverse for cortical extract). During the second step, spectra that did not satisfy these criteria were exported using Scaffold software 2.2.0 (Proteome Software, Portland, USA) and searched against the restricted decoy database with the same variable modifications as before. Protein identification were validated when one peptide was detected (less than 22 points below Mascot's threshold ion score and ion score more than 15, 14 proteins, 0 reverse for total extract, 25 proteins, 0 reverse for cortical extract).

4.2 2-DE gel LC-MS/MS data

Protein identifications were considered valid when one peptide with high quality MS/MS spectra (less than 8.1 points below Mascot's threshold score of identity at 95% confidence level and ion score more than 20) was detected (37 proteins, 0 reverse). All spectra that did not satisfy these criteria were exported using Scaffold and searched in the same target-decoy database restricted to keratin and intermediate filament components (606 entries) with the same criteria described below in addition to the following variable modifications: methylation (H, K), dimethylation (K), trimethylation (K) and oxidation into sulfonic acid (C). Protein identifications were validated with the same previously set threshold (23 proteins, 0 reverse).

5 2-DE gel LC-MS peptide mapping and relative abundance measurements

12 spots of interest were reanalyzed in nanoLC-MS and MS/MS for peptide mapping on a nanoACQUITY UPLC coupled to a SYNAPT hybrid quadrupole orthogonal acceleration time-of-flight tandem mass spectrometer (Waters, Milford, MA), equipped with nano-spray ion source and a lock mass system in the positive ion mode. The digests were trapped on a 20 x 0.18 mm, 5 µm Symmetry C18 precolumn (Waters, Milford, MA), and the peptides were separated on a ACQUITY UPLC® BEH130 C18 column (Waters, Milford, MA), 75 µm x 200 mm, 1.7 µm particle size. The solvent system consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile. Trapping was performed during 3 min at 5 µL/min with 99% A and 1% B. Elution was performed at a flow rate of 400 nL/min, using 6-40% B over 45 min at 45°C followed by 65% B over 5 min. The capillary voltage was set at 3.5 kV and the cone voltage at 35 V. Mass calibration of the TOF was achieved using phosphoric acid (H_3PO_4) on the [50-2000] m/z range. Online correction of this calibration was performed with a lock-mass. For LC-MS experiments, a scan was acquired each 600 ms. Peak capacity for the LC system was greater than 200. For tandem MS experiments, the system was operated with automatic switching between MS and MS/MS modes (MS 0.5 s/scan on m/z range [250;1500] and MS/MS 0.7 s/scan on m/z range [50;2000]). The two most abundant ions (intensity threshold 40 counts/s), preferably doubly and triply charged ions, were selected on each MS spectrum for further isolation and CID fragmentation with two energies set using collision energy profile, then were excluded for 20 seconds. Peptide identifications obtained after Mascot search on MS/MS data as previously described, were used to extract ion intensity of corresponding peptides on LC-MS runs.

RESULTS and DISCUSSION

1. Study of digested human hair protein extracts by offline two dimensional chromatography (SCX/RP/MS/MS)

The MudPit approach was applied to proteins extracted using the procedure described by Lee et al. [89]. Strategy showed a specific extraction of the cortex leaving the cuticle intact as shown in the scanning electron microscopy view (Figure 1). As expected, two fractions were obtained, a soluble and an insoluble one respectively equivalent to cortical and cuticular extracts. Both were submitted to tryptic digestion and were run for the first dimension of separation through a SCX column followed by the second dimension on reversed phase column coupled to tandem mass spectrometry analysis.
The fractions were analyzed by nanoLC-MS/MS leading to the identification of fifty six proteins including 34 keratins and KAPs proteins listed in Table 1. The major proteins were Type I and Type II keratins with 13 of 15 hair keratin from intermediate filaments identified: 5 Type II keratins (K81, K82, K83, K85 and K86) and 8 Type I keratins (K31, K32, K33a, K33b, K34, K35, K36 and K37). The highest sequence coverage was obtained for K81, K85, K86, K31, K33a and K33b, suggesting that these keratins are the most abundant in hair. Regarding KAPs, we identified 13 of them with specific peptides, KAP 24.1, KAP 11.1, KAP 3.1, 3.2 and 3.3, KAP 13.2, KAP 19.5, KAP 9.6 and 9.7, KAP 4.3, 4.4, 4.7 and KAP 4.9 and other 10 with peptides common to a KAP family (KAP 4.6/4.12 with 2 peptides, KAP 9.2/9.9 with 2 peptides, KAP 10.1/10.3/10.7 with 2 peptides, KAP 2.1 or 2.3 or 2.4 with 4 peptides).
LC-2D-MS approach between soluble and insoluble extract gave approximately the same protein identification results . Nevertheless, we can suppose that this protocol led to the extraction of some cuticle proteins as suggested by the presence of cuticular K32, K82, KAPs 10 and 24.1 [78, 140], while all other identified KAPs have a cortical origin. We also note the presence of more peptides from these proteins in the insoluble extract.
In addition, we identified 22 minor proteins which could be classified in different groups (Table 2): desmosomal proteins (desmoglein-4, plakoglobin), which play a crucial role in keratinocyte adhesion to neighbouring cells, structural proteins such as histones H2A, H2B and H4, calcium-binding proteins such as S100-A3 and calmodulin-like protein 3, which bind calcium ions which are needed for the function of desmosomal proteins, stabilization proteins (Heat shock cognate 71 kDa), protein involved in signalling (14-3-3 protein), proteolytic enzymes (lysozyme g-like, bleomicin hydrolase), an antimicrobial peptide (dermicin), which is known to protect epithelial surface, selenium-binding protein involved in intra-Golgi protein transport, sialydase-2 which hydrolyzes sialylated compounds, lyzosome-associated membrane glycoproteins, protein present in epithelial cells (Serpin B5), Nesprin-1 involved in the maintenance of nuclear organization and structural integrity. These proteins were identified with little peptides (1 to 3), which is not surprising. The latter are probably non-residual fragments reflecting past cellular activity. Finally, the identification of TGase-3 with one peptide could be linked to some interprotein cross-links described in the hair shaft [154, 155, 304, 305].
Results from mass spectrometry showed that all localized PTMs were identified from keratin peptides (Table 3). Oxidative alteration of cysteine to sulfonic acid (+ 47.985) was unambiguously identified whereas methylation type modification, previously described on hair keratin [89], was difficult to demonstrate unequivocally.
The target decoy strategy employed to assign PTMs did not exempt us from careful examination of MS/MS spectra in order to demonstrate the validity of our interpretation. For every methylated peptide, additional steps of spectra analysis were performed, so as to consider many ambiguities from isoforms and isobaric amino acid leading to close molecular weight for ion parent.
In conclusion, the MudPit approach allows three different PTMs to be observed: oxidation of cysteine to sulfonic acid, dimethylation of lysine and methylation of histidine. We identified 15 peptides with sulfonic acids, 9 peptides with dimethylated lysines, 2 peptides with trimethylated lysine and 2 peptides with methylated histidines. However, all PTMs observed could not be assigned to a given protein because several modified peptides were shared with some keratins. In some cases, we identified different modifications of the same peptide sequence such as for example FLEQQNKLLETK and GLTGGFGSHSVCGGFR (Table 3). Modifications were mostly observed on Type II basic keratins with more oxidized cysteine than on Type I acidic keratins. In fact, most non-PTM oxidation of cysteine is known to occur after hair chemical treatments [164, 166]. Considering that the study was performed on untreated hairs, oxidation may only result from UV exposure which is known to induce an oxidative stress.

2 Human hair proteins identified by two dimensional gel electrophoresis (2-DE) followed by nanoLC-MS/MS analysis

Despite its use as a versatile method, many challenges had to be overcome to get 2-DE gels of acceptable quality from samples such as hair. Difficulty of extraction of structural proteins, insolubility, large dynamic ranges, extensive cross-links were features that had to be considered. Usually, migration of human hair proteins is characterized by a pattern of hair keratins described by Langbein et al. [82, 83] showing a long train of proteins in the 62 kDa area between pI 4-7 which correspond to the Type II keratins and a cluster of proteins at low pI which correspond to Type I keratins. The KAPs identification with 2-DE gel remains more difficult. It explains why as far we know, no clear 2-DE gel map profile of human KAPs has been done yet.
By improving migration parameters and increasing the loading quantity in order to potentially detect other proteins identified with MudPit approach, particularly KAPs families, we have been able to obtain a 2-DE gel pattern with a new

fingerprint of a rich group of spots (Figure 2). These spots were not detected for sample loadings below 50 µg (not shown). In this area, some background was observed probably due to the overloaded quantity of protein material. We have chosen a sampling of 59 spots (from 27 to 85 in Figure 2) principally in the 28 to 42 kDa zone and 26 spots (from 1 to 26) in the zone below 28 kDa presumably where the KAPs might be localized. Twenty six proteins were identified by nanoLC-MS/MS and are listed in Table 4.

The results reveal that the major identified proteins were Type II and Type I keratins. We identified 4 out of the 6 Type II keratins (K81, K83, K85 and K86) and 7 out of the 9 Type I keratins (K31, K32, K33a, K33b, K34, K35 and K36). The five most prominent proteins identified were K81, K85 and K86 for Type II keratins and K31 and K33a for Type I keratins, respectively. Because they were present also below their expected molecular weight, we interpreted their presence as the result of degradation that could occur inside the fibre or during the extraction procedure. To confirm this interpretation, sequence coverages obtained in each spot were investigated. Interestingly, the coverage profile of keratin peptides identified in several spots seemed to be a function of the spot location. Consequently, spots were grouped into 6 families according to sequence coverage. These families show a distribution pattern in some region of the gel as illustrated in Figure 3a. Nevertheless, the presence of peptides from the N-terminal and C-terminal part of protein sequence in the majority of spots tends to refute the hypothesis of cleaved species and needs to be investigated.

Regarding Type I keratins (Figure 3c), we found the majority of peptides in the low acidic region investigated below the keratins Type I cluster (spots 13-14 and 23-26), which could also indicate cleaved species. In spite of the acid pI of these proteins, several Type I keratins peptides were identified in the basic region previously described.

Regarding KAPs (Figure 3b), only 6 could be identified (1.5/ 3.1/ 3.2/ 3.3/ 11.1/ 13.1) with one or two specific peptides and 3 KAPs from the families 2, 4 and 9, with one peptide common to one family. Although less covered and less abundant than MudPit results, these identifications show the possibility of identifying these proteins with 2D-gel based strategy. Nevertheless, an unexpected result which can be emphasized was the identification of KAPs in the spots localized on the highest area of the gel where Type II keratins were identified. In fact, only KAP 3.1, 1.5 and 4 families were found in regions which are consistent with their molecular respectively as spots 10-15, 16 and 6.

The absence of focalized spots for these proteins could be a result of several factors. First, staining with current techniques such as silver stain or Coomassie blue seems to be unfavourable to the detection of KAPs, presumably because of the unusual amino acid composition of these proteins. The best conditions to reveal HSPs and UHSPs is the labelling of the high quantity of cysteine in these protein sequences with ^{14}C radioactive labelled iodoacetamide or iodoacetic acid [108] but it requires particular equipment and precaution. Second, several families such as KAP 4 and 9 are members of multigenic families. The slight heterogeneity induced by the dozen of proteins with low differences in sequence within family could also lead to dispersion of protein signal comparable to clusters for Type I or Type II keratins on 2-DE gel. Third, the amino acid composition of several families could explain their extreme pI values at which they are difficult to resolve by electrofocusing. Furthermore, the alkylating reagent used to avoid reoxidation of the high number of cysteine could induce strong variation of isoelectric point. As an example of this issue, we have computed pI of each expected keratin and KAP protein with the common alkylating reagent used in this study, iodoacetamide. To simulate the acid or basic nature of the alkylation group, we have substituted for the cysteines with amino acids of similar properties such as glutamic acid for iodoacetic acid and glutamine for iodoacetamide reagent. For example, KAP 4.2 has a theoretical pI of 8.30 without alkylation which increases to 12.18 with glutamine substitution and decreases to 3.90 with acid glutamic substitution. Theoretical 2-D maps of keratins and KAPs obtained without or with alkylation are shown in Figure 4 as an image of this concept. Fourth, a partially incomplete alkylation of about 40 to 70 cysteines in protein sequence could induce charge heterogeneity of lateral chains, thus increasing the dispersion of these proteins in the gel. Finally, alterations such as deamidation, trioxidation of cysteine to cysteic acid or others could also increase this heterogeneity.

In these conditions, electrophoresis at a constant pH, such as previously employed before introducing immobilized pH gradient in first dimension and fluorimetric detection, seemed to be more adapted to carboxymethylated HSPs and UHSPs migration and detection. This strategy leaded to large spots corresponding to the expected proteins [59, 64, 108].

The same PTMs search that was used for the MudPit approach was performed on MS/MS data of spot analysis. Mascot search showed that only Type II keratins K81, K83, K85 and K86 were modified with the identification of 7 PTMs in the gel spots of the 28-42 kDa region and 2 PTMs in the gel spots of the 14-28 kDa region. The modified peptides which were identified are listed in Table 5. Five of them could be unambiguously assigned to the K85 keratin because of their specificity; the four others could not be ascribed to a specific keratin. The fact that no PTMs were identified on Type I keratins could be explained by the fact that these spots are less widespread within the gel compared to Type II keratins spots, even though PTM identification is still a challenging task in proteomic studies. Therefore, it might be possible that more PTMs exist within these keratins, particularly with acidic pI. Even if we identified some PTMs on Type II keratins, these results did not explain the high mass and pI shifts of these proteins on the 2-DE gel.

The biological significance of methylations detected on keratins is difficult to understand. No N-methyltransferases were found in all our analyses and activity of this enzyme is rather described on histones and located into nucleus [306]. We could make hypothesis of minor non specific activity of nucleus N-methyltransferases on major hair keratins during the last stage of the cortical cell death to explain this unexpected protein alteration.

Regarding the pI and MW shifts, KAPs located in the same spots as keratins could suggest that intermolecular bonds exist between keratins and KAPs, thereby generating these shifts. Indeed, KAPs are basic proteins with low molecular weight that could also generate the pI shifts. For example, the KAP 2.4 which has a molecular weight of 13.5 kDa with a pI ~8.3 was identified in many spots in the 30 kDa regions. If this KAP binds to a keratin, a higher pI shift in the basic region could occur for the keratin fragment accompanied by a MW shift in higher masses for the KAP. Although keratins and KAPs are known to self associate through disulfide linkages, it cannot be attributed to reducible disulfide bonds as these extracts were treated with DTT and iodoacetamide before migration on 2-DE gel in order to cleave disulfide linkages and prevent their formation. Other linkages have been described on hair and wool protein such as lanthionine or lysinoalanine crosslinks generated after alkaline treatment [166] or dityrosine [165]. Transglutaminase activity is also described in the hair and could induce links between lysine and glutamine of hair proteins [154]. These unreducible crosslinks could generate the abnormal features of spots and unexpected identification of this gel. Nevertheless, evidence of this potential crosslink has not been found by these MS/MS experiments only. Therefore, further investigations will have to be performed in order to understand these unexpected attributions of this gel.

3 Peptide mapping and ion abundance analysis on 2-DE gel spots

To obtain more comprehensive information about identification and sequence coverage obtained on spots from the 2-DE gel associated to tandem mass spectrometry, 12 digested spots were reanalyzed with a nanoLC-QTOF system. This coupling takes high advantage of mass accuracy compared to the ion trap system. After MS/MS runs, a second LC-MS analysis was performed to obtain a peptide mapping of a previously analyzed spot. The intensity of each chromatographic peak for each peptide identified by MS/MS was assessed to bring supplementary information on the major protein species contained in each spot. Thanks to the MS intensity of each peptide, we obtained an estimation of their abundance in the corresponding spot. The study of retention time and accurate mass measurement allowed the identification of potentially unselected peptides during autoMS/MS.

Only Type II keratins but no KAP peptides were identified with these MS/MS experiments. We have supposed that KAP peptides were present in too low abundance and that the corresponding ions were unable to trigger the autoMS/MS experiment. In order to confirm this hypothesis, we have searched the ion KAP's signal previously identified with nanoLC-Chip-MS/MS experiments as KAP 3.1, thanks to the extraction of the traces of the corresponding ion mass. Extracted ion chromatograms showed a very low signal compared to the signals of major ions peptides from Type II keratins (Figure 5a, 5b and 5c). These results were confirmed in the other investigated spots, which indicated the presence of KAPs at a very low level of abundance compared to Type II keratins peptides in focalized spot. If the majority of spot intensity is widely attributed to keratins peptides, then KAPs could be seen as a minor contaminant present in the migration background.

The same strategy was conducted on Type I keratins. In fact, ions for some peptides of Type I keratins were found with a low signal (Figure 5d). Signal from modified peptides with methylation were not found probably because of their low abundance.

A nomenclature of Type II keratins tryptic peptides was elaborated for mapping and ion extractions for the most intense peptides were carried out for Type II keratins K85, K86, K81 and K83. In the case of multi identification of Type II keratins in the same spot, the intensity of unique peptides gave information of the major protein in the spot as illustrated in the comparison of specific peptides from spot 47 and 52 (Figure 6b). After the unambiguous detection of specific peptide signals, other expected tryptic peptides were searched. Abnormally low intense ions from adjacent peptides for keratins demonstrated the presence of chain disruption, which were interpreted as a fragmentation of proteins in spots 69, 78 and 80 (Figure 6a). In several cases, we identified the cleavage site thanks to MS/MS of cleaved species (data not shown).

In the case of spots 47, 51, 52 and 60, it was more difficult to understand the peptide mapping because of the simultaneous identification of Nter and Cter segments from several keratins despite their chain disruption detection. To assign unambiguously the keratins fragments in these spots, we have investigated the presence of the corresponding keratin proteotypic peptides and performed their abundance estimation, which was based on intensity signal measurement. Consequently, we were able to establish the co-migration of several fragments containing simultaneously Nter and Cter fragments with a similar abundance in the same spot. Information about fragments found in investigated spots gives a view of cleaved species preferentially occurring on Type II keratins (Figure 7). We have found evidence of a preferential cleavage site between the middle of Coil 1B and the middle of Coil 2 according to subsegments described by [132]. The natural occurrences of these minor fragments into the untreated hair fibre cannot be explained with certainty and this behaviour could be a consequence of fibre damage before or during the extraction. Nevertheless, it suggests a zone of weakness in these helical segments. These properties seem to be identical to those of Type I keratins, which may explain stained clusters below the usual Type I cluster. The particular location of cleavages could be a consequence of keratin arrangement into microfibrils constituted by keratins dimer assembling. Head to tail bonds between keratins with disulphide linkage could induce more resistance of these regions than helical parts.

The interpretation of most fragments explains the number of observed spots. Nevertheless, distinct spots showed fragments with the same peptide distribution leading to the question of these different pI. In fact, cleavage could be localized on the first peptide not detected in the peptide succession of the chain analyzed. Cuts could occur between each peptidic bond in

this peptide sequence. If the sequence contains several basic or acid amino acid, pI shifts of the different species could be sufficiently different to obtain a separation thanks to first dimension electrophoresis, although chain lengths are almost similar. The observation of deamidation modification in several fragments leads to a possible additional explanation of charge heterogeneity. Another explanation for the different spots with the same peptide distribution could be the presence of conformational isomers as suggested by the work of Paton et al. [132]. Results exclude the hypothesis of potential crosslinks between proteins species previously identified. The abnormal migration of these proteins in the gel might be explained by a streaking effect. This effect could have occurred both in the first and second dimension as a result of the use of overloaded sample amount in our experimental procedure. This could explain the gel background on analyzed spots.

CONCLUSION

Both approaches described in this work give complementary information about structural proteins from human hair and especially of the cortex separated from the cuticle.
The MudPit approach seems to be more adapted to the study of KAPs than gel based approaches which probably disperse their signal spatially because of the particular polymorphism, as well as physical and chemical properties of these proteins. 2D-LC method allowed us to identify more KAPs with specific peptides (21 different proteins) compared to 9 when only performing 2-DE gel experiment. This strategy seems to be more appropriated to the future studies on these multigenic families whose existence at the protein level has not been fully established.
Nevertheless, 2-DE gels have led to the study of type II keratin migration heterogeneity and suggested preferential weakness region on the α-helix parts of the hard-keratin. This information could not be reachable with MudPit approach. The results show the significance of using peptide ion intensity in addition to MS/MS experiments in order to establish the major protein species which contribute to the focalization and detection on the 2-DE gel spot. Using MS/MS data only could bring equivocal information especially when the mass spectrometer provides enhanced detection dynamics.
This preliminary study offers us interesting perspectives for future investigation. It provides us with new insights of the structure and organization of hair proteins. Improvements would have to be brought to the qualitative and quantitative aspect of keratins and KAPs in order to understand their prevalence and to study whether a relation exists with the macromolecular aspect of hair quality or shape.

Acknowledgments: The authors thank Dr Bruno A Bernard and Dr Christine Carapito for critical reading of this article, Franck Zerbib for performing 2D gels and Marcelle Huart for the Scanning Electron Microscopy image. We also thank Abel Bernot and Yann Barilly for editing the English text.

REFERENCES

[1] C. Popescu, and H. Hocker, Hair--the most sophisticated biological composite material. Chem Soc Rev 36 (2007) 1282-91.
[2] R. Dawber, Hair: its structure and response to cosmetic preparations. Clin Dermatol 14 (1996) 105-12.
[3] L.J. Wolfram, Human hair: a unique physicochemical composite. J Am Acad Dermatol 48 (2003) S106-14.
[4] R.C. Marshall, D.F. Orwin, and J.M. Gillespie, Structure and biochemistry of mammalian hard keratin. Electron Microsc Rev 4 (1991) 47-83.
[5] L. Langbein, M.A. Rogers, H. Winter, S. Praetzel, U. Beckhaus, H.R. Rackwitz, and J. Schweizer, The catalog of human hair keratins. I. Expression of the nine type I members in the hair follicle. J Biol Chem 274 (1999) 19874-84.
[6] L. Langbein, M.A. Rogers, H. Winter, S. Praetzel, and J. Schweizer, The catalog of human hair keratins. II. Expression of the six type II members in the hair follicle and the combined catalog of human type I and II keratins. J Biol Chem 276 (2001) 35123-32.
[7] L.N. Jones, Hair structure anatomy and comparative anatomy. Clin Dermatol 19 (2001) 95-103.
[8] D.A. Parry, S.V. Strelkov, P. Burkhard, U. Aebi, and H. Herrmann, Towards a molecular description of intermediate filament structure and assembly. Exp Cell Res 313 (2007) 2204-16.
[9] J.W. Hearle, Proteins fibers: structural mechanics and future opportunities. J. Mater. Sci. 42 (2007) 8010-8019.
[10] M.A. Rogers, L. Langbein, S. Praetzel-Wunder, H. Winter, and J. Schweizer, Human hair keratin-associated proteins (KAPs). Int Rev Cytol 251 (2006) 209-63.
[11] J.W. Hearle, A critical review of the structural mechanics of wool and hair fibres. Int J Biol Macromol 27 (2000) 123-38.
[12] R.D. Fraser, and D.A. Parry, Macrofibril assembly in trichocyte (hard alpha-) keratins. J Struct Biol 142 (2003) 319-25.
[13] D.A. Parry, T.A. Smith, M.A. Rogers, and J. Schweizer, Human hair keratin-associated proteins: sequence regularities and structural implications. J Struct Biol 155 (2006) 361-9.
[14] L. Langbein, and J. Schweizer, Keratins of the human hair follicle. Int Rev Cytol 243 (2005) 1-78.
[15] T.A. Smith, and D.A. Parry, Sequence analyses of Type I and Type II chains in human hair and epithelial keratin intermediate filaments: promiscuous obligate heterodimers, Type II template for molecule formation and a rationale for heterodimer formation. J Struct Biol 158 (2007) 344-57.
[16] J.E. Plowman, W.G. Bryson, L.M. Flanagan, and T.W. Jordan, Problems associated with the identification of proteins in homologous families: the wool keratin family as a case study. Anal Biochem 300 (2002) 221-9.

[17] J.E. Plowman, L.M. Flanagan, L.N. Paton, A.C. Fitzgerald, N.I. Joyce, and W.G. Bryson, The effect of oxidation or alkylation on the separation of wool keratin proteins by two-dimensional gel electrophoresis. Proteomics 3 (2003) 942-50.
[18] J.E. Plowman, S. Deb-Choudhury, A. Thomas, S. Clerens, C.D. Cornellison, A.J. Grosvenor, and J.M. Dyer, Characterisation of low abundance wool proteins through novel differential extraction techniques. Electrophoresis 31 (2010) 1937-46.
[19] M.L. Fournier, J.M. Gilmore, S.A. Martin-Brown, and M.P. Washburn, Multidimensional separations-based shotgun proteomics. Chem Rev 107 (2007) 3654-86.
[20] D.A. Wolters, M.P. Washburn, and J.R. Yates, 3rd, An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 73 (2001) 5683-90.
[21] Y.J. Lee, R.H. Rice, and Y.M. Lee, Proteome analysis of human hair shaft: from protein identification to posttranslational modification. Mol Cell Proteomics 5 (2006) 789-800.
[22] M.A. Rogers, H. Winter, L. Langbein, A. Wollschlager, S. Praetzel-Wunder, L.F. Jave-Suarez, and J. Schweizer, Characterization of human KAP24.1, a cuticular hair keratin-associated protein with unusual amino-acid composition and repeat structure. J Invest Dermatol 127 (2007) 1197-204.
[23] M.A. Rogers, L. Langbein, H. Winter, I. Beckmann, S. Praetzel, and J. Schweizer, Hair keratin associated proteins: characterization of a second high sulfur KAP gene domain on human chromosome 21. J Invest Dermatol 122 (2004) 147-58.
[24] J.E. Folk, and J.S. Finlayson, The epsilon-(gamma-glutamyl)lysine crosslink and the catalytic role of transglutaminases. Adv Protein Chem 31 (1977) 1-133.
[25] K. Yoneda, M. Akiyama, K. Morita, H. Shimizu, S. Imamura, and S.Y. Kim, Expression of transglutaminase 1 in human hair follicles, sebaceous glands and sweat glands. Br J Dermatol 138 (1998) 37-44.
[26] S. Thibaut, E. Candi, V. Pietroni, G. Melino, R. Schmidt, and B.A. Bernard, Transglutaminase 5 expression in human hair follicle. J Invest Dermatol 125 (2005) 581-5.
[27] S. Thibaut, N. Cavusoglu, E. de Becker, F. Zerbib, A. Bednarczyk, C. Schaeffer, A. van Dorsselaer, and B.A. Bernard, Transglutaminase-3 enzyme: a putative actor in human hair shaft scaffolding? J Invest Dermatol 129 (2009) 449-59.
[28] H. Zahn, and H.G. Gattner, Hair sulfur amino acid analysis. EXS 78 (1997) 239-58.
[29] J.M. Gillespie, and R.C. Marshall, A comparison of the proteins of normal and trichothiodystrophic human hair. J Invest Dermatol 80 (1983) 195-202.
[30] R.C. Marshall, Characterization of the proteins of human hair and nail by electrophoresis. J Invest Dermatol 80 (1983) 519-24.
[31] Y. Zhang, and D. Reinberg, Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. Genes Dev 15 (2001) 2343-60.
[32] K. Stewart, P.L. Spedding, M.S. Otterburn, and D.M. Lewis, Surface Layer of Wool. I. Dityrosine Synthesis and Characterization. Journal of Applied Polymer Science 66 (1997) 2359-2363.
[33] P. Coulombe, and M. Omary, Hard and Soft principles defining the structure, function and regulation of keratin intermediate filaments. Current Opinion in Cell Biology 14 (2002) 110-122.
[34] L.N. Paton, J.A. Gerrard, and W.G. Bryson, Investigastions into charge heterogeneity of wool intermediate filament proteins. Journal of Proteomics 71 (2008) 513-529.

FIGURES

Figure 1

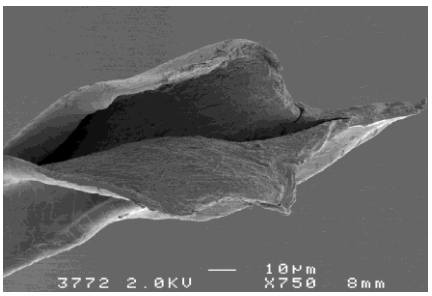


Figure 1: Scanning Electron Microscopy of hair fiber after protein extraction shows soluble cortical proteins are extracted, leaving the cuticle intact.

Figure 2

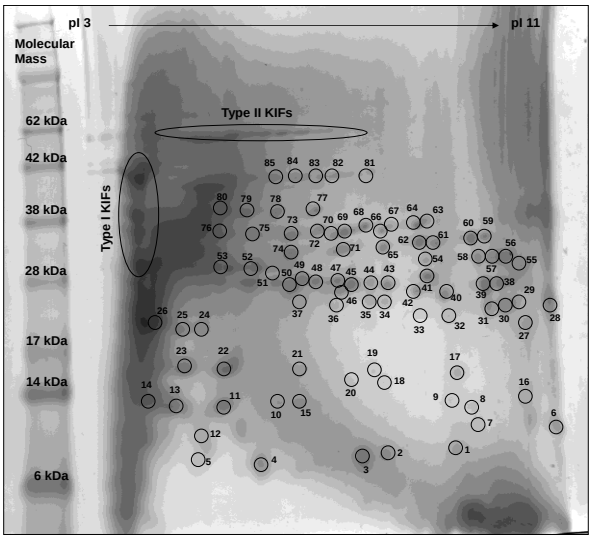


Figure 2 : 2-DE gel of hair cortical extract and sampled spots selected for the study. This gel is representative of triplicate.

Figure 3

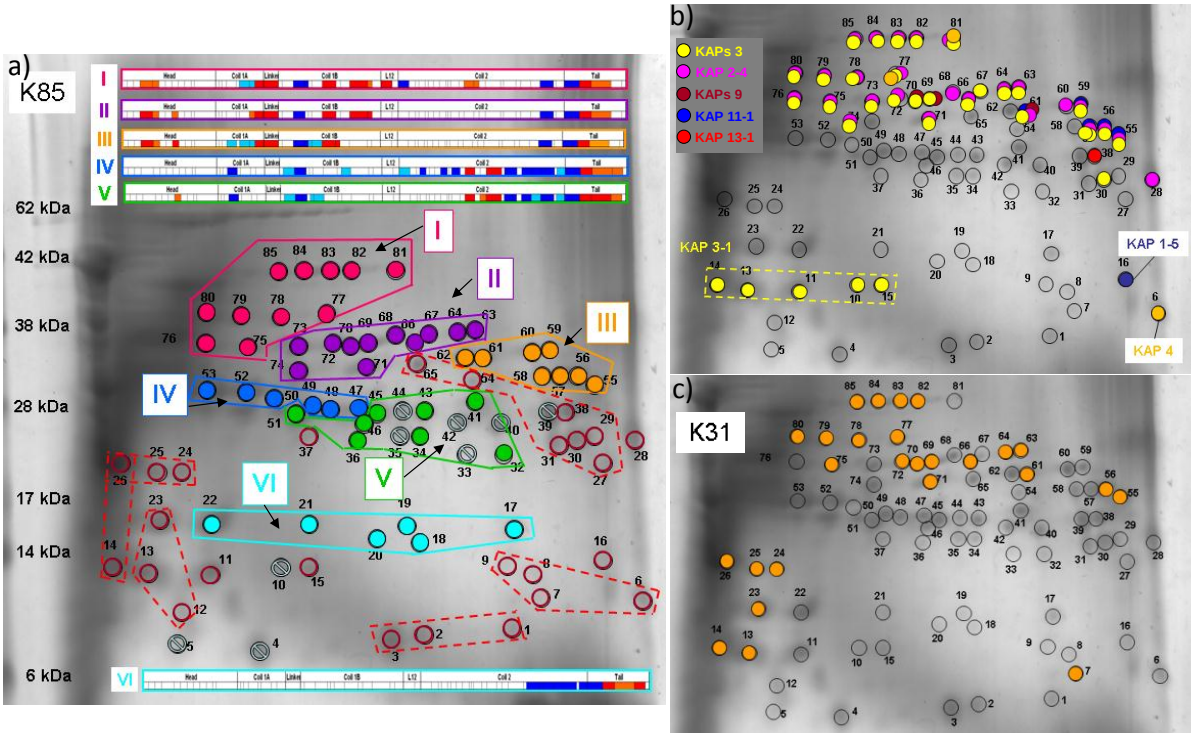


Figure 3 : Distribution of protein identifications on 2-DE gel for K85, KAPs and K31. Figure 3a: Identification for K85 showed a specific peptide pattern depending on the spot location. 6 sequence coverage patterns have been described (I to VI) and indicated with distinct color. Peptides identified in different region of keratin sequence are systematically missing or present in a family pattern. Figure 3b:: Summary of KAPs identification on 2-DE gel. KAPs 3 (in yellow) were expected around 10 kDa and found in spots 10-11 and 13-15. Other coloured spots: KAP 1-5 was identified (spot 16) below expected MW (around 18 kDa) and at unexpected pI value (expected 6,6) ; KAP 4 family identified in spot 6 was expected from 13 to 22 kDa ; highest gel zone (above spot 28) shows a large number of spots mainly containing KAPs 3 (yellow), KAPs 2 (purple), KAPs 9 (brown).

Figure 4

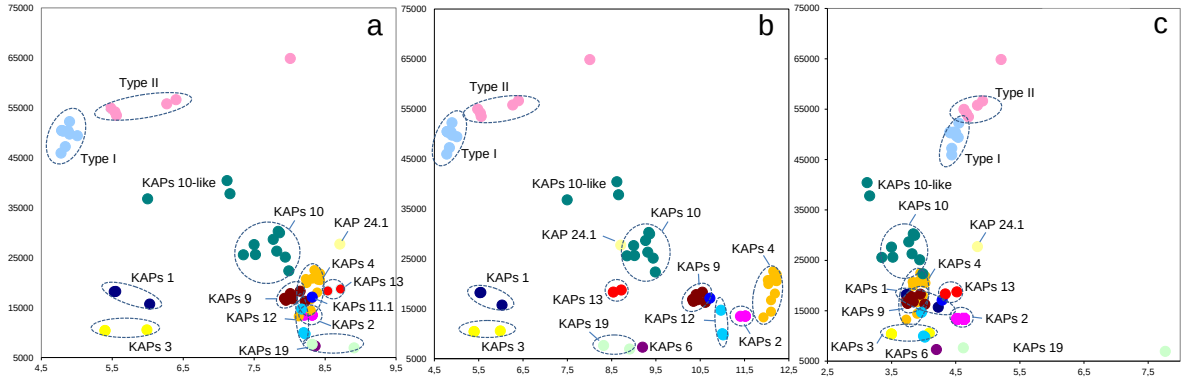


Figure 4 : Theoretical 2-DE maps of cortical and cuticular KIFs and KAPs expected after gel-MS analysis. Predicted 2-DE patterns (computing with pI/MW tool, available online on ExPASy Proteomics Server): a) without alkylation on cysteine. b) cysteine alkylation with iodoacetamide (used in this study). c) cysteine alkylation with iodocetic acid. X axis, isoelectric point; Y axis, molecular weight.

Figure 5

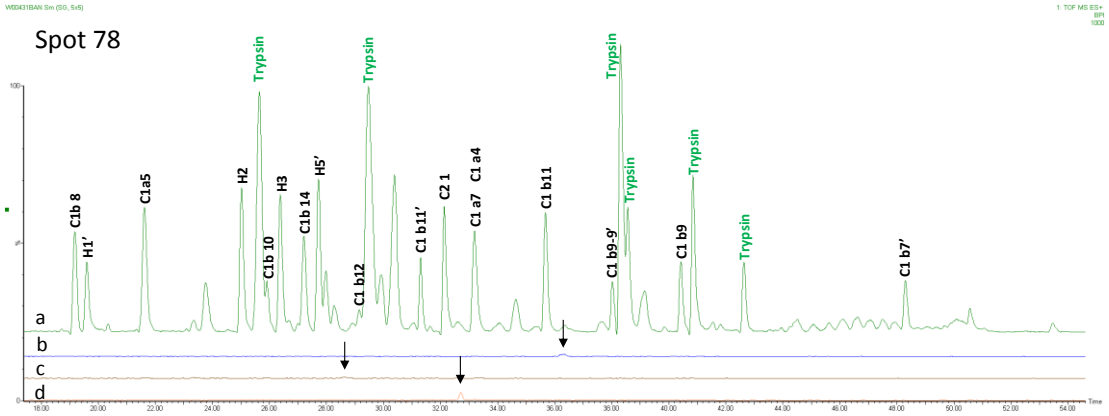


Figure 5: Illustration of ion abundance on several ion currents from proteins simultaneously identified in the spot 78. The nanoLC-MS experiment was performed on a Q-TOF mass spectrometer. The first analysis of this spot with nanoLC-MS/MS and ion trap spectrometer has identified simultaneously K81, K85, K86, K31, K33a, KAP 2 and KAP 3. Intensity axes were the same for all chromatograms.

a: Base peak ion chromatogram of tryptic digest of spot 78 reanalysed in LC-MS for peptide mapping. Noted peaks correspond to 2+ or 3+ charged peptides for K85 according to mapping results.

b: Extracted ion chromatogram (EIC) for m/z=931.409. The arrow indicates position of 2+ charged ion corresponding to KAP 3.1 peptide SCamSVPTGPATTFCamSFDK.

c: EIC for m/z=869.394. The arrow indicates position of 2+ charged ion corresponding to KAP 3.3 peptide GCamSVPTGPATTICamSSDK.

d: EIC for m/z=622.337. The arrow indicates position of 2+ charged ion corresponding to K31 peptide QLVESDINGLR.

EIC examination suggests K85 as the major protein in this spot. Specific ions from K83, KAP2 and K33a proteins were below the detection limits. Methylation observed with the ion trap on K85 was not detected suggesting the low level of this modification.

Figure 6

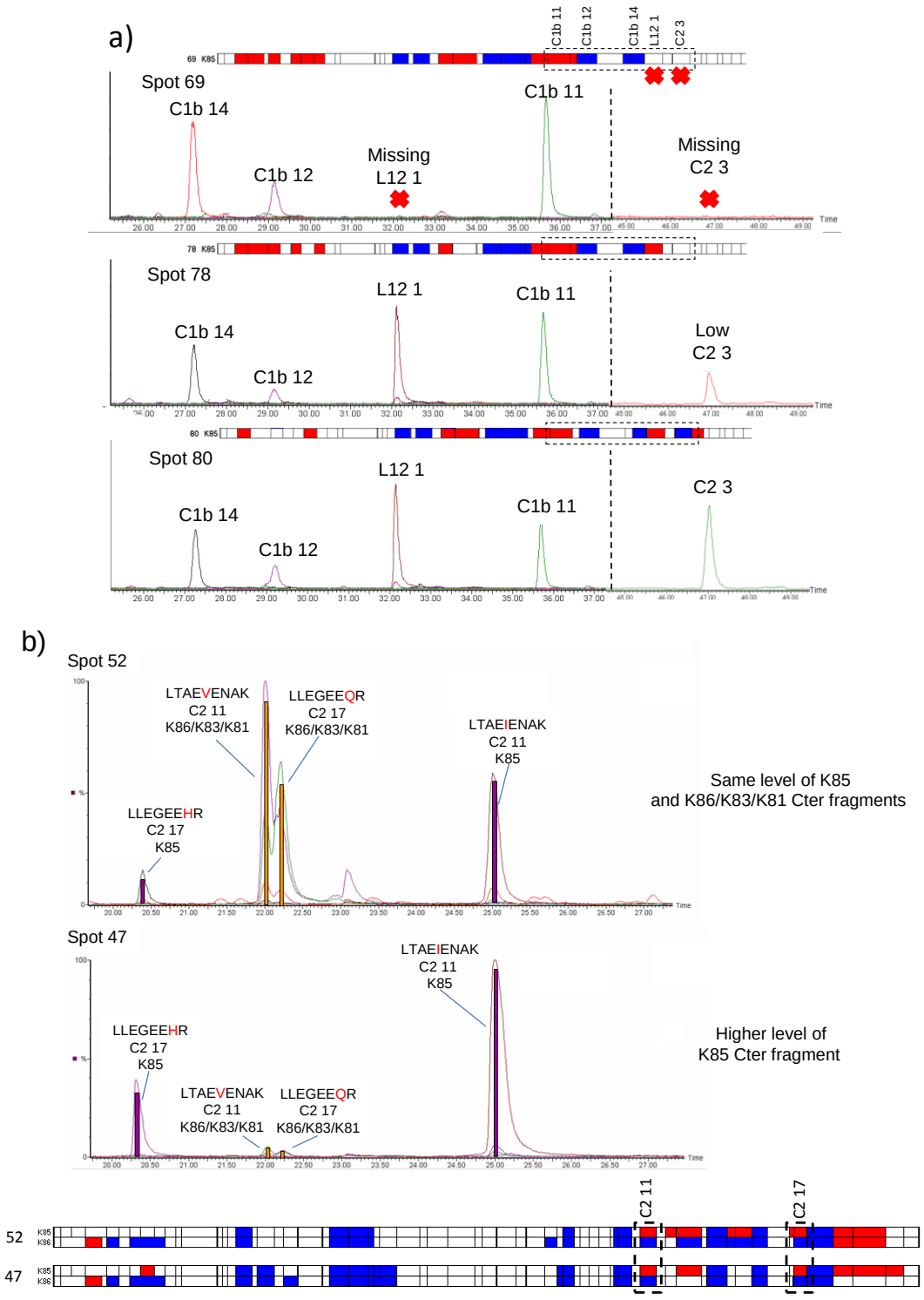


Figure 6: Extracted ion chromatograms used for spot composition understanding.
a) Five extracted ion chromatograms (EIC) with a mass tolerance window of +/- 0.05 Da corresponding to K85 peptides near the Linker 12 in spots 69, 78 and 80. Respectively and according to our nomenclature: C1b 11= ATAENEFVVLK m/z=610.83 (z=2); C1b 12= DVDCAYLR m/z=506.23 (z=2); C1b 14= LYEEEIR m/z=476.24 (z=2); L12 1= VLQAHISDTSVVVK m/z=503.96 (z=3); C2 3= DLNMDCIIAEIK m/z=717.85 (z=2).. In this case, peptides from the head to the end of the Coil 1 of K85 are identified in each spot. In spot 80, the five peptides are detected as abundant: relative ion intensities from these peptides are used as reference to compare with the corresponding ions in other spots. In spot 78, the lower relative intensity for C2 3 peptide indicates the breaking of the keratin fragment near this segment which is confirmed by the other missing Cter contiguous peptides. In spot 69, the lack of signal for L12 1 and C2 3 peptides indicates a cleavage site on L12 confirmed by the other missing Cter peptides.

b) Comparison between the EIC of two proteotypic peptides of K85 sequence and the EIC of the 2 equivalent peptides of isoforms K86, K83 or K81 between spots 52 and 47. Decrease in the intensity of K86/K83/K81 peptides indicates the low level of the corresponding protein fragment in the spot 47.

Figure 7

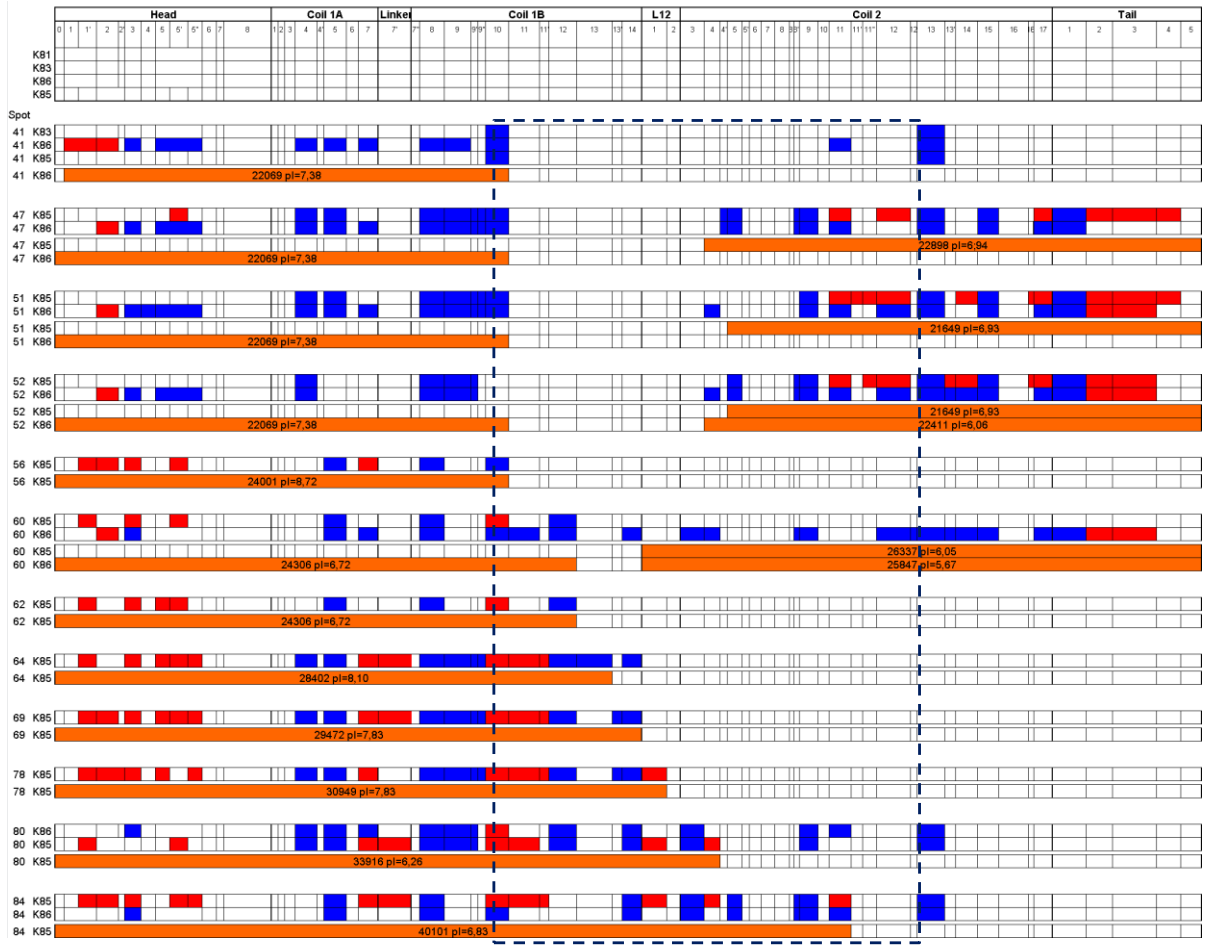


Figure 7: Summary of the mapped peptides for K85 and K86 in the investigated spots. After consideration of specific peptide ion intensity for each isoform we have established which fragments mainly contribute to the spot intensity (underlined zone). Red colour indicates specific peptide (proteotypic) and blue colour indicates peptides shared by two or more isoforms.
Expected isoelectric point and molecular mass of the corresponding fragments were calculated. Boxed zone corresponds to the observed zone of weakness of the keratin helical segments.

TABLE LEGENDS

Table 1: Keratin proteins identified after offline bidimensionnal chromatography. With the accession numbers (AC), protein names, sequence coverages and the number of unique peptides.

Table 2: Other proteins identified after offline bidimensionnal chromatography. With accession numbers (AC), proteins names, sequence coverages and the number of unique peptides.

Table 3: List of peptides identified with post-translational modifications by the offline bidimensionnal chromatography approach performed on the digested total hair extract. We showed to which proteins these sequences could match. For each identified peptide, the modified amino acids were underlined with modification Type in superscript. Legend: tox= trioxidation; dm = dimethylation; am = carbamidomethylation; ox = oxidation; m = methylation.; tm = trimethylation.

Table 4: Distribution of the 20 proteins identified in the 85 spots excised of the 2-DE gel of the hair cortical extract. Numbers indicate unique peptides for the identified protein. Asterisks indicate the number of modified peptides for the protein. Major proteins identified were the human Type II and Type I keratins and the KAPs.

Table 5: List of peptides identified with post-translational modifications and the spots numbers of the hair cortical extract 2-DE gel in which they have been detected. For each identified peptide, modified amino acids were underlined with modification type in superscript. Legend: tox= trioxydation; dm = dimethylation; am = carbamidomethylation; ox = oxydation; m = methylation.

Table 1

AC	Proteins	Insoluble		Soluble	
		Sequence coverage (%)	Number of unique peptides	Sequence coverage (%)	Number of unique peptides
O43790	Keratin Type II K86	74	18	70	19
P78386	Keratin Type II K85	73	38	70	38
Q14533	Keratin Type II K81	65	3	59	2
P78385	Keratin Type II K83	63	4	58	4
Q9NSB4	Keratin Type II K82	25	7	9	1
Q15323	Keratin Type I K31	64	8	68	11
O76009	Keratin Type I K33a	75	28	76	29
Q14525	Keratin Type I K33b	68	7	71	6
O76011	Keratin Type I K34	48	8	48	9
Q92764	Keratin Type I K35	33	7	36	7
O76013	Keratin Type I K36	7	1	0	0
Q14532	Keratin Type I K32	18	2	16	2
O76014	Keratin Type I K37	16	4	0	0
Q9BYU5	KAP 2.1 or 2.3 or 2.4	42	4	27	4
Q8IUC1	KAP 11.1	29	4	16	3
Q07627	KAP 1.1	0	0	7	1
Q9BYR8	KAP 3.1	17	1	17	2
Q9BYR7	KAP 3.2	17	1	17	1
Q9BYR6	KAP 3.3	17	1	17	1
Q52LG2	KAP 13.2	7	1	7	1
Q9BYQ4	KAP 9.2 or 9.9	19	2	24	3
A8MTY7	KAP 9.7	19	1	0	0
A8MVA2	KAP 9.6	9	1	0	0
Q3LI72	KAP 19.5	22	1	21	1
Q9BYR4	KAP 4.3	7	1	13	2
Q9BYR3	KAP 4.4	0	0	15	2
Q9BYQ5	KAP 4.6/4.12	10	2	27	5
Q9BYR0	KAP 4.7	0	0	12	1
Q9BYQ8	KAP 4.9	22	2	0	0
P60331	KAP 10.1	5	2	0	0
P60369	KAP 10.3	10	3	0	0
P60409	KAP 10.7	6	3	0	0
A8MUX0	KAP 10-like	2	1	0	0
Q3LI83	KAP 24.1	5	1	5	1

Table 2

AC	Proteins	Sequence coverage (%)	Number of unique peptides
P27482	Calmodulin-like protein 3	19	2
Q86SJ6	Desmoglein-4	7	5
P33778, …	Histone H2B	12	2
P31947	14-3-3 protein sigma	9	2
P14923	Junction plakoglobin	6	2
P04908, …	Histone H2A	22	3
P81605	Dermcidin precursor	9	1
P33764	Protein S100-A3	11	1
P62805	Histone H4	21	4
Q13228	Selenium-binding protein 1	4	1
P11142	Heat shock cognate 71 kDa protein	2	1
Q9Y3R4	Sialidase-2	3	1
Q13867	Bleomycin hydrolase	3	1
P36952	Serpin B5 precursor	4	1
P63244	Guanine nucleotide-binding protein subunit beta 2-like 1	7	1
Q86SG7	Lysozyme g-like protein 2 r	6	1
Q8TDC3	BR serine/threonine-protein kinase 1	1	1
Q08188	Protein-glutamine gamma-glutamyltransferase E (TGase 3)	4	1
P11279	Lysosome-associated membrane glycoprotein 1	2	1
P13473	Lysosome-associated membrane glycoprotein 2	2	1
Q6MZM0	Hephaestin-like protein 1	4	3
Q8NF91	Nesprin-1	2	1

Table 3

Total Hair Extract	
Proteins	**Modified peptides identified**
K81/K83/K86	K.C[am]QLSK[dm]LEAAVAQSEQQGEAALSDAR.C
K81/ K83/ K85/ K86	K.K[dm]DVDC[am]AYLR.K
K81/K83/85/K86	K.K[tm]DVDC[am]AYRL.K
K81/ K86	R.AFSC[am]ISAC[tox]GPRPGR.C
K86	R.GGVVC[am]GDLC[tox]ASTTAPVVSTR.V
K86	R.VLQSH[m]ISDTSVVV.K
K85/K86	R.LC[tox]EGCGSVNVC[am]VSSSR.G
K81/ K83/ K85/ K86	R.TK[dm]EEINELNR.M
K85	R.GGVSC[tox]GGLSYSTTPGR.Q
K85	R.SLC[am]NLGSC[tox]GPR.I
K81/ K83/ K85/ K86	K.LAELEGALQK[dm]AK.Q
K81/ K83/ K85/ K86	R.C[tox]KLAELEGALQK.A
K81/ K85/ K86	R.DLNM[ox]DC[tox]IIAEIK.A
K81/ K83/ K85/ K86	R.EAEC[tox]VEADSGR.L
K81/ K83/ K85/ K86	R.FAAFIDK[dm]VR.F
K81/ K83/ K85/ K86	R.FLEQQNK[dm]LLETK.W
K81/ K83/ K85/ K86	R.FLEQQNK[tm]LLETK.W

K81/ K83/ K85/ K86	R.LASELNHVQEVLEGYK^{dm}K.K
K81/ K83/ K85/ K86	K.AK^{dm}QDMAC^{am}LIR.E.
K81/ K83/ K85/ K86	K.AKQDMAC^{tox}LIR.E.
K81/ K83/ K86	R.GLTGGFGSH^{m}SVC^{am}GGFR.A
K81/ K83/ K86	R.GLTGGFGSHSVC^{tox}GGFR.A
K31/ K33a/ K33b/ K34/ K35	R.ARLEC^{tox}EINTYR.S
K31/ K33a/ K33b/ K34/K35	R.LEC^{tox}EINTYR.S
K31/ K32/K34	R.ILDELTLC^{tox}K.S
K31/ K33a/ K33b/ K34	R.LASYLEK^{dm}VR.Q
K33a/ K33b	R.ILDELTLC^{tox}R.S
K33a/K35	R.SDEAQVESLKEELLC^{tox}LK.Q

Table 5

Proteins	Sequences of modified peptides	Identified in spot numbers
K81/ K86	K.LAELEGALQK^{dm}AK.Q	23
K81/ K86	R.GLTGGFGSH^{m}SVC^{am}GGFR.A	29/30/31/32/41/43
K81/ K83/ K86	R.K^{tm}SDLEANVEALIQEIDFLR.R	76/80
K81/ K85/ K86	R.FLEQQNK^{dm}LLETK.W	56/58/85
K85	R.NFSSC^{am}SAVAPK^{dm}TGNR.C	55/56/57/58/59/60/61/62/64/69/70/71/72/73/76/78/80
K85	R.NFSSC^{am}SAVAPK^{tm}TGNR.C	56/57/62/71/76
K85	K.LLETK^{tm}WQFYQNQR.C	58/73
K85	K.K^{tm}YEEEVALR.A	62
K85	R.LTAEIENAK^{dm}C^{am}QR.A	83/84/85

Annexe 5 : Protéines identifiées dans les extraits cuticulaires, corticaux et unguéaux

Listes des protéines identifiées lors de l'analyse du protéome de la cuticule (majoritairement endocuticulaire) telle que décrite dans le chapitre II de la troisième partie et lors de l'analyse du protéome des ongles telle que décrite dans le chapitre III de la troisième partie.

Les protéines ainsi identifiées sont comparées avec celles identifiées dans le cortex au cours d'une analyse témoin d'un digest trypsique en LC-LC-MS/MS effectuée avec trois répétitions (différentes de celles décrites dans le premier chapitre de la troisième partie).

Annexe 5 : Protéines identifiées dans les extraits cuticulaires, corticaux et unguéaux

					angle							cortex			endocuticule riche								
					Tryps (x3)		Chymo (x3)		V8 (x1)			Tryps (x3)			Tryps (x3)		Chymo (x1)		NTCB Tryps (x2)		NTCB Chymo (x2)		
Protéine	Symbole	N° Acession	MW	Prob	SC	pu	SC	pu	SC	pu	Tot	SC	pu	Tot	SC	pu	SC	pu	SC	pu	SC	pu	Tot
10 kDa heat shock protein, mitochondrial OS=Homo sapiens G	HSPE1	sp|P61604	10914	81%							0			0	2	1			2	1			4
14-3-3 protein beta/alpha OS=Homo sapiens GN=YWHAB PE=1	YWHAB	sp|P31946	28065	100%							0			0	1	2	2	1	1	1			4
14-3-3 protein epsilon OS=Homo sapiens GN=YWHAE PE=1 SV=:	YWHAE	sp|P62258	29157	100%							0			0	2	5			2	2			4
14-3-3 protein gamma OS=Homo sapiens GN=YWHAG PE=1 SV=	YWHAG	sp|P61981	28285	81%							0			0	3	1			2	1			5
14-3-3 protein sigma OS=Homo sapiens GN=SFN PE=1 SV=1	SFN	sp|P31947	27757	100%	3	2	6		1	2	10	3	2	3	4	2	1	5	2	2	2	3	9
14-3-3 protein theta OS=Homo sapiens GN=YWHAQ PE=1 SV=1	YWHAQ	sp|P27348	27747	100%							0			0	1	1	1	1	1	1			3
14-3-3 protein zeta/delta OS=Homo sapiens GN=YWHAZ PE=1 S	YWHAZ	sp|P63104	27728	100%	3	1					3			0	2	3	2	1	2	3			6
1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase	PLCD1	sp|P51178	85649	100%	3	1					3			0	2	4			9	1	1	1	12
26S proteasome non-ATPase regulatory subunit 12 OS=Homo s	PSMD12	sp|O00232	52888	81%			1	1			1			0	1	1	1	1			1	1	3
3-hydroxyacyl-CoA dehydrogenase type-2 OS=Homo sapiens G	HSD17B10	sp|Q99714	26905	81%							0			0	1	1							1
40S ribosomal protein S11 OS=Homo sapiens GN=RPS11 PE=1 S	RPS11	sp|P62280	18413	100%							0			0	2	2							2
40S ribosomal protein S12 OS=Homo sapiens GN=RPS12 PE=1 S	RPS12	sp|P25398	14497	79%							0			0					1	1			1
40S ribosomal protein S14 OS=Homo sapiens GN=RPS14 PE=1 S	RPS14	sp|P62263	16255	100%							0			0	2	2							2
40S ribosomal protein S15a OS=Homo sapiens GN=RPS15A PE=	RPS15A	sp|P62244	14822	100%							0			0	1	1	1	1	1	1			3
40S ribosomal protein S16 OS=Homo sapiens GN=RPS16 PE=1 S	RPS16	sp|P62249	16428	100%							0			0	2	1			6	1			8
40S ribosomal protein S18 OS=Homo sapiens GN=RPS18 PE=1 S	RPS18	sp|P62269	17701	100%							0			0	4	3			4	2			7
40S ribosomal protein S2 OS=Homo sapiens GN=RPS2 PE=1 SV=	RPS2	sp|P15880	31307	100%							0			0			2	2			1	1	3
40S ribosomal protein S20 OS=Homo sapiens GN=RPS20 PE=1 S	RPS20	sp|P60866	13355	81%							0			0	1	1							1
40S ribosomal protein S21 OS=Homo sapiens GN=RPS21 PE=1 S	RPS21	sp|P63220	9094	81%							0			0	4	1			2	1			6
40S ribosomal protein S23 OS=Homo sapiens GN=RPS23 PE=1 S	RPS23	sp|P62266	15790	81%							0			0	2	1			4	1			6
40S ribosomal protein S24 OS=Homo sapiens GN=RPS24 PE=1 S	RPS24	sp|P62847	15406	81%							0			0	1	1							1
40S ribosomal protein S25 OS=Homo sapiens GN=RPS25 PE=1 S	RPS25	sp|P62851	13725	100%							0			0	1	3	1	1	2	2			4
40S ribosomal protein S28 OS=Homo sapiens GN=RPS28 PE=1 S	RPS28	sp|P62857	7823	81%							0			0	1	1			1	1			2
40S ribosomal protein S3 OS=Homo sapiens GN=RPS3 PE=1 SV=	RPS3	sp|P23396	26671	100%	1	1	4	1			5			0	2	6			2	1			4
40S ribosomal protein S3a OS=Homo sapiens GN=RPS3A PE=1 S	RPS3A	sp|P61247	29927	85%							0			0			1	1			3	1	4
40S ribosomal protein S4, X isoform OS=Homo sapiens GN=RPS	RPS4X	sp|P62701,s	29581	98%							0			0	1	1			2	1			3
40S ribosomal protein S6 OS=Homo sapiens GN=RPS6 PE=1 SV=	RPS6	sp|P62753	28664	100%							0			0	2	2							2
40S ribosomal protein S7 OS=Homo sapiens GN=RPS7 PE=1 SV=	RPS7	sp|P62081	22110	85%							0			0			1	1					1
40S ribosomal protein S8 OS=Homo sapiens GN=RPS8 PE=1 SV=	RPS8	sp|P62241	24188	100%							0			0	1	2			3	2			4
40S ribosomal protein S9 OS=Homo sapiens GN=RPS9 PE=1 SV=	RPS9	sp|P46781	22575	100%							0			0	1	2			1	1			2
40S ribosomal protein SA OS=Homo sapiens GN=RPSA PE=1 SV=	RPSA	sp|P08865	32836	100%							0			0	2	5	2	2	3	3	2	3	9
4-trimethylaminobutyraldehyde dehydrogenase OS=Homo sa	ALDH9A1	sp|P49189	53784	81%							0			0	1	1							1
60S acidic ribosomal protein P0 OS=Homo sapiens GN=RPLP0 F	RPLP0	sp|P05388,s	34347	81%	6	1					6			0	1	1	2	1	2	1	2	1	7
60S acidic ribosomal protein P2 OS=Homo sapiens GN=RPLP2 F	RPLP2	sp|P05387	11647	100%							0			0	2	4			4	1			6
60S ribosomal protein L10 OS=Homo sapiens GN=RPL10 PE=1 S	RPL10	sp|P27635	24587	81%							0			0	2	1			1	1			3
60S ribosomal protein L11 OS=Homo sapiens GN=RPL11 PE=1 S	RPL11	sp|P62913	20235	100%							0			0	1	4			3	1			4
60S ribosomal protein L12 OS=Homo sapiens GN=RPL12 PE=1 S	RPL12	sp|P30050	17801	100%	1	1					1			0	1	2			1	1	1	1	3
60S ribosomal protein L13 OS=Homo sapiens GN=RPL13 PE=1 S	RPL13	sp|P26373	24244	81%							0			0	1	1			1	1			2
60S ribosomal protein L13a OS=Homo sapiens GN=RPL13A PE=	RPL13A	sp|P40429	23560	79%							0			0					1	1			1
60S ribosomal protein L14 OS=Homo sapiens GN=RPL14 PE=1 S	RPL14	sp|P50914	23414	85%							0			0	2	1	2	1			2	1	6
60S ribosomal protein L15 OS=Homo sapiens GN=RPL15 PE=1 S	RPL15	sp|P61313	24129	81%							0			0	2	1							2
60S ribosomal protein L18 OS=Homo sapiens GN=RPL18 PE=1 S	RPL18	sp|Q07020	21617	100%							0			0	1	3			3	1			4
60S ribosomal protein L21 OS=Homo sapiens GN=RPL21 PE=1 S	RPL21	sp|P46778	18547	81%							0			0	1	1							1
60S ribosomal protein L22 OS=Homo sapiens GN=RPL22 PE=1 S	RPL22	sp|P35268	14769	100%							0			0	2	2			1	2			3
60S ribosomal protein L27a OS=Homo sapiens GN=RPL27A PE=	RPL27A	sp|P46776	16543	81%							0			0	1	1			1	1			2
60S ribosomal protein L3 OS=Homo sapiens GN=RPL3 PE=1 SV=	RPL3	sp|P39023	46092	100%							0			0	2	2	1	1					3
60S ribosomal protein L30 OS=Homo sapiens GN=RPL30 PE=1 S	RPL30	sp|P62888	12767	98%							0			0	1	1							1
60S ribosomal protein L31 OS=Homo sapiens GN=RPL31 PE=1 S	RPL31	sp|P62899	14445	85%							0			0			1	1			1	1	2
60S ribosomal protein L35 OS=Homo sapiens GN=RPL35 PE=1 S	RPL35	sp|P42766	14535	81%							0			0	2	1			1	1			3
60S ribosomal protein L35a OS=Homo sapiens GN=RPL35A PE=	RPL35A	sp|P18077	12520	81%							0			0	1	1							1
60S ribosomal protein L4 OS=Homo sapiens GN=RPL4 PE=1 SV=	RPL4	sp|P36578	47681	81%	2	1					2			0	2	1			1	1			3
60S ribosomal protein L5 OS=Homo sapiens GN=RPL5 PE=1 SV=	RPL5	sp|P46777	34346	81%							0			0	1	1							1
60S ribosomal protein L6 OS=Homo sapiens GN=RPL6 PE=1 SV=	RPL6	sp|Q02878	32711	100%							0			0	1	1	1	1	1	1			3
60S ribosomal protein L7 OS=Homo sapiens GN=RPL7 PE=1 SV=	RPL7	sp|P18124	29210	99%							0			0	2	1			2	1			4
60S ribosomal protein L7a OS=Homo sapiens GN=RPL7A PE=1 S	RPL7A	sp|P62424	29978	81%							0			0	2	1			3	1			5
60S ribosomal protein L9 OS=Homo sapiens GN=RPL9 PE=1 SV=	RPL9	sp|P32969	21846	81%			1	1			1			0							1	1	1
6-phosphogluconate dehydrogenase, decarboxylating OS=Hor	PGD	sp|P52209	53124	81%							0			0	1	1							1
6-phosphogluconolactonase OS=Homo sapiens GN=PGLS PE=1	PGLS	sp|O95336	27530	100%							0			0	1	1			1	1			2
78 kDa glucose-regulated protein OS=Homo sapiens GN=HSPA!	HSPA5	sp|P11021	72317	100%							0			0	2	4	1	1	1	3			4
Absent in melanoma 1 protein OS=Homo sapiens GN=AIM1 PE=	AIM1	sp|Q9Y4K1	2E+05	100%							0			0	1	3	1	2	2	5	2	5	6
Acid ceramidase OS=Homo sapiens GN=ASAH1 PE=1 SV=5	ASAH1	sp|Q13510	44643	81%							0			0	1	1							1
Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1	ACTB	sp|P60709	41720	100%	4	2	1	1	1	1	6			0	1	2							1

Actin, cytoplasmic 2 OS=Homo sapiens GN=ACTG1 PE=1 SV=1	ACTG1	sp|P63261	41776	100%			2	1			2			0																
Actin-related protein 2 OS=Homo sapiens GN=ACTR2 PE=1 SV=1	ACTR2	sp|P61160	44744	81%							0			0	1	1													1
Actin-related protein 2/3 complex subunit 3 OS=Homo sapiens	ARPC3	sp|O15145	20530	81%							0			0	1	1													1
Acyl-CoA-binding protein OS=Homo sapiens GN=DBI PE=1 SV=2	DBI	sp|P07108	10027	81%							0			0	2	1													2
ADP-ribosylation factor 3 OS=Homo sapiens GN=ARF3 PE=1 SV=	ARF3	sp|P61204,s	20680	100%							0			0	10	1			3	2									13
ADP-ribosylation factor 5 OS=Homo sapiens GN=ARF5 PE=1 SV=	ARF5	sp|P84085	20512	81%							0			0	1	1							1	1	2				
ADP-ribosylation factor 6 OS=Homo sapiens GN=ARF6 PE=1 SV=	ARF6	sp|P62330	20065	81%							0			0	1	1													1
ADP-sugar pyrophosphatase OS=Homo sapiens GN=NUDT5 PE=	NUDT5	sp|Q9UKK9	24310	81%							0			0	1	1							1	1	2				
Alpha-2-macroglobulin-like protein 1 OS=Homo sapiens GN=A2	A2ML1	sp|A8K2U0	2E+05	99%			2	1	1	1	3			0																
Alpha-actinin-4 OS=Homo sapiens GN=ACTN4 PE=1 SV=2	ACTN4	sp|O43707,s	1E+05	81%							0			0	1	1			2	1									3
Alpha-crystallin B chain OS=Homo sapiens GN=CRYAB PE=1 SV=	CRYAB	sp|P02511	20141	100%			4	2			4			0																
Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2	ENO1	sp|P06733	47152	100%	3	1	5	1			8			0	1	4			1	2	1	1	3							
Annexin A1 OS=Homo sapiens GN=ANXA1 PE=1 SV=2	ANXA1	sp|P04083	38698	100%	3	2			2	1	5			0																
Annexin A2 OS=Homo sapiens GN=ANXA2 PE=1 SV=2	ANXA2	sp|P07355	38588	100%	3	4					3			0																
Annexin A4 OS=Homo sapiens GN=ANXA4 PE=1 SV=4	ANXA4	sp|P09525	35866	100%							0			0	1	3													1
Annexin A8-like protein 2 OS=Homo sapiens GN=ANXA8L2 PE=2	ANXA8L2	sp|Q5VT79	36849	81%			1				1			0																
AP-2 complex subunit alpha-1 OS=Homo sapiens GN=AP2A1 PE	AP2A1	sp|O95782	1E+05	81%							0			0	2	1			1	1									3
Armadillo repeat-containing protein 10 OS=Homo sapiens GN=	ARMC10	sp|Q8N2F6	37523	85%							0			0			1	1											1
Arylsulfatase F OS=Homo sapiens GN=ARSF PE=1 SV=4	ARSF	sp|P54793	65923	98%							0			0	1	1													1
Aspartate aminotransferase, mitochondrial OS=Homo sapiens	GOT2	sp|P00505	47501	100%							0			0	1	3			3	2									4
Aspartyl-tRNA synthetase, cytoplasmic OS=Homo sapiens GN=	DARS	sp|P14868	57119	85%							0			0			1	1			1	1	2						
ATP synthase subunit alpha, mitochondrial OS=Homo sapiens	ATP5A1	sp|P25705	59734	100%	3	1	1	1			4			0	2	4			1	1			3							
ATP synthase subunit beta, mitochondrial OS=Homo sapiens C	ATP5B	sp|P06576	56543	100%							0			0	1	2	1	1	2	1			4						
ATP-dependent RNA helicase A OS=Homo sapiens GN=DHX9 PE	DHX9	sp|Q08211	1E+05	100%							0			0	2	2			2	2	1	1	4						
Beta-1,4-galactosyltransferase 6 OS=Homo sapiens GN=B4GAL	B4GALT6	sp|Q9UBX8	44897	81%							0			0	1	1													1
Beta-actin-like protein 2 OS=Homo sapiens GN=ACTBL2 PE=1 SV	ACTBL2	sp|Q562R1	41986	81%	13	1					13			0	2	1			2	2	2	2	6						
Beta-catenin-interacting protein 1 OS=Homo sapiens GN=CTNI	CTNNBIP1	sp|Q9NSA3	9152	81%							0			0	1	1	1	1	2	1			4						
Beta-hexosaminidase subunit beta OS=Homo sapiens GN=HEX	HEXB	sp|P07686	63095	81%							0			0	1	1	1	2					2						
Bleomycin hydrolase OS=Homo sapiens GN=BLMH PE=1 SV=1	BLMH	sp|Q13867	52545	100%							0			0	1	2							1						
BRCA1-associated protein required for ATM activation protein	BAAT1	sp|Q6PJG6	88101	81%			15	1			15			0			1	1			1	1	2						
Calmodulin-like protein 3 OS=Homo sapiens GN=CALML3 PE=1	CALML3	sp|P27482	16873	100%	6	4	3	2	1	2	10	3	3	3	6	5	1	1	6	2	1	1	13						
Calmodulin-like protein 5 OS=Homo sapiens GN=CALML5 PE=1	CALML5	sp|Q9NZT1	15875	100%	3	1	2	1			5			0	1	3	1	1	3	2	2	1	7						
Carbonyl reductase [NADPH] 1 OS=Homo sapiens GN=CBR1 PE=	CBR1	sp|P16152	30357	81%							0			0	2	1							2						
Carnitine O-acetyltransferase OS=Homo sapiens GN=CRAT PE=	CRAT	sp|P43155	70841	100%							0			0	2	4	1	3	1	4	1	2	5						
Casein kinase I isoform delta OS=Homo sapiens GN=CSNK1D PI	CSNK1D	sp|P48730,s	47299	81%							0			0	1	1	1	1	2	1	1	2	5						
Catenin beta-1 OS=Homo sapiens GN=CTNNB1 PE=1 SV=1	CTNNB1	sp|P35222	85479	100%							0			0	2	2			2	2			4						
Cathepsin D OS=Homo sapiens GN=CTSD PE=1 SV=1	CTSD	sp|P07339	44535	100%							0			0	1	1			1	1			2						
Chloride intracellular channel protein 1 OS=Homo sapiens GN	CLIC1	sp|O00299	26905	100%							0			0	1	1			2	1			3						
Citrate synthase, mitochondrial OS=Homo sapiens GN=CS PE=	CS	sp|O75390	51696	81%							0			0	1	1	1	1					2						
Clathrin heavy chain 1 OS=Homo sapiens GN=CLTC PE=1 SV=5	CLTC	sp|Q00610	2E+05	81%			1	1			1			0	1	5			1	1	1	2	3						
CMP-N-acetylneuraminate-poly-alpha-2,8-sialyltransferase O	ST8SIA4	sp|Q92187	41278	100%	2	1					2			0																
Coatomer subunit beta OS=Homo sapiens GN=COPB1 PE=1 SV=	COPB1	sp|P53618	1E+05	100%							0			0	1	1							1						
Coatomer subunit beta' OS=Homo sapiens GN=COPB2 PE=1 SV=	COPB2	sp|P35606	1E+05	81%							0			0	1	1							1						
Coatomer subunit delta OS=Homo sapiens GN=ARCN1 PE=1 SV=	ARCN1	sp|P48444	57193	81%							0			0	1	1							1						
Coatomer subunit gamma-2 OS=Homo sapiens GN=COPG2 PE=	COPG2	sp|Q9UBF2	97607	100%							0			0	1	2							1						
Coatomer subunit zeta-1 OS=Homo sapiens GN=COPZ1 PE=1 SV	COPZ1	sp|P61923	20181	81%							0			0	1	1							1						
Constitutive coactivator of PPAR-gamma-like protein 2 OS=Homo sapiens G		sp|Q9NX05,s	1E+05	85%							0			0			3	1					3						
Creatine kinase U-type, mitochondrial OS=Homo sapiens GN=(	CKMT1A	sp|P12532	47020	100%							0			0	2	3			3	1			5						
Cullin-associated NEDD8-dissociated protein 1 OS=Homo sapi	CAND1	sp|Q86VP6	1E+05	81%							0			0	1	1							1						
Cysteinyl-tRNA synthetase, cytoplasmic OS=Homo sapiens GN	CARS	sp|P49589	85458	100%							0			0	1	1							1						
Cytochrome b561 domain-containing protein 1 OS=Homo sapi	CYB561D1	sp|Q8N8Q1	25407	88%							0			0							1	1	1						
Cytochrome c oxidase subunit 2 OS=Homo sapiens GN=MT-CO2	MT-CO2	sp|P00403	25548	81%							0			0	1	1							1						
Cytoskeleton-associated protein 4 OS=Homo sapiens GN=CKAI	CKAP4	sp|Q07065	66004	100%							0			0	1	2							1						
Cytosol aminopeptidase OS=Homo sapiens GN=LAP3 PE=1 SV=1	LAP3	sp|P28838	56150	81%			1	1			1			0	2	9	1	3	1	2	2	2	6						
Cytosolic endo-beta-N-acetylglucosaminidase OS=Homo sapie	ENGASE	sp|Q8NFI3	83968	100%							0			0	1	3	1	1	1	1			3						
Cytosolic non-specific dipeptidase OS=Homo sapiens GN=CND	CNDP2	sp|Q96KP4	52862	99%			5	1			5			0							3	1	3						
D-3-phosphoglycerate dehydrogenase OS=Homo sapiens GN=I	PHGDH	sp|O43175	56633	100%							0			0	2	2			2	2	1	1	5						
Dermcidin OS=Homo sapiens GN=DCD PE=1 SV=2	DCD	sp|P81605	11266	100%							0			0	2	2							2						
Desmocollin-1 OS=Homo sapiens GN=DSC1 PE=1 SV=2	DSC1	sp|Q08554	99970	100%	2	1					2			0																
Desmocollin-2 OS=Homo sapiens GN=DSC2 PE=1 SV=1	DSC2	sp|Q02487	99945	92%	3	1					3			0																
Desmocollin-3 OS=Homo sapiens GN=DSC3 PE=1 SV=3	DSC3	sp|Q14574	99953	100%	2	1	3	2	1	1	6			0																
Desmoglein-1 OS=Homo sapiens GN=DSG1 PE=1 SV=2	DSG1	sp|Q02413	1E+05	100%	2	2	2	6	2	2	6			0																
Desmoglein-3 OS=Homo sapiens GN=DSG3 PE=1 SV=2	DSG3	sp|P32926	1E+05	100%			3	1			3			0																
Desmoglein-4 OS=Homo sapiens GN=DSG4 PE=1 SV=1	DSG4	sp|Q86SJ6	1E+05	81%	2	1	3	3	1	1	6			0	1	1	2	1	1	1	2	1	6						
Desmoplakin OS=Homo sapiens GN=DSP PE=1 SV=3	DSP	sp|P15924	3E+05	100%	5	30	3	18	3	21	11	2	4	2	2	32	1	3	2	9	1	2	6						
Destrin OS=Homo sapiens GN=DSTN PE=1 SV=3	DSTN	sp|P60981	18488	81%			3	1			3			0																
DNA (cytosine-5)-methyltransferase 1 OS=Homo sapiens GN=D	DNMT1	sp|P26358	2E+05	85%							0			0			1	1					1						
DNA-dependent protein kinase catalytic subunit OS=Homo sa	PRKDC	sp|P78527	5E+05	100%							0			0	1	2							1						
DnaJ homolog subfamily B member 1 OS=Homo sapiens GN=DI	DNAJB1	sp|P25685	38028	100%							0			0	1	2					1	1	2						
Dual specificity protein phosphatase 14 OS=Homo sapiens GN	DUSP14	sp|O95147	22237	92%	2	1	2	1			4	6	1	6					1	1			1						

Annexe 5 : Protéines identifiées dans les extraits cuticulaires, corticaux et unguéaux

Dynamin-1-like protein OS=Homo sapiens GN=DNM1L PE=1 SV=	DNM1L	sp|O00429	81861	81%						0			0	1	1	1	1			1	1	3
E3 ubiquitin-protein ligase BRE1A OS=Homo sapiens GN=RNF2(	RNF20	sp|Q5VTR2	1E+05	79%						0			0					1	1			1
Elongation factor 1-alpha 1 OS=Homo sapiens GN=EEF1A1 PE=1	EEF1A1	sp|P68104,s	50123	100%	3	1	5	1		8			0	2	5	1	2	1	1	1	3	5
Elongation factor 1-beta OS=Homo sapiens GN=EEF1B2 PE=1 SV	EEF1B2	sp|P24534	24746	81%						0			0	1	1							1
Elongation factor 1-delta OS=Homo sapiens GN=EEF1D PE=1 SV	EEF1D	sp|P29692	31104	100%						0			0	2	3			3	3			4
Elongation factor 1-gamma OS=Homo sapiens GN=EEF1G PE=1	EEF1G	sp|P26641	50101	100%						0			0	5	2	1	1	1	1	1	1	8
Elongation factor 2 OS=Homo sapiens GN=EEF2 PE=1 SV=4	EEF2	sp|P13639	95322	100%			1	1		1			0	2	7	1	4	3	2	1	1	7
Endoplasmin OS=Homo sapiens GN=HSP90B1 PE=1 SV=1	HSP90B1	sp|P14625	92454	100%						0			0	1	1			1	1			2
Endothelial protein C receptor OS=Homo sapiens GN=PROCR P	PROCR	sp|Q9UNN8	26653	81%						0			0	1	1			1	1			2
Enoyl-CoA hydratase domain-containing protein 1 OS=Homo s	ECHDC1	sp|Q9NTX5	33681	100%						0			0	2	7	1	1	2	4	2	1	7
Envoplakin OS=Homo sapiens GN=EVPL PE=1 SV=3	EVPL	sp|Q92817	2E+05	81%			1	1		1			0									
Epidermal growth factor receptor kinase substrate 8-like prot	EPS8L2	sp|Q9H6S3	80603	81%						0			0	1	1							1
Epiplakin OS=Homo sapiens GN=EPPK1 PE=1 SV=2	EPPK1	sp|P58107	6E+05	100%	2	5	8	3	1	3	11		0									
Erlin-1 OS=Homo sapiens GN=ERLIN1 PE=1 SV=1	ERLIN1	sp|O75477,s	38909	81%						0			0	1	1							1
Eukaryotic initiation factor 4A-I OS=Homo sapiens GN=EIF4A1 P	EIF4A1	sp|P60842	46137	100%						0			0	2	6	2	2	1	5	1	1	5
Eukaryotic peptide chain release factor subunit 1 OS=Homo s	ETF1	sp|P62495	49015	100%						0			0			1	1					1
Eukaryotic translation initiation factor 2 subunit 1 OS=Homo s	EIF2S1	sp|P05198	36095	100%						0			0	2	2	1	1					3
Eukaryotic translation initiation factor 2 subunit 2-like protein	A6NK07	sp|A6NK07,s	37770	81%						0			0	1	1							1
Eukaryotic translation initiation factor 2 subunit 3 OS=Homo s	EIF2S3	sp|P41091	51092	100%						0			0	1	1							1
Eukaryotic translation initiation factor 3 subunit A OS=Homo s	EIF3A	sp|Q14152	2E+05	100%						0			0	1	2			2	2			3
Eukaryotic translation initiation factor 3 subunit B OS=Homo s	EIF3B	sp|P55884	92465	81%						0			0	1	1							1
Eukaryotic translation initiation factor 3 subunit C OS=Homo s	EIF3C	sp|Q99613	1E+05	81%						0			0	1	1			1	1			2
Eukaryotic translation initiation factor 3 subunit E OS=Homo s	EIF3E	sp|P60228	52205	100%						0			0	2	1							2
Eukaryotic translation initiation factor 3 subunit F OS=Homo s	EIF3F	sp|O00303	37546	100%						0			0	1	3			2	2			3
Eukaryotic translation initiation factor 3 subunit H OS=Homo s	EIF3H	sp|O15372	39913	81%						0			0	1	1							1
Eukaryotic translation initiation factor 3 subunit K OS=Homo s	EIF3K	sp|Q9UBQ5	25042	81%						0			0	1	1							1
Eukaryotic translation initiation factor 3 subunit L OS=Homo s	EIF3L	sp|Q9Y262	66711	81%						0			0	1	1							1
Eukaryotic translation initiation factor 3 subunit M OS=Homo s	EIF3M	sp|Q7L2H7	42486	81%						0			0	1	1			1	1			2
Eukaryotic translation initiation factor 5A-1 OS=Homo sapiens	EIF5A	sp|P63241,s	16815	81%						0			0	1	1							1
Eukaryotic translation initiation factor 5B OS=Homo sapiens G	EIF5B	sp|O60841	1E+05	81%						0			0	1	1							1
Eukaryotic translation initiation factor 6 OS=Homo sapiens GN	EIF6	sp|P56537	26580		1	1				1			0									
F-actin-capping protein subunit beta OS=Homo sapiens GN=CA	CAPZB	sp|P47756	31334	100%						0			0	2	2	1	1					3
Fatty acid synthase OS=Homo sapiens GN=FASN PE=1 SV=3	FASN	sp|P49327	3E+05	100%						0			0	2	10			2	4	1	2	4
Fatty acid-binding protein, adipocyte OS=Homo sapiens GN=FA	FABP4	sp|P15090	14701	100%			3	2	1	1	4		0									
Fatty acid-binding protein, epidermal OS=Homo sapiens GN=F	FABP5	sp|Q01469	15146	81%	5	1	4	2			8		0	4	1			1	1			5
FCH and double SH3 domains protein 1 OS=Homo sapiens GN=FCHSD1 PE=1		sp|Q86WN1	76925	81%						0			0	1	1							1
Filaggrin OS=Homo sapiens GN=FLG PE=1 SV=3	FLG	sp|P20930	4E+05	81%			2	1			2		0									
Fructose-1,6-bisphosphatase 1 OS=Homo sapiens GN=FBP1 PE	FBP1	sp|P09467	36825	100%						0			0	1	2			1	1			2
Fructose-bisphosphate aldolase A OS=Homo sapiens GN=ALDO	ALDOA	sp|P04075	39403	81%						0			0	1	1							1
Fructose-bisphosphate aldolase C OS=Homo sapiens GN=ALDO	ALDOC	sp|P09972	39438	81%			1	1			1		0	1	1							1
Fumarate hydratase, mitochondrial OS=Homo sapiens GN=FH	FH	sp|P07954	54620	81%						0			0	1	1							1
Galectin-7 OS=Homo sapiens GN=LGALS7 PE=1 SV=2	LGALS7	sp|P47929	15057	100%	3	2	9	1			12	1	1	1	1	1		2	2			3
Galectin-related protein OS=Homo sapiens GN=GRP PE=1 SV=2	GRP	sp|Q3ZCW2	18968	81%						0			0	1	1							1
Gamma-glutamylcyclotransferase OS=Homo sapiens GN=GGC	GGCT	sp|O75223	20990	100%						0			0	2	4	1	1	3	2	1	1	6
Gasdermin-A OS=Homo sapiens GN=GSDMA PE=2 SV=4	GSDMA	sp|Q96QA5	49348	100%						0			0			1	1			2	1	3
Gelsolin OS=Homo sapiens GN=GSN PE=1 SV=1	GSN	sp|P06396	85680	100%	3	1	1	1			4		0									
General vesicular transport factor p115 OS=Homo sapiens GN	USO1	sp|O60763	1E+05	81%						0			0	2	1			1	1			3
Gigaxonin OS=Homo sapiens GN=GAN PE=1 SV=1	GAN	sp|Q9H2C0	67622	100%			3	2			3		0									
Glucose-6-phosphate isomerase OS=Homo sapiens GN=GPI PE	GPI	sp|P06744	63130	81%			1	1			1		0	1	2	1	1					2
Glutamine-dependent NAD(+) synthetase OS=Homo sapiens G	NADSYN1	sp|Q6IA69	79268	81%						0			0	1	1							1
Glyceraldehyde-3-phosphate dehydrogenase OS=Homo sapie	GAPDH	sp|P04406	36035	100%	5	2	1	1	1	1	7		0	1	3	1	2	3	1			5
Glycerophosphodiester phosphodiesterase domain-containir	GDPD3	sp|Q7L5L3	36580	81%	2	1					2		0	2	1			3	1			5
Glycyl-tRNA synthetase OS=Homo sapiens GN=GARS PE=1 SV=3	GARS	sp|P41250	83150	100%						0			0	1	3	1	1			1	1	3
Glyoxalase domain-containing protein 4 OS=Homo sapiens GN	GLOD4	sp|Q9HC38	34776	81%						0			0	1	1			1	1			2
GTP-binding nuclear protein Ran OS=Homo sapiens GN=RAN PE	RAN	sp|P62826	24405	81%						0			0	2	1					2	1	4
Guanine nucleotide-binding protein subunit beta-2-like 1 OS=	GNB2L1	sp|P63244	35059	81%			1	1			1		0	1	4	1	1			1	1	3
Heat shock 70 kDa protein 1A/1B OS=Homo sapiens GN=HSPA1	HSPA1A	sp|P08107	70036	100%						0			0	2	2			4	1			6
Heat shock 70 kDa protein 4 OS=Homo sapiens GN=HSPA4 PE=	HSPA4	sp|P34932	94314	81%						0			0	1	1							1
Heat shock 70 kDa protein 6 OS=Homo sapiens GN=HSPA6 PE=	HSPA6	sp|P17066	71011	100%	3	1					3		0									
Heat shock cognate 71 kDa protein OS=Homo sapiens GN=HSP	HSPA8	sp|P11142	70882	98%						0			0	1	1	2	2	1	3			4
Heat shock protein beta-1 OS=Homo sapiens GN=HSPB1 PE=1 S	HSPB1	sp|P04792	22765	100%	9	3					9		0			1	1					1
Heat shock protein HSP 90-alpha OS=Homo sapiens GN=HSP90	HSP90AA1	sp|P07900	84645	100%						0			0	3	4	1	3	2	4	1	1	7
Heat shock protein HSP 90-beta OS=Homo sapiens GN=HSP90A	HSP90AB1	sp|P08238	83249	100%						0			0	4	1							4
Heat shock-related 70 kDa protein 2 OS=Homo sapiens GN=HS	HSPA2	sp|P54652	70005	100%	4	3	3	3			7		0	1	2					1	3	2
Helicase-like transcription factor OS=Homo sapiens GN=HLTF	HLTF	sp|Q14527	1E+05	81%						0			0	1	1							1
Hephaestin-like protein 1 OS=Homo sapiens GN=HEPHL1 PE=2	HEPHL1	sp|Q6MZM0	1E+05	100%						0			0	1	2							1
Heterogeneous nuclear ribonucleoprotein D0 OS=Homo sapie	HNRNPD	sp|Q14103	38417	81%						0			0	1	1							1
Heterogeneous nuclear ribonucleoprotein K OS=Homo sapien	HNRNPK	sp|P61978	50961	81%						0			0	1	1			1	1			2
Heterogeneous nuclear ribonucleoprotein M OS=Homo sapier	HNRNPM	sp|P52272	77499	79%						0			0					1	1			1
Histone H2A type 1-B/E OS=Homo sapiens GN=HIST1H2AB PE=1	HIST1H2AB	sp|P04908,s	14118	100%	9	2					9	4	1	4								

Histone H2A type 2-C OS=Homo sapiens GN=HIST2H2AC PE=1 S	HIST2H2AC	sp|Q16777,s	14078	100%							0			0	1	1									1
Histone H2A.Z OS=Homo sapiens GN=H2AFZ PE=1 SV=2	H2AFZ	sp|P0C0S5,s	13491	92%	1	1					1			0	1	1									1
Histone H2B type 1-K OS=Homo sapiens GN=HIST1H2BK PE=1 S	HIST1H2BK	sp|O60814,s	13933	100%							0	10	1	10	2	1			4	1					6
Histone H4 OS=Homo sapiens GN=HIST1H4A PE=1 SV=2	HIST1H4A	sp|P62805	11350	100%	3	1					3	4	3	4	3	2	2	1	1	1	1	1	1	7
Homeobox protein Hox-B8 OS=Homo sapiens GN=HOXB8 PE=2	HOXB8	sp|P17481	27556	94%					2	1	2			0										
Huntingtin-interacting protein 1-related protein OS=Homo sap	HIP1R	sp|O75146	1E+05	100%							0			0	1	3								1
Ig gamma-1 chain C region OS=Homo sapiens GN=IGHG1 PE=1 S	IGHG1	sp|P01857	36087	100%	1	1	3	3			4			0										
Ig kappa chain C region OS=Homo sapiens GN=IGKC PE=1 SV=1	IGKC	sp|P01834	11591	100%	1	1					1			0										
Importin subunit beta-1 OS=Homo sapiens GN=KPNB1 PE=1 SV=	KPNB1	sp|Q14974	97153	100%							0			0	1	2								1
Inorganic pyrophosphatase 2, mitochondrial OS=Homo sapien	PPA2	sp|Q9H2U2	37903	81%							0			0	1	1								1
Interleukin enhancer-binding factor 2 OS=Homo sapiens GN=IL	ILF2	sp|Q12905	43045	81%							0			0	1	1			1	1				2
Involucrin OS=Homo sapiens GN=IVL PE=1 SV=2	IVL	sp|P07476	68457	98%	1	1					1			0										
Junction plakoglobin OS=Homo sapiens GN=JUP PE=1 SV=3	JUP	sp|P14923	81728	100%	2	5	7	3			9	5	1	5	1	8	1	1	3	4	1	2	6	
Keratin, type I cuticular Ha1 OS=Homo sapiens GN=KRT31 PE=1	KRT31	sp|Q15323	47219	100%	22	26	11	9	10	4	42	20	36	20	3	3								3
Keratin, type I cuticular Ha2 OS=Homo sapiens GN=KRT32 PE=1	KRT32	sp|Q14532	50324	100%	1	1					1	2	2	2	9	10	3	10	11	11	3	11	26	
Keratin, type I cuticular Ha3-I OS=Homo sapiens GN=KRT33A PI	KRT33A	sp|O76009	45922	100%	16	12			1		17	23	18	23	2	4								2
Keratin, type I cuticular Ha3-II OS=Homo sapiens GN=KRT33B P	KRT33B	sp|Q14525	46196	100%	36	7	24	11	8	7	68	28	10	28	2	3	1	1	2	1	2	1	7	
Keratin, type I cuticular Ha4 OS=Homo sapiens GN=KRT34 PE=1	KRT34	sp|O76011	44702	100%	18	11	11	11	4	8	33	17	14	17	3	4	3	1	3	1			9	
Keratin, type I cuticular Ha5 OS=Homo sapiens GN=KRT35 PE=1	KRT35	sp|Q92764	50343	100%	1	3	1	1			2	5	11	5	4	13	2	2	4	6	1	1	11	
Keratin, type I cuticular Ha6 OS=Homo sapiens GN=KRT36 PE=1	KRT36	sp|O76013	52229	100%	23	13	5	18	2	3	30			0	2	4	1	3	3	4	1	4	7	
Keratin, type I cuticular Ha7 OS=Homo sapiens GN=KRT37 PE=2	KRT37	sp|O76014	49729	100%			8	2			8			0										
Keratin, type I cuticular Ha8 OS=Homo sapiens GN=KRT38 PE=2	KRT38	sp|O76015	50462	100%							0	3	6	3										
Keratin, type I cytoskeletal 10 OS=Homo sapiens GN=KRT10 PE	KRT10	sp|P13645	58811	100%	4	6	2	2	2	4	7			0	3	8					1	2	4	
Keratin, type I cytoskeletal 12 OS=Homo sapiens GN=KRT12 PE	KRT12	sp|Q99456	53495	81%			1	1	2	1	3			0										
Keratin, type I cytoskeletal 14 OS=Homo sapiens GN=KRT14 PE	KRT14	sp|P02533	51545	100%	9	9	5	6	8	2	22			0										
Keratin, type I cytoskeletal 15 OS=Homo sapiens GN=KRT15 PE	KRT15	sp|P19012	49195	92%	1	1					1			0										
Keratin, type I cytoskeletal 16 OS=Homo sapiens GN=KRT16 PE	KRT16	sp|P08779	51251	100%	14	20	4	8	3	7	21			0										
Keratin, type I cytoskeletal 17 OS=Homo sapiens GN=KRT17 PE	KRT17	sp|Q04695	48089	100%	12	7	3	4	2	1	16			0										
Keratin, type I cytoskeletal 39 OS=Homo sapiens GN=KRT39 PE	KRT39	sp|Q6A163	55633	100%	4	6	9	2	2	4	15	2	7	2	2	13	2	3	3	5	1	3	8	
Keratin, type I cytoskeletal 40 OS=Homo sapiens GN=KRT40 PE	KRT40	sp|Q6A162	48121	100%							0			0	3	8	2	6	8	5	2	3	15	
Keratin, type I cytoskeletal 9 OS=Homo sapiens GN=KRT9 PE=1	KRT9	sp|P35527	62048	100%	3	4	1	1	1	1	5			0	2	5							2	
Keratin, type II cuticular Hb1 OS=Homo sapiens GN=KRT81 PE=	KRT81	sp|Q14533	54910	85%	48	2	7	3	1	1	56	32	3	32	1	1	1	1					2	
Keratin, type II cuticular Hb2 OS=Homo sapiens GN=KRT82 PE=	KRT82	sp|Q9NSB4	56635	100%			1	1			1	2	1	2	7	17	1	10	3	27	3	7	14	
Keratin, type II cuticular Hb3 OS=Homo sapiens GN=KRT83 PE=	KRT83	sp|P78385	54177	100%	22	6	1	1	2	2	24	21	6	21	1	3					1	1	2	
Keratin, type II cuticular Hb4 OS=Homo sapiens GN=KRT84 PE=	KRT84	sp|Q9NSB2	64825	100%	15	20	10	9	5	6	30			0					1	1			1	
Keratin, type II cuticular Hb5 OS=Homo sapiens GN=KRT85 PE=	KRT85	sp|P78386	55784	100%	29	35	12	10	10	4	50	32	26	32	3	10	3	2	4	8	4	2	13	
Keratin, type II cuticular Hb6 OS=Homo sapiens GN=KRT86 PE=	KRT86	sp|O43790	53482	100%	47	21			7	3	54	74	43	74	3	7			2	5			5	
Keratin, type II cytoskeletal 1 OS=Homo sapiens GN=KRT1 PE=1	KRT1	sp|P04264	66022	100%	9	7	4	3	3	5	15	3	1	3	2	7			2	4			4	
Keratin, type II cytoskeletal 2 epidermal OS=Homo sapiens GN	KRT2	sp|P35908	65416	100%	1	4					1			0	1	4							1	
Keratin, type II cytoskeletal 3 OS=Homo sapiens GN=KRT3 PE=1	KRT3	sp|P12035	64401	99%	1	1					1			0										
Keratin, type II cytoskeletal 5 OS=Homo sapiens GN=KRT5 PE=1	KRT5	sp|P13647	62362	100%	19	20	15	14	10	4	44			0										
Keratin, type II cytoskeletal 6A OS=Homo sapiens GN=KRT6A PI	KRT6A	sp|P02538	60028	100%	12	14	2	2	2	1	15	1	1	1										
Keratin, type II cytoskeletal 6B OS=Homo sapiens GN=KRT6B PI	KRT6B	sp|P04259	60051	100%	29	1	16	2			45			0										
Keratin, type II cytoskeletal 6C OS=Homo sapiens GN=KRT6C PI	KRT6C	sp|P48668	60009	81%			11	1			11			0										
Keratin, type II cytoskeletal 7 OS=Homo sapiens GN=KRT7 PE=1	KRT7	sp|P08729	51369								0	18	2	18										
Keratin, type II cytoskeletal 80 OS=Homo sapiens GN=KRT80 PI	KRT80	sp|Q6KB66	50508	100%	6	4	3	1	1	2	10	2	1	2	2	8	1	1	1	3	1	1	5	
Keratin-associated protein 10-1 OS=Homo sapiens GN=KRTAP1	KRTAP10-1	sp|P60331,s	28638	81%							0			0	18	1					2	2	20	
Keratin-associated protein 10-10 OS=Homo sapiens GN=KRTAI	KRTAP10-10	sp|P60014	25549	100%							0			0	1	3			1	1			2	
Keratin-associated protein 10-11 OS=Homo sapiens GN=KRTAI	KRTAP10-11	sp|P60412	30221	100%							0			0	5	1							5	
Keratin-associated protein 10-3 OS=Homo sapiens GN=KRTAP1	KRTAP10-3	sp|P60369	22327	100%							0			0					1	2			1	
Keratin-associated protein 10-5 OS=Homo sapiens GN=KRTAP1	KRTAP10-5	sp|P60370	27604	99%							0			0					1	1			1	
Keratin-associated protein 10-7 OS=Homo sapiens GN=KRTAP1	KRTAP10-7	sp|P60409	37792	100%							0			0					1	1			1	
Keratin-associated protein 10-9 OS=Homo sapiens GN=KRTAP1	KRTAP10-9	sp|P60411	30015	100%							0			0					1	2			1	
Keratin-associated protein 1-1 OS=Homo sapiens GN=KRTAP1-	KRTAP1-1	sp|Q07627	18215	100%			6	1			6	5	2	5					1	1	2	1	3	
Keratin-associated protein 11-1 OS=Homo sapiens GN=KRTAP1	KRTAP11-1	sp|Q8IUC1	17066	100%	12	4	3	2			15	15	4	15										
Keratin-associated protein 12-1 OS=Homo sapiens GN=KRTAP1	KRTAP12-1	sp|P59990	9718	81%							0			0	6	1	1	1	2	2			9	
Keratin-associated protein 12-2 OS=Homo sapiens GN=KRTAP1	KRTAP12-2	sp|P59991	14669	100%							0			0					1	1	1	1	2	
Keratin-associated protein 12-3 OS=Homo sapiens GN=KRTAP1	KRTAP12-3	sp|P60328	9927	100%							0			0					2	1			2	
Keratin-associated protein 13-1 OS=Homo sapiens GN=KRTAP1	KRTAP13-1	sp|Q8IUC0	18302	100%	4	2	8	5			12	1	2	1										
Keratin-associated protein 13-2 OS=Homo sapiens GN=KRTAP1	KRTAP13-2	sp|Q52LG2	18709	100%	12	5	18	2			29	11	7	11										
Keratin-associated protein 13-4 OS=Homo sapiens GN=KRTAP1	KRTAP13-4	sp|Q3LI77	17737	100%	3	1	1	1			4			0										
Keratin-associated protein 1-5 OS=Homo sapiens GN=KRTAP1-	KRTAP1-5	sp|Q9BYS1	17991	100%							0	1	1	1										
Keratin-associated protein 19-4 OS=Homo sapiens GN=KRTAP1	KRTAP19-4	sp|Q3LI73	9089	81%							0			0	1	1							1	
Keratin-associated protein 19-5 OS=Homo sapiens GN=KRTAP1	KRTAP19-5	sp|Q3LI72	7607	100%							0	5	2	5										
Keratin-associated protein 2-4 OS=Homo sapiens GN=KRTAP2-	KRTAP2-4	sp|Q9BYR9	13460	100%							0	3	6	3										
Keratin-associated protein 24-1 OS=Homo sapiens GN=KRTAP2	KRTAP24-1	sp|Q3LI83	27701	100%							0			0	4	3	2	1	3	3	3	2	11	
Keratin-associated protein 26-1 OS=Homo sapiens GN=KRTAP2	KRTAP26-1	sp|Q6PEX3	22536	100%							0			0	3	2	2	4	2	13	1	7	8	
Keratin-associated protein 28 family		sp|X0X281,s	8588	81%							0			0	2	1							2	
Keratin-associated protein 3-1 OS=Homo sapiens GN=KRTAP3-	KRTAP3-1	sp|Q9BYR8	10521	100%	29	1	7	3			36	17	3	17	1	1							1	
Keratin-associated protein 3-2 OS=Homo sapiens GN=KRTAP3-	KRTAP3-2	sp|Q9BYR7	10389	100%	21	1	5	4			26	26	1	26					2	2	4	1	6	

Annexe 5 : Protéines identifiées dans les extraits cuticulaires, corticaux et unguéaux

Keratin-associated protein 3-3 OS=Homo sapiens GN=KRTAP3-	KRTAP3-3	sp|Q9BYR6	10347	100%	25	1			4	1	29	15	1	15								
Keratin-associated protein 4-1 OS=Homo sapiens GN=KRTAP4-	KRTAP4-1	sp|Q9BYQ7	13221	100%							0	1	1	1								
Keratin-associated protein 4-11 OS=Homo sapiens GN=KRTAP4	KRTAP4-11	sp|Q9BYQ6	20906	100%					1	1	1	1	3	1								
Keratin-associated protein 4-12 OS=Homo sapiens GN=KRTAP4	KRTAP4-12	sp|Q9BQ66,s	21386	100%							0	5*	3*	5*								
Keratin-associated protein 4-2 OS=Homo sapiens GN=KRTAP4-	KRTAP4-2	sp|Q9BYR5	14442	100%							0	1	1	1								
Keratin-associated protein 4-3 OS=Homo sapiens GN=KRTAP4-	KRTAP4-3	sp|Q9BYR4	20484	100%	1	1					1	3	3	3								
Keratin-associated protein 4-4 OS=Homo sapiens GN=KRTAP4-	KRTAP4-4	sp|Q9BYR3	18002	100%							0	2	1	2								
Keratin-associated protein 4-7 OS=Homo sapiens GN=KRTAP4-	KRTAP4-7	sp|Q9BYR0	22469	94%					1	1	1			0								
Keratin-associated protein 4-9 (Corrected) OS=Homo sapiens	KRTAP4-9	sp|Q9BYQ8	20569	100%	2	1					2			0								
Keratin-associated protein 6-1 OS=Homo sapiens GN=KRTAP6-	KRTAP6-1	sp|Q3LI64	7261	100%	1	1					1	2	1	2								
Keratin-associated protein 9-2 OS=Homo sapiens GN=KRTAP9-	KRTAP9-2	sp|Q9BYQ4	18267	100%							0	8	3	8								
Keratin-associated protein 9-3 OS=Homo sapiens GN=KRTAP9-	KRTAP9-3	sp|Q9BYQ3	16833	100%							0	8	1	8								
Keratin-associated protein 9-6 OS=Homo sapiens GN=KRTAP9-	KRTAP9-6	sp|A8MVA2	16780	100%							0	1	1	1								
Keratin-associated protein 9-7 OS=Homo sapiens GN=KRTAP9-	KRTAP9-7	sp|A8MTY7,s	18267	100%	9	1					9	2	1	2					2	1	1	5	3
Leucine-rich repeat-containing protein 15 OS=Homo sapiens G	LRRC15	sp|Q8TF66	64380	100%	2	4	9	12	2	2	14			0	1	5	1	5					2
L-lactate dehydrogenase A chain OS=Homo sapiens GN=LDHA	LDHA	sp|P00338	36671	85%	1	1					1			0			1	1	1	1			2
L-lactate dehydrogenase B chain OS=Homo sapiens GN=LDHB	LDHB	sp|P07195	36621	100%							0			0	1	4	1	1	2	1			4
Long-chain fatty acid transport protein 6 OS=Homo sapiens GN	SLC27A6	sp|Q9Y2P4	70096	79%							0			0									
Low molecular weight phosphotyrosine protein phosphatase	ACP1	sp|P24666	18025	81%							0			0	1	1			1	1			2
L-xylulose reductase OS=Homo sapiens GN=DCXR PE=1 SV=2	DCXR	sp|Q7Z4W1	25894	79%							0			0							1	1	1
Lysophospholipid acyltransferase 5 OS=Homo sapiens GN=LPC	LPCAT3	sp|Q6P1A2	56020	88%							0			0									
Lysosome-associated membrane glycoprotein 1 OS=Homo sap	LAMP1	sp|P11279	44865	100%							0	1	1	1					2	3			2
Lysozyme g-like protein 2 OS=Homo sapiens GN=LYG2 PE=1 SV=	LYG2	sp|Q86SG7	23481	100%							0			0	10	2	1	1	1	1	1	1	13
Macrophage migration inhibitory factor OS=Homo sapiens GN=	MIF	sp|P14174	12459	81%							0			0	2	1			3	1			5
Malate dehydrogenase, mitochondrial OS=Homo sapiens GN=	MDH2	sp|P40926	35486								0	1	1	1	2	4							2
Minor histocompatibility antigen H13 OS=Homo sapiens GN=H	HM13	sp|Q8TCT9	41473	81%							0			0	1	1							1
Molybdenum cofactor biosynthesis protein 1 OS=Homo sapien	MOCS1	sp|Q9NZB8	70088								0	1	1	1							1	1	1
Motile sperm domain-containing protein 2 OS=Homo sapiens	MOSPD2	sp|Q8NHP6	59730	85%							0			0			1	1					1
Myosin regulatory light chain 12B OS=Homo sapiens GN=MYL1	MYL12B	sp|O14950,s	19777	81%							0			0	1	1							1
Myosin-14 OS=Homo sapiens GN=MYH14 PE=1 SV=1	MYH14	sp|Q7Z406	2E+05	99%					1	1	1			0									
Myosin-15 OS=Homo sapiens GN=MYH15 PE=1 SV=5	MYH15	sp|Q9Y2K3	2E+05	94%					1	1	1			0									
Nascent polypeptide-associated complex subunit alpha OS=H	NACA	sp|Q13765	23365	81%							0			0	1	1							1
Neuroblast differentiation-associated protein AHNAK OS=Hom	AHNAK	sp|Q09666	6E+05	100%	2	4					2			0									
Neutrophil defensin 1 OS=Homo sapiens GN=DEFA1 PE=1 SV=1	DEFA1	sp|P59665,s	10183	81%							0			0	1	1							1
NFU1 iron-sulfur cluster scaffold homolog, mitochondrial OS=H	NFU1	sp|Q9UMS0	28445	94%					1	1	1			0									
Niemann-Pick C1 protein OS=Homo sapiens GN=NPC1 PE=1 SV=	NPC1	sp|O15118	1E+05	81%							0			0	1	1							1
Nuclear receptor corepressor 1 OS=Homo sapiens GN=NCOR1	NCOR1	sp|O75376,s	1E+05								0	3	1	3									
Nuclear transport factor 2 OS=Homo sapiens GN=NUTF2 PE=1 S	NUTF2	sp|P61970	14461	85%							0			0			1	1					1
Nucleosome assembly protein 1-like 1 OS=Homo sapiens GN=I	NAP1L1	sp|P55209	45357	81%							0			0	1	1					1	1	2
Peptidyl-prolyl cis-trans isomerase A OS=Homo sapiens GN=PF	PPIA	sp|P62937	17995	100%			1	2			1			0	3	4	2	1	2	1			7
Peptidyl-prolyl cis-trans isomerase B OS=Homo sapiens GN=PF	PPIB	sp|P23284	23725	100%							0			0	2	2	1	1					3
Peroxiredoxin-1 OS=Homo sapiens GN=PRDX1 PE=1 SV=1	PRDX1	sp|Q06830	22093	81%							0			0	1	1			1	1			2
Peroxiredoxin-2 OS=Homo sapiens GN=PRDX2 PE=1 SV=5	PRDX2	sp|P32119	21874	100%							0			0	2	2							2
Peroxiredoxin-6 OS=Homo sapiens GN=PRDX6 PE=1 SV=3	PRDX6	sp|P30041	25018	100%	2	5			2	1	4			0	1	2			3	2			4
Peroxisomal multifunctional enzyme type 2 OS=Homo sapiens	HSD17B4	sp|P51659	79670	100%							0			0	3	2	1	2	1	2	1	1	6
Phosphatidylethanolamine-binding protein 1 OS=Homo sapie	PEBP1	sp|P30086	21039	100%							0			0	2	1	1	2	1	1			4
Phosphoglycerate kinase 1 OS=Homo sapiens GN=PGK1 PE=1 S	PGK1	sp|P00558	44597	100%	1	1					1			0	2	4			3	2	1	1	5
Plakophilin-1 OS=Homo sapiens GN=PKP1 PE=1 SV=2	PKP1	sp|Q13835	82845	100%	2	2	1	2			3			0	2	2	1	1			1	1	4
Plakophilin-3 OS=Homo sapiens GN=PKP3 PE=1 SV=1	PKP3	sp|Q9Y446	87067	100%	3	2	2	2			5			0					2	1			2
Plectin OS=Homo sapiens GN=PLEC PE=1 SV=3	PLEC	sp|Q15149	5E+05	100%							0			0	1	5			1	1	1	1	3
Poly(rC)-binding protein 1 OS=Homo sapiens GN=PCBP1 PE=1 SV	PCBP1	sp|Q15365	37480	100%							0			0	2	2							2
Poly(U)-specific endoribonuclease OS=Homo sapiens GN=END	ENDOU	sp|P21128	46854	92%	3	1					3			0					1	1	1	1	2
Polyadenylate-binding protein 1 OS=Homo sapiens GN=PABPC	PABPC1	sp|P11940	70653	100%							0			0	1	5	1	1					2
Polyubiquitin-B OS=Homo sapiens GN=UBB PE=1 SV=1	UBB	sp|P0CG47,s	14711	92%	1	1					1			0	1	1	1	1					2
Prelamin-A/C OS=Homo sapiens GN=LMNA PE=1 SV=1	LMNA	sp|P02545	74123	100%	2	2					2			0	2	2			1	1			3
Programmed cell death 6-interacting protein OS=Homo sapien	PDCD6IP	sp|Q8WUM4	96007	81%			1	1			1			0			1	1					1
Proliferation-associated protein 2G4 OS=Homo sapiens GN=PA	PA2G4	sp|Q9UQ80	43769	100%							0			0	1	1			3	1	2	2	6
Prolyl endopeptidase OS=Homo sapiens GN=PREP PE=1 SV=2	PREP	sp|P48147	80684	100%							0			0	2	2	2	2					3
Prostatic acid phosphatase OS=Homo sapiens GN=ACPP PE=1 S	ACPP	sp|P15309	44550	81%							0			0	1	1			1	1	1	1	3
Proteasome subunit beta type-7 OS=Homo sapiens GN=PSMB7	PSMB7	sp|Q99436	29948	85%							0			0	1	1	1	1					2
Protein arginine N-methyltransferase 6 OS=Homo sapiens GN=	PRMT6	sp|Q96LA8	41919	92%										0	1	1							1
Protein disulfide-isomerase A3 OS=Homo sapiens GN=PDIA3 P	PDIA3	sp|P30101	56767	100%							0			0	2	2	2	2	1	1			5
Protein disulfide-isomerase A4 OS=Homo sapiens GN=PDIA4 P	PDIA4	sp|P13667	72916	79%							0			0					1	2			1
Protein disulfide-isomerase A6 OS=Homo sapiens GN=PDIA6 P	PDIA6	sp|Q15084	48104	81%							0			0	1	1	1	1	1	1			3
Protein disulfide-isomerase OS=Homo sapiens GN=P4HB PE=1	P4HB	sp|P07237	57100	81%							0			0	2	1							2
Protein dopey-2 OS=Homo sapiens GN=DOPEY2 PE=1 SV=5	DOPEY2	sp|Q9Y3R5	3E+05	81%							0			0	1	1			1	1			2
Protein FAM26D OS=Homo sapiens GN=FAM26D PE=1 SV=1	FAM26D	sp|Q5JW98	35043	100%							0			0	1	2			2	1	2	1	5
Protein FAM83H OS=Homo sapiens GN=FAM83H PE=1 SV=3	FAM83H	sp|Q6ZRV2	1E+05	79%							0			0									
Protein RIC-3 OS=Homo sapiens GN=RIC3 PE=1 SV=1	RIC3	sp|Q7Z5B4	41075	85%			2	1			2			0			1	1					1
Protein S100-A14 OS=Homo sapiens GN=S100A14 PE=1 SV=1	S100A14	sp|Q9HCY8	11644	92%	2	1					2			0					5	3	1	3	6

Protein S100-A3 OS=Homo sapiens GN=S100A3 PE=1 SV=1	S100A3	sp|P33764	11695	100%					1	1	1	5	1	5	17	1							17
Protein S100-A4 OS=Homo sapiens GN=S100A4 PE=1 SV=1	S100A4	sp|P26447	11711	81%							0			0	1	1							1
Protein S100-A6 OS=Homo sapiens GN=S100A6 PE=1 SV=1	S100A6	sp|P06703	10162	100%							0			0	4	2							4
Protein transport protein Sec23A OS=Homo sapiens GN=SEC23	SEC23A	sp|Q15436	86145	81%							0			0	1	1			2	1			3
Protein transport protein Sec23B OS=Homo sapiens GN=SEC23	SEC23B	sp|Q15437	86463	81%							0			0	2	1			1	1			3
Protein unc-45 homolog A OS=Homo sapiens GN=UNC45A PE=1	UNC45A	sp|Q9H3U1	1E+05	79%							0			0					2	7	1	1	3
Protein-arginine deiminase type-3 OS=Homo sapiens GN=PADI	PADI3	sp|Q9ULW8	74725	100%							0			0	2	4			1	5	2	3	5
Protein-glutamine gamma-glutamyltransferase E OS=Homo sa	TGM3	sp|Q08188	76615	100%	3	4	3	5	4	1	11			0	4	8	1	6	2	3	1	2	7
Protein-glutamine gamma-glutamyltransferase K OS=Homo sa	TGM1	sp|P22735	89768	100%							0			0	2	3	1	3					3
Proteolipid protein 2 OS=Homo sapiens GN=PLP2 PE=1 SV=1	PLP2	sp|Q04941	16673	81%							0			0	2	1							2
Putative annexin A2-like protein OS=Homo sapiens GN=ANXA2	ANXA2P2	sp|A6NMY6,s	38643						1	1	1	1	1	1	2	1							2
Putative basic leucine zipper and W2 domain-containing prot	BZW1L1	sp|P0C7T9,s	31924	81%							0			0	2	1			3	10	1	2	6
Putative keratin-associated protein 10-like ENSP0000037514	A8MUX0	sp|A8MUX0	53889	100%							0			0	5	7			1	3	1	1	7
Putative keratin-associated protein 10-like ENSP0000037514	A8MX34	sp|A8MX34	35190	100%							0			0					1	4			1
Pyruvate kinase isozymes M1/M2 OS=Homo sapiens GN=PKM2	PKM2	sp|P14618	57920	100%	3	2			1	1	4			0	2	5					2	1	4
Rab GDP dissociation inhibitor beta OS=Homo sapiens GN=GD	GDI2	sp|P50395	50648	81%			1	1			1			0	1	2	1	1	5	1			7
Ras GTPase-activating protein 2 OS=Homo sapiens GN=RASA2	RASA2	sp|Q15283	96511	79%							0			0					2	1			2
Ras-related protein Rab-13 OS=Homo sapiens GN=RAB13 PE=1	RAB13	sp|P51153,s	23007	79%							0			0									
Ras-related protein Rab-14 OS=Homo sapiens GN=RAB14 PE=1	RAB14	sp|P61106	23880	81%							0			0	2	1							2
Ras-related protein Rab-1B OS=Homo sapiens GN=RAB1B PE=1	RAB1B	sp|Q9H0U4	22154	100%							0			0	2	2							2
Regulation of nuclear pre-mRNA domain-containing protein 2	RPRD2	sp|Q5VT52,s	2E+05	94%					1	1	1			0							1	1	1
Reticulon-4 OS=Homo sapiens GN=RTN4 PE=1 SV=2	RTN4	sp|Q9NQC3	1E+05	88%							0			0									
Retinal dehydrogenase 1 OS=Homo sapiens GN=ALDH1A1 PE=1	ALDH1A1	sp|P00352	54845	99%	3	1					3			0							1	1	1
Ribonuclease inhibitor OS=Homo sapiens GN=RNH1 PE=1 SV=2	RNH1	sp|P13489	49956	92%	2	1					2			0	1	1			1	1			2
Ribonuclease T2 OS=Homo sapiens GN=RNASET2 PE=1 SV=2	RNASET2	sp|O00584	29463	79%							0			0									
rRNA/tRNA 2'-O-methyltransferase fibrillarin-like protein 1 OS=	FBLL1	sp|A6NHQ2,s	34657	81%							0			0	1	1							1
Selenium-binding protein 1 OS=Homo sapiens GN=SELENBP1 PI	SELENBP1	sp|Q13228	52374	100%	6	6	14	4	1	1	20	3	4	3	1	1							1
Serine/threonine-protein kinase receptor R3 OS=Homo sapien	ACVRL1	sp|P37023	56106	92%	1	1					1			0					1	1			1
Serine/threonine-protein phosphatase 2A 65 kDa regulatory s	PPP2R1B	sp|P30154	66197	98%							0			0	3	1			1	1			4
Serine/threonine-protein phosphatase 2A activator OS=Homo	PPP2R4	sp|Q15257	40650	100%							0			0	1	1							1
Serpin B12 OS=Homo sapiens GN=SERPINB12 PE=1 SV=1	SERPINB12	sp|Q96P63	46260	92%	2	1					2			0									
Serpin B13 OS=Homo sapiens GN=SERPINB13 PE=2 SV=2	SERPINB13	sp|Q9UIV8	44259	81%			2	1			2			0									
Serpin B3 (B4) OS=Homo sapiens GN=SERPINB3 PE=1 SV=2	SERPINB3	sp|P29508,s	44548	94%					1	1	1			0					1	1	1	1	2
Serpin B5 OS=Homo sapiens GN=SERPINB5 PE=1 SV=2	SERPINB5	sp|P36952	42084	100%	4	1	3	3			7			0	3	1	1	1					4
Serpin B7 OS=Homo sapiens GN=SERPINB7 PE=2 SV=1	SERPINB7	sp|O75635	42888	92%	2	1	1				3			0									
Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2	ALB	sp|P02768	69349	81%			1	1			1			0	1	1			1	1			2
Seryl-tRNA synthetase, cytoplasmic OS=Homo sapiens GN=SAR	SARS	sp|P49591	58761	79%							0			0					1	1	1	1	2
Sialidase-2 OS=Homo sapiens GN=NEU2 PE=1 SV=2	NEU2	sp|Q9Y3R4	42235	100%							0	3	1	3									
Small nuclear ribonucleoprotein Sm D3 OS=Homo sapiens GN=	SNRPD3	sp|P62318	13899	81%							0			0	3	1							3
Sodium/potassium/calcium exchanger 6 OS=Homo sapiens GN	SLC24A6	sp|Q6J4K2,s	1E+05	81%			1	1			1			0	2	1	2	1	1	1			5
Solute carrier family 35 member F4 OS=Homo sapiens GN=SLC	SLC35F4	sp|A4IF30	57793	79%							0			0									
Sterol-4-alpha-carboxylate 3-dehydrogenase, decarboxylating	NSDHL	sp|Q15738	41883	98%							0			0	1	1			1	1			2
Thioredoxin OS=Homo sapiens GN=TXN PE=1 SV=3	TXN	sp|P10599	11719	100%	2	2	6	1			8			0	3	2	1	1	2	2	1	3	7
Thioredoxin reductase 1, cytoplasmic OS=Homo sapiens GN=T	TXNRD1	sp|Q16881	70889	100%							0			0	2	4	1	4			1	1	5
Transcription factor BTF3 OS=Homo sapiens GN=BTF3 PE=1 SV=	BTF3	sp|P20290,s	23356	85%							0			0	1	1	1	1					2
Transgelin-2 OS=Homo sapiens GN=TAGLN2 PE=1 SV=3	TAGLN2	sp|P37802	22374	100%	3	2					3			0					1	4	1	1	2
Transitional endoplasmic reticulum ATPase OS=Homo sapiens	VCP	sp|P55072	89307	100%							0			0	2	6	1	2					3
Transketolase OS=Homo sapiens GN=TKT PE=1 SV=3	TKT	sp|P29401	67861	81%							0			0	2	1							2
Translation initiation factor eIF-2Bsubunit delta OS=Homo sa	EIF2B4	sp|Q9UI10	57539	100%							0			0	1	3							1
Translational activator GCN1 OS=Homo sapiens GN=GCN1L1 P	GCN1L1	sp|Q92616	3E+05	85%							0			0			1	1					1
Transmembrane emp24 domain-containing protein 9 OS=Hom	TMED9	sp|Q9BVK6	27260	81%							0			0	1	1							1
Transmembrane protein 144 OS=Homo sapiens GN=TMEM144	TMEM144	sp|Q7Z5S9	37637	81%			1	1			1			0	1	1							1
Trifunctional enzyme subunit alpha, mitochondrial OS=Homo s	HADHA	sp|P40939	82984	81%	1	1					1			0									
Triosephosphate isomerase OS=Homo sapiens GN=TPI1 PE=1 S	TPI1	sp|P60174	26651	100%	2	1			1	1	3			0	2	4			2	2	5	1	8
Tubulin alpha-1B chain OS=Homo sapiens GN=TUBA1B PE=1 SV=	TUBA1B	sp|P68363,s	50118				5	1			5	1	1	1	1	2	2	2	1	2	2	1	6
Tubulin alpha-4A chain OS=Homo sapiens GN=TUBA4A PE=1 SV=	TUBA4A	sp|P68366	49907	100%	2	1	12	1			14			0	1	1							1
Tubulin beta chain OS=Homo sapiens GN=TUBB PE=1 SV=2	TUBB	sp|P07437	49653	100%	1	1	1	1			2			0					3	6	1	6	4
Tubulin beta-2A chain OS=Homo sapiens GN=TUBB2A PE=1 SV=	TUBB2A	sp|Q13885,s	49889	100%			4	2			4			0	3	5	1	2					4
Tumor protein D52 OS=Homo sapiens GN=TPD52 PE=1 SV=2	TPD52	sp|P55327	24310	92%	1	1					1			0	1	1							1
Tumor protein D54 OS=Homo sapiens GN=TPD52L2 PE=1 SV=2	TPD52L2	sp|O43399	22220	81%							0			0	1	1							1
Ubiquitin carboxyl-terminal hydrolase 2 OS=Homo sapiens GN	USP2	sp|O75604,s	68056				3	1			3	1	1	1							1	1	1
Ubiquitin thioesterase OTUB2 OS=Homo sapiens GN=OTUB2 PE	OTUB2	sp|Q96DC9	27196	85%							0			0			1	1			1	1	2
Ubiquitin-conjugating enzyme E2 D1 OS=Homo sapiens GN=UB	UBE2D1	sp|P51668,s	16585	88%							0			0							1	1	1
Ubiquitin-conjugating enzyme E2 E1 OS=Homo sapiens GN=UB	UBE2E1	sp|P51965,s	16718	88%							0			0									
Ubiquitin-conjugating enzyme E2 K OS=Homo sapiens GN=UBE	UBE2K	sp|P61086	22389	81%							0			0	1	1			1	1			2
Ubiquitin-like modifier-activating enzyme 1 OS=Homo sapiens	UBA1	sp|P22314	1E+05	79%							0			0									
UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransfera	B3GNT3	sp|Q6UX72	43735	81%							0			0	1	1							1
Uncharacterized protein C3orf62 OS=Homo sapiens GN=C3orf62 PE=2 SV=1		sp|Q6ZUJ4	30177	81%							0			0	1	1							1
Uncharacterized protein C6orf47 OS=Homo sapiens GN=C6orf47 PE=1 SV=2		sp|O95873,s	31693								0	2	1	2									
UV excision repair protein RAD23 homolog B OS=Homo sapiens	RAD23B	sp|P54727	43153	81%							0			0	2	1							2

Annexe 5 : Protéines identifiées dans les extraits cuticulaires, corticaux et unguéaux

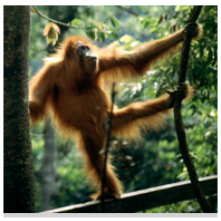
Vacuolar protein sorting-associated protein 35 OS=Homo sapi	VPS35	sp|Q96QK1	91692	85%							0			0			1	1			1	1	2
Valyl-tRNA synthetase OS=Homo sapiens GN=VARS PE=1 SV=4	VARS	sp|P26640	1E+05	88%							0			0							1	1	1
Very long-chain specific acyl-CoA dehydrogenase, mitochondri	ACADVL	sp|P49748	70374	88%							0			0									
Vesicle-fusing ATPase OS=Homo sapiens GN=NSF PE=1 SV=3	NSF	sp|P46459	82579	99%							0			0	1	1							1
Voltage-dependent anion-selective channel protein 3 OS=Hom	VDAC3	sp|Q9Y277	30642	81%							0			0	1	1			2	5	2	9	4
V-set and immunoglobulin domain-containing protein 8 OS=Ho	VSIG8	sp|Q5VU13	43872	100%	2	2	10	5	2	3	14	1	1	1	3	3	2	9	1	1			7
V-type proton ATPase catalytic subunit A OS=Homo sapiens GN	ATP6V1A	sp|P38606	68287	100%							0			0	1	1							1
V-type proton ATPase subunit B, brain isoform OS=Homo sapie	ATP6V1B2	sp|P21281	56484	100%							0			0	1	3							1
V-type proton ATPase subunit E 1 OS=Homo sapiens GN=ATP6V	ATP6V1E1	sp|P36543	26128	81%							0			0	1	1							1
keratin-associated protein 2 family	A8MTN3	sp|A8MTN3,s	13494	100%	14	1					14	7	1	7					1	2			1

Annexe 6 : Les motifs penta peptidiques des séquences des KAP 4 de mammifères dont le génome a été séquencé

(Pan Troglodytes, Pongo Abelii, Macaca Mulatta, Rattus Norvegicus, Mus Musculus, Bos Taurus, Equus Cabalus, Monodelphis Domestica, Ornithorhynchus Anatinus).

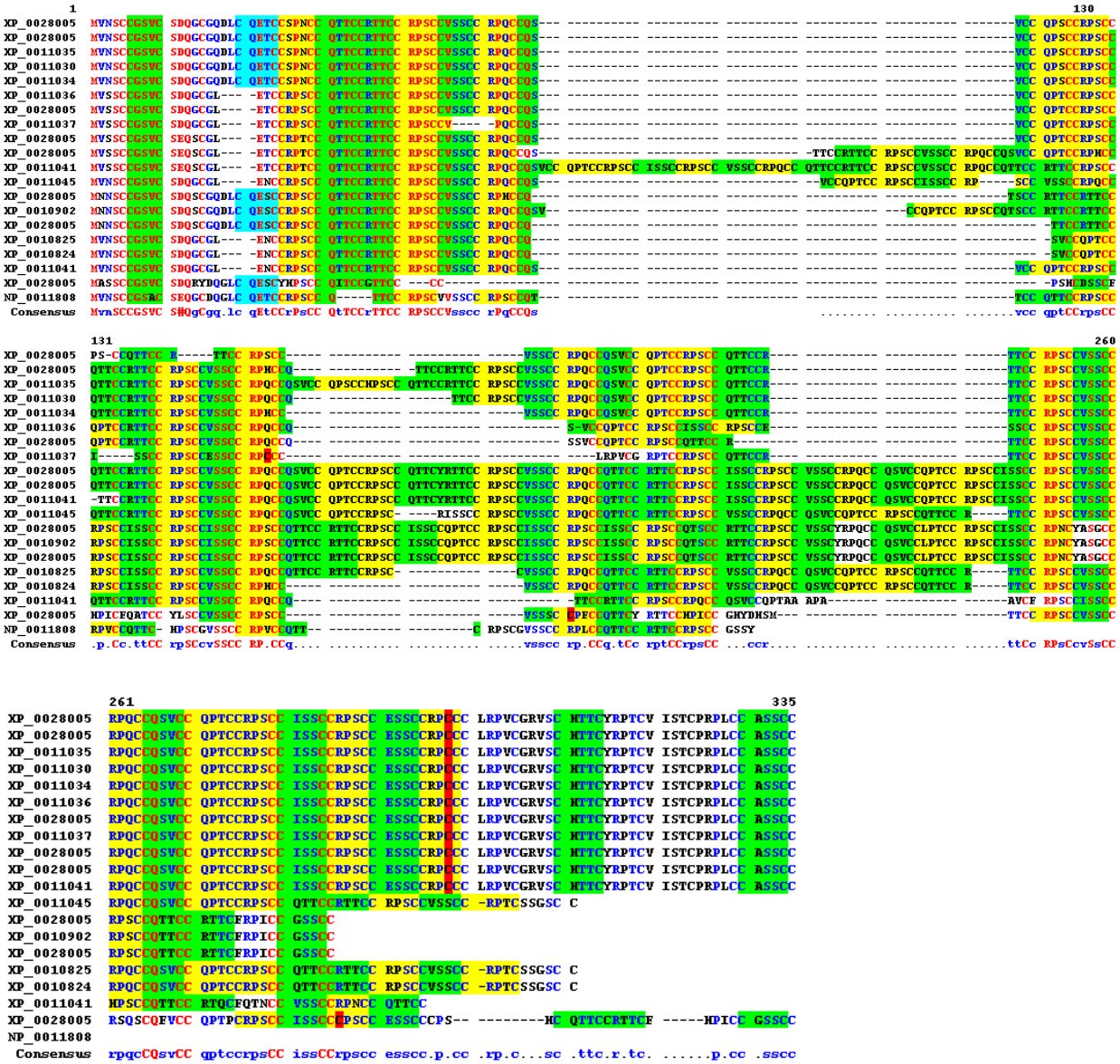
Source NCBI, séquences identifiées après Blast de séquences de KAP 4 humaines.

Pan troglodytes (Chimpanzé)



Pongo Abelii (Orang Outan)

Macaca mulatta (Macaque)



Rattus norvegicus (Rat)

```
                                                                                                                                                130
             1
XP_577126.                                 MVN SCCGSVCSEQ GCDQSPCQES CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS C-QTTCCRPV CCQTTC---- --RPSCGVSS CCRPVCCQTT C-RPSCGVSS
XP_0010561                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCV------- ---------- CCVSSCCRPQ CCQSVCCQPT CCRPSCCQAT CCRTTCCRPS
XP_0027245                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCVSSCCRPS CCVSSCCRPQ CCQSVCCQPT CCRPSCCQAT CCRTTCCRPS
XP_0010564                                 MVS SHCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCVSSCCRPQ CCQSVCCQPT CCRPSCCISS CCRPSCCVSS CCRPSCCISS
EDM05991.1                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCQSLCCRPS CCISSCCRPP CCRPSCCISS CCRPSCCRPT CCISSCCRPS
XP_0027246                                 MVN SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCQSVCCQPT CCRPSCCISS CCRPSCCRPS
XP_0027246                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCQSLCCRPS CCISSCCRPP CCRPSCCISS CCRPSCCRPS CCISSCCRPS
XP_0027246                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCQSFCCRPS CC-------- --RPTCCISS CCRPSCCRPS CCISSCCRPS
EDM05993.1                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPQ CCQSVCCQPT CCRPTCCRPS CCRPSCCVSS CCRPSCCVSS CCRPSCCTSS
XP_0027246                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCVSSCCRPS CCVS------ ---------S CCRPSCCVSS CCRPQCCQSV
XP_0010560                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPQ CCQSVCCQPT CCRPSCCRPS CCISSCCRPS CCRPSCCVSS CCRPQCCQSV
XP_0027246                                 MVN SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCISSCCRPQ CCQSVCCQPT CC-----RPS CCRPSCCVSS CCRPQCCQSV
EDM05990.1                                 MVS SCCGSVCSEE GCGQG----- CCQPSCCQTT CCRTTCCRPS CCVSSCCRPS CCVS------ ---------- ---------S CCRPSCCVSS CCRPQCCQSV
EDM05994.1 XXXXXXXXXX XXXXXXXXXX XXXXXXXMVS SCCGSVCSEE GCGPG----- CCQPSCCQTT CCQTTCCKPT CCHPSCCVSS CCRPQCCQSV CCQPTCCRPS CCISSCCRPS CCRPSCCVSS CCRPSCCVSS
Consensus                                  MVs SCCGSVCSEE GCGQG      CCQPSCCQTT CCRTTCCRPS CCvsSCCrps Cc.s.ccrp. CC.s.ccrp. cc..scc..s CCRPSCCvss CCrpsCC.ss

           131                                                                                                                                  260
XP_577126. CCRPVCCQTT CCQSSCCRPV CCQTTC---- --RPSCGVSS CCRPVCCQTT C--------- --RPSCGVSS CCRPVCCQTT CRPSCGVSSC CRPVCCQTTC CRTTSCGPSC
XP_0010561 CCVSSCCRPS CCQPSCCRPS CCQPSCCRPS CCQPSCCVSS CCRPSCCQPS CCVSSCCRPS CCRPQCCQSA CCQPTCC--- -RPSCCISSC CQPSCGGSSC CGSSCCRP-C CRPCCCLRPV CGQVCCQTTC
XP_0027245 CCVSSCCRPS CCQPSCCRPS CCQPSCCRPS CCQPSCCVSS CCRPSCCQPS CCV------- ---PQCCQSA CCQPTCC--- -RPSCCISSC CQPSCGGSSC CGSSCCRP-C CRPCCCLRPV CGQVCCQTTC
XP_0010564 CCRPCCGSSS CCGSSCCRPS CCISSCCRPS CCEG------ -CGQGCCQPS CCQTTCCRTT CCRPSCCVSS CCRPQCC--- -Q-SGCGQGC CQPSCCQTTC CRTTCCRPSC CVSSCCRPSC CVSSCCRPSC
EDM05991.1 CCHPTCCISS CCRPSCCRPT CCISSCCRPS CCRPSCCISS CCRPSCCRPS CCISSCCRPS CCRPTCCISS CCRPSCC--- -RPSCCISSC CRPSCCRPSC CRPSCCISSC CRPSCCRPQC CISSCCRPIC
XP_0027246 CCVSSCCRPQ CCQSVCCQPT CCRPSCCRPS CCRPSCCVSS CCRPSCCISS C--------- --RPCCGSSS CCGSSCC--- -RPACCISSC CRPSCCRPSC CVSSCCRPQC CTPTCCQTTC CRTTCCRPSC
XP_0027246 CCRPSCCISS CCRPSCCRPT CCISSCCRPS CCRPSCCISS CCRPSCCRPS CC-------- --RPSCCISS CCRPSCC--- -RPQCCISSC CRPICCQTTC CRTTCCRPAC SSASCC
XP_0027246 CCRPSCCISS CCRPSCCRPT CCISSCCRPS CCRPSCCISS CCRPSCCRPS CC-------- --RPSCCISS CCRPSCC--- -RPQCCISSC CRPICCQTTC CRTTCCRPAC SSASCC
EDM05993.1 CCRPCCTSSS CCGSSCCRPS CCISSCCRPS CCISSCCRPS CCRP-CCGSS SC-------- --GSSCCRPS CCVSSCC--- -RPQCCISTC CRPTCCQTTC CRATCCRPAC SSGSCC
XP_0027246 CCQP-----T CCRPSCCRPS CCISSCCQPS CGGSSCC-RP CCQP-CCRPC CCL------- --RPVCG-QV CCQSTCY--- -RPTCVISTC PRPMCCATPC C
XP_0010560 CCQP-----T CCRPSCCRPS CCISSCCQPS CGGSSCC-RP CCRP-CCRPC CCL------- --RPVCG-QV GCQSTCY--- -RPTCVISTC PRPMCCATPC C
XP_0027246 CCQP-----T CCRPSCCRPS CCISSCCQPS CGGSSCC-RP CCRP-CCRPC CCL------- --RPVCG-QV GCQSTCY--- -RPTCVISTC PRPMCCATPC C
EDM05990.1 CCQP-----T CCRPSCCRPS CCISSCCQPS CGGSSCCGSS CCRP-CCRPC CCL------- --RPVCG-QV CCQSSCY--- -RPTCVISTC PRPMCCATPC C
EDM05994.1 CCRP-----S CCVSSCCRPS CCVSSCCRPS CCVSSCCRPS CCRPSCCVSS CY-------- --RPXXXXXX XXAA------ -RAPVVAQLA LVVLVAKN
Consensus  CC.p.c...s CCrpSCCRPs CCisSCCrPS Cc..SCc.ss CCrP.CCrps Cc........ ..rP.C..s. cC...Cc    Rp.c.is.C crP.Cc.t.C C....c.p.c .......... ..........

           261                                                  319
XP_577126.
XP_0010561 YRPTCVISTC PRPMCCATPC C
XP_0027245 YRPTCVISTC PRPMCCATPC C
XP_0010564 CQSVCCQPTC CRPSCCLLPP HLLPDHLLQD HLLPPSMLQW VLLLSTVTHS LLFFITQHS
EDM05991.1 CQTTCCRTTC CRPACSSASC C
XP_0027246 SSGSCC
XP_0027246
```

Mus musculus (Souris)

```
           1                                                                                                                                    130
NP_0010138 MVNSCCGSVC SEQGCDQSPC QESCGQPSCS QPSCSQPSCC QTTCCRTTCC RPSCCVSSCC RPTC-QTTCC RPVCCQ---- ------TTCR PSCGVSSCCR PVCCQ----- -TTCRPSCGV PSCCRPVCCQ
XP_0030858 MVSSCCGSVC SEEGCGQSCC Q--------- -TTCCRPSCC -----VSSCC RPQCCQSVCC QPTCCHPSCC ISSCR----- ----VSSCCC PSCVNSSCCG SSSCRPTCCV SSCCRPQCCP SVCYQPSCCR
NP_0011823 MVSSCCGSVC SEEGCGQSCC Q--------- -TTCCRPSCC -----VSSCC RPQCCQSVCC QPTCCHPSCC ISSCR----- ----VSSCCC PSCVNSSCCG SSSCRPTCCV SSCCRPQCCP SVCYQPSCCR
NP_0010790 MVSSCCGSVC SEEGCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPSCCVSSCC RPQCCQSVCC QPTCCRP--- ------SCCR PSCCISSCCR P-CCRPSCCV SSCCRPQCCQ SVCCQPTCCR
NP_0011197 MVSSCCGSVC SEEGCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPQCCQSVCC QPTCCRPSCC ISSCCRPCCR PTCCVSSCCR PSCCISSCCR PSCCRPSCCV SSCCRPQCCQ SLCCQPTCCR
NP_0010927 MVSSCCGSVC SEEGCGQSCC QPSCCQTTCC RPSCCVSSCC RPSCCVSSCC RPQCCQSVCC QPTCCRPSCC ISSCCRP--- ------SCCR PSCCVS---- ---------- -SCCRPQCCQ SVCCQPTCCR
NP_0010927 MVSSCCGSVC SEEGCGQSCC QPSCCQTTCC RPSCCVSSCC RP-------- --QCCQSVCC QPTCCRPSCC ISSCCRP--- ------SCCR PSCCVS---- ---------- -SCCRPQCCQ SVCCQPTCCR
NP_0010790 MVSSCCGSVC SEEGCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPQCCQSVCC QPTCCRPSCC ISSCCRP--- ------SCCR PSCCVSSCCR PQCCQSVCCQ PTCCRPSCCI SSCCRPSCCR
NP_0010927 MVSSCCGSVC SEEGCGQGCC QPSCCQPTCC RTTCCRPSCC -----VSSCC RPQCCQSVCC QPTCCRPSCC ISSCCRP--- ------SCCV PSCC-----R PQCCIS---- -SCCRPTCCV SNCCRPSCCS
NP_0010940 MVSSCCGSVC SEEGCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPQCCQSVCC QPTCCRPSCC ISSCCRP--- ------SCCR PSCCRPSCCR PSCCQPSCCV SSCCKPSCCV SSCCRPSCCQ
NP_081110. MANSCCGSVC SEESCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPQCCQSLCC QPTCCRPSCC ISSCCRP--- ------TCCR PTCCISSCCR P--------- -TCCRPSCCI SSCCRPTCCR
NP_081363. MVSSCCGSVC SEEGCGQSCC QPSCCQ---- -TTCCRPSCC -----VSSCC RPSCCRPSCC VSSCCRPQCC QSVCCQP--- ------TCCR PSCCISSCCR PSCCRPSCCV SSCCRPQCCQ SVCCQPTCCR
NP_0011197 MVSSCCGSVC SEEGCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPQCCQSVCC QPTCCRPSCC RPSCCVS--- ------SCCR PSCCQSVCCQ P--------- -TCCRPSCCR PSCCISSCCQ
NP_083889. MVSSCCGSVC SEEGCSQGCC QPSCCVSSCC RPQCCQSVCC -----QPTCC RPSCCISSCC RPSCCRPSCC RPSCCVS--- ------SCCR PQCCQSACCQ P--------- -TCCRPSCCR PSCCISSCCQ
NP_0010416 MVNSCCGSVC SEESCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPSCCVSSCC RPSCCRPQCC QSVCCQP--- ------TCCR PSCCISSCCR P--------- -SCGSSSCCG SSCC-RPCCR
NP_081083.  MVSSCGSVC SEEGCGQGCC QPSCCQTTCC RTTCCRPSCC -----VSSCC RPSCCVSSCC RPSCCRPQCC QSVCCQP--- ------TCCR PSCCISSCCQ P--------- -SCGSSSCCG SSCC-RPCCR
Consensus  MvsSCCGSVC SEEgCgQgCC QpSCcqttcc rttCCrpSCC .....vssCC RPqCCqSvCC qPtCCrpsCC issCC.p          ..sCCr PsCc.ssccr p..c...... .sCcrpsCC. ssCC.p.CCr

           131                                                                                                                                  260
NP_0010138 TTC------- ---------- ----RPSCGV SSCCRP---- -------VCC QTTCRPSC-- ----GVSSCC RPVCCQTTC- RPSCGVSSCC RPVCCQTTC- -----RPSCG VSSCCRPVCC QTTCCRTTCC
XP_0030858 SSCR------ ---------- SSRCRLCCGS SSCCGS---- -SYYRPSCCI SSYRRPTCCI -SSYCRPSCC QSTCYHSYCS RPSCCASSCY RPSCCSPTCC VPSCCRPSCF QSVCCKPTCY VSSCCRPNCC
NP_0011823 SSCR------ ---------- SSRCRLCCGS SSCCGS---- -SYYRPSCCI SSYRRPTCCI -SSYCRPSCC QSTCYHSYCS RPSCCASSCY RPSCCSPTCC VPSCCRPSCF QSVCCKPTCY VSSCCRPNCC
NP_0010790 PSCC------ ---------- ----RPTCCI SSCCRP---- -CCGSSSCCG SSCCRPTCCI -SSCCRPSCC RPSCCVSSCC RPQCCISSCC RPCCGSS--- ---------- --SCCGSSCC RPSCCRPSCC
NP_0011197 PSCCISSCCR PSCCRPSCCV SSCCRPSCCI SSCCRP---- -CCGSSSCCG SSCCRPSCCI -SSCCRPSCC RPSCCVSSCC RPSCCVSSCC RPQCCQSLCC QPTCCRPSCC RPSCCISSCC RPSCCRPSCC
NP_0010927 PSCC------ ---------R PCCGSSSCCG SSCCRP---- ------SCCT SSCCRPIC-- ----CQATCS RPGCCVSSCC RPQCCISSCC RPICCQTTCC RTTCCRPACS SGSCC
NP_0010927 PSCC------ ---------R PCCGSSSCCG SSCCRP---- ------SCCT SSCCRPIC-- ----CQATCC RPGCCVSSCC RPQCCISSCC RPICCQTTCC RTTCCRPACS SGSCC
NP_0010790 PSCC------ ---------V SSCCRPSCCI SSCCRP-CCR PCCGSSSCCG SSCCRPTCCI -SSCCRPSCC RPSCCVSSCC RPQCCISSCC RPTCCQTTCC R-----PACS SGSCC
NP_0010927 PSCC------ ---------V SSCCRPQCCQ SLCCQPTCCR PCCDSSSCCG SSCCRPTCCI -SSCCRPSCC RPSCCVSSCC RPQCCISSCC RPTCCQTTCC RTTCCRPACS SGSCC
NP_0010940 SVCC------ ---------Q PTCCRPSCCI SSCCRP---- ------SCCV SSCCRPICSG GSSCCGSSCC RPSCCISSCC RPSCCVSSCC RPTCCQSTCC R-----PTCS SCSCC
NP_081110. PSCC------ ---------I SSCCRPSCCR PSCCI----- -----SSCCR PSCCRPSCCR -PSCCISSCC RPSCCVSSCC RPQCCISSCC RPICCQTTCC RTTCYRPACS SGSCC
NP_081363. PSCC------ ---------R PCCGSSSCCV SSCCRP---- ------QCCI SSCCRPICCQ -TTCCRTTCC RPACSSGSCC
NP_0011197 PSCG------ ---------- ---SS-SCCG SSCCRP---- ---------C C RPCCRPCC-- ----CLRPVC GQVCCQTTCY RPTCVISTCP RPMCCATPCC
NP_083889. PSCG------ ---------- ---GS-SCCG SSCCRP---- ---------C C RPCCRPCC-- ----CLRPVC GQVCCQTTCY RPTCVISTCP RPMCCATPCC
NP_0010416 PCCS------ ---------- ---PCCSPCC RPCCRP---- ---------C C RPCCRPCC-- ----CLRPVC GQVCCQTTCY RPTCVISTCP RPMCCATPCC C
NP_081083. PCCS------ ---------- ---PCCSPCC RPCCRP---- ---------C C RPCCRPCC-- ----CLRPVC GQVCCQTTCY RPTCVISTCP RPMCCAIPCC
Consensus  psCc                 . ....r.scc. ssCCrp ... ......scC. sscCRP.C..  ...C..scC rp.Cc.ssCc RP.CciSsCc RP.Ccqttcc ......p.c. ..scc..... ..........

           261                                      308
NP_0010138 GPSC
XP_0030858 ---------- ISSCCR---- -----PVCCR PTCCCPSCSG GSCC
NP_0011823 RPTSCHPTSS ASSYCRSNCL PYYFHPTCCV VSCCRPSCFQ SVCCKPTC
NP_0010790 VSSCCRPQCC ISSCCRPTCC ----QTTCCR TTCCRPSCSS GSCC
NP_0011197 VSSCCRPQCC ISSCCRPTCC ----ETTCCR TTCCRPACSS GSCC
NP_0010927
```

Annexe 6 : Les motifs penta peptidiques des séquences des KAP 4 de mammifères dont le génome a été séquencé

Bos Taurus

```
1                                                                                                                           130
XP_0027022  MVSSCCGSVC SDQSCGRSLC QETCCRPSCC QTTCCRTTCC RPSCCVSSCC RPVCCQPQC- ---------- -------CQP TCCPPSCCQT TCCRTTCCR- --------PS CGVSSCCRPV CCQPICPRPT
XP_0027022  MVSSCCGSVC SDQSCGRSLC QETCCRPSCC QTTCCRTTCC RPSCCVSSCC RPVCCQPTCP RPTCCISSCY RPSCCGLCQE TCCPPSCCQT TCCRTTCCR- --------PS CGVSSCCRPV CCQPICPRPT
XP_0027022  MVSSCCGSVC SDQSCGRSLC QETCCQPSCC QTTCCRTTCC RPSCGV---- ---CCRPQCC QPVC------ -------CQP TCPRPTCCIS SCCRPSCCG- ---------- ---------- -----LCQET
XP_0017904  MVSSCCGSIC SDQSCGRSLC QETCCRPRCC QTTCCRTTCC RPSCGV---- ---CCRPQCC QPVC------ -------CQP TCSRPACCIS SCCRPSCCG- ---------- ---------- -----LCQET
XP_0017904  MVSSCCGSVC SDQSCGRSLC QETCCQPSCC QTTCCRTTCC RPSCGVPTCC ISSCCRPQCC QPVC------ -------CQP TCPRPTCCIS SCCRPSCCG- --------PT CCISSCCRPR CCQSVLCQET
XP_0027022  MVSSCCGSIC SDQSCGRSLC QETCCRPRCC QTTCCRTTCC RPSCGVPTCC ISSCCRPQCC QPVC------ -------CQP TCSRPACCIS SCCRPSCCGS SCGSSCCRPT CCISSCCRPR CCQSVLCQET
XP_0026960  MVSSCCGSVC SDQSCGRSLC QETCCPPSCC QTTCCRTTCC RPSCGVPTCC ISSCCRPQCC QPVC------ -------CQP TCPRPTCCIS SCYRPSCC-- --------GS SCGSSCCRPQ CCQPVCCQPT
XP_0026960  MVSSCCGSVC SDQSCGRSLC QETCCQPSCC RTTCCRTTCC RPSCGVSSCC RPVCCQPTCP RPTC------ -------CIS SCYRPSCCVS SCGSTCCR-- --------PT CCISSCCRPQ CCQPVCCQPT
NP_0010947  MVSSCCGSVC SDQSCGRSLC QETCCQPSCC QTTCCRTTCC RPSCGVSSCC RPVCCQPTCP RPTC------ -------CIS SCYRPSCCVS SCGSSCCR-- --------PT CCISSCCRPQ CCQPVCCQPT
XP_0012541  MVSSCCGSVC SDQSCGRSLC LETCCRPTCC RTTCCRTTCC RPSCGVSSCC RPVCCQPTCP RPTC------ -------CIS SCYRPSCCVS SCGSSSCR-- --------PT CCISSCCRPQ CCQPVCCQPI
XP_0026960  MVSSCCGSVC SDQSCGRSLC QETCCQPSCC RTTCCRTTCC RPSCGVSSCC RPVCCQPTCP RPTC------ -------CIS SCYRPSCCVS SCGSTCCR-- --------PT CCISSCCRPQ CCQPVCCQPT
XP_0012541  MVSSCCGSVC SDQSCGRSLC QETCCQPSCC RTTCCRTTCC RPSCGVSSCC RPVCCQPTCP RPTC------ -------CIS SCYRPSCCVS SCGSTCCR-- --------PT CCISSCCRPQ CCQPVCCQPT
XP_0027022  MVSSCCGSVC SDQSCGRSLC QETCCQPSCC RTTCCRTTCC RPSCGVSSCC RPVCCQPTC- ---------- -------CIS SCCRPQCC-- ---------- ---------- ---------- --QPVCCQPT
XP_0026960  MVSSCCGSIC SDQSCGRSLC QETCCRPSCC QTTCCRTTCC RPSCGVPTCC ISSCCRPQC- ---------- -------CQP VCCQPTCS-- ---------- ---------- ---------- --RPACCISS
XP_0026960  MVSSCCGSVC SDQSCGRSLC QETCCRPSCC QTTCCRTTCC RPSCCVSSCC RPVCCQPTCP RPTC------ -------CIS SCYRPSCCGS S--------- ---------- ---------- ----------
XP_0012540  MVSSCCGSVC SDQSYGQSLC QETCCQPSCC QTTCCRTTCC RPSCGVSSCC RPVCCQPTCP RPTC------ -------CIS SCYRPSCCVS S--------- ---------- ---------- ----------
XP_0027022  MVSSCCGSIC SDQGCG---- QQTCCSPSCC QTTCCRTTCC RPSCGVSSCR RPICCQTTC- RPSC------ -------AVS SCCRPICCQT TC-------- ---------- ---------- ----------
 Consensus  MVSSCCGS!C SDQsCGrslc Q#TCC.PsCC qTTCCRTTCC RPSCgVssCc rpvCCqptC. rp.c       c.s sC.rP.CC.s sc....c..  ..  .......... ..qp.ccq.t

            131                                                                                                                     260
XP_0027022  CCISSCYRPS CC-VSSCGSS CCR----PTC CISSCCRPQC CQPVCCQPTC PRPTCCISSC YRPSCCGSSC GSSCCRP--- --HCC----- LRPVCGRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0027022  CCISSCYRPS CC-VSSCGSS CCR----PTC CISSCCRPQC CQPVCCQPTC PRPTCCISSC YRPSCCGSSC GSSCCRP--- --QCCQPVCC LRPVCGRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0027022  CCQPSCCQTT CCRTTCCRPS CGV----PTC CISSCCRPRC CQPVCCQPTC SRPACCISSC CRPSCCVSSC GSSCCRPTCC ISSCCRPRCC LRPVCGRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0017904  CCRPSCCQTT CCRTTCCRPS CGV----PTC CISSCCRPQC CQPVCCQPTC SRPACCISSC YRPSCCGSSC GSSCCRPTCC ISSCCRPRCC LRPVCGRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0017904  CCQPSCCQTT CCRTTCCRPS CGV----PTC CISSCCRPRC CQPVCCQPTC SRPACCISSC CRPSCCVSSC GSSCCRPTCC ISSCCRPRCC LRPVCGRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0027022  CCRPSCCQTT CCRTTCCRPS CGV----PTC CISSCCRPQC CQPVCCQPTC SRPACCISSC YRPSCCGSSC GSSCCRPTCC ISSCCRPRCC LRPVCGRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0026960  CPRPTCCISS CYRPSCCGSS CGSSCCRPTC CISSCCRPRC CQSVCCQPTC SR----ISSC CRPSCCGSSC ---------- ---CLRPVC- -----GRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0026960  CSRPTCCISS CCRPSCCVSS CGSSCCRPAC CISSCCRPRC CQSVCCQPTC SR----ISSC CRPSCCGSSC ---------- ---CLRPVC- -----GRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
NP_0010947  CSRPACCISS CCRPSCCVSS CGSSCCRPAC CISSCCRPRC CQSVCCQPTC SH----ISSC CRPSCCGSSC ---------- ---CLRPVC- -----GRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0012541  CPRPTCCISS CCRPSCCGSS CGSNCCRPAC CISSCCRPRC CQSVCCQPTC SR----ISSC CRPSCCGSSC ---------- ---CLRPVC- -----GRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0026960  CSRPTCCISS CCRPSCCVSS CGS------- ---SCCRPRC CQSVCCQPTC SR----ISSC CRPSCCGSSC ---------- ---CLRPVCC LRPVCGRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0012541  CSRPTCCISS CCRPSCCVSS CGS------- ---SCCRPRC CQSVCCQPTC SR----ISSC CRPSCCGSSC ---------- ---CLRPVC- -----GRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0027022  CSRPTCCISS CCRPSCCVSS CG-------- --SSCCRPRC CQSVCCQPTC SR----ISSC CRPSCCGSSC ---------- ---CLRPVC- -----GRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0026960  CYRPSCCGSS CGSSCCRPTC CI-------- --SSCCRPRC CQSVCCQPTC SR----ISSC CRPSCCGSSC ---------- ---CLRPVC- -----GRVSC HTTCYRPTCV ISTCPRPVCC PSSCC
XP_0026960  ------YGSS CCRPTCC--- ---------- -ISSCCRPQC CQPVCCQPTC SL----ISSC CRPSCCGSSS CGS------- --SCCRPSCC LRPVCGRVSC HTTCYRPTCV ISTCPRLGFI VLADLLMKLA
XP_0012540  ------CGSS CCRPTCC--- ---------- -ISSC-RPQC CQPVCCQPTC SH----ISSC CRPSCCGSSS CGS------- --SCCRPSCC LRPVCGRVSC HTTCYRLTCV ISTCPRPVCC PSSCC
XP_0027022  --RPSCGVSS CCRPVCCQTT CR-----PSC GVSSCCRPVC CQTTCCRTTC ---------- CRPSCGGSFC
 Consensus  c.rp.cc.ss Ccrp.CC..s cg.....p.c .iSSCCRPrC CQ.vCCqpTC sr....issc cRPSCcGSsc .......... ...c.rp.c. .....grvsc httcyrptcv istcprpvcc psscc
```

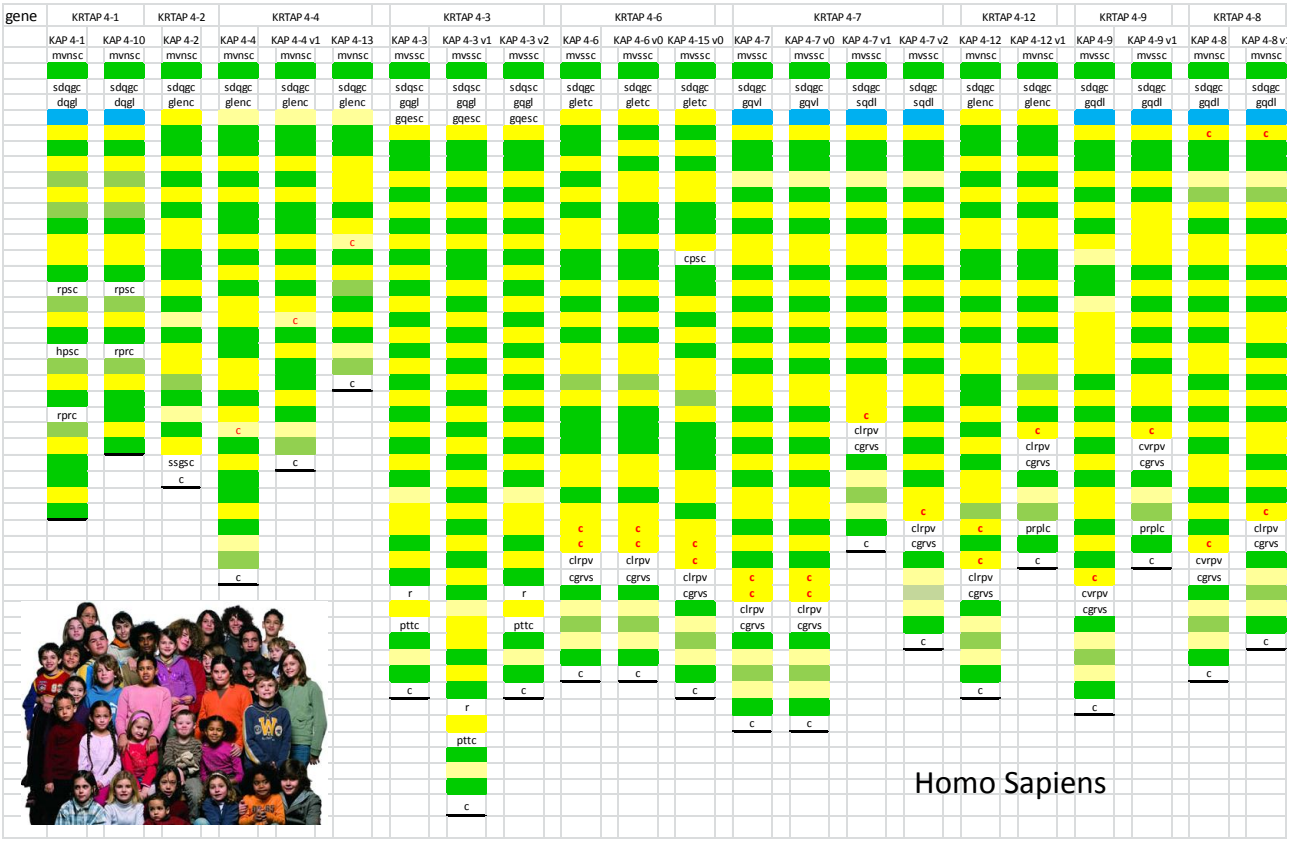
Equus Caballus

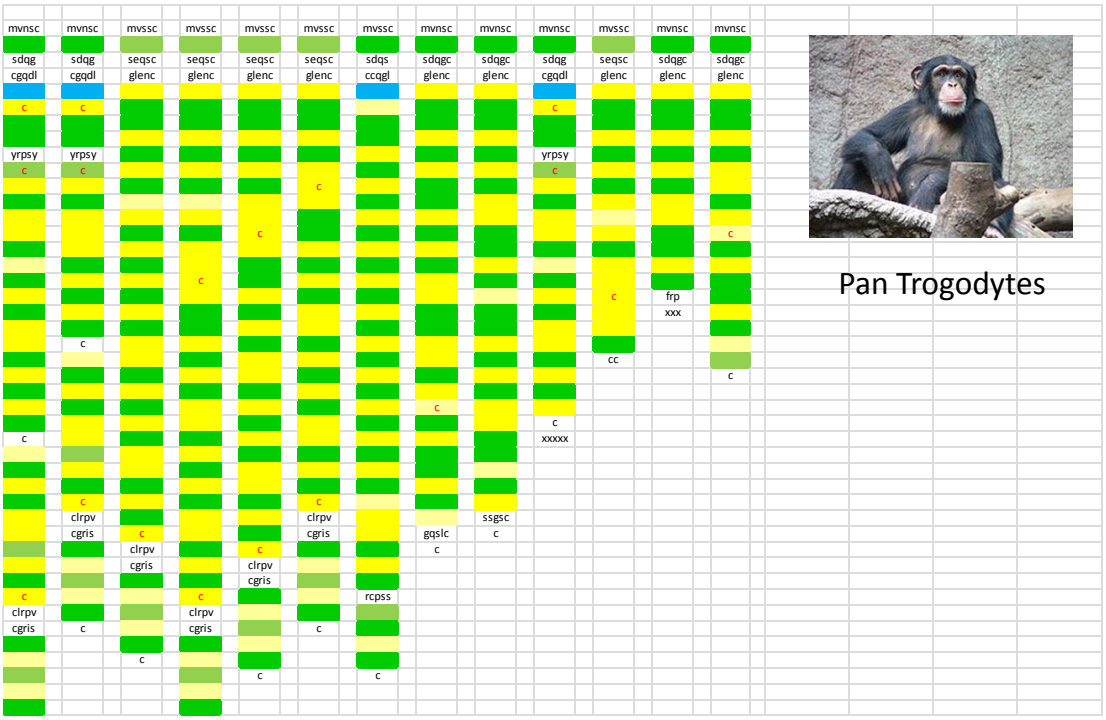
```
1                                                                                                                           130
XP_0014991  MVSSCCGSVC SEHSCGQGLC QETCCRP--- -------SCD ASSCCRPQCC ISSCSRPSCS QSVCCQPTC- ---------- ----CRPSCS VSSCCRPQCC ISSCCRP--- --GCSV---- -SSCCRPQCC
XP_0019179  MVSSCCGSVC SEHSCGQGLC QETCCRPRC- ----CQTTCC RTTCCRPSCC VSSCCRPQCC QSVCCRPTC- ---------- ----CRPSCC ISSCCRPSCC QTTCCRTTCC RPSCCI---- -SSCCHPSCC
XP_0014983  MVSSCCGSVC SEHSCGQGLC QETCCRPRC- ----CQTTCC RTTCCRPSCC VSSCCRPQCC QSVCCRPTC- ---------- ----CRPSCC ISSCCRPSCC VSSCCRPQCC QSVCCQ---- -PTCCHPSCC
XP_0014990  MVSSCCGSVC SDQGCGQ--- -ETCCRPSC- ----CQITCC RTTCYQPTCS VPSCCRPQCC ISSCCRPSCS VSSCCRPQCC ISSCCRPSCS VSSCCRPQCC ISSCCHPSCS GSSSCG---- -SSCCRPSCS
XP_0019179  MVSSCCGSIC SDQGCGQ--- -ESCCRPSC- ----CQITCC RTTCLRPSCS VSSCCRPQCC ISSCCQPSCS SSSCCG---- -SSCCRPSCS VSSCCRPSCC ISSCCRPSCS GSSSCG---- -STCCRPSCS
XP_0019179  MVSSCCGSIC SDQGCGQ--- -ESCCRPSC- ----CQITCC RTTCLRPSCS VSSCCRPQCC ISSCCQPSCS SSSCCG---- -SSCCRPSCS VSSCCRPSCC ISSCCRPSCS GSSSCG---- -STCCRPSCS
XP_0019179  MVSSCCGSIC SDQGCGQ--- -ESCCRPSC- ----CQITCC RTTYLRPTCS VSSCCRPQCC ISSCCQPSCS SSSCCG---- ---------- -SSCCRPSCC ISRCCRPSCS GSSSCG---- -STCCRPSCS
XP_0019179  MVSSCCGSIC SDQGCGQ--- -ESCCRPSC- ----CQITCC RTTCLRPSCS VSSCCRPQCC ISSCCQPSCS SSSCCG---- -SSCCRPSCS VSSCCRPSCC ISSCCRPSCS GSSSCG---- -STCCRPSCS
XP_0014991  MVSSCCGSIC SENSCGQGLC QETCCRPSCD VSSCCRPQCC ISSCSRPSCS VSSCCRPQCC ISSCSRPSCS VSSCCRPQCC ISSCCRPSCS VSSCCRPQCC ISSCCRPSCS VSSCCRPQCC ISSCCRPSCS
XP_0014991  MVSSCCGSIC SENSCGQGLC QETCCRPSCD VSSCCRPQCC ISSCSRPSCS VSSCCRPQCC ISSCSRPSCS VSSCCRPQCC ISSCCRPSCS VSSCCRPQCC ISSCCRPSCS VSSCCRPQCC ISSCCRPSCS
XP_0014989  MVSSCCGSVC SDQGCGQ--- -ETCCRPSC- ----CQIACC RTTCLRPSCS VSSCCRPQCC ---------- ---------- ---------- ISSCCRPQCQ QSVCCQP--- ---------- --TFLRPSCS
XP_0014983  MVSSCCGSVC SEHSCGQGLC QETCCRPRC- ----CRTTCC RTTCCRPSCC VSSCCRPQCC QSVCCQPTS- ---------- ----CHPSCC ISSCCRPSCC RPQGCQSVCC QPICCR---- -PSCCQTTCC
 Consensus  MVSSCCGS!C S#.sCGQglc qEtCCRPsC. ....Cq.tCC rttC.RPsCs VSSCCRPQCC .s.ccqp.c. .......... ....crpsc. !SSCCRPsCc isscCrp.c. .sscc..... .ssccrpsCs

            131                                                                                                         250
XP_0014991  ISSCCRPSCD V--------- -SSCCRP--- -------QCY ISSCCRPSCS VSNCCQPQCC ISSCCQPSCS SSSCCDSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0019179  ISSCCSPSCC G--------- -SSCCRP--- -------QCC QSVCCQPTC- ----CHPSCC ISSCCRPSC- ----CGSSCC RPTCCIRPVC GWVSCHTTCY RPTCVISTCP RPVCCASSCC
XP_0014983  ISSCCRPSCC G--------- -SSCCRP--- -------TCC ISSCCRPSCS ISSCCRPQCC QSECCQPTC- ----CIPSCC RPSCCLRPVC GWVSCHTTCY RPTCVISTCP RPVCCASSCC
XP_0014990  VSSCCHPSCC ---------- ISSCCRPSFS VSSCCRPSCD VSSCCRPSCG ISSCCRPSCC ISSCCQPSCS SSSCCGSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0019179  VSSCCRPSSC GSTCCRPRCD VSSCCRPQCY ISSCCRPRCY ISSCCRPSCS VSNCCQPQCC ISSCCQPSCS SSSCCGSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0019179  VSSCCRPSSC G--------- -STCCRP--- ---------- --SCCRPSCS VSNCCQPQCC ISSCCQPSCS SSSCCGSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0019179  VSSCCRPSC- ---------D VSSCSRP--- -------RCY ISSCCRPSCS VSNCCQPQCC ISSCCQPSCS SSSCCGSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0019179  VSSCCRPSSC G--------- -STCCRP--- ---------- --SCCRPSCS VSNCCH---- ---------- -SSCCGSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0014991  VSSCSRPQCC ---------- ISSCSRP--- --SCSQSVCC QPTCCRPTCS VSSCCRPHCC ---------- -----DSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0014991  VSSCSRPQCC ---------- ISSCSRP--- --SCSQSVCC QPTCCRPTCS VSSCCRPQCC ISSCCQPSCS SSSCCDSSCC RPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSSCC
XP_0014989  VSSCCRPRCC ---------- ISSCCRP--- -------SCS VSSCCGPSCD VSSCCRPQCC ISSCCQPSCS NSSCCGSSCC HPSCCLRPVC GRVSCHTTCY RPTCVISTCP RPVCCPSTCC
XP_0014983  RTTCCRPSCC ---------- ISSCCRP--- -------SCC GSSCCRPTCC ISSCCRPSCC PTICCRPICS SGSCC
 Consensus  vssCcrPscc .         . .SsCcRP... ........c. .ssCCrPsCs !SsCCrPqCC isscCqpscs ssscCgsscc rpscclrpvc grvschttcy rptcvistcp rpvccpsscc
```

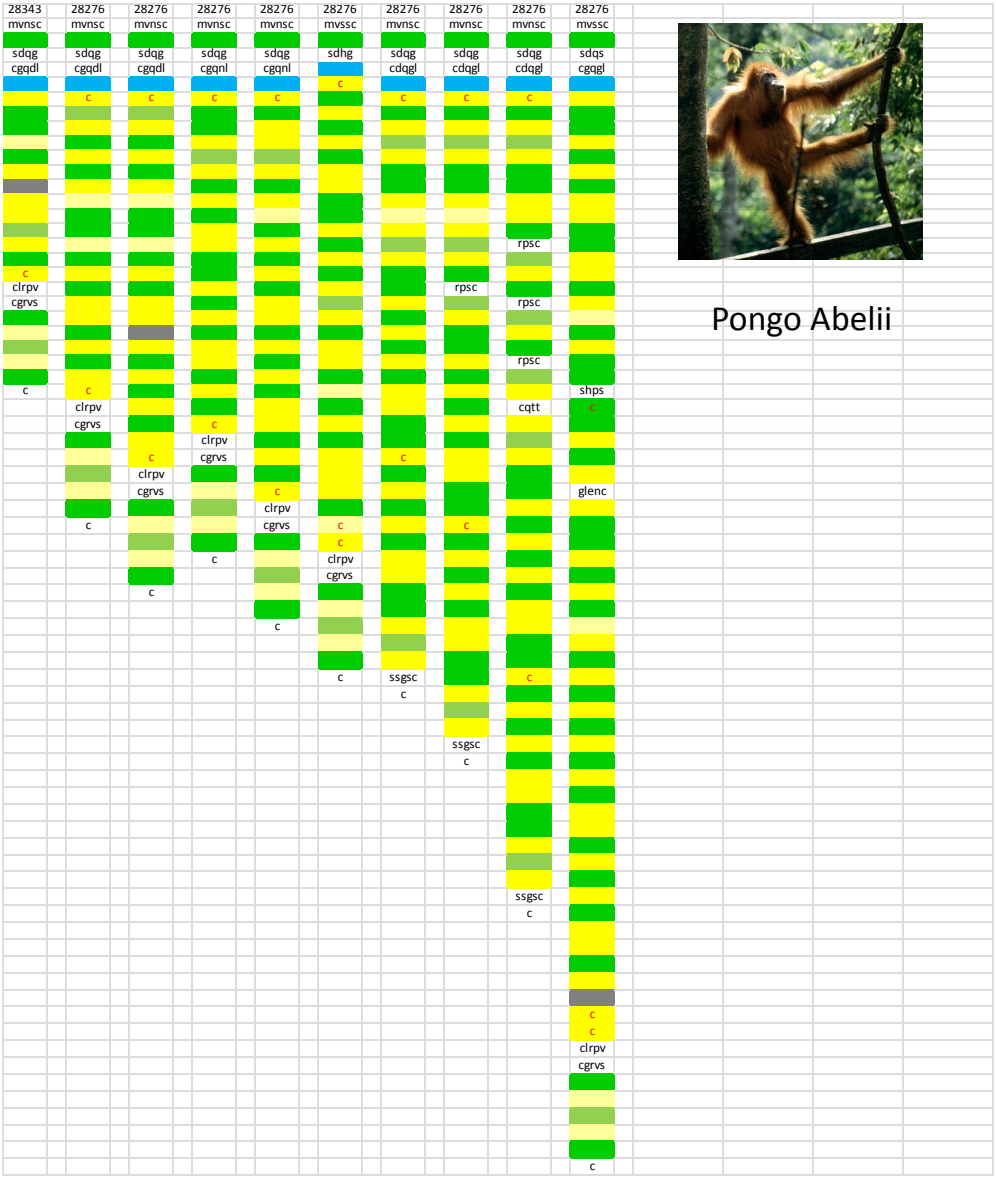
Ornithorhynchus anatinus (Ornithorynque)

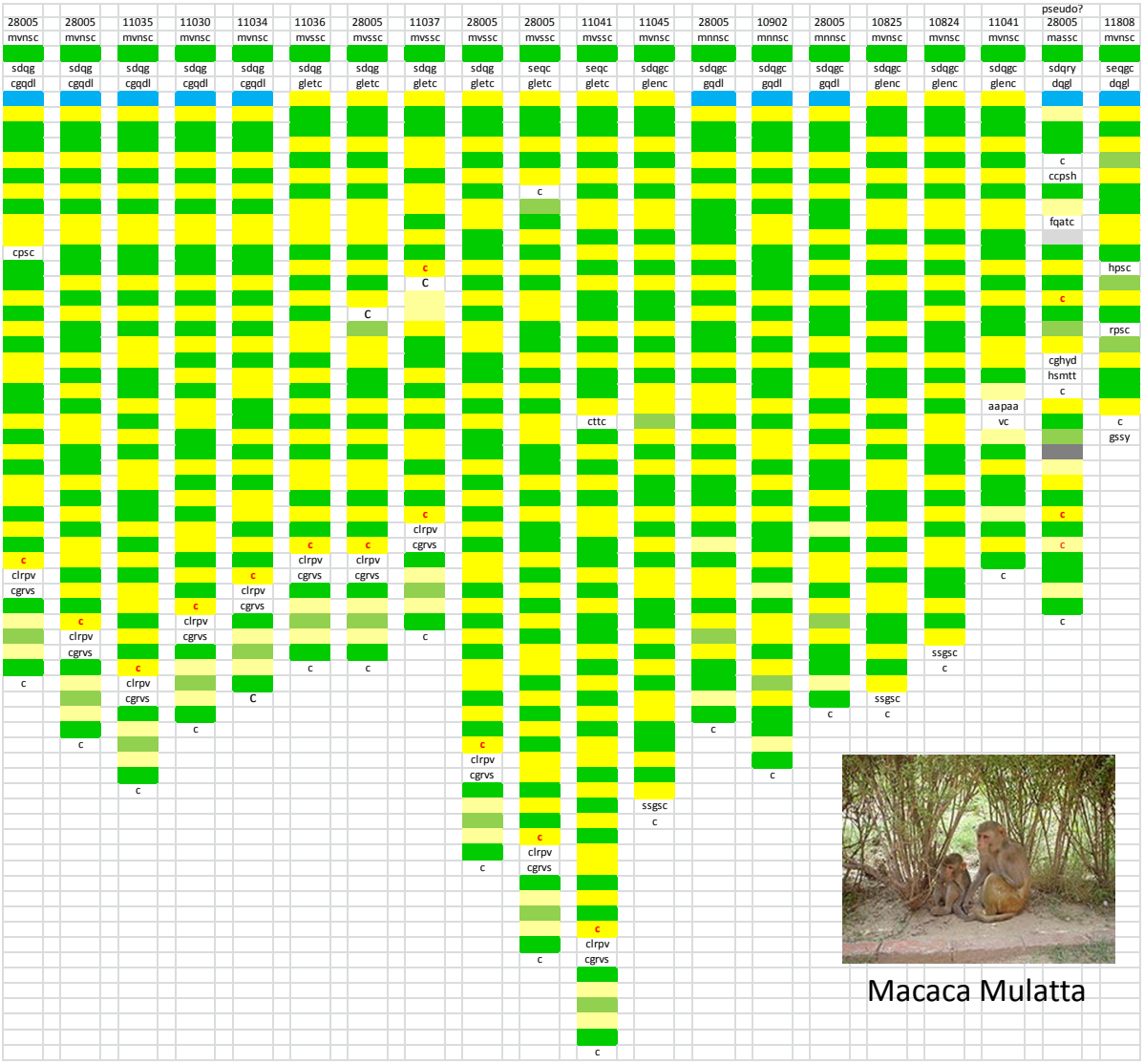
Annexe 6 : Les motifs penta peptidiques des séquences des KAP 4 de mammifères dont le génome a été séquencé

Monodelphis Domestica (Opossum)

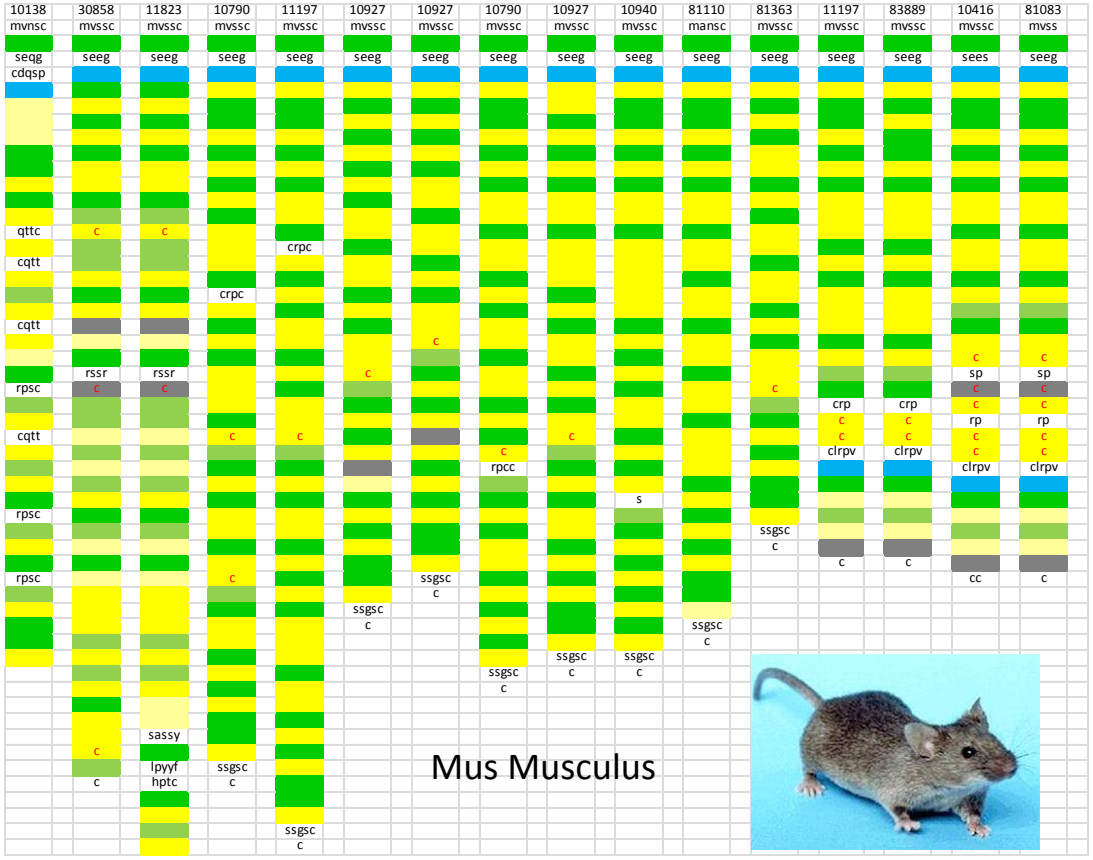

gene
KRTAP 4-1
KRTAP 4-2
KRTAP 4-4
KRTAP 4-3
KRTAP 4-6
KRTAP 4-7
KRTAP 4-12
KRTAP 4-9
KRTAP 4-8
KAP 4-1
KAP 4-10
KAP 4-2
KAP 4-4
KAP 4-4 v1
KAP 4-13
KAP 4-3
KAP 4-3 v1
KAP 4-3 v2
KAP 4-6
KAP 4-6 v0
KAP 4-15 v0
KAP 4-7
KAP 4-7 v0
KAP 4-7 v1
KAP 4-7 v2
KAP 4-12
KAP 4-12 v1
KAP 4-9
KAP 4-9 v1
KAP 4-8
mvnsc
mvssc
sdqgc
dqgl
glenc
gqesc
gletc
gqvl
sqdl
gqdl
rpsc
hpsc
rprc
ssgsc
clrpv
cgrvs
cvrpv
prplc
pttc
Homo Sapiens

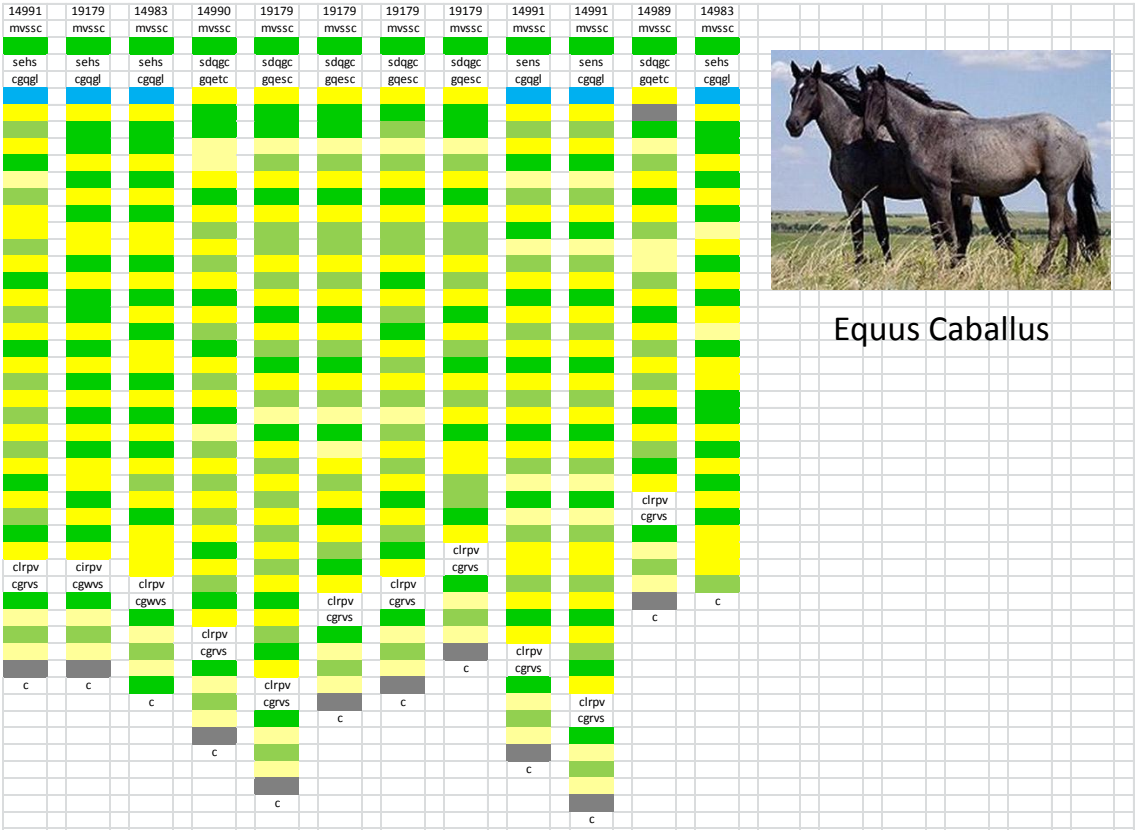

mvnsc
mvssc
seqsc
sdqs
ccqgl
sdqg
cgqdl
glenc
yrpsy
clrpv
cgris
rcpss
gqslc
ssgsc
xxxxx
frp
xxx
cc
Pan Trogodytes


28343 mvnsc sdqg cgqdl c clrpv cgrvs c
28276 mvnsc sdqg cgqdl c c clrpv cgrvs clrpv cgrvs c
28276 mvnsc sdqg cgqdl c c clrpv cgrvs c
28276 mvnsc sdqg cgqnl c c c clrpv cgrvs c
28276 mvnsc sdqg cgqnl c c clrpv cgrvs clrpv cgrvs c
28276 mvssc sdhg c c c clrpv cgrvs c ssgsc c
28276 mvnsc sdqg cdqgl c c c ssgsc c
28276 mvnsc sdqg cdqgl c c rpsc rpsc rpsc ssgsc c
28276 mvnsc sdqg cdqgl c rpsc rpsc rpsc cqtt c shps c glenc c ssgsc c
28276 mvssc sdqs cgqgl c c clrpv cgrvs c
Pongo Abelii


pseudo?
28005 mvnsc sdqg cgqdl cpsc c clrpv cgrvs
28005 mvnsc sdqg cgqdl c clrpv cgrvs c
11035 mvnsc sdqg cgqdl c clrpv cgrvs c
11030 mvnsc sdqg cgqdl c clrpv cgrvs c
11034 mvnsc sdqg cgqdl c clrpv cgrvs C
11036 mvssc sdqg gletc c clrpv cgrvs c
28005 mvssc sdqg gletc C c clrpv cgrvs c
11037 mvssc sdqg gletc c C c clrpv cgrvs c
28005 mvssc sdqg gletc c clrpv cgrvs c
28005 mvssc seqc gletc c c clrpv cgrvs c
11041 mvssc seqc gletc cttc c clrpv cgrvs c
11045 mvnsc sdqgc glenc ssgsc c
28005 mnnsc sdqgc gqdl c
10902 mnnsc sdqgc gqdl c
28005 mnnsc sdqgc gqdl c
10825 mvnsc sdqgc glenc ssgsc c
10824 mvnsc sdqgc glenc ssgsc c
11041 mvnsc sdqgc glenc aapaa vc c
28005 massc sdqry dqgl c ccpsh fqatc hpsc c rpsc cghyd hsmtt c c gssy c c c c c
11808 mvnsc seqgc dqgl
Macaca Mulatta

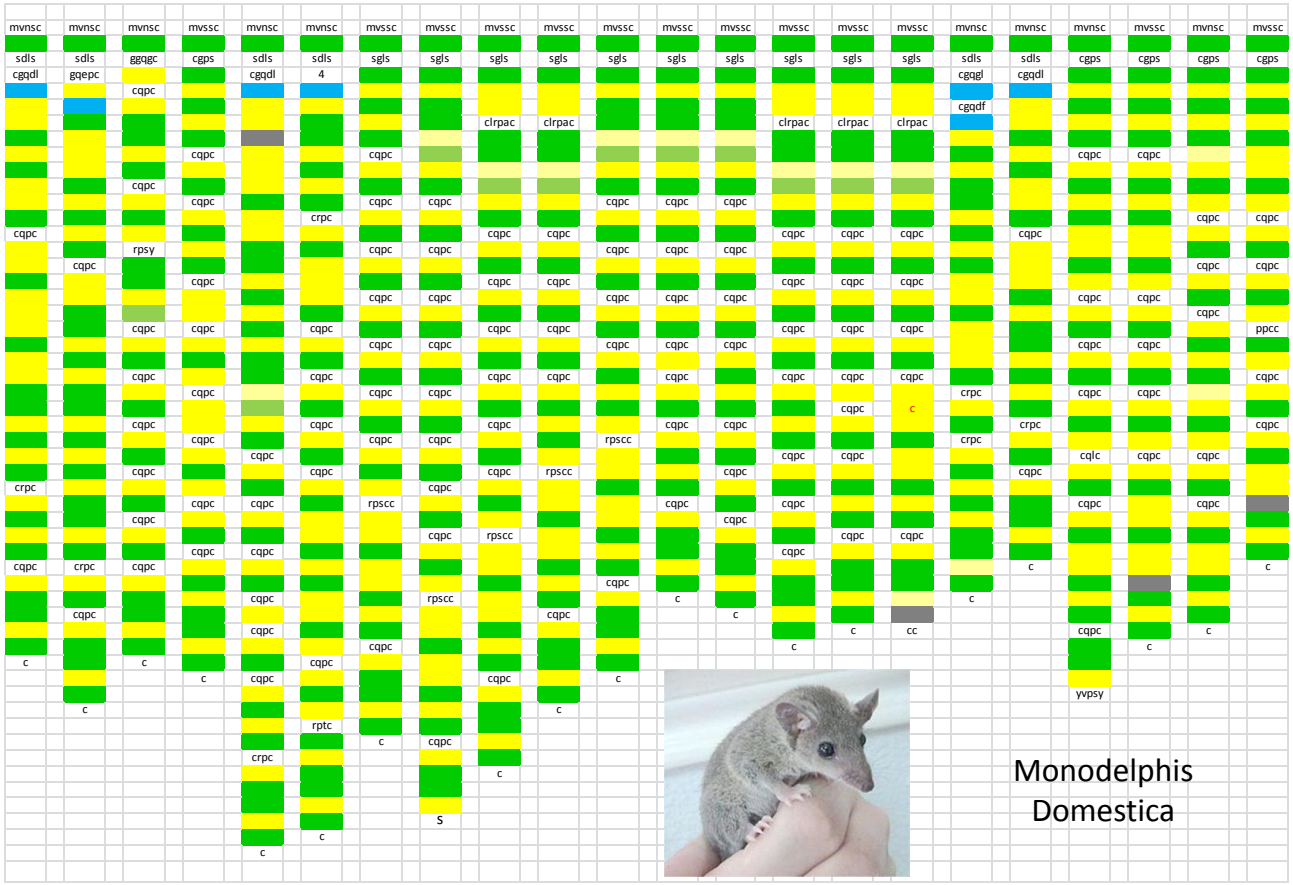

mvnsc seqgc cdqsp qttc rpsc rpsc rpsc rps rps
mvssc seeg cvc rpcc clrpv c
mvssc seeg rpcc clrpv c
mvssh seeg ceg qsg c xxxxx
mvssc seeg c ssasc c
mvnsc seeg c rpcc c
mvssc seeg c c
mvssc seeg c c
mvssc seeg c gsss c
mvssc seeg ggss c c c clrpv c
mvssc seeg ggss c c c clrpv cgqvg c
mvnsc seeg ggss c c c clrpv cgqvg c
mvssc seeg cgss c c clrpv c
mvssc seeg yrp xxxxxx
Rattus Norvegicus





Equus Caballus


Bos Taurus


Monodelphis
Domestica

Annexe 6 : Les motifs penta peptidiques des séquences des KAP 4 de mammifères dont le génome a été séquencé

mvnsc	mvnsc	mvnsc	mvnsc	mvnsc	mvnsc	mvn-c	mvnsc	mvnsc	mvnsc
sdls	sdls	sdls	sdls	sals	sals	sdls	sdls	sdls	sdls
c			cqpc			c	c	c	
				prpc				c	
sc								sppc	
cqpc			rptc	pqpc				cqpc	
		svasc			cvst	pqpc			
				prpc		yvsi			
		pqpc		cvss					pqpc
			cqpc						
								cqpc	
cqpc	cqpc	prpc	gispc		rpac		cqpc		prpc
						pqpc			
				prpc	cqpc				
		prpc				rqpc			prpc
				cqpc					
xxxxx	xxxxx							c	
			c				c		c
		prpc	xxxxx						xxxxx
cqpc	cqpc				c				
		c			sppc	c			
		rpac		c		rptc			
c	yppts								
xxxxx									
	c				gssac				
	lptc				y				
		cqpc				c			
	xxxxx								
		xxxxx							

Ornithorhynchus anatinus