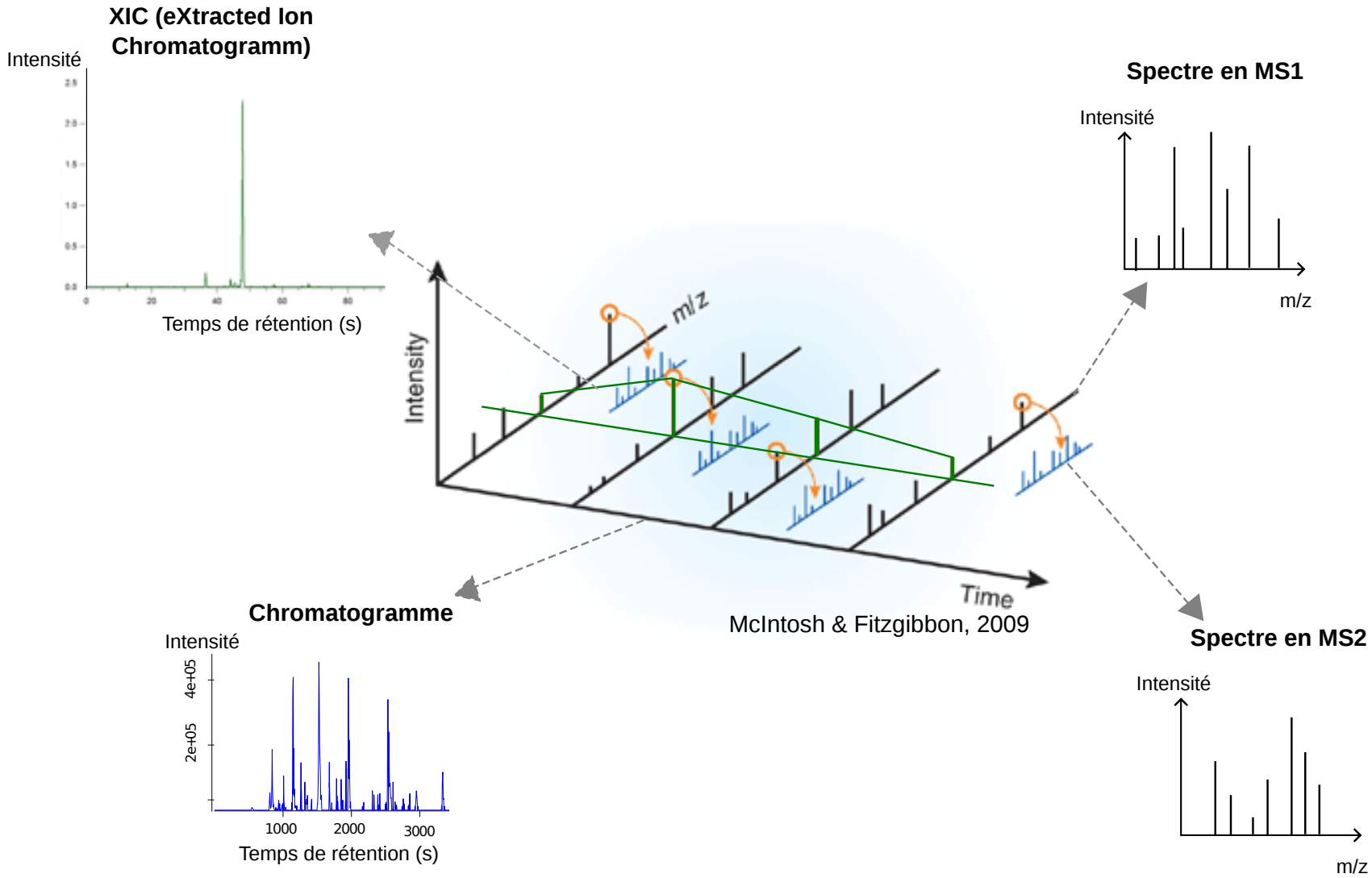
Traitement et analyse des données en protéomique

Mélisande Blein-Nicolas

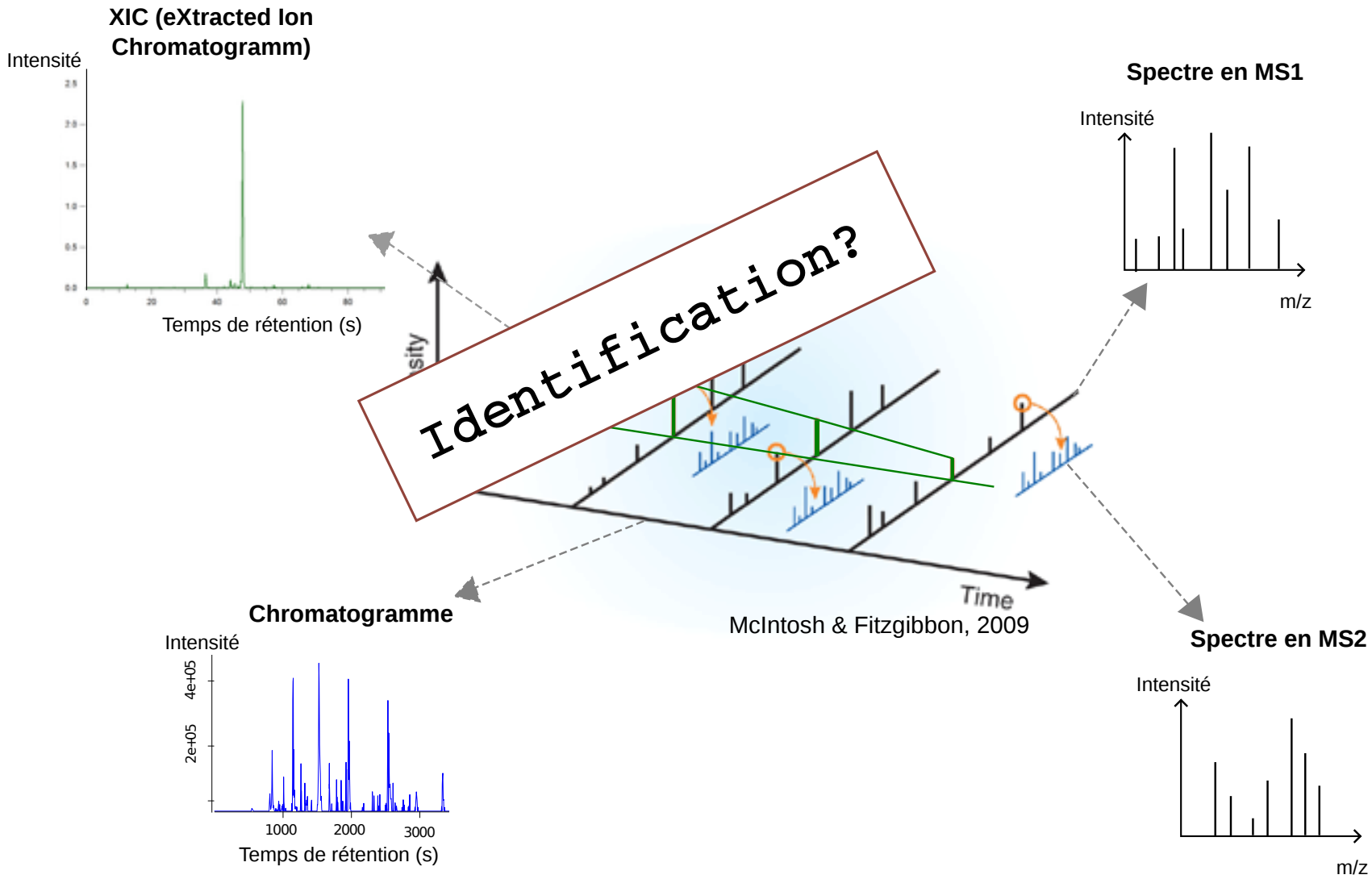
Journée commune SFEAP - RFMF

28 mars 2018

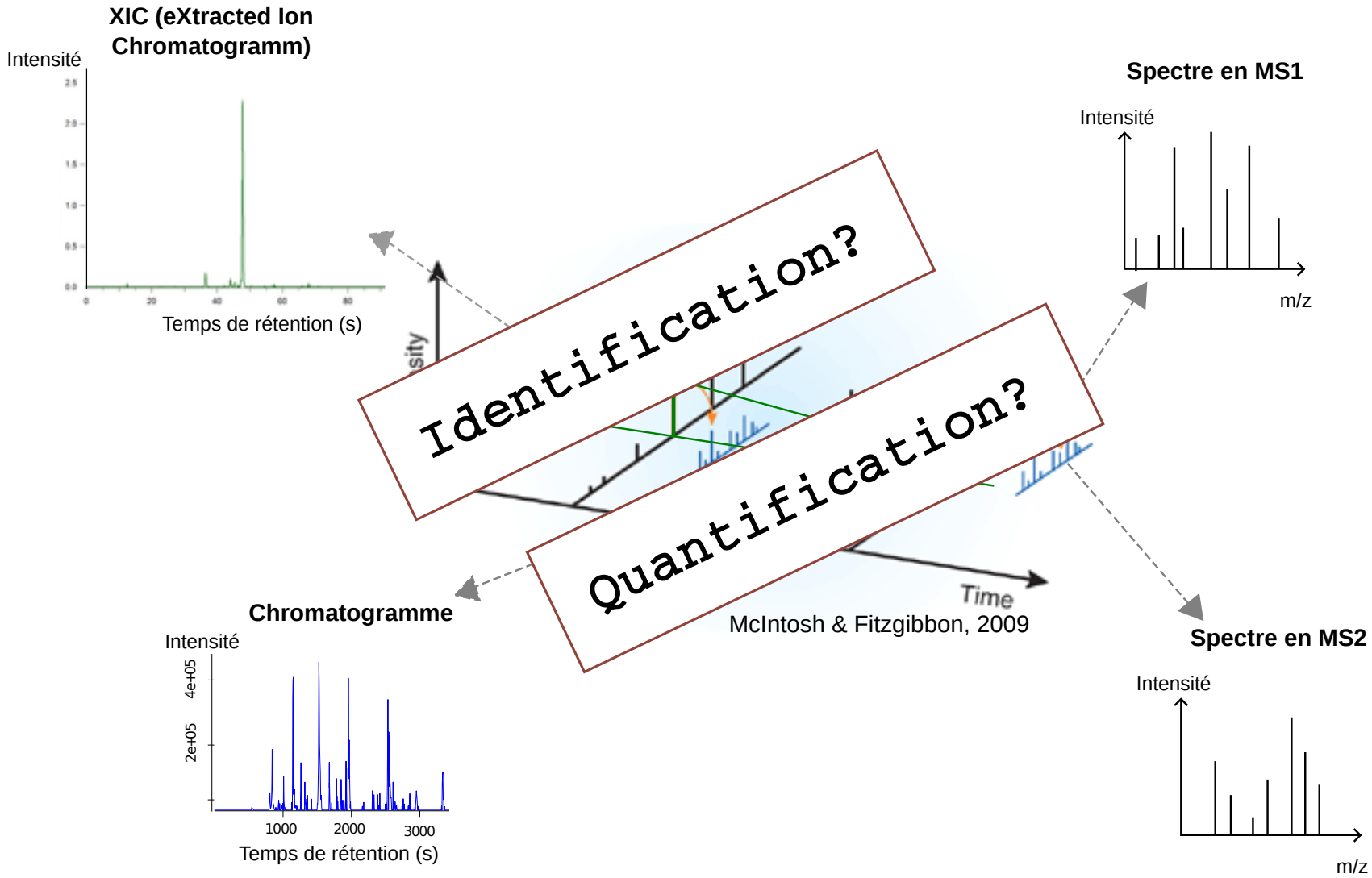
Données acquises en LC-MS/MS



Données acquises en LC-MS/MS



Données acquises en LC-MS/MS



Identification des peptides par peptide spectrum match

1. Digestion in silico

>protein1
MANNHGSKFVSVNLNKLYAQPSHHNYHSHT
>protein2
MLQSEVPPKISFINGVIAVRPLENIEEPLSV

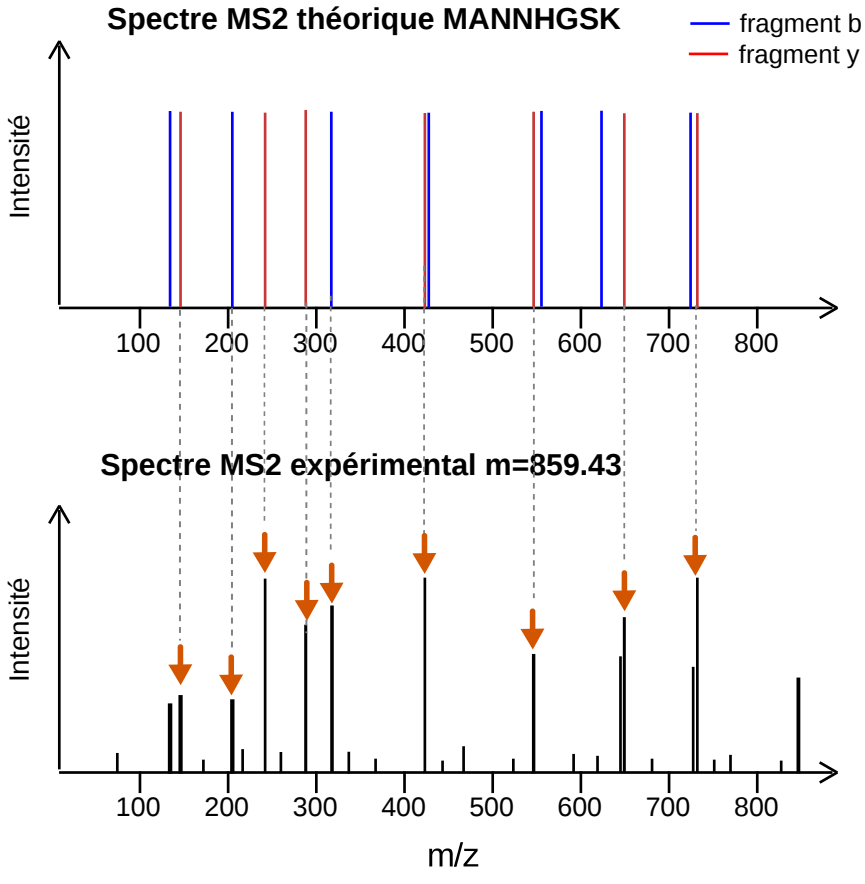
2. Création d'une liste de peptides théoriques et sélection de ceux dont la masse est proche de celle du peptide

Protéine	Peptide	Séquence	Masse (Da)
prot1	pep1	MANNHGSK	859.36
prot1	pep2	FVSVNLNK	808.54
prot1	pep3	LYAQPSHHNYHSHT	1005.26
prot2	pep4	MLQSEVPPK	859.86
prot2	pep5	ISFINGVIAVR	679.62
prot2	pep6	PLENIEEPLSV	983.76

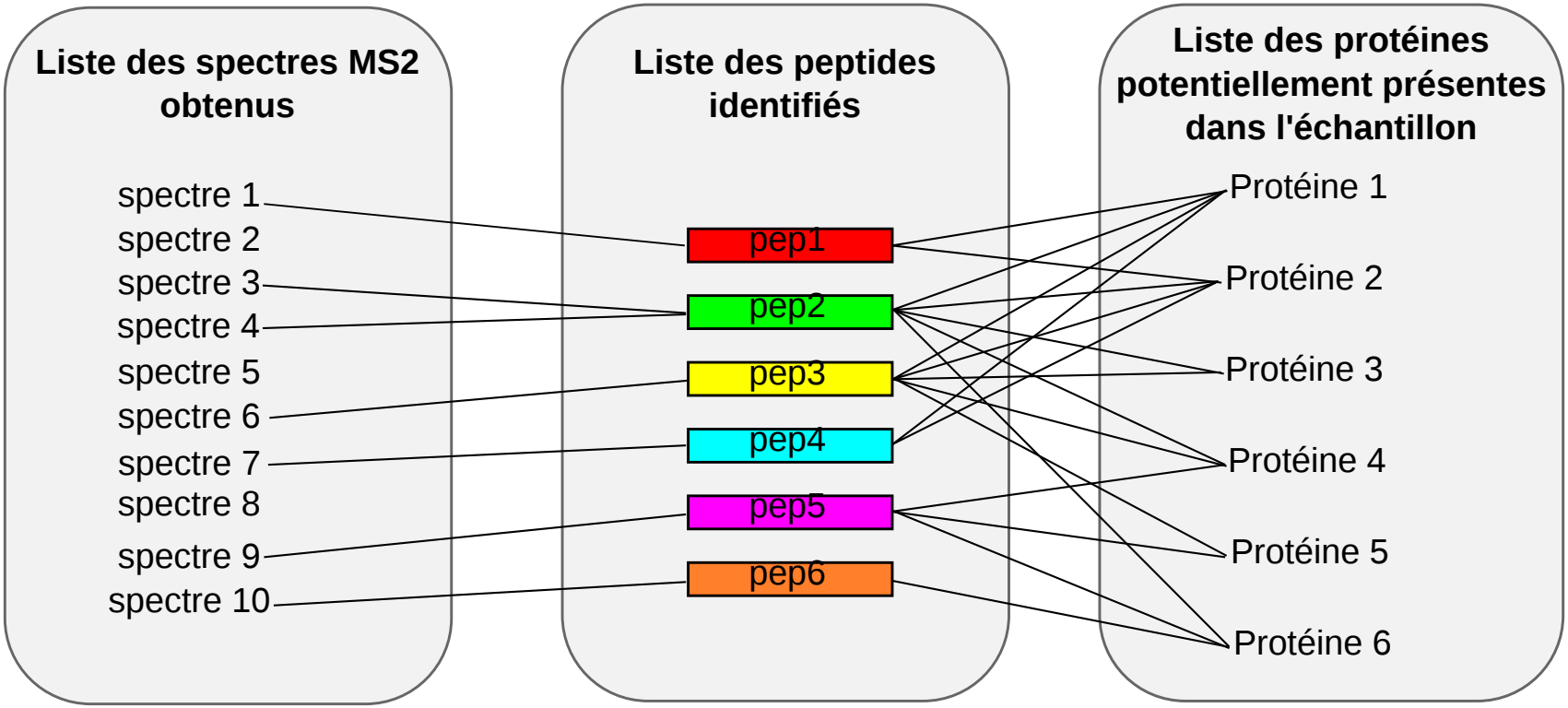
3. Fragmentation in silico (ex. avec MANNHGSK)

Fragment b	Masse (Da)	Fragment y	Masse (Da)
MANNHGS	712.27	ANNHGSK	727.32
MANNHG	625.24	NNHGSK	656.28
MANNH	568.22	NHGSK	542.24
MANN	431.16	HGSK	428..20
MAN	317.12	GSK	291.14
MA	203.08	SK	234.12
M	132.04	K	147.09

4. Comparaison des spectres observés et théoriques et calcul d'un score de similarité



Reconstruction de la liste exhaustive des protéines potentiellement présentes dans l'échantillon



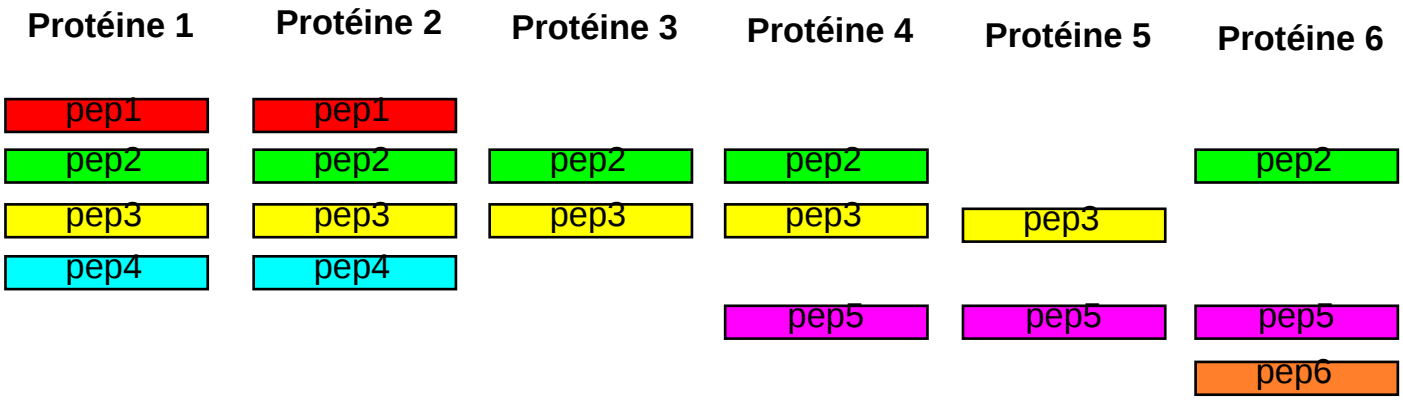
Inférence des protéines

- Objectif: ne garder que les protéines dont la présence est biologiquement pertinente
- Application du principe de parcimonie pour éliminer les protéines pour lesquelles il n'y a pas de preuve de présence

Pour plus d'info:
Li & Radivojac 2012, BMC Bioinformatics, 13:S4
Huang et al. 2012, Briefings in Bioinformatics, 13, 586-614

Inférence des protéines

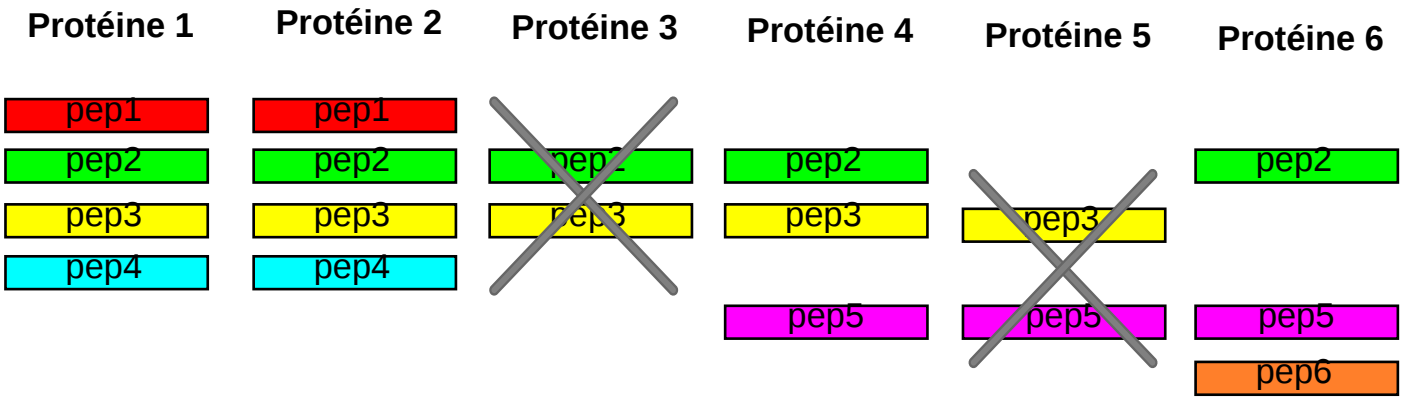
- Objectif: ne garder que les protéines dont la présence est biologiquement pertinente
- Application du principe de parcimonie pour éliminer les protéines pour lesquelles il n'y a pas de preuve de présence



Pour plus d'info:
Li & Radivojac 2012, BMC Bioinformatics, 13:S4
Huang et al. 2012, Briefings in Bioinformatics, 13, 586-614

Inférence des protéines

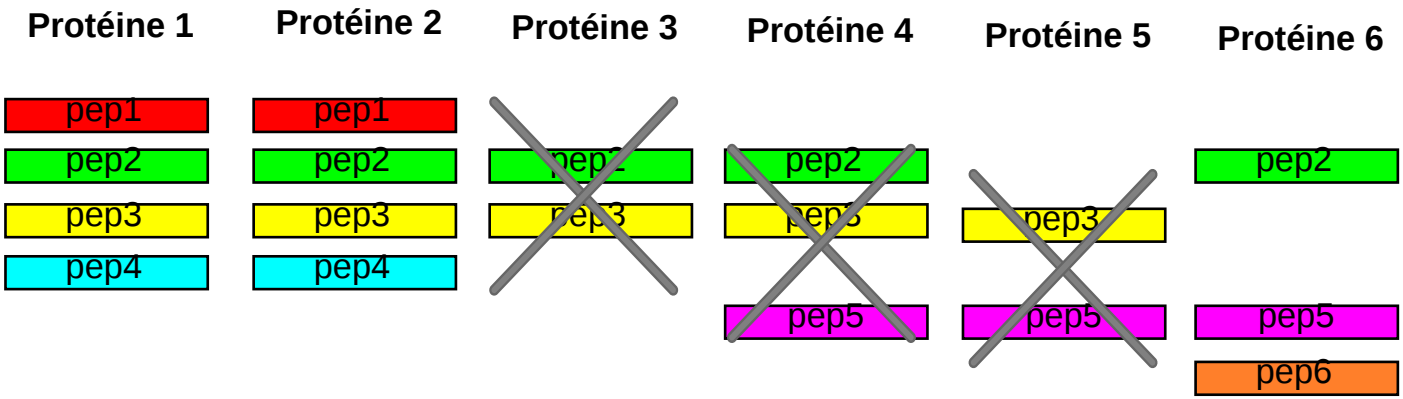
- Objectif: ne garder que les protéines dont la présence est biologiquement pertinente
- Application du principe de parcimonie pour éliminer les protéines pour lesquelles il n'y a pas de preuve de présence



Pour plus d'info:
Li & Radivojac 2012, BMC Bioinformatics, 13:S4
Huang et al. 2012, Briefings in Bioinformatics, 13, 586-614

Inférence des protéines

- Objectif: ne garder que les protéines dont la présence est biologiquement pertinente
- Application du principe de parcimonie pour éliminer les protéines pour lesquelles il n'y a pas de preuve de présence



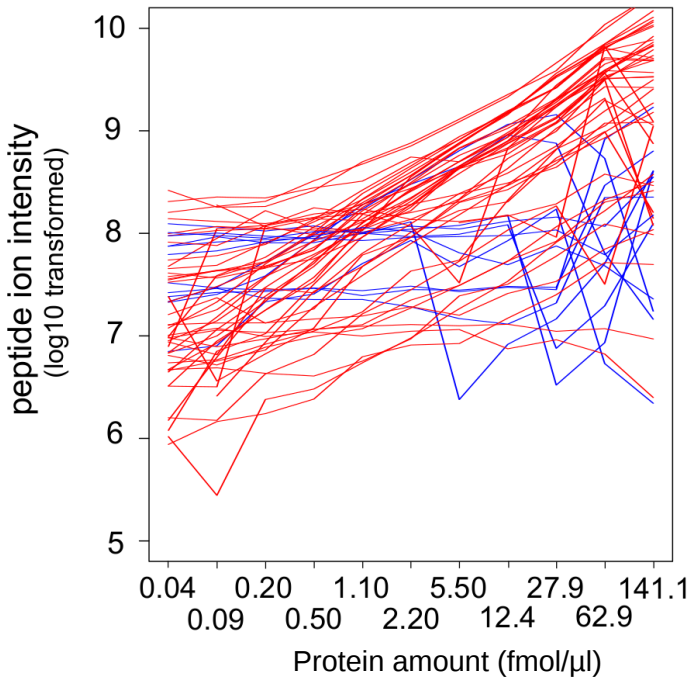
Pour plus d'info:
Li & Radivojac 2012, BMC Bioinformatics, 13:S4
Huang et al. 2012, Briefings in Bioinformatics, 13, 586-614

Principe

- Mesure de l'aire sous les pics chromatographiques en MS1
- Nécessite l'utilisation d'un logiciel de quantification qui va :
  - définir et isoler les pics
  - aligner les temps de rétention entre les échantillons
  - calculer l'aire sous les pics chromatographiques
  - relier les infos de quantification (MS1) et d'identification (MS2)
- Les données obtenues sont des intensités de peptide

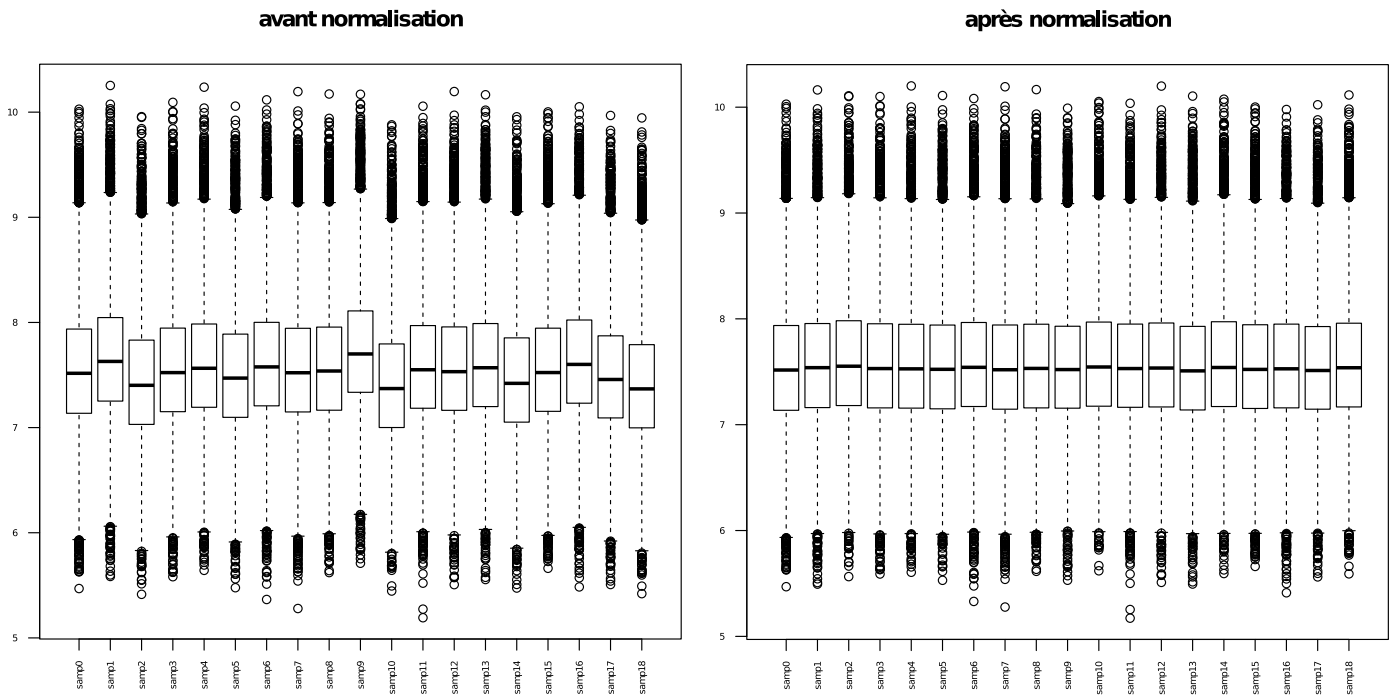
Post-processing 1: nettoyage des données

- Objectif : éliminer les données abérrantes
  - peptides dont le temps de rétention varie d'un échantillon à l'autre
  - peptides dont le pic chromatographique est anormalement large
  - peptides dont le profil d'intensité est atypique



Post-processing 2: normalisation

- Objectif : corriger les variations de quantité totale de protéines qui peuvent exister entre les échantillons (biais expérimental)



Pour plus d'info:
Valikangas et al., 2018, Briefings in Bioinformatics, 19, 1–11
Callister et al., 2006 J. Proteome Res., 5, 277-286

Post-processing 3: traitement des données manquantes

- Les données manquantes peuvent avoir plusieurs origines:
  - le hasard (problèmes d'alignement ou de détection des pics)
  - le seuil de détection du spectro
- Plusieurs possibilités pour traiter les données manquantes:
  - produire un jeu de données sans données manquantes en ne conservant que les peptides quantifiés dans tous les échantillons
  - limiter la quantité de données manquantes en supprimant les peptides quantifiés dans un petit nombre d'échantillon
  - imputer les données manquantes

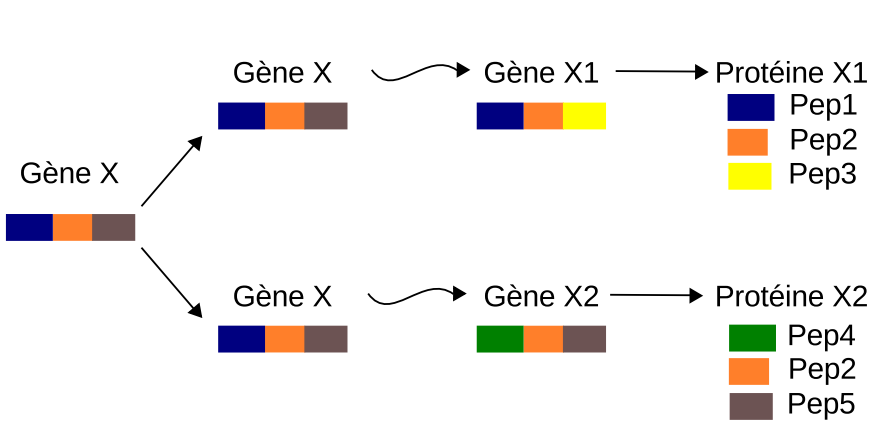
Pour plus d'info:
Webb-Robertson et al., 2015, J. Proteome Res. 14, 1993-2001

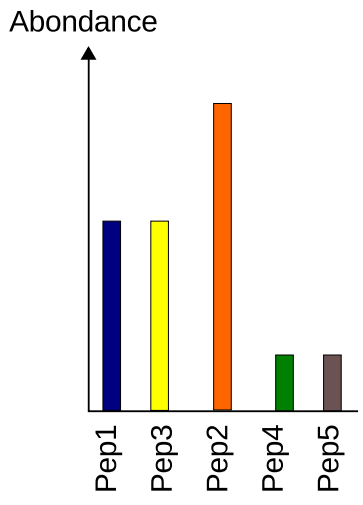
Post-processing 4: estimation des abondances de protéine

- Deux difficultés majeures:
  - les peptides communs à plusieurs protéines. Issus de l'épissage alternatif ou des gènes dupliqués, sont généralement supprimés car difficile de déconvoluer l'information qu'ils portent

Post-processing 4: estimation des abondances de protéine

- Deux difficultés majeures:
  - les peptides communs à plusieurs protéines. Issus de l'épissage alternatif ou des gènes dupliqués, sont généralement supprimés car difficile de déconvoluer l'information qu'ils portent





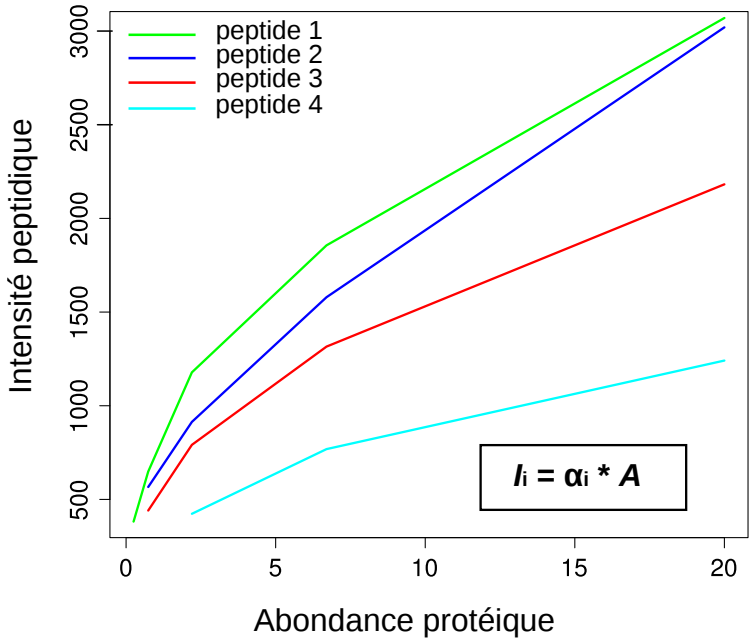
Abondance_Pep1 = Abondance_X1
Abondance_Pep3 = Abondance_X1
Abondance_Pep2 = Abondance_X1 + Abondance_X2

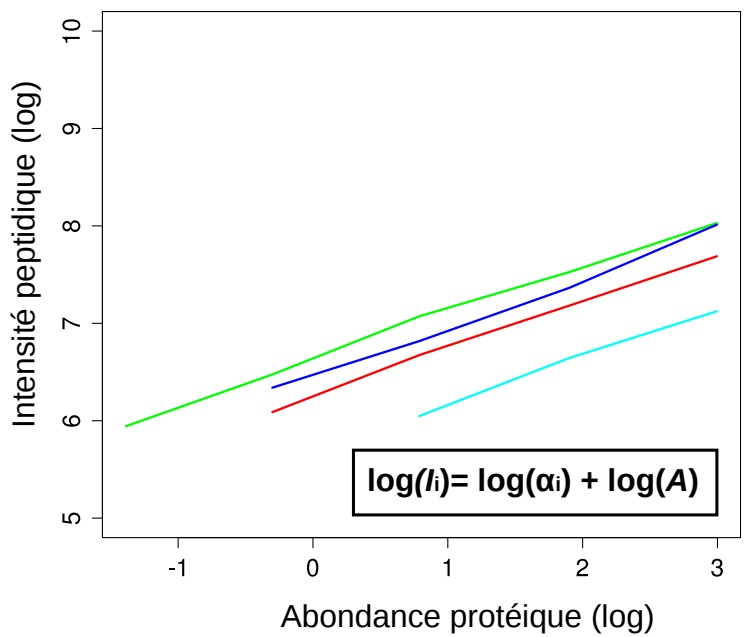
Post-processing 4: estimation des abondances de protéine

- Deux difficultés majeures:
  - les peptides communs à plusieurs protéines. Issus de l'épissage alternatif ou des gènes dupliqués, sont généralement supprimés car difficile de déconvoluer l'information qu'ils portent
  - variabilité du potentiel d'ionisation des peptides => difficulté de comparer les abondances entre échantillons en cas de données manquantes et entre protéines

Post-processing 4: estimation des abondances de protéine

- Deux difficultés majeures:
  - les peptides communs à plusieurs protéines. Issus de l'épissage alternatif ou des gènes dupliqués, sont généralement supprimés car difficile de déconvoluer l'information qu'ils portent
  - variabilité du potentiel d'ionisation des peptides => difficulté de comparer les abondances entre échantillons en cas de données manquantes et entre protéines





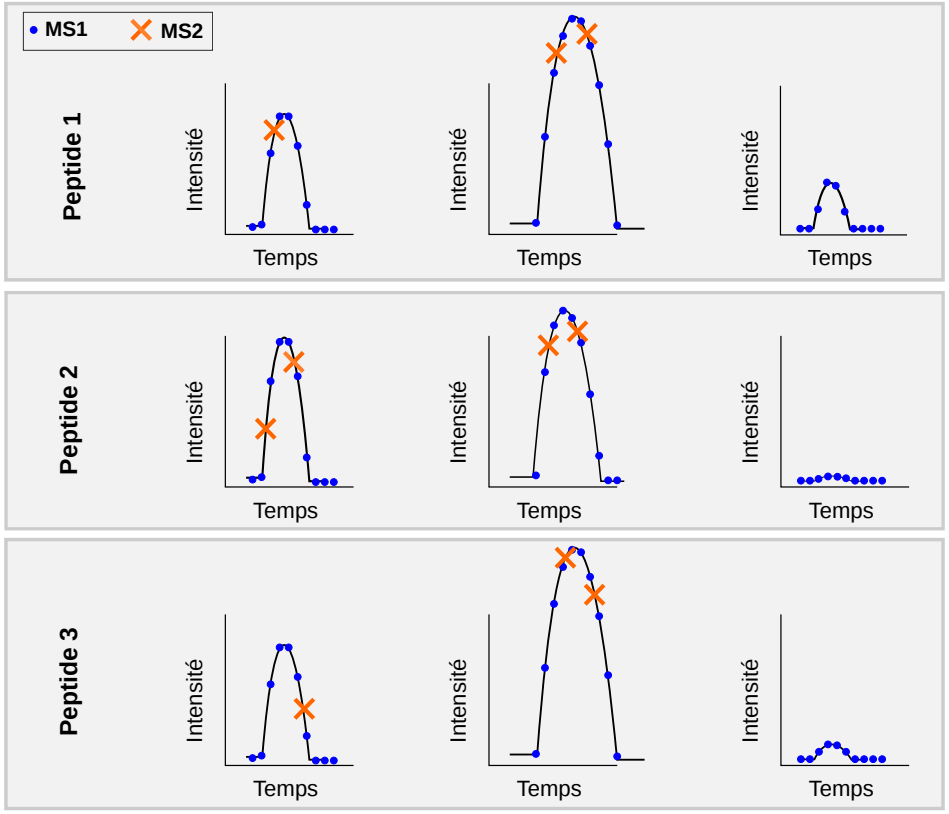
Post-processing 4: estimation des abondances de protéine

- Les méthodes “simples ”: somme ou moyenne des intensités de peptide
  - Avantage: faciles à mettre en oeuvre
  - Inconvénients: pas de prise en compte de l’effet peptide ni des peptides communs
- Les méthodes basées sur de la modélisation statistique : les mesures sur les peptides appartenant à une même protéine sont considérées comme des répétitions de la mesure sur la protéine
  - Avantages: prise en compte de l’effet peptide et possibilité de prendre en compte les peptide communs
  - Inconvénients: selon le modèle employé, la mise en oeuvre peut être complexe et les temps de calcul longs

Pour plus d’info:
Blein-Nicolas et al., 2016, Biochim. Biophys. Acta. 1864, 883-895

Principe

- Comptage du nombre de spectres MS2 ou de pics chromatographiques attribués à chaque protéine



Total pour la protéine	Sample 1	Sample 2	Sample 3
	4 spectres MS2 3 pics	6 spectres MS2 3 pics	0 spectre MS2 1 pic

Peak counting vs spectral counting

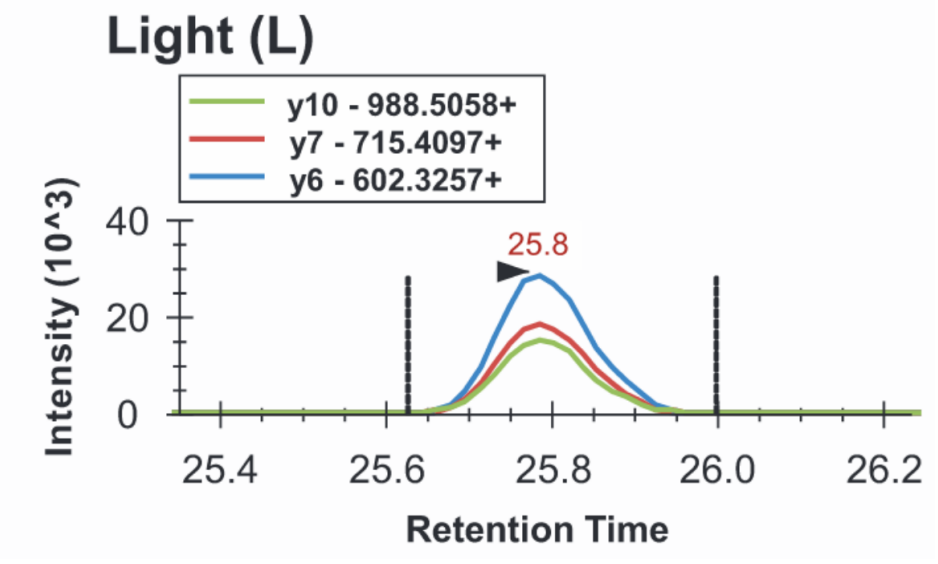
- Peak counting (PC) plus complexe à mettre en oeuvre (besoin d'un logiciel de quanti) mais possibilité de filtrer les données peptidiques (post-processing)
- Le PC compte les peptides identifiés par appariement : données plus complètes que pour le SC
- Il ne compte qu'une fois un pic quelque soit sa taille : moins dynamique que le SC

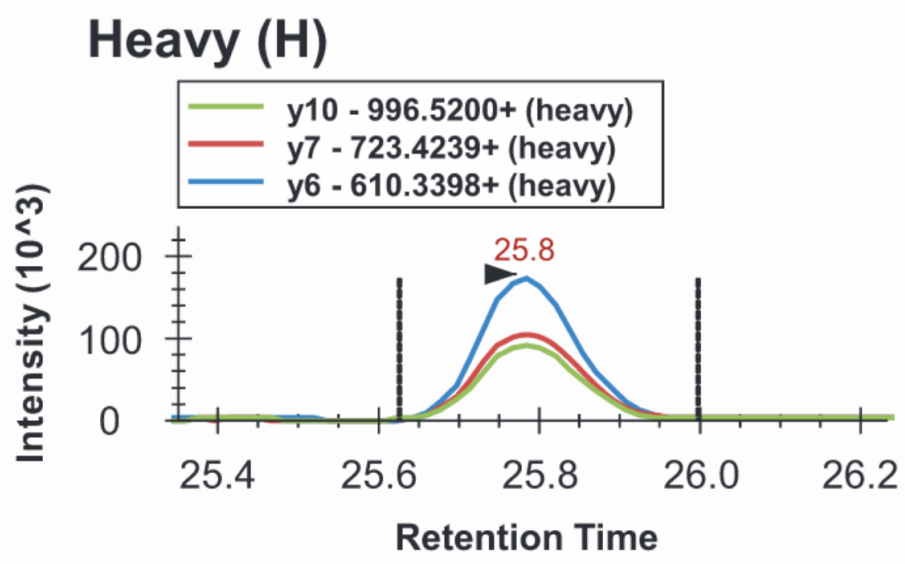
Approches semi-quantitatives vs approche quantitative

- Approches semi-quantitatives:
  - pas de données manquantes mais des zéros, ce qui permet d'analyser les variations de présence/absence
  - pas de difficulté lié au potentiel d'ionisation des peptides
  - quantification peu précise, qui ne permet pas de détecter les variations d'abondance fines
- Complémentarité des deux types d'approches: les protéines présentant trop de données manquantes pour être analysées sur la base des XIC peuvent quand même être analysées par les approches semi-quantitatives

Principe

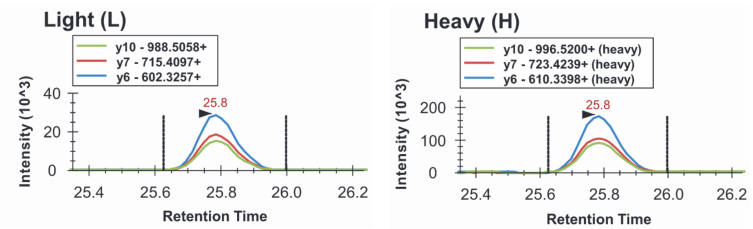
- Mesure de l'aire sous les pics chromatographiques en MS2 pour les différentes versions isotopiques des fragments d'un peptide
- Les données obtenues sont des intensités de fragments peptidiques





Post-processing

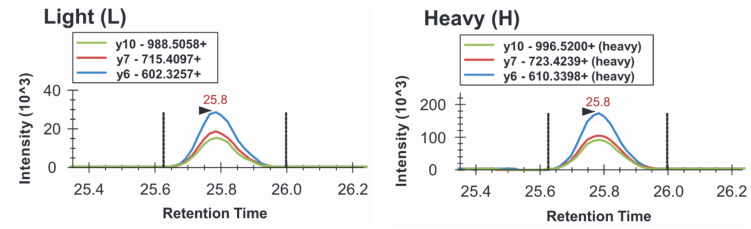
- pas de filtre sur les transitions si sélectionnées en amont
- pour chaque protéine ciblée et chaque échantillon, calcul du ratio L/H



Transition	Intensity light	Intensity heavy
y10	30000	170000
y7	18000	100000
y6	16000	90000

Post-processing

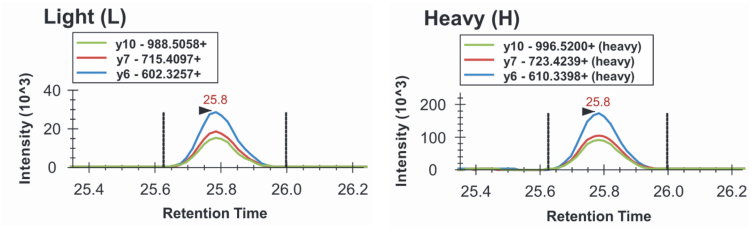
- pas de filtre sur les transitions si sélectionnées en amont
- pour chaque protéine ciblée et chaque échantillon, calcul du ratio L/H



Transition	Intensity light	Intensity heavy	Ratio L/H
y10	30000	170000	0.177
y7	18000	100000	0.18
y6	16000	90000	0.178

Post-processing

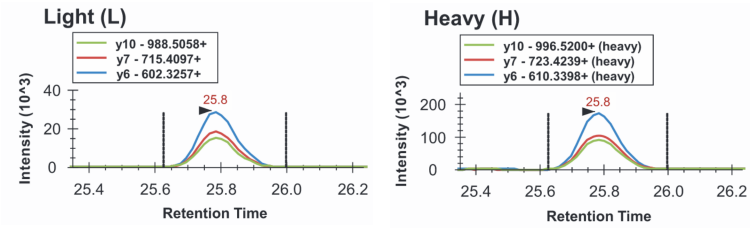
- pas de filtre sur les transitions si sélectionnées en amont
- pour chaque protéine ciblée et chaque échantillon, calcul du ratio L/H



Transition	Intensity light	Intensity heavy	Ratio L/H	Mean ratio
y10	30000	170000	0.177	
y7	18000	100000	0.18	0.1783
y6	16000	90000	0.178	

Post-processing

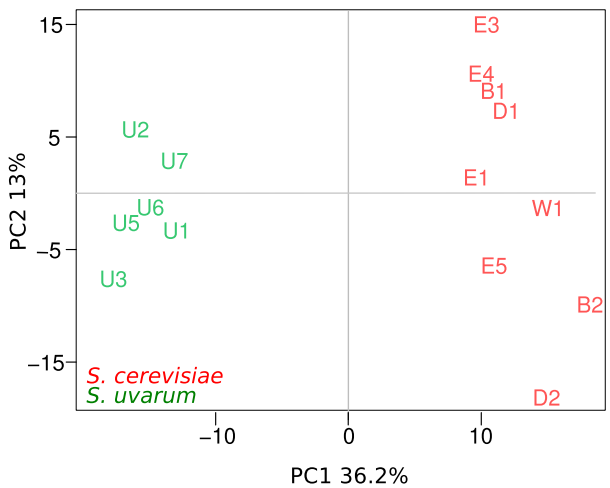
- pas de filtre sur les transitions si sélectionnées en amont
- pour chaque protéine ciblée et chaque échantillon, calcul du ratio L/H

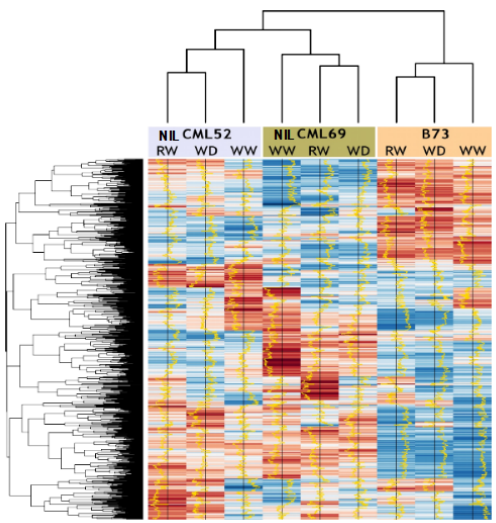


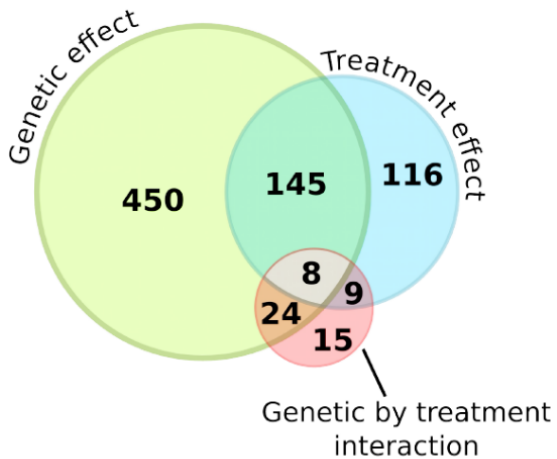
Transition	Intensity light	Intensity heavy	Ratio L/H	Mean ratio
y10	30000	170000	0.177	
y7	18000	100000	0.18	0.1783
y6	16000	90000	0.178	
sum	64000	360000	0.178	

Analyse statistique des données

- Statistiques descriptives: ACP, clustering, heatmap
- Statistiques décisionnelles: ANOVA, t-test
- Interpréation biologique: lien avec la fonction des protéines, données phénotypiques, transcriptomiques, génétiques, environnementales, etc.







Outils pour l'analyse statistique

- Plusieurs logiciels d'analyse des données de protéomique vont jusqu'aux statistiques : MaxQuant(Perseus), Skyline
- Paquets R développés spécifiquement pour l'analyse des données de protéomique:
  - MSstats (Choi et al., 2014, Bioinformatics, 30: 2524-2526)
  - DAPAR & ProStar (Wieczorek et al., 2017, Bioinformatics, 33, 135–13)
  - SafeQuant (Ahrné, unpublished, https://github.com/eahrne/SafeQuant/)
  - GiaPronto (Weiner et al., 2017, Mol. Cell. Prot. doi: 10.1074/mcp.TIR117.000438)
  - MCQ (PAPPSO, unpublished)
- Paquet R 'mixOmics' (http://mixomics.org/a-propos/publications/) pour l'intégration de données omics

Merci de votre attention!