Vigrid/MSDA: la grille de calcul au service de l'analyse protéomique

Christine CARAPITO, Jérôme PANSANEL, Patrick GUTERL, Alexandre BUREL,
Fabrice VARRIER, Fabrice BERTILE, Stéphane GENAUD, Alain VAN DORSSELAER, Christelle ROY

Journées Scientifiques Mésocentres et France Grilles, Paris, 1 octobre 2012

Du génome au protéome

Génome → **Transcriptome** → **Protéome**

ADN — **Transcription** — ARNm — **Traduction** — Protéines

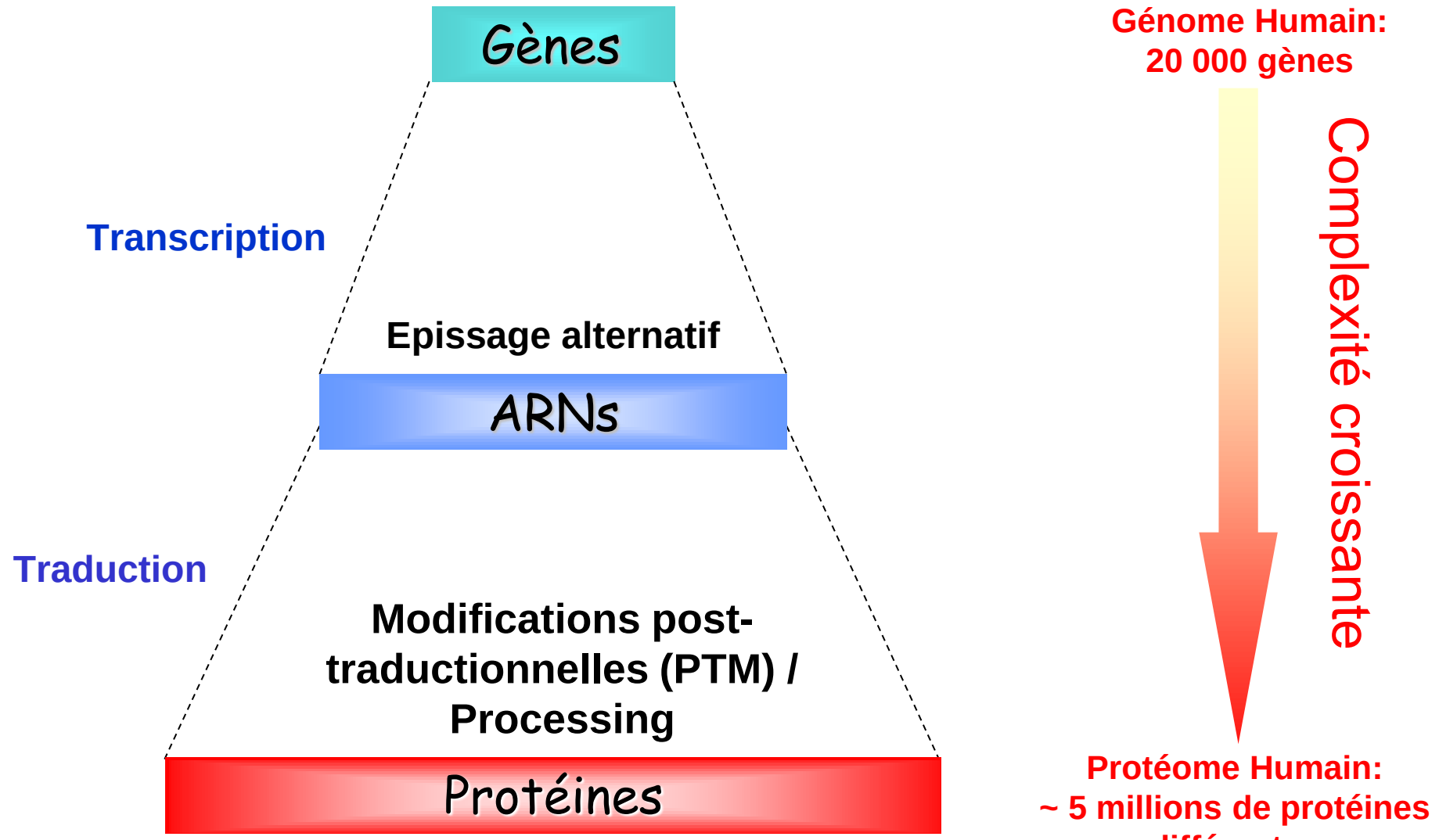
Code à 4 lettres
A, T, G, C

→ Le code génétique →

Code à 20 lettres
20 acides aminés

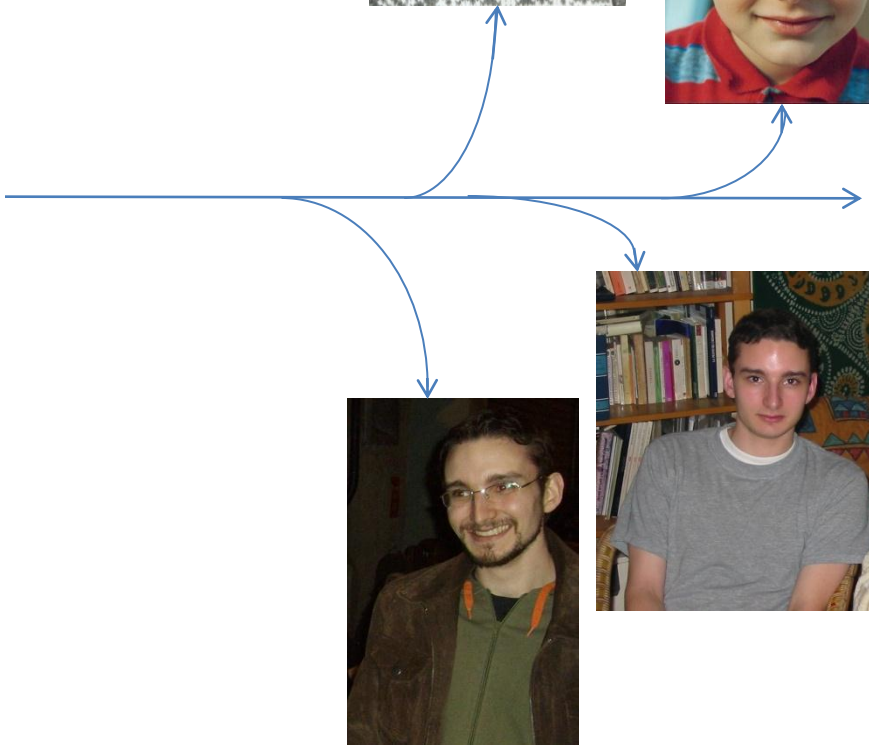
1er nucléotide	2e nucléotide T	2e nucléotide C	2e nucléotide A	2e nucléotide G	3e nucléotide
T	TTT phénylalanine	TCT sérine	TAT tyrosine	TGT cystéine	T
	TTC phénylalanine	TCC sérine	TAC tyrosine	TGC cystéine	C
	TTA leucine	TCA sérine	TAA codon-stop	TGA codon-stop	A
	TTG leucine	TCG sérine	TAG codon-stop	TGG tryptophane	G
C	CTT leucine	CCT proline	CAT histidine	CGT arginine	T
	CTC leucine	CCC proline	CAC histidine	CGC arginine	C
	CTA leucine	CCA proline	CAA glutamine	CGA arginine	A
	CTG leucine	CCG proline	CAG glutamine	CGG arginine	G
A	ATT isoleucine	ACT thréonine	AAT asparagine	AGT sérine	T
	ATC isoleucine	ACC thréonine	AAC asparagine	AGC sérine	C
	ATA isoleucine	ACA thréonine	AAA lysine	AGA arginine	A
	ATG méthionine	ACG thréonine	AAG lysine	AGG arginine	G
G	GTT valine	GCT alanine	GAT acide aspartique	GGT glycine	T
	GTC valine	GCC alanine	GAC acide aspartique	GGC glycine	C
	GTA valine	GCA alanine	GAA acide glutamique	GGA glycine	A
	GTG valine	GCG alanine	GAG acide glutamique	GGG glycine	G

Du génome au protéome



Le protéome est évolutif et dynamique

Un génome …



… plusieurs protéomes !

L’analyse protéomique

Définition: C’est l’analyse de l’ensemble des protéines exprimées par un type cellulaire, un tissu ou un fluide biologique (sang, urine), à un instant donné et avec un historique donné.

Technique: L’analyse protéomique repose sur l’interprétation des données de Spectrométrie de Masse.

Instruments de type
Quadrupole-TOF

Instruments de type
Trappe ionique

Nature des données: Des tableaux de masses moléculaires très précises (7 digits).

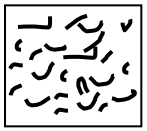
L’analyse protéomique

Mélange de protéines

10^4 molécules (protéines) de 500 acides aminés

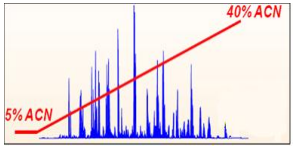
Coupure enzymatique

Mélange de peptides



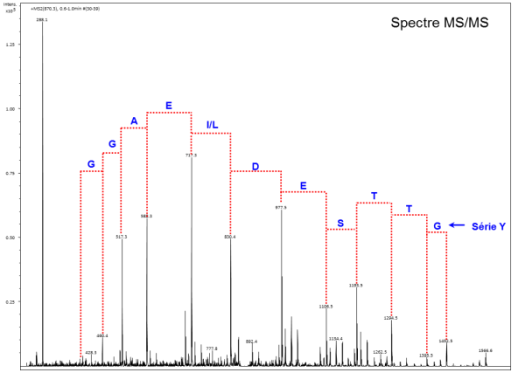
5.10^5 molécules (peptides) de 10 acides aminés

Séparation



Electrophorèse, Chromatographie, …

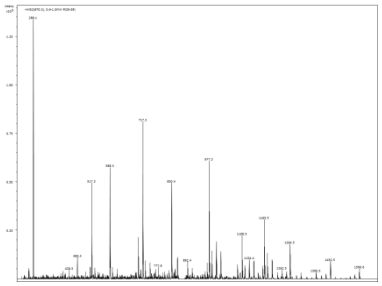
Analyse MS/MS



Spectrométrie de masse

L'interprétation des données de l'analyse protéomique

Spectre MS/MS



Liste de masses expérimentales

Banques de données de séquences protéiques

>Albumin
ILPMVCCDEKTISHEDAVFRPMLVC
KHFDIYTREHPKSQDCWATTPMKF
HLKETIPRHVVCDETR...

Séquences connues ou « théoriques »:
Jusqu'à 1Go de fichier texte

Listes de masses théoriques

MS	MS/MS
546,45	789,67 876,43 999,12 1018,98 1342,34 1597,09 1678,95 2202,22

Algorithmes d'identification

Comparaison (confrontation) des listes de masses expérimentales/théoriques

MS	MS/MS
546,43	789,69 876,41 987,50 999,14 1018,97 1342,30 1597,11 1678,99 1987,60 2202,24

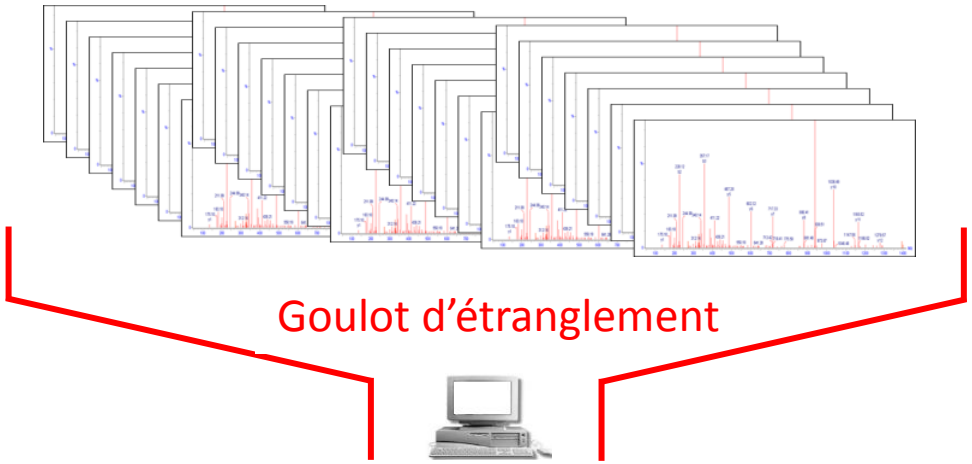
L'interprétation des données de l'analyse protéomique

Quelques chiffres:

Un spectromètre de masse génère 8000 spectres MS/MS par heure

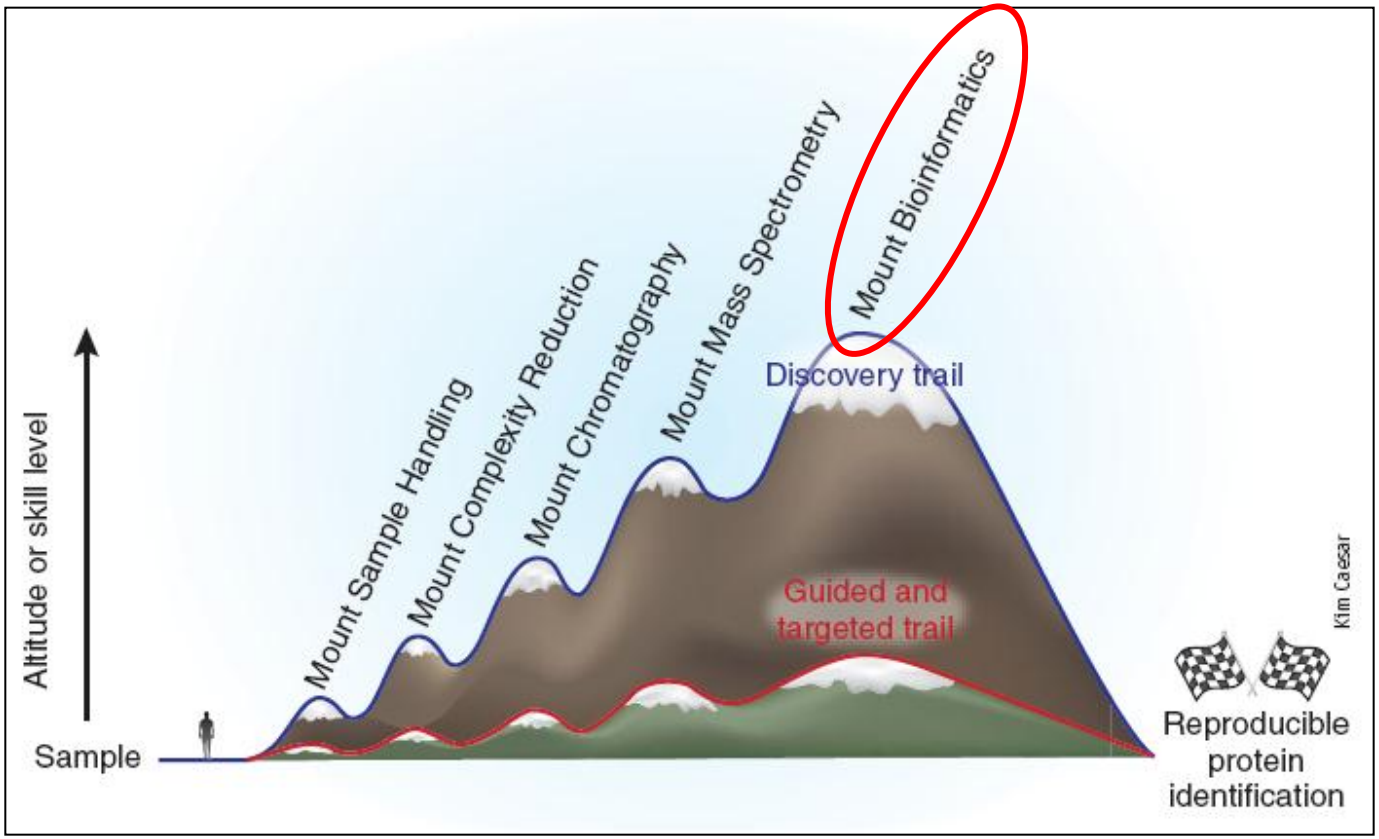
10 instruments au laboratoire travaillant 24h/24h

≅ 2 millions de spectres MS/MS acquis par jour



Interprétation par confrontation des données expérimentales/théoriques

L’interprétation des données de l’analyse protéomique



D’après R. Aebersold, « A stress test for mass spectrometry-based proteomics », Nature Methods, 6 (6), 411-412, June 2009.

L’interprétation des données est reconnue comme un verrou majeur de l’analyse protéomique!

L'interprétation des données de l'analyse protéomique

Un avantage:

Chaque spectre MS/MS peut être interprété individuellement dans la banque de données de séquences protéiques choisie.

- **Utilisation de la grille adaptée**

Avec néanmoins une exigence:

Un unique spectre MS/MS peut être à l'origine de l'identification de LA protéine d'intérêt (le biomarqueur par exemple).

- **Développement de Vigrid et MSDA**

MSDA est une plateforme d'outils d'interprétation des données de protéomique.
Vigrid est une application indépendante développée pour permettre aux outils de MSDA de s'exécuter sur la grille.

Vigrid

Dans la mythologie nordique:

- Vigrid est le dieu des histoires
- Marié à Sága, déesse des contes et des légendes

Trois objectifs pour Vigrid:

1. Optimisation des temps de latence des jobs
2. Supervision des jobs
3. Tolérance aux pannes

➢ Utilisation de la librairie JSAGA (développée au Centre de Calcul de Lyon)

Java implementation of the Simple API for Grid Applications (SAGA) specification from the Open Grid Forum (OGF).

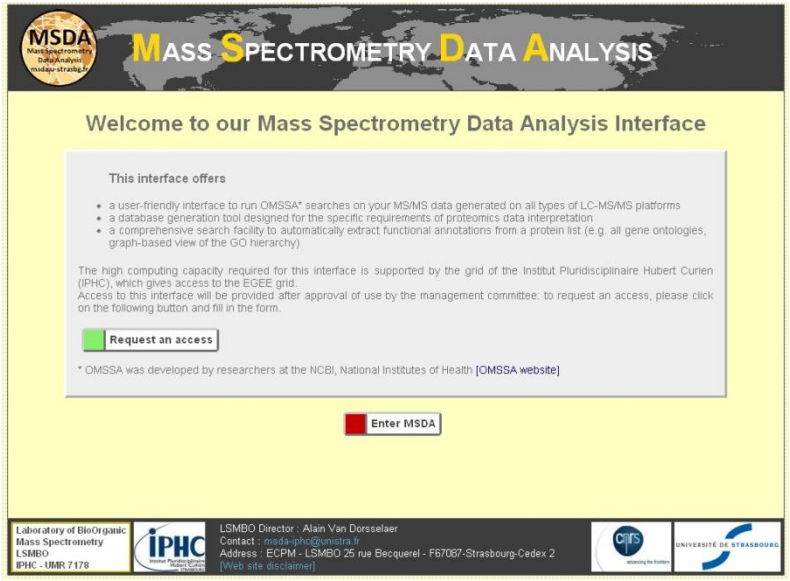
Vigrid

Pour les détails techniques, voir démo durant la session de poster (Patrick Guterl et Alexandre Burel)

- Vigrid permet d'optimiser le temps d'exécution par une gestion dynamique et de garantir l'interprétation de l'ensemble des données de spectrométrie de masse.

MSDA (Mass Spectrometry Data Analysis)

La suite logicielle MSDA https://msda.unistra.fr



MSDA contient déjà 4 modules:

1. Une boite à outils de génération des banques de données
2. Une interface conviviale de lancement de requêtes OMSSA sur la grille
3. Une boite à outils d'annotations fonctionnelles
4. Une interface conviviale pour l'interprétation *de novo* des données MS/MS sur la grille

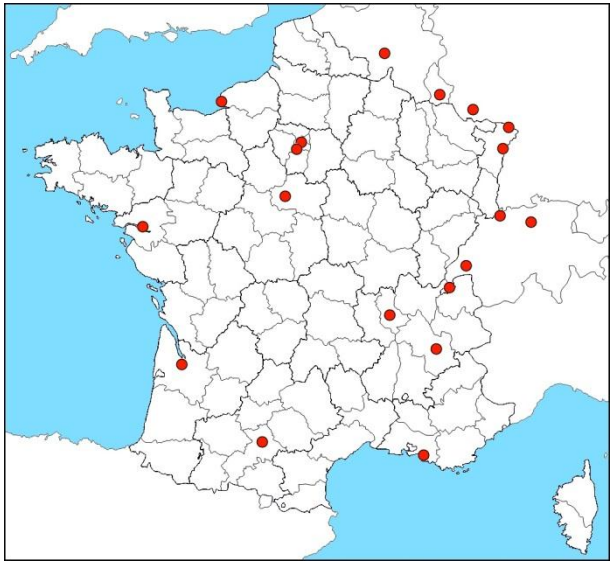
MSDA (Mass Spectrometry Data Analysis)

Depuis sa présentation au congrès de Spectrométrie de Masse et d'Analyse Protéomique en Avignon en Septembre 2011, MSDA compte :

≅ 30 utilisateurs internes à l'IPHC

et ≅ 50 laboratoires extérieurs autorisés

(laboratoires et plateformes de protéomique français et européens, industriels)



Répartition des utilisateurs de MSDA en juillet 2012,
Et depuis de nouveaux utilisateurs en Allemagne et au Brésil.

MSDA (Mass Spectrometry Data Analysis)

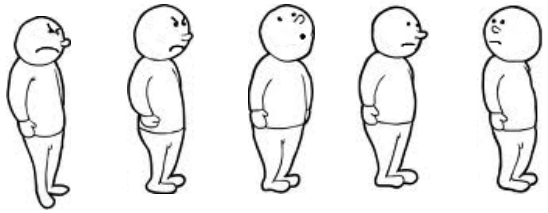
L'utilisation de la grille nous a permis:

Un gain de temps net

Type d'expérience de spectrométrie de masse	Lancement sur un serveur local (h)	Lancement sur la grille (h)	Facteur de gain de temps
1.Spectrométrie de masse à haute résolution et spécificité enzymatique complète	0,3	0,2	1,5
2.Spectrométrie de masse à haute résolution et semi-spécificité enzymatique	5,8	0,5	12
3.Spectrométrie de masse à basse résolution et spécificité enzymatique complète	3,6	0,7	5
4.Spectrométrie de masse à basse résolution et semi-spécificité enzymatique	74,6	0,95	79

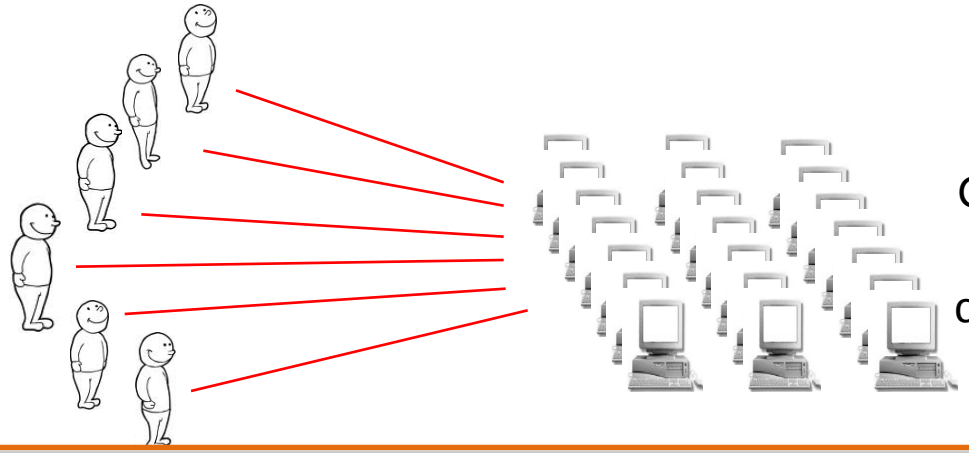
L'utilisation de la grille nous a permis:

Un gain de temps global



Serveur local

- Plus de file d'attente
- Plus il y a d'utilisateurs, plus le gain de temps est important
- Des études trop gourmandes en ressources auparavant deviennent envisageables

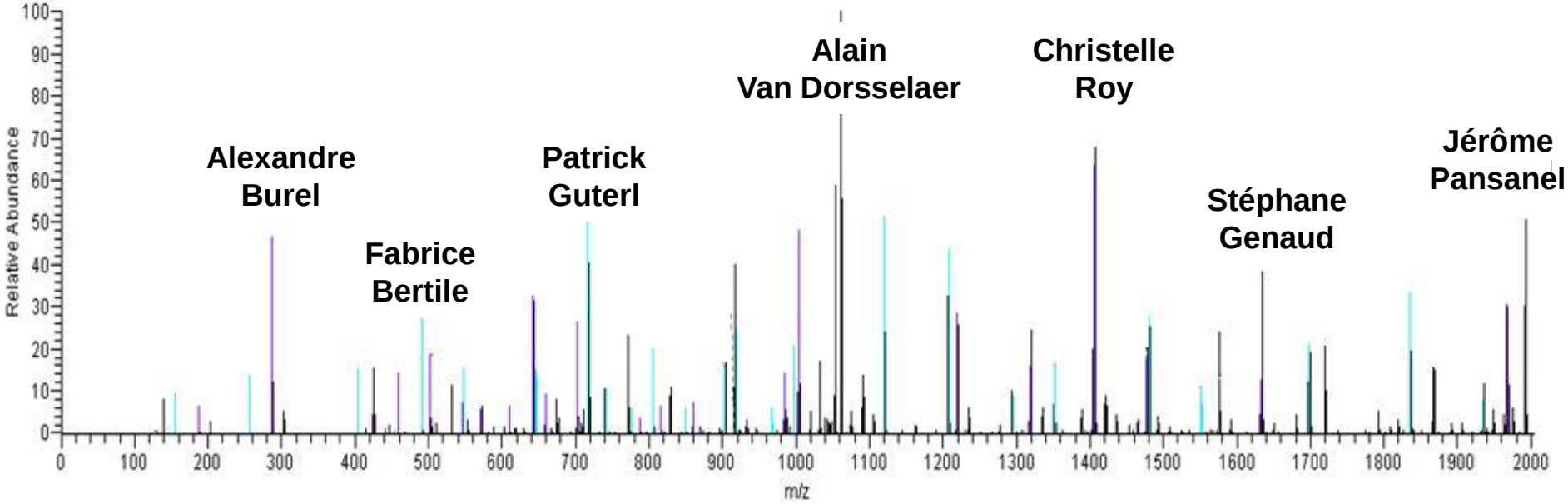


Grille de calcul

Conclusions

- La grille de calcul est adaptée pour l'interprétation des données de protéomique.
- Le goulot d'étranglement de l'interprétation des données de protéomique est désengorgé.
- L'utilisation de la grille ouvre de nouvelles possibilités d'interprétation des données de protéomique, non envisageables jusqu'ici: des projets censurés auparavant deviennent possibles!

Merci!


Alexandre Burel
Fabrice Bertile
Patrick Guterl
Alain Van Dorsselaer
Christelle Roy
Stéphane Genaud
Jérôme Pansanel
Relative Abundance
m/z