De la fiabilité des données d'identification et de quantification de protéines par MS

Myriam Ferro

BGE/EDyP Laboratory

CEA/Grenoble

29/11/2012

Atelier Prospectom 29-30 novembre 2012, Grenoble

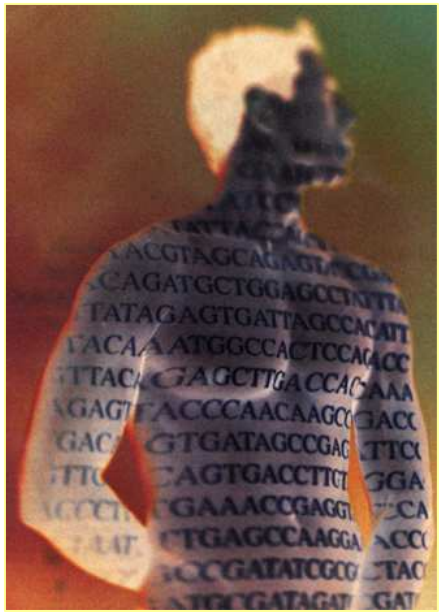
Definition of proteomics

GENES → TRANSCRIPTS (RNA) → PROTEINS

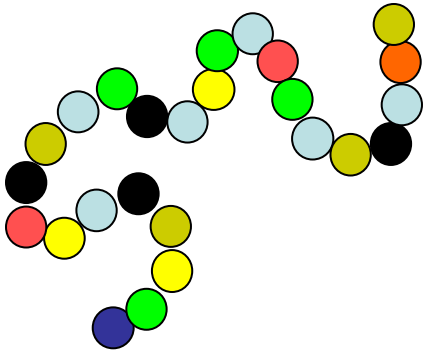
Genomics

Transcriptomics

Proteomics



Dynamics
(differential expression: localization, time)



- Processing
- PTMs
- Localization
- Partners
- Etc.

Proteomics: what's for ?

Some questions
- mutation or environmental conditions
- protein-protein interactions
- searching for biomarkers
- etc.

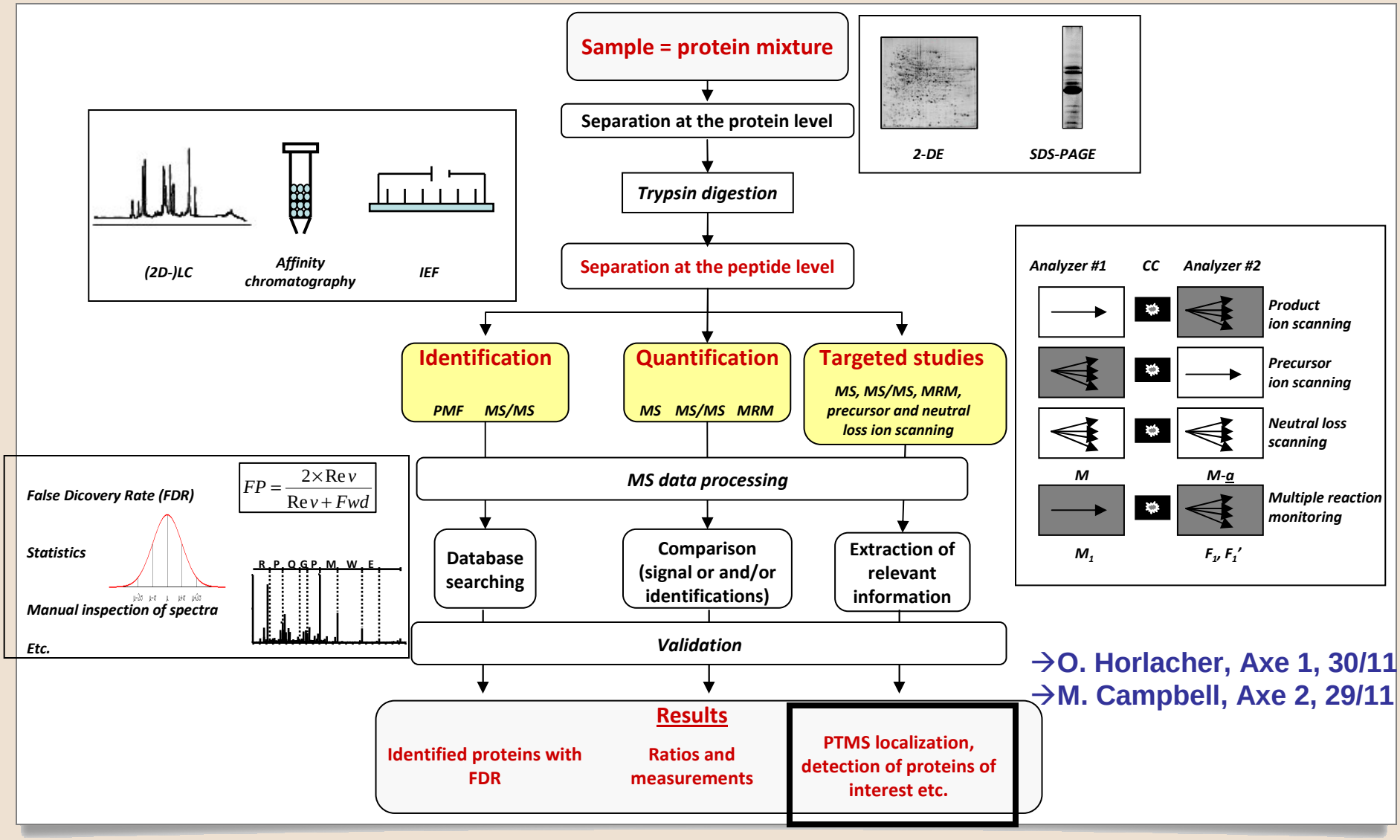
Data visualization

New knowledge and questions !

From Käll & Vitek, 2011

Käll & Vitek, Computational Mass Spectrometry-Based Proteomics, 2011, PLOS Computational Biology 7 (12)

Proteomics analyses using MS



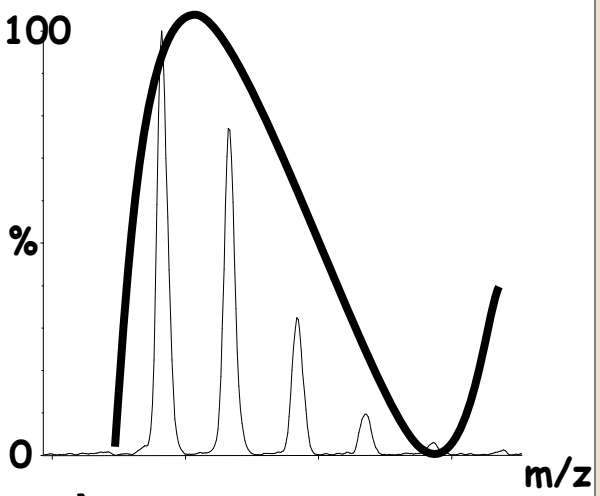
→O. Horlacher, Axe 1, 30/11
→M. Campbell, Axe 2, 29/11

Agrawal GK, Bourguignon J, Rolland N, Ephritikhine G, Ferro M, Jaquinod M, Alexiou KG, Chardot T, Chakraborty N, Jolivet P, Doonan JH, Rakwal R. Plant organelle proteomics: Collaborating for optimal cell function. Mass Spectrom Rev. 2010 Oct 29

Some reminders about mass spectrometry measurements

Reminder # 1: the molecular mass

- **Ex: EWMPGQPR = C44 H65 N13 O12 S**



- **<u>Nominal mass</u> (approximation)**

C=12; H=1; N= 14; O=16; S=32

→ M = 999

- **<u>Monoisotopic mass (the more stable isotopes)</u>**

C=12; H=1.007825; N= 14.003074; O=15.9949146; S=31.9720718

→ M = 999.4596

- **<u>Average mass (barycentre of all masses)</u>**

C=12.011; H=1.007994; N=14.006674; O=15.9994; S=32.066

→ M= 1000.1464

Reminder # 3: m/z ratio

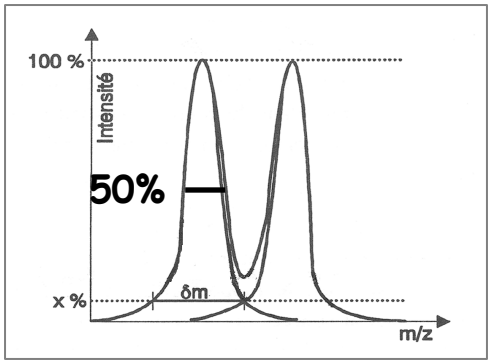
- M = molecular mass (ex: peptide)
- H+= proton
- z = state of charge

$$m/z = \frac{M + z\,H^+}{z}$$

Reminder # 4: resolution

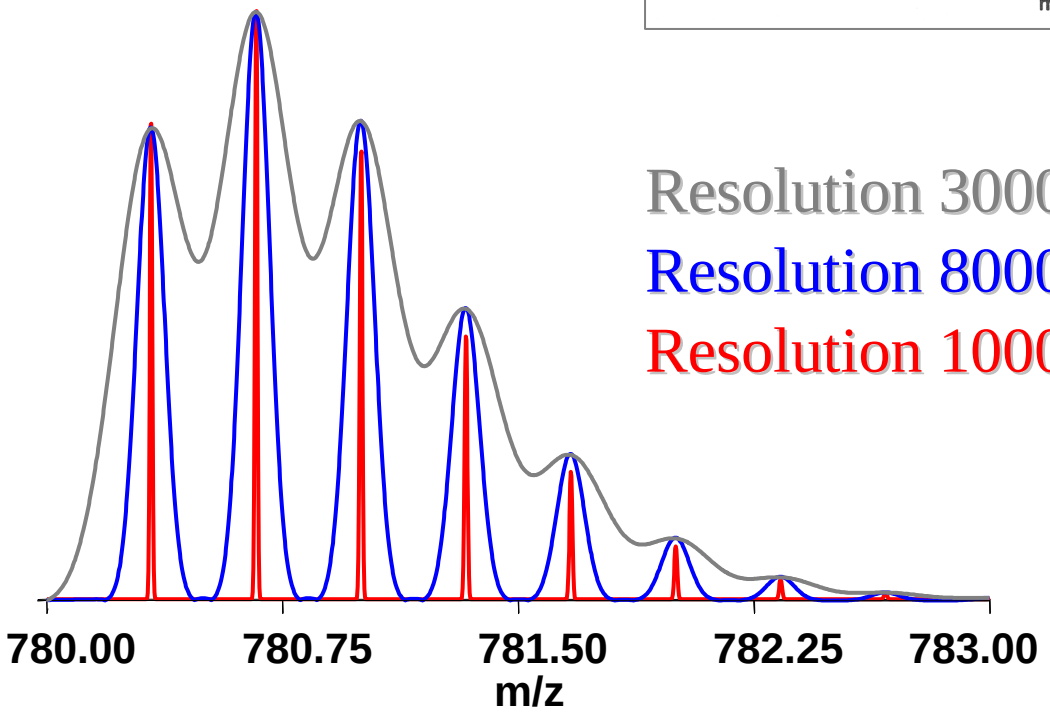
- **Characterize peaks separation**

(FWHM: Full Width at Half Maximum)



$$R = \frac{\Delta m}{m}$$

Insulin (Chain A) 3+



Resolution 3000

Resolution 8000

Resolution 100000

Accuracy and precision

High precision but low accuracy

High precision but high accuracy

Precision is related to reproducibility → careful control of instrument settings

Accuracy is related to calibration and resolution

Instrument with high resolving power (e.g OrbiTrap VELOS)
Resolution = 60 000
Mass accuracy (Orbitrap analyser) ~ 1-5 ppm

$$\frac{\left| M_{exp} - M_{th} \right|}{M_{th}}$$

ppm = part per million

Features of mass spectrometers

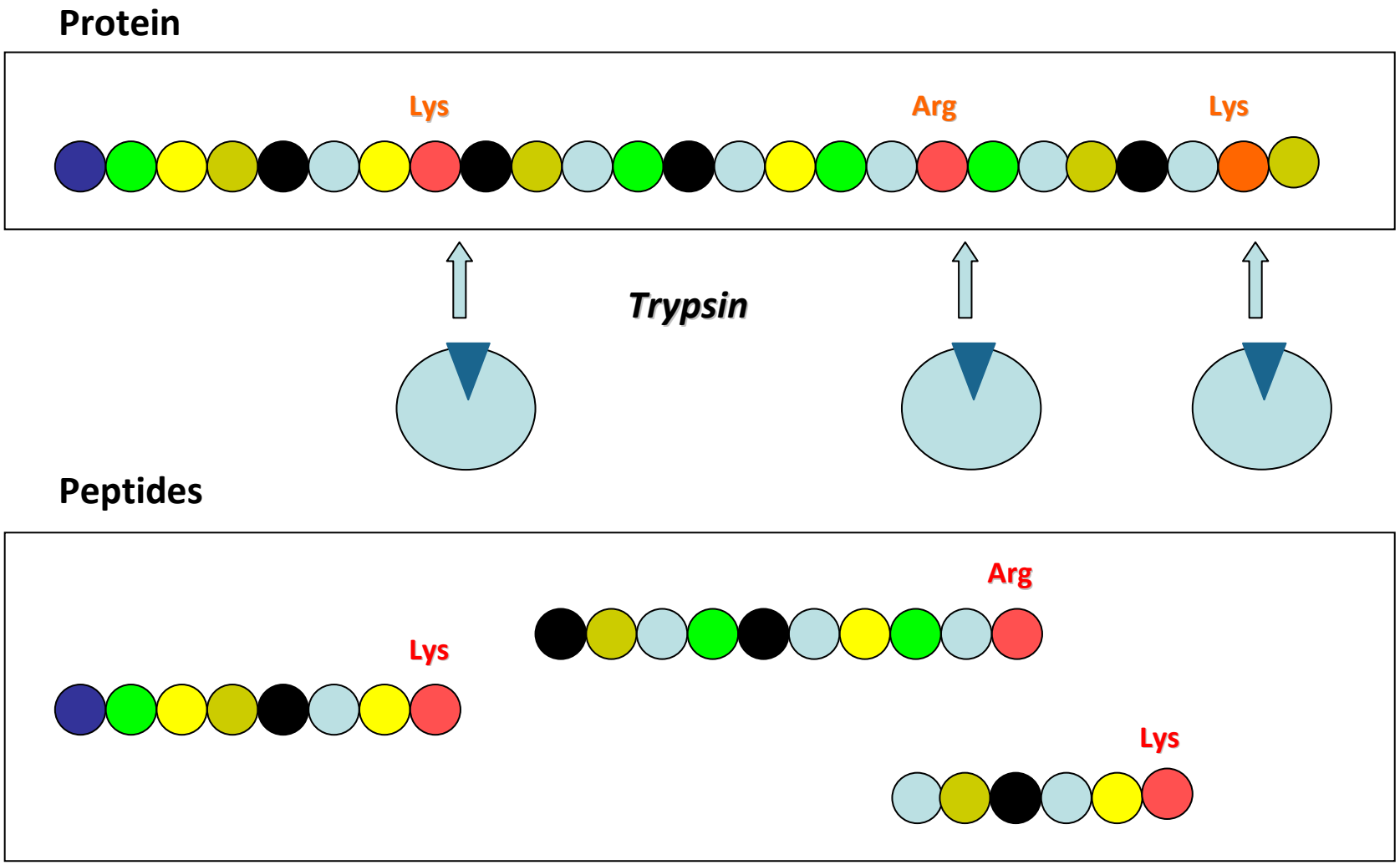
Table 1 Performance comparisons of the mass spectrometry instruments

Instrument	Applications	Resolution	Mass accuracy	Sensitivity	Dynamic range	Scan rate
LIT (LTQ)	Bottom-up protein identification in high-complexity, high-throughput analysis, LC-MS^n capabilities	2000	100 ppm	Femtomole	1e4	Fast
TQ (TSQ)	Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)	2000	100 ppm	Attomole	1e6	Moderate
LTQ-Orbitrap	Protein identification, quantification, PTM identification	100,000	2 ppm	Femtomole	1e4	Moderate
LTQ-FTICR, Q-FTICR	Protein identification, quantification, PTM identification, top-down protein identification	500,000	<2 ppm	Femtomole	1e4	Slow, slow
Q-TOF, IT-TOF	Bottom-up, top-down protein identification, PTM identification	10,000	2–5 ppm	Attomole	1e6	Moderate, fast
Q-LIT	Bottom-up peptide and protein quantification; medium complexity samples, peptide and protein quantification (SRM, MRM, precursor, product, neutral fragment monitoring)	2,000	100 ppm	Attomole	1e6	Moderate, fast

Yates et al., Annu. Rev. Biomed. Eng. 2009. 11:49–79

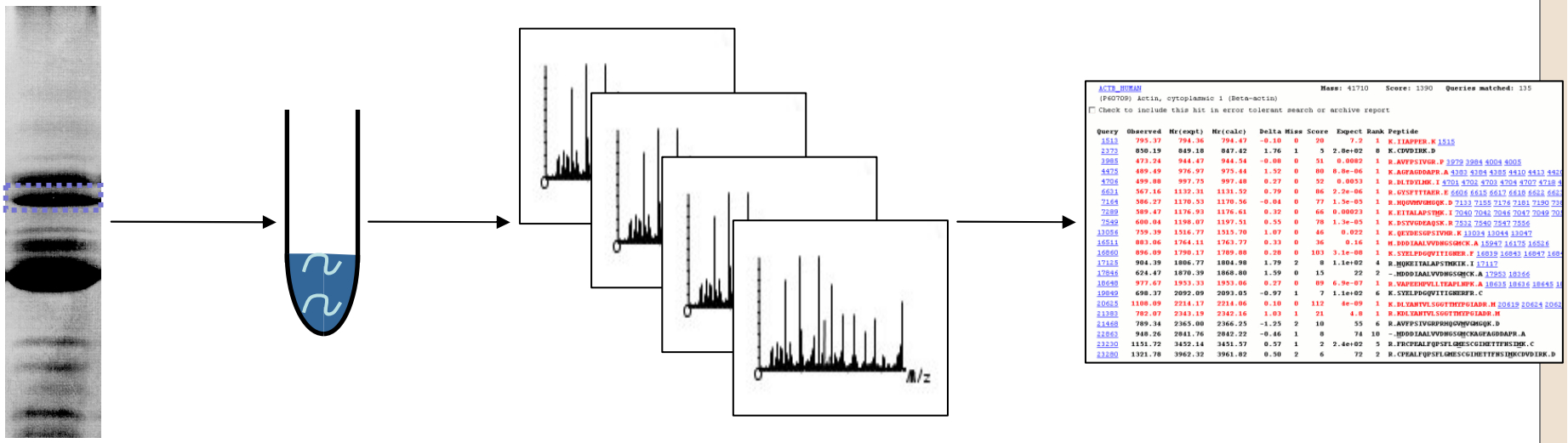
The strategies for protein identification

Trypsin digestion



A proteomics workflow

Proteins *Tryptic peptides* *MS(/MS) spectra (peak picking)* *Identification & quantification results*



Trypsin Digestion

LC-MS(/MS) analysis

Database searching & data analysis

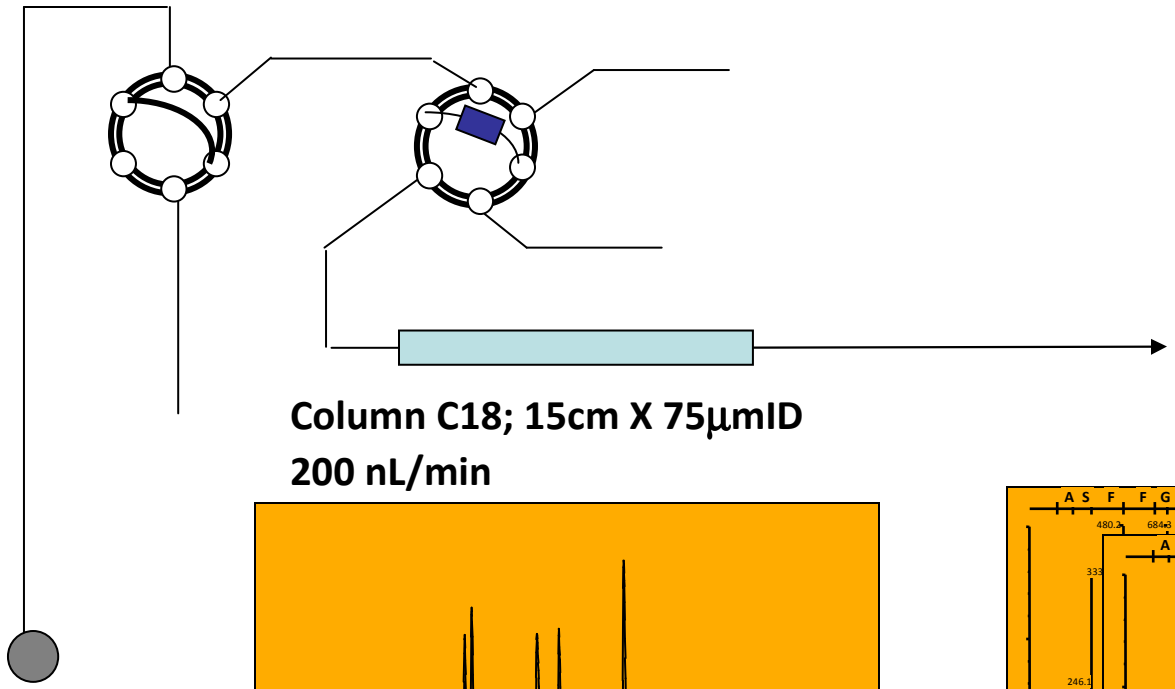
Sample preparation
Biochemistry

Mass spectrometry
Liquid chromatography

Informatics
Bioinformatics

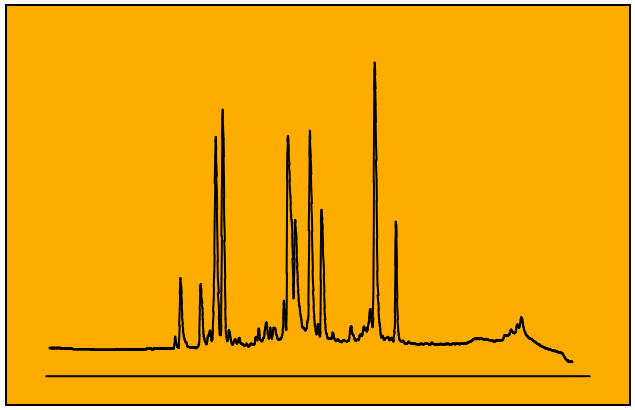
Cottrell JS. Protein identification using MS/MS data. J Proteomics. 2011 Sep 6;74(10):1842-51.

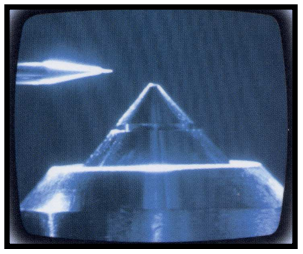
Shotgun analyses

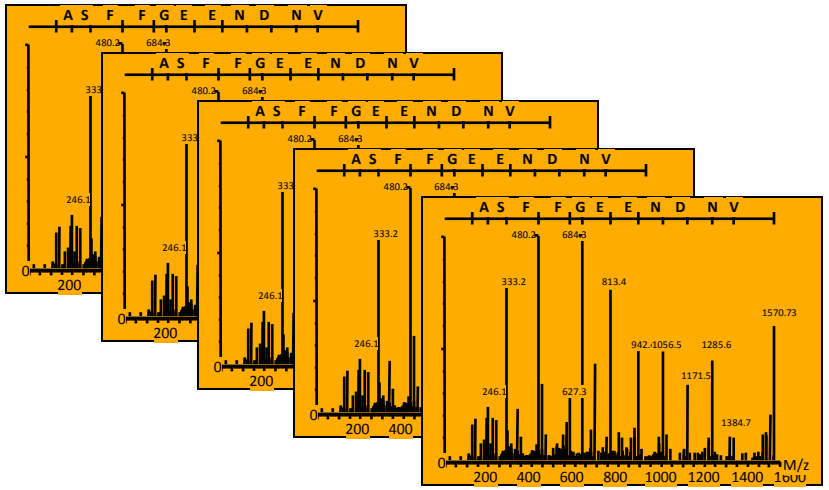


Column C18; 15cm X 75µmID

200 nL/min



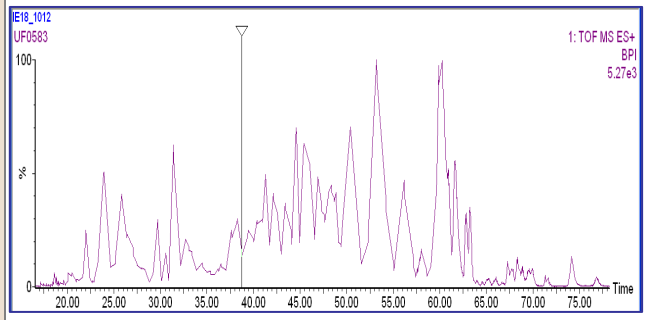


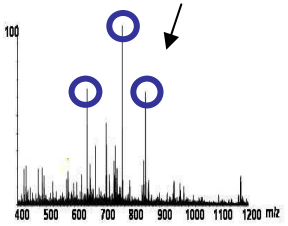


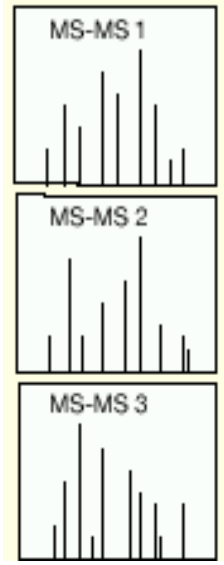
Ex: Orbitrap Velos ~ 5000 spc/hr

Separation and concentration **Automated MS/MS**

Data dependent acquisition

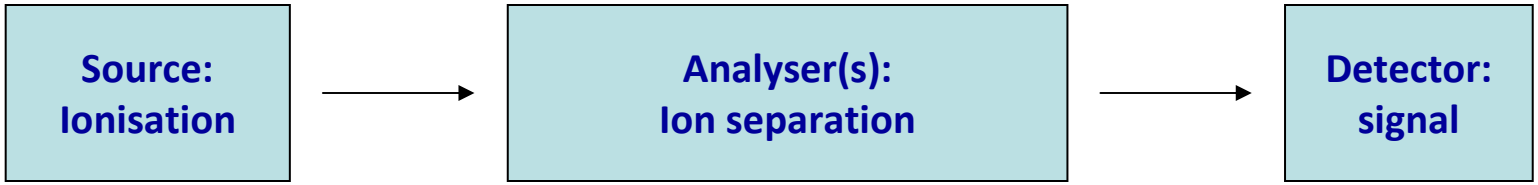




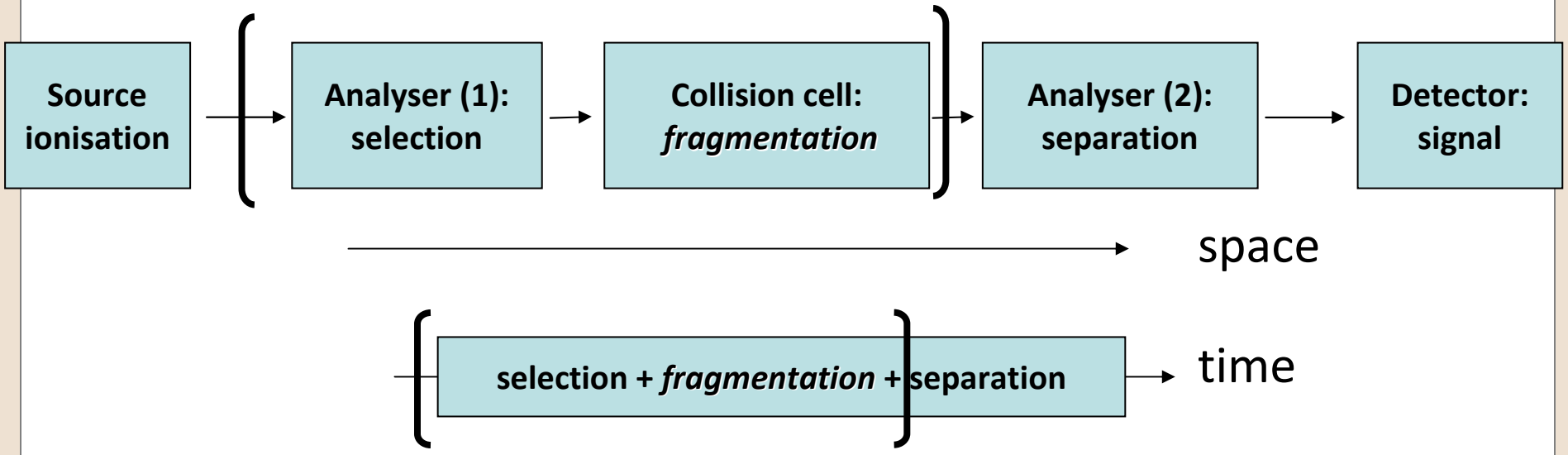


1-2 seconds (for the Orbitrap)

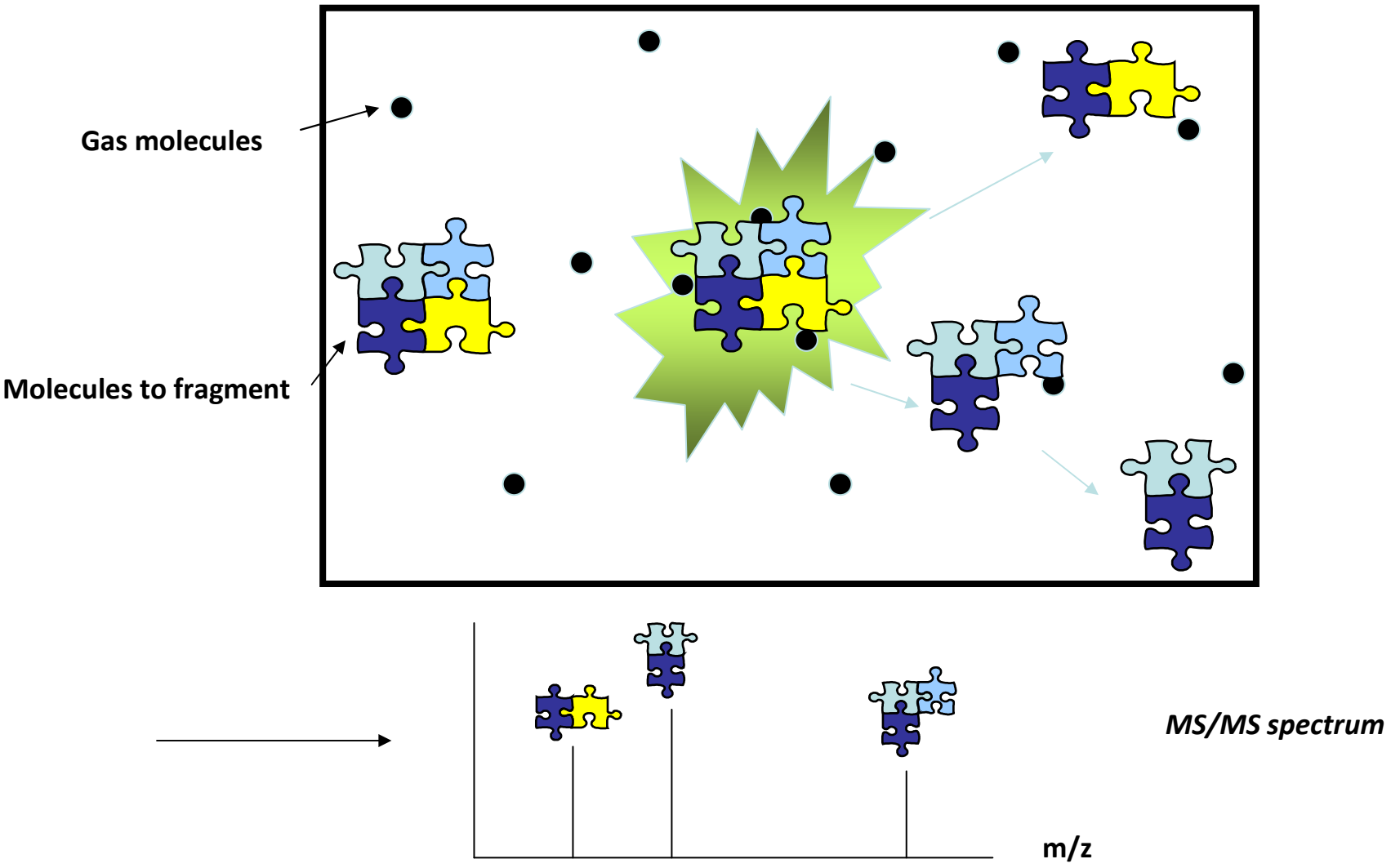
Tandem mass spectrometry



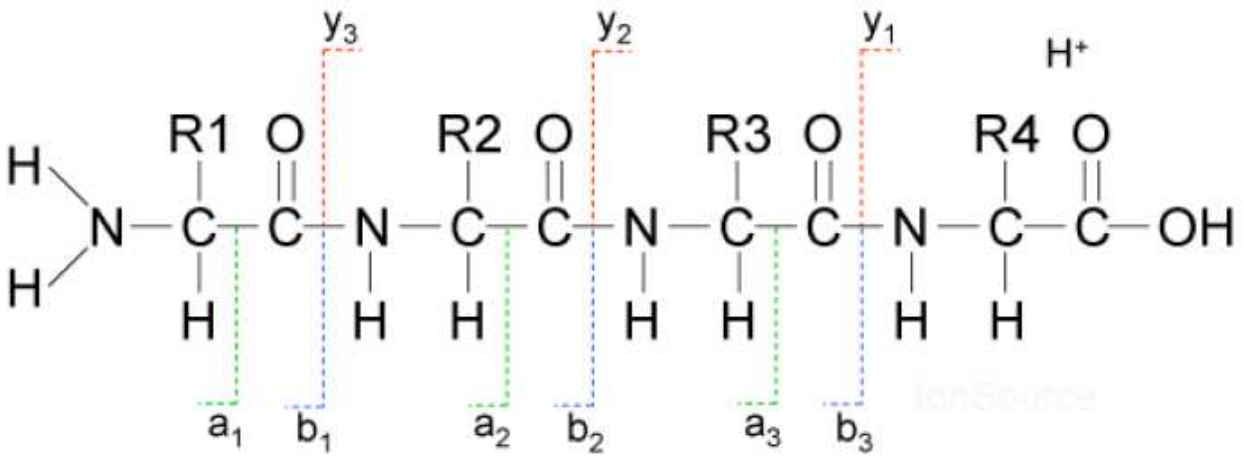
Tandem mass spectrometry (MS/MS) → peptide fragmentation

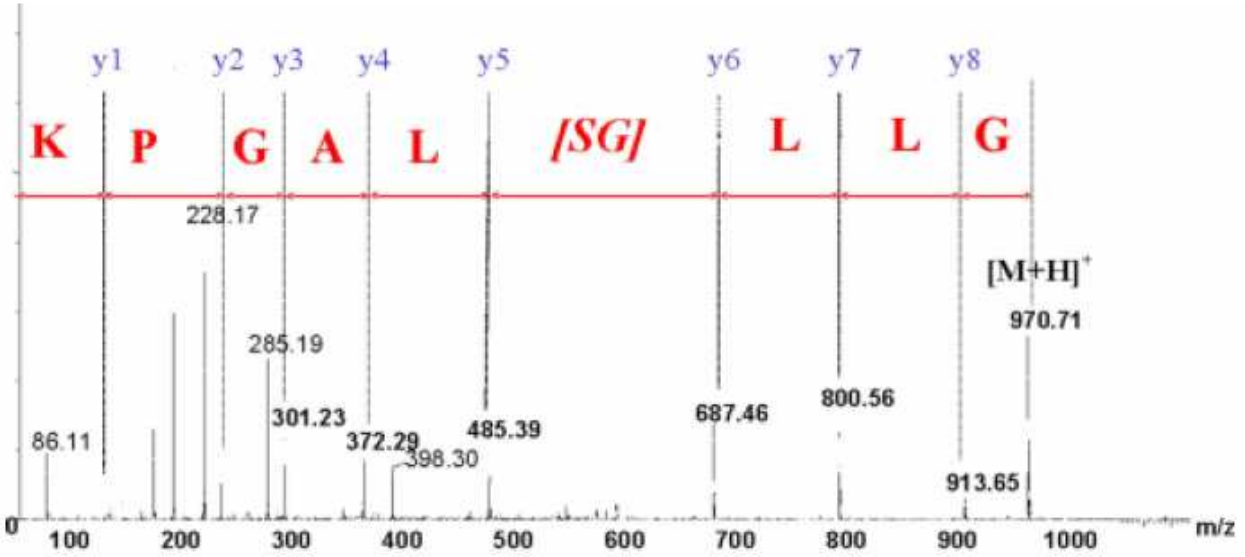


Peptide fragmentation in the mass spectrometer



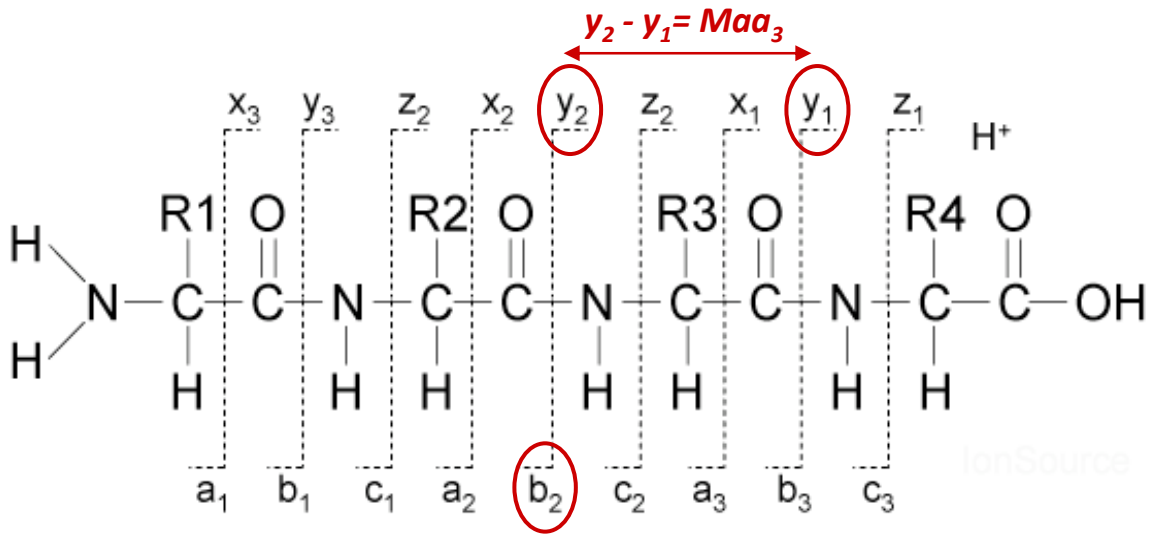
A typical MS/MS spectrum



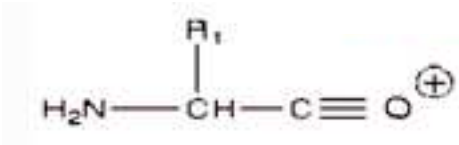


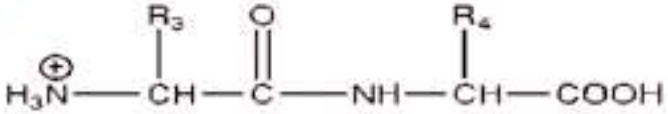
Roepstorff, P and Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* 1984; **11(11)**:601

Rules of fragmentation



Peptide bond cleaved: complementary ions y'' et b





b ion → peptide N-ter

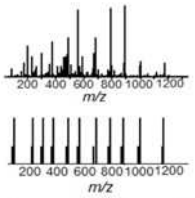
y ion → peptide C-ter

Other ions

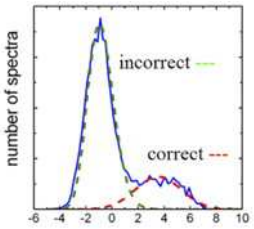
Immonium ions : specific of each amino acid ex: m/z 120 --> Phe
Internal fragments
Specific fragmentations of PTMs (cf M. Jaquinod presentation)
Neutral loss ex: - 18 Da --> perte d'H_2O

Schematic overview of a typical workflow of the proteomics informatics processing of a data set

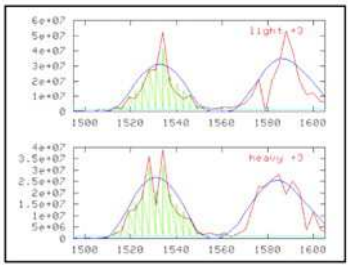
1. Conversion to and use of open data formats



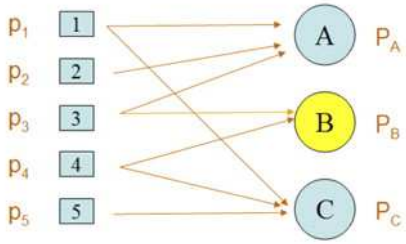
2. Spectrum identification with a search engine



3. Validation of identifications

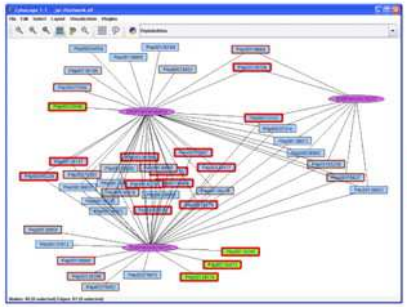


5. Quantification

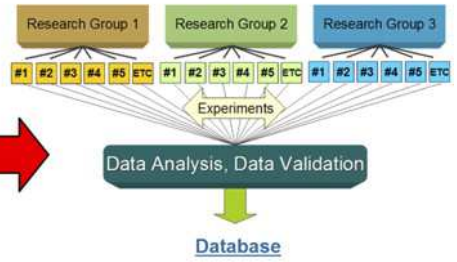


4. Protein inference

6. Organization in local data management systems



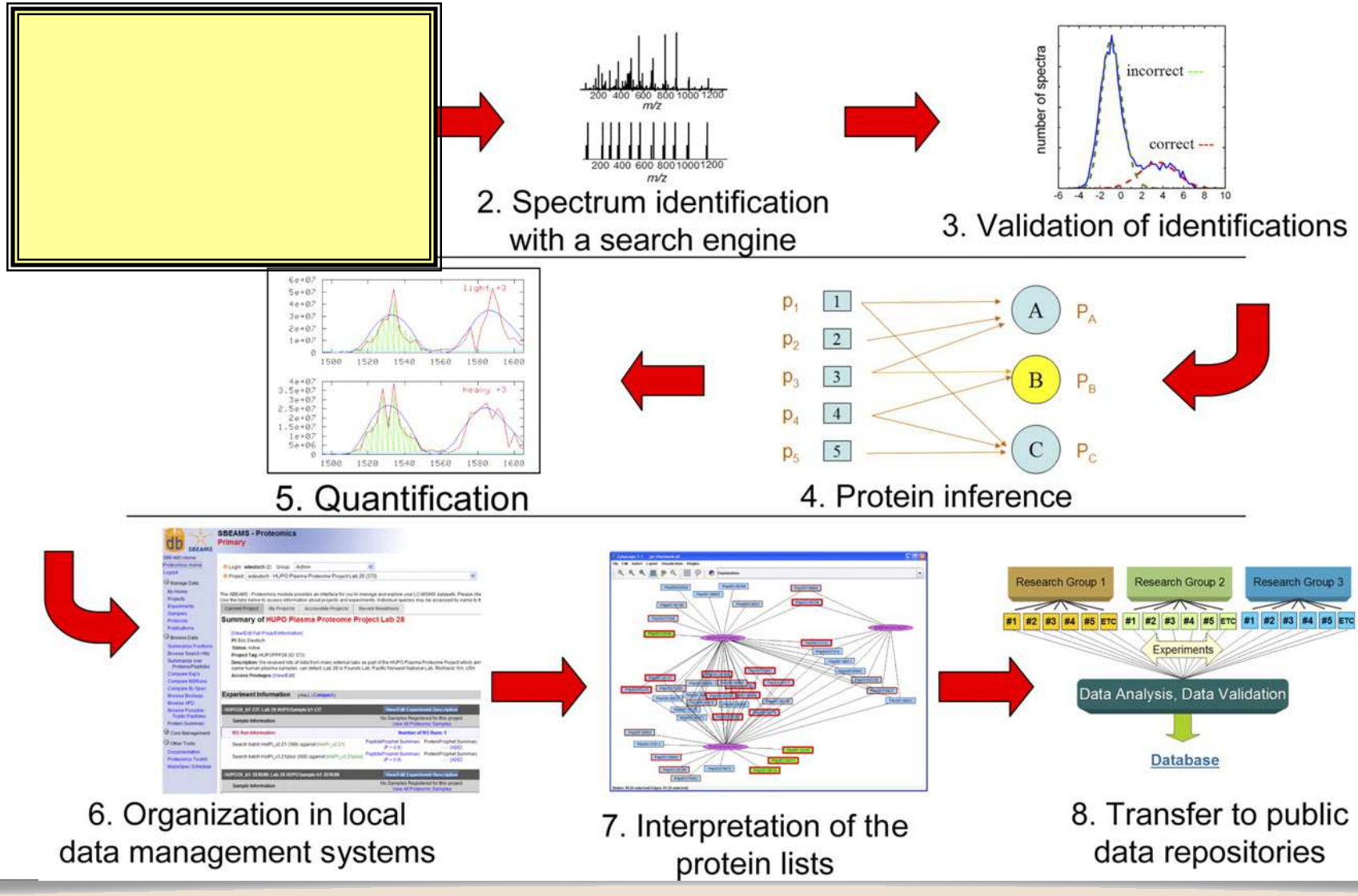
7. Interpretation of the protein lists



8. Transfer to public data repositories

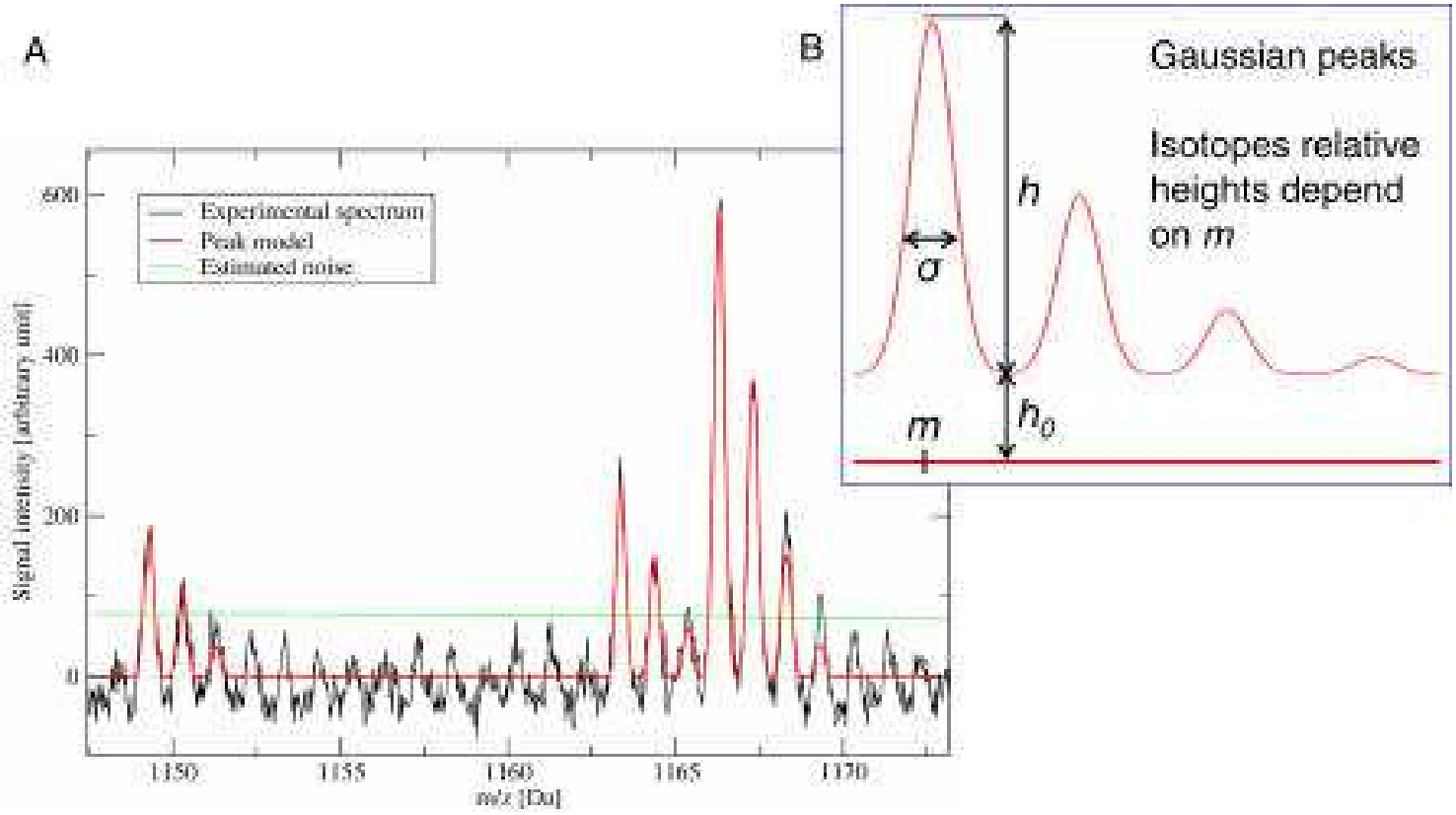
Deutsch E W et al. Physiol. Genomics 2008;33:18-25

Schematic overview of a typical workflow of the proteomics informatics processing of a data set



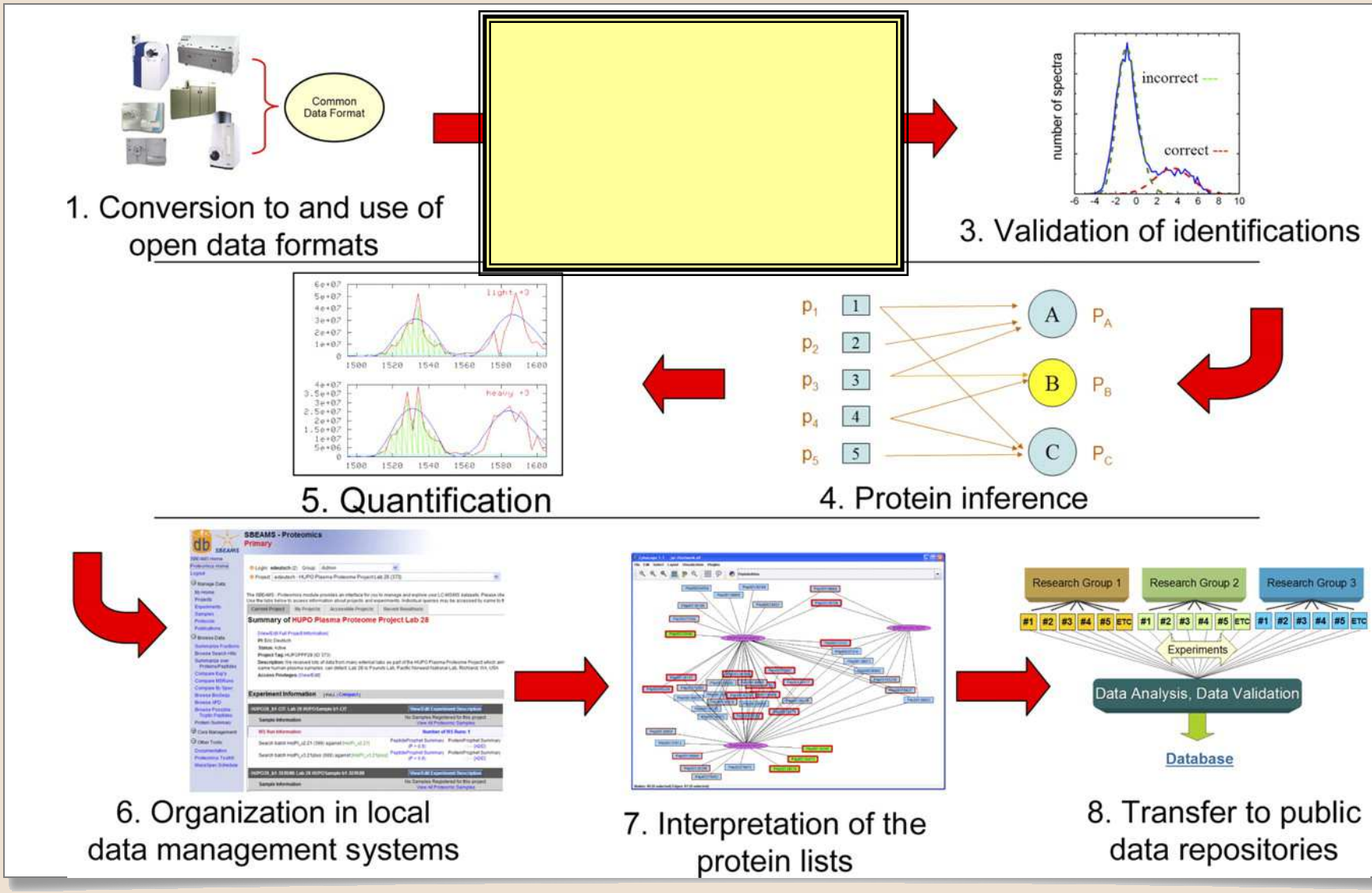
Deutsch E W et al. Physiol. Genomics 2008;33:18-25

Peak detection



From Colinge & Bennett, 2007

Schematic overview of a typical workflow of the proteomics informatics processing of a data set



Deutsch E W et al. Physiol. Genomics 2008;33:18-25

The paradigms for protein identification using MS/MS data

- **Database searching (protein sequences)**
- **Interpretation of MS/MS spectra**
  - De novo sequencing
  - Peptide Sequence Tags (PSTs)
- **Using a reference database (analytical data)**
  - AMT database
  - Spectra libraries

The paradigms for protein identification using MS/MS data

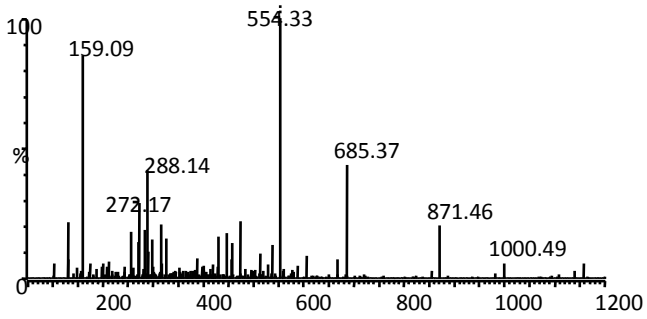
- **Database searching (protein sequences)**
- **Interpretation of MS/MS spectra**
  - De novo sequencing
  - Peptide Sequence Tags (PSTs)
- **Using a reference database (analytical data)**
  - AMT database
  - Spectra libraries

Searching uninterpreted MS/MS spectra

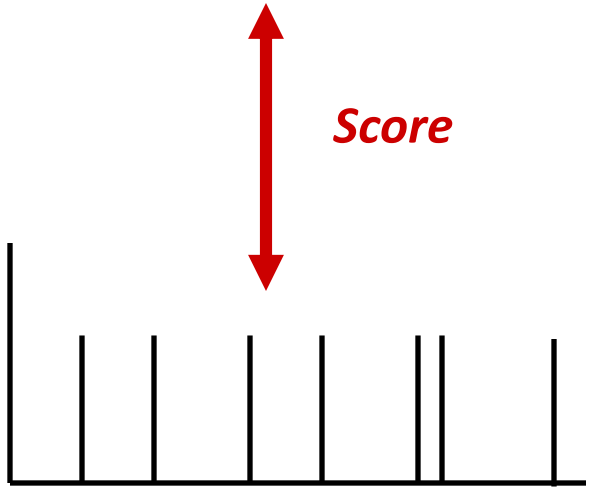
Experimental data :

1- M = mass of the peptide

2- MS/MS spectrum



Mascot, Sequest, X! Tandem, OMSSA...



Virtual data:

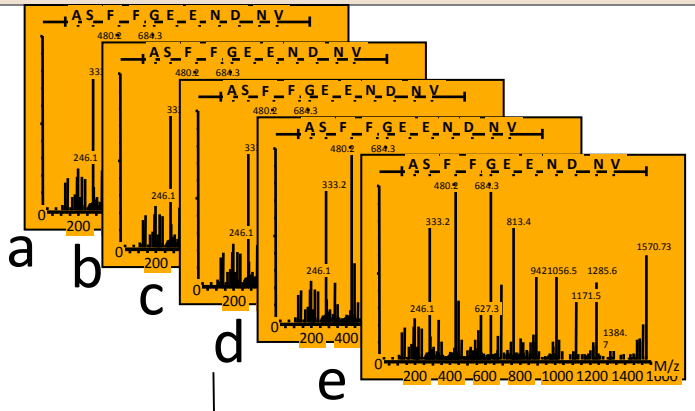
1- trypsin digestion n M

2- n MS/MS spectra

Protein databanks

Yates JR 3rd, Eng JK, McCormack AL, Schieltz D., **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database.** Anal Chem. 1995 Apr 15;67(8):1426-36.

Database searching for shotgun analyses



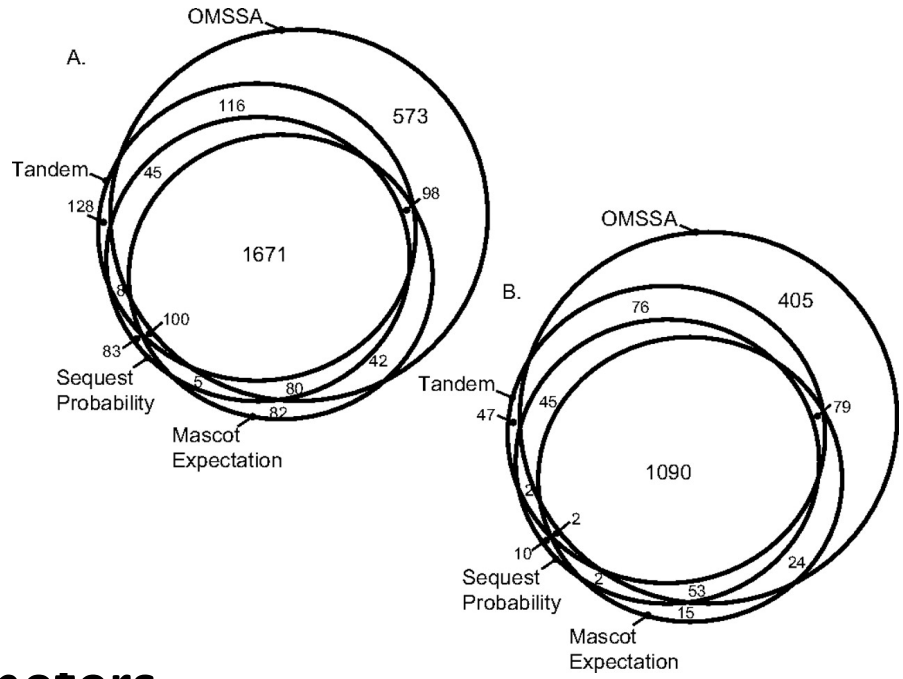
Database searching

(uninterpreted MS/MS spectra)

Clustering

Search engines for uninterpreted MS/MS spectra

- **Open source tools**
  - X! Tandem
  - OMSSA
  - Comet …
- **Commercial softwares**
  - Mascot
  - Phenyx …
- **Available with Mass spectrometers**
  - Sequest
  - ProteinLynx Global Server
  - ProteinPilot …



→ A. Cornuéjols, Axe 1, 30/11

Balgley B M et al. Mol Cell Proteomics 2007;6:1599-1608

Search constraints

Chloroplast proteins

A. thaliana

Reduction/alkylation

Trypsin digestion

Peptides

LC-MS/MS

Orbitrap

Raw database searching results

Choices to be done

Protein databank (species)
(contaminants)

Chemical modifications
(e.g. alkylated methionine

Enzyme specificity
Missed cleavage ?

Mass accuracy
For Orbitrap instruments
Peptide: 1-10 ppm
Fragments: 0.6 Da

Constraints in database searching

MASCOT MS/MS Ions Search

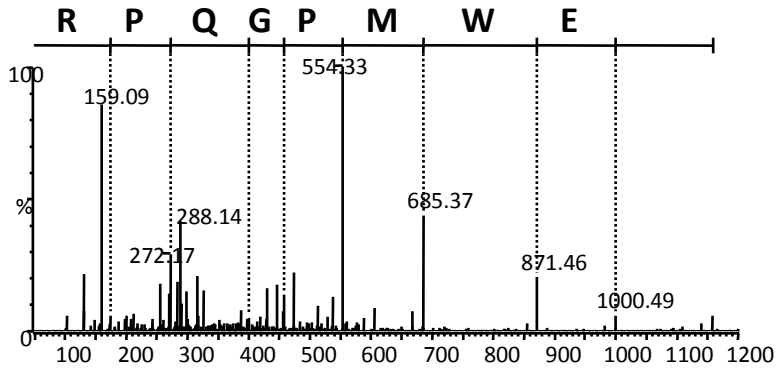
Your name	Myriam	**Email**	myriam.ferro@cea.fr
Search title			
Database	ATH_Cplet_AMT		
Taxonomy	All entries		
Enzyme	Trypsin	**Allow up to**	1 missed cleavages
Fixed modifications	Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term)	**Variable modifications**	Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term)
Quantitation	None		
Peptide tol. ±	20 ppm # ^{13}C 0	**MS/MS tol. ±**	1.2 Da
Peptide charge	2+	**Monoisotopic**	**Average**
Data file	D:\pkl_files\EUJU780b.mgf Browse...		
Data format	Mascot generic	**Precursor**	m/z
Instrument	ESI-QUAD-TOF	**Error tolerant**	
Decoy		**Report top**	AUTO hits
	Start Search ...		Reset Form

http://www.matrixscience.com/

The paradigms for protein identification using MS/MS data

- Database searching (protein sequences)
- **Interpretation of MS/MS spectra**
  - De novo sequencing
  - Peptide Sequence Tags (PSTs)
- Using a reference database (analytical data)
  - AMT database
  - Spectra libraries

De novo sequencing



PepNovo; PEAKS etc.

MS/MS interpretation

EWMPGQPR → whole or partial sequence

MSBlast, BLASTP, TBLASTN ...

Alignment with protein, EST and genomic databases

EWMPGQPR
....VGRRHGSRVSKSAEWMPGQPRPPHLDGSAPGD....

Searching for similar proteins

EWMPGQPR
....VGATNSMSRFSMSADWMPGQPRPSYLDGSAPGD....

Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. **De novo peptide sequencing via tandem mass spectrometry. J Comput Biol.** 1999 Fall-Winter;6(3-4):327-42.

Shevchenko A, Chernushevic I, Shevchenko A, Wilm M, Mann M.**"De novo" sequencing of peptides recovered from in-gel digested proteins by nanoelectrospray tandem mass spectrometry.** Mol Biotechnol. 2002 Jan;20(1):107-18.

The paradigms for protein identification

- **Interpretation of MS/MS spectra**
  - De novo sequencing
  - Peptide Sequence Tags (PSTs)
- **Database searching (protein sequences)**
- **Spectral libraries**
  - Spectra libraries
    - → Prerequisite: MS/MS spectra already “annotated”

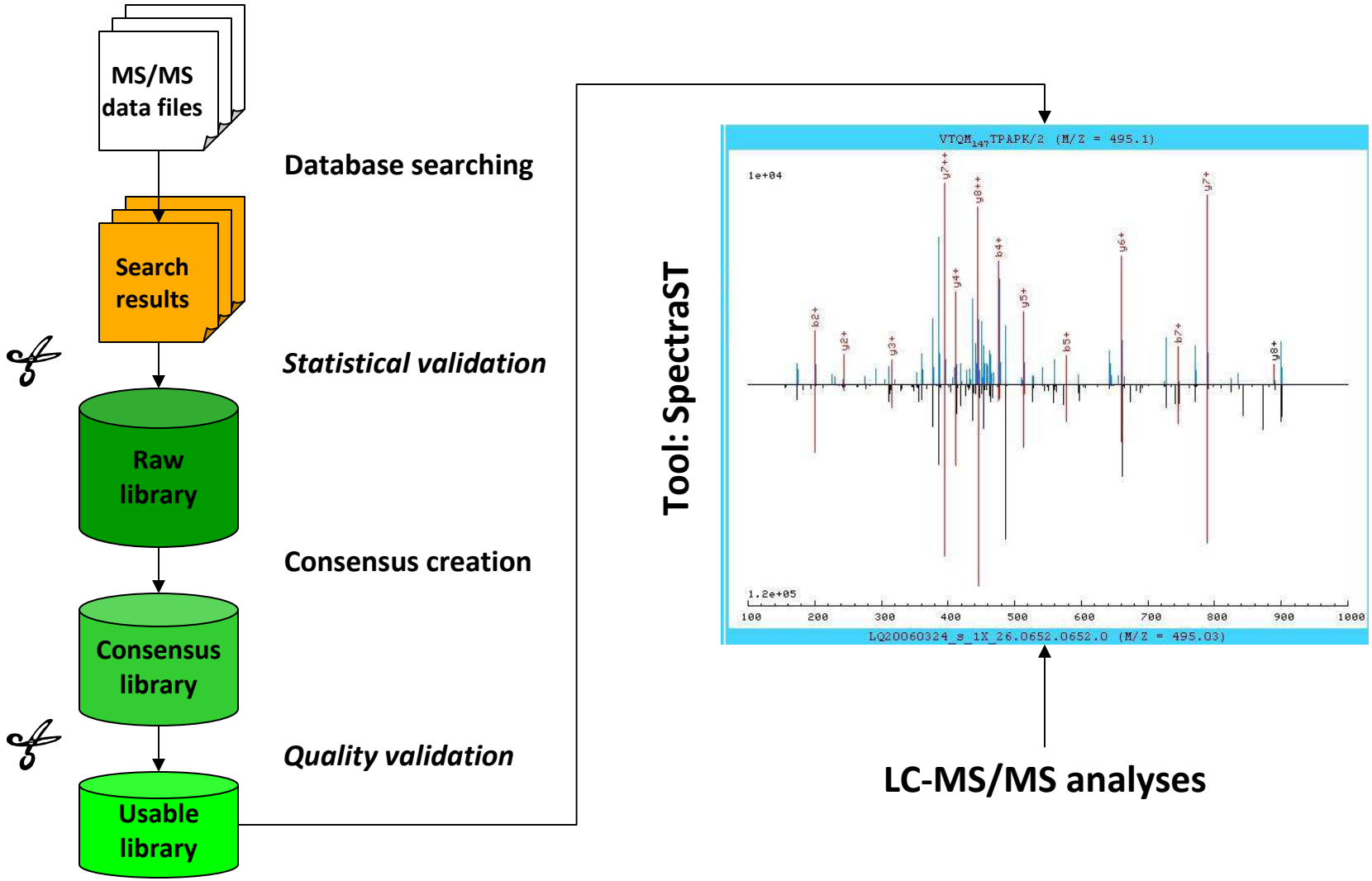
Spectral libraries

- **Concept**
  - Capitalize on validated experimental data
    - Use peptide fragmentation « pattern »

- **Principle**
  - Create a « consensus » spectrum for all peptides identified from experimental spectra
  - Compare this « consensus » spectrum with new experimental spectrum

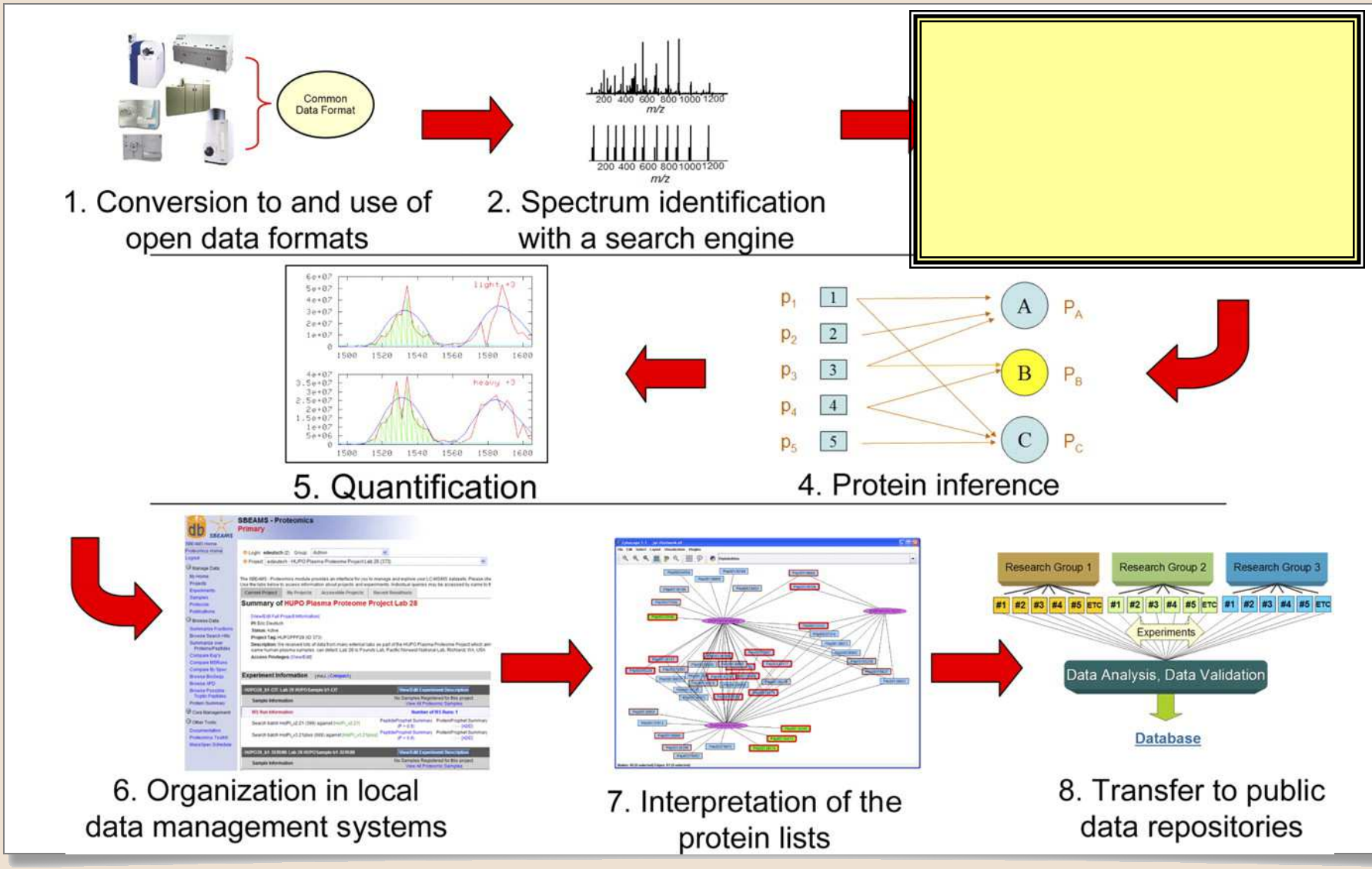
Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold R. Building consensus spectral libraries for peptide identification in proteomics. Nat Methods. 2008 Oct;5(10):873-5

Pipeline of a spectral library search



Lam H. Building and searching tandem mass spectral libraries for peptide identification. Mol Cell Proteomics. 2011 Sep 6. [Epub ahead of print]

Schematic overview of a typical workflow of the proteomics informatics processing of a data set

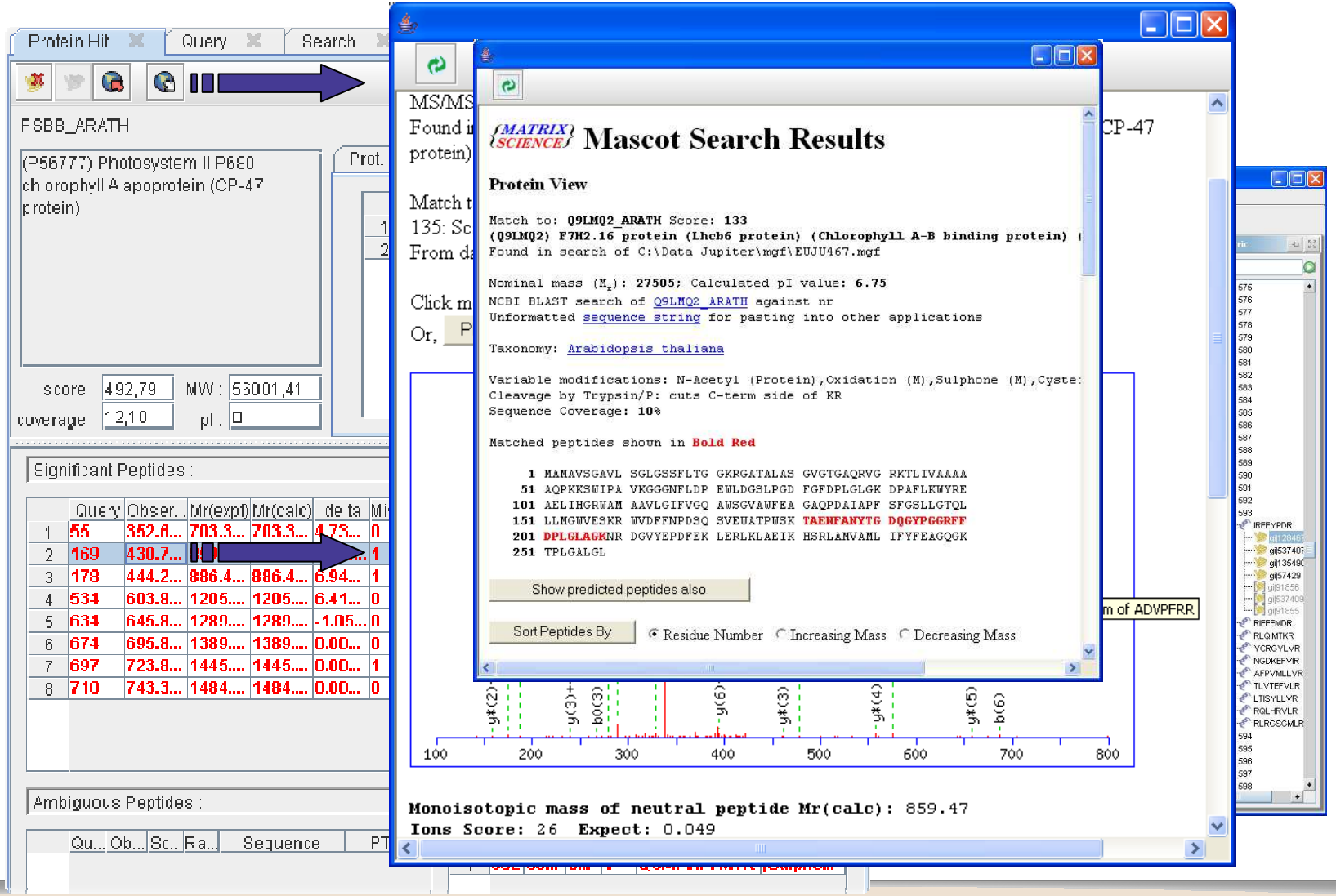


Deutsch E W et al. Physiol. Genomics 2008;33:18-25

Validation of database searching results

- **Tools**
  - Many!
- **Needs**
  - Results filtering
    - Using predefined rules
  - Manual filtering
    - Validate spectrum-peptide match
      - View « matched » spectrum on peptide sequence
    - Access protein description (NCBI, SwissProt, etc)
  - Refine a result by re-submitting spectra
  - Save on-going validation
  - Generate reports
    - Filtered … but consistent

Validation of the results: an example using IRMa



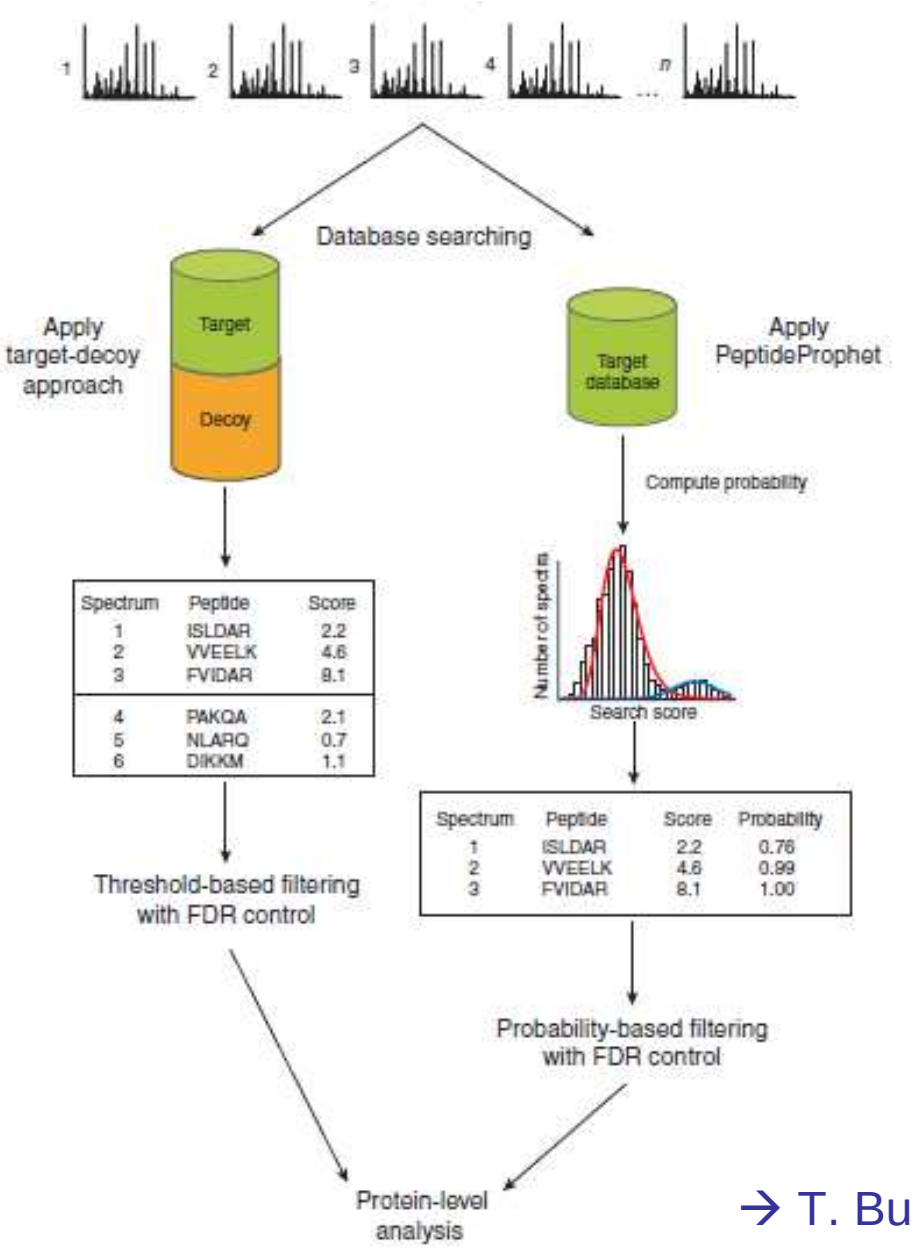
Dupierris V, Masselon C, Court M, Kieffer-Jaquinod S, Bruley C. A toolbox for validation of mass spectrometry peptides identification and generation of database: IRMa. Bioinformatics. 2009 Aug 1;25(15):1980-1

False discovery rate (FDR)

- **False discovery rate**
  - Percentage of incorrect peptide-spectra matches in identification result
- **Problem**
  - How to distinguish « true » and « false » PSMs

FDR

$$FP = \frac{2 \times \mathrm{Re}v}{\mathrm{Re}v + Fwd}$$

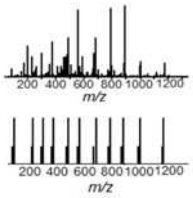


→ T. Burger, Axe1, 29/11

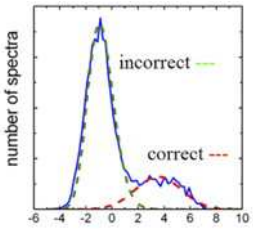
Alexey I Nesvizhskii, Olga Vitek & Ruedi Aebersold,4NATURE METHODS | VOL.4 NO.10 | OCTOBER 2007

Schematic overview of a typical workflow of the proteomics informatics processing of a data set

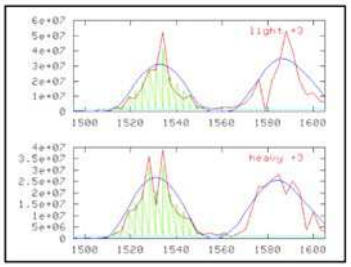
1. Conversion to and use of open data formats



2. Spectrum identification with a search engine

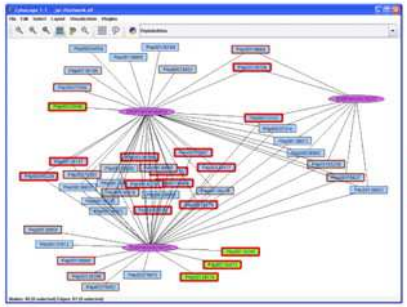


3. Validation of identifications

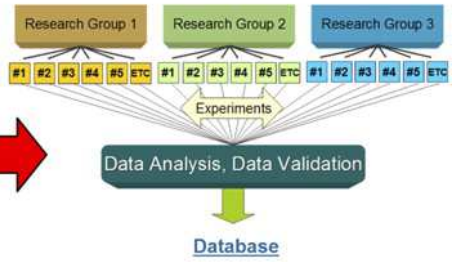


5. Quantification

6. Organization in local data management systems



7. Interpretation of the protein lists

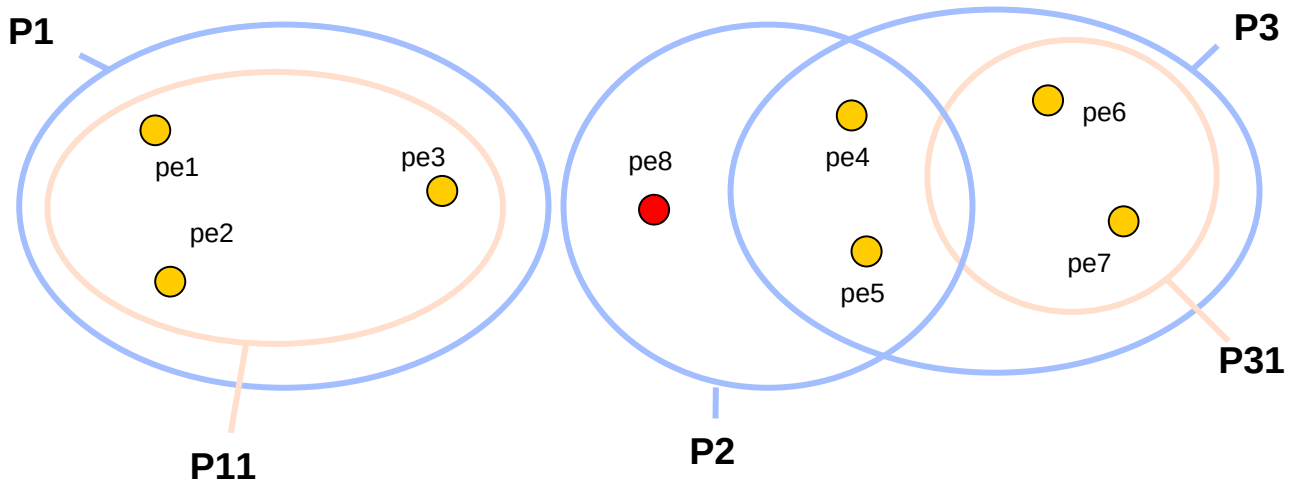


8. Transfer to public data repositories

Deutsch E W et al. Physiol. Genomics 2008;33:18-25

The protein inference problem

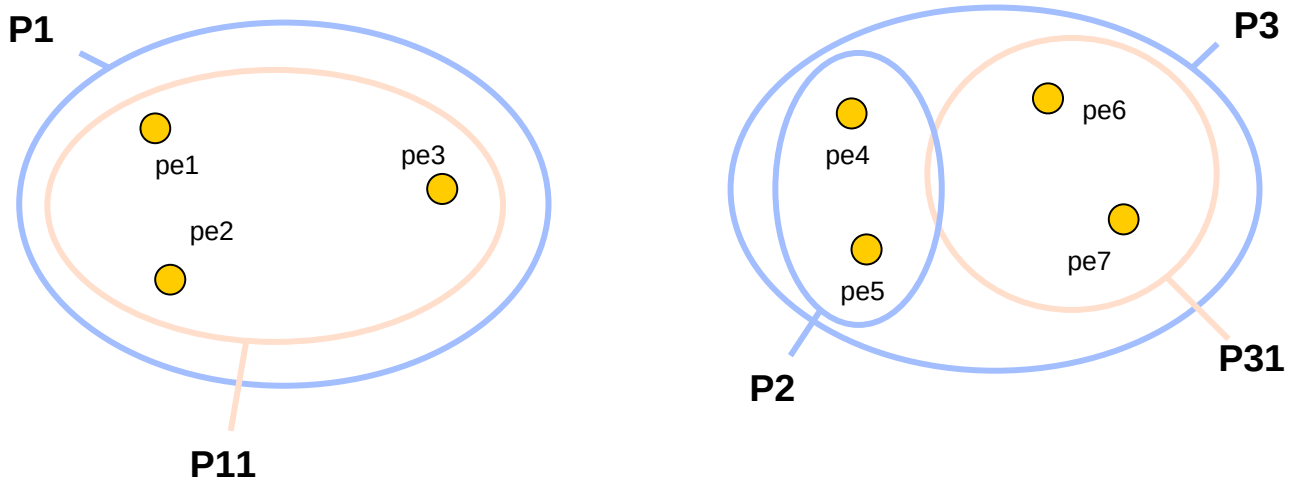
- **The result is a list of protein groups and not proteins**
  - Proteins that share a same set of peptides (P1 and P11)
  - Proteins that share a subset of peptides (P3 and P31)



Nesvizhskii, A.I. and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics 4*, 1419-1440

The protein inference problem

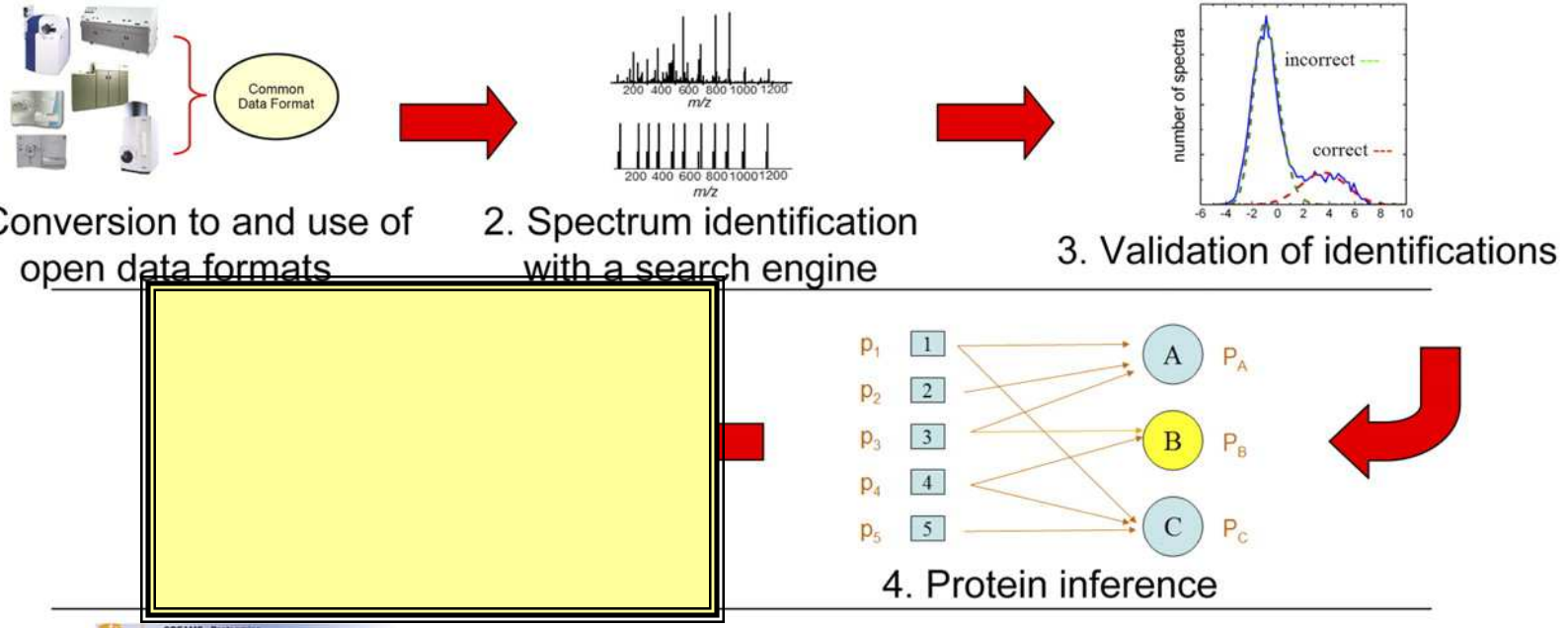
- **List of protein groups and not proteins**
  - Proteins that share a same set of peptides (P1 and P11)
  - Proteins that share a subset of peptides (P3 and P31)
  - After validation the list of protein groups can be modified (consistency to be assured)



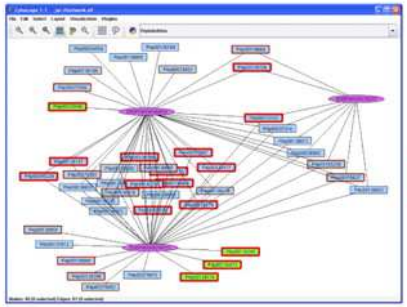
→ B. Valot, S. Bouveret, axe 1, 29/11

Nesvizhskii, A.I. and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics 4*, 1419-1440

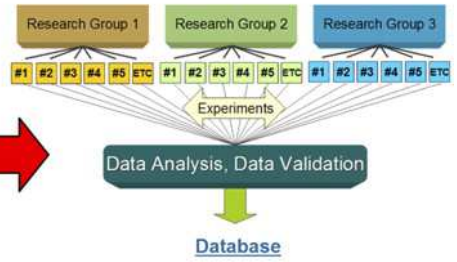
Schematic overview of a typical workflow of the proteomics informatics processing of a data set



6. Organization in local data management systems



7. Interpretation of the protein lists



8. Transfer to public data repositories

Deutsch E W et al. Physiol. Genomics 2008;33:18-25

Quantitative proteomics

-Comparison of peptide/protein levels in 2, 3 … n samples

→ Relative quantification

(up- or down-regulation of a protein in a sample relative to an other, results expressed as a 'fold' increase or decrease)

Determination of an exact amount (concentration) of peptide/protein in a sample

→ Absolute quantification

(e.g. ng or nmoles per gram of tissue, or ng or nmoles/ml of plasma ; use of an internal standard)

Even Absolute quantification is Relative – relative to an internal standard

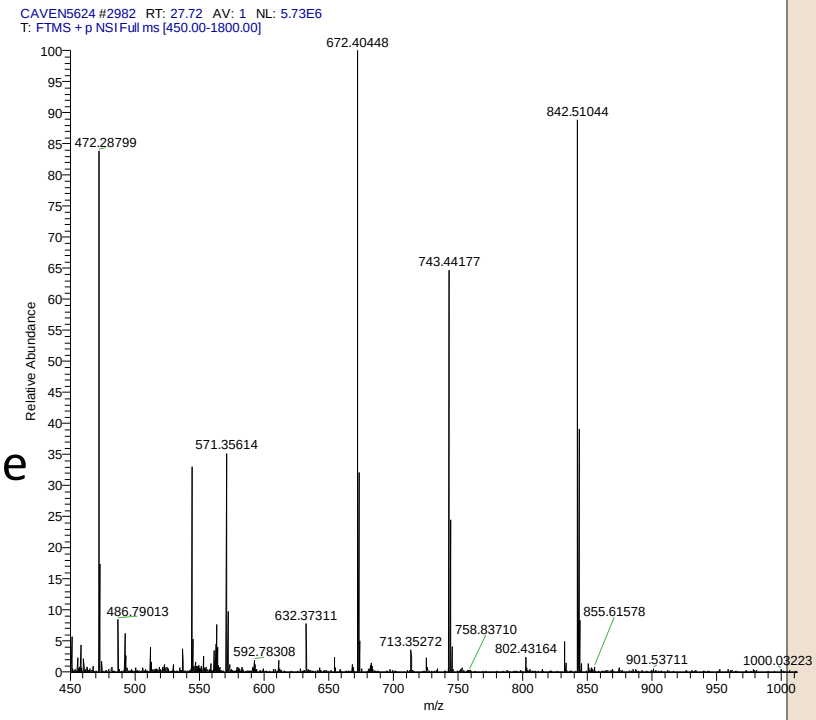
MS-based quantitative proteomics

!!! For several reasons, MS is not a quantitative tool !!!

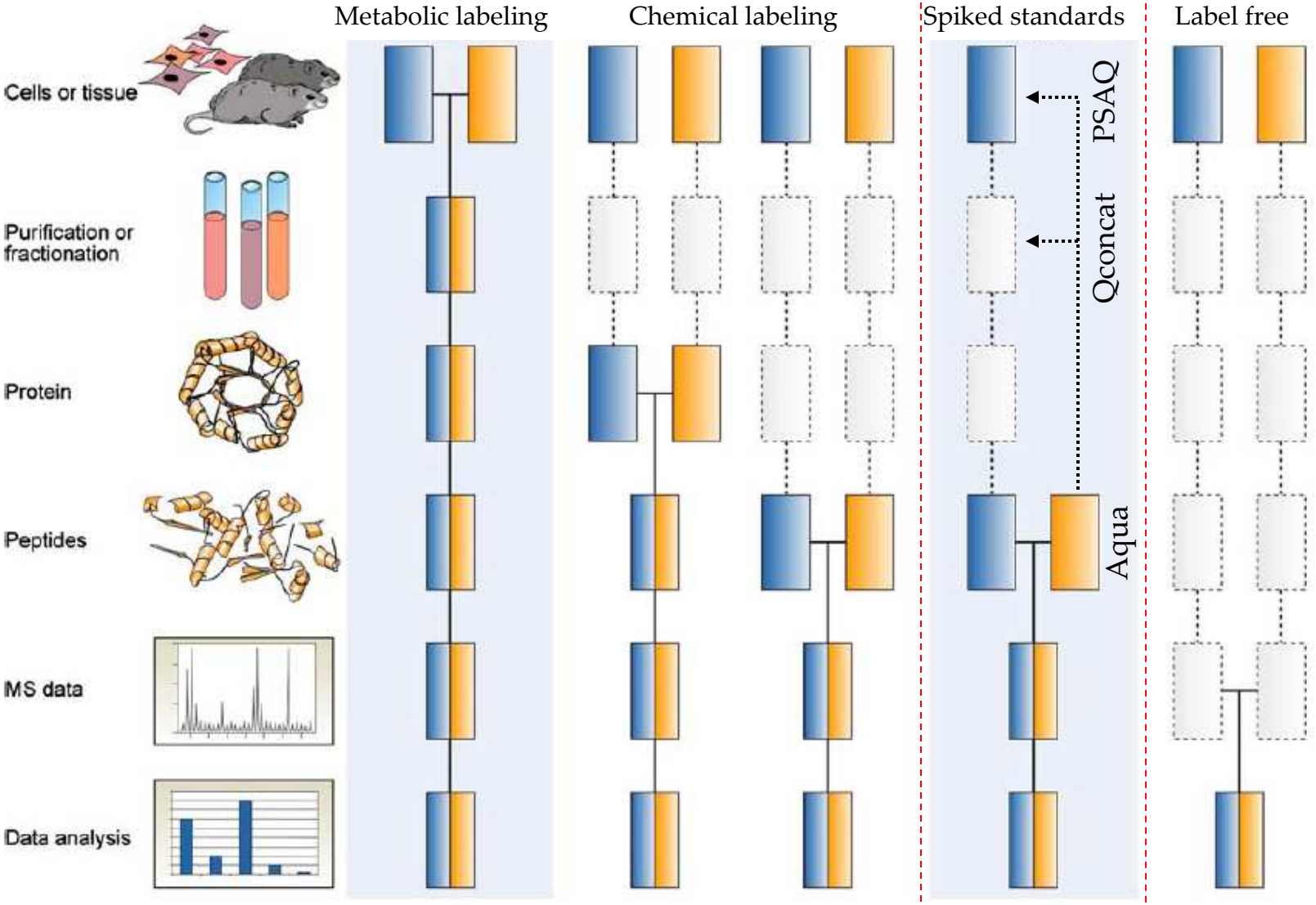
- Ionization is function of the structure
  - Physico-chemical properties of amino acids present in peptide

- Ionization suppression effect
  - Competition between peptides
  - Presence of one peptide may suppress the ionization of another one

- Ionization depends of the sample matrix
  (salt, detergents, contaminants, …)
  - Variation in samples will result in variation in signal intensities

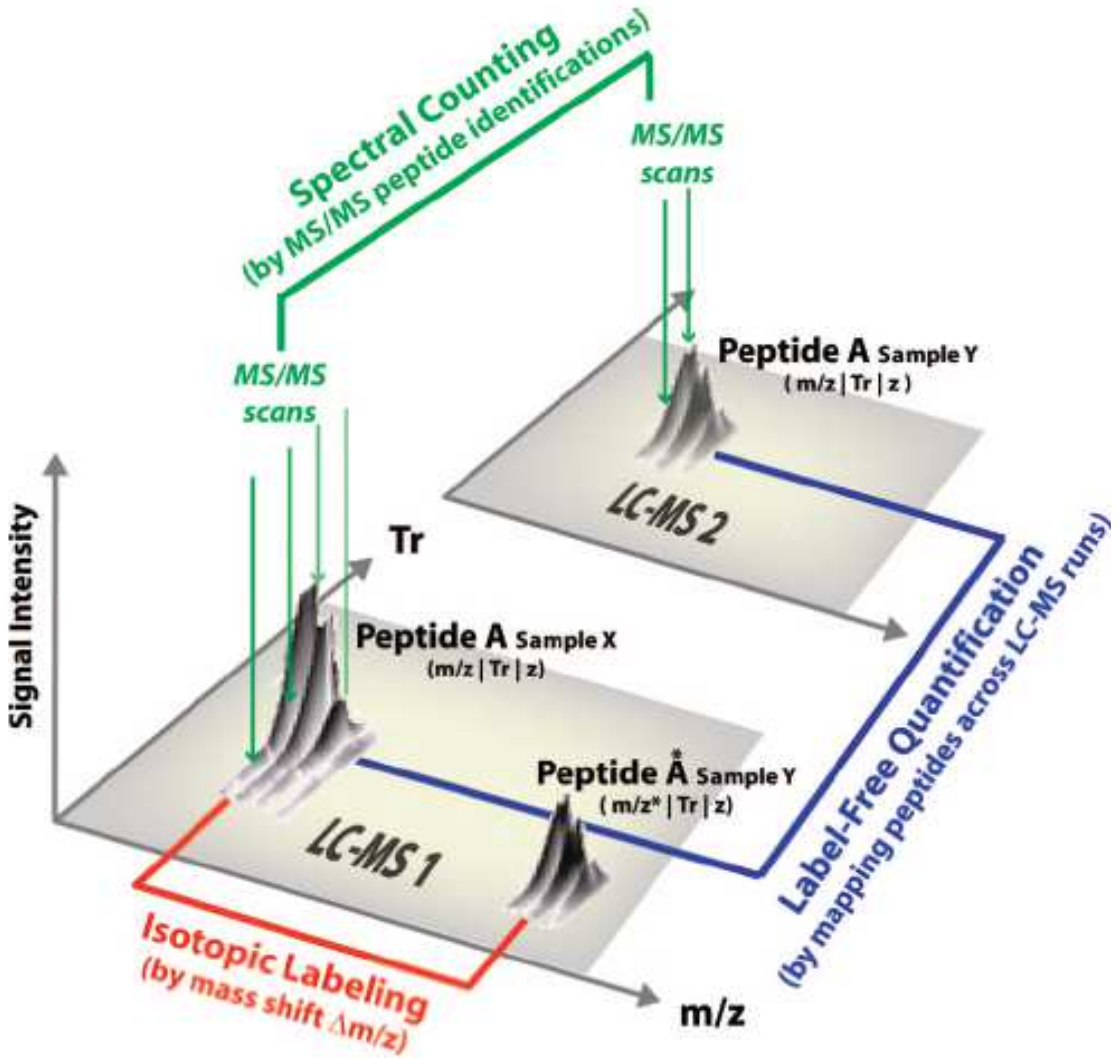


Comparative proteomics workflows



From Bantscheff et al. *Anal Bioanal Chem* (2007) 389:1017–1031

Relative quantification



From Mueller et al., 2008, JPR

Spectral counts

- **Implementations:**
  - Number of matched MS/MS spectra for a given protein
- **Normalization factors**
  - Protein length
  - Total number of spectra
  - etc.

- **Derived metrics used to determine sample relative composition**

$$PAI = \frac{N_{Obsd}}{N_{Obsvbl}} \qquad emPAI = 10^{PAI} - 1$$

$$\text{Protein i molar content} = \frac{emPAI_i}{\sum emPAI}$$

$$APEX_i = \frac{n_i \times p_i}{O_i \times \sum_{k=1}^{N} \frac{n_k \times p_k}{O_k}} \times C$$

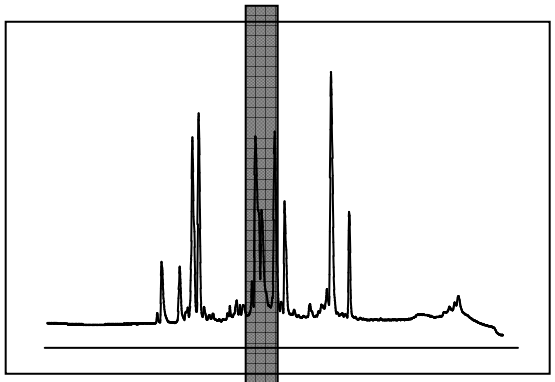
Gao et al. *J Am Soc Mass Spectrom* **2005**, 16: 1231–38

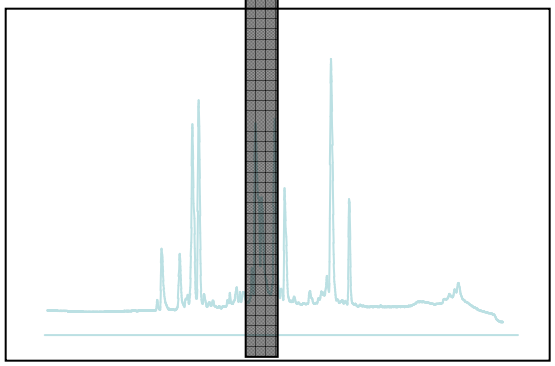
Spectral counts : a semi-quantitative approach

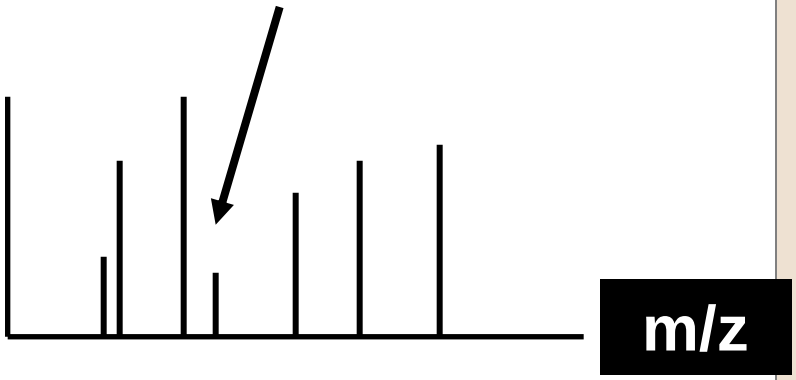
- **Advantages**
  - Easy to implement.
  - Do not require observation of the same peptides between experiments.

- **Drawbacks**
  - Requires many spectra per protein (MudPIT based).
  - Saturation at high SC
  - Complicated in case of shared peptides (isoforms)

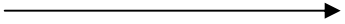
Label-free quantification

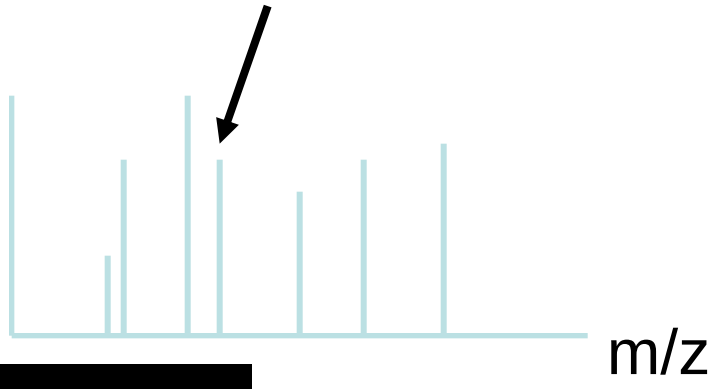






Mass accuracy, calibration, resolution



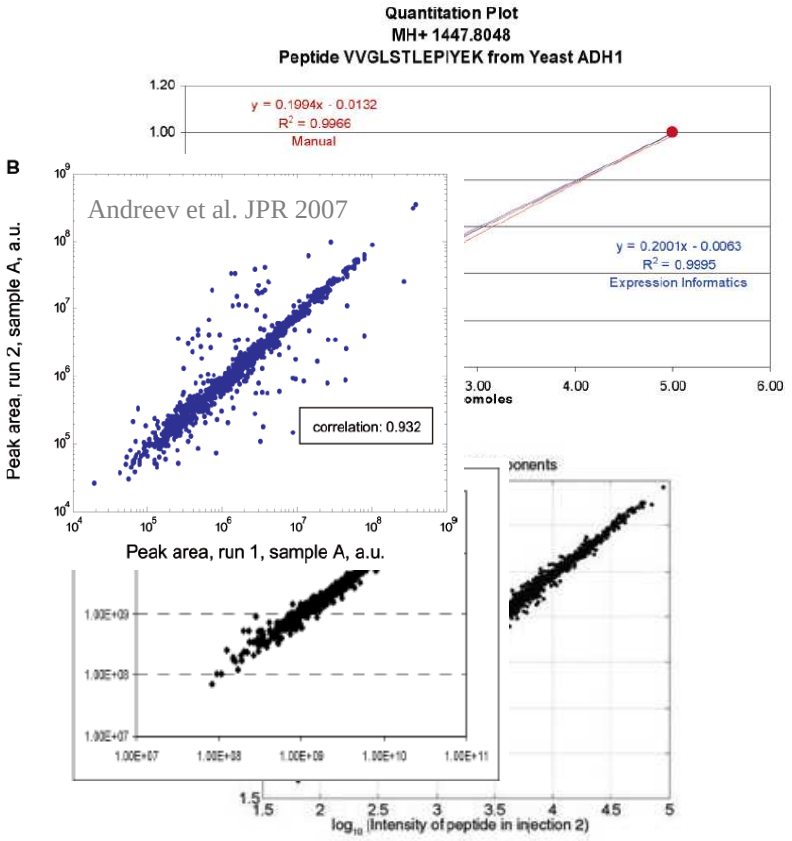


Retention time

LC reproductibility

EXtracted Ion Chromatogram (XIC) alias « MS trace »

- **Peptides abundances scale linearly with concentration**
- **Controlled sample handling and analytical procedures allow repeatable peptide abundance measurements**

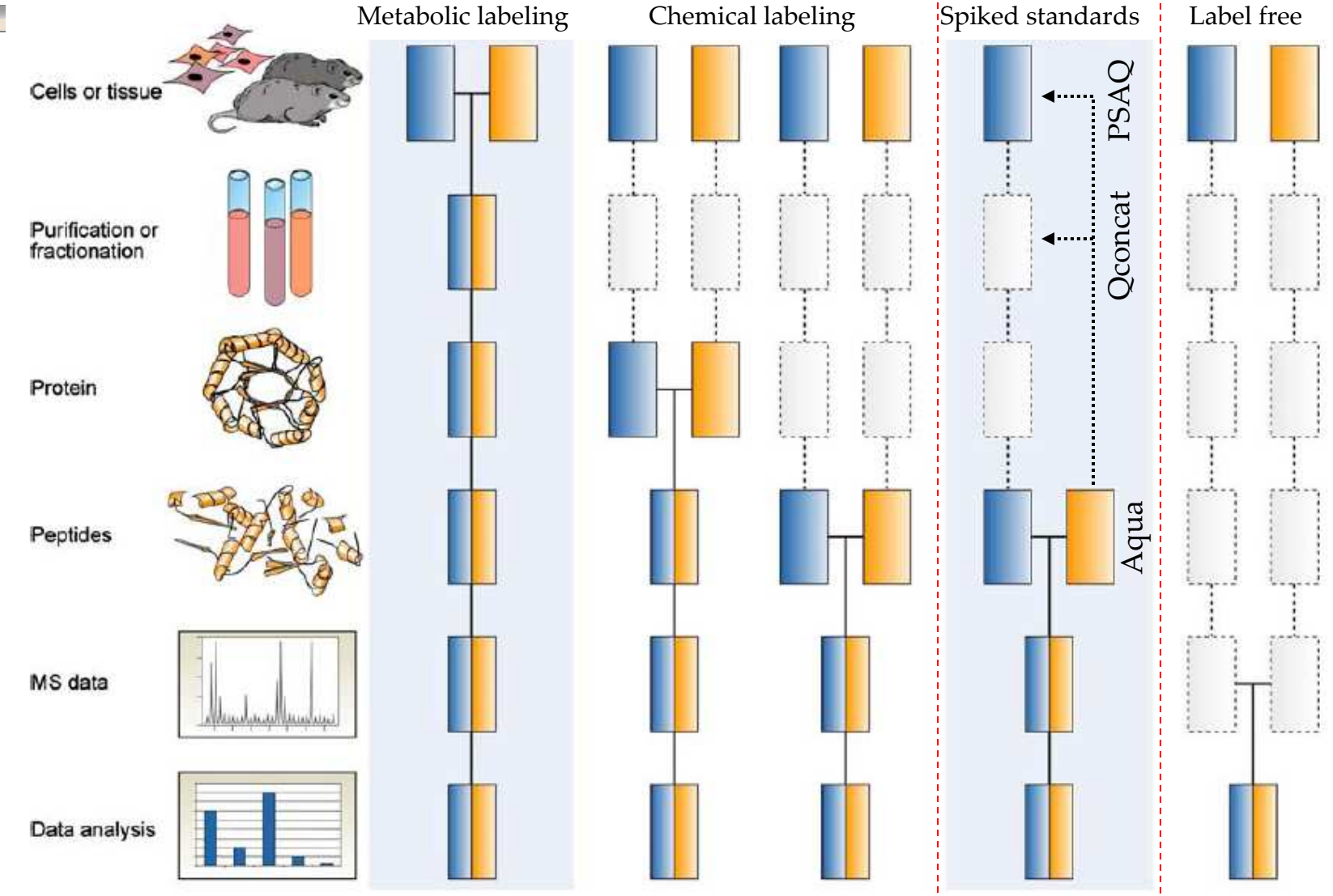


Silva *Anal. Chem.* **2005,** 77: 2187-2200

XIC: challenges

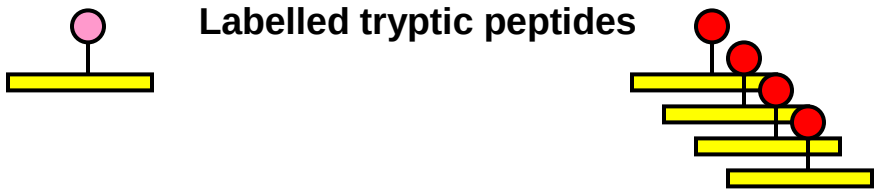
- **Quantitative readings must be extracted from MS or MS/MS spectra → intensities to be extracted**
- Peptide and protein identification must be performed
- The two types of information must be merged and quality controlled
- **Applicable statistical methods have to be identified**
- Individual steps have to be combined in an automated workflow bridging the gaps between commercially available software and custom-built tools.

Labeling approaches

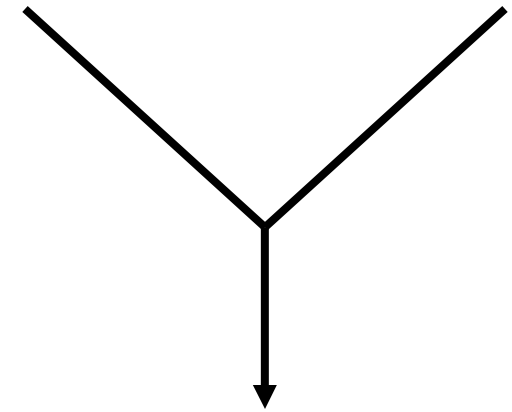


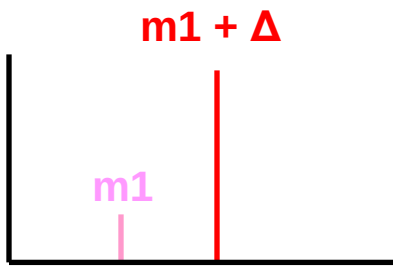
From Bantscheff et al. *Anal Bioanal Chem* (2007) 389:1017–1031

Labeling approaches : general principle



Differential labelling

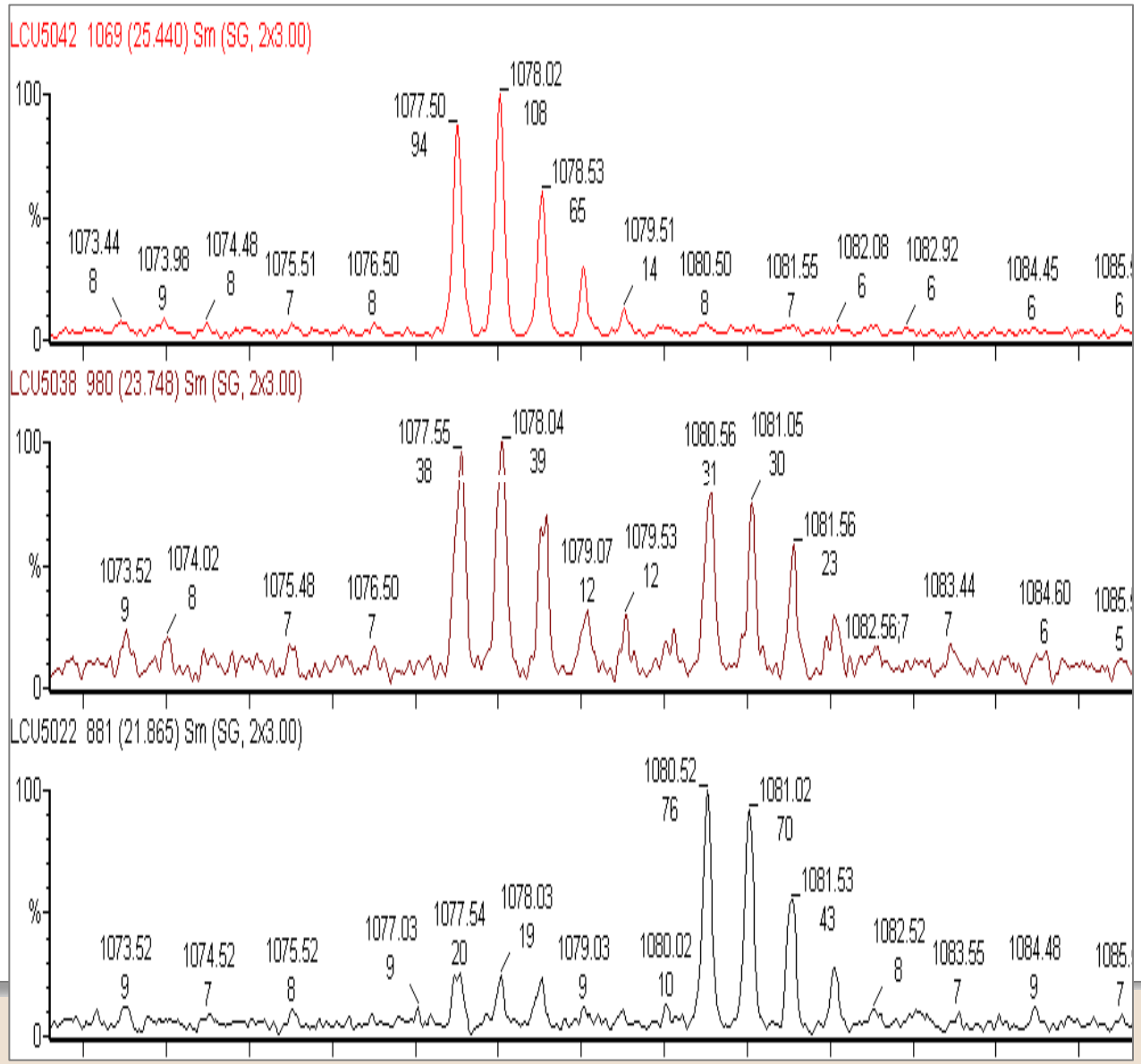




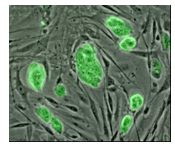
LC-MS/MS

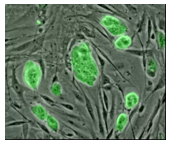
Quantification (pairs detection)

SILAC: example

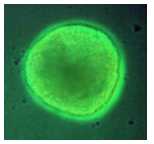


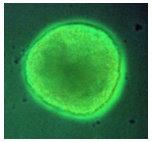
ES (^{12}C-Arg)





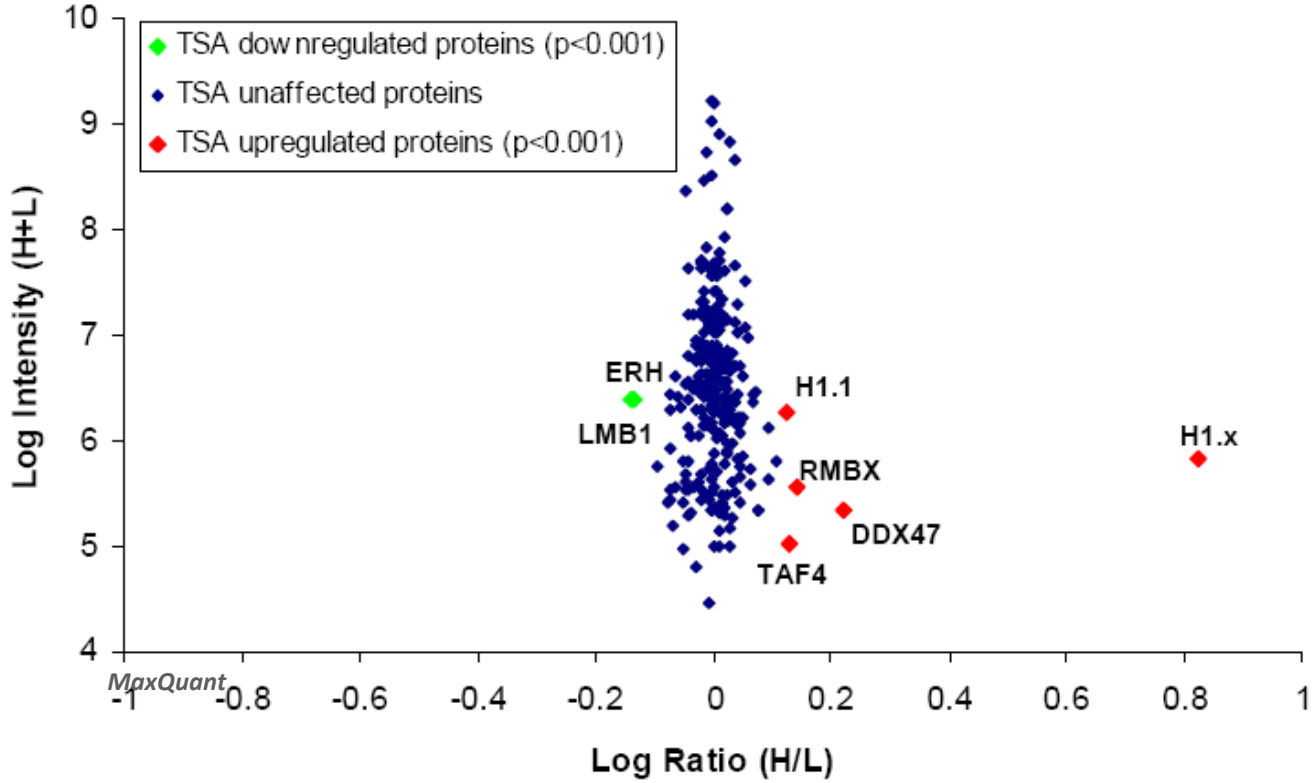
+





ESd (13C-Arg)

SILAC: application exemple



Biological duplicates + Cross-labelling

Identification : 1% False positive – **Quantification** : At least 2 peptides / protein

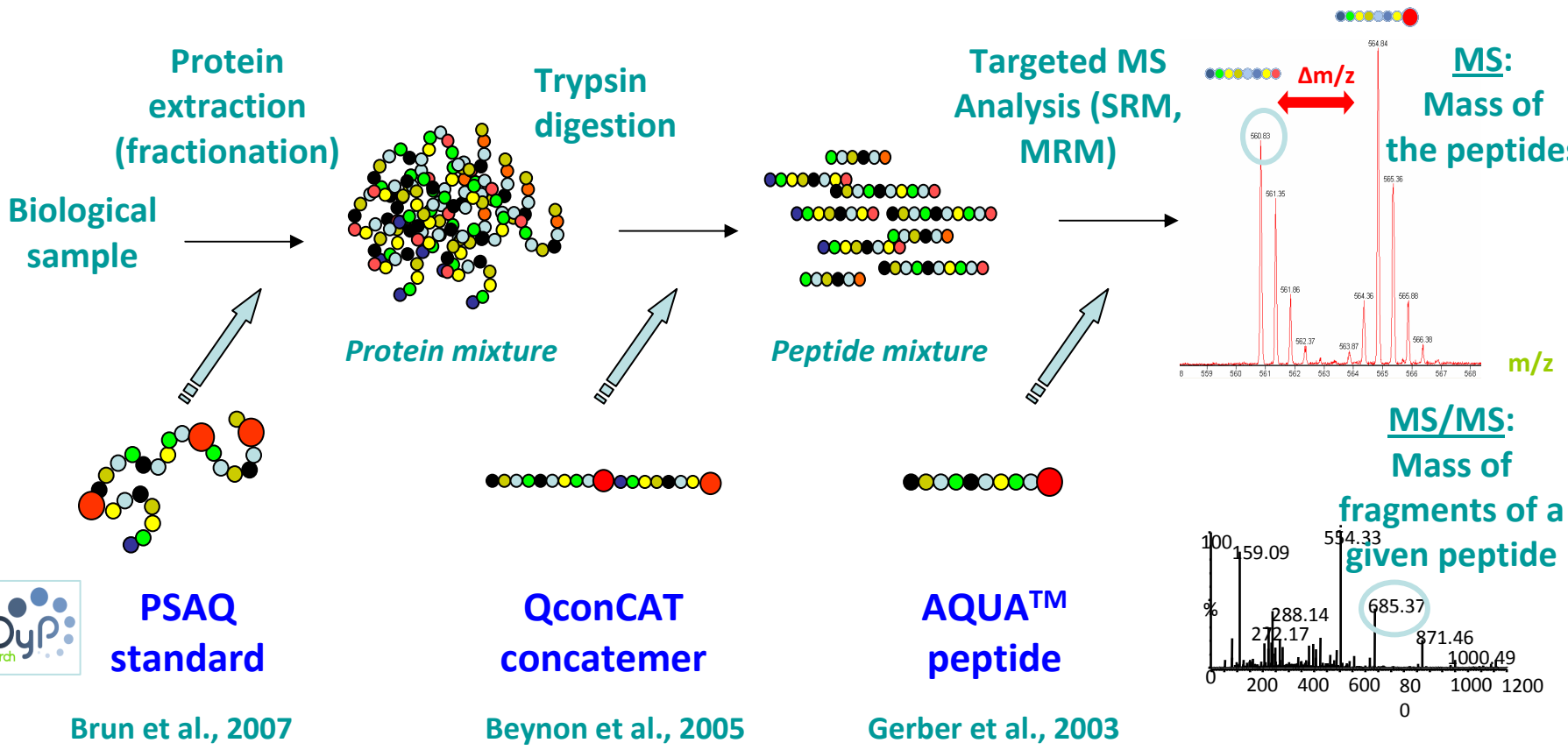
From A. Emadali

Methods with standards

- Towards « absolute » (accurate) quantification : measuring the amount of a given protein in a complex mixture

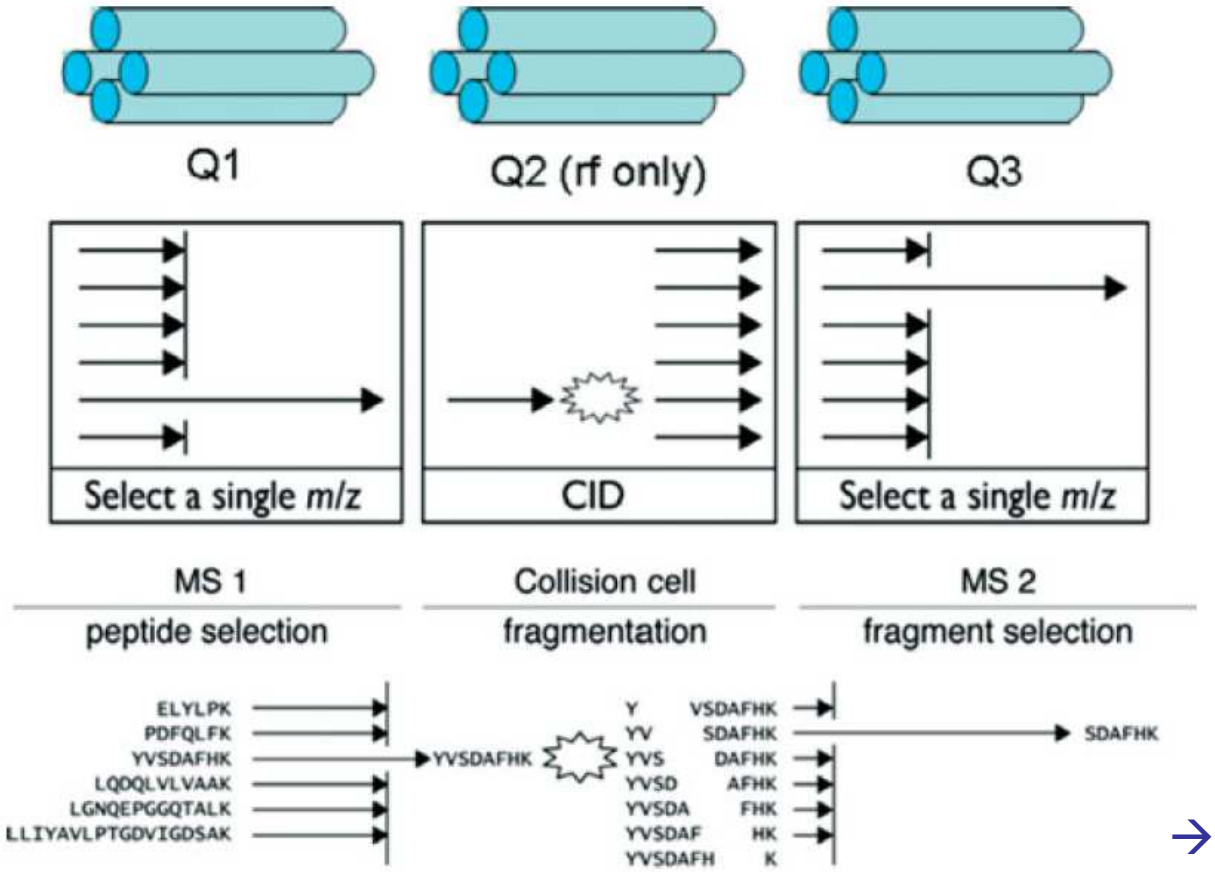
- The proteins to be quantified are already identified (ex: validation de biomarkers)

Absolute quantification of proteins: use of labelled standards (isotope dilution)



SRM

Monitoring of selected parent and specific fragment ions (enhanced accuracy and specificity).



Selection of proteotypic peptides and of good transitions.

Less automated process than discovery proteomics.

→ JF Giovanelli, Axe 1, 30/11

Quantitative proteomics: a question of choice

Label-free, labeling	Orbitrap	T-tests
Absolute, relative	Qtrap	Logistic regression
Limited replicates	Etc.	ML
		Etc.

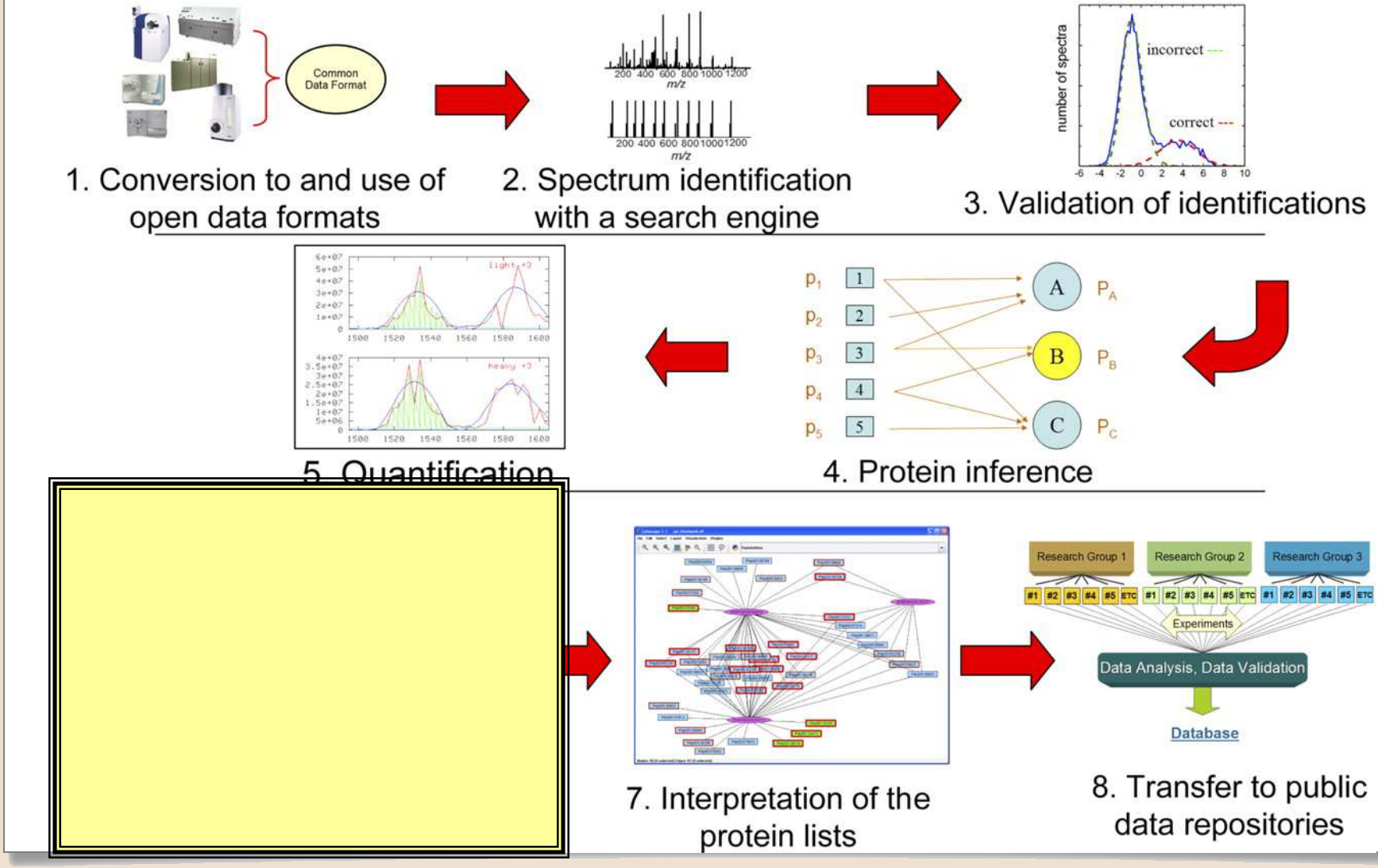
→A. Klich, Axe1, 29/11

→ L. Gerfault, Axe1, 30/11

→Not necessarily the optimal choice but the most pragmatic one

From Käll & Vitek, 2011

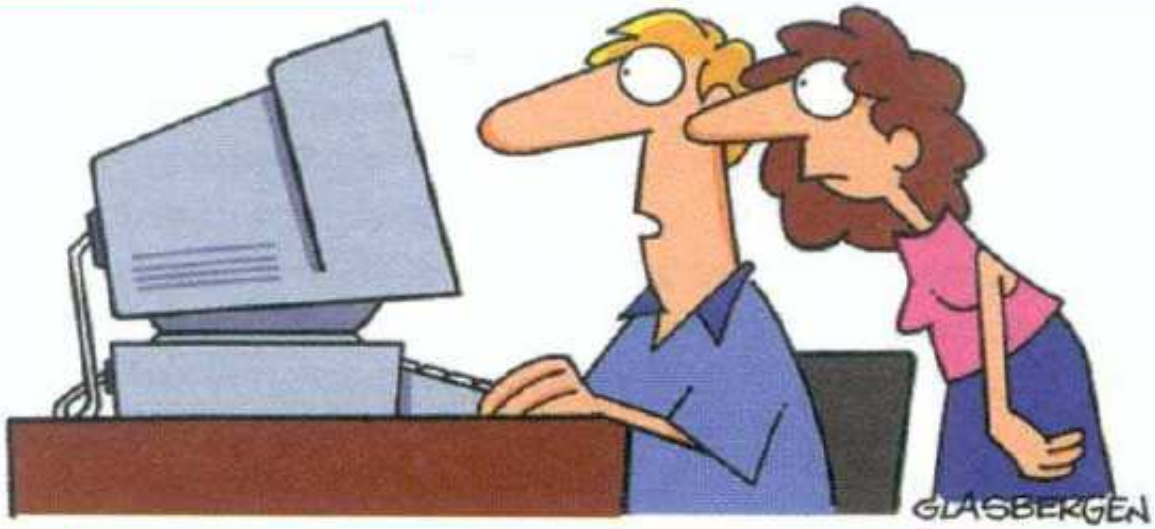
Schematic overview of a typical workflow of the proteomics informatics processing of a data set



Deutsch E W et al. Physiol. Genomics 2008;33:18-25

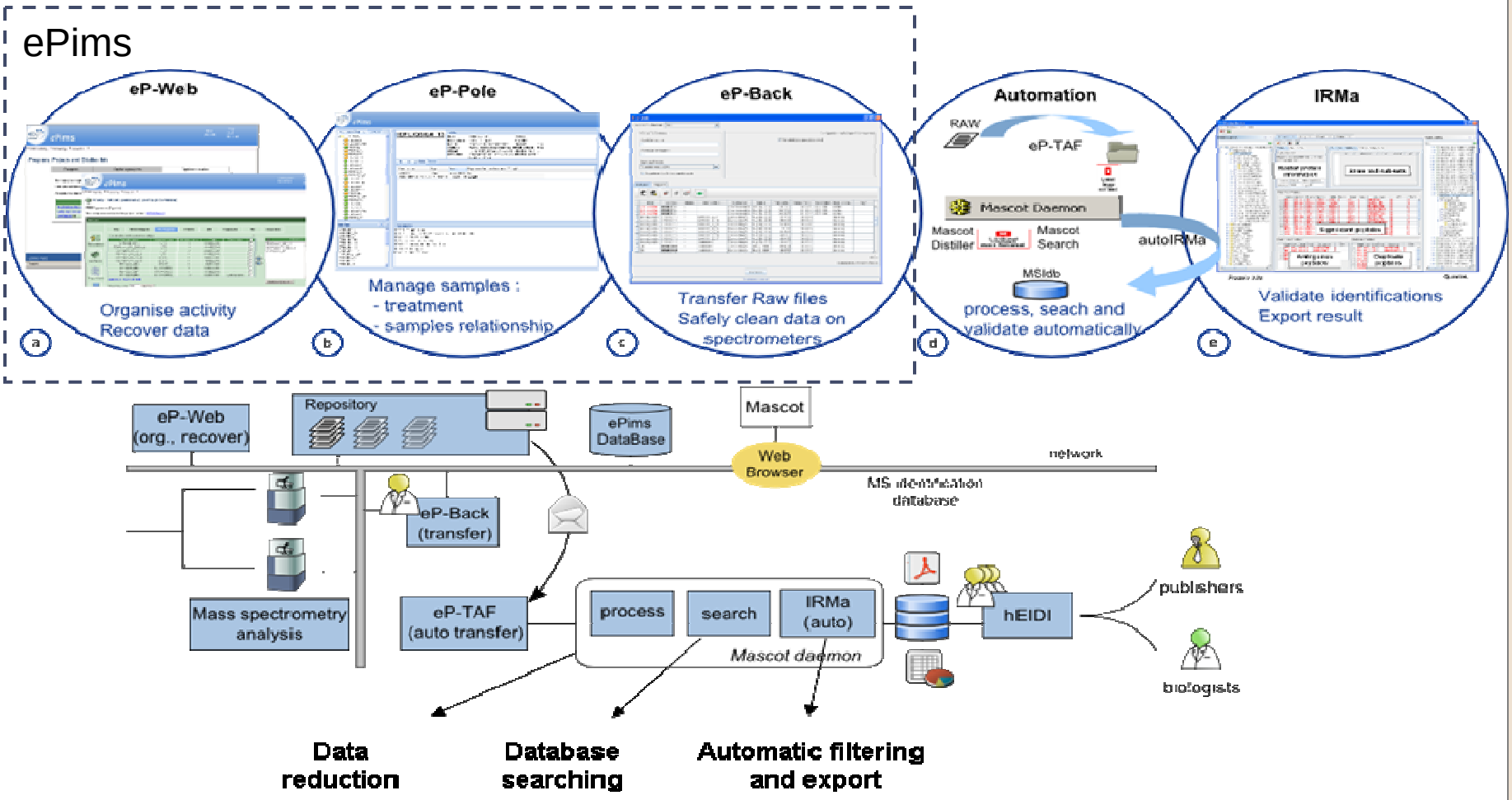
Need for computing assistance !

600 proteins/h/ instrument
Up to 20 000 MS-MS spectra/day/instrument
Up to 50 Go de données/ day



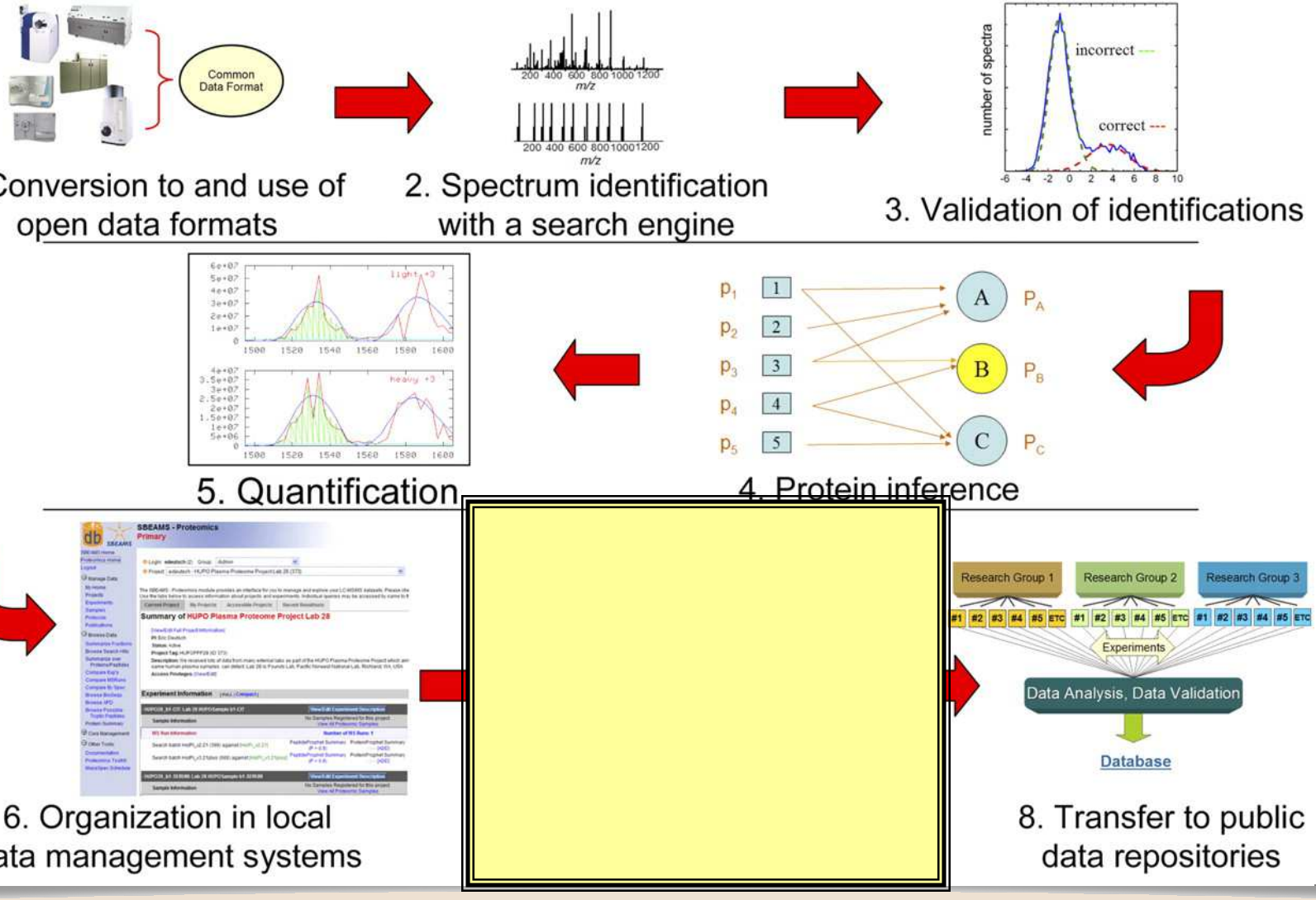
« The computer says I need to upgrade my brain
to be compatible with proteomic data analysis »

An IT workflow



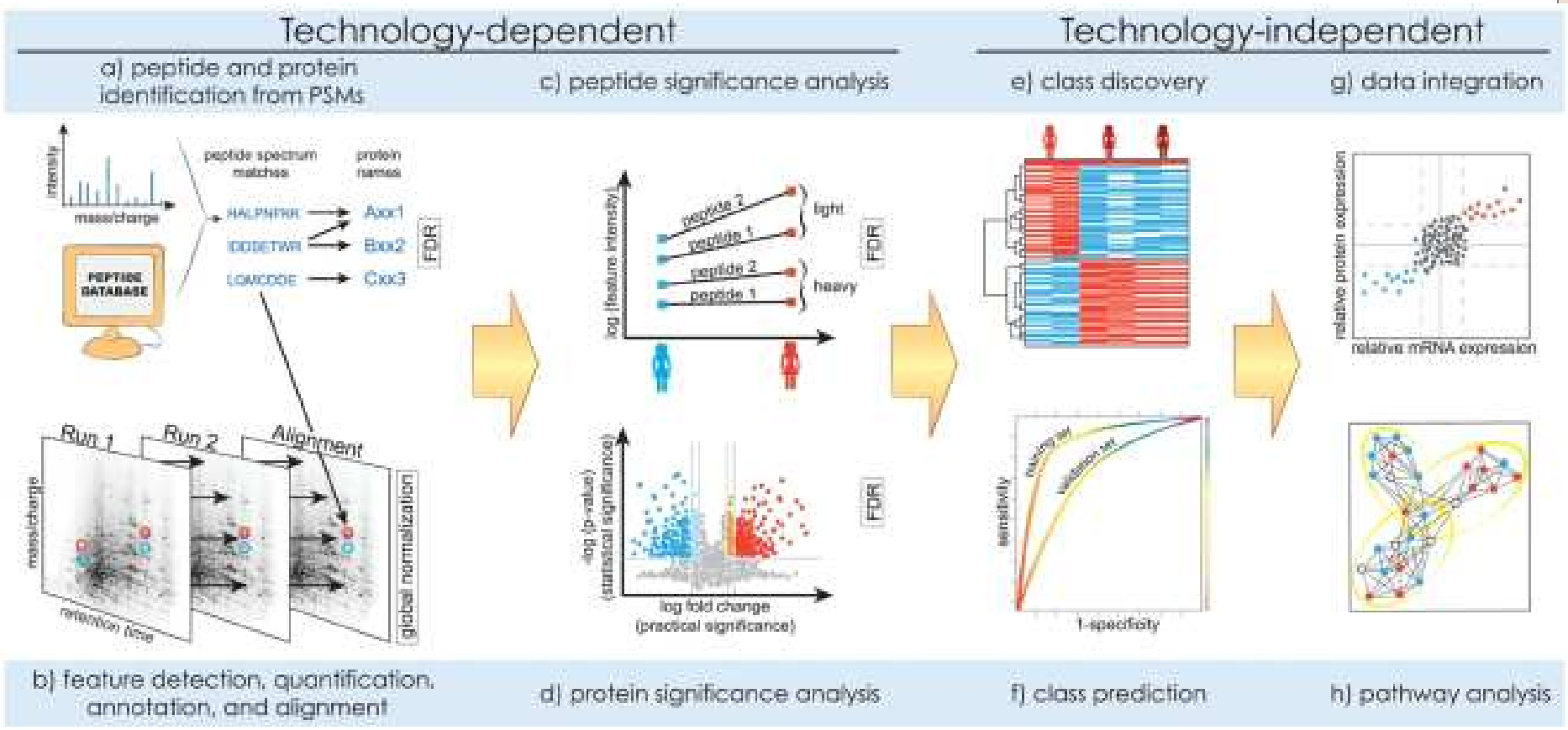
ePIMS™: a dedicated LIMS for proteomics, open source (EDyP)
IRMa: automatic validation of Mascot results. *Dupierris et al. 2009, Bioinformatics*
MSIdb: relational database for the storage of proteomics data (identification, quantification)

Schematic overview of a typical workflow of the proteomics informatics processing of a data set



Deutsch E W et al. Physiol. Genomics 2008;33:18-25

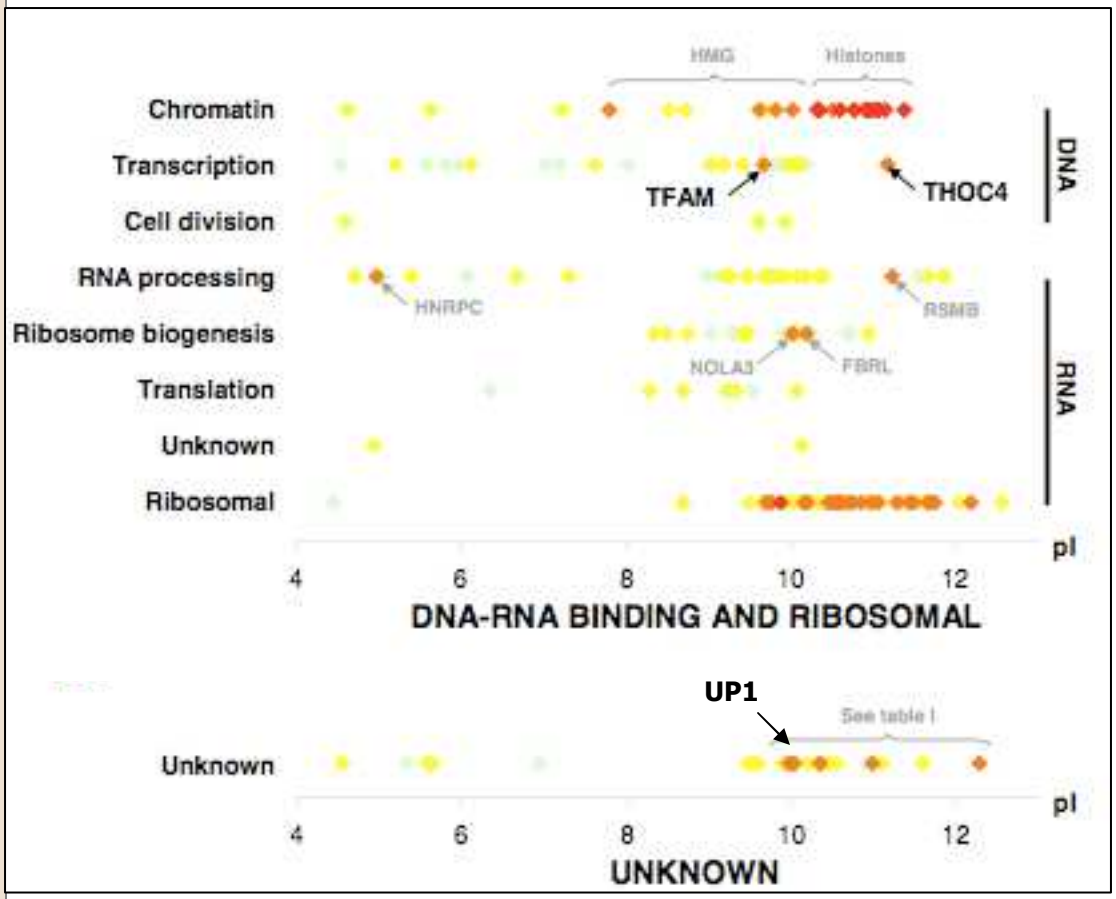
Looking for additional information

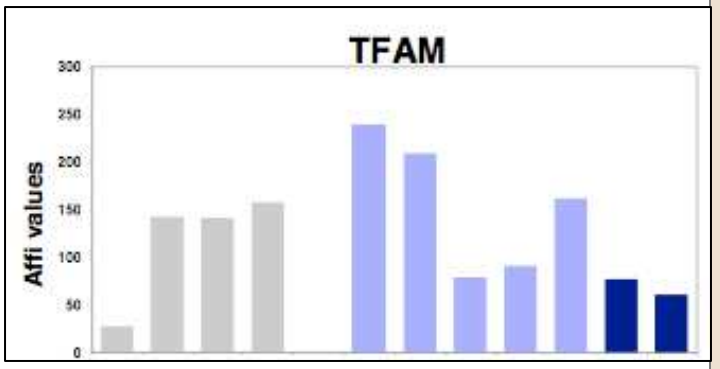


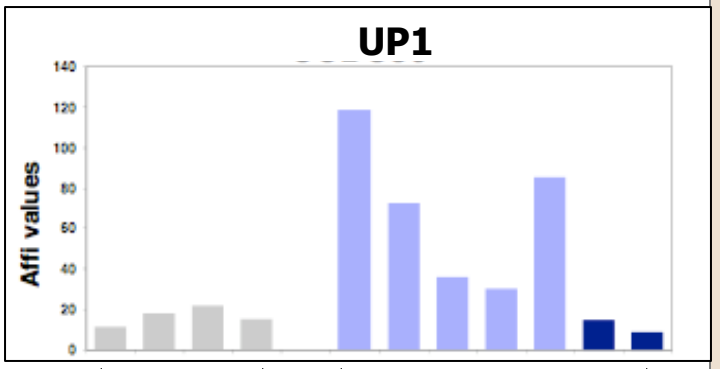
→ L. Gatto, Axe 2, 29/11

From Käll & Vitek, 2011

Integration of OMICS data





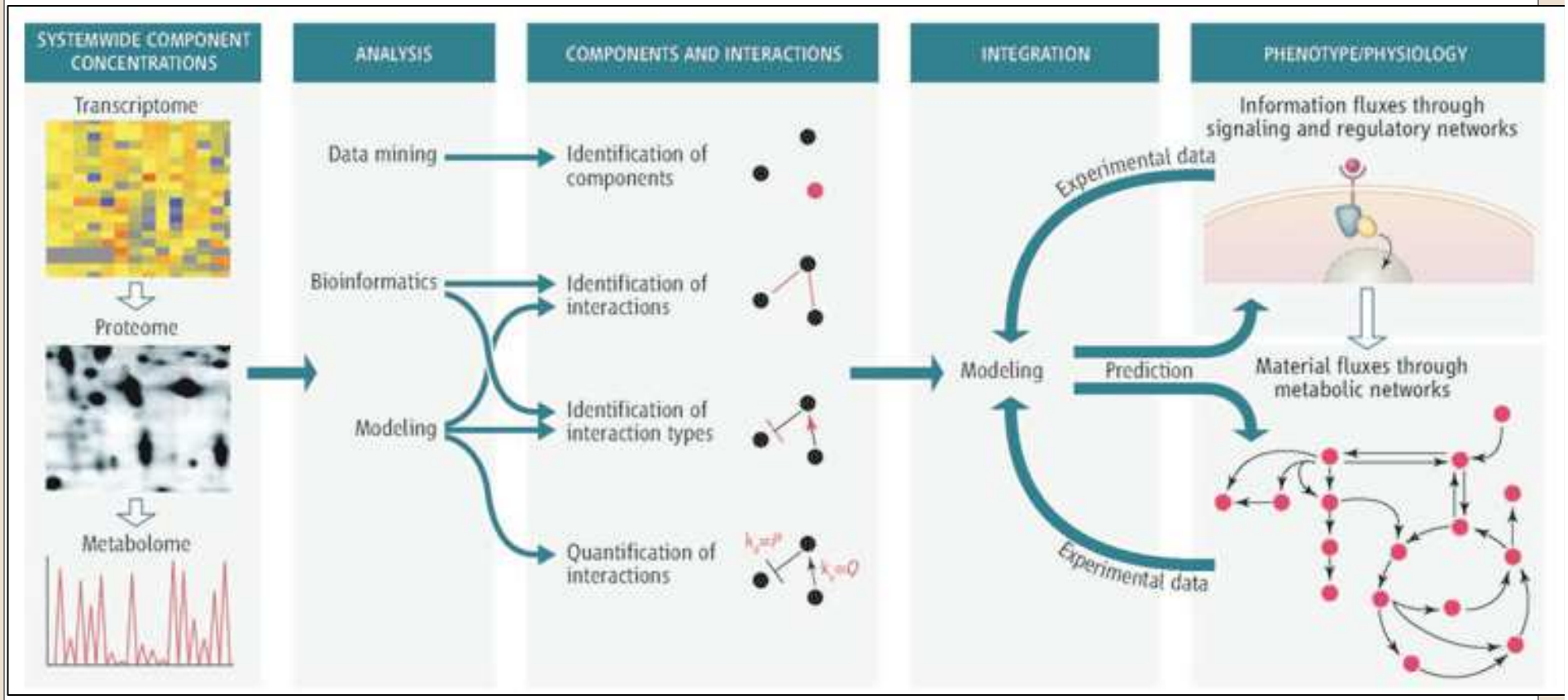


normal lymphoma

Proteomics: detection of proteins with high expression levels

Transcriptomics: differential expression between normal and lymphoma cells (GEO database; Basso set)

Systems Biology

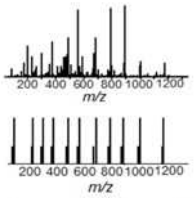


→G. Launay, Axe 2, 29/11
→ Axe 2, 30/11
→ L. Tichit, Axe 3, 29/11

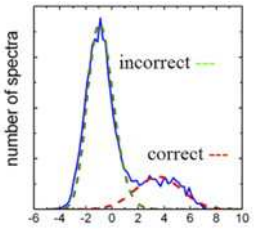
From Sauer, Heinemann & Zamboni, 2007, Science 316:550

Schematic overview of a typical workflow of the proteomics informatics processing of a data set

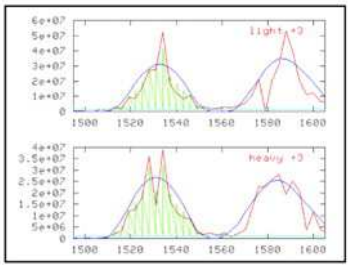
1. Conversion to and use of open data formats



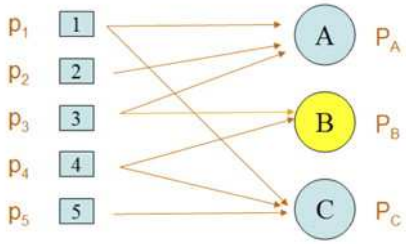
2. Spectrum identification with a search engine



3. Validation of identifications

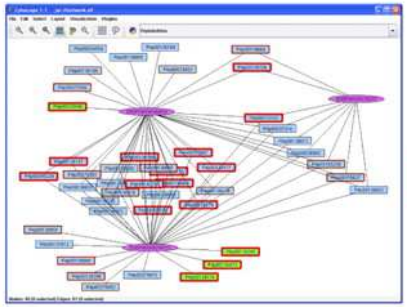


5. Quantification

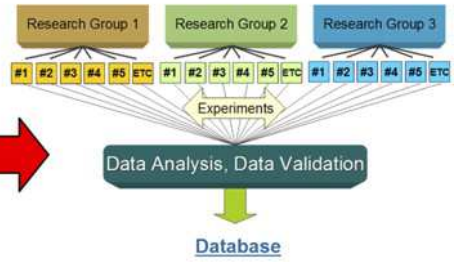


4. Protein inference

6. Organization in local data management systems



7. Interpretation of the protein lists



8. Transfer to public data repositories

Deutsch E W et al. Physiol. Genomics 2008;33:18-25

Some resources and references

- http://www.proteomicstutorials.org
- Cottrell JS. Protein identification using MS/MS data. J Proteomics. 2011 Sep 6;74(10):1842-51.
- Beck M, Claassen M, Aebersold R. Comprehensive proteomics. Curr Opin Biotechnol. 2011 Feb;22(1):3-8.
- Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiol Genomics. 2008 Mar 14;33(1):18-25.
- Walther TC, Mann M. Mass spectrometry-based proteomics in cell biology. J Cell Biol. 2010 Aug 23;190(4):491-500.