Introduction to Proteomics Analysis and Databases

Jon Ambler: jon@ambler.co.za
Nicky Mulder: nicola.mulder@uct.ac.za

What is Proteomics?

- Large-scale study of proteins to determine their function
- Proteome is protein complement of the genome
- Includes the study of:
  - Protein structure and function
  - Protein-protein interactions
  - Protein expression
  - Protein localisation
  - Protein modifications
  - Etc.

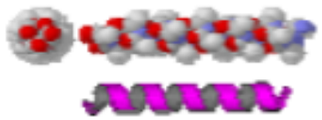
Proteomics studies

- Primary structure (*sequence*)

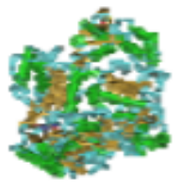
…YSFVATAER…

Mass spectrometry

- Secondary structure (*structural elements*)



- Tertiairy structure (*3D shape*)



Xray, NMR

- Modifications (*dynamic, function*)
  phosphorylation

Mass spectrometry

- Processing (*targetting, activation*)
  trypsin
  platelet activity

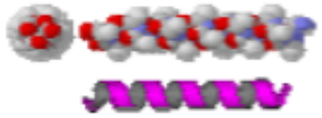
Localisation studies

Determine function at:

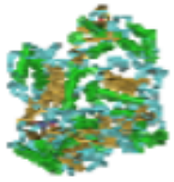
- Primary structure (*sequence*)

…YSFVATAER…

- Secondary structure (*structural elements*)



- Tertiairy structure (*3D shape*)



- Modifications (*dynamic, function*)
  phosphorylation
- Processing (*targetting, activation*)
  trypsin
  platelet activity

Through alignment based methods

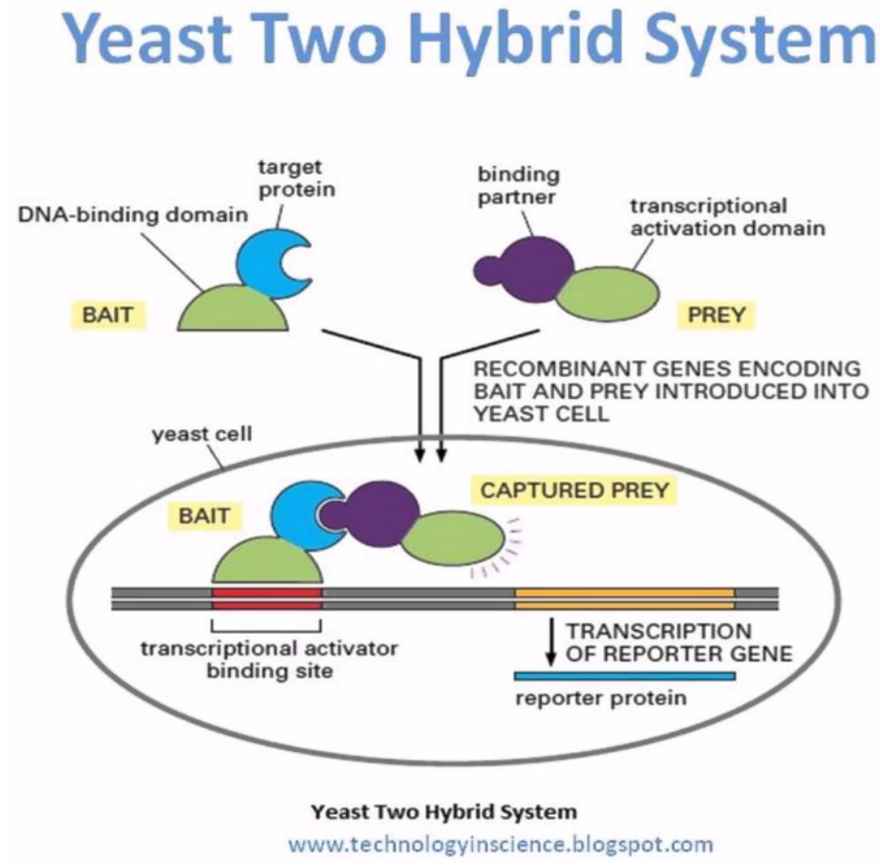
Identification of conserved motifs

Structural alignments

Assign function based on similarity

Physical interactions

- Experiments to identify physical interactions between DNA and proteins (for example TFs) or between two proteins:
  - Yeast two hybrid
  - Protein arrays



Protein-protein interaction databases

- Protein-protein interaction databases store pairwise interactions or complexes
    - IntAct
    - DIP (Database of Interacting Proteins)
    - BIND (Biomolecular Interaction Network Database)
    - STRING

Expression

- Differences in protein quantity between samples
- Timing
    - Is it always expressed?
        - Only expressed under certain conditions?
- Enzyme kinetics
    - Is it a rate limiting protein?
- Protein turnover rate

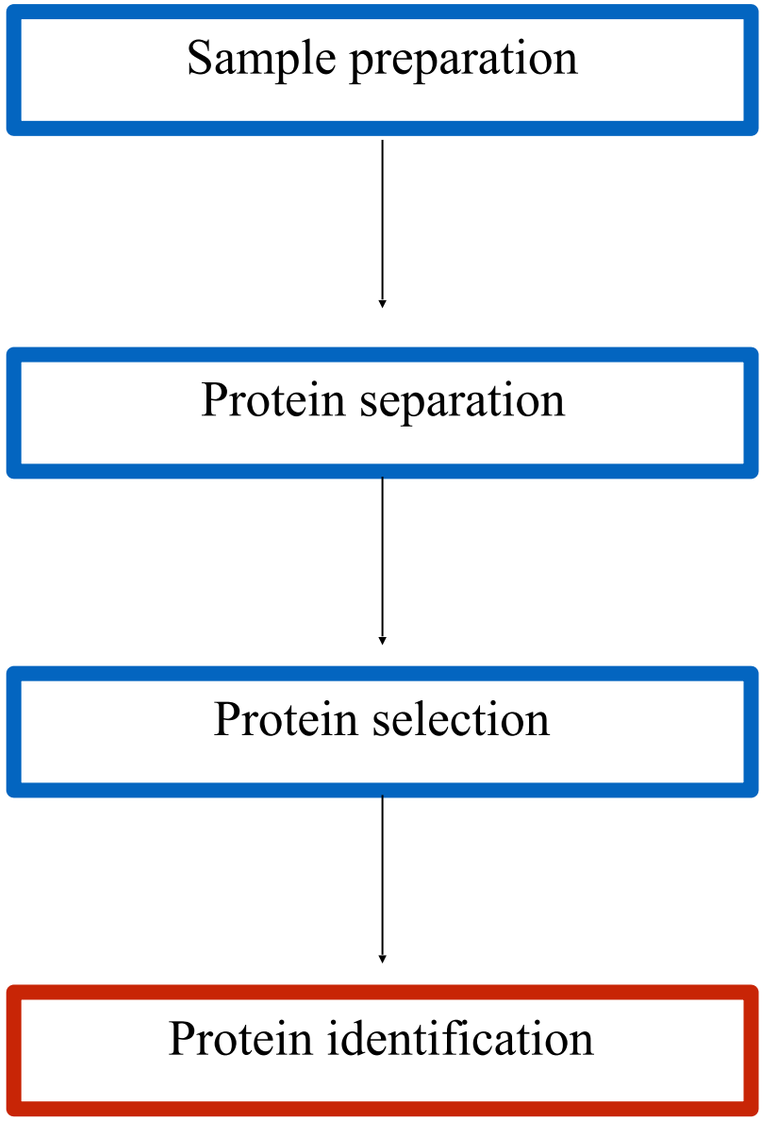
Localisation

- Relates to function
- Can be inferred by looking at signal sequence if present
- Determined experimentally from labelling and imaging techniques
- Co-localised proteins may share functional relationships
    - Not always case, e.g. in cytoplasm
- Localisation can change with environment

Modifications

- Various modifications
- Are mostly not “visible” at sequence level
- Have various effects on proteomic analysis
  - Phosphorylation, glycosylation, ubiquitination, etc…

Proteomics workflow



Isolation of tissue, growing isolates

Based on properties of the proteins (size, charge, etc…) 2-D PAGE, HPLC, ICAT, etc.

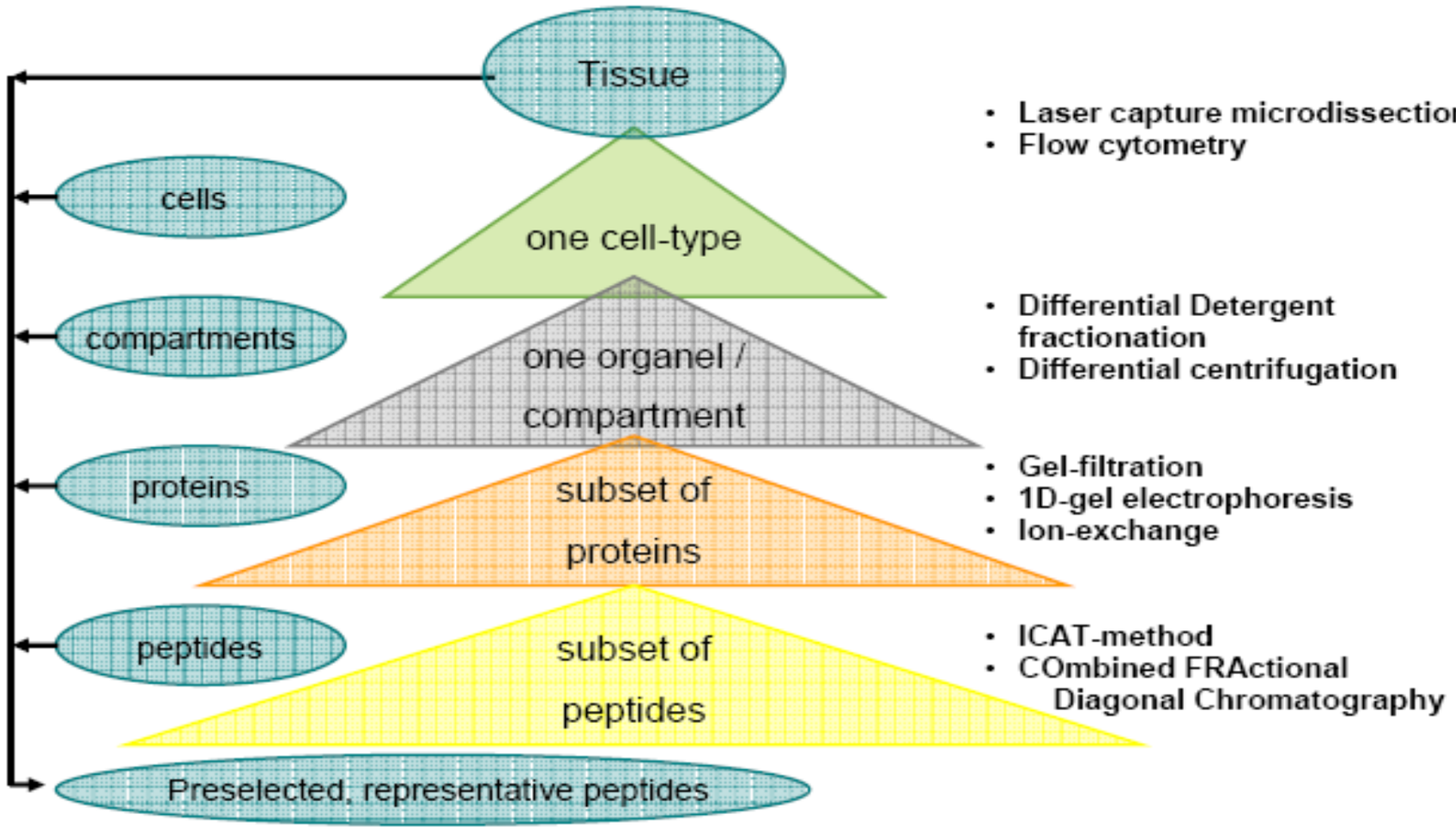
Dependant on seperation method

Mass spectrometry

Wet lab

In silico

Protein seperation



Protein seperation

2-D PAGE

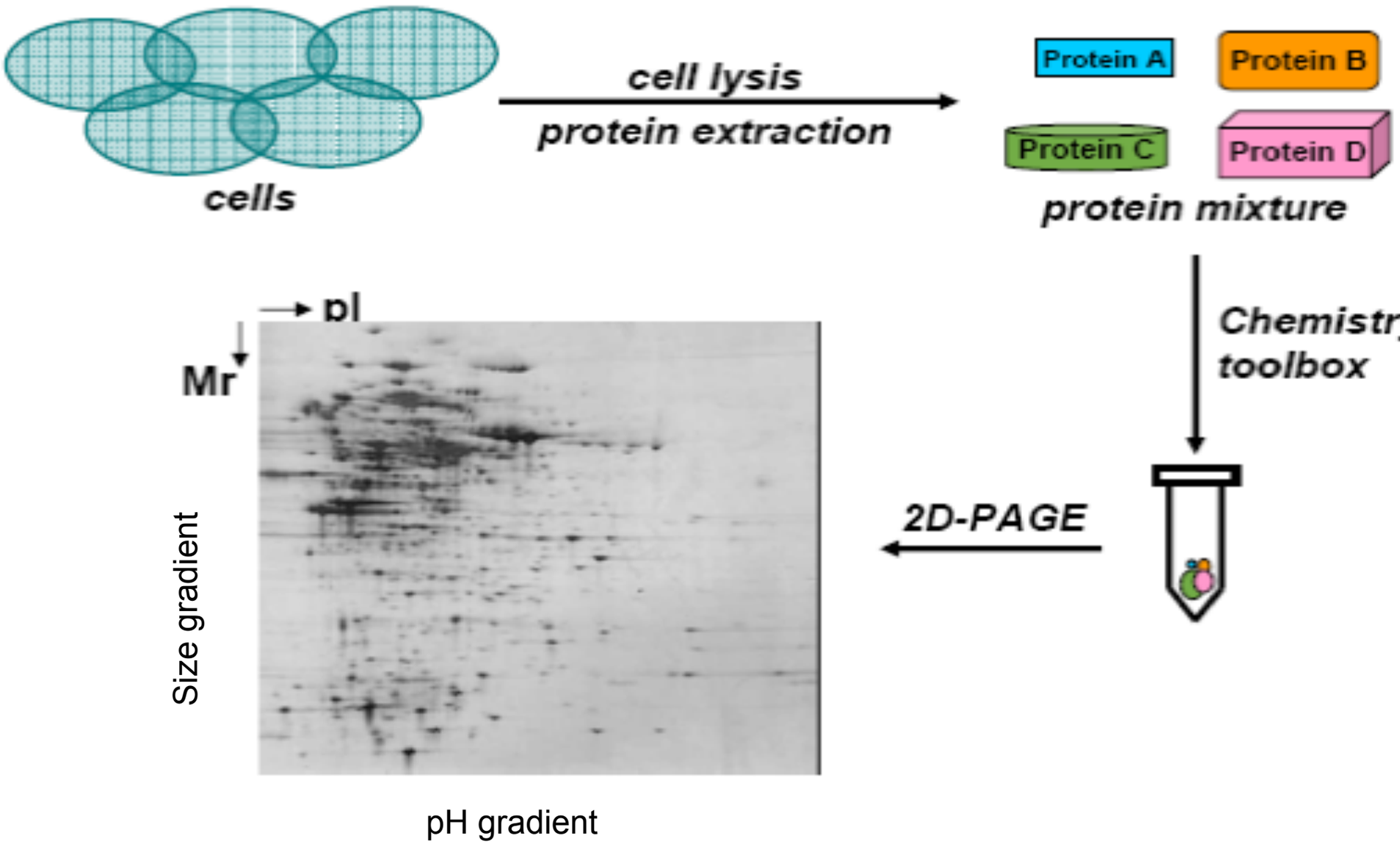
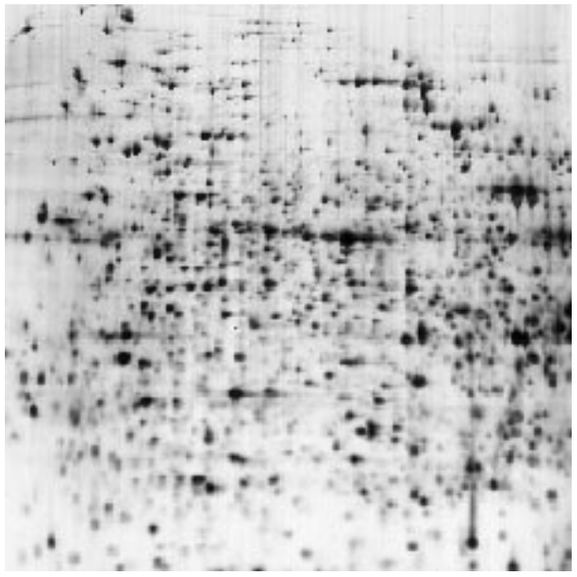


Image analysis

- Image capture and precessing
  - **Removing background noise**
  - Thresholding
    - Identification of centroids
- Image comparison:
  - Measuring intensities
  - Finding difference between gels



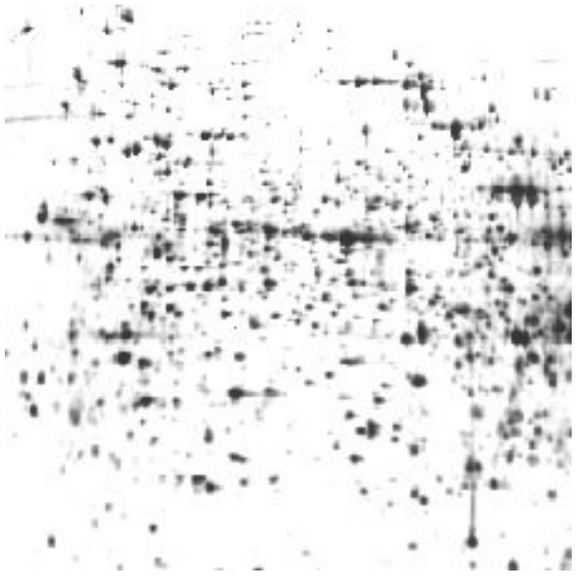


Image analysis

- Image capture and precessing
  - Removing background noise
  - **Thresholding**
  - Identification of centroids
- Image comparison:
  - Measuring intensities
  - Finding difference between gels

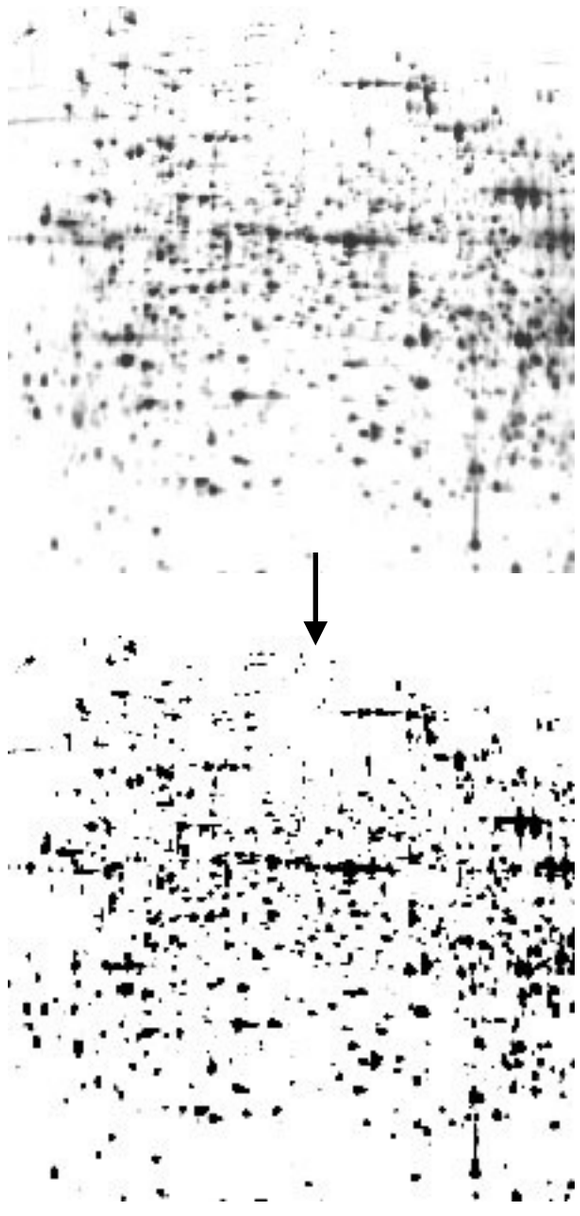


Image analysis

- Image capture and precessing
  - Removing background noise
  - Thresholding
  - **Identification of centroids**
- Image comparison:
  - Measuring intensities
  - Finding difference between gels

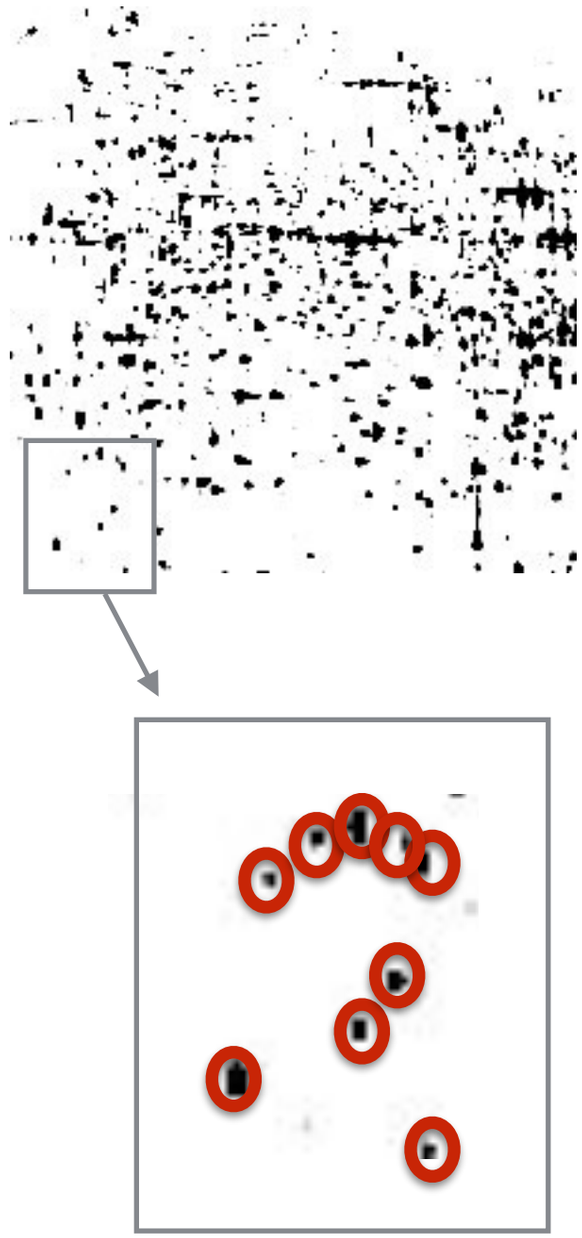
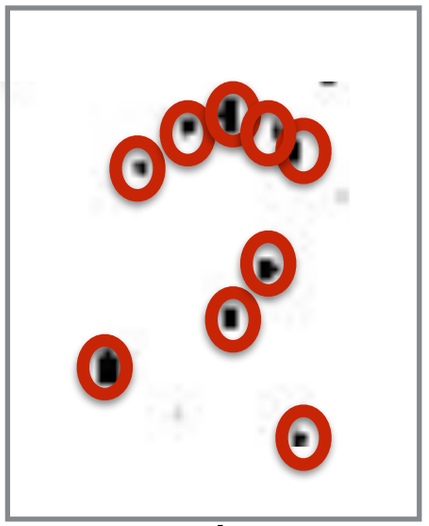


Image analysis

- Image capture and precessing
  - Removing background noise
  - Thresholding
  - **Identification of centroids**
- Image comparison:
  - Measuring intensities
  - Finding difference between gels



0	0	0	0	0	0	0	0
0	0	0	0	**3**	0	0	0
0	0	0	**2**	0	**1**	**2**	0
0	0	**1**	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	**2**	0	0
0	0	0	0	**1**	0	0	0
0	0	**3**	0	0	0	0	0
0	0	0	0	0	**3**	0	0
0	0	0	0	0	0	0	0

Image analysis

- Image capture and precessing
  - Removing background noise
  - Thresholding
  - Identification of centroids
- **Image comparison:**
  - **Measuring intensities**
  - **Finding difference between gels**

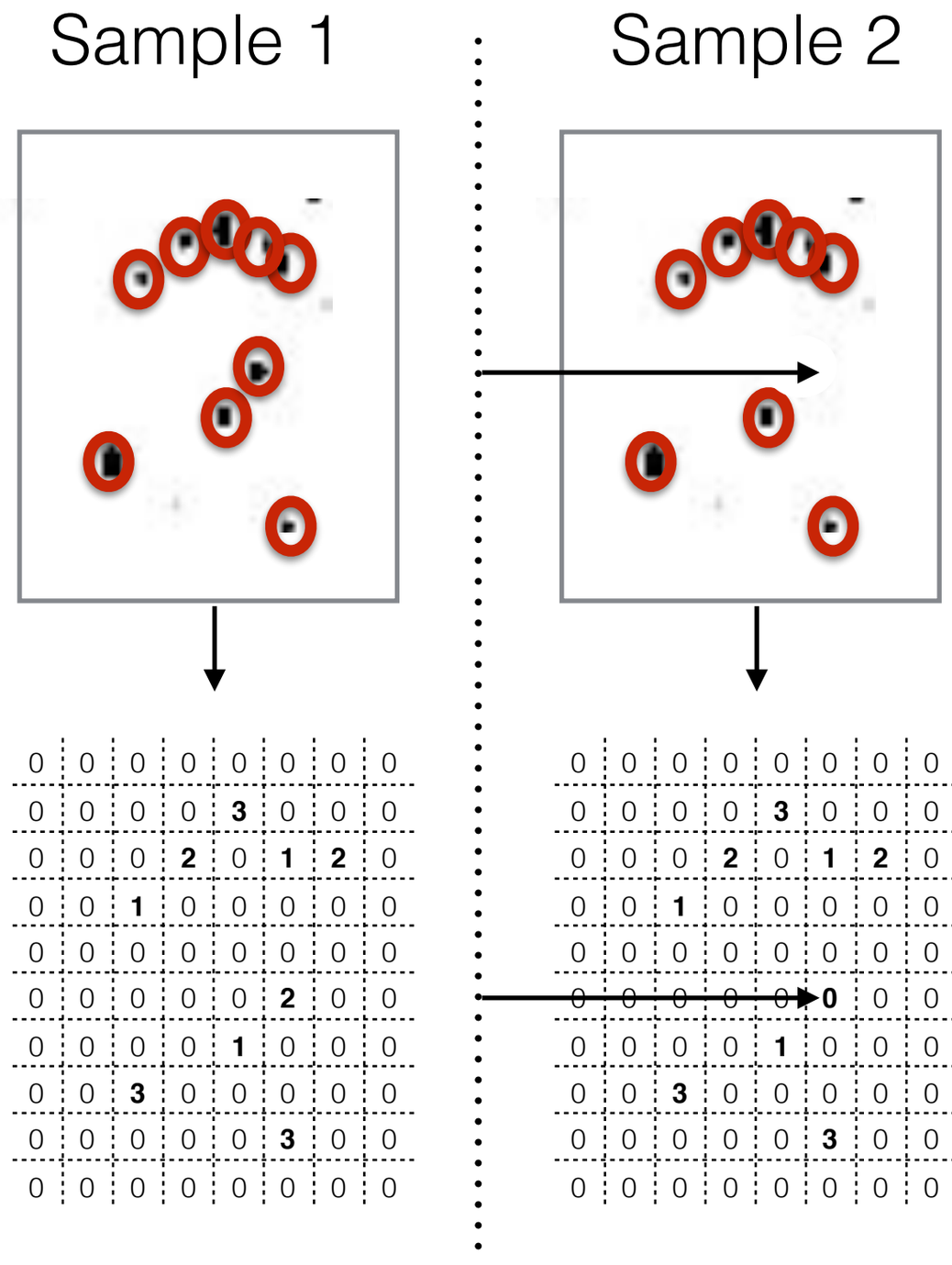
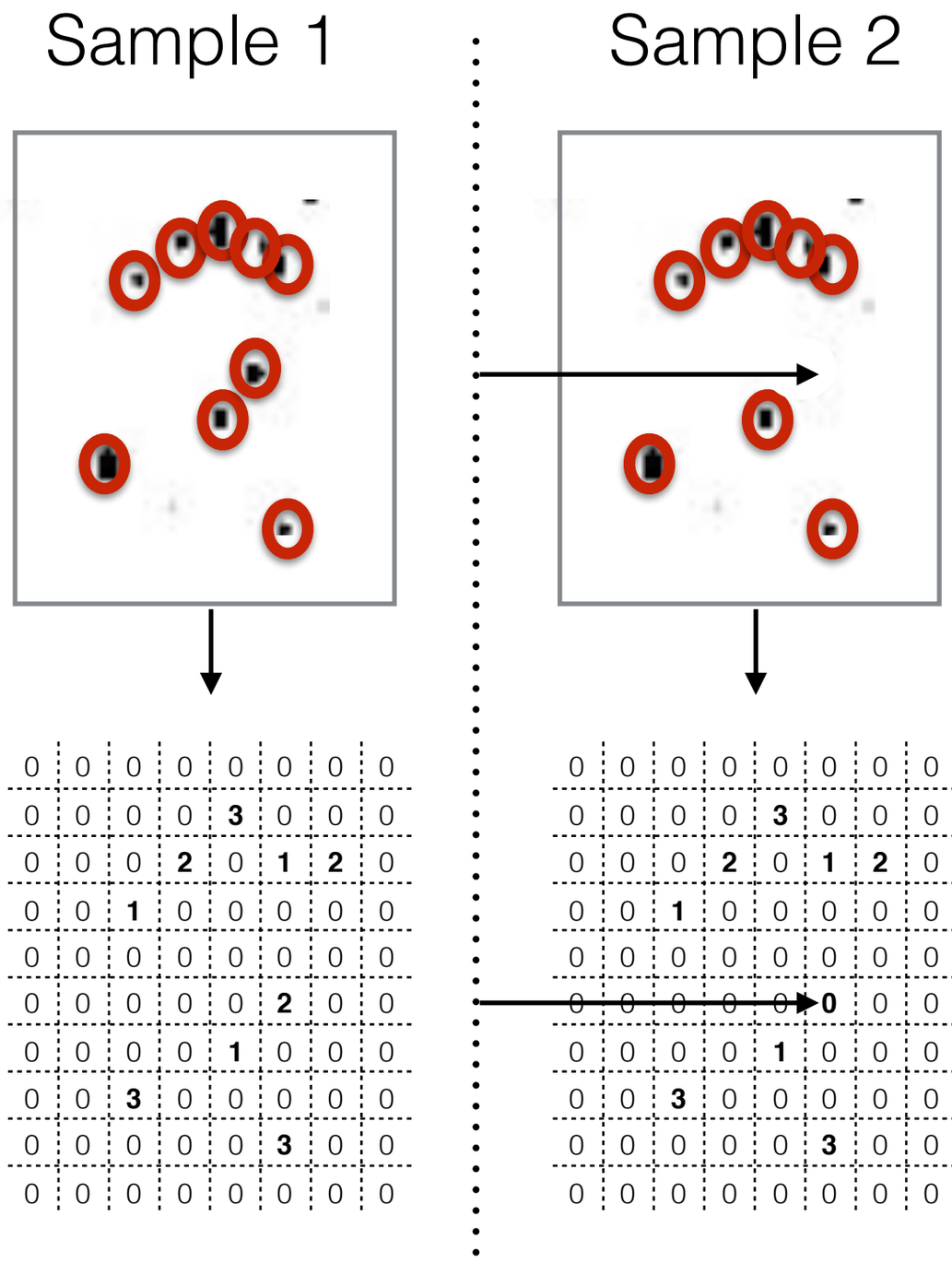
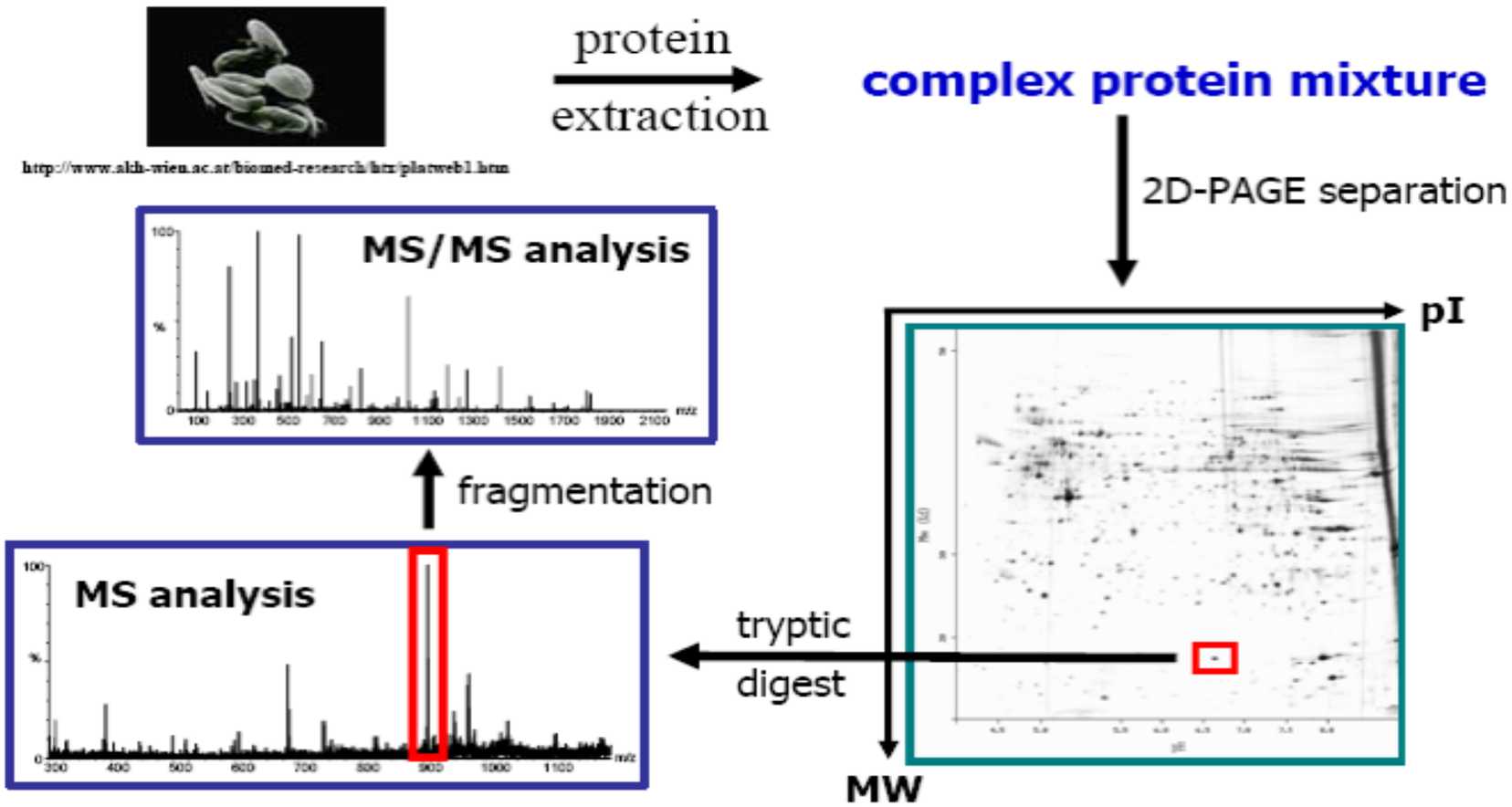


Image analysis

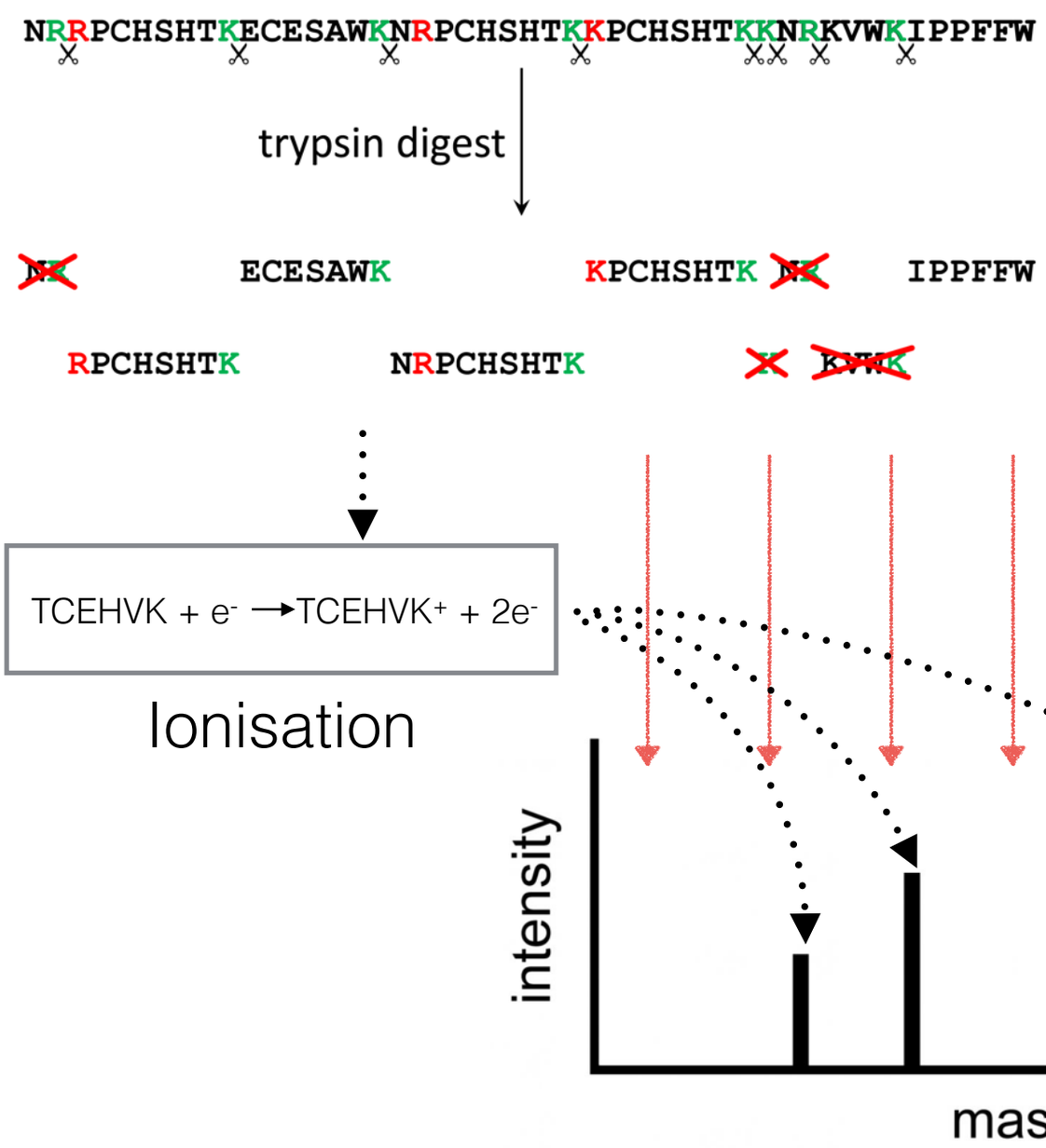
- Challenges:
  - Some proteins can't be detected
    - Low abundance
    - Highly charged (run off gel)
  - Imperfect separation –multiple proteins in a spot
  - Reproducibility (Replicates needed ideally)



Mass Spectrometry (MS)

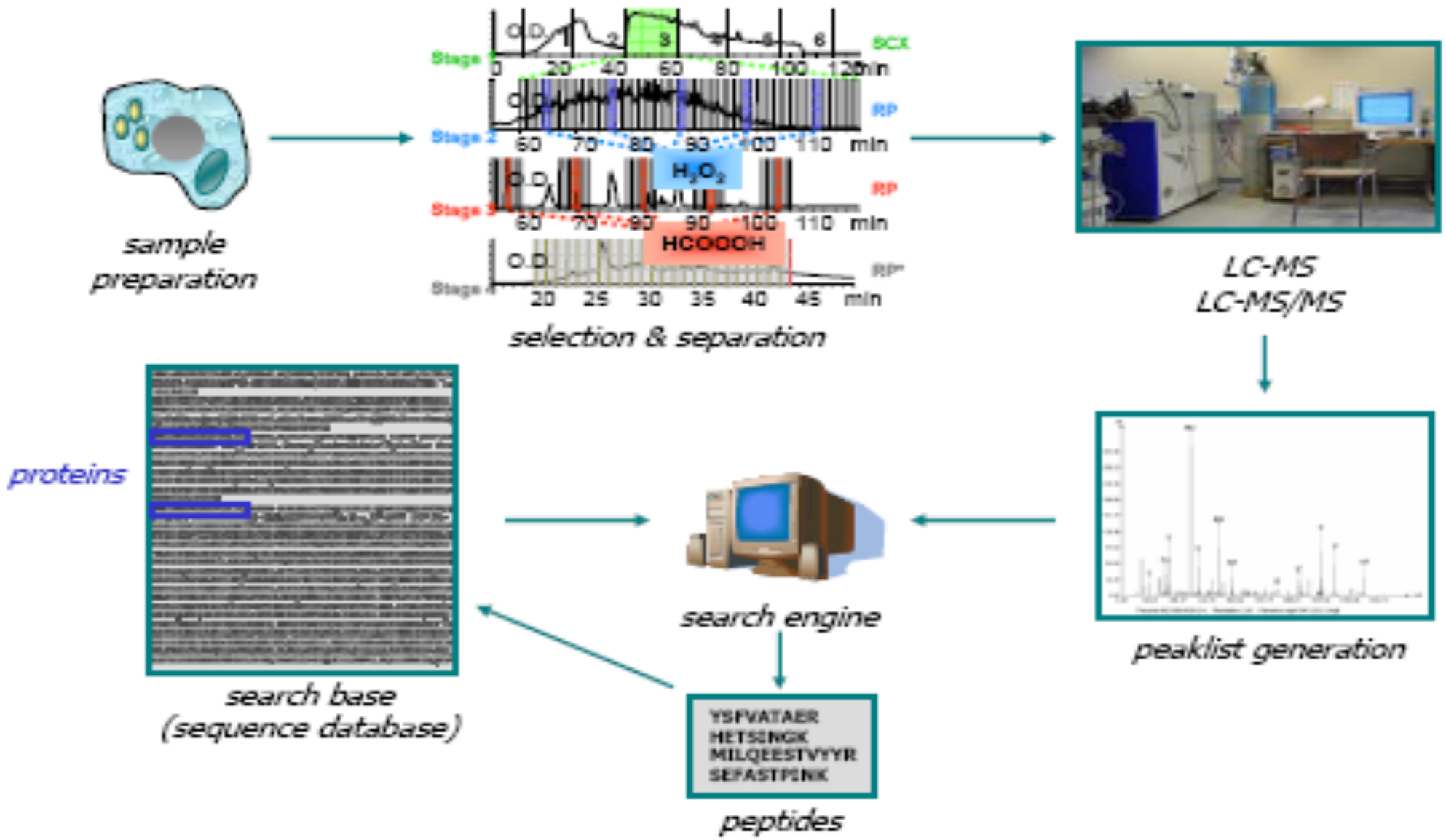


Mass Spectrometry (MS)



- Protein digestion with trypsin (Cleaves Arginine [R] and Lysine [K])
- F = m.a

Protein identification with Mass Spectrometry



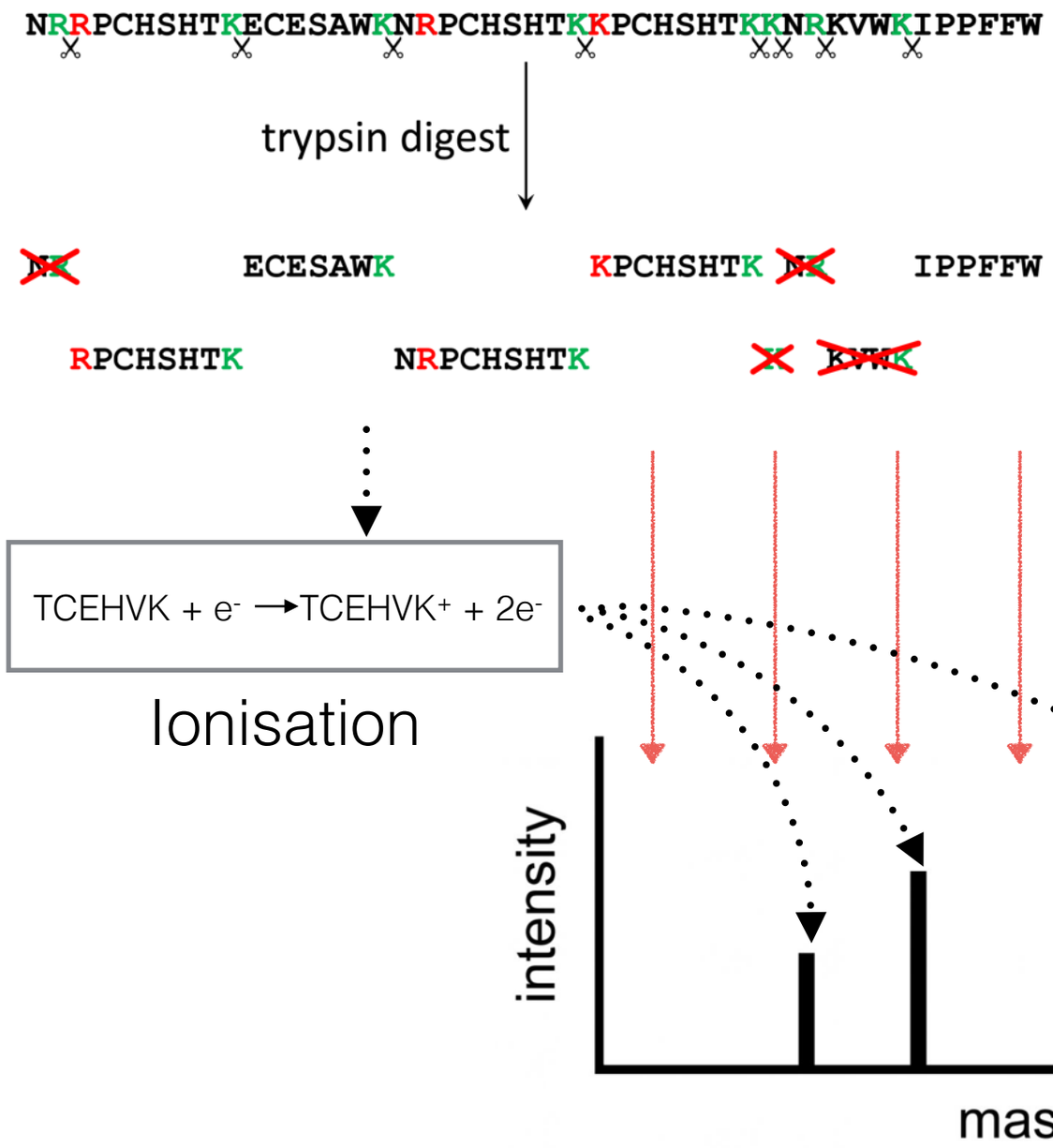
Problems with mass spec ID

- Protein samples often contain a mixture of proteins
- Digestion/fragmentation isn't always complete
- Not all proteins get ionized
- Background noise in spectra
- Proteins can contain modifications, which will change mass
    - (Phosphorylation, glycosylation, ubiquitination, etc…)

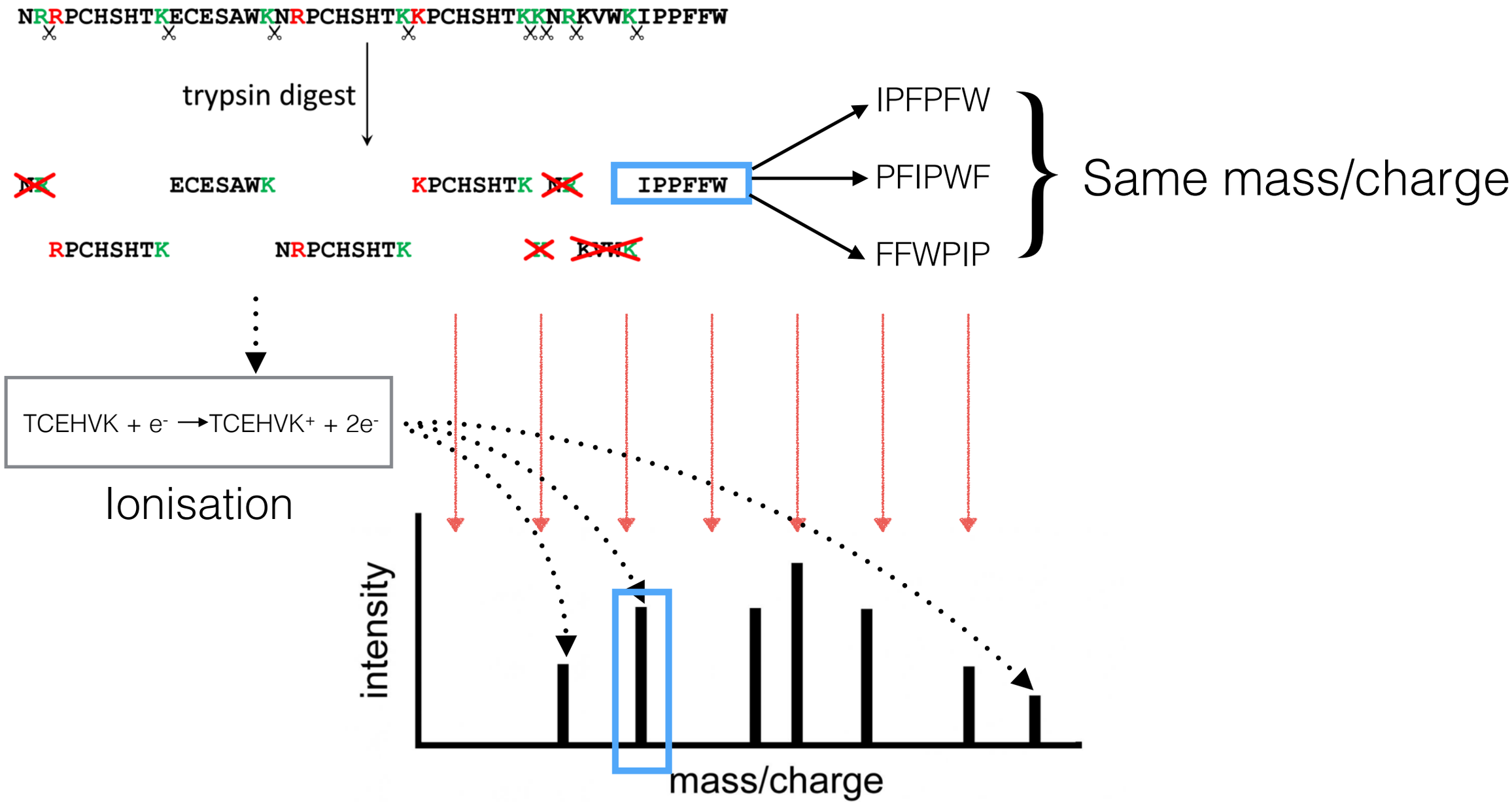
Solving problems –Tandem MS

- Two rounds of mass spec
- Fragment peptides and obtain spectrum
- Select peak you want then fragment this again
- Able to better separate peptides/proteins

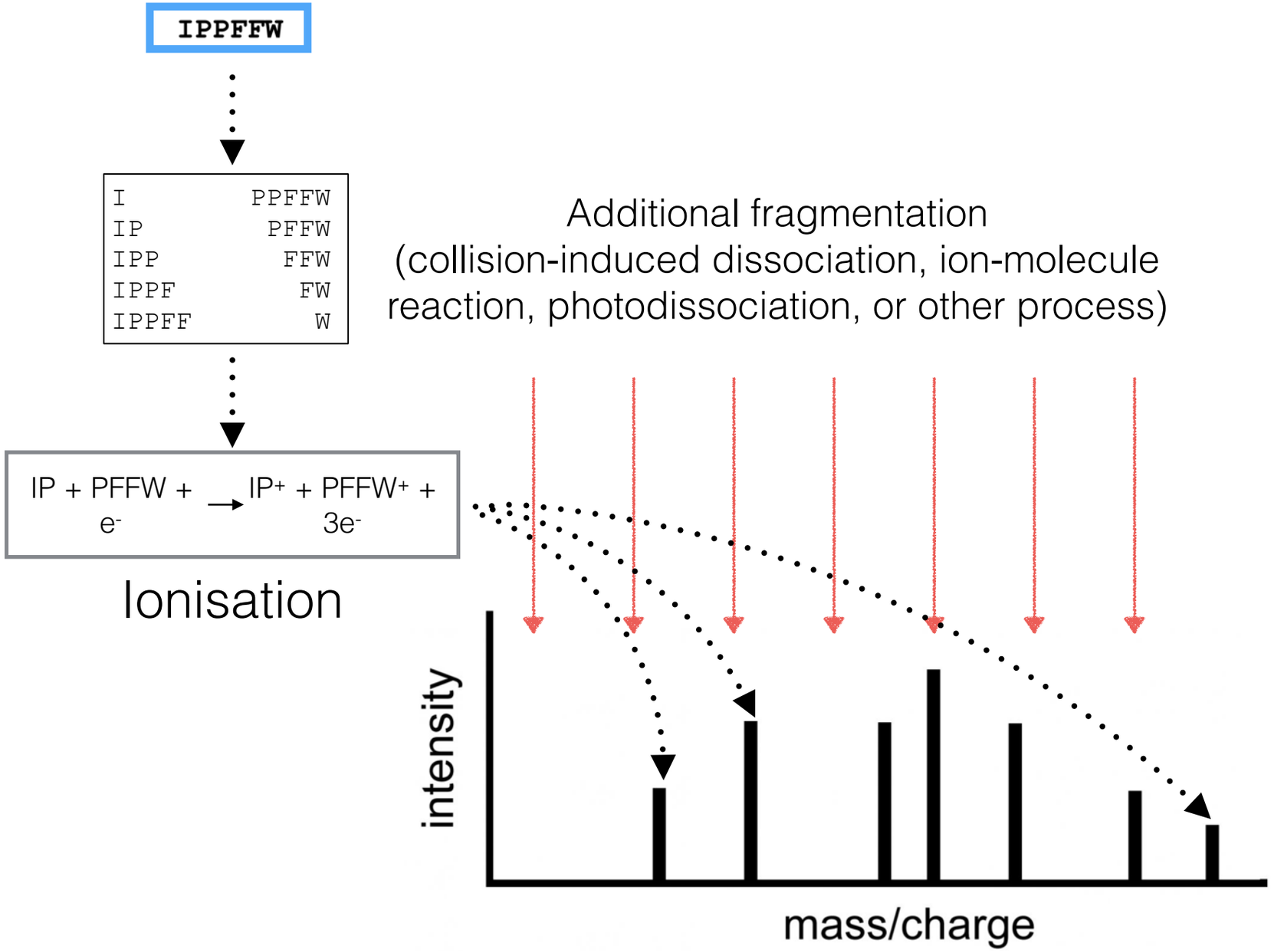
Tandem Mass Spectrometry (MS/MS)



Tandem Mass Spectrometry (MS/MS)

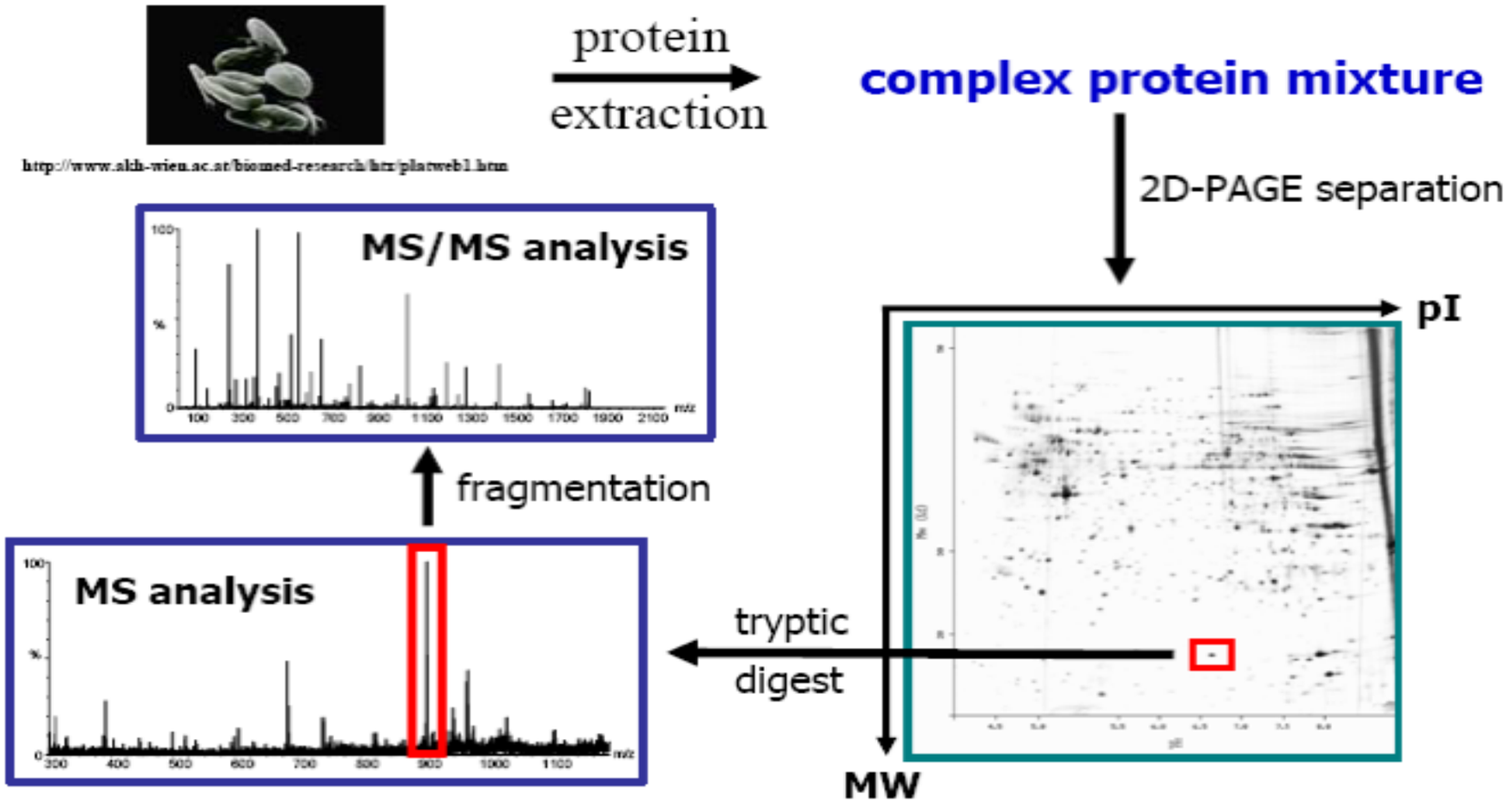


Tandem Mass Spectrometry (MS/MS)



(New spectra produced)
compare with theoretical
peptide spectra;
ID = best similarity

In summary:


protein
extraction
http://www.akh-wien.ac.at/biomed-research/htr/plantweb1.htm
complex protein mixture
2D-PAGE separation
pI
MW
MS/MS analysis
fragmentation
MS analysis
tryptic
digest

In addition:

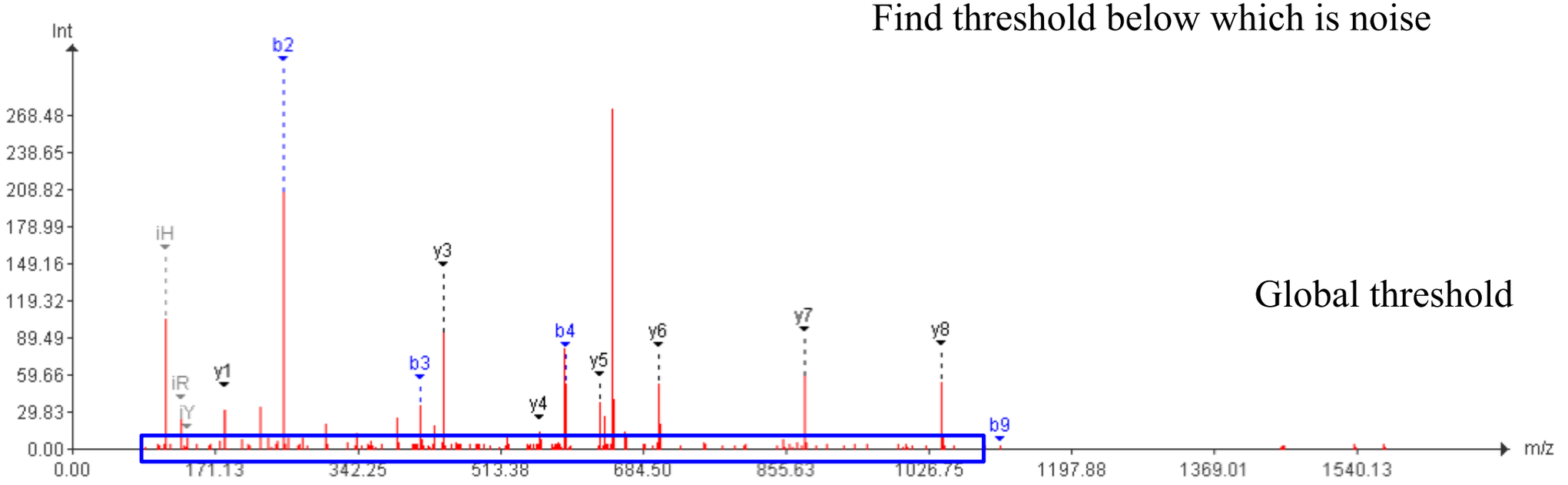
- Matrix-assisted laser desorption/ionization time of flight (MALDI-TOF)
- LASER = Light Amplification by Stimulated Emission of Radiation

Data analysis of MS

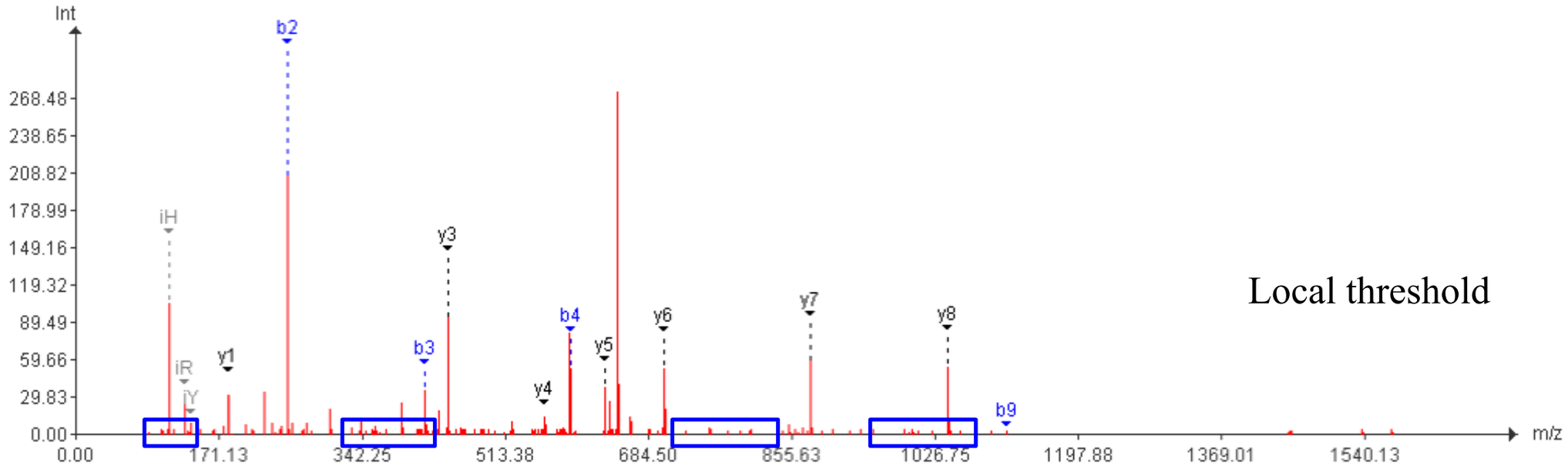
- Pre-processing
    - Noise reduction
    - Charge deconvolution
        - One peptide may have multiple charge states
    - Peak picking
- Spectrum filtering and clustering
- Protein identification
- Pathway analysis
- Enrichment analysis

Pre-processing: noise reduction

Find threshold below which is noise



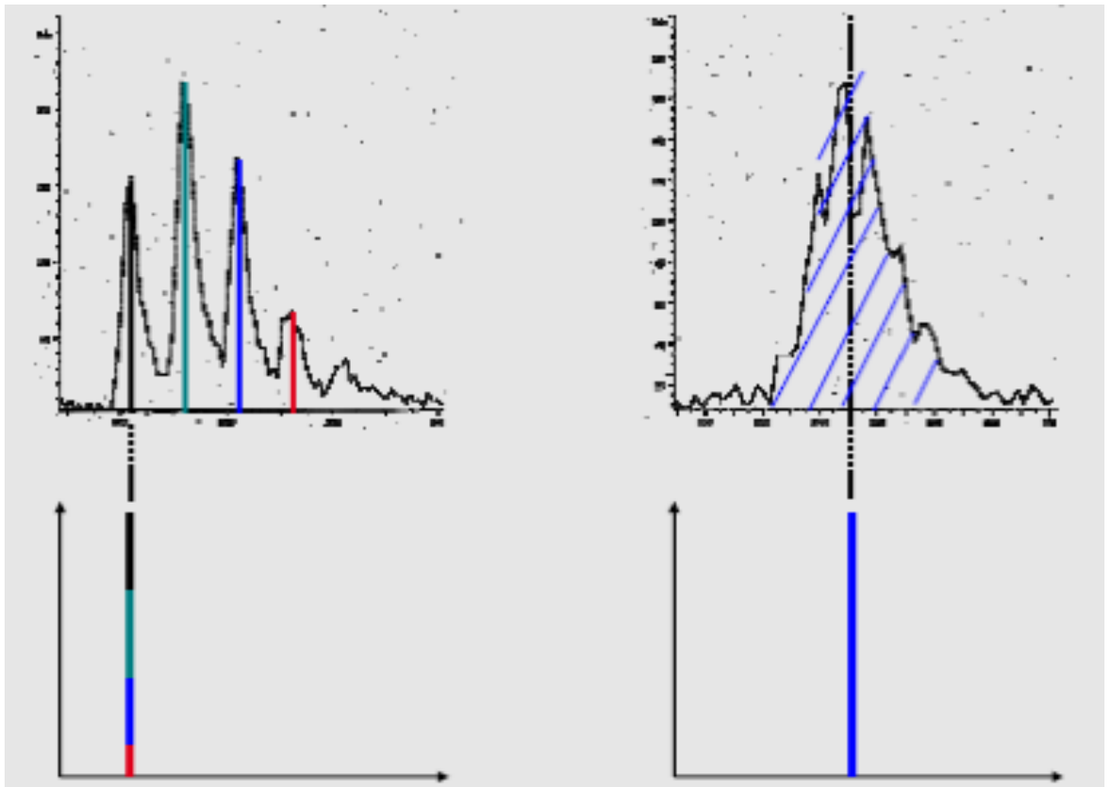
Global threshold



Local threshold

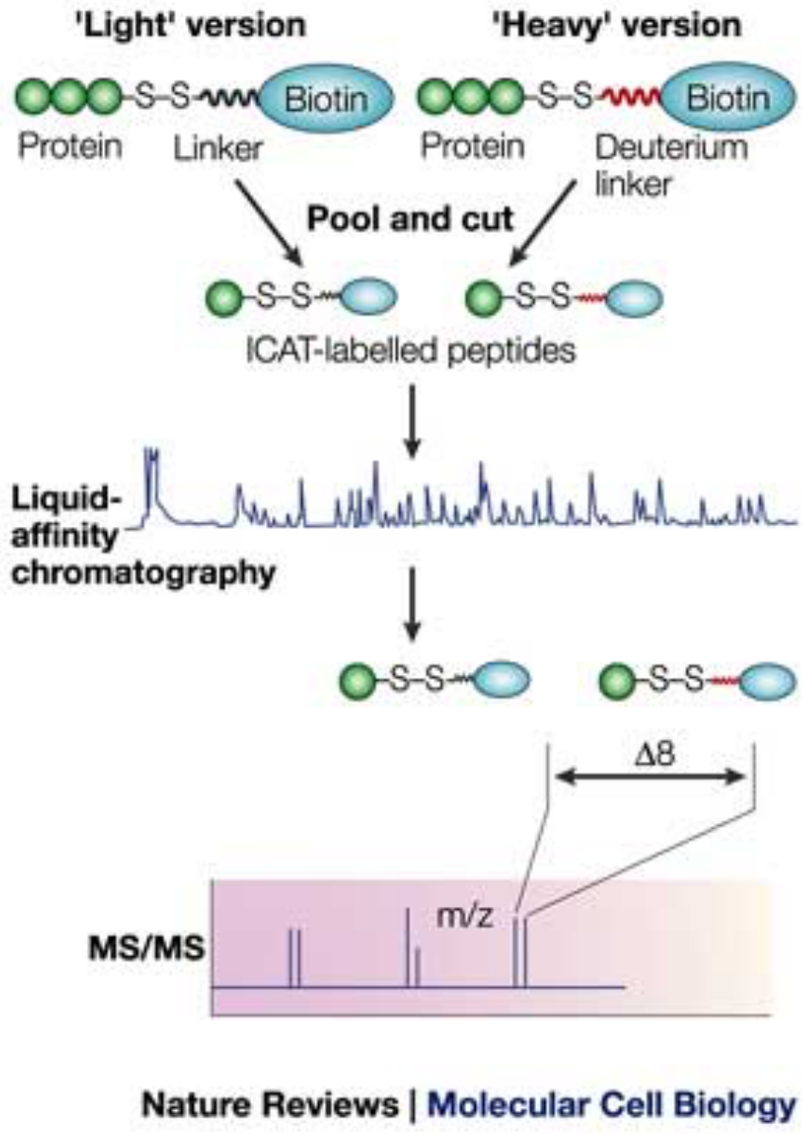
Pre-processing: peak picking

- The process of extracting this information, that means the conversion of the ”raw” ion count data acquired by the mass spectrometer into peak lists for further processing



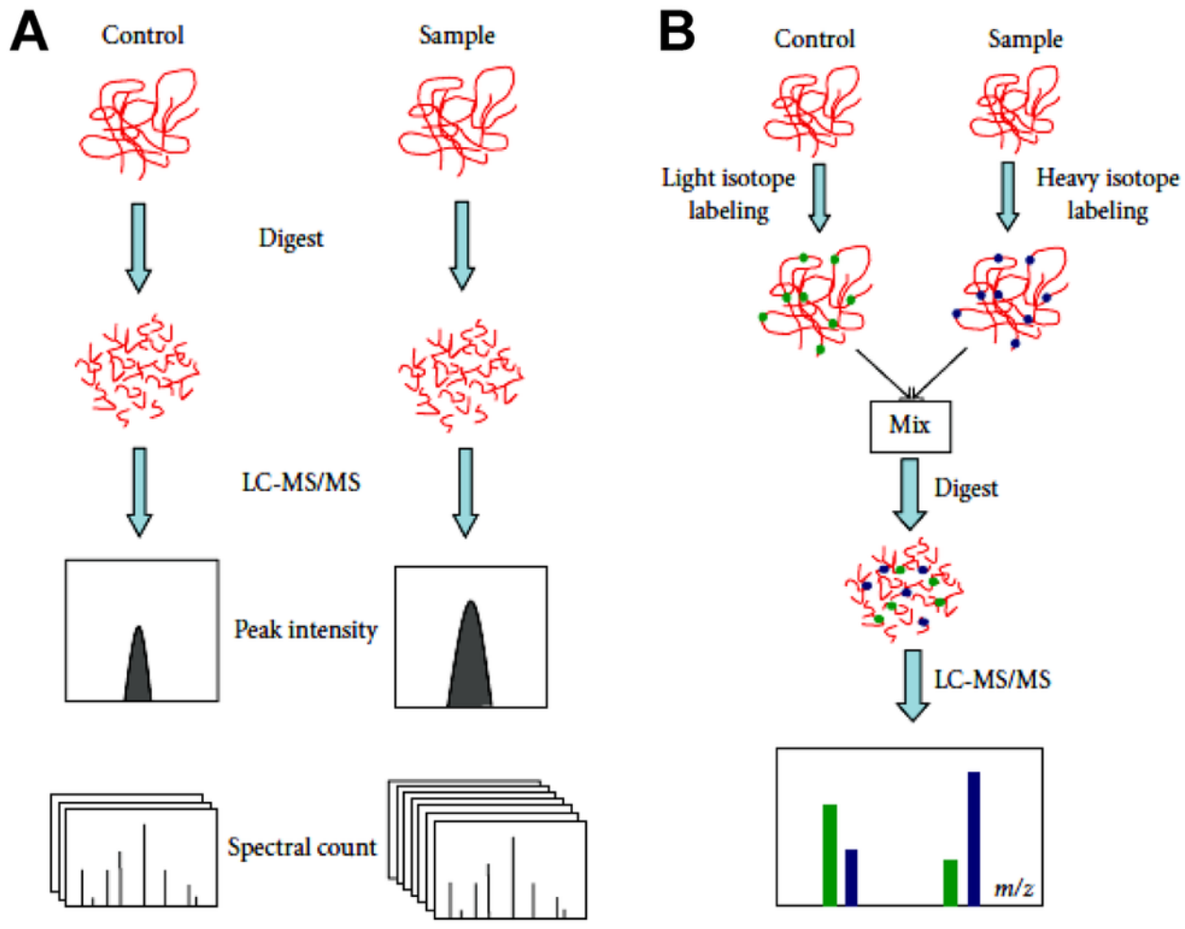
Comparative proteomics

- Quantification of protein differential expression:
  - Isotope-coded affinity tags (ICAT)
  - Label-free quantification
    - By spectral counting



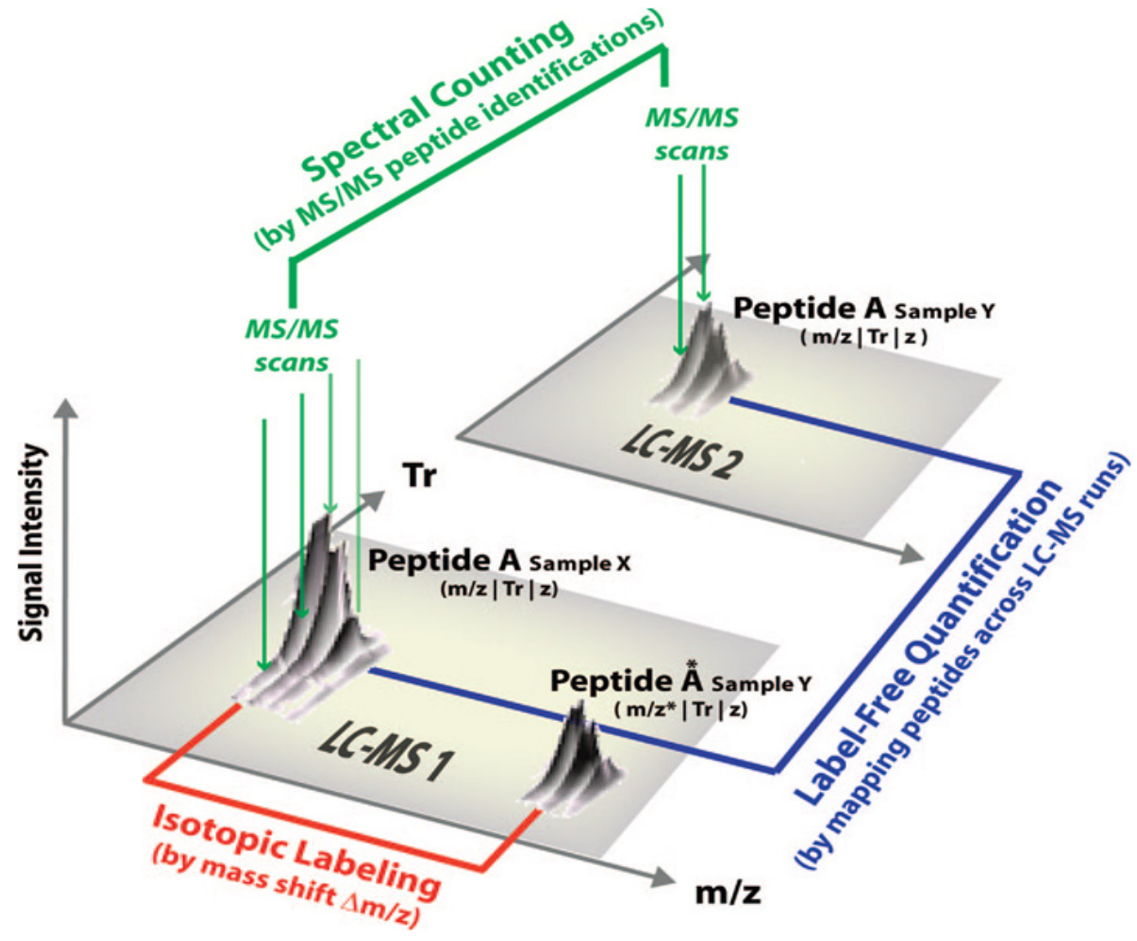
Comparative proteomics

- Quantification of protein differential expression:
  - Isotope-coded affinity tags (ICAT)
  - Label-free quantification
    - By spectral counting



Comparative proteomics

- Quantification of protein differential expression:
  - Isotope-coded affinity tags (ICAT)
  - Label-free quantification
    - By spectral counting

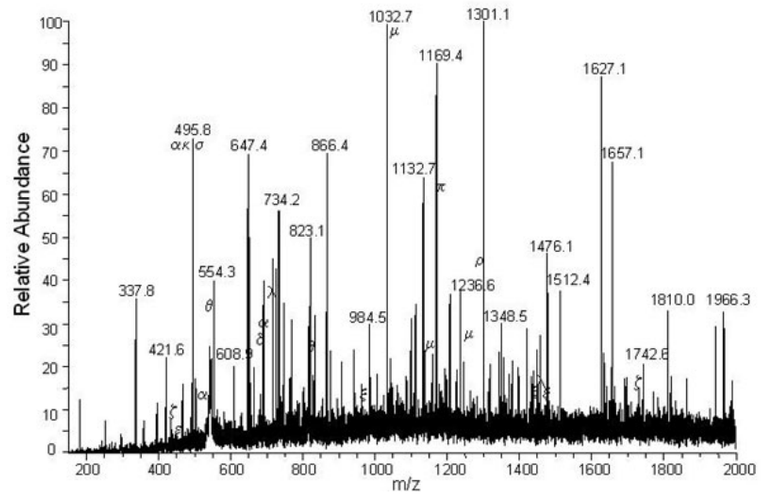


Protein ID

- Protein identification through mass spectrometry can be done in many ways:
    - Peptide Mass Fingerprinting
    - Tandem MS
        - Peptide Fragment Fingerprinting
- Summarised as:
    - Fragment
    - Generate spectra
    - Compare to database

Protein identification from Tandem MS spectra

- Compare to peptide spectra from databases
  - Use existing DB
    - NCBI
    - EMBL - UniProt
  - Create your own
    - In silico predicted spectra
- Sequence databases differ
  - Content
  - Redundant / non-redundant



DB

1: YFVAT → VAT, YFV, FVA

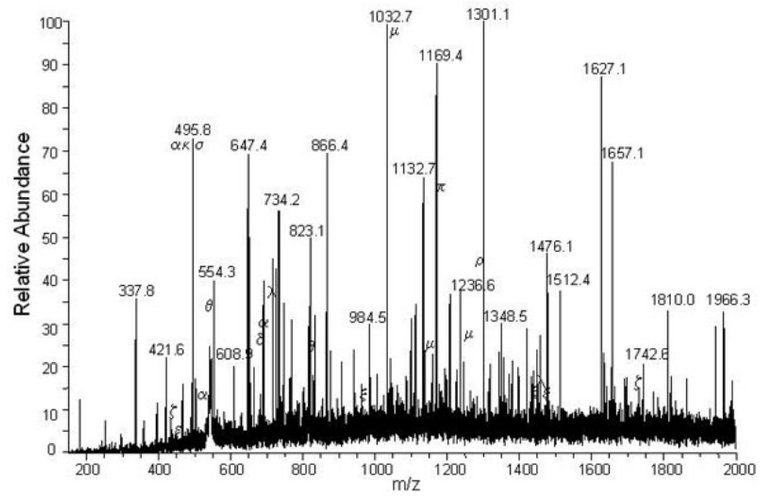
2: FVAD → VAD, FVA

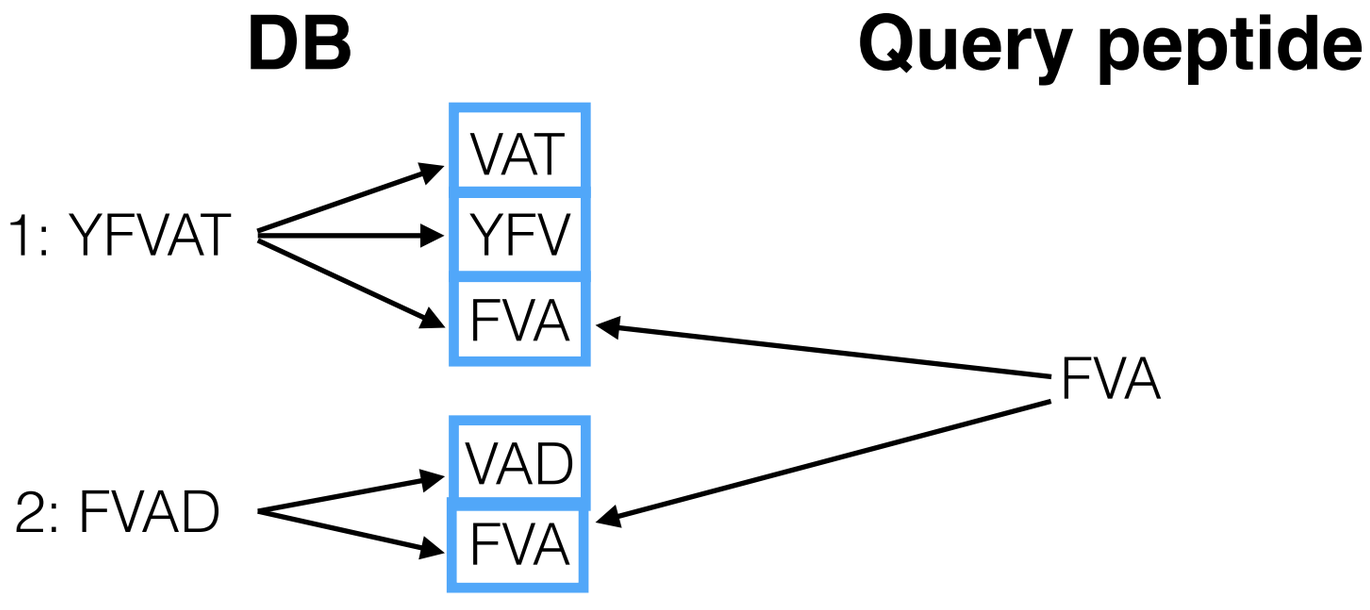
Query peptide

FVA

Protein identification from Tandem MS spectra

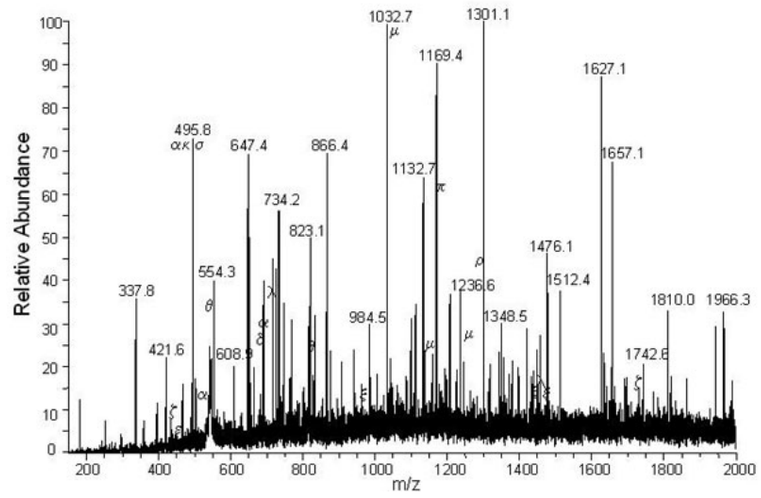
- Compare to peptide spectra from databases
  - Use existing DB
    - NCBI
    - EMBL - UniProt
  - Create your own
    - In silico predicted spectra
- Sequence databases differ
  - Content
  - Redundant / non-redundant

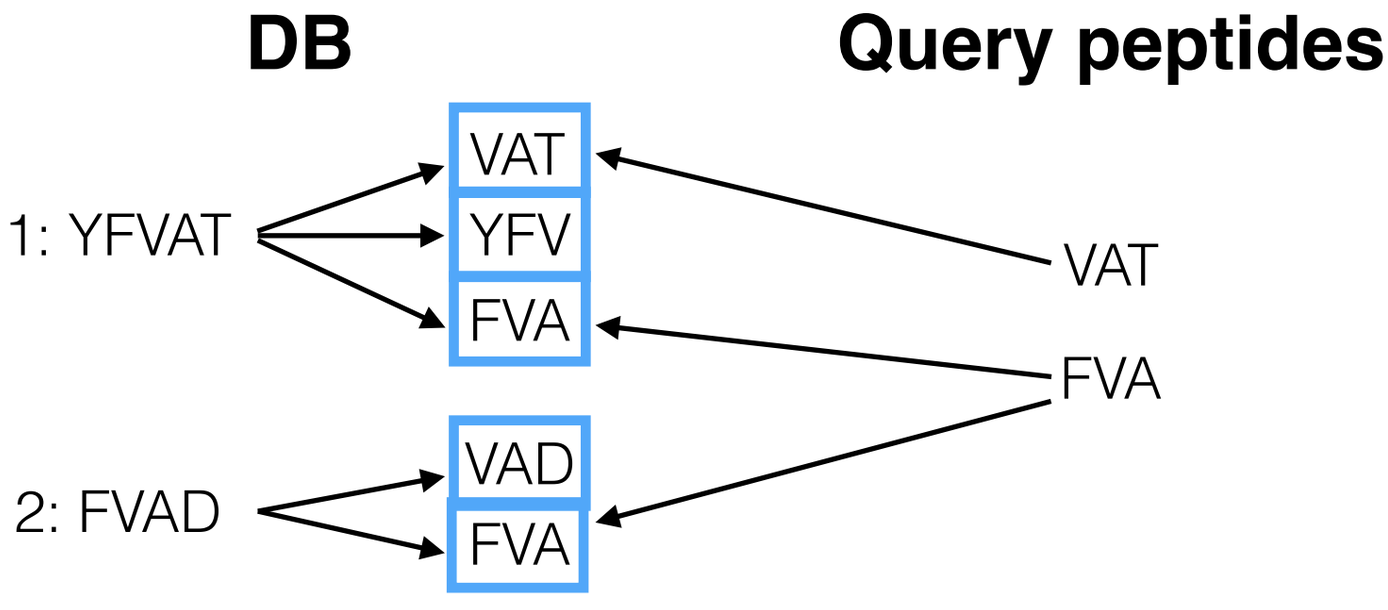




Protein identification from Tandem MS spectra

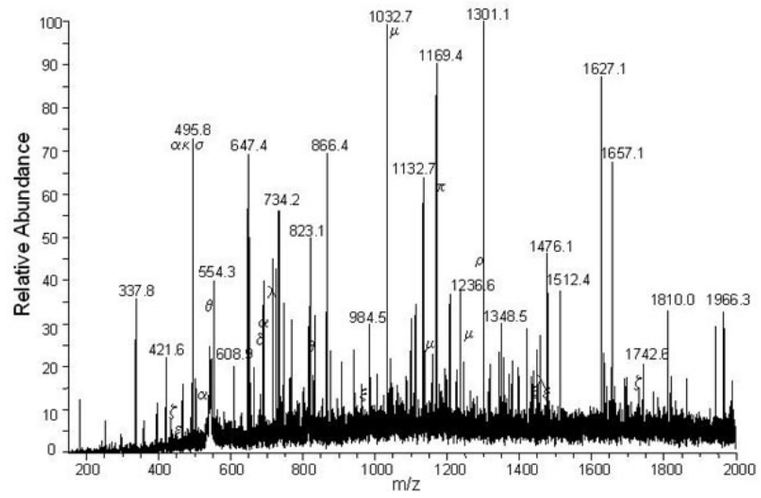
- Compare to peptide spectra from databases
  - Use existing DB
    - NCBI
    - EMBL - UniProt
  - Create your own
    - In silico predicted spectra
- Sequence databases differ
  - Content
  - Redundant / non-redundant

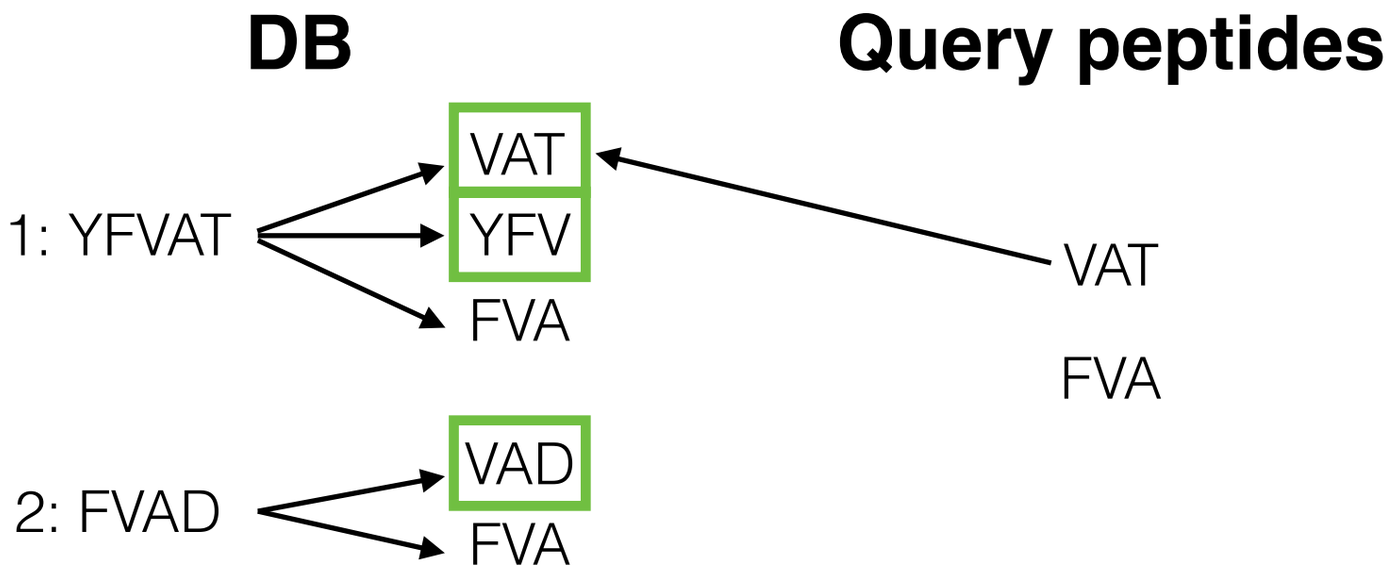




Protein identification from Tandem MS spectra

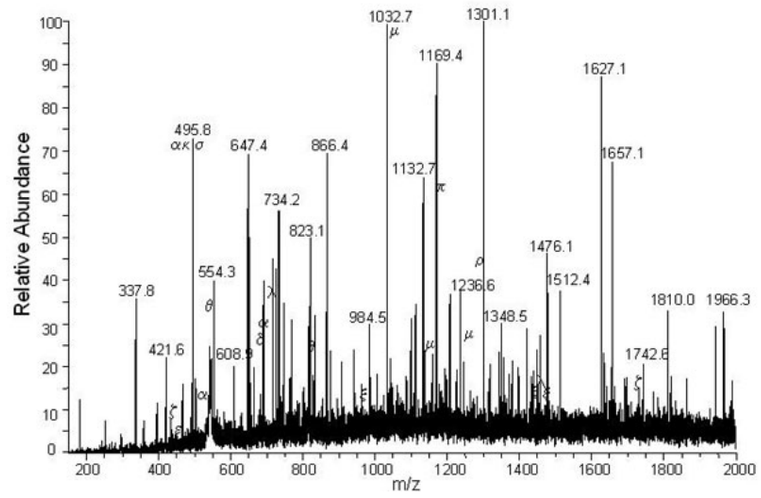
- Compare to peptide spectra from databases
  - Use existing DB
    - NCBI
    - EMBL - UniProt
  - Create your own
    - In silico predicted spectra
- Sequence databases differ
  - Content
  - Redundant / non-redundant



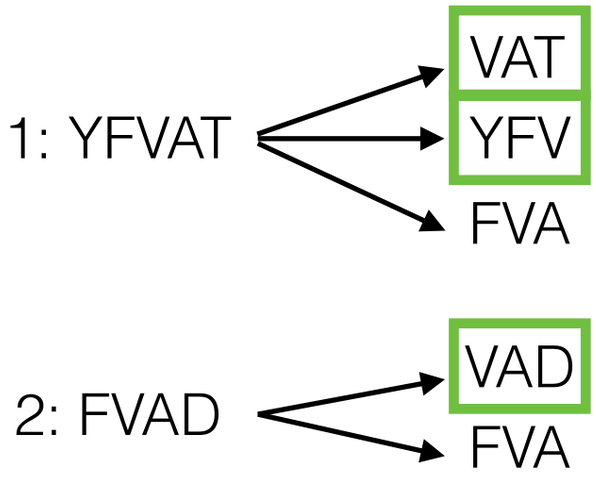


Protein identification from Tandem MS spectra

- Compare to peptide spectra from databases
  - Use existing DB
    - NCBI
    - EMBL - UniProt
  - Create your own
    - In silico predicted spectra
- Sequence databases differ
  - Content
  - Redundant / non-redundant



DB



Query peptide

FVA?

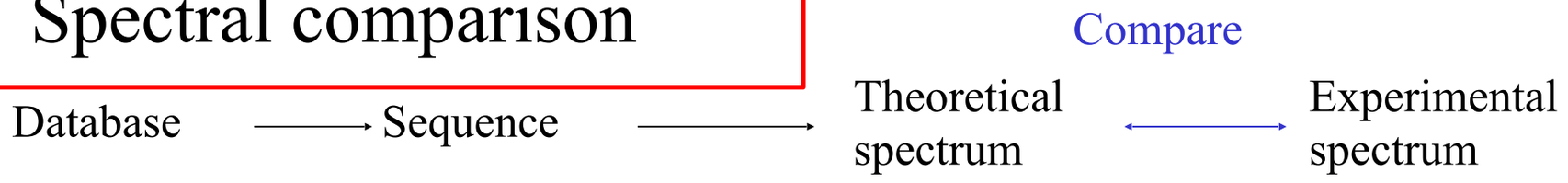
Peptide Fragment Fingerprinting (PFF)

- Identification of a protein based on the peptide fragmentation pattern after enzymatic digestion

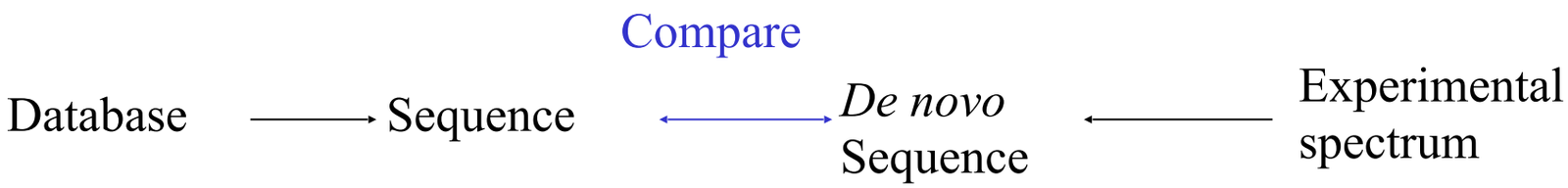
- Assumptions:
  - All peaks in spectrum are from the same protein
  - The protein is in the same form as it is in the database
  - Protein is completely digested
  - All pieces produce a signal

Peptide Fragment Fingerprinting (PFF)

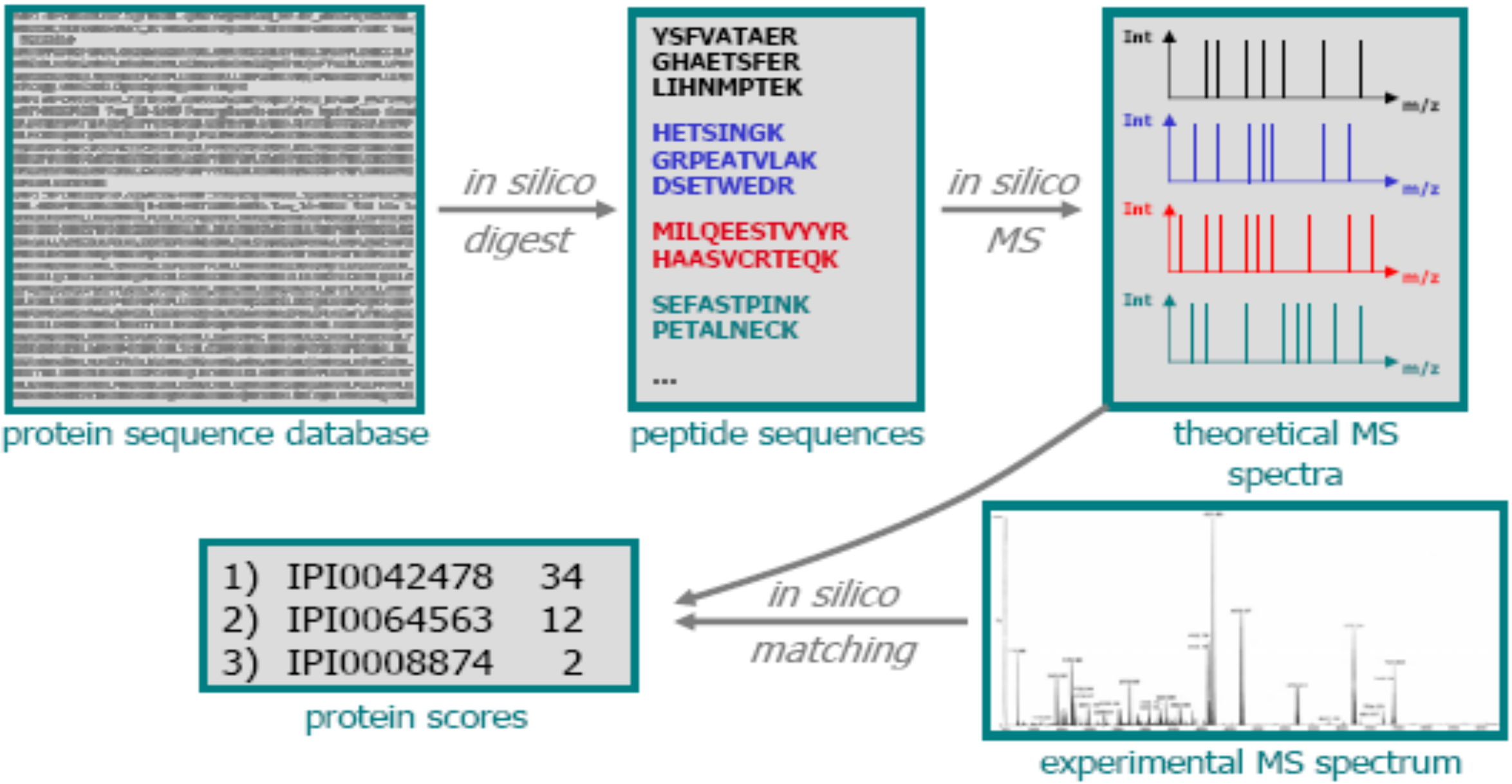
- Spectral comparison



- Sequence comparison



Peptide Fragment Fingerprinting (PFF)



Software

- MASCOT (http://www.matrixscience.com)
  - Predicts threshold score that needs to be passed
  - Provides rank, score and threshold
- SEQUEST (http://fields.scripps.edu/sequest)
  - User decides on threshold
  - Provides rank and score
- XTandem (http://www.thegpm.org/TANDEM)

Problems with peptide ID

- Does not give you the actual sequence
- Problematic when using an unsequenced genome
- Ambiguity with protein families
- False positive and false negative matches

Potential solutions

- Combining search algorithms
    - Diff programs have different strengths
    - All give some Fs and Ns
    - Run a combination of search engines then:
        - Union of results –extends identifications –fewer Ns
        - Intersection of results –stricter set of results –fewer Fs
- What is your research question?

Validation: Peptide- and ProteinProphet

- Validation tools:
    - PeptideProphet & ProteinProphet
    - Calculate the probability the ID is correct
- Use of decoy databases:
    - Three main types: reversed, shuffled and randomised
    - Use to calculate probability of identifications and FP rate
    - Reversed databases – reverse all sequences
        - e.g. RKLYWSML -> LMSWYLKR

Types of identification

- Spectral comparison

Database → Sequence → Theoretical spectrum ↔ Experimental spectrum

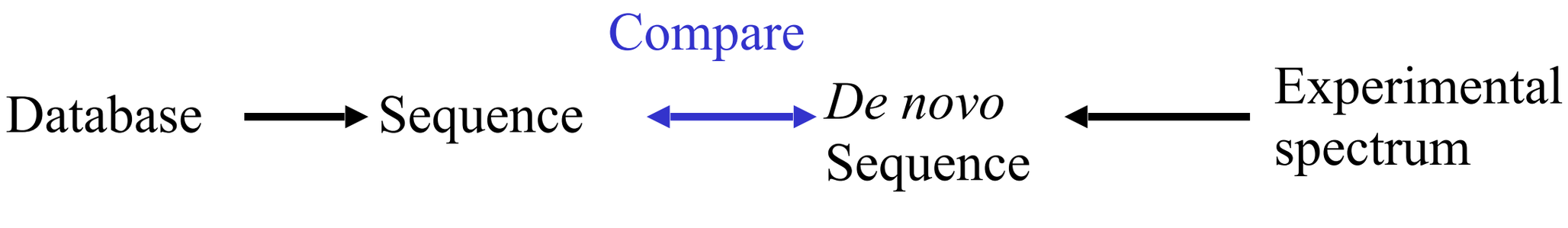
Compare

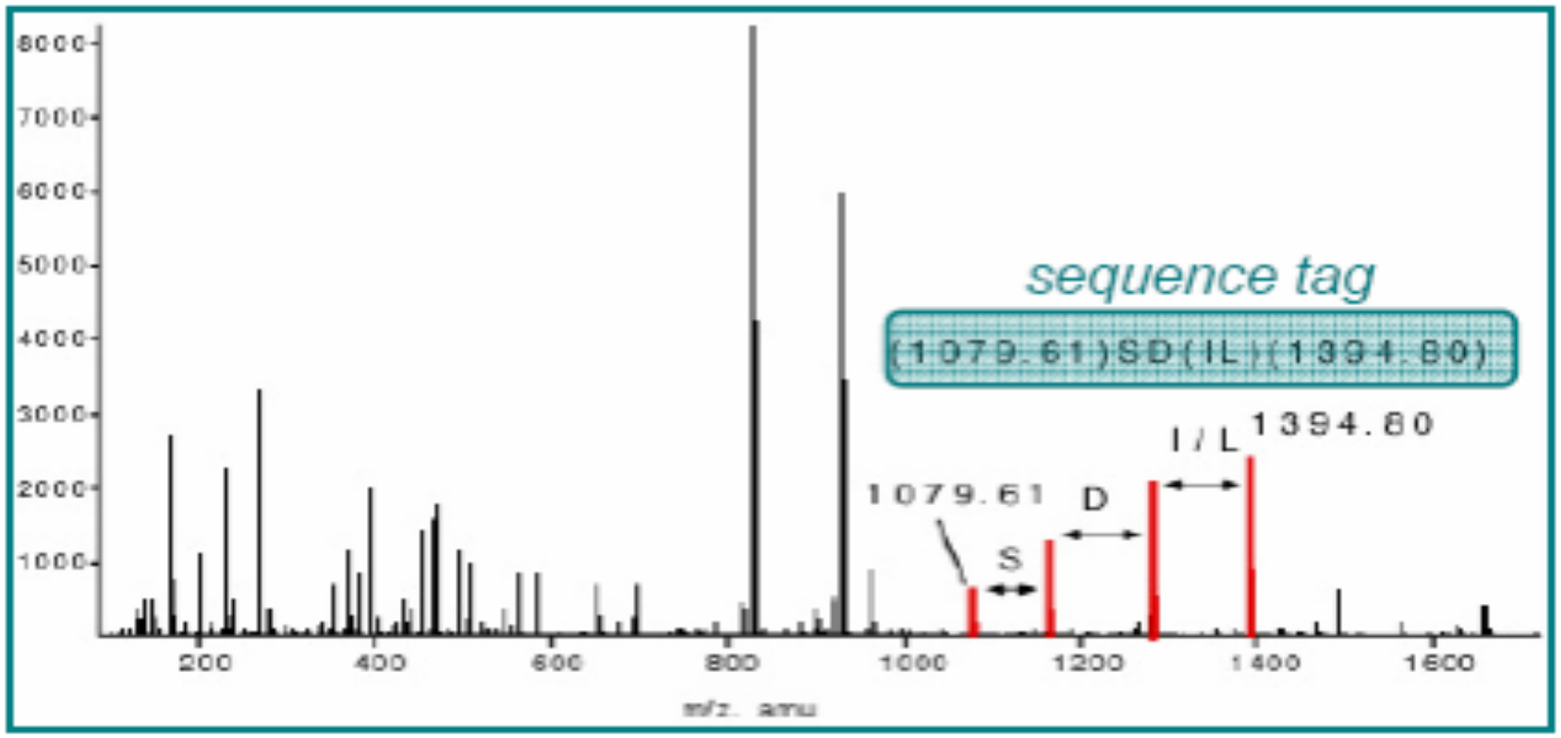
- Sequence comparison

Database → Sequence ↔ *De novo* Sequence ← Experimental spectrum

Compare

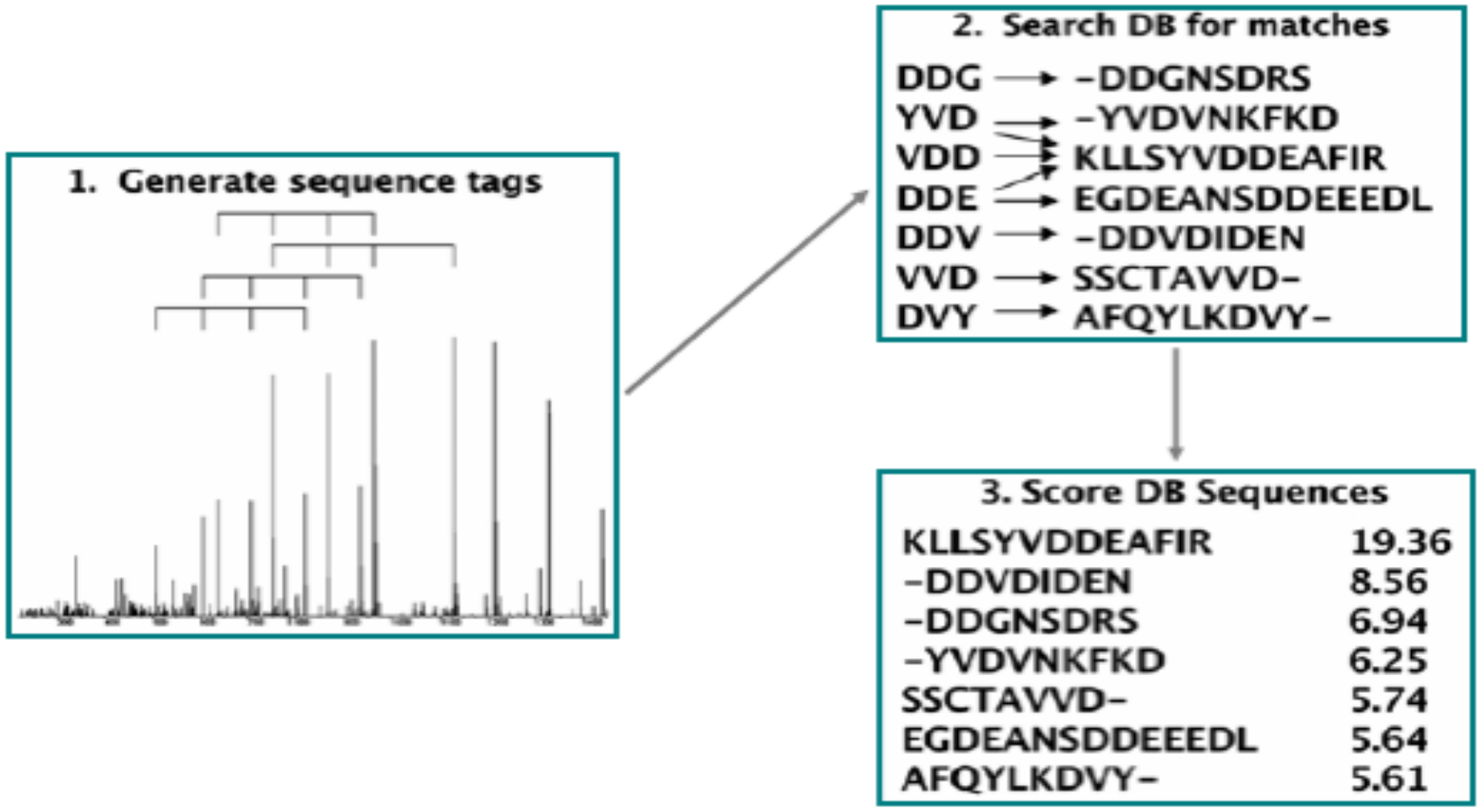
Sequencing with mass spec





Use known weights of amino acids

Sequencing with mass spec



From: Tabb et al., Anal. Chem., 2003

Summary

- Proteomics allows:
  - Identification of proteins
  - Comparison of protein levels between samples
- Requires:
  - Careful choice of databases
  - Careful selection of parameters
  - Selection of the right technique that will allow you to answer the research question