Modern mass spec based proteomics

(Because nucleic acids are overrated)

Presentation outline

- What is "proteomics" ?
- Historical overview over development of the technology
- Applications of proteomics
- Data processing and analysis
- Future perspectives

What is proteomics?

Dictionary definition:

- Proteomics is the systematic characterization of all the proteins in an organism, their abundance, localization, structure, modifications, function and interactions.

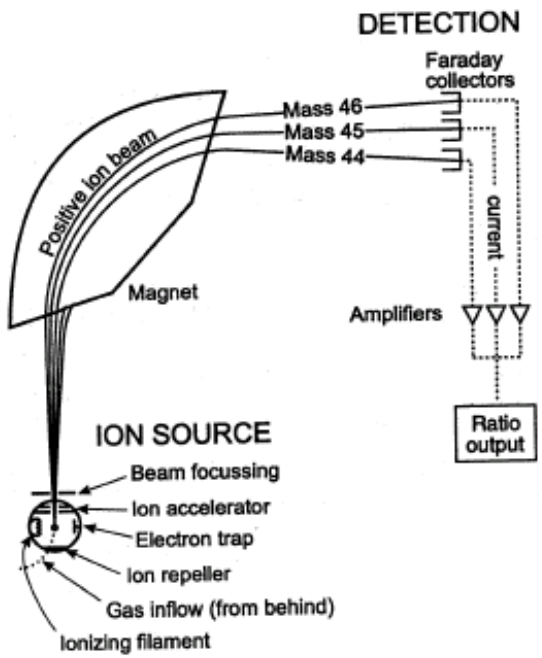
- Most researchers take a narrower view
  - Protein-protein interactions
  - Quantitative proteomics
  - Functional proteomics

- Various technogogies can be applied
  - Our focus: LC-MS/MS

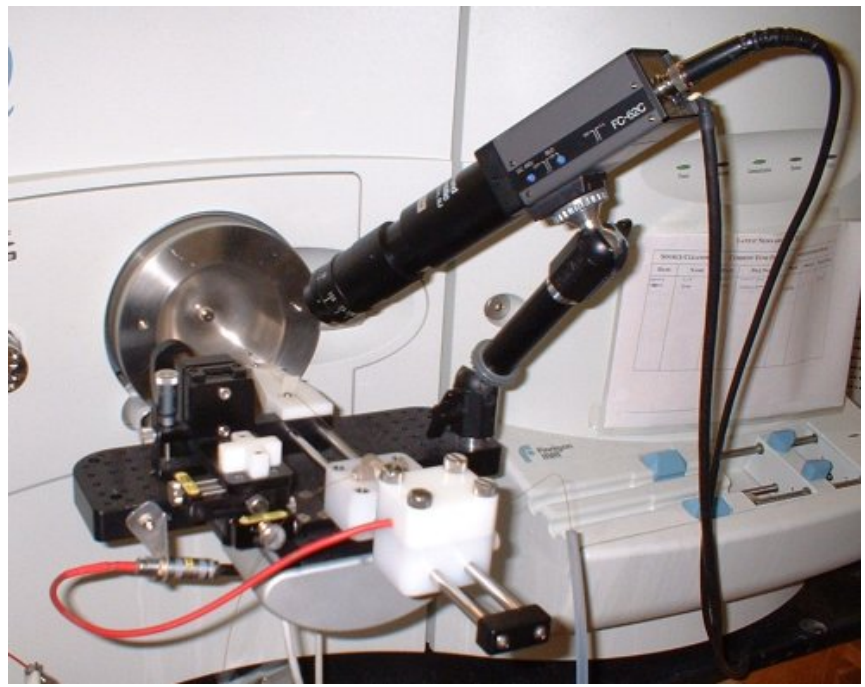
Development of the technology

(From the deflection of "canal rays" to MudPIT)

- Protein mass spectrometry
- Protein separation
- Data analysis



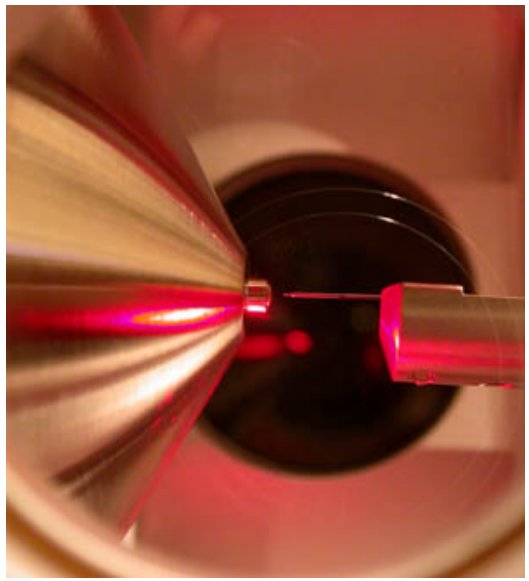
->



Protein mass spectrometry

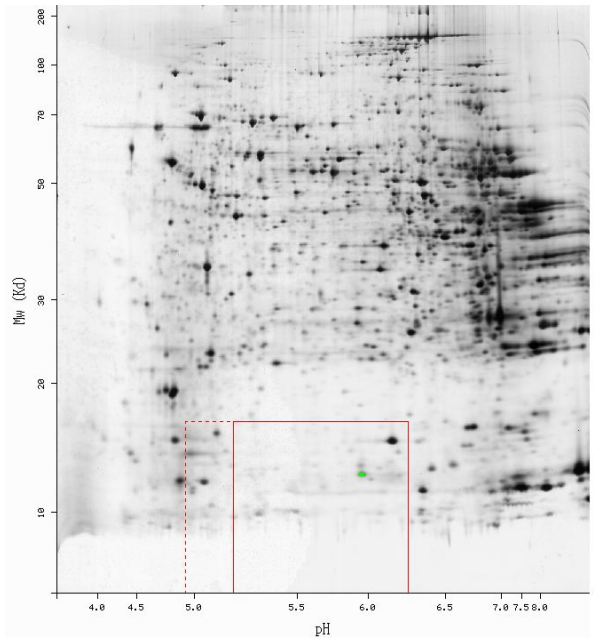
- Mass spec
  - Wilhelm Wien (Foundation), 1898
  - Sir Joseph Thomson (Neon isotopes) , 1913

- Beginning of protein mass spec
  - Problem of protein ionization
    - Koichi Tanaka (SLD), 1988
    - John Fenn (ESI), 1989

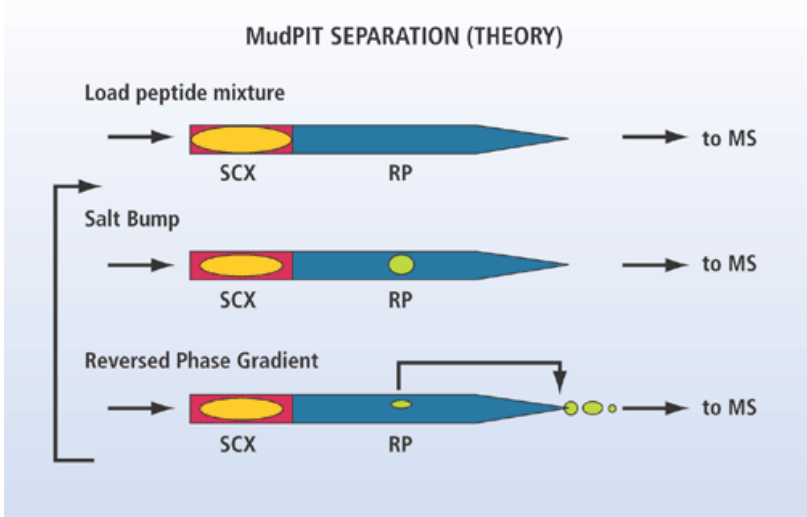


Protein separation

- 2D gel based approaches
  - low sensitivity (staining)
  - extensive sample handling
  - difficult to reproduce
  - no sympathy for the gel

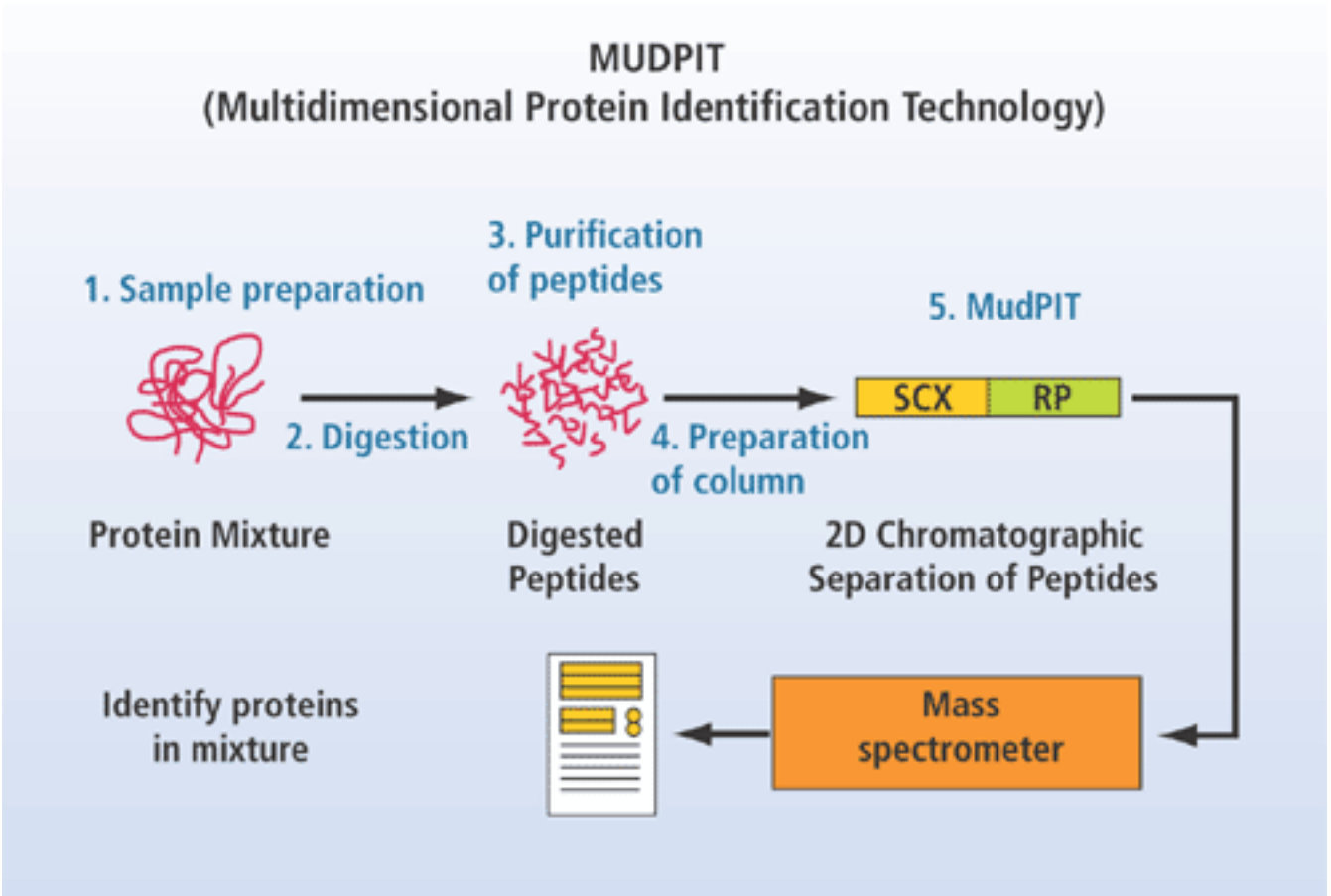


- Chromatography based approaces
  - Washburn *et al.* (MudPIT), 2001
    - on-line
    - semi quantitative
    - more sensitive
    - high throughput

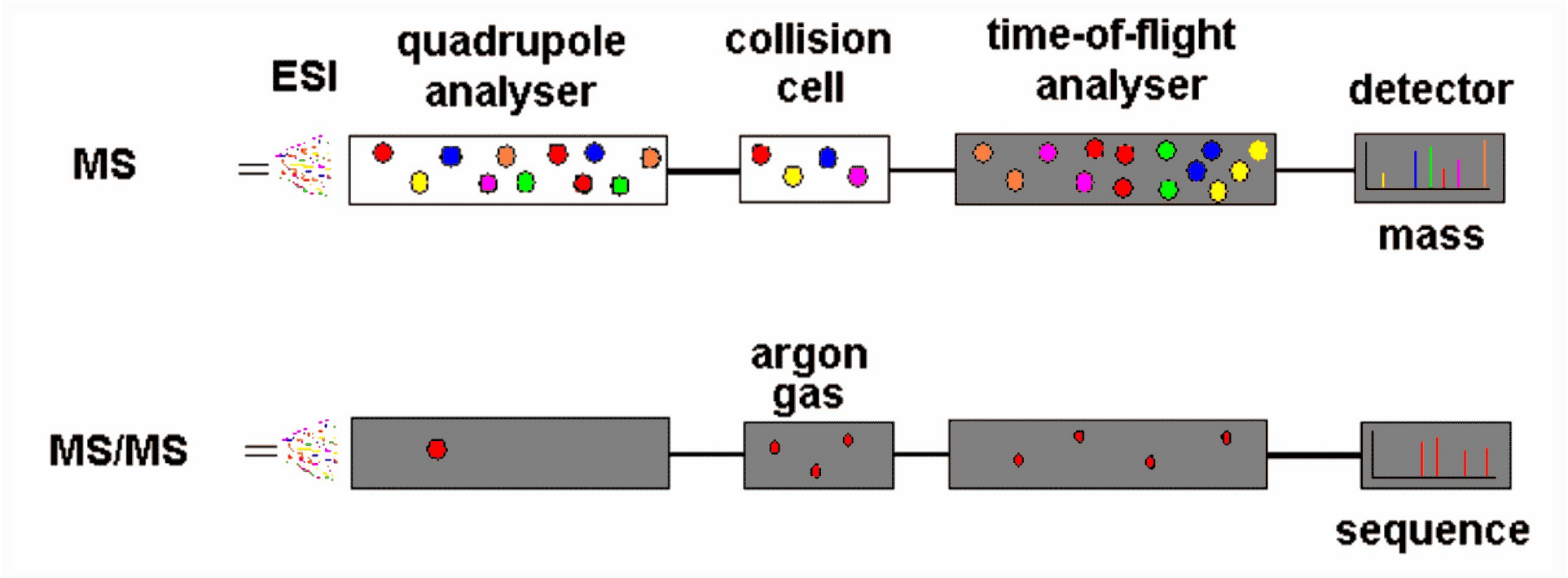


A state of the art setup

- MudPIT (multi-dimensional protein identification technology)
- Originally developed at Yates lab



Methodological background
Quadrupole-TOF (MS/MS)



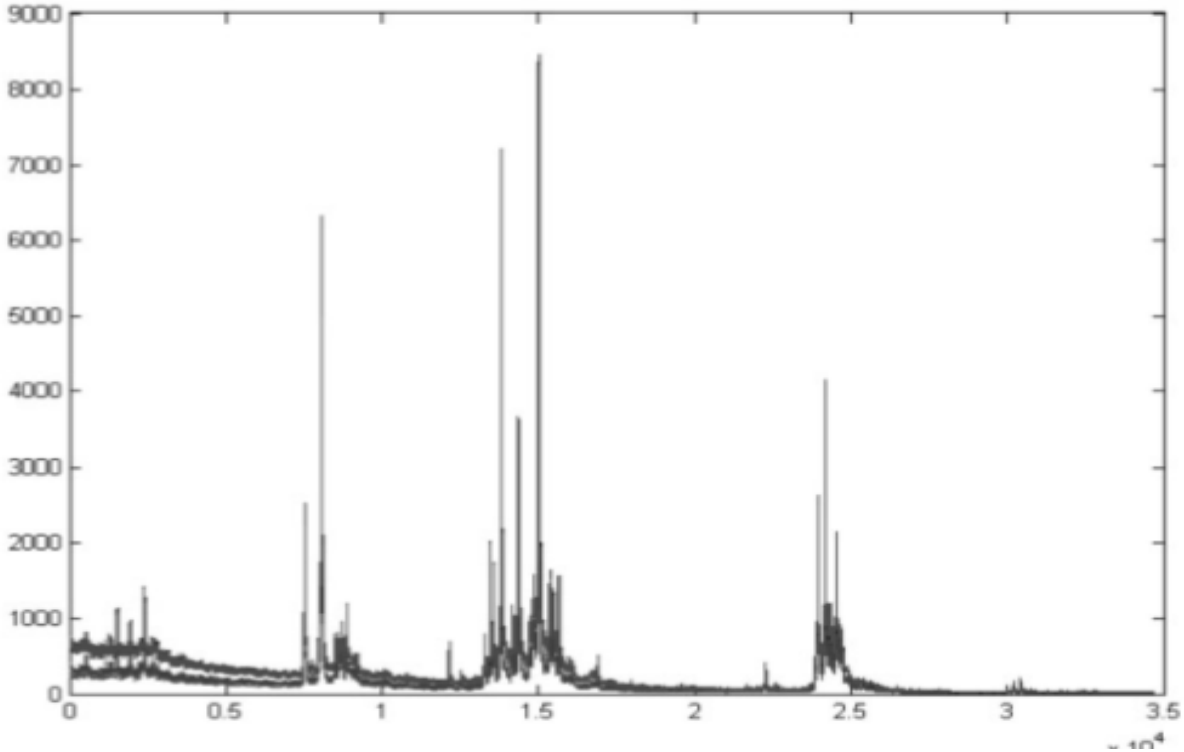
Operates on either MS or MS/MS mode

Data Analysis

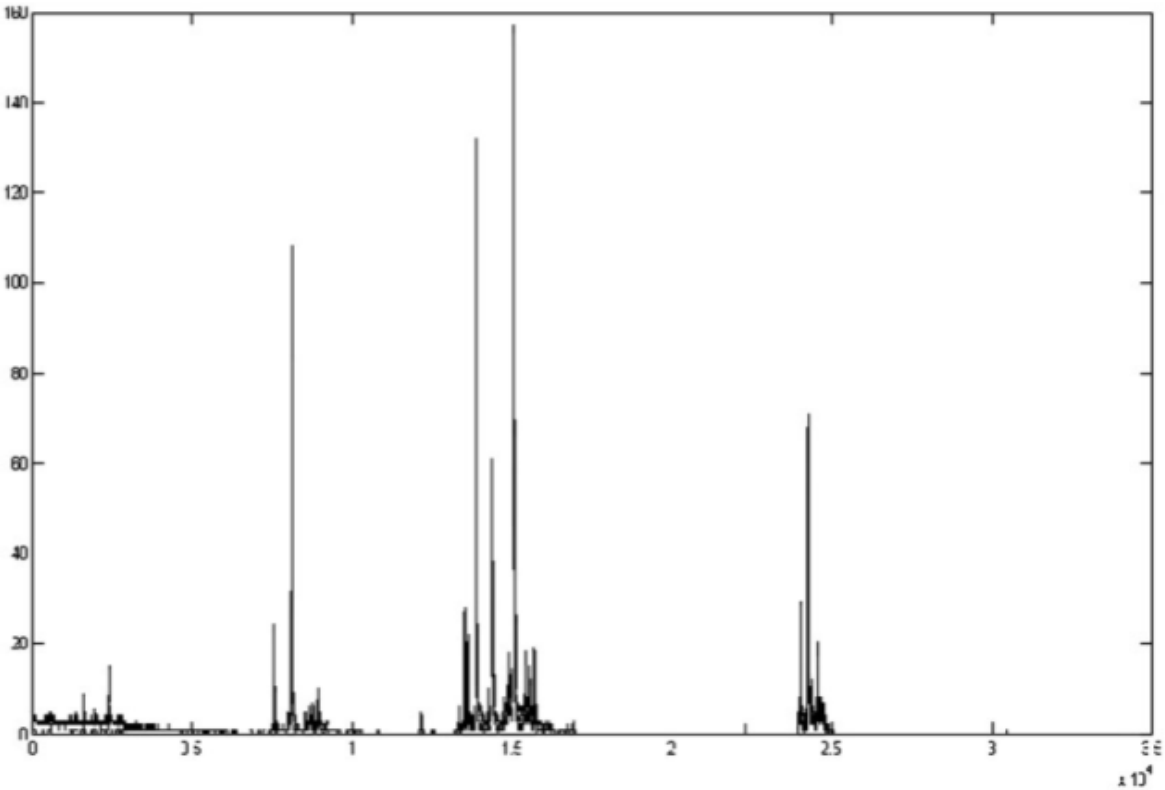
- Reducing raw data to manageable levels.
- Analysis
- Algorythms
- How to estimate the quality of data

Reducing raw data to manageable levels

- Preprocessing
  - Peak detection, peak labeling, baseline correction
  - Data reduction
    - noise removal, smoothing
  - Normalization
  - Deconvolution
    - Ion charge state recognition (isotope patterns)
  - Peak alignment



Before preprocessing

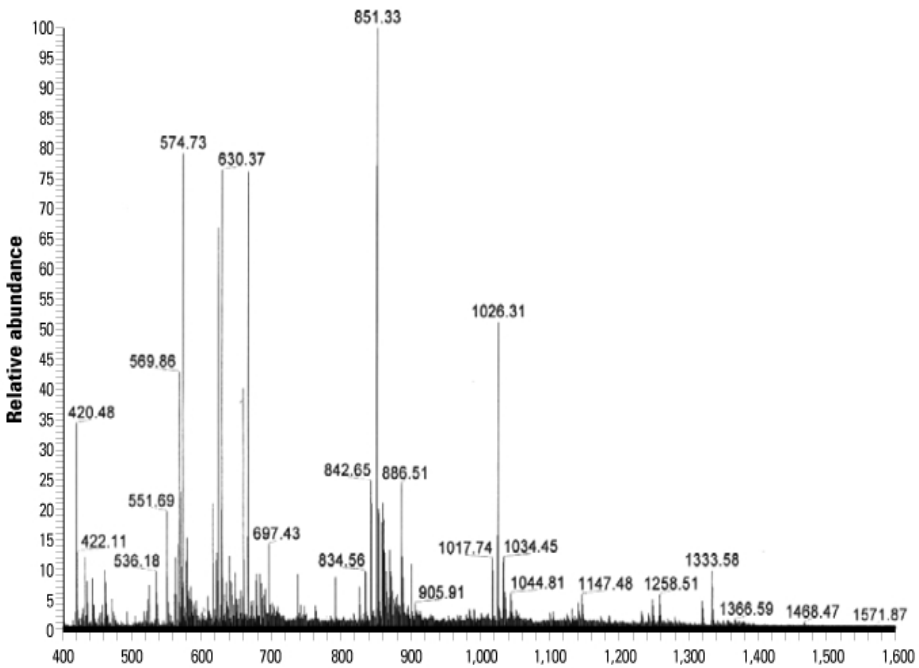


After preprocessing

Images from Veltri et al

Analysis

- Database search, Mann and Yates
- High throughput data
- High noise
- Computationally intense
- Variety of software



Algorithms

Examples:

- SEQUEST (Yates 1995)
- Mascot
- ProLuCID
- Specral network analysis (Bandeira 2007)

SEQUEST

Basic concept published by Yates et al. in 1995.

- Reverse pseudospectral library search.
- Protein sequences analysed sequentially through entire database.
- Preliminary scoring equation:

$$S_p = (\sum i_m) n_i (1 + \beta)(1 + \varrho)/\eta_\tau$$

- Cross correlation by Fourier transforming gives final score.
- Detects modified amino acids by testing alternative masses for all possible modification sites.
- Descriptive model.

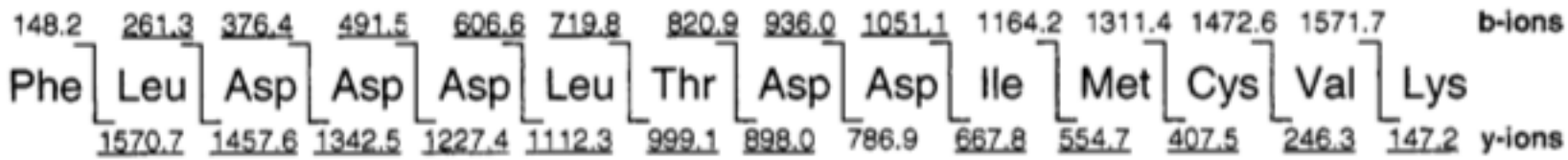


Figure 1. Schematic of the approach used by the computer algorithm to match tandem mass spectra of modified peptides to sequences in the protein database.

Mascot

- Incorporates a probability based implementation of Mowse, molecular weight search.
- Mowse assigns a statistical weight to each peptide match.
- Mowse factor matrix M:

$$m_{i,j} = \frac{f_{i,j}}{\left|f_{i,j}\right|_{\max \text{ in column } j}}$$

- Scoring equation:

$$\text{Score} = \frac{50{,}000}{M_{Prot} \times \prod_{n} \prod m_{i,j}}$$

- The total score is the absolute probability that the observed match is a random event.
- High score = low probability.
- Presented as -Log(P).
- Probability-based model.

http://www.matrixscience.com/help/scoring_help.html

ProLuCID

- Combines descriptive and probability-based models.
- Binomial probability preliminary scoring.
- Introduces a ProLuCID Z score.
- Algorithm description:
  - Candidate peptides selected from databases based on the precursor mass and peptide mass tolerance.
  - Binomial probability computed for each candidate:

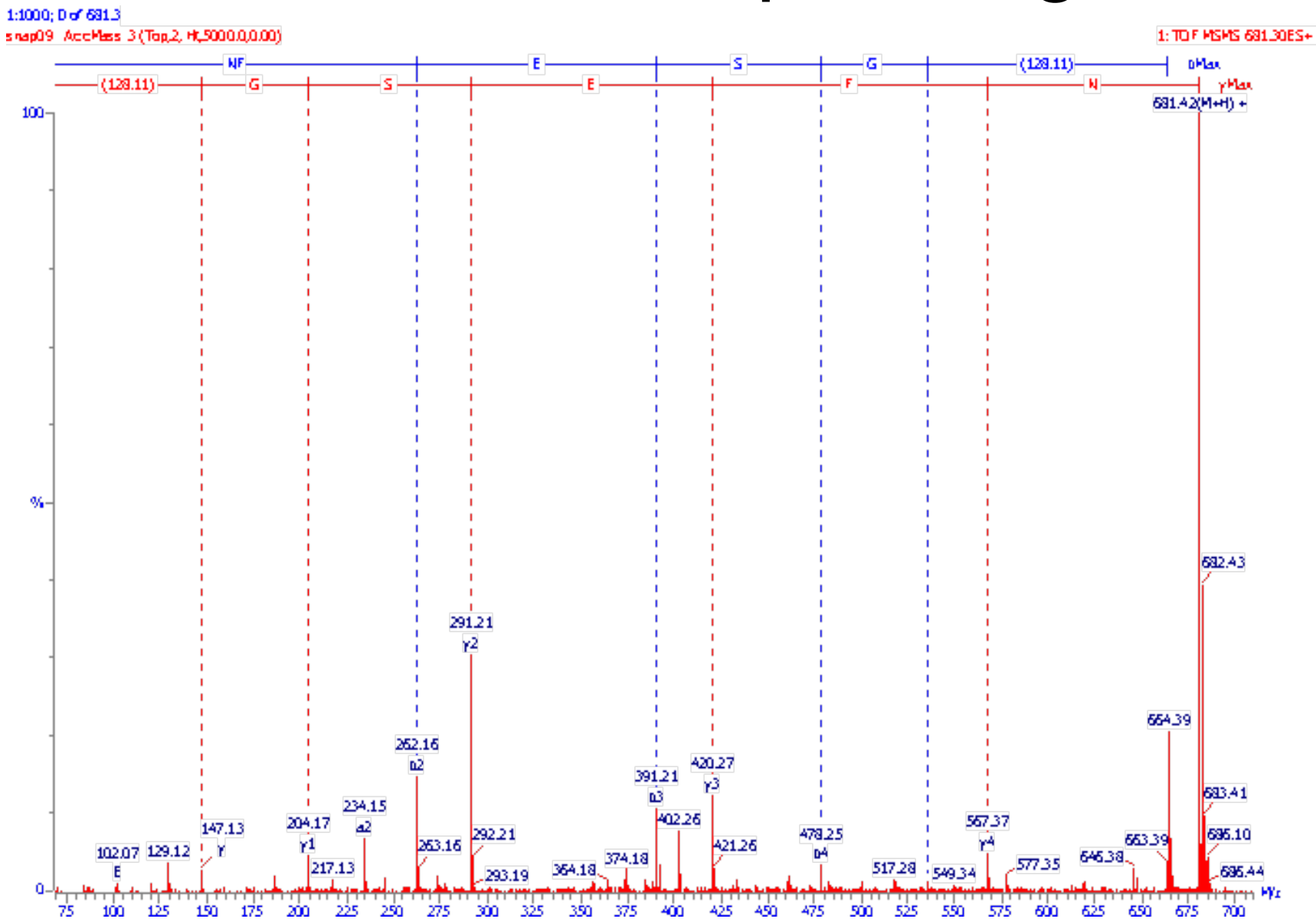
$$P(x>=m)=\sum_{i=m}^{n}P(x=k) \quad \text{where} \quad p(x=k)=\frac{n!}{k!(n-k)!}p^{k}(1-p)^{(n-k)}$$

  - XCorr computed with modified cross-correlation algorithm.
  - ProLuCID Z score computed:

$$ZScore=\frac{X-\mu}{SD} \quad \text{where} \quad SD=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}$$

Ref. Poster by Tao Xu et al.

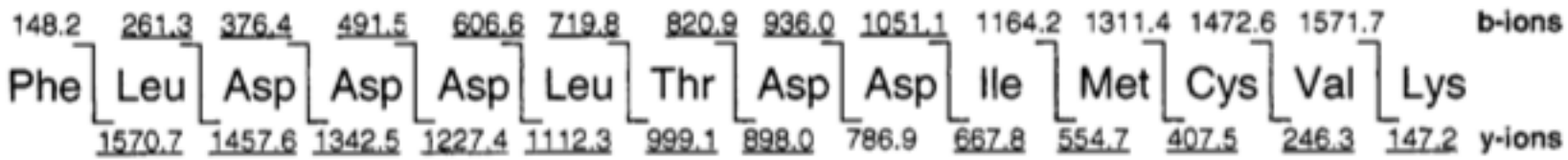
De novo sequencing

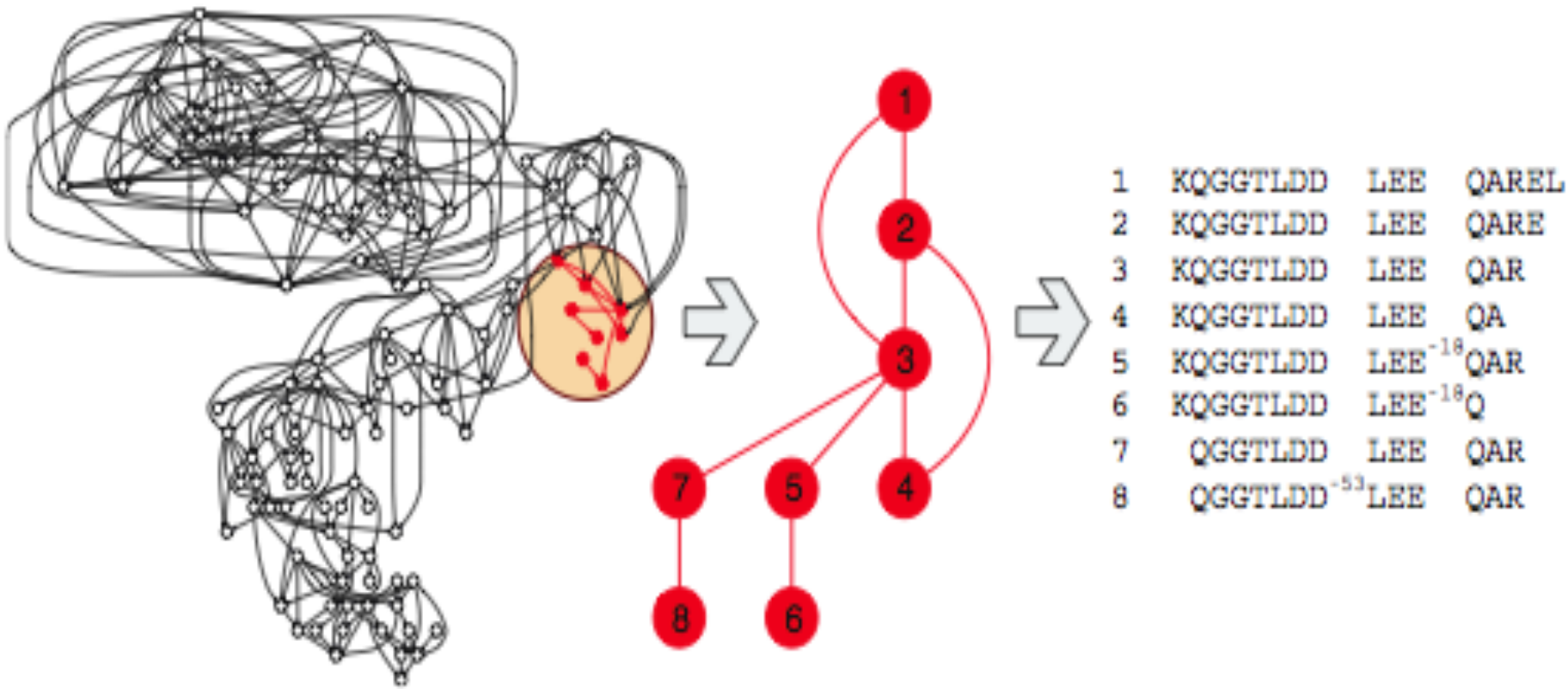


http://www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm

Spectral Network analysis

- Described by Bandeira et al. in 2007.
- Combination of *de novo* and spectral alignment techniques.
- Spectral pairs:
  - Overlapping peptides.
  - Modified vs. unmodified peptides.
- Spectral paires usually avoided due to higher running times.
- Generates covering sets of peptides 7-9 aa. long.
  - Most often a single hit in database.
  - Easily found using a hash function.
  - No need for a database comparison.
- Spectral networks.




1
2
3
7
5
4
8
6
1 KQGGTLDD LEE QAREL
2 KQGGTLDD LEE QARE
3 KQGGTLDD LEE QAR
4 KQGGTLDD LEE QA
5 KQGGTLDD LEE-18QAR
6 KQGGTLDD LEE-18Q
7 QGGTLDD LEE QAR
8 QGGTLDD-53LEE QAR

How to estimate quality of data?

- Compare to scrambled or reversed databases.
  - A peptide from the database is scrambled or reversed and compared to the spectral data.
  - Has the same aa ratios but different sequences.
  - Many scrambled or reversed hits means bad data.

Applications off protein mass spec

- Post translational modifications
- Protein interactions
- Disease genes and Biomarkers
- Stem cell characterization
- Alternative to microarrays
  - mRNA changes may not be physiologically relevant
  - mRNA may not be present in tissue of interest (blood)

Future perspectives

- Functional proteomics
- Quantitative proteomics
- Systems biology
  - Integration with other -omics datasets
- Standardization of protocols and analysis
  - Databases "ProteomeExpress"
  - The minimum information about a proteomics experiment (MIAPE)

Difficulties and bottlenecks

- Digestion (poor Km, few and inefficenient proteases)
- Peptide separation
- Masking by abundant proteins
  - Difficult to mass spec transcription factors and other low abundant proteins
- Not all peptides fly
- Isomer identification difficult
- There is hope
  - Field is young and moves fast
  - MudPIT setups are becoming commercially available
  - High demand (everybody wants so be friends with the mass spec guy)