Proteomics: Large-Scale Identification of Proteins

Ulf Leser

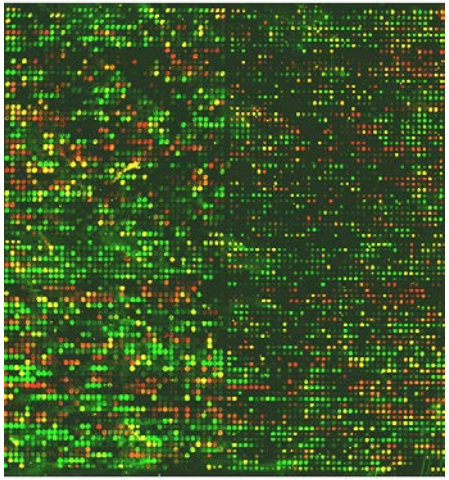
This Lecture

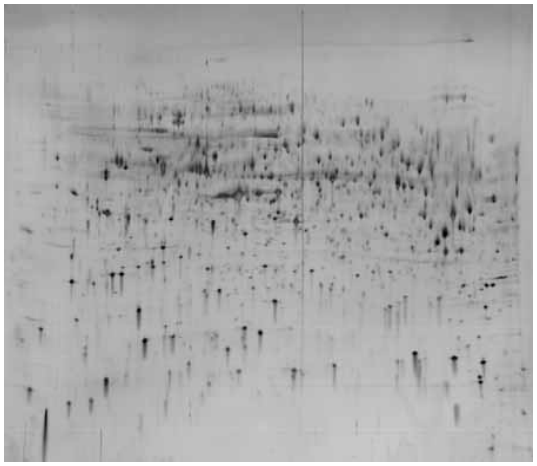
- Proteomics
- Separation: 2D Gels
- Identification: Mass Spectrometry

Proteomics

- Genomics =
  Determining the genome of a species
- Transcriptomics =
  Determining the mRNA of a cell / a tissue / a status
- Proteomics =
  Determining the proteins in a cell / a tissue / a status
- Proteomics and transcriptomics have mostly identical goals
  - Understanding the processes happening in a cell
  - Differentiate between species, tissues, developmental state, ...
  - Biomarker: Finding protein (forms, concentrations) that are characteristic for a certain diseases (state)
- Metabolomics, interactomics, bibliomics, cellomics, ...

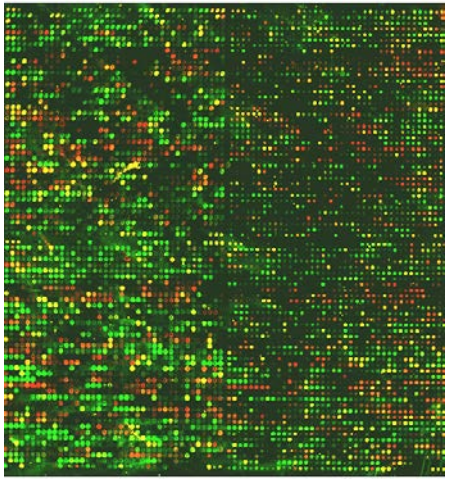
Proteomics (versus Transcriptomics)

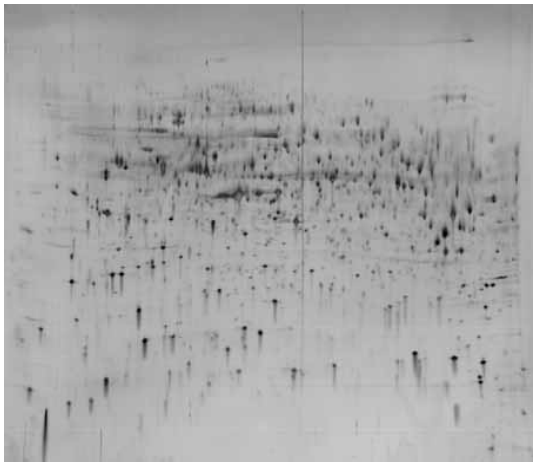




- Advantages
  - Proteins make you live, not mRNA
  - mRNA is only a indirect evidence with non-linear relationship
    - Regulation by miRNA, alternative splicing, ...
  - Protein survive (some time), mRNA is transient
  - Proteins may be drug targets

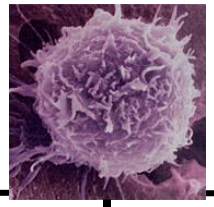
Proteomics versus Transcriptomics





- Disadvantages
  - Scale: 25K genes, 100K proteins, 500K protein forms
  - Handling: No PCR, no hybridization, no sequencing, no long-term „storage“ as clones, high reactivity with everything in contact, ...
  - Reactivity much more context-dependent: temperature, solution, pH, ...

Typical Proteomics Workflow



Extraction

Separation

Isolation

Identification

Analysis

From a cell mixture

2D gel electrophoresis / LC/GC

From the gel / from the flow

Mass spectrometry

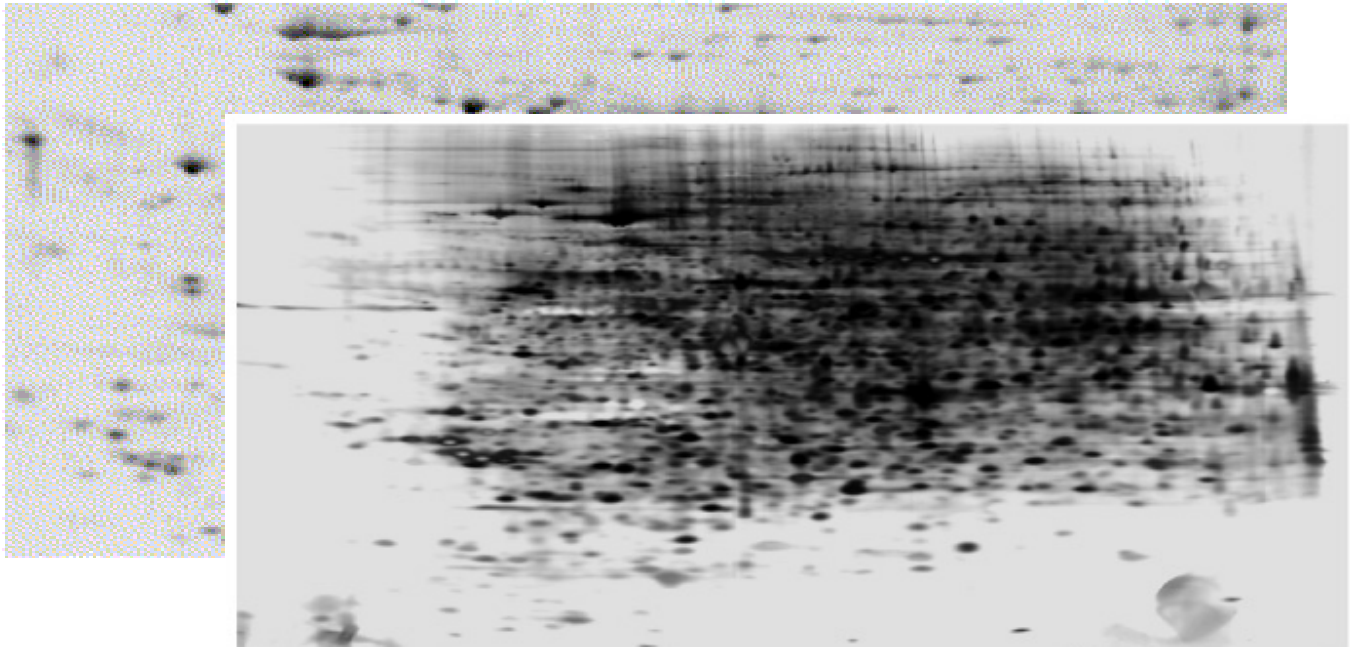
Quantification, clustering, ...

This Lecture

- Proteomics
- Separation: 2D Gels
- Identification: Mass Spectrometry

2D Gel Elektrophoresis

- Separation of proteins in two dimensions
  - Mass
  - Charge
- Every spot one protein (hopefully)



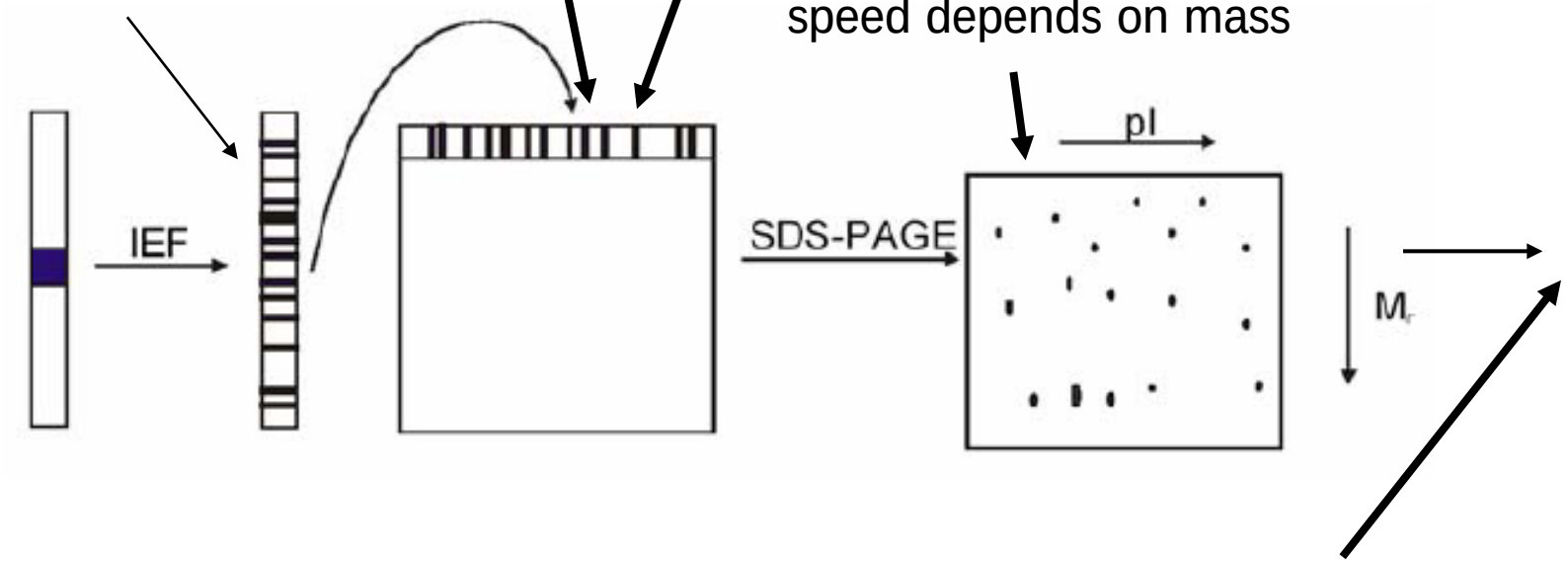
Method

2. Charging of proteins with SDS (Sodiumdodecylsulphate)

4. Place on slide of polyacrylamide gel (PAGE)

1. Separation in pH-gradient: Proteins move to their isoelectric points

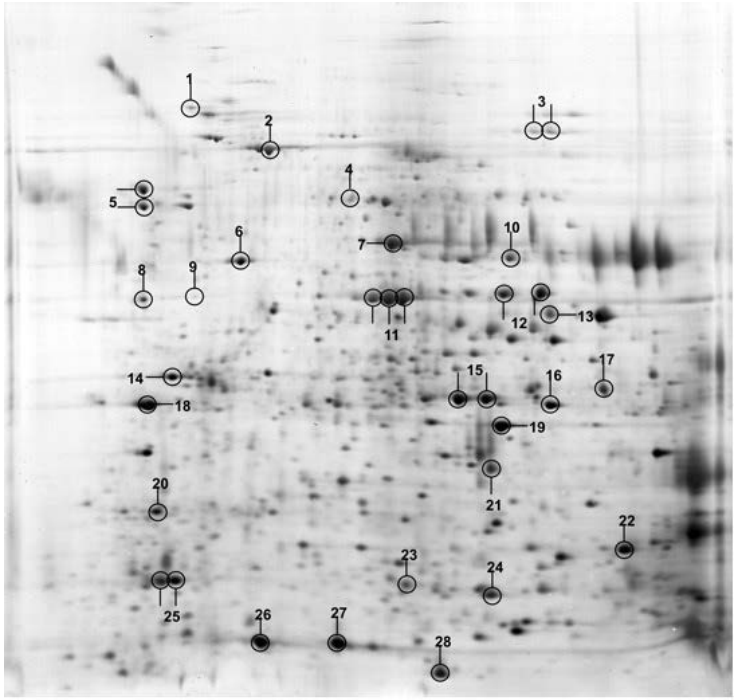
4. Proteins move in an electric field; speed depends on mass



5-. Staining; photo; image analysis; excision

Analysis

- 2D-Page may separate up to 10.000 proteins
- Under identical conditions, the position of a particular protein is fairly stable
- Software for identification of proteins by position
  - After photo and image analysis
  - Compared to a reference – how?
- Various databases of 2D-Gels
  - E.g. Swiss 2D-Page: Federation of 11 databases



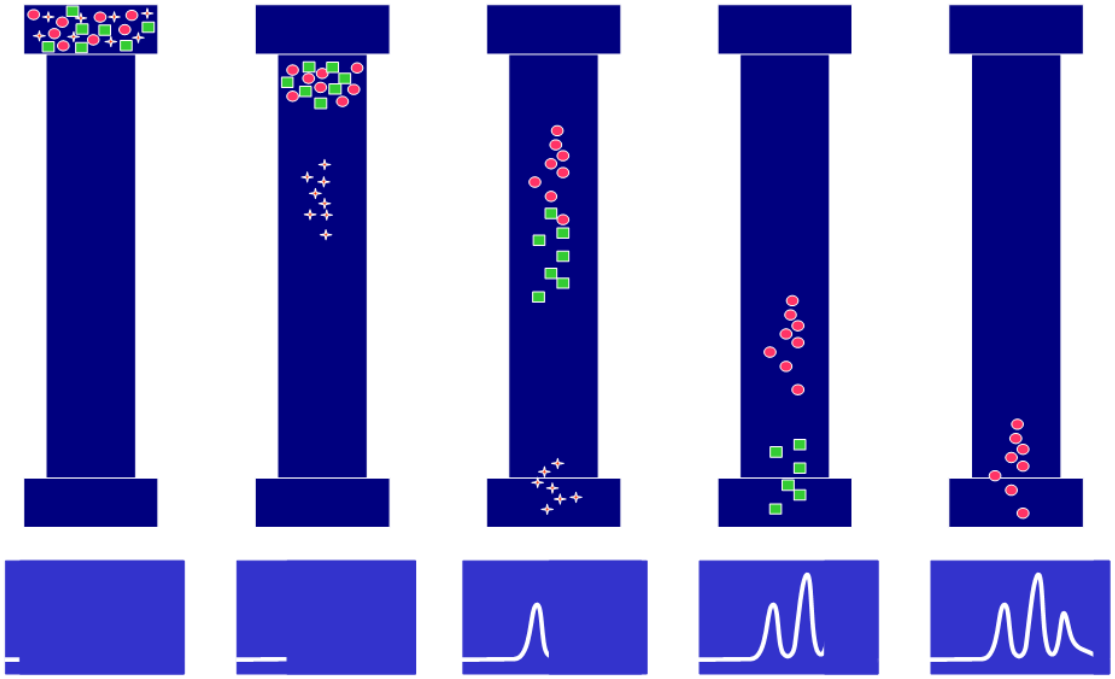
1 HSP86
2 HSP70
3 ATP:Guanidino Kinase
4 Adenylate Dehydrogenase
5 Calreticulin
6 Actin
7 Enolase
8 Tropomyosin
9 Serpin-like
10 Phosphoglycerate kinase
11 p40
12 Aldolase
13 GAPDH
14 14-3-3 e
15 GST28
16 Triose Phosphate Isomerase
17 Elongation Factor 1a
18 14-3-3 homolog 1
19 GST26
20 Calpain
21 Myosin Light Chain
22 Cyeophilin
23 Superoxide Dismutase
24 Fatty Acid Binding Protein (Sm14)
25 SME16
26 Thioredoxin
27 Dynein Light Chain
28 Ubiquitin
29 Adenylate Kinase

Pro / Contra

- Comparably simple and cheap method, well established
- Disadvantages
  - No high-throughput – much manual work
  - No robust quantification (spot intensity, depends on staining)
  - Similar proteins (e.g. protein forms) build highly overlapping spots
  - Many restrictions
    - No proteins with <20KD or >200KD
    - No highly charged proteins
    - No detection of low concentrations
    - No membrane proteins
  - No de-novo protein identification
  - Limited accuracy in comparative identification

Liquide / Gas Chromatography

- 2D-Page once often used as first step before MS-based identification
- Today: Mostly GC/LC

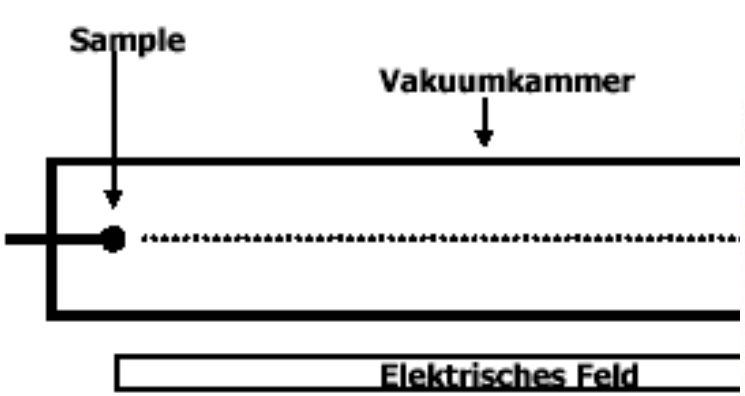


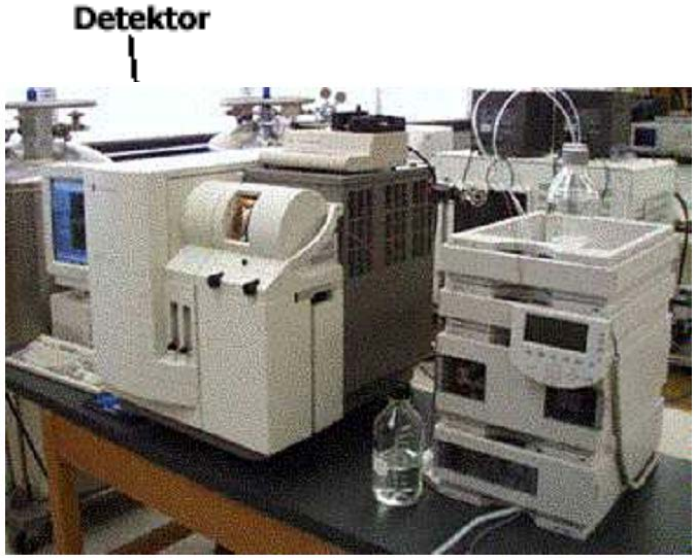
This Lecture

- Proteomics
- 2D Gels
- Mass Spectrometry
  - Method
  - Algorithms: Naïve, heuristic, probabilistic
  - De-Novo sequencing and quantification

Mass Spectrometry

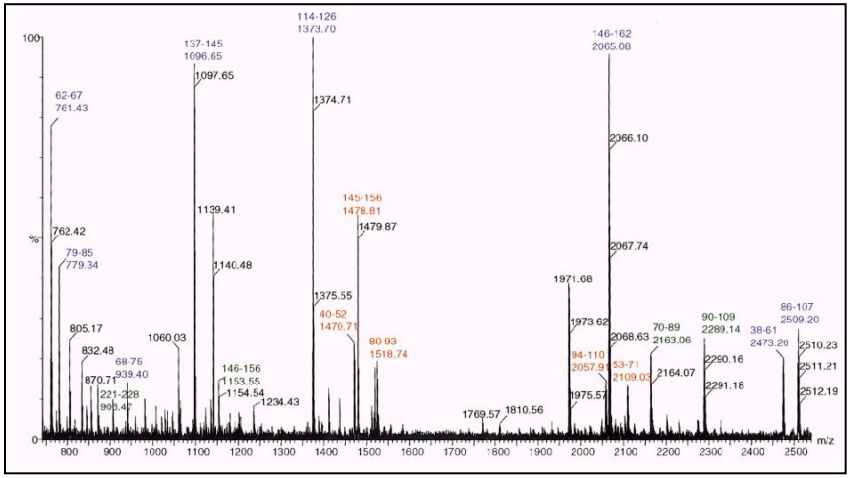
- Accelerate particles (must be charged) in an electric field
- Detector measures ion hits at back wall
- Flight time proportional to mass
  - ToF – other techniques exist (magnetic drift, ...)
- Spectrum of mass peaks is used to identify particle





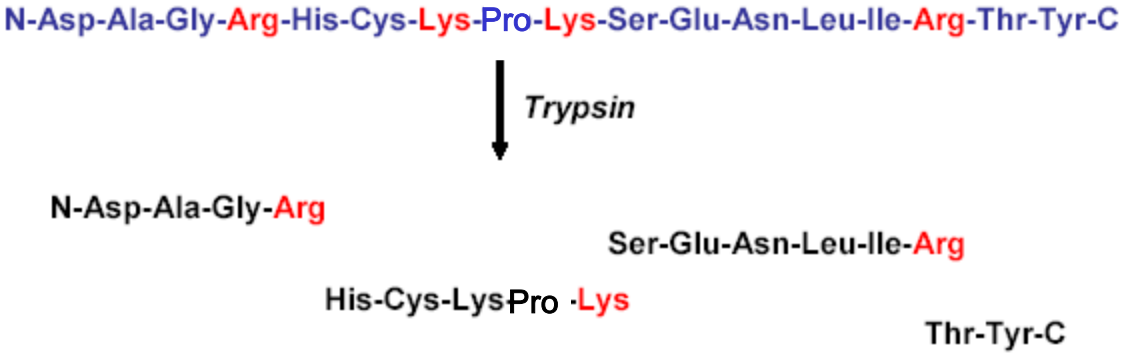
Using Proteins

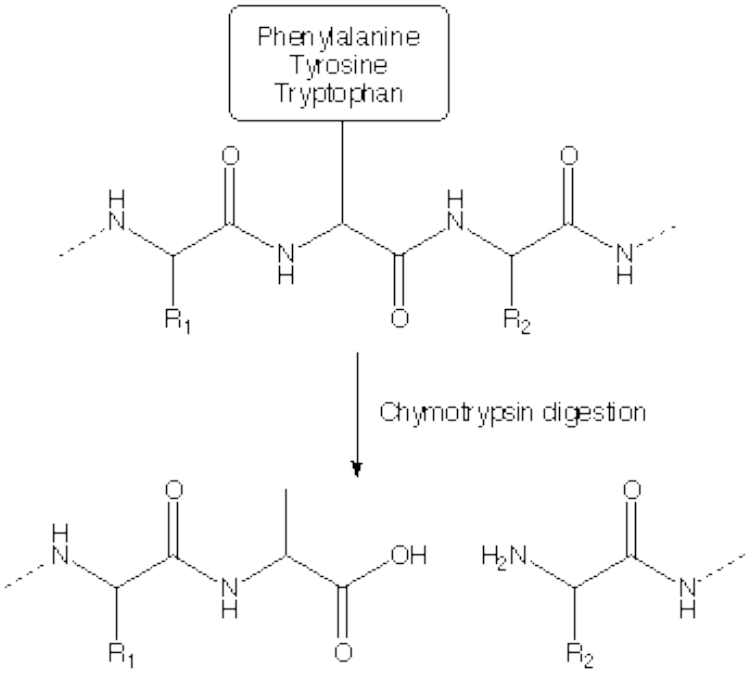
- Problem: Proteins are too fragile – they break
- Solution
  - Break proteins into peptides before acceleration
    - Enzymatic digestion
  - Measure peptides (each peptide one hit)
  - Identify protein based on spectrum of peptide hits
- In theory, every protein has an almost unique spectrum
  - Using modern MS/MS, even different protein forms are separable



Digestion

Trypsin:
Cleaves after Arginine und Lysine if next AA is not Proline





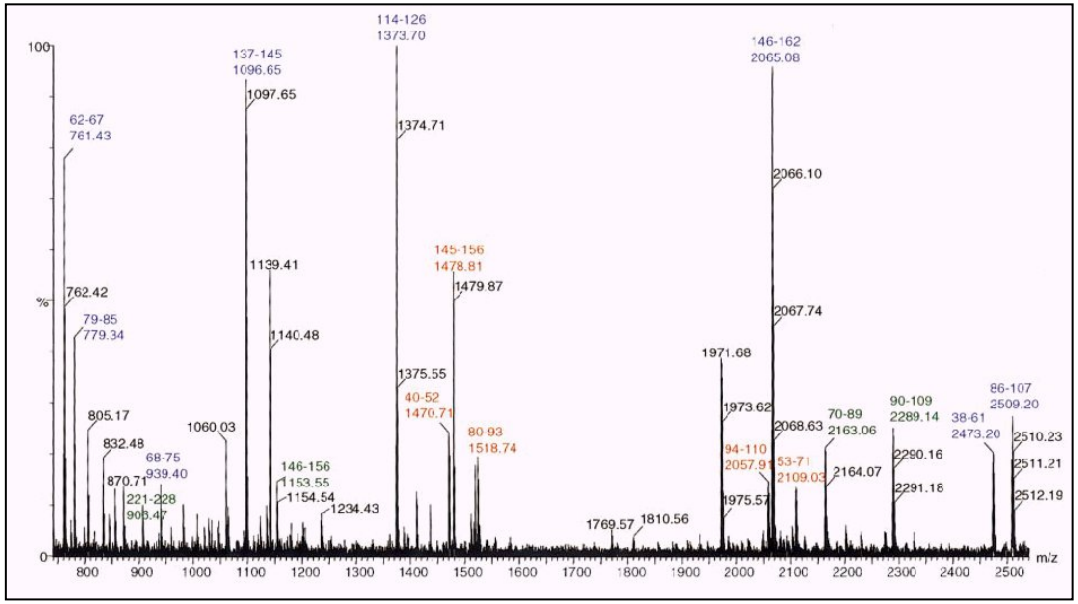
Chymotrypsin:
After Tyr, Trp, Phe, Met

Ionization

- Problem: Peptides often are uncharged – no acceleration
- Solution
  - MALDI – Matrix Assisted Laser Desorption / Ionization
  - Peptide are embedded in a „matrix“
    - Crystallization with charged, light-sensitive molecules
  - Fire on crystal with laser
  - Light-sensitive molecules vaporize and carry peptides with them
  - Accelerate
- Other techniques known (ESI: electrospray ionization)

From Measurement to Peaks

- Detecting peaks and assigning them to peptides is difficult
  - Systematic bias in runs / machines
  - Noise
  - Inaccuracies of measures
  - Inhomogeneous sample preparation
  - Different quantities of peptides
  - ...



- Signal processing – not covered here

This Lecture

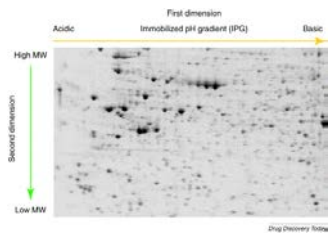
- Proteomics
- Separation: 2D Gels
- Identification: Mass Spectrometry
  - Method
  - Algorithms: Naïve, heuristic, probabilistic
  - De-Novo sequencing

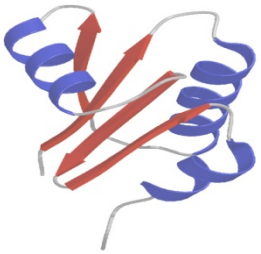
Algorithms for Protein Identification from Spectra

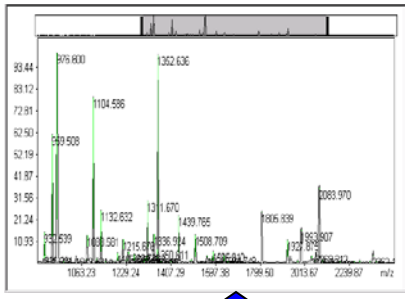
- We focus on database-based identification
- Idea
  - We have a database of protein sequences
    - Each is subjected to electronic digestion – set of peptides per protein
    - For each peptide, we know its theoretic flight time
    - One theoretic spectrum per protein in the database
  - Measure spectrum of unknown protein
  - Compare spectra
- Again, we can only discover what we already know
  - No novel proteins

Illustration

Real experiment





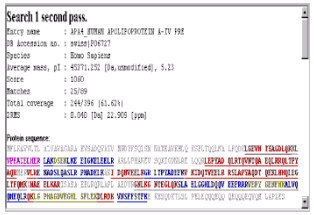


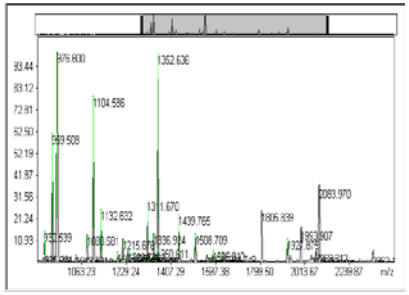
Comparison

Theoretical experiment

ttctcatcaaagaaatcaaagggaagc
ttattaccgagtggagttcaaggtcttgg
caaaaggtttccaaagaggatcatgttg
tcaaaggattggaatccatttcttctaag
aaacaattcgaggatcttgctggatcta

ANSHTDDLDELLDSALDDFKDLNLSHQ
QGLGMGLPDMRSKKRGKQKVSKEDHV
LPASDDDGMVEDFLKQFEDLAGSKDLE
RYPKWLKENEASLSEDYKRYSQQYKLI
ECGQPPSDIVKEIDPGFDFASLGQISPEM





Naive Algorithm: Hitcount

- Compare peptides of measurement P with all S_i in DB
- Sequence which has the most peptides in common wins
- Algorithm
  - Input: $P=\{p_1,...p_m\}$, $S_i=\{p_{i1},...p_{im(i)}\}$, $i < n$
  - Compute an array A storing for each peptide k all sequences containing it: $A[k] = \{S_i \mid k \in S_i\}$
  - Initialize a counter for sequence $M[i] = 0$, $i < n$
  - For all $k \in P$, for all i: If $S_i \in A[k]$: $M[i] = M[i] + 1$
  - Sequence S_i with $M[i]=MAX$ wins
- Complexity?
  - Theoretical worst-case $O(|P|*n)$
  - Average-case is $O(|P|)$

Example

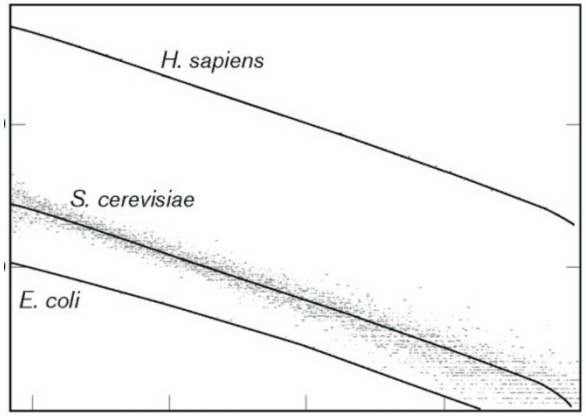
- Input
  - S_1 = [5,8,9,14,18]
  - S_2 = [3,5,9,12]
  - S_3 = [4,8,16,17,20]
  - S_4 = [1,7,9,17]
  - P = [7,8,14,16,17]
- A

1	3	4	5	7	8	9	12	14	16	17	18	20
4	2	3	1,2	4	1,3	1,2,4	2	1	3	3,4	1	3

- Score
  - $sim(S_1,P) = 1\ (8) + 1\ (14)$ 2
  - $sim(S_2,P) = 0$ 0
  - $sim(S_3,P) = 1\ (8) + 1\ (16) + 1\ (17)$ 3
  - $sim(S_4,P) = 1\ (7) + 1\ (17)$ 2

Why “Naïve”?

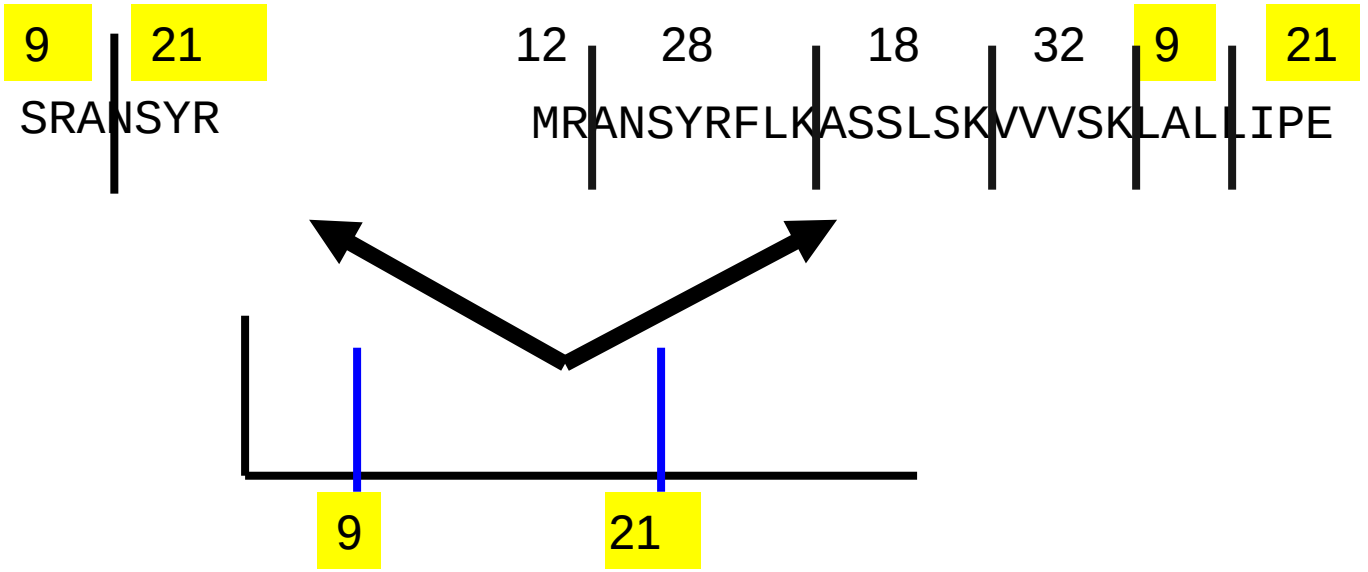
- Peptide masses are not really equal
  - Always small deviation – nearest hit – need not be unique
- Some (short) peptides are more frequent than others
  - Some peptides appear in almost all proteins
  - Should have a lower impact
- Proteins have different lengths
  - Longer proteins have an a-priori higher chance for high scores



X: Peptide mass (1000-5000 Dalton)

Y: Peptide count (log)

Example



- Which one would you prefer?

More Problems

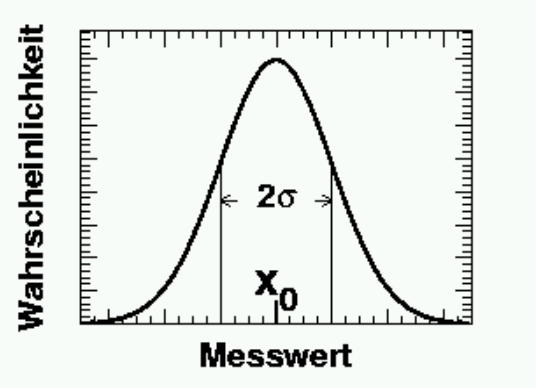
- Enzymes don‘t work 100% correct
  - Some peptides that should be there are missing, others that should not be there are present
- Protein sequences in DB contain errors
  - Especially when directly translated from genome
  - Especially bad when frameshifts occur
- Ignores posttranslational modifications
- Peptide mass not constant – isotopes
- MS is not perfect – spurious hits, shifted hits, missing hits
- Some protein always has the highest count – what if real sequence is not in the database?
  - No confidence scores

Practically Relevant Algorithms

- Heuristic: MOWSE
  - Considers total protein mass and peptide frequencies
  - Generates a score (but not a confidence)
- Probabilistic algorithm: Profound
  - Bayes' statistics
  - Can cope with measurement errors, protein mass and peptide frequencies
  - Generates a probability of match for each protein
- Many more (and newer) algorithms have been published
  - MASCOT, PeptIdent, ProteinProspector, SEQAN, ...

ProFound

- Zhang, W. and Chait, B. T. (2000). "ProFound: an expert system for protein identification using mass spectrometric peptide mapping information." *Anal Chem 72(11): 2482-9.*
- Probabilistic method
- Computes, for a given spectrum D (P) and each protein k (S_i), the probability that D was produced by k
- The formula is complex; its derivation is even more complex and skipped here
- Assumption: Measured peptide masses are normally distributed around the “canonical” value
  - Most probable isotopes



ProFound Formula

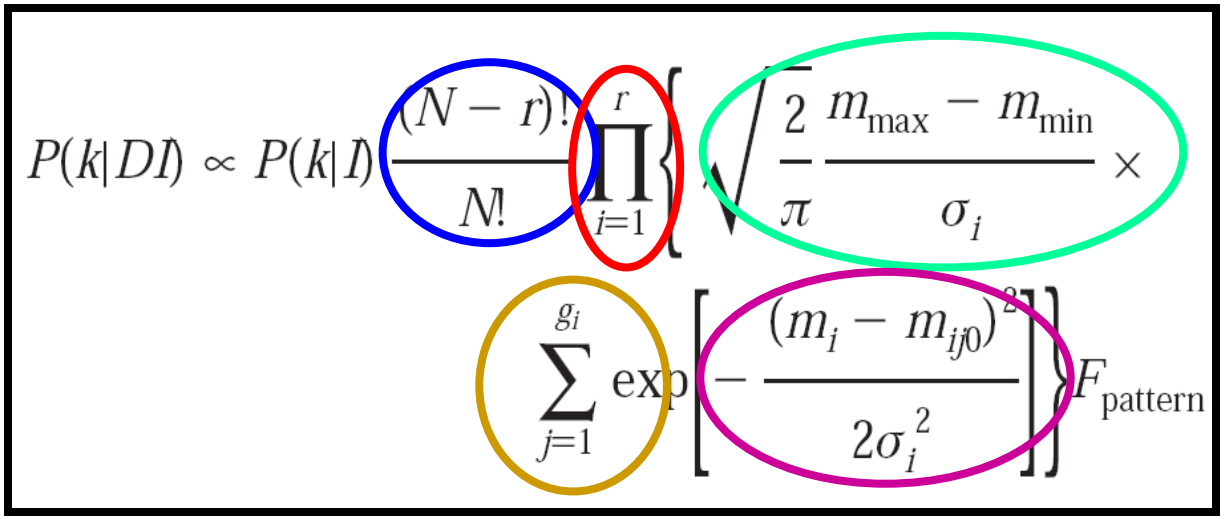
$$P(k|DI) \propto P(k|I)\,\frac{(N-r)!}{N!}\prod_{i=1}^{r}\left\{\sqrt{\frac{2}{\pi}}\,\frac{m_{\max}-m_{\min}}{\sigma_i}\times \sum_{j=1}^{g_i}\exp\left[-\frac{(m_i-m_{ij0})^2}{2\sigma_i^{\,2}}\right]\right\}F_{\text{pattern}}$$

Legend

$$P(k|DI) \propto P(k|I)\,\frac{(N-r)!}{N!}\prod_{i=1}^{r}\left\{\sqrt{\frac{2}{\pi}}\,\frac{m_{\max}-m_{\min}}{\sigma_i}\times \sum_{j=1}^{g_i}\exp\left[-\frac{(m_i-m_{ij0})^2}{2\sigma_i^2}\right]\right\}F_{\text{pattern}}$$

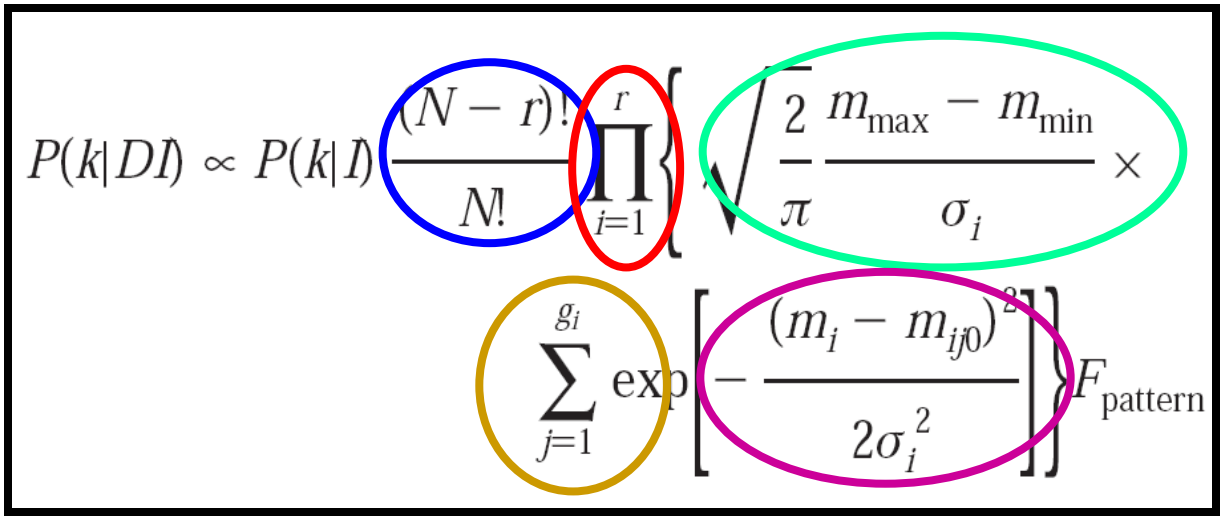
- p(k|D,I) = prob. that protein k was measured given the spectrum D and background information I
- N: Number of peptides of protein k
- r: Number of hits (with a certain fuzzy'ness)
- m_{max}, m_{min} – range of observed masses for current hit
- σ_i – standard deviation of i'th hit
- m_i: mass of the hit (must be between m_{max} and m_{min})
- g_i: How often is the peptide contained in k?
- m_{ij}: Theoretical mass of j'th occurrence of this peptide in k
- p(k|I): A-priori probability of k in the given species / cell / tissue
- $F_{pattern}$: Heuristic factor dealing with "overlapping peaks"

ProFound Explanation



- How many of the expected peptides for k did we observe?
- Multiply probabilities of all hits
- "Freedom" of measurements of hits for this peptide
- One observed peak may stem from various predicted peaks, each with slightly different mean m_{ij0}
- Probability of the deviation of the canonical mass to the measured mass

ProFound Intuition



- Many hits (r ~ N) – score goes up (outweighs influence of more small factors in the red part)
- Hits have in narrow range – score goes up
- Observed peak matches many theoretical peaks – score goes up
- Observed peak close to canonical peak – score goes up
- Theoretical peak as high stddev – scores go down (also green)

Critique

- Score assumes that protein is in the database
  - Better: formulate „null“ hypothesis, compute prob. of the spectrum given the null hypothesis, and report the log-odds ratio as score
  - But this is not as simple done as spoken out
- Assumes that every peak comes from “the” protein
  - But measurements might be contaminated with peptides from other proteins
- Assumes that observed peaks can be assigned clearly to theoretical peaks
  - This problem is tried to be covered by $F_{pattern}$
- Many more suggestions since 2000

Further Reading

- Basics on proteomics: Every Bioinformatics book
- Spectrum-analysis algorithms: Original papers
- Survey: Colinge J, Bennett KL (2007) Introduction to Computational Proteomics. PLoS Comput Biol 3(7): e114