

Proteomics: Large-Scale Identification of Proteins

Ulf Leser

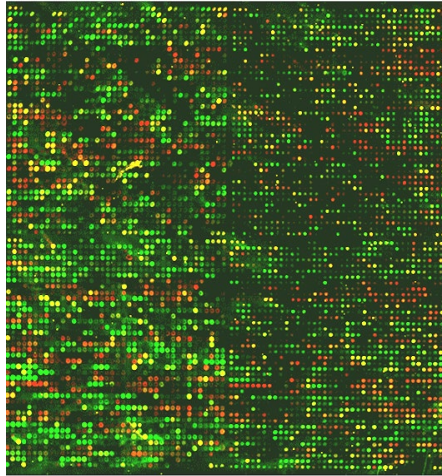
This Lecture

- Proteomics
- Separation
- Identification: Mass Spectrometry

Proteomics

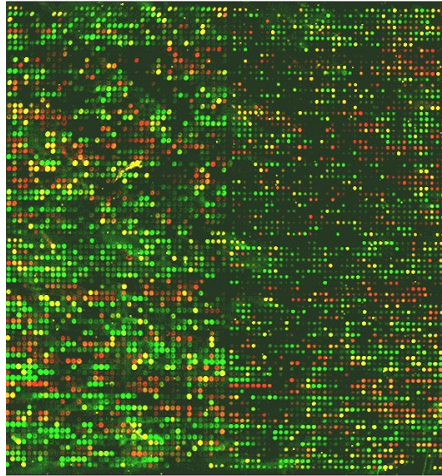
- **Genomics** =
Determining the genome of a species
- **Transcriptomics** =
Determining the mRNA of a cell / tissue / state
- **Proteomics** =
Determining the proteins in a cell / tissue / state
- Proteomics and transcriptomics have mostly identical goals
 - **Understanding** the processes happening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

Proteomics versus Transcriptomics

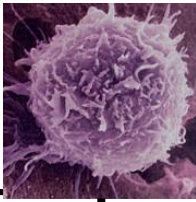


- Advantages
 - Proteins make you live, not mRNA
 - mRNA is only **indirect evidence** with little correlation with proteome
 - Regulation by miRNA, post-translation modifications, decay, ...
 - Protein survive (some time), mRNA is (mostly) transient
 - Proteins are favorite **drug targets**

Proteomics versus Transcriptomics



- Disadvantages
 - Scale: ~20K genes, ~500K proteins, >1M protein forms
 - Handling: No PCR, no hybridization, no simple synthesis, no fast sequencing, no long-term „storage“ as clones, high reactivity, ...
 - Behavior highly context-dependent: Temperature, solution, pH, ...



Typical Proteomics Workflow

Proteome Extraction

From a cell mixture

Protein Separation

2D gel electrophoresis / LC/GC

Sample Isolation

From the gel / from the flow

Protein Identification

Mass spectrometry

Analysis

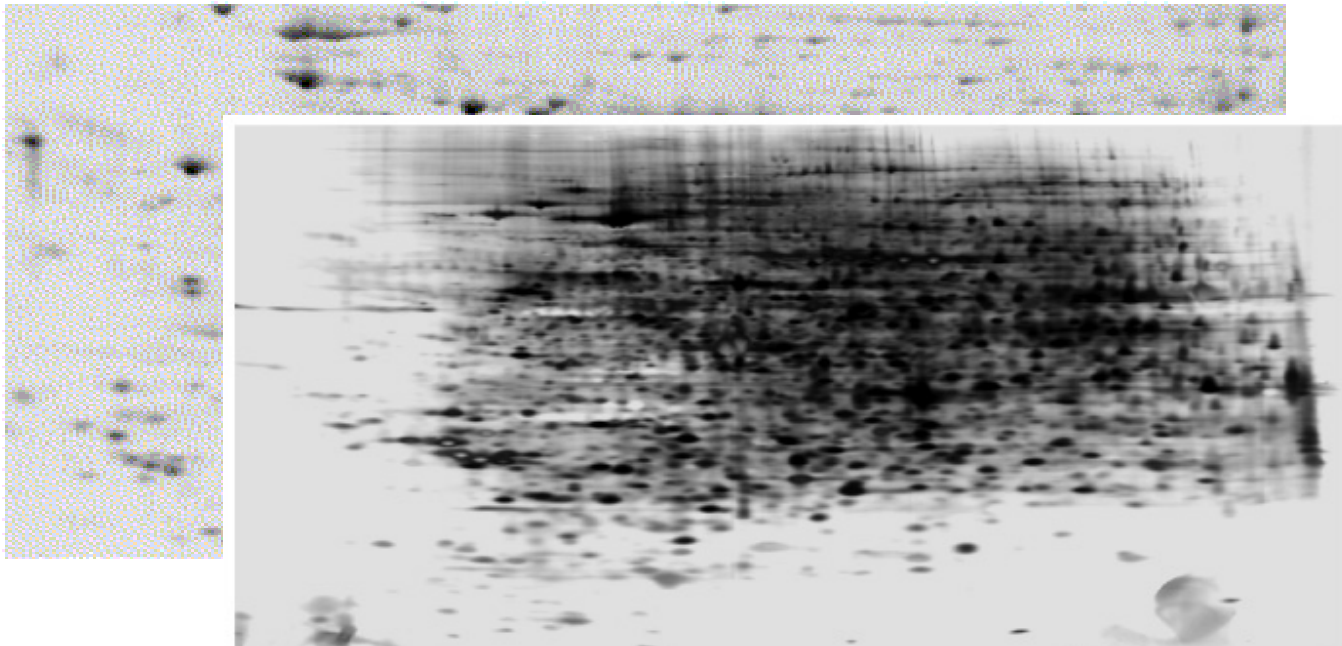
Quantification, clustering, ...

This Lecture

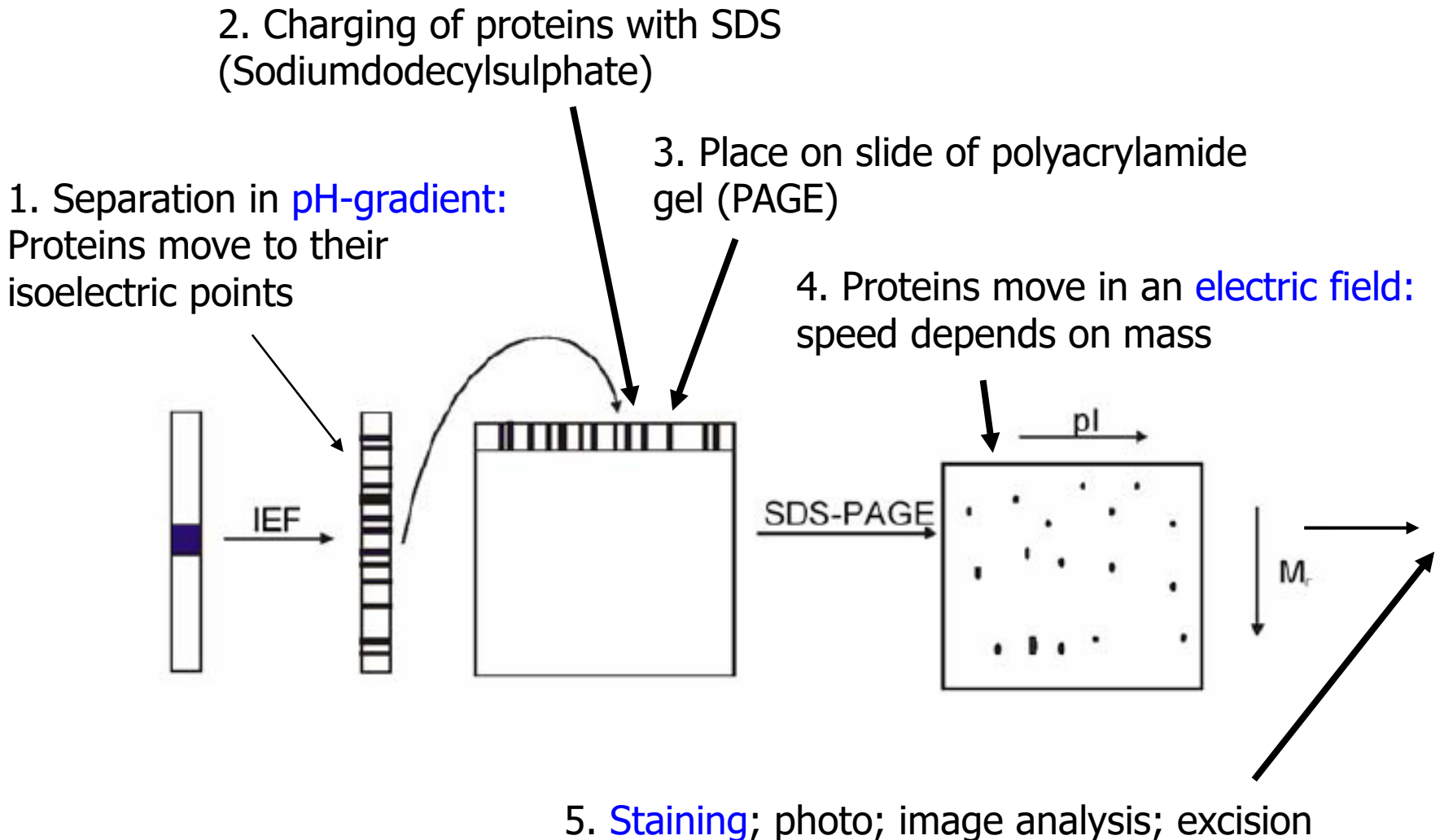
- Proteomics
- Separation
- Identification: Mass Spectrometry

2D Gel Elektrophoresis

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

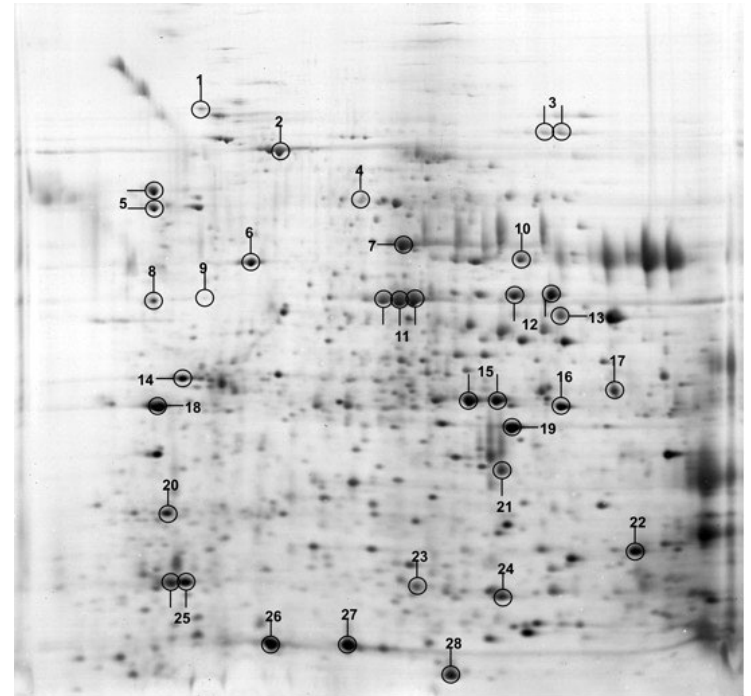


Method



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
 - Align image to reference
- Various databases of 2D-Gels



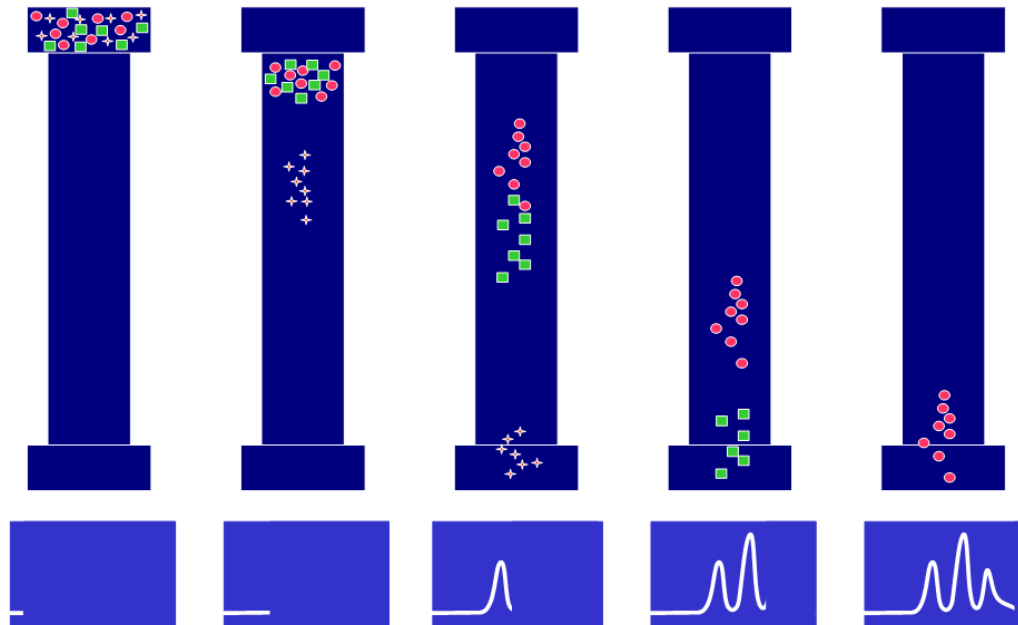
1 HSP86	11 p40	21 Myosin Light Chain
2 HSP70	12 Aldolase	22 Cycophilin
3 ATP:Guanidino Kinase	13 GAPDH	23 Superoxide Dismutase
4 Adenylate Dehydrogenase	14 14-3-3 e	24 Fatty Acid Binding Protein (Sm14)
5 Calreticulin	15 GST28	25 SME16
6 Actin	16 Triose Phosphate Isomerase	26 Thioredoxin
7 Enolase	17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

Pro / Contra

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

Liquide / Gas Chromatography

- Other option: GC/LC
 - Chamber contains two phases (liquid / liquid, liquid/gas)
 - Different speeds depending on mass/charge ratio
 - Separation by retention times

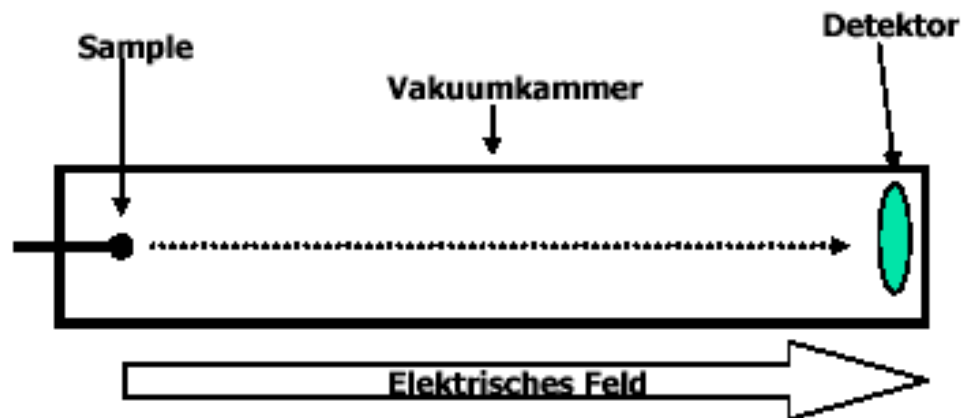


This Lecture

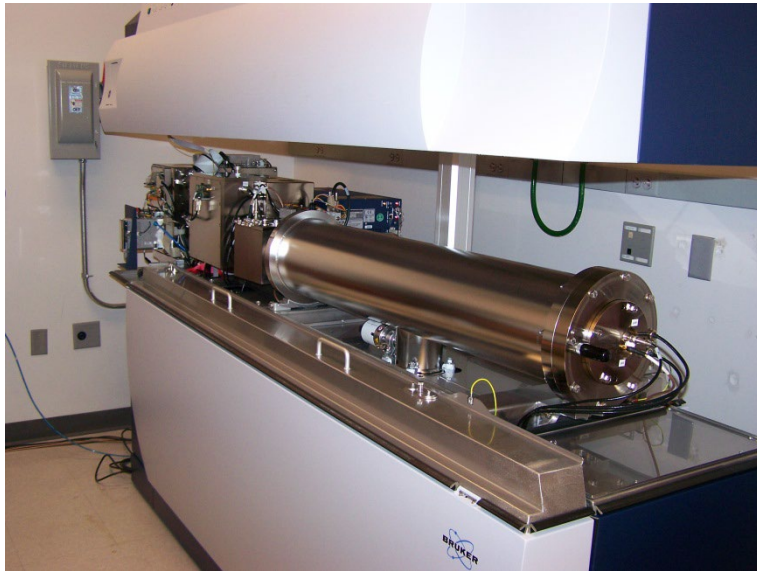
- Proteomics
- Separation
- Identification: Mass Spectrometry
 - Method
 - Algorithms: Naïve, probabilistic

Mass Spectrometry

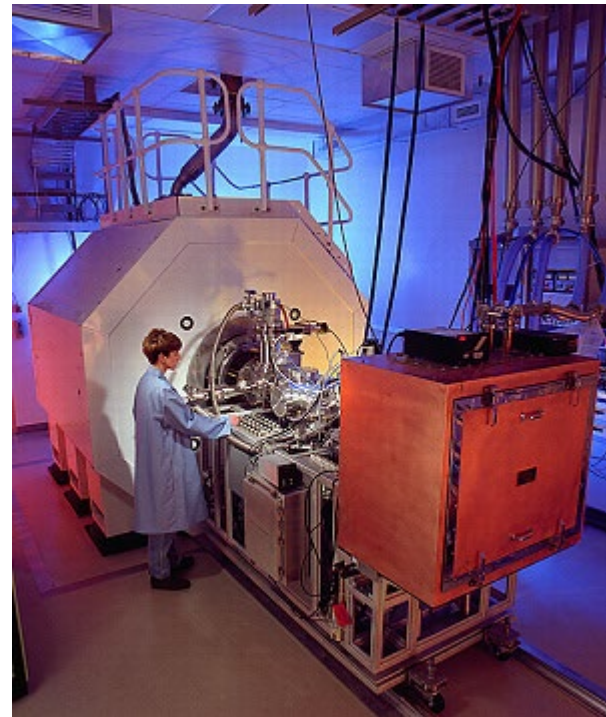
- Accelerate particles (must be charged) in an electric field
- Detector measures hits at back wall
- **Time of flight** (ToF) proportional to mass
 - Other techniques exist (magnetic drift, ...)
- **Spectrum of mass peaks** is used to identify particle



Mass Spectrometry



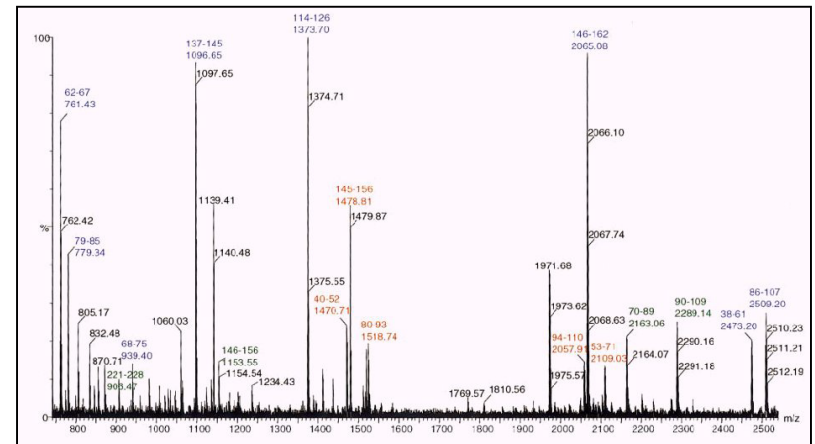
Source: <http://imr.osu.edu>



Source: <http://www.sysbio.org>

MS for Protein Identification

- Problem: Proteins are fragile and break during acceleration
- Solution
 - Break proteins into **peptides** before acceleration (digestion)
 - Measure peptides ToF (each **peptide one signal**)
 - Identify protein based on **spectrum of peptide signals**
- In theory, every protein has an almost **unique spectrum** due to its peptide composition
 - Using modern MS/MS, even different forms of the same protein are separable

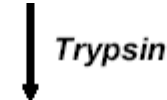


Digestion

Trypsin:

Cleaves after Arginine
und Lysine if next AA
is not Proline

N-Asp-Ala-Gly-Arg-His-Cys-Lys-Pro-Lys-Ser-Glu-Asn-Leu-Ile-Arg-Thr-Tyr-C

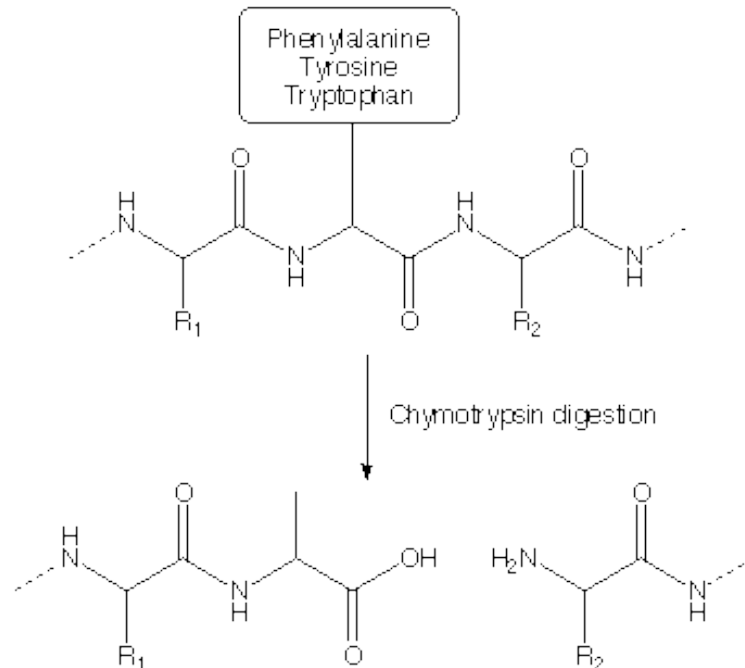


N-Asp-Ala-Gly-Arg

Ser-Glu-Asn-Leu-Ile-Arg

His-Cys-Lys-Pro-Lys

Thr-Tyr-C



Chymotrypsin:

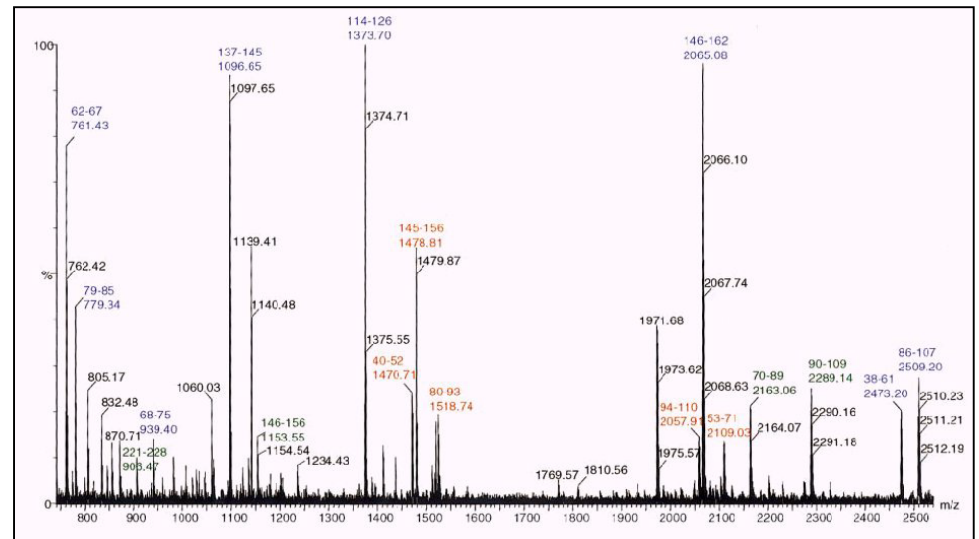
After Tyr, Trp, Phe, Met

Ionization

- Problem: Peptides often are uncharged – no acceleration
- Solution
 - MALDI – Matrix Assisted Laser Desorption / Ionization
 - Peptides 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
 - E.g. ESI: electrospray ionization

From Spectra to Peaks

- **Detecting peaks** and assigning them to peptides is difficult
 - Technical bias in runs / machines
 - Inaccuracies of measurement
 - Inhomogeneous sample preparation
 - Matrix etc.
 - Different quantities of peptides
- Creating a spectrum: **Signal processing** (not covered here)
 - Peak detection, peak disambiguation, noise filtering, ...



This Lecture

- Proteomics
- Separation
- Identification: Mass Spectrometry
 - Method
 - Algorithms: Naïve, probabilistic

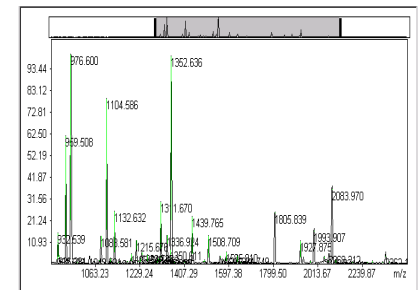
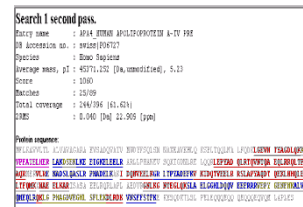
Algorithms for Protein Identification from Spectra

- We focus on **database-based** identification
- Idea
 - We have a database D of protein sequences d_1, d_2, \dots
 - Each d_i is subjected to **electronic digestion** – set of peptides
 - For each peptide, we know its theoretical ToF
 - Compute a **theoretical spectrum** s_i for each d_i
 - Measure real spectrum s of unknown protein k
 - Compare **empirical spectrum** s with all **theoretical spectra** s_i
- We can only **find what we already know**
 - I.e. proteins contained in D

Comparison



ANSHTDDLDELDSALDDFKDLNLSHO
QGLGMGLPDMRSKKRGKQKVSKEHDV
LPASDDDGMMVEDFLKQFEDLAGSKDLE
RYPKWLKENEASLSEDKYRSQYKLI
ECGQPPSDIVKEIDPGDFASLGOISPEM



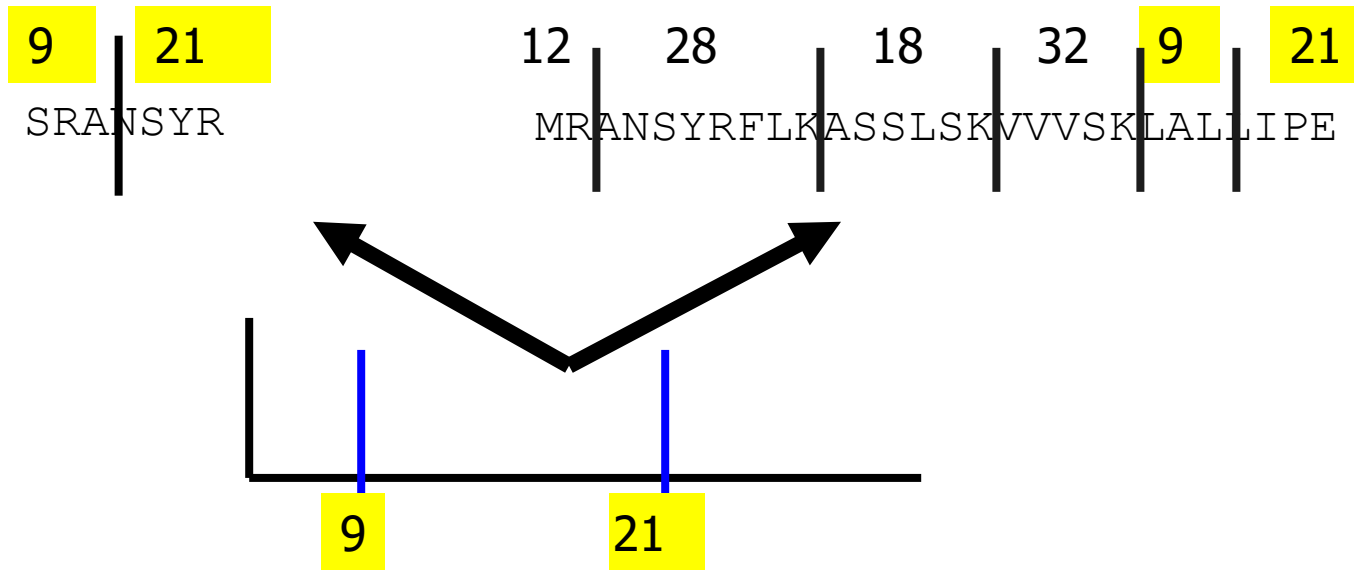
Naive Algorithm: Hitcount

- Compare measured spectrum s with all s_i in DB
- Protein d_i which has the **most peaks in common** wins
 - Input: $s = \{p_1, \dots, p_m\}$, database D with each $s_i = \{q_{i1}, \dots, q_{ij}\}$
 - For each s_i : Compute $|s \cap s_i|$
 - Protein d_i where s_i has **maximal overlap** wins
- Complexity?
 - Keep peak lists s and s_i sorted
 - We need to compare $|s|$ hits with $|D|$ proteins in DB
 - Let q be the average number of peaks in a database spectrum
 - Together: **$\sim(|s| + q) * |D|$ comparisons**
 - Can be sped-up further (indexing)

Why “Naïve”?

- Peptide masses are not really equal (e.g. isotopes)
 - Small deviation – **nearest peak**; match might not be unique
- Some (short) peptides are **more frequent** than others
 - Some peptides appear in almost all proteins – little signal
 - Smaller peptides are much more frequent but **much less specific**
 - And peptide length is stochastic
 - Frequent peptides should have a lower impact
- Proteins have **different lengths**
 - Longer proteins have a higher a-priori chance for more peak matches

Example; Which one would you prefer?



More Problems

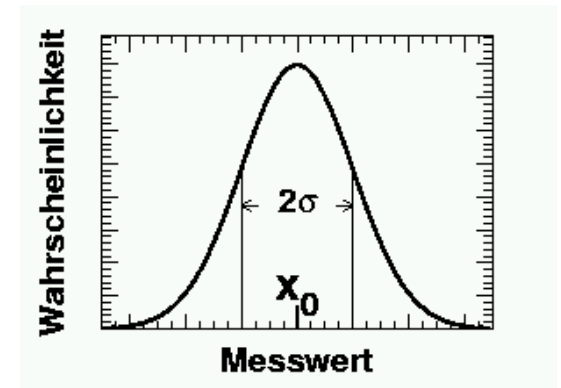
- Many sources of error
 - Enzymes don't work 100%
 - Theoretical spectra don't match
 - Protein sequences in **DB contain errors**
 - Especially when directly translated from genome
 - Leads to theoretical spectra not existing in nature
 - Posttranslational modifications modify real spectra
 - MS is not perfect – spurious, shifted, missing peaks
 - Lead to **false positive and false negative** peak matches
- Closed-world assumption
 - What if real sequence is not in the database?
 - Some protein always has the highest count – high enough?
 - **No confidence scores**

Some Relevant Algorithms

- Heuristic: MOWSE (outdated)
 - Considers total protein mass and peptide frequencies
 - Generates a score
- Probabilistic algorithm: **Profound**
 - Copes with measurement errors, deviation in protein mass, and different peptide frequencies
 - Generates a **probability of match** for each protein (~ confidence)
- Many more (and newer) algorithms
 - MASCOT, PeptIdent, ProteinProspector, SEQAN, ...
 - Many are commercial – inner work unknown

Example of a Probabilistic Method: ProFound [zc00]

- Given: Measured spectrum D and a protein k
 - D : Previously s ; k : previously s_i
- ProFound computes prob. $p(k|D)$ that D was produced by k
- The formula is complex; its derivation is even more complex and skipped
- Basic assumption: Measured peptide masses are normally distributed around the “canonical” value
 - Most probable isotope composition



ProFound Approach

- First step: Assign peaks from k to closest peaks from D
 - A-priori assignment is a strong first filter; errors are propagated
- Then compute probabilities using

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

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 observed by spectrum D given the background information I
- $p(k|I)$: A-priori probability of k in the given species / cell / tissue
- N: Predicted number of peptides of database protein k
- r: Number of hits between D and k (results from initial assignment)
- m_{\max}, m_{\min} – range of observed masses for current peak (background)
- σ_i – standard deviation of current peak (background)
- g_i : How often is the i'th peptide contained in k?
- m_i : Mean mass of the DB peak (background)
- m_{ij0} : Measured mass of j'th occurrence of this peptide
- F_{pattern} : Heuristic factor dealing with “overlapping peaks”

ProFound Explanation

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

- How many of the expected peptides of k did we observe?
- Multiply probabilities of all hits
- “Freedom” of measurements of hits for this peptide
- Many predicted peaks may create only one measured peak
- Probability of the difference between the expected mass and the measured mass (assuming normal distribution)

ProFound Intuition

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

- Many hits ($r \sim N$) – score goes down (outweighs influence of more factors in the red product)
- Hits with a small stddev or a broad range – score goes up
- Many observed peaks match the predicted peaks – score goes up
- Observed peaks close to canonical peaks – 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 said
- 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 predicted peaks
 - This problem is tried to be covered by F_{pattern}

Further Reading

- Basics on proteomics: Every Bioinformatics book
- 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.
- Pappin, D. J. C., Hojrup, P. and Bleasby, A. J. (1993). "Rapid identification of proteins by peptide-mass fingerprinting." *Current Biology* 3(327-332).
- Survey: Colinge J, Bennett KL (2007) Introduction to Computational Proteomics. PLoS Comput Biol 3(7): e114