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 disease (state)
- Metabolomics, interactomics, bibliomics, cellomics, …
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 transient
  - Proteins are favorite 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, …
  - Reactivity highly context-dependent: temperature, solution, pH, …
Typical Proteomics Workflow

1. **Extraction**
   - From a cell mixture

2. **Separation**
   - 2D gel electrophoresis / LC/GC

3. **Isolation**
   - From the gel / from the flow

4. **Identification**
   - Mass spectrometry

5. **Analysis**
   - 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

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

2. Charging of proteins with SDS (Sodiumdodecylsulphate)

4. Place on slide of polyacrylamide gel (PAGE)

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
Pro / Contra

- Comparably simple and cheap, 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 $<20$KD or $>200$KD
    - No highly charged proteins
    - No detection of low concentrations
    - No membrane proteins (depending on method)
    - ...
  - No de-novo protein identification
  - Limited accuracy in comparative identification
Liquide / Gas Chromatography

- Separation today: **Mostly GC/LC**
  - Chamber contains two phases (liquid / liquid, liquid/gas)
  - Proteins travel with **different speed** depending on mass/charge ratio
This Lecture

• Proteomics
• 2D Gels
• **Mass Spectrometry**
  - Method
  - Algorithms: Naïve, heuristic, probabilistic
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 (enzym. digestion)
  - Measure peptides (each peptide one signal)
  - Identify protein based on spectrum of peptide hits

• In theory, every protein has an almost unique spectrum
  - 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

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
  - E.g. 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
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 **theoretical spectrum** per protein in the database
- Measure spectrum of unknown protein
  - Compare spectra

- Again, we can only **discover what we already know**
Illustration

Real experiment

Theoretical experiment

Comparison
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: $P = \{p_1, \ldots, p_m\}$, $S_i = \{k_1, \ldots, k_{m(i)}\}$
- Compute for each $i$: $|P \cap S_i|$  
- Sequence $S_i$ with maximal overlap wins

Complexity?
- Keep all hit lists sorted
- We need to compare $|P|$ hits with $n$ proteins in DB
- Together: $\sim |P| \times n$ comparisons
Indexed version (Inverted list, inverted files)

- **Input:** \( P = \{p_1, ... p_m \} \), \( S_i = \{k_1, ..., k_{m(i)} \} \)

- **Indexing:** Compute an array \( A \) storing for each peptide \( k \) all sequences in \( DB \) containing it: \( A[k] = \{S_i | k \in S_i \} \)

- **Search**
  - Initialize a counter for each protein in \( DB \): \( M[i] = 0 \)
  - For all \( k \in P \), for all \( i \): If \( S_i \in A[k] \): \( M[i] = M[i] + 1 \)
  - Sequence \( S_i \) with \( M[i]!\neq \text{MAX} \) wins

- **Complexity?**
  - Theoretical worst-case: \( \sim|P|*n \) increments
  - Realistic number: \( \sim k*|P| \) increments (\( k \) small)
    * \( P \) has hits only with few proteins
Example

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

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

\[ X: \text{Peptide mass (1000-5000 Dalton)} \]
\[ Y: \text{Peptide count (log)} \]
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

- **Probabilistic algorithm: Profound**
  - Bayes’ statistics
  - Copes with measurement errors, deviation in protein mass, and different peptide frequencies
  - Generates a probability of match for each protein

- **Many more (and newer) algorithms**
  - MASCOT, PeptIdent, ProteinProspector, SEQAN, ...
ProFound [ZC00]

• 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

\[ P(k|DI) \propto P(k|I) \frac{(N - r)!}{N!} \prod_{j=1}^{r} \left\{ \sqrt{\frac{2}{\pi}} \frac{m_{\text{max}} - m_{\text{min}}}{\sigma_i} \times \sum_{j=1}^{g_i} \exp \left[ - \frac{(m_i - m_{i0})^2}{2\sigma_i^2} \right] \right\} \]
Legend

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

- \( p(k | D, I) \) = prob. that protein \( k \) was observed by spectrum \( D \) and background information \( I \)
- \( N \): Number of peptides of protein \( k \)
- \( r \): Number of hits between \( D \) and \( k \) (with a certain fuzzy’ness)
- \( m_{\text{max}}, m_{\text{min}} \) – range of observed masses for current hit
- \( \sigma_i \) – standard deviation of \( i \)'th hit
- \( m_i \): mass of the hit (must be between \( m_{\text{max}} \) and \( m_{\text{min}} \))
- \( g_i \): How often is the peptide contained in \( k \)?
- \( m_{ij0} \): 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
- \( F_{\text{pattern}} \): 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

\[ P(k|Dl) \propto P(l|D) \frac{(N - r)!}{N!} \prod_{i=1}^{r} \sqrt{\frac{2}{\pi} \frac{m_{\text{max}} - m_{\text{min}}}{\sigma_i}} \times \sum_{j=1}^{g_i} \exp \left[-\frac{(m_i - m_{j0})^2}{2\sigma_i^2}\right] F_{\text{pattern}} \]

- Many hits \((r \sim N)\) – score goes down (outweighs influence of more factors in the red part)
- Hits with a narrow range – score goes up
- Observed peak matches many expected 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 $F_{\text{pattern}}$

• Many more suggestions since 2000
Further Reading

• Basics on proteomics: Every Bioinformatics book