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 disease (state)

- 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, ~100K proteins, ~500K protein forms
  - Handling: No PCR, no hybridization, no simple synthesis, no sequencing, no long-term „storage“ as clones, high reactivity, ...
  - Reactivity 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 Electrophoresis

- 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
  - Align image to reference

• Various databases of 2D-Gels
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 overlapping spots
  - Many restrictions
    - No proteins with \(<20\text{KD}\) or \(>200\text{KD}\)
    - 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
Liqueide / 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 at defined points before acceleration (digestion)
  - Measure peptides (each peptide one signal – time of flight)
  - 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 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 $d_1, d_2, \ldots$
    
    • Each is subjected to electronic digestion – set of peptides per protein
    
    • For each peptide, we know its theoretic ToF
    
    • One theoretical spectrum $s_i$ per protein $d_i$ in the database

  - Measure spectrum $s$ of unknown protein $k$

  - Compare *empirical* spectrum $s$ of $k$ with all *theoretical* spectra $s_i$

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

Real experiment

Theoretical experiment

Comparison
Naive Algorithm: Hitcount

- Compare measured s with all $s_i$ in DB
- Sequence which has the most peptides in common wins
  - Input: $s = \{p_1, ..., p_m\}$, $s_i = \{q_1, ..., q_j\}$
  - For each $s_i$: Compute $|s \cap s_i|$
  - Sequence $s_i$ with maximal overlap wins

- Complexity?
  - Keep all peak lists 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: $\sim (|s| + q) \times |D|$ 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
  - Should have a lower impact
- Proteins have different lengths
  - Longer proteins have higher chances for high scores
Example

- Which one would you prefer?

SRANSYR  
MRANSYRFKLASSLSKVVVSKLALLIPE

9  21  12  28  18  32  9  21

9  21
More Problems

• Enzymes don’t work 100% correct
• Protein sequences in DB contain errors
  – Especially when directly translated from genome
  – Leads to theoretical spectra not existing in nature
• Ignores posttranslational modifications
• MS is not perfect – spurious, shifted, missing peaks
• Some protein always has the highest count – what if real sequence is not in the database?
  – No confidence scores
Practically 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, ...
ProFound [ZC00]

- **Probabilistic method**
- Computes, for a given spectrum $D(s)$ 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
- Assumption: Measured peptide masses are normally distributed around the “canonical” value
  - Most probable isotope composition
- First step: Assign peaks from $k$ to peaks from $D$
ProFound Formula

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

- \( p(k|D,I) = \text{prob. that protein } k \text{ was observed by spectrum } D \text{ with background information } I \)
- \( p(k|I) = \text{A-priori probability of } k \text{ in the given species / cell / tissue} \)
- \( N \): Number of peptides of database 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 peak (background)
- \( \sigma_i \) – standard deviation of current peak (background)
- \( m_i \): Mean mass of the DB peak (background)
- \( g_i \): How often is the peptide contained in \( k \)?
- \( m_{ij0} \): Empirical mass of \( j \)’th occurrence of this peptide
- \( F_{\text{pattern}} \): Heuristic factor dealing with “overlapping peaks”
ProFound Explanation

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

- How many of the expected peptides for \( k \) 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 deviation of the canonical mass to the measured mass (assuming normal distribution)
ProFound Intuition

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

- Many hits \((r \sim N)\) – score goes down (outweighs influence of more factors in the red product)
- Hits with a small stddev (or 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 $F_{\text{pattern}}$
Further Reading

- Basics on proteomics: Every Bioinformatics book