Mass spec basics: MS-based strategies for protein identification in proteomics

Niels Lion
Service Régional Vaudois de Transfusion Sanguine
niels.lion@mavietonsang.ch
Course syllabus

• April 14: Introduction to proteomics

• April 21: Two-dimensional gel electrophoresis

• May 5: Liquid-chromatography-mass spectrometry

• May 12: Bioinformatics and protein quantitation

• May 19: The case of abundant proteins / complex samples

• May 27: Case study in transfusion medicine
Outline

• Introduction: what does it mean to identify a protein in proteomics?
• Bottom-up proteomics
  • Peptide mass fingerprinting
    • Principle
    • Importance of mass accuracy
    • Procedure
  • Shotgun LC-MS/MS
    • Principle
    • Performances
    • The protein inference problem
• Top-down proteomics
Why do proteomics?

Same individual, same genome, somehow different proteomes....
Why do proteomics?

Biological functions are embedded in proteins:

- structure: actin, spectrin...
- energy conversion: glycolytic enzymes (glucose to ATP)
- signaling: Fas receptor
- homeostasis: Gardos channels
- Enzymatic activity control: protein kinases / phosphatases
- transport: haemoglobin

One fundamental goal of proteomics is to understand how proteins are expressed from genes, how they interact, how they are regulated...
Example: RBC metabolism
Why do proteomics (2)?

• The search for disease biomarker has been driven for years by physiology and biochemistry (rational approach).

• This research is intrinsically incremental and cumulative, and the discovery of a few biomarkers can take decades.

• On the contrary, if one can analyse the entire proteome of patients and compare it the one of healthy individuals, one might be able to identify what makes the difference (see later lecture for a critical appraisal of this approach).
Importance of biomarkers

Early cancer diagnosis is key to life expectancy!
Position of the analytical challenge

- In human beings, 30-50'000 genes
- Each gene gives 5-20 proteins = 150’000 - 1’000’000 proteins
- Dynamic range is at least $10^5$ (more probably $10^6$ to $10^{12}$)
- Protein expression is temporal
- Proteins can be a few tens of amino acids up to a few thousands
- Proteins can be very hydrophobic (membrane proteins)
- Proteins can be very acidic or very basic

Niels Lion, Bioanalytics and Biosensors, Spring 2011
Example of concentration dynamic range: Human plasma

Leigh Anderson, The Plasma Proteome Institute
Available information

For some organisms, the genome is sequenced (< 200 organisms):

- 22 archaea
- 142 bacteria
- 35 eukaryotes, including man, monkey, dog, mice...

From the genome, you have Open Reading Frames (ORFs), Expressed Sequenced Tags (ESTs), or genes

+ protein databases
## Organisms complexity

<table>
<thead>
<tr>
<th></th>
<th>Klebsiella pneumoniae</th>
<th>E Coli O6</th>
<th>Homo Sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrEMBL entries</td>
<td>3684</td>
<td>3664</td>
<td>85495</td>
</tr>
<tr>
<td>SwissProt entries</td>
<td>223</td>
<td>1672</td>
<td>20233</td>
</tr>
</tbody>
</table>
Organisms coverage

<table>
<thead>
<tr>
<th>Organism</th>
<th>% of known proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>K. Pneumoniae</td>
<td></td>
</tr>
<tr>
<td>E. Coli</td>
<td>30</td>
</tr>
<tr>
<td>H. Sapiens</td>
<td></td>
</tr>
</tbody>
</table>
What does it mean to identify a protein?

• Affinity definition: protein recognition by a specific antibody (Western blotting)

• Gene-product definition: the proteinaceous specie comes from a specific gene

• Sequence definition: amino acid sequence

• Post-translational modifications definition: amino acid sequence + phosphorylation, acetylation, glycosylation....
# Post-translational modifications (PTMs)

<table>
<thead>
<tr>
<th></th>
<th>disulfide bond</th>
<th>oxidation</th>
<th>glutathionylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulphydryls (C)</td>
<td>-2.0159</td>
<td></td>
<td>+15.9994</td>
</tr>
<tr>
<td>cysteinylation</td>
<td>+119.1442</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulphydryls (C)</td>
<td></td>
<td>oxidation</td>
<td></td>
</tr>
<tr>
<td>cysteinylation</td>
<td>+119.1442</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glutathionylation</td>
<td>+305.3117</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amines (K/N-terminus)</td>
<td>methylation</td>
<td>formylation</td>
<td></td>
</tr>
<tr>
<td>acetylation</td>
<td>+41.0373</td>
<td>lipoic acid</td>
<td>+188.3147</td>
</tr>
<tr>
<td>farnesylation</td>
<td>+204.3556</td>
<td>myristoylation</td>
<td>+210.3598</td>
</tr>
<tr>
<td>biotinylation</td>
<td>+226.2994</td>
<td>palmitoylation</td>
<td>+238.4136</td>
</tr>
<tr>
<td>stearoylation</td>
<td>+266.4674</td>
<td>geranylation</td>
<td>+272.4741</td>
</tr>
<tr>
<td>Amines (K/N-terminus)</td>
<td>pyroglutamic acid (Q)</td>
<td>-17.0306</td>
<td>deamidation (Q/N)</td>
</tr>
<tr>
<td>carboxylation (E/D)</td>
<td>+44.0098</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acids &amp; amides (E/D/Q/N)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hydroxyl groups (S/T/Y)</td>
<td>phosphorylation</td>
<td>+79.9799</td>
<td>sulphation</td>
</tr>
<tr>
<td>carbohydrates (S/T/N)</td>
<td>pentoses</td>
<td>deoxyhexoses</td>
<td>+146.1430</td>
</tr>
<tr>
<td>hexosamines</td>
<td>+161.1577</td>
<td>hexoses</td>
<td>+162.1424</td>
</tr>
<tr>
<td>N-acetylhexosamines</td>
<td>+203.1950</td>
<td>sialic acid</td>
<td>+291.2579</td>
</tr>
</tbody>
</table>

From Prowl, Rockefeller University, [http://prowl.rockefeller.edu/aainfo/postmod.html](http://prowl.rockefeller.edu/aainfo/postmod.html)

Niels Lion, Bioanalytics and Biosensors, Spring 2011
Various kind of MS are available
## Summary of mass spectrometer performances

<table>
<thead>
<tr>
<th>Mass spectrometer type</th>
<th>Mass range</th>
<th>Resolving power</th>
<th>Accuracy (ppm)</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quad or ion traps</td>
<td>&lt;2000 or 4000</td>
<td>3000-5000</td>
<td>50-100</td>
<td>low</td>
</tr>
<tr>
<td>TOF</td>
<td>unlimited</td>
<td>$10^4$-$10^5$</td>
<td>5-50</td>
<td>medium</td>
</tr>
<tr>
<td>Orbitrap</td>
<td>&lt;2000 or 4000</td>
<td>$10^5$-$10^7$</td>
<td>1-50</td>
<td>high</td>
</tr>
<tr>
<td>FT-ICR</td>
<td>&lt;2000 or 4000</td>
<td>$10^5$-$10^7$</td>
<td>0.1-10</td>
<td>high</td>
</tr>
</tbody>
</table>
Why can’t we identify proteins from mass alone?

<table>
<thead>
<tr>
<th>Mass accuracy</th>
<th>Number of protein candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ppm</td>
<td>1</td>
</tr>
<tr>
<td>10 ppm</td>
<td>1</td>
</tr>
<tr>
<td>20 ppm</td>
<td>1</td>
</tr>
<tr>
<td>100 ppm</td>
<td>4</td>
</tr>
</tbody>
</table>

But most proteins carry post-translational modifications that are not taken into account in the calculation of the molecular weight in databases!

For example, glycophorin A (red blood cell membrane protein) has a theoretical MW of 11 kDa but appears at 30 kDa on an SDS-PAGE.

Theoretical mass of Human Serum Albumin
Number of candidates in SwissProt
Different workflows

Bottom-up

Top-down

Niels Lion, Bioanalytics and Biosensors, Spring 2011
Bottom-up approaches: 2D-GE with peptide mass fingerprinting
The golden approach: 2D-GE/MALDI-TOF (or 2D-GE-LC-ESI-MS)

Human platelets, 3-10, silver staining
Features of 2D-GE

• Pros:
  - Analysis of whole functional proteins
  - Separation of thousands of proteins

• Cons:
  - Loss of very acidic, very basic, very large, very small, very hydrophobic proteins
  - Requires time and experimental expertise
Subsequent procedure

- Spot excision
- In-gel digestion by trypsin
- Gel piece solubilization
- Sample prep for MALDI (co-cristallization with matrix)
- Analysis by MALDI-TOF or LC-ESI-MS
- Database interrogation with peptide masses
The golden approach: 2D-GE/MALDI-TOF (or 2D-GE-LC-ESI-MS)

Overexpressed proteins in Intercept-treated platelets
- glutaredoxin-related protein 5 (spot 1)
- Protein DJ-1
- Ras related protein Rap-11A

Underexpressed proteins in Intercept-treated platelets
- Glutaredoxin-related protein 5 (spot 2)
- FGFR1 oncogene partner 2 (spot 3)
- Talin-1
- Twinfilin-2
- Drebrin-like protein
- Prefoldine subunit 3
- Fermitin family homolog 3 (Kindlin-3)
Important notions for peptide mass fingerprinting

- Identification score: gives a measure of the probability that the identification is purely random.

- Sequence coverage: percentage of the protein sequence that is covered by the identified peptides
### Identification quality

<table>
<thead>
<tr>
<th>Low sequence coverage</th>
<th>High sequence coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor score</td>
<td>Ideal case</td>
</tr>
<tr>
<td>Gene product identification</td>
<td>Low reliability</td>
</tr>
</tbody>
</table>

**Thursday, April 14, 2011**
Exemple of HSA

<table>
<thead>
<tr>
<th>Human serum albumin</th>
<th>193 peptides:</th>
</tr>
</thead>
<tbody>
<tr>
<td>MKWVTFISLLFLFSSAYSRGVFRR</td>
<td>(KYICENQDSISSKLKECCEKPLEKS)</td>
</tr>
<tr>
<td>DAHKSEVAHRFKDLGEENFKAL</td>
<td>(KYLYEIARRHPFYAYEPELLFFAKR)</td>
</tr>
<tr>
<td>VLLIAFAQYLQQCPF</td>
<td>(KLCTVATLRETYGEMADCCAKQEPERN)</td>
</tr>
<tr>
<td>EDHVKLVNEVTTEFAKTCVADES</td>
<td>(KEFNAETFTFADICTLSEKERQIKK)</td>
</tr>
<tr>
<td>AENCSDKSLHTLFGDKLCTVAT</td>
<td>(KCCTESLVNNRPFCFSALEVDETYVPKE)</td>
</tr>
<tr>
<td>LRETYGEMADCCAKQEP</td>
<td>(KSHCIAEVENDEMPADLPPLAADFVESKD)</td>
</tr>
<tr>
<td>ERNECFLQHKDDNPNLPRVLR</td>
<td>(KQNCHELFEQLGHEYKFQNALVRYTKK)</td>
</tr>
<tr>
<td>PEVDVMCTAFHDNEETFLKKYL</td>
<td>(RETYGEMADCCAKQEPERNECFLQHKD)</td>
</tr>
<tr>
<td>YEIARRHPFYAYEPELLF</td>
<td>(RMPCAEDYLCSVNLQCLVLHEKTPVSDRV)</td>
</tr>
</tbody>
</table>

Niels Lion, Bioanalytics and Biosensors, Spring 2011
Identification on an ion trap (200 ppm)
Identification on an ion trap in the presence of noise (200 ppm)
Identification on a TOF instrument (50 ppm)
Identification on FT-ICR (1 ppm)

Niels Lion, Bioanalytics and Biosensors, Spring 2011
Comparison of different instruments

Niels Lion, Bioanalytics and Biosensors, Spring 2011
PMF practical approach

1. Acquire the spectrum

Interrogation with unprocessed data: no identification!
PMF practical approach

1. Acquire the spectrum
2. Remove peaks from trypsin autolysis and non-peptidic peaks

Peptide masses are not random and fall at precise masses.
PMF practical approach

1. Acquire the spectrum
2. Remove peaks from trypsin autolysis and non-peptidic peaks

Submit only true peptide masses: score of 84.
PMF practical approach

1. Acquire the spectrum
2. Remove peaks from trypsin autolysis and non-peptidic peaks
3. Recalibrate data.

Recalibrate data (statistical or a posteriori): score of 137.
Features of 2D-GE-PMF

- Analysis of individual proteins (discrimination of different isoforms on the gel)
- Good sequence coverage (30-80%)
- Low throughput
- Manual work
Bottom-up approaches: shotgun LC-MS/MS
Peptides tandem mass spectrometry

One can fragment peptides in the gas phase by increasing their energy. Energy uptake = breakage of the peptide backbone (one peptidic bond = three breakage positions).
Collision induced dissociation of peptides
Resulting fragments

etc........
Ordered fragments
Peptide sequencing by CID
Integration in a proteomic workflow

Protein mixture → Peptide mixture → SCX → RP → n MS/MS spectra

Generation of hundreds of MS spectra, thousands of MS/MS spectra

Importance of automatic identification tools
Data-dependent acquisition parameters

1. Acquire 1 spectrum (100 ms)

2. Select the most abundant peak

3. Fragment it by CID (generic parameters) and acquire product ion spectrum (100 ms)

4. Add parent mass in the exclusion list (mass, time)

5. Return to step 2 till the predefined number of MS/MS spectra is reached.

6. Return to step 1.
Shotgun, bottom up: data dependent acquisition

- large peaks are automatically selected from MS/MS in each spectrum (most often 3).
- All other peaks are ignored.
- Exclusion lists allow not to resequence the same peptides in the following spectra.

The automatically selected peptides are not necessarily the more interesting ones!
Database interrogation

• The dataset comprises typically 1000 MS spectra and 3000-6000 MS/MS spectra.

• Each peak list (MS/MS) is made of the mass of the parent ion and masses of fragments.

• The search engine (Phenyx, Sequest, Mascot...) generates from each protein in the database the list of tryptic peptides, and a theoretical MS/MS spectrum. It then compares it with the experimental peaks list, and scores it. The best score is the identification.

Phenyx example: see exercise #2
Shotgun, bottom-up: problems and limitations

Undersampling in MS/MS mode; in most instruments, only 2-5 MS/MS spectra can be acquired per second:

In shotgun proteomics, sequence coverages are very low!!!
Shotgun, bottom-up: poor MS/MS identification

Distribution of XCorr scores over 10000 MS/MS spectra of human proteins.

Alternative strategy: offline selection

1st LC-MS run

Then selection of interesting spots for targeted MS/MS
DDA versus targeted proteomics

Analysis of the tryptic digest of beta-lactoglobulin

Identified peptides

The protein inference problem: how to identify proteins from peptide identifications?
Protein isoforms

F-actin capping protein beta subunit

Niels Lion, Bioanalytics and Biosensors, Spring 2011
Protein isoforms

Epithelial protein lost in neoplasm isoforms are truncated at the N-ter
Database redundancy

There are six entries in Entrez Protein Database for the same protein, heat shock 70-kDa protein 9B (with small variations)
Peptide grouping scenarios

- **Distinct proteins**
  - Peptides: 1, 2, 3, 4
  - Proteins: A, B

- **Differentiable proteins**
  - Peptides: 1, 2, 3, 4
  - Proteins: A, B

- **Indistinguishable proteins**
  - Peptides: 1, 2, 3, 4
  - Proteins: A, B

- **B: Subsumable protein**
  - Peptides: 1, 2
  - Proteins: A, B, C

- **A group of proteins identified by shared peptides only**
  - Peptides: 1, 2, 3, 4
  - Proteins: A, B, C
Protein ID from peptide ID

peptides

proteins

protein summary list

minimal list of proteins:
1. Protein A
   peptides 1, 2
2. Protein B
   peptides 3, 4*
3. Protein C
   peptides 4*, 5
4. Protein E
   peptides 6*, 7, 8
5. Protein F, Protein G
   peptides 9*, 10*
6. Protein group:
   (1) Protein H
       peptides 11*, 12*, 13*
   (2) Protein I
       peptides 11*, 12*
   (3) Protein J
       peptides 11*, 13*

"protein" count: 6

no conclusive evidence:
7. Protein D
   peptides 6*
Take-home message

Protein ID from MS and MS/MS data must be considered with precaution, due to a lot of instrumental, methodological and fundamental limitations.

One has to be conscious of what type of identification was performed (gene product, protein isoforms, protein with PTMs...)