Introduction to Proteomics

- Mass Spectrometry for proteomics
- Protein identification by MS/MS
- Proteomics workflows and applications
• Mass Spectrometry for proteomics
Mass spectrometry: essential functions

**Sample** -> **Ion Source** -> **Mass Analyzer** -> **Detector**

**ION SOURCE**
- ESI: Electrospray Ionisation
- MALDI: Matrix Assisted Laser Desorption/Ionization

**MASS ANALYZER**
- Quadrupoles
- Ion traps
- Time-of-flight with reflectron
- TOF/TOF
- Orbitrap
- FT-ICR

**DETECTOR**
- Faraday cup
- Scintillation counter
- Electromultiplier
- High-energy dynodes with electron multiplier
- Array (detector)
- FT-MS
MS of peptides vs. proteins

• Small peptides « fly » much better than big proteins
  – Higher ionisation efficiency → sensitivity
  – Signal intensity inversely proportional to mass

• MS in proteomics is mostly (but not only) MS of peptide fragments after protein digestion (typically w. TRYPsin)
MS of peptides: general concepts

• MS: measures charged species (ions)

• MS: we ALWAYS measure M/Z, not \( M_r \)

• Molecular mass \( (M_r) \Leftrightarrow M/Z : \)

\[
\frac{M}{Z} = \frac{M_r + (M_a \times Z)}{Z}
\]

\( M_a \): mass of the ionizing adduct (typically H+ for positive MS)
\( M_a(H^+) = 1.0072 \text{ u} \)

• Work mostly in positive (+) ion mode:
  => peptides protonated (MH+) due to acidic conditions and ionisation process

• Single or multiple \((Z>1)\) charge states can be observed
  f (compound, ionisation mode, instrument)
MALDI-TOF of a tryptic digest of BSA

+TOF MS 50 MDAs ans from Sample 1 (BSADigest 100 fmol) of BSADigest 100 fmol MS... Max 13050 counts

a=3.56217430068478150e-004, t0=3.64725878201043440e+001, Thresholded

YLYEIAR

LSQKFPK

LVNELTEFAK

FKDLGEEHFK

HPEYAVSVLLR

HLVDEPQNLIK

LGEYGFQNALIVR

DAFLGSFLYEYSR

KVPQVSTPCLVEYSR
1+ versus 2+ ions

LVNELTEFAK  Mr = 1162.6234

MH⁺

MH₂²⁺
Modes of measurement: MS & MS/MS

**Ion production (ionisation)**

**Ion separation**

**Ion detection**

---

**Ion production (ionisation)**

**Ion separation — isolation of “parent” ion**

**Ion fragmentation (CID)**

**Ion separation — separate fragment ions**

**Ion detection — measure fragment ions**

---

**MS**

*Detect all ions present*

---

**Tandem MS (MS/MS)**

*Fragment a specific ion and measure fragments*
CID = collision induced dissociation

Precursor ion = parent ion : the one being fragmented

Daughter ions = fragment ions produced by CID

Tandem mass spectrometry = MS/MS
  - here : the combination of ion selection / CID / fragment analysis

ESI of tryptic peptides typically generates 2+ - 3+ charged ions due to the presence of Lys or Arg at the C terminal end of the peptides

y- and b- ion series fragments are usually observed in MS/MS fragmentation spectra of tryptic peptides.
Modes of measurement: MS & MS/MS

**MS**
- Ion production
- Ion separation
- Detection

**MS/MS** (Tandem MS)
- Ion production
- Ion separation
- Fragmentation (CID)
- Detection
Peptide backbone fragmentation in the gas phase (1)
Peptide Fragmentation
In the Gas phase (2)
Ion structure
y-, b- fragment ion series

\[ \text{AA}_1 \rightarrow \text{AA}_2 \rightarrow \text{AA}_3 \rightarrow \text{AA}_4 \rightarrow \text{AA}_5 \rightarrow \text{OH} \]
ESI- MS/MS-capable instruments used in proteomics

Q: quadrupole
IT: ion trap
LIT: linear ion trap
TOF: time-of-flight
OT: orbitrap

ESI- MS/MS-capable instruments used in proteomics:

- **QQQ**
- **3D-IT**
- **LIT**
- **QQ-TOF**
- **LIT-OT**
Experimental set-up: nanoLC-MS/MS

- **HPLC pump**
- **T-splitter**: > 99% (waste), < 1% sample
- **C_{18} Column**: L = 15-25 cm, ID = 75 µm
- **Mass spectrometer**: 200-300 nl/min
LC-MS/MS data acquisition steps - 1

1. Determination of peptide mass

Chromatographic separation of peptides (time)

K.LVNELTEFAK.T

Measured: 1162.6253 Da
Calculated: 1162.6234 Da
2. Isolation, fragmentation, fragment analysis

Ion isolation
Collision Induced Dissociation (CID)
Data Dependent Acquisition (DDA)

Peptide LC elution

Survey scan
Detect precursors

MS/MS of 1
Exclude 1

MS/MS of 2
• Protein identification by MS/MS
MS/MS-based protein identification: concept

**Experimental**

Protein sample → Protein fragments (5-30 AA peptides) → MS → Exact masses of peptides → Fragmentation (MS/MS) spectrum of each peptide

**In silico**

Protein sequence(s) → Protein fragment sequences (same protease specificity) → software → Calculated exact masses of peptides → Calculated fragmentation spectrum of each peptide → software → Best Match(es)
Matching a peptide: F A D L S E A A N R

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>148.1</td>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td>1093.5</td>
<td>547.3</td>
<td>1076.5</td>
<td>1075.5</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>219.1</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>946.5</td>
<td>473.7</td>
<td>929.4</td>
<td>928.5</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>334.1</td>
<td></td>
<td>316.1</td>
<td>D</td>
<td></td>
<td>875.4</td>
<td>438.2</td>
<td>858.4</td>
<td>857.4</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>447.2</td>
<td></td>
<td>429.2</td>
<td>L</td>
<td></td>
<td>760.4</td>
<td>380.7</td>
<td>743.4</td>
<td>742.4</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>534.3</td>
<td>516.3</td>
<td>S</td>
<td></td>
<td></td>
<td>647.3</td>
<td>324.2</td>
<td>630.3</td>
<td>629.3</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>663.3</td>
<td>332.2</td>
<td></td>
<td></td>
<td>E</td>
<td>560.3</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>734.3</td>
<td>367.7</td>
<td>716.3</td>
<td>A</td>
<td>431.2</td>
<td>414.2</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>805.4</td>
<td>403.2</td>
<td>787.4</td>
<td>A</td>
<td>360.2</td>
<td>343.2</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>919.4</td>
<td>460.2</td>
<td>902.4</td>
<td>N</td>
<td>289.2</td>
<td>272.1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>1093.5</td>
<td>547.3</td>
<td>1076.5</td>
<td>R</td>
<td>175.1</td>
<td>158.1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Mascot Score Histogram
Ions score is \(-10\times \log(P)\), where \(P\) is the probability that the observed match is a random event. Individual ions scores > 34 indicate identity or extensive homology (p<0.05). Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.

« Threshold »
Mascot peptide (ion) score

- All steps of score calculation are not known (proprietary algorithm)
- Based on MOWSE algorithm
- Measures degree of matching of MS/MS spectrum vs. theoretical spectrum (# matched fragments, % of matched fragments,…)
- Also takes into account natural distribution of masses and thus « uniqueness » of a peptide mass in database

**LIMITATIONS**

- Not corrected for multiple testing problem
- Bias against small peptides

More in-depth information:
For those interested, a more technical description of the calculation of the MASCOT score is given in:
## Mascot output

<table>
<thead>
<tr>
<th>Query</th>
<th>Observed</th>
<th>Mr(expt)</th>
<th>Mr(calc)</th>
<th>Delta</th>
<th>Miss</th>
<th>Score</th>
<th>Expect</th>
<th>Rank</th>
<th>Unique</th>
<th>Peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>627.7522</td>
<td>1253.4899</td>
<td>1253.5598</td>
<td>-0.0699</td>
<td>0</td>
<td>80</td>
<td>1e-06</td>
<td>1</td>
<td>U</td>
<td>R.LGDLYEEEMR.E</td>
</tr>
<tr>
<td>86</td>
<td>662.2811</td>
<td>1322.5476</td>
<td>1322.6102</td>
<td>-0.0626</td>
<td>0</td>
<td>86</td>
<td>2.4e-07</td>
<td>1</td>
<td>U</td>
<td>R.EEAENTLQSFR.Q</td>
</tr>
<tr>
<td>95</td>
<td>714.8277</td>
<td>1427.6408</td>
<td>1427.7045</td>
<td>-0.0637</td>
<td>0</td>
<td>84</td>
<td>3.7e-07</td>
<td>1</td>
<td>U</td>
<td>R.SLYASSPGGVYATR.S</td>
</tr>
<tr>
<td>115</td>
<td>556.9347</td>
<td>1667.7823</td>
<td>1667.8366</td>
<td>-0.0543</td>
<td>0</td>
<td>27</td>
<td>0.17</td>
<td>1</td>
<td>U</td>
<td>R.ETNLDSLPLVDTHSK.R</td>
</tr>
<tr>
<td>117</td>
<td>578.9162</td>
<td>1733.7267</td>
<td>1733.8076</td>
<td>-0.0809</td>
<td>1</td>
<td>49</td>
<td>0.00081</td>
<td>1</td>
<td>U</td>
<td>R.LQDEIQNMEEMAR.H</td>
</tr>
</tbody>
</table>

Vimentin OS=Homo sapiens GN=VIM PE=1 SV=4
Check to include this hit in error tolerant search
Mascot Search Results

Peptide View

MS/MS Fragmentation of LGEYGFQNALIRV

Found in ALBU_BOVIN, Serum albumin O= Bos taurus G= ALB F= 1 SV= 4

Match to Query 2: 1478.828768 from(740.420660,2+) index(0)

Title: 181206_QS_hsal wiff, Sample 181206_QS_hsal (sample number 1), Elution 58.079 to 58.532 min, Period 1, Cycle(s): 5323-5325 (E)

Data file 461_740_only.mgf

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from [ ] to [ ] Da [ ] Full range
Label all possible matches [ ] Label matches used for scoring [ ]

All matches to this query

<table>
<thead>
<tr>
<th>Score</th>
<th>Mr(calc)</th>
<th>Delta</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
<tr>
<td>1.26</td>
<td>1478.7821</td>
<td>0.0586</td>
<td>LGEYGFQNALIRV</td>
</tr>
</tbody>
</table>

Nonstatistic range of neutral peptide M(r) (calc): 1470.7681

Fixed modifications: Carbamidomethyl (C) (apply to specified residues or terminal only)

Ion Scores: 1.26 Peptide: 1.26-40

Matches: 1.26/10 fragment ions using 24 most intense peaks

---

26
Many solutions proposed for validation: manual, semi-manual, training sets, mass accuracy, **statistical validation**
A good hit  (Mascot score=69)

MS/MS Fragmentation of VMLAANIGTPK
Match to Query 525: 1113.616556 from(557.815554,2+)

Monoisotopic mass of neutral peptide Mr(calc): 1113.622
Fixed modifications: Carbamidomethyl (C)
Ions Score: 69  Expect: 3.8e-07
Matches (Bold Red): 17/90 fragment ions using 32 most intense peaks
A bad hit (Mascot score=13)

MS/MS Fragmentation of **VSIALSSHWINPR**
Found in **KLOT_MOUSE**, Klotho precursor - Mus musculus (Mouse)
Match to Query 2: 1478.838768 from(740.426660,2+)

- Monoisotopic Mr(calc): 1478.7994
- Fixed modifications: Carbamidomethyl (C)
- Ions Score: 13 Expect: 18
- Matches (Bold Red):
  10/110 fragment ions using 26 most intense peaks

---

**Strong unmatched peaks**

---

<table>
<thead>
<tr>
<th>#</th>
<th>b</th>
<th>b''</th>
<th>b''</th>
<th>b''</th>
<th>b''</th>
<th>Seq</th>
<th>y</th>
<th>y''</th>
<th>y''</th>
<th>y''</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.0757</td>
<td>50.5415</td>
<td></td>
<td></td>
<td></td>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>187.1077</td>
<td>94.0575</td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td>1380.7383</td>
<td>690.5728</td>
<td>1363.7117</td>
<td>682.3595</td>
<td>1362.7277</td>
</tr>
<tr>
<td>3</td>
<td>300.1918</td>
<td>150.5995</td>
<td></td>
<td></td>
<td></td>
<td>I</td>
<td>1293.7062</td>
<td>647.3568</td>
<td>1276.6797</td>
<td>638.8435</td>
<td>1275.6957</td>
</tr>
<tr>
<td>4</td>
<td>371.2280</td>
<td>186.1181</td>
<td></td>
<td></td>
<td></td>
<td>T</td>
<td>1180.6222</td>
<td>590.8147</td>
<td>1163.5956</td>
<td>582.2014</td>
<td>1162.6116</td>
</tr>
<tr>
<td>5</td>
<td>484.3130</td>
<td>242.6601</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>533.2183</td>
<td>177.1128</td>
<td>1156.6346</td>
<td>524.1228</td>
<td>1155.6444</td>
</tr>
<tr>
<td>6</td>
<td>484.3139</td>
<td>242.6601</td>
<td></td>
<td></td>
<td></td>
<td>E</td>
<td>1109.2831</td>
<td>325.2962</td>
<td>1092.2382</td>
<td>316.7829</td>
<td>1091.2543</td>
</tr>
<tr>
<td>7</td>
<td>571.3450</td>
<td>286.1761</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>553.3344</td>
<td>277.1708</td>
<td>996.6610</td>
<td>498.7541</td>
<td>979.4744</td>
</tr>
<tr>
<td>8</td>
<td>658.3770</td>
<td>329.6921</td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td>909.4690</td>
<td>455.2381</td>
<td>892.4424</td>
<td>446.7248</td>
<td>891.4584</td>
</tr>
<tr>
<td>9</td>
<td>715.4459</td>
<td>358.2216</td>
<td></td>
<td></td>
<td></td>
<td>H</td>
<td>777.4254</td>
<td>389.2163</td>
<td>822.4369</td>
<td>411.7221</td>
<td>805.4104</td>
</tr>
<tr>
<td>10</td>
<td>795.4359</td>
<td>398.2116</td>
<td></td>
<td></td>
<td></td>
<td>W</td>
<td>963.5047</td>
<td>482.2560</td>
<td>685.3768</td>
<td>343.1926</td>
<td>668.2515</td>
</tr>
<tr>
<td>11</td>
<td>1094.5993</td>
<td>547.8033</td>
<td></td>
<td></td>
<td></td>
<td>G</td>
<td>1076.5887</td>
<td>538.7930</td>
<td>1069.5827</td>
<td>520.1530</td>
<td>1062.5762</td>
</tr>
<tr>
<td>12</td>
<td>1048.6422</td>
<td>604.8248</td>
<td>1191.6157</td>
<td>596.3115</td>
<td>1196.6317</td>
<td>593.8195</td>
<td>N</td>
<td>386.2146</td>
<td>193.6110</td>
<td>369.1831</td>
<td>185.0977</td>
</tr>
<tr>
<td>13</td>
<td>1305.6950</td>
<td>653.3511</td>
<td>1289.6684</td>
<td>644.3379</td>
<td>1287.6944</td>
<td>644.3459</td>
<td>P</td>
<td>272.1717</td>
<td>136.5095</td>
<td>255.1452</td>
<td>128.0762</td>
</tr>
<tr>
<td>14</td>
<td>1355.7414</td>
<td>700.3658</td>
<td>1339.7184</td>
<td>691.3527</td>
<td>1337.7455</td>
<td>691.3605</td>
<td>R</td>
<td>175.1199</td>
<td>88.0631</td>
<td>158.0924</td>
<td>79.5498</td>
</tr>
</tbody>
</table>
**Protein inference problem**

! **We are not identifying proteins, but peptides!!**

1) Apply principle of parsimony (Occam’s razor): within a family, list protein sequence which can explain the most of the identified peptides.

2) To highlight the presence of a member of a protein family, at least one discriminating (unique) peptide must be present (what if it is a borderline hit?)

---

**Protein inference problem**

<table>
<thead>
<tr>
<th>Peptide A</th>
<th>Peptide B</th>
<th>Peptide C</th>
<th>Peptide D</th>
<th>Peptide E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein 1</td>
<td>Protein 2</td>
<td>Protein 3</td>
<td>Protein 4</td>
<td></td>
</tr>
</tbody>
</table>

- Both hits are reported
- Reported
- Not reported!

_**Worst case**: two distinct homologous members of the same family found in two samples, each with a weak discriminating peptide….what to say?_
Further processing of Mascot results

(shotgun experiment, one or more samples)
Data analysis and distribution software: Scaffold

<table>
<thead>
<tr>
<th>Protein ID probability</th>
<th>Percentage of total spectra</th>
<th>Nr. assigned spectra</th>
<th>Nr. unique peptides</th>
<th>Nr. unique spectra</th>
<th>Unweighted spectrum count</th>
<th>Quantitative value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Name</td>
<td>Accession Number</td>
<td>Molecular Weight</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneous nuclear 1</td>
<td>ROA2_HUMAN</td>
<td>37 kDa</td>
<td>11</td>
<td>14</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>ribonucleo...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tubulin, beta polypeptide</td>
<td>Q5R532_HUMAN</td>
<td>46 kDa</td>
<td>12</td>
<td>12</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Keratin, type I cytoskeletal 19 (Cyt...</td>
<td>K1C10_HUM...</td>
<td>50 kDa</td>
<td>16</td>
<td>14</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>Keratin, type II cytoskeletal 1 (Cyto...</td>
<td>K1C6_HUMAN...</td>
<td>64 kDa</td>
<td>14</td>
<td>16</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>Alpha tubulin</td>
<td>Q5R532_HUMAN</td>
<td>50 kDa</td>
<td>8</td>
<td>11</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>Alpha tubulin</td>
<td>Q5R532_HUMAN</td>
<td>50 kDa</td>
<td>8</td>
<td>11</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>Actin, cytoplasmic 1 (Beta-actin)</td>
<td>ACTB_HUM...</td>
<td>42 kDa</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Cullin-3 (CUL-3)</td>
<td>Q5R532_HUMAN</td>
<td>32 kDa</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Kelch-like protein 13 DBT and kelhec...</td>
<td>KLH110_HUMAN</td>
<td>66 kDa</td>
<td>11</td>
<td>21</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Heat shock cognato 71 kDa protein...</td>
<td>HSF7C_HUM...</td>
<td>71 kDa</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>Protein L-asparagin (L-asparagin...</td>
<td>PMF_HUM...</td>
<td>25 kDa</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Kelch-like protein 32</td>
<td>KIK22_HUMAN</td>
<td>72 kDa</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>NIV/E protein hypothetic...</td>
<td>ORG1B2_HUM...</td>
<td>51 kDa</td>
<td>14</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keratin, type I cytoskeletal 9 (Cyt...</td>
<td>K1C9_HUM...</td>
<td>52 kDa</td>
<td>7</td>
<td>12</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Keratin, type II cytoskeletal 9 (Cyo...</td>
<td>K1C9_HUM...</td>
<td>52 kDa</td>
<td>7</td>
<td>12</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Scaffold attachment factor A2</td>
<td>CHMM1_HUM...</td>
<td>85 kDa</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Carboxy-methane synthase [co...</td>
<td>GSPM_HUM...</td>
<td>166 kDa</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneous nuclear ribonucleo...</td>
<td>HHNPD0_HUM...</td>
<td>38 kDa</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Keratin, type II cytoskeletal 2 opide...</td>
<td>K2E2_HUM...</td>
<td>66 kDa</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>TATA binding protein interacting pe...</td>
<td>G53Kh_HUM...</td>
<td>50 kDa</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Free Scaffold viewer: www.proteomesoftware.com
Spectral counting and quantification

- multiple spectra matches for abundant peptides
- spectral counting reflects relative abundance of proteins between samples
- poorly reliable at low spectral count

Spectral counting: Scaffold®

http://www.proteomesoftware.com/

Ex. Scaffold report

- Normalisation needed
- Oversampling necessary
- Accurate at medium/high spectral counts, unreliable at low spectral counts
- Samples must be reasonably similar

Probability of hitting a protein by unbiased sampling in MS/MS runs ~ concentration

Relative quantitation

Spiked myoglobin (ratio 1:2) in E.coli lysate

CAVEATS:

- Normalisation needed
- Oversampling necessary
- Accurate at medium/high spectral counts, unreliable at low spectral counts
- Samples must be reasonably similar

Improvements possible using:

- low-scoring spectra matched to confidently identified peptide sequences (Zhou et al. J Proteome Res. 2010 9:5698-704.)
- MS/MS TIC (Scaffold: average, total or TOP 3)
**Summary: Typical Analytical Workflow**

1. **Biological question**
2. Protease digestion
3. Peptide extraction
4. **Nano-HPLC**
5. **MS/MS**
6. **Chromatographic Separation (reversed-phase)**
7. **Tandem mass spectra of 50-2000 peptides**

**Database matches**
- DHX9_HUMAN
- NFM_HUMAN
- MYO6_HUMAN
- TP2A_PIG
- Q7Z5Y2
- FLIH_HUMAN
- TP2B_MOUSE
- S3B1_HUMAN
- Q8VCW5
- Q8CHF9

**Output:**
- Protein identification in simple/complex mixtures
- Extensive sequence coverage and peptide mapping
- Analysis of modified peptides possible

**Database searching**

**Software (MASCOT)**

**Protein sequence database**

**Table:**

- **Q7Z5Y2**
  - Mass: 118789
  - Total score: 178
  - Peptides matched: 6
  - Rho-interacting protein 3.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Mr(calc)</th>
<th>Score</th>
<th>Peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>930.48</td>
<td>42</td>
<td>EGLTVQER</td>
<td></td>
</tr>
<tr>
<td>1032.54</td>
<td>11</td>
<td>NWIQTIMK</td>
<td></td>
</tr>
<tr>
<td>1206.63</td>
<td>29</td>
<td>FSLCILTEPK</td>
<td></td>
</tr>
<tr>
<td>1369.75</td>
<td>24</td>
<td>LSTHELTSLEK</td>
<td></td>
</tr>
<tr>
<td>1406.77</td>
<td>55</td>
<td>FFILYEHGLLR</td>
<td></td>
</tr>
<tr>
<td>1775.88</td>
<td>16</td>
<td>QVPIAPVHLSSEDGGR</td>
<td></td>
</tr>
</tbody>
</table>

**Protease digestion**

**Peptide extraction**

**Nano-HPLC**

**MS/MS**

**Chromatographic Separation (reversed-phase)**

**Tandem mass spectra of 50-2000 peptides**
• Proteomics workflows and applications
Expression proteomics: analysis of protein expression levels and their changes

Typical questions:
* What distinguishes a lymphocyte from a neuron? 
* Which proteins are newly induced in a cell after a specific stimulus?

- Protein levels: main end product of gene activation, functionally active molecules
- Transcriptomics (cDNA, Affymetrix oligo chips, RNAseq,...) vs. proteomics
  - Comprehensive
  - Higher throughput, fast(er)
  - More sensitive
  - Assumption: [mRNA] ~ [protein]

2) **Functional proteomics**: studies on subsets of the proteome to map interactions and functional relationships

*Typical question:*

* which groups of proteins bind to each other and function together ?
* how does the protein composition of an organelle change in determined conditions

- Major proteomics application
- Protein-protein interactions $\Rightarrow$ active complexes (molecular machines)
- In vivo (keep PTMs, consider natural abundances, occurrences).
- Strategy: guilty by association
- Others: 2-hybrid screens (genetic), FACS, fluorescence
3) **Modification proteomics**: the analysis of post-translational modifications (PTMs).

*Typical question: how is protein activity modulated by covalent chemical modification?*

- Very numerous and varied PTMs
- Affect activity, targeting, degradation, …
- Combinatorial
- Dynamic
- Heterogeneous
- Various stoichiometry
- Proteomics and particularly MS are the sole “universal” techniques to study them
• Classical proteomics:
  – 2D-PAGE and MS (MALDI-TOF)

• More and more:
  LC-MS/MS-based “shotgun” approaches for large scale protein identification and quantification
  – Compared to 2D-PAGE: greater dynamic range and higher proteome coverage
  – Can be combined with isotope labelling to achieve relative protein quantification
WORKFLOW 1:

„classical“

2D-PAGE

+

Protein ID by MALDI-TOF-MS
Example: adaptation of bacteria to growth conditions

Normal medium  Low Glucose

E.Coli adaptation to low glucose by modulation of 15 proteins


(+)  ID and quantification
PTM separated

(-)  Reproducibility: poor
Dynamic Range limited
PTM separated
WORKFLOWS 2:

General shotgun protein ID
Shotgun sequencing from complex mixtures

Multiprotein complex

Denaturation, Proteolytic digestion

Complex peptide mixture (1000-20000 species)

Nano rp-LC-MSMS

Db search

List of identified proteins
1. P45218
2. P21543
3. Q12588
4. P32651
5. Q01245
6. ....
A VERY complex mixture – direct analysis (no separation)
Mascot Search Results
User: MQ
Email: 
Search title: 151002_ACO_B4strep.wiff: Angelos frac B4 IP strept
MS data file: C:\DOCUME~1\paf\LOCALS~1\Temp\mas5D.tmp
Database: Sprot 4028 (114033 sequences; 41888693 residues)
Taxonomy: Mammalia (mammals) (23838 sequences)
Timestamp: 17 Oct 2002 at 08:09:30 GMT

Significant hits:

- ALBU_BOVIN (P02769) Serum albumin precursor (Allergen Bos d 6).
- DNM1_HUMAN (P26358) DNA (cytosine-5)-methyltransferase 1 (EC 2.1.1.37)
- AC15_HUMAN (P35251) Activator 1 140 kDa subunit (Replication factor C
- IF16_HUMAN (Q16666) Gamma-interferon-inducible protein Ifi-16 (Interferon stimulus response gene)
- K1CJ_HUMAN (P13645) Keratin, type I cytoskeletal 10 (Cytokeratin 10) (KRT10)
- K22E_HUMAN (P35908) Keratin, type II cytoskeletal 2 epidermal (Cytokeratin 2)
- ACF7_HUMAN (Q9UPN3) Actin cross-linking family protein 7 (Macrophin) (MPR)
- AC14_HUMAN (P35250) Activator 1 40 kDa subunit (Replication factor C 4)
- ALBU_FELCA (P49064) Serum albumin precursor (Allergen Fel d 2).
- AC15_MOUSE (P35601) Activator 1 140 kDa subunit (Replication factor C
- DYHC_MOUSE (Q9JHU4) Dynein heavy chain, cytosolic (DYHC) (Cytoplasmic dynein)
- AC12_HUMAN (P35249) Activator 1 37 kDa subunit (Replication factor C 3
- EF11_CRIGR (P20001) Elongation factor 1-alpha 1 (EF-1-alpha-1) (Elongation factor 1)
- RYR3_HUMAN (Q15413) Ryanodine receptor 3 (Brain-type ryanodine receptor)
- K2C1_HUMAN (P04264) Keratin, type II cytoskeletal 1 (Cytokeratin 1) (KRT1)
- PLE1_RAT (P30427) Plectin 1 (PLTN) (PCN).
- AHNK_HUMAN (Q09666) Neuroblast differentiation associated protein AHNA
- TRYP_PIG (P00761) Trypsin precursor (EC 3.4.21.4).
- ACF7_MOUSE (Q9QXZ0) Actin cross-linking family protein 7 (Microtubule cross-linking protein)
- CENF_HUMAN (P49454) CENP-F kinetochore protein (Centromere protein F)
- ALBU_HUMAN (P02768) Serum albumin precursor.
- PLE1_HUMAN (Q15149) Plectin 1 (PLTN) (PCN) (Hemidesmosomal protein 1)
- TRI4_HUMAN (Q15650) Activating signal cointegrator 1 (ASC-1) (Thyroid transcription factor)
- NF1_HUMAN (P21359) Neurofibromin (Neurofibromatosis-related protein N)
- NEBU_HUMAN (P20929) Nebulin

A VERY complex mixture still gives results, but...
....... how deep are we going?
Fractionation to reduce complexity (1)

GeLC-MS workflow

Complex mixture

Trypsin digestion

LC-MS/MS

Db search

Protein IDs
1. P45218
2. P21543
3. Q12588
4. P32651
5. Q01245

....

1545. Q34258
Ex: protein-protein interactions analysis by affinity purification

- Bait + -
- IgG HC 52 kDa
- IgG LC 24 kDa

Trypsin digestion
- A
- B
- C
- D
- E
- F

LC-MS/MS

db search

(+)
- AHNK
- ANXA2
- GNAI2
- 5NTD
- ACTG
- GBB1
- GNAI3
- CD44
- CALM
- HSP7C
- LYN
- STOM
- CD59
- ECHB
- RAB35
- PLEC1
- VPP1
- CAD13
- CLH1
- SNP23

(-)
- AHNK
- ANXA2
- GNAI2
- 5NTD
- ACTG
- GBB1
- GNAI3
- CD44
- CALM
- HSP7C
- LYN
- STOM
- CD59
- ECHB
- RAB35
- PLEC1
- VPP1
- CAD13
- CLH1
- SNP23
Why is SDS-PAGE such a good preparation method?

- Ideal interface to biology
- Analytical and micropreparative
- Robust
- Solid phase chemistry of proteins
- Easy, low-tech
- Removal of contaminants:
  - At the loading point
  - After migration during fix / staining steps

Disadvantages

- Protein digestion in gel: non quantitative
- Peptide sequence recovery: usually incomplete
- Whole protein recovery: poor
Fractionation to reduce complexity (2)

Peptide-based fractionation

Complex mixture

Trypsin digestion

Peptide fractionation

LC-MS/MS

Db search

Protein IDs
1. P45218
2. P21543
3. Q12588
4. P32651
5. Q01245
....
1545. Q34258

SCX LC or Peptide IEF
WORKFLOWS 3:

quantification strategies
Comparison:  A ↔ B

Which proteins change in amount and how much?

Applications:
- Healthy vs. diseased tissues
- Healthy vs. diseased body fluids
- Drug treated / untreated cells
- Stimulated / unstimulated cells
- Mutants / wt cells

......
Techniques for large scale quantitative proteomics

- **2D Electrophoresis**
  - Classical (Label-free)
  - DIGE (Labelling)

- **MS-based methods**
  - **Label-free methods**
    - MS-based
    - Spectral counting
  - **Labelling methods**
    - Chemical Labelling
      - ICAT
      - iTRAQ
      - ICPL
      - TMT
      - Di-ME
      - $^{18}$O
      - ... 
    - Metabolic Labelling
      - SILAC
      - ISIS
      - $^{15}$N
      - ... 

- quantification at protein level

- quantification at peptide level
Relative quantification by stable isotope labelling

Sample A
Light

Sample B
Heavy

mix
analyse

Labelling strategies:

• Chemical (side chains: C, K, N-term)
  
  ICAT, iTRAQ, ICPLP, ...

• Metabolic (K, R, all)
  
  SILAC, ...

• Enzymatic
  
  Trypsin + 18O, ...

Co-analyse
Eliminate analytical variability

$\Delta m$
SILAC experiment workflow

Data analysis software!

MaxQuant
SILAC peaks

\[ \Delta m = 3.0 \text{ Da} \]

Heat shock protein
HSP 90β
Chemical labelling: Isobaric Tags (iTRAQ)- multiplex quantification

**Control**

Harvest cells

Trypsin digestion

**Treated**

**↔ in vitro chemical labeling**

**↔ Same peptide from 4 samples has same mass (isobaric)**

**↔ quantification by tags in MS/MS spectra at fixed M/Z**

*Figure 1. The concept of iTRAQ™ Reagent chemistry (example of a 4-plex experiment) Each sample is labeled with one of the four iTRAQ Reagents and then pooled prior to MS analysis.*

Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents
How to label? Pros and cons

• **Metabolically** (during protein synthesis)
  → Incorporation of one or more labelled amino acid
    (+) “native” proteins
    (+) compatible w. purifications
    (+) accurate
    (-) need cultivatable organism
    (-) limited multiplexing (max. 3)

• **Chemically** (post protein synthesis)
  → “specific” chemical modification of AA side chain
    (+) any sample can be done
    (+) higher multiplex (iTRAQ max 8-plex)
    (-) side (or incomplete) reactions
    (-) separate purifications
    (-) less accurate
**Labeling**

▷ Analytical variability minimized

▷ Number of samples limited (2-8)

**Label free**

▷ Number of samples unlimited

▷ Simpler sample preparation

▷ Analytical variability

Computationally heavy (XIC)
Signal processing in MS quantification

## Quantitation summary

<table>
<thead>
<tr>
<th>Application</th>
<th>Multiplexing</th>
<th>Accuracy (process)</th>
<th>Quantitative proteome coverage</th>
<th>Linear dynamic range&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Ease of use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolic protein labeling</td>
<td>2-3</td>
<td>+++</td>
<td>++</td>
<td>1–2 logs</td>
<td>+</td>
</tr>
<tr>
<td>• Complex biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Cell culture systems only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical protein labeling (MS)</td>
<td>2-3</td>
<td>+++</td>
<td>++</td>
<td>1–2 logs</td>
<td>+</td>
</tr>
<tr>
<td>• Medium to complex biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical peptide labeling (MS)</td>
<td>2-3</td>
<td>++</td>
<td>++</td>
<td>2 logs</td>
<td>+</td>
</tr>
<tr>
<td>• Medium complexity biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical peptide labeling (MS/MS)</td>
<td>2-8</td>
<td>++</td>
<td>++</td>
<td>2 logs</td>
<td>+</td>
</tr>
<tr>
<td>• Medium complexity biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enzymatic labeling (MS)</td>
<td>2</td>
<td>++</td>
<td>++</td>
<td>1–2 logs</td>
<td>++</td>
</tr>
<tr>
<td>• Medium complexity biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spiked peptides</td>
<td>multiple</td>
<td>++</td>
<td>+</td>
<td>2 logs</td>
<td>++</td>
</tr>
<tr>
<td>• Medium complexity biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted analysis of few proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label free (ion intensity)</td>
<td>multiple</td>
<td>+</td>
<td>+++</td>
<td>2–3 logs</td>
<td>++</td>
</tr>
<tr>
<td>• Simple biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Whole proteome analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label free (spectrum counting)</td>
<td>multiple</td>
<td>+</td>
<td>+++</td>
<td>2–3 logs</td>
<td>+++</td>
</tr>
<tr>
<td>• Simple biochemical workflows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Whole proteome analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> In MRM mode, dynamic range may be extended to 4–5 logs

Adapted from:
WORKFLOWS 4: "Modificomics"

f.ex. Phosphoproteomics
Caveat

Protein identification

IS NOT

protein characterisation

Two peptides are enough to identify a protein but we are still identifying two peptides, not the entire protein

Highly similar sequences cannot be distinguished

For finding PTMs extensive sequence coverage is essential !!!
## Some common PTMs

<table>
<thead>
<tr>
<th>Modification</th>
<th>Δ Mass</th>
<th>Residue</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteolysis</td>
<td>Various</td>
<td>Any</td>
<td>PTM, artefact</td>
</tr>
<tr>
<td>Dehydration</td>
<td>- 18.0106</td>
<td>N, Q, S, T, Y</td>
<td>PTM, artefact</td>
</tr>
<tr>
<td>Glycosylation (N-, O-, simple/complex)</td>
<td>Various</td>
<td>N, S, T, (Q)</td>
<td>PTM</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>+ 79.9663</td>
<td>S, T, Y</td>
<td>PTM</td>
</tr>
<tr>
<td>Sulfonation</td>
<td>+ 79.9568</td>
<td>S, T, Y, C</td>
<td>PTM</td>
</tr>
<tr>
<td>Acetylation</td>
<td>+ 42.0106</td>
<td>N-term or K</td>
<td>PTM, derivative</td>
</tr>
<tr>
<td>Carboxydimethylation</td>
<td>+ 57.0215</td>
<td>C</td>
<td>Derivative</td>
</tr>
<tr>
<td>Methylation</td>
<td>+ 14.0156</td>
<td>K, R, D, E, ...</td>
<td>PTM, artefact</td>
</tr>
<tr>
<td>Ubiquitination (mono, di-, poly, K48, K63, ..)</td>
<td>Various / + 114.043</td>
<td>K</td>
<td>PTM</td>
</tr>
<tr>
<td>Sumoylation (SUMO-1, -2, -3)</td>
<td>Various</td>
<td>K</td>
<td>PTM</td>
</tr>
<tr>
<td>Oxidation</td>
<td>+ 15.9949</td>
<td>C, M, W</td>
<td>PTM, artefact</td>
</tr>
<tr>
<td>ADP-ribosylation</td>
<td>+ 541.0611</td>
<td>R, C, N, S, E</td>
<td>PTM</td>
</tr>
<tr>
<td>Myristoylation</td>
<td>+ 210.1984</td>
<td>N-term G, K, C</td>
<td>PTM</td>
</tr>
<tr>
<td>Palmitoylation</td>
<td>+ 238.2297</td>
<td>C, K, S, T, N-term</td>
<td>PTM</td>
</tr>
<tr>
<td>Prenylation (farnesyl-, geranylgeranyl-)</td>
<td>Various</td>
<td>CaaX (C-term)</td>
<td>PTM</td>
</tr>
<tr>
<td>Nitrosylation</td>
<td>+ 28.9902</td>
<td>C</td>
<td>PTM</td>
</tr>
</tbody>
</table>

….. Almost 100 known…..
PTM characterization by MS

M

\[ b_5 \]
\[ b_4 \]
\[ b_3 \]
\[ b_2 \]
\[ b_1 \]

H \[ AA_1 \] \[ AA_2 \] \[ AA_3 \] \[ AA_4 \] \[ AA_5 \] OH

\[ y_5 \]
\[ y_4 \]
\[ y_3 \]
\[ y_2 \]
\[ y_1 \]

M' = M + \Delta M

\[ b_5^* = b_5 + \Delta M \]
\[ b_4^* = b_4 + \Delta M \]
\[ b_3^* = b_3 + \Delta M \]

\[ y_5^* = y_5 + \Delta M \]
\[ y_4^* = y_4 + \Delta M \]
\[ y_3^* = y_3 + \Delta M \]
Common issues in PTM analysis

• Protein sequence coverage
  → can be increased by multi-enzyme digestion, linked to abundance issue

• Labile PTMs / MS suitability
  → enzyme inhibitors, PTM derivatisation, use of alternative MS fragmentation (for ex. ETD)

• Abundance
  → PTM enrichment

• Artefacts
  → appropriate sample preparation, control experiments

• Isobaric PTMs
  → high resolution MS, specific fragments (for ex. immonium ions)

• Unknown (untargeted) PTMs
  → error-tolerant search, blind search

• Localization
  → use of alternative MS fragmentation, localization algorithms

• Connectivity
  → middle-down / top-down analysis
Phosphoproteomics

- Kinase cascades ⇔ all aspects of cell regulation
- Functional assay of the proteome
Questions in phosphorylation analysis

• Is a protein of interest phosphorylated?

• Which proteins are phosphorylated in a cell (or in a precise pathway?)

• Mapping phosphorylation sites: exact residues

• Quantitation of changes in response to a stimulus

• Effect on physiological protein activity
Problems with phosphopeptide analysis

1) **Quantity** problem : abundance of the protein to analyse is often low and phosphorylation is substoichiometric, especially when purifying from in vivo. 
   → Scale up prep, P-peptide enrichment

2) **Binding** of phosphopeptides to metal and other surfaces : what is its real impact ? Probably sequence dependent
   → Inert HPLC systems ; injection with Phosphoric acid

3) Low ( or no ) **signal intensity** due to acidic nature : highly variable depending on seq.
   → Derivatisation

4) Are we digesting with the good **enzyme** ? Phosphorylated regions are sometimes (often?) in problematic regions of proteins : very acidic, K/R-poor sequences (example plectin)
   → Digestion with multiple proteases

5) **Bad fragmentation** due to neutral loss : highly variable depending on peptide sequence
   → Choice of MS instrument, neutral loss MS, MS^n, multi-stage activation
Phosphoproteomics by TiO$_2$ P-peptide enrichment

Trypsin

Raw digest

Peptide fractionation (SCX, IEF, …)

Phosphopeptide Enrichment (TiO$_2$ column)

LC-MS/MS (all fractions)

ID: P-proteins

P-sites

Variants / options

• SILAC label
• $\alpha$-P-Tyr protein enrichment
• Asp, Glu esterification
Metal affinity enrichment makes P-peptides detectable

MALDI-TOF analysis of a digested phosphoprotein
Multi-enzymatic strategy

- Protein sequence coverage can be improved by using different digestion enzymes:
  - Trypsin (K,R)
  - Chymotrypsin (F, L, W, Y)
  - Lys-C (K)
  - Gluc-C (D, E)
  - Arg-C (R)
  - Combination of 2 enzymes

Ex: POM1_SCHPO \((S.\ pombe,\ fission\ yeast)\)

**Semi-specific search, 4 missed cleavages allowed:**

1. SEQ: sequence covered with trypsin digestion
2. SEQ: additional sequence covered with chymotrypsin digestion
3. SEQ: additional sequence covered with Lys-C digestion
4. SEQ: additional sequence covered with Glu-C digestion

=> Total sequence coverage: 95.9 %

**S**: phosphosite found with trypsin

**S**: additional phosphosite found with chymotrypsin

**S**: additional phosphosite found with Lys-C

**SS**: ambiguous phosphosite localization

=> Total number of phosphosites: 41

```
MGYLQSQKAV SLGDENTDAL FKLH[SNRS] ANMFGIKSEL LN[SEL][AVG
SYSNDICP[QSSSU] AADT SPSTNASNTN IS[PEQEHKD ELF[NEP[PKG
V[SSMNDHAI TIIHSTCNGL LR[SFHDHYR Q[NSPRNSIH RLSNISIGNN
PIDFESSQON NE[SLTSS] HRTSSINSK S[FC][SLSYN R[SSP][DWQ
QNNGGHLSGV[ISISQDVSSV PLQSSVFS[NG NHAYHASMAP K[EGSW[RTN
IHSTSHPRAA SIGNR[S] GIPP V[PPTPNP[CH TDHQP[AN ISGSLT[SSS
APSVDNKNKP VSSDHNNT[ETS SQ[HPD][SR NPDPA[PKA VSQKTN[VDGH
RNHEAKHGNT VQNESKSQ[S] SNKEGRSSRG G[FT][LSFSR SS[RM][KKGSK
AKHEDAPDV[AIHPAYIAA[S] SYRNGK RTPTRK[SM Q[FINWFKPS
KERSSN[NSD S][EPVPVPL HITRSQV[S]E PEK[EE[ESV PPLPSNKDK
GIVPO[QST][YS] Y[PPKR[SD] ES[LOP][LSFA SS[NL][SEPFD RKVADLAMKA
INSKRINKLL DD[AKVMQ[LL DRACIT[Pr VR NT[VQ[INTA PLTEYEQ[DEI
NNYDNIFTG LRNVKR[RSA DENTSSNFG DDERG[DYKV LG[DIAYRE
VVD[FLKGS[GF QQVLR[CIDYE T[GLVALKII R[NKRFH[MQA LVE[TKI[QKI
REWDPLDEC[YS MVQXTDFHYF RDHLCVATEL LG[KNL[ELIK SNG[FPGLP[IV
VIKSTRQ[LI QCLTLLNEKH VIHCDLP[EN ILLCHPFK[S VKVIDFG[S[CC
FEGC[NT[QI QSRFHYR[PEV ILGMRGTP[ DCWSSLGC[I AEMTGFLF[PP
GENEQ[QLAC I[MEIFGPPDH SLIDKCRSRK V[FDSSG[KKR PFVSSG[VD[SR
RPFSKSLHQV LOCK[DSF[LS FISDCLKWD[ DER[QPOQA QHD[FTGKQD
VRRPA[APAR QKFARPPNIE L[APIP[LP[LPN LPMEYN[HTL P[PEPSN[QA
SNLV[S]DKF PN[TL]NLDYS IISDNGFLRK P[VEKSR
```
Unknown PTMs – mining unassigned MS/MS spectra

- A large proportion of MS/MS spectra are not assigned in proteomics samples:


- All possible unknown PTMs cannot be searched in the classical way (too large search space)

⇒ error-tolerant search (Mascot) can be used on proteins identified by a first pass search
  - The selected enzyme becomes semi-specific
  - The complete list of modifications is tested, serially
  - For a protein, the set of substitutions that can arise from single base substitutions is tested
  - Only one of the above is allowed per peptide.

- Other tools: MS-Alignment, InsPecT, MODa, …(Note: some tools use de novo sequencing)
**Unknown PTMs – error-tolerant search (Mascot)**

- Results of error-tolerant search must be interpreted with caution: many artefacts (PTM identity or position)!
- Many PTMs can be explained by sample preparation artefacts (oxidation, carbamylation, propionamide, ...)
- Ex (PTMs from error-tolerant search in italics):

<table>
<thead>
<tr>
<th>Glycerol-3-phosphate O-acyltransferase 2 OS=Saccharomyces cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>617 – 624</td>
</tr>
<tr>
<td>617 – 624</td>
</tr>
<tr>
<td>617 – 624</td>
</tr>
<tr>
<td>617 – 624</td>
</tr>
<tr>
<td>617 – 624</td>
</tr>
<tr>
<td>617 – 625</td>
</tr>
<tr>
<td>625 – 640</td>
</tr>
<tr>
<td>625 – 640</td>
</tr>
<tr>
<td>625 – 640</td>
</tr>
<tr>
<td>625 – 640</td>
</tr>
<tr>
<td>625 – 640</td>
</tr>
<tr>
<td>625 – 640</td>
</tr>
<tr>
<td>625 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>626 – 640</td>
</tr>
<tr>
<td>649 – 661</td>
</tr>
<tr>
<td>649 – 665</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>670 – 675</td>
</tr>
<tr>
<td>688 – 708</td>
</tr>
<tr>
<td>688 – 710</td>
</tr>
</tbody>
</table>
Key concepts

- Proteome: complexity, plasticity, dynamic range
- Proteomics: more challenging than genomics but direct access to cell functions
- 2D-PAGE: differential display, then MS for ID
- LC & MS: many workflows to ID and quantify proteomes to depths of 4000-5000 proteins

Applications:
- Expression analysis
- PTMs
- Protein-protein interactions
- Protein trafficking, compartments, tissues,…
- Biomarker discovery
- ….
Take home message-1

- Many new possibilities in large scale protein analysis

**PTM’s**

- PTM are one of the most exciting and difficult « new » fields
- Huge variety and complexity of PTMs; no general workflow exists

**Quantitative proteomics**

- Quantitation is now feasible on a significant fraction of the proteome
- Several methods available; data quality and throughput are variable. Choice is often based on the experimental system and design
Take home message-2

• Some choices crucial for success:
  - Biological question: what are we looking for?
  - Model system
  - Sample preparation (!)
    - Abundance of protein of interest
    - Complexity of mixture
    - Enrichment mechanism
  - Data analysis: *not sooooooooooooo easy!*
  - If we get results, can we interpret them?
  - If we get results, are they going to be useful?
Some good reviews -1


Some good articles to start reading


  Discusses limits and potential of proteomics and compares proteomics with genomics (incl. mRNA ⇔ protein correlation)


  Very good overview presenting examples of proteomics studies which had a significant impact on a field of biology


  Focus on MS technology and its applications
Some good reviews - 2


*Give an overview of proteomics techniques used for PTM characterization in cells*


*Review of databases and other tools useful for study of PTMs*


*General introduction to quantitative proteomics*


*Review and comparison of analytical techniques used in quantitative proteomics*