PROTEOME DISCOVERER
Workflow concept

- Data goes through the workflow
  - Spectra
  - Peptides
  - Quantitation
- A “Node” contains an operation
- An edge represents data flow
- The results are brought together in tables
  - Protein (group) table
  - Peptide (group) table
Two aspects of the analysis

- Identification
  - MS₁ + MS₂

- Quantification
  - MS₁
Identification
Decision tree based MS

precursor charge state

2+

<750 m/z

HCD

ETD

3+

<1000 m/z

HCD

ETD

4+

ETD FT

ETD

≥ 5+

ETD
How to analyze?

- Decision tree is implemented in acquisition software
- Data analysis is slightly different for each fragmentation / detector method
- Rebuild the decision tree in the workflow!
Overview of the steps

1. Select the file(s) to use
2. Select only the MS2 spectra
3. Choose the different fragmentation methods
4. Create a specific workflow for each fragmentation type
   1. Spectrum filtering
   2. Database search
5. PSM Validation, PTM scoring
Example dataset 1

- Human cancer cell-line with 3 treatments
- Fractionated with SCX
- Decision tree: HCD / ETD-FT / ETD
1. Select files to use
2. Extract the ms2 spectra
3. Select spectrum origins
Side note:
For ETD: precursor removal

1. Precursor Peaks
   - Remove Precursor Peak: True
   - Mass Window Offset: 4 Da

2. Charge Reduced Precursors
   - Remove Charge Reduced: True
   - Mass Window Offset: 2 Da

3. Neutral Losses From Charge Reduced Precursors
   - Remove Neutral Loss Peak: True
   - Mass Window Offset: 2 Da
   - Remove Only Known Mass: True
   - Maximum Neutral Loss Mass: 120 Da

4. FT Overtones
   - Remove Precursor Overton: True
   - Mass Window Offset: 0.5 Da
For all: size reduction
### Identifications: Mascot

#### 1. Input Data
- **Protein Database**: UniProt_Human_NoRev
- **Enzyme Name**: Trypsin
- **Maximum Missed Cleavage**: 2
- **Instrument**: ESI-QUAD
- **Taxonomy**: All entries

#### 2. Tolerances
- **Precursor Mass Tolerance**: 50 ppm
- **Fragment Mass Tolerance**: 0.05 Da
- **Use Average Precursor M**: False

#### 3. Modification Groups

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxidation (M)</td>
<td></td>
</tr>
<tr>
<td>Phospho_STY (STY)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl (K)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl (N-term)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(4) (K)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(4) (N-term)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(6)13C(2) (K)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(6)13C(2) (N-term)</td>
<td></td>
</tr>
</tbody>
</table>

#### 4. Dynamic Modifications

<table>
<thead>
<tr>
<th>Dynamic Modification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxidation (M)</td>
<td></td>
</tr>
<tr>
<td>Phospho_STY (STY)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl (K)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl (N-term)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(4) (K)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(4) (N-term)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(6)13C(2) (K)</td>
<td></td>
</tr>
<tr>
<td>Dimethyl:2H(6)13C(2) (N-term)</td>
<td></td>
</tr>
</tbody>
</table>

#### 5. Static Modifications

<table>
<thead>
<tr>
<th>Static Modification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbamidomethyl (C)</td>
<td></td>
</tr>
</tbody>
</table>
Proteomics Mass Spectrometry

Trypsin Digest
Single Stage MS
Tandem Mass Spectrometry (MS/MS)

Precursor selection
Tandem Mass Spectrometry (MS/MS)

Precursor selection + collision induced dissociation (CID)
How do search engines work?

- **Input**
  - Spectrums
  - Sequence database (proteins)
  - Search parameters

- Match spectrum to possible peptides
  - Score
A peptide database

- We now know a range where the peptide mass has to be in
- Which peptides confirm to this criterium?
  - Taken from protein database
  - Modifications taken into account
  - Miscleavages taken into account
Protein database

Digest, modify

Peptide database with modifications indexed by mass

Fixed modifications

Variable modifications

Allowed missed cleavages
Matching the spectra

- Now we’re able to match the spectrum
- For that, *theoretical spectra* are composed
- This is done with the information from the peptide database
Peptide database
With modifications
Indexed by mass

Range of possible peptide masses

Theoretical spectra

Compare and score

MS/MS-spectrum

Ranked list of peptides
How to build a theoretical spectrum

- Information is needed about
  - Amino acid masses
  - Modification masses
  - Methods of fragmentation
### Peptide Fragmentation

**Peptide:** S-G-F-L-E-E-D-E-L-K

<table>
<thead>
<tr>
<th>MW</th>
<th>ion</th>
<th>Peptide</th>
<th>ion</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>b₁</td>
<td>S</td>
<td>y₉</td>
<td>1080</td>
</tr>
<tr>
<td>145</td>
<td>b₂</td>
<td>SG</td>
<td>y₈</td>
<td>1022</td>
</tr>
<tr>
<td>292</td>
<td>b₃</td>
<td>SGF</td>
<td>y₇</td>
<td>875</td>
</tr>
<tr>
<td>405</td>
<td>b₄</td>
<td>SGFL</td>
<td>y₆</td>
<td>762</td>
</tr>
<tr>
<td>534</td>
<td>b₅</td>
<td>SGFLE</td>
<td>y₅</td>
<td>633</td>
</tr>
<tr>
<td>663</td>
<td>b₆</td>
<td>SGFLEE</td>
<td>y₄</td>
<td>504</td>
</tr>
<tr>
<td>778</td>
<td>b₇</td>
<td>SGFLEED</td>
<td>y₃</td>
<td>389</td>
</tr>
<tr>
<td>907</td>
<td>b₈</td>
<td>SGFLEEDE</td>
<td>y₂</td>
<td>260</td>
</tr>
<tr>
<td>1020</td>
<td>b₉</td>
<td>SGFLEEDEL</td>
<td>y₁</td>
<td>147</td>
</tr>
</tbody>
</table>
## Identifications: Mascot

### 1. Input Data
- **Protein Database**: UniProt_Human_NoRev
- **Enzyme Name**: Trypsin
- **Maximum Missed Cleavage**: 2
- **Instrument**: ESI-QUAD
- **Taxonomy**: All entries

### 2. Tolerances
- **Precursor Mass Tolerance**: 50 ppm
- **Fragment Mass Tolerance**: 0.05 Da
- **Use Average Precursor Mass**: False

### 3. Modification Groups
- **From Quan Method**

### 4. Dynamic Modifications
1. Dynamic Modification: Oxidation (M)
2. Dynamic Modification: Phospho_STY (STY)
3. Dynamic Modification: Dimethyl (K)
4. Dynamic Modification: Dimethyl (N-term)
5. Dynamic Modification: Dimethyl:2H(4) (K)
6. Dynamic Modification: Dimethyl:2H(4) (N-term)
7. Dynamic Modification: Dimethyl:2H(5)13C(2) (K)
8. Dynamic Modification: Dimethyl:2H(5)13C(2) (N-term)
9. Dynamic Modification: Dimethyl:2H(6)13C(2) (K)
10. Dynamic Modification: Dimethyl:2H(6)13C(2) (N-term)

### 5. Static Modifications
1. Static Modification: Carbamidomethyl (C)
2. Static Modification: COOCH2CH3
3. Static Modification: Carbamidomethyl (C)
4. Static Modification: Carbamidomethyl (C)
5. Static Modification: Carbamidomethyl (C)
6. Static Modification: Carbamidomethyl (C)
PSM Thresholds
PSM thresholds

- Several options:
  - Target/decoy databases
  - Percolator
False Discovery Rate

- This is a number indicating how many of your hits are ‘false’
- If the FDR is 0.01 (1%), we expect 1 out of 100 peptide hits to be a false identification
- The number is estimated by the number of hits in the decoy database search result:
  \[\text{FDR} = \frac{\text{‘decoy hits’}}{\text{‘target hits’}}\]
Decoy hits

- Too much “noise”
  - Mixed spectrum
  - Overall low signal
- MS\(_2\) peaks do match with a peptide
Two ways of FDR calculation

- Concatenated database
  - One search
  - A spectrum matches either one or the other ("Competitive")

- Separate decoy database (in Mascot)
  - Two searches
  - A spectrum may match both decoy or target seq ("non-competitive")

- FDR = decoy hits / total hits
Limitation: one dimension
Introducing Percolator

- Percolator (Lukas Käll, McCoss Lab 2007)
- Machine learning
- Multiple features
What is a support vector machine

- Classification algorithm
- Best hyperplane between two groups
What is a support vector machine

- Classification algorithm
- Best hyperplane between two groups
Soft edges for SVMs

Non-separable training sets
Use linear separation, but admit training errors.

Penalty of error: distance to hyperplane multiplied by error cost $C$.  

Separating Hyperplane
Using several features

- Score
- Mass delta (ppm)
- Delta score
- Number of matched ions
- ...

Higher dimensions

Separation may be easier in higher dimensions

complex in low dimensions  simple in higher dimensions
PTM scoring

- PhosphoRS node
- Parallel to percolator node

- PhosphoRS version 2 names change
  - PhosphoRS score ➔

- Mao will explain in the afternoon
Quantification
Labeled Quantification

- Peptide modifications
- Extracted Ion Chromatrogram (XIC)
- Node to extract the intensities
How to interpret the XICs

- The XICs need to be grouped
  - Isotopes
  - Known modification (e.g. dimethylation, SILAC)
- The node needs details
  - Quantification method
Manage quant methods
PD Output
Understanding PD output

- PD run log
- Node result columns
  - Protein output
  - Peptide output

NB: the “modified peptide” column comes from the Mascot node, whilst the phosphoRS site localizations are from that node
  - Modified peptide does not represent “best” localizations
“Peptides” tab

- List Peptides
- Grouping (right click)
  - Grouped
  - Ungrouped (PTMs)
  - Grouping parameters are in the “Result Filters” tab
    - Based on sequence
    - Based on sequence and mass (i.e. modifications)
- PD will filter the PTMs based on the settings in the “Result Filters” tab
Peptide output

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Peptides</th>
<th>Search Input</th>
<th>Result Filters</th>
<th>Peptide Confidence</th>
<th>Search Summary</th>
<th>Quantification Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“Protein” tab

- List proteins
- Grouping (right click)
  - Grouped
  - Ungrouped
- Grouping based on shared peptides
What’s Grouped?

- Grouping allows for the calculation of standard deviations
- Show the highest scores
All PSMs

Step 1
Collect PSMs meeting criteria specified for protein grouping

PSMs relevant to protein grouping

Step 2
Group all proteins that share the same set or subset of identified peptides.

Preliminary protein groups

Step 3
Filter out protein groups that have no unique peptides among the considered peptides.

Step 4
Iterate through all spectra and select which PSM to use in ambiguous cases.

Step 5
Resolve cases where protein groups form circular rings of identified peptides.

Final protein groups

Steps 3–5 are performed only if you select the Apply Strict Maximum Parsimony Principle option in the Protein Grouping area of the Result Filters page.
Filters
Example dataset 2

- Yeast time course experiment
- Labeled with TMT 6-plex
- Fractionated with SCX
... build this!
Alternative TMT analysis

- Isobar package
- Nice
  - Statistical analysis
  - Nice output
  - Freely available
- Disadvantages
  - Written in R
  - Somewhat hard to use
  \(\rightarrow\) Ask Bas