Tutorial for tutors
Bioinformatics for MS analysis

MS identification

Patricia M. Palagi, PIG, SIB, Geneva
Part One:
Understanding MS data
Data Acquisition

1. Acquire full (MS) scan

2. Select an ion

3. Isolate ion

4. Fragment ion

MS/MS scan

Source: Peter James
Raw data: how it looks like

\[ \text{spectrum} = \text{baseline} + \text{peptide_signal} + \text{noise} \]

\[ I(m) = B(m) + S(m) + N(m) \]

Source: Markus Müller
What is a baseline?

- It is an offset of the intensities of masses, which happens mainly at low masses, and varies between different spectra.
- It results mainly from molecules of the matrix.
- The baseline has to be subtracted from the spectrum before any further processing takes place.

Example from MALDI-TOF

Correcting the baseline

Image source: MS Facility @ OCI HD
What is noise?

- Rapidly varying part of spectrum (< 1 Dalton)
- Mass spectra are affected by two types of noise:
  - electrical, a result of the instrument used
  - chemical, a result of contaminants in the sample or matrix molecules
- Noise can only be described by means of statistics.
Isotopes and Natural Abundance

Any given element has one or more naturally occurring isotopes:

- **Carbon-**
  - 98.89% $^{12}\text{C}$ (6 p + 6 n + 6 e, mass=12.000000 a.m.u.)
  - 1.11% $^{13}\text{C}$ (6 p + 7 n + 6 e, mass=13.003354 a.m.u.)
  - Very small* $^{14}\text{C}$ (6 p + 8 n + 6 e, mass=14.003241 a.m.u.)
  - *The average mass is then 12.01115 (in the periodic table)*

- **Oxygen-** $^{16}\text{O}$ (99.8%), $^{17}\text{O}$ (0.04%), $^{18}\text{O}$ (0.2%)

- **Sulphur-** $^{32}\text{S}$ (95.0%), $^{33}\text{S}$ (0.8), $^{34}\text{S}$ (4.2%)

- **Bromine-** $^{79}\text{Br}$ (50.5%), $^{81}\text{Br}$ (49.5%)

*1 part per trillion
*amu=atomic mass units
What is Mass?

- **Mass** is given as m/z which is the mass of the molecule divided by its charge.

- **Monoisotopic mass** is the mass of a molecule for a given empirical formula calculated using the exact mass of the most abundant isotope of each element (C=12.00000, H=1.007825 etc).

- **Average mass** is the mass of a molecule for a given empirical formula calculated using the weighted average mass for each element weighted by the isotopic abundances (C=12.01115, H=1.00797 etc).
Mono and average: example

C48 H82 N16  O17

Monoisotopic mass = 1155.6

Average mass = 1156.3
Mass Spectrometer Resolution

Is the power of separation of two ions of similar mass, defined as: \[ R = \frac{m_1}{\Delta m} \]

For example:

- if the spectrometer has a resolution of 1000, it can resolve a mass-to-charge ratio (m/z) of 1, i.e. distinguishes m/z 1000 from m/z 1001 (called unit resolution)

- TOF analyser (high mass resolution, reflectron mode): resolves at \(10^4\), i.e. distinguishes m/z 1000 from m/z 1000,1.
- Quadrupole: unit mass resolution \(10^3\)
- Ion Traps \(10^3\) to \(10^6\) (FTICR)
Mass Accuracy

- Accurate: consistent with the theory
- Mass accuracy is measured in *parts per million* (ppm)
  \[
  \text{Mass accuracy} = \frac{(\text{measured mass} - \text{theoretical mass})}{\text{true mass}} \times 10^6
  \]
- Measured mass
  - 545.4200
- Theoretical (calculated) mass
  - 545.3234
- Mass accuracy = \((545.4200-545.3234)/545 = 0.00011724 \text{ Da}\)
  \[
  = 0.00011724 \times 10^6 = 117 \text{ ppm}
  \]

A monoisotopic mass can be measured as accurately as the instrument allows, as long as the monoisotopic peak has been correctly identified. If the wrong peak is selected, the mass value will be false by one or two Daltons.

Adapted from: Peter James
High Mass Accuracy Measurements

3 different compounds
Same nominal mass
Low resolution

3 different compounds
3 different exact masses
High resolution, high accuracy

Source: Peter James
Determining charge states

Singly charged Ion: Distance between Peak and Isotope 1 Da

$\Delta = 1.0$ Da

524.3

525.3

526.3

Source: ?
Doubly charged Ion: Distance between Peak and Isotope 0.5 Da

\[ \Delta = 0.5 \text{ Da} \]

Source: ?
Electrospray Ionisation Principles – multiple charges

Source: Peter James
Part Two: Bioinformatics tools
Raw file formats

- Heterogeneity: each constructor, even each instrument has its own file format
- Different information and type of data across formats
- Format often not known
- Libraries for reading formats often not available
Processing spectral data

• Peak detection
  – Baseline removal
  – Discriminate signal from noise
  – Precise determination of monoisotopic masses
  – Separation of overlapping peaks
  – Charge state deconvolution
The processed data: list of m/z values

<table>
<thead>
<tr>
<th>Peptide mass values and intensities</th>
<th>Fragment mass values</th>
<th>MS/MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>840.6950 13.75</td>
<td>1163.7008 2</td>
<td></td>
</tr>
<tr>
<td>1676.9606 26.1</td>
<td>86.1105 220.1429</td>
<td></td>
</tr>
<tr>
<td>1498.8283 128.9</td>
<td>86.1738 13.7619</td>
<td></td>
</tr>
<tr>
<td>1045.564  845.2</td>
<td>102.0752 4.3810</td>
<td></td>
</tr>
<tr>
<td>2171.9670 2.56</td>
<td>147.1329 57.3333</td>
<td></td>
</tr>
<tr>
<td>861.1073  371.2</td>
<td>185.1851 649.0953</td>
<td></td>
</tr>
<tr>
<td>842.51458 53.7</td>
<td>185.3589 5.3810</td>
<td></td>
</tr>
<tr>
<td>1456.7274 12.9</td>
<td>186.1876 81.4286</td>
<td></td>
</tr>
<tr>
<td>863.268365 3.1</td>
<td>213.0791 1.4286</td>
<td></td>
</tr>
</tbody>
</table>

Parent mass value

Parent mass charge

fragment intensities

fragment intensities
Main file formats

• DTA (data)
• Company: Thermo Electron + Sequest
• Precursor mass unit: [M+H]+
• Precursor Intensity: no
• Format:
  – An original DTA file has a single spectrum
  – Concatenated dta files can have multiple spectra (accepted by various tools)

Example

1603.9204 2
101.0909 15.4762
202.1079 21.4762
203.1045 5.3333
244.1280 14.3810
254.1056 3.4286
255.1910 7.2381
270.2388 2.1905
...
962.6160 2
70.0560 2.1224
86.0947 5.1565
115.0842 8.4263
Main file formats

- **MGF (Mascot generic format)**
- **Company**: Matrix Science
- **Precursor mass unit**: \( m/z \)
- **Precursor Intensity**: optional
- **Format**:
  - Single or Multiple spectra

BEGIN IONS
TITLE=A1.1013.1013.2
CHARGE=2+
PEPMASS=715.940915
218.251 1.6
259.403 1.7
271.122 1.2
284.268 1.4
287.317 2.3
297.139 1.2
326.877 1.9
...
END IONS

BEGIN IONS
TITLE=A1.1013.1013.2
Precursor value in DTA and Mascot

- In a DTA file, the precursor peptide mass is an MH\(^+\) value independent of the charge state.

- In Mascot generic format, the precursor peptide mass is an observed m/z value, from which M\(_r\) or MH\(_n^{n+}\) is calculated using the prevailing charge state.

- For example, in Mascot:
  - PEPMASS=1000
  - CHARGE=2+

- ... means that the relative molecular mass M\(_r\) is 1998. This is equivalent to a DTA file which starts by:
  1999 2

Source: Matrix Science site
Tentative standard file formats

- **mzML**
- Merge of: **mzData** from HUPO PSI (Proteomic Standards Initiative) and **mzXML** from Institute for Systems Biology (Seattle)

Source: http://www.psidev.info/
Protein identification tools
One direct access to all: ExPASy

ExPASy Proteomics tools


The tools marked by ⚖ are local to the ExPASy server. The remaining tools are developed and hosted on other servers.

[Protein identification and characterization] [DNA -> Protein] [Similarity searches] [Pattern and profile searches] [Post-translational modification prediction] [Topology prediction] [Primary structure analysis] [Secondary structure prediction] [Tertiary structure] [Sequence alignment] [Phylogenetic analysis] [Biological text analysis]

<table>
<thead>
<tr>
<th>Identification and characterization with peptide mass fingerprinting data</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Aldeite - Identify proteins with peptide mass fingerprinting data. A new, fast and powerful tool that takes advantage of Hough transformation for spectra recalibration and outlier exclusion</td>
</tr>
<tr>
<td>- FindMod - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest</td>
</tr>
<tr>
<td>- FindPept - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artifactual chemical modifications, post-translational modifications (PTMs) and protease autolytic cleavage</td>
</tr>
<tr>
<td>- GlycoMod - Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses (can be used for free or derivatized oligosaccharides and for glycopeptides)</td>
</tr>
<tr>
<td>- Mascot - Peptide mass fingerprint from Matrix Science Ltd., London</td>
</tr>
<tr>
<td>- PopMAPPER - Peptide mass fingerprinting tool from UMIST, UK</td>
</tr>
<tr>
<td>- PFMLITS - Shows the possible single and double mutations of a peptide fragment from MALDI peptide mass fingerprinting</td>
</tr>
<tr>
<td>- ProFound - Search known protein sequences with peptide mass information from Rockefeller and NY Universities [or from Genomic Solutions]</td>
</tr>
<tr>
<td>- ProteinProspector - UCSF tools for peptide masses data (MS-Fit, MS-Pattern, MS-Digest, etc.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identification and characterization with MS/MS data</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Popitam - Identification and characterization tool for peptides with unexpected modifications (e.g., post-translational modifications or mutations) by tandem mass spectrometry</td>
</tr>
<tr>
<td>- Phenyx - Protein and peptide identification/characterization from MS/MS data from GenoBio, Switzerland</td>
</tr>
<tr>
<td>- Mascot - Sequence query and MS/MS ion search from Matrix Science Ltd., London</td>
</tr>
<tr>
<td>- OMSA - MS/MS peptide spectra identification by searching libraries of known protein sequences</td>
</tr>
<tr>
<td>- PopFrag - Search known protein sequences with peptide fragment mass information from Rockefeller and NY Universities [or from Genomic Solutions]</td>
</tr>
<tr>
<td>- ProteinProspector - UCSF tools for fragmentation mass data (MS-Tag, MS-Seq, MS-Product, etc.)</td>
</tr>
<tr>
<td>- SearchXLinks - Analysis of mass spectra of modified, cross-linked, and digested proteins whose amino acid sequence is known, from Caeser, Germany</td>
</tr>
</tbody>
</table>

Identification with isoelectric point, molecular weight and/or amino acid composition
Automatic protein identification

- Peptide mass fingerprinting – PMF
- Sequence TAG
- Peptide fragment fingerprinting - PFF
- de novo sequencing
Peptide mass fingerprinting = PMF
MS database matching

Protein(s) → Enzymatic digestion → Peptides → Mass spectra → Peaklist

Sequence database entry

In-silico digestion

Theoretical proteolytic peptides

Theoretical peaklist

Result: ranked list of protein candidates

...MAILAGGHSVRFGPKAF AEVNETFYSRVITLESTNM FNEIIIINSTNAQLATQFKNYPN VVIDDENHNDKGPLAGIYTI MKQHPEELFFVSVDTPM ITGKAVSTLYQFLV ...

- MAILAGGHSV
- FGPK
- AFAEVNETFYSR
- VITLESTNMFNEIIISTNAQLATQF
- YPNVVIDDENHNDK ...

840.695086
1676.96063
1498.8283
1045.564
2171.967066
861.107346
842.51458
1456.727405
863.268365

861.107346
838.695086
1676.96063
1498.8283
1045.564
2171.967066
861.107346
842.51458
1457.827405
863.268453
Peptide mass fingerprinting

What you have:
- Set of peptide mass values
- Information about the protein: molecular weight, pI, species.
- Information about the experimental conditions: mass spectrometer precision, calibration used, possibility of missed-cleavages, possible modifications
- Biological characteristics: post-translational modifications, fragments

What to do:
- Match between this information and a protein sequence database

What you get:
- a list of probable identified proteins
What is the expected information in a submission form?

- Place to upload a spectrum (many spectra)
- Description of the sample process used
  - Chemical process such as alkylation/reduction,
  - Cleavage properties (enzyme),
  - Mass tolerance (m/z tolerance)
- Search space
  - Sequence databank,
  - taxonomy restriction
  - Mw, pI restriction
- Scoring criteria and filters
Data about experimental conditions

- Accepted mass tolerance
  - due to imprecise measures and calibration problems

Table 1
Effect of Mass Accuracy and Mass Tolerance on Peptide Mass Fingerprinting Search Result

<table>
<thead>
<tr>
<th>Search m/z</th>
<th>Mass tolerance (Da)</th>
<th># Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1529</td>
<td>1</td>
<td>478</td>
</tr>
<tr>
<td>1529.7</td>
<td>0.1</td>
<td>164</td>
</tr>
<tr>
<td>1529.73</td>
<td>0.01</td>
<td>25</td>
</tr>
<tr>
<td>1529.734</td>
<td>0.001</td>
<td>4</td>
</tr>
<tr>
<td>1529.7348</td>
<td>0.0001</td>
<td>2</td>
</tr>
</tbody>
</table>

*Searches were done with the MS-FIT program at [http://prospector.ucsf.edu/](http://prospector.ucsf.edu/)

# Summary of PMF tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Source website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aldente</td>
<td><a href="http://www.expasy.org/cgi-bin/aldente">www.expasy.org/cgi-bin/aldente</a></td>
</tr>
<tr>
<td>Mascot</td>
<td><a href="http://www.matrixscience.com/">www.matrixscience.com/</a></td>
</tr>
<tr>
<td>MS-Fit</td>
<td>prospector.ucsf.edu/</td>
</tr>
<tr>
<td>ProFound</td>
<td>prowl.rockefeller.edu/profound_bin/WebProFound.exe</td>
</tr>
<tr>
<td>PepMAPPER</td>
<td>wolf.bms.umist.ac.uk/mapper/</td>
</tr>
<tr>
<td>PeptideSearch</td>
<td><a href="http://www.mann.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html">www.mann.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html</a></td>
</tr>
<tr>
<td>PepFrag</td>
<td>prowl.rockefeller.edu/prowl/pepfragch.html</td>
</tr>
</tbody>
</table>

Non exhaustive list!
Scoring systems

• Essential for the identification! Gives a confidence value to each matched protein

• Three types of scores

  • **Shared peaks count (SPC):** simply counts the number of matched mass values (peaks)

  • **Probabilistic scores:** confidence value depends on probabilistic models or statistic knowledge used during the match (obtained from the databases)

  • **Statistic-learning:** knowledge extraction from the influence of different properties used to match the proteins (obtained from the databases)
MS-Fit and Mowse*

- NCBInr and other databases, index of masses (many enzymes).
- Considers chemical and biological modifications.
- Statistic score which considers the mass frequencies.

Calculates the frequency of peptide masses in all protein masses for the whole database. The frequencies are then normalization.

The protein score is the inverse of the sum of the normalized frequencies of matched masses.
- The pFactor reduces the weight of masses with missed-cleavages in the frequency computation.

*MOlecular Weight SEarch
- Mowse score considers the peptide frequencies.
Mascot  
http://www.matrixscience.com/

• Internet free version in the above website (commercial versions available too)  
• Choice of several databases.  
• Considers multiple chemical modifications.  
• 0 to 9 missed-cleavages.  
• Score based on a combination of probabilistic and statistic approaches (is based on Mowse score).  
• Considers Swiss-Prot annotations for Splice Variants (only with locally installed versions).
Mascot - principles

- Probability-based scoring
- Computes the probability $P$ that a match is random
- Significance threshold $p < 0.05$ (accepting that the probability of the observed event occurring by chance is less than 5%)
- The significance of that result depends on the size of the database being searched.
- Mascot shades in green the insignificant hits
- Score: $-10\log_{10}(P)$
Decoy

Hints about the significance of the score

Proteins that share the same set of peptides

Probability Based Mowse Score

Protein score is \(-10 \log(P)\), where \(P\) is the probability that the observed match is a random event. Protein scores greater than 55 are significant (\(p<0.05\)).

Concise Protein Summary Report

1. **PDI13 HUMAN** Mass: 57246
   Score: 75
   Expect: 0.00052
   Queries matched: 14
   Protein disulfide-isomerase K2 precursor (EK5.3.7.1) (Disulfide isomerase ER-60) (ERp60) (58 kDa microsomal protein) (p58)

2. **UBQ1 HUMAN** Mass: 5012
   Score: 20
   Expect: 76
   Queries matched: 3
   Putative ubiquitin-conjugating enzyme K2 82-like protein (Ubiquitin-protein ligase D2-like) (Ubiquitin carrier protein D2-like)

3. **KRT1 HUMAN** Mass: 53671
   Score: 47
   Expect: 0.33
   Queries matched: 12
   Keratin, type II cytoskeletal 8 (Cytokeratin-8) (CK-8) (Keratin-8) (K8) - Homo sapiens (Human)
Matched peptides shown in Bold Red

1. ERLLELALFP GVALILAR LADALEVLEL TDDNPFSPIS D76SAGLMLV
2. EEPFGDDCHK KRPETQRA KTRQGTVPL AKYDCTATNN TKNGYIVS
3. PIILKATQE EAKSEDGEK AQDVCALQ RQQPSVLR TLEEKVFIS
4. DOAEVYFF LDOSEPHEH EFLGKAMED WDEEDFVSU SIYKVDFVAG
5. EKILIEPSL TLLNWIEKTV TTEVQKQDE KEGQTIQRL PKPSQEPNG
6. EQL10EGK LWEFDVTVE RRAHRATK YRHEPAAS ERDELEHR
7. VASRTSHE LEQCELST AEIIPVAK ERGECVYQK EPOCQEALE
8. FLLYFDYFRLKQILESPI FSPNDGFKV YAQEGDNL VRNKGKVLEK
9. FYAFYCOHEK LEPEXKLEK EKLSDPDEIV IAARDTAND WTPYVGRF
10. PTIVCRFPR KLPKETHSG HELSDEYSL QREAINFPVI QEEFFQGK

Show predicted peptides also

Cho Peptides By:  Residue Number  Increasing Mass  Decreasing Mass

<table>
<thead>
<tr>
<th>Start - End</th>
<th>Observed</th>
<th>Mr (exp)</th>
<th>Mr (calc)</th>
<th>Delta</th>
<th>Miss</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>63 - 73</td>
<td>1191.6060</td>
<td>1190.5987</td>
<td>1190.5931</td>
<td>0.0056</td>
<td>0</td>
<td>R.LAPEYKAAAR.L</td>
</tr>
<tr>
<td>108 - 119</td>
<td>1236.5289</td>
<td>1235.5207</td>
<td>1235.5054</td>
<td>0.0135</td>
<td>0</td>
<td>R.EDEWYGDAPR.T</td>
</tr>
<tr>
<td>131 - 140</td>
<td>993.1668</td>
<td>994.3527</td>
<td>994.3580</td>
<td>-0.0053</td>
<td>0</td>
<td>K.UGDPYFVLP.T</td>
</tr>
<tr>
<td>174 - 183</td>
<td>1179.5850</td>
<td>1178.5777</td>
<td>1178.5752</td>
<td>-0.0025</td>
<td>1</td>
<td>R.KASHLREGMR.E</td>
</tr>
<tr>
<td>259 - 271</td>
<td>1619.7919</td>
<td>1618.7837</td>
<td>1618.7766</td>
<td>-0.0051</td>
<td>0</td>
<td>K.LLIAKIPVDYKEK.H</td>
</tr>
<tr>
<td>287 - 304</td>
<td>878.4837</td>
<td>876.4837</td>
<td>876.4837</td>
<td>-0.0088</td>
<td>0</td>
<td>K.LHYDPWYK.X</td>
</tr>
<tr>
<td>306 - 326</td>
<td>2575.2500</td>
<td>2574.2427</td>
<td>2574.2366</td>
<td>-0.0059</td>
<td>0</td>
<td>K.EPSHELSDFQESTAEIPWAVK.T</td>
</tr>
<tr>
<td>336 - 344</td>
<td>1188.9568</td>
<td>1187.9567</td>
<td>1187.9581</td>
<td>0.0024</td>
<td>0</td>
<td>K.EVQDEEFSR.D Oxidation (M)</td>
</tr>
<tr>
<td>352 - 362</td>
<td>1399.6950</td>
<td>1398.6877</td>
<td>1398.6807</td>
<td>0.0311</td>
<td>0</td>
<td>R.FLLYFYGDLNL.R</td>
</tr>
<tr>
<td>367 - 379</td>
<td>1368.6680</td>
<td>1367.6527</td>
<td>1367.6527</td>
<td>-0.0041</td>
<td>0</td>
<td>K.SIFPESHEGPIPK.V</td>
</tr>
<tr>
<td>434 - 448</td>
<td>1680.7530</td>
<td>1679.7457</td>
<td>1679.7461</td>
<td>0.0004</td>
<td>0</td>
<td>K.RDAPADVLPVYQYR.L Oxidation (M)</td>
</tr>
<tr>
<td>449 - 460</td>
<td>1341.7030</td>
<td>1340.6957</td>
<td>1340.6765</td>
<td>0.0192</td>
<td>0</td>
<td>K.GEPYHPYKPJK.K</td>
</tr>
<tr>
<td>472 - 482</td>
<td>1370.7150</td>
<td>1369.7077</td>
<td>1369.6878</td>
<td>0.0208</td>
<td>0</td>
<td>R.FLSDFSTLQX.E</td>
</tr>
<tr>
<td>483 - 496</td>
<td>1579.9540</td>
<td>1578.9467</td>
<td>1578.8253</td>
<td>0.0234</td>
<td>0</td>
<td>R.EAKHINPVQOEKFK.K</td>
</tr>
</tbody>
</table>

No match to: 1045.5530, 1056.5050, 1075.5310, 1125.6930, 1129.6200, 1277.7260, 1293.6770, 1305.6820, 1344.6580, 1419.7650, 1462.20
Aldente

- SwissProt/TrEMBL db, indexed masses (trypsin and many others).
- Considers chemical modifications and user specified modifications.
- Considers biological modifications (annotations SWISS-PROT).
- 0 or 1 missed-cleavages.
- Use of robust alignment method (Hough transform):
  - Determines deviation function of spectrometer
  - Resolves ambiguities
  - Less sensitive to noise
Aldente – summary

- The Hough Transform estimates from the experimental data the deviation function of the mass spectrometer (the calibration error function).
- The program optimizes the set of best matches, excluding noise and outliers, to find the best alignment.
Aldente - Input

Enzyme: Trypsin
Missed cleavage: 1
Resolution: Monoisotopic
Ion mode: [M+H]

PTMs: Use PTM

Modifications:
- Carboxyamidomethyl (CAM)
- Add to list

Table:
<table>
<thead>
<tr>
<th>Label</th>
<th>Locus</th>
<th>Formula</th>
<th>Mode</th>
<th>Max</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM</td>
<td>C</td>
<td>C2H3ON</td>
<td>Fixed</td>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>MSO</td>
<td>M</td>
<td></td>
<td>Variable</td>
<td>2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Theoretical Peptides:
- Enzyme: Specify the enzyme (see list and cleavage rules) you used to generate your peptides.
- Missed cleavage: Select the number of missed cleavages allowed.
- Resolution: Specify the isotopic resolution of the experimental masses. The theoretical masses of the peptides will be calculated accordingly.
- Ion mode: Specify the charge state of the peptides: Protonated molecular ions, [M+H]^+.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Hits</th>
<th>AC</th>
<th>ID</th>
<th>DE</th>
<th>Mw</th>
<th>pI</th>
<th>Cov %</th>
<th>TaxId</th>
<th>Validate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>77.89</td>
<td>14</td>
<td>S30101</td>
<td>PDIA3_HUMAN (C_1)</td>
<td>Protein disulfide-isomerase A3</td>
<td>54</td>
<td>5.6</td>
<td>37</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>1</td>
<td>28.31</td>
<td>10</td>
<td>P05787</td>
<td>K2C8_HUMAN (C_1)</td>
<td>Keratin, type II cytoskeletal 8</td>
<td>54</td>
<td>5.5</td>
<td>24</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>2</td>
<td>5.92</td>
<td>5</td>
<td>Q5E298</td>
<td>TTY12_HUMAN (C_1)</td>
<td>Transcript Y 12 protein.</td>
<td>10</td>
<td>9.6</td>
<td>37</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>3</td>
<td>4.60</td>
<td>8</td>
<td>O60423</td>
<td>AT8B3_HUMAN (C_1)</td>
<td>Probable phospholipid-transporting ATPas...</td>
<td>148</td>
<td>8.0</td>
<td>8</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>4</td>
<td>4.11</td>
<td>8</td>
<td>Q4UYV3</td>
<td>TRIM1_HUMAN (C_1)</td>
<td>Microln-2.</td>
<td>81</td>
<td>7.2</td>
<td>12</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>5</td>
<td>4.07</td>
<td>7</td>
<td>Q0TDD1</td>
<td>DDX54_HUMAN (C_1)</td>
<td>ATP-dependence helicase DDX54.</td>
<td>99</td>
<td>10.0</td>
<td>11</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>6</td>
<td>3.86</td>
<td>4</td>
<td>Q9H0U3</td>
<td>IAG2_HUMAN (C_1)</td>
<td>Implantation-associated protein</td>
<td>35</td>
<td>9.7</td>
<td>24</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>7</td>
<td>3.59</td>
<td>9</td>
<td>Q15154</td>
<td>PCTM1_HUMAN (C_1)</td>
<td>Pericentriolar material 1 protein</td>
<td>229</td>
<td>5.0</td>
<td>5</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>8</td>
<td>3.47</td>
<td>6</td>
<td>Q52841</td>
<td>DDX17_HUMAN (C_1)</td>
<td>Probable ATP-dependent RNA helicase DDX17</td>
<td>72</td>
<td>8.8</td>
<td>13</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>9</td>
<td>3.41</td>
<td>7</td>
<td>Q6IA69</td>
<td>NDE1_HUMAN (C_1)</td>
<td>Glutamine-dependent NAD(+) synthetase.</td>
<td>79</td>
<td>6.0</td>
<td>10</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>10</td>
<td>3.29</td>
<td>7</td>
<td>P43920</td>
<td>CRK6_HUMAN (C_1)</td>
<td>G protein-coupled receptor kinase 6</td>
<td>66</td>
<td>8.3</td>
<td>14</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>11</td>
<td>3.23</td>
<td>7</td>
<td>Q0U7EB</td>
<td>PALD_HUMAN (C_1)</td>
<td>Paladin</td>
<td>97</td>
<td>6.1</td>
<td>7</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>12</td>
<td>3.21</td>
<td>8</td>
<td>P57679</td>
<td>EVC_HUMAN (C_1)</td>
<td>Ellis-van Creveld syndrome protein.</td>
<td>112</td>
<td>6.3</td>
<td>9</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>13</td>
<td>3.19</td>
<td>3</td>
<td>P39545</td>
<td>DUX5_HUMAN (C_1)</td>
<td>Double homeobox protein 5</td>
<td>22</td>
<td>10.7</td>
<td>31</td>
<td>9606</td>
<td>Validate</td>
</tr>
<tr>
<td>14</td>
<td>3.04</td>
<td>8</td>
<td>P82094</td>
<td>TMF1_HUMAN (C_1)</td>
<td>TATA element modulatory factor.</td>
<td>123</td>
<td>4.9</td>
<td>9</td>
<td>9606</td>
<td>Validate</td>
</tr>
</tbody>
</table>

Hints about the significance of the score:

- Score is greater than the best random score.
- Score is lower than the best random score.
Protein disulfide-isomerase A3.

Swiss-Prot annotation. Processed protein (signal peptide is cleaved).

Score: 77.89  Mw: 54264  pI: 5.51  Hits: 14  Coverage: 37%  Shift (Da): 0.045  Slope (ppm): -2.6
BioGraph

ALDentE results for: sample1_txt

Score | Matches | AC | ID
--- | --- | --- | ---
77.89 | 14 | P30101 | Pdia3_Human
52.31 | 10 | P05787 | K2C8_Human
4.69 | 8 | Q8INQ2 | ATBB3_Human
4.07 | 7 | Q8TCD1 | Ddx54_Human
3.59 | 9 | Q15154 | Pcm1_Human
3.47 | 6 | Q82841 | Ddx17_Human
3.41 | 7 | Q6LA69 | Nade1_Human
3.25 | 7 | Q9AR76 | Fald_Human
3.21 | 6 | P57679 | Evc_Human
3.04 | 6 | P52094 | Tmfl_Human
2.85 | 4 | Q9Y67D | Hes7_Human
2.76 | 7 | Q9LGE5 | Mti1_Human
2.64 | 6 | Q9P7N6 | Kef12_Human
2.51 | 5 | Q9M4N5 | Hs128_Human
2.45 | 7 | Q95271 | Tms1_Human

Graphical visualization of the results: BioGraph

Score is greater than the best random score
Score is lower than the best random score

P30101: Pdia3_Human 677.499 / LMFAYASR
What is the expected information in an identification result?

- A summary of the search parameters
- A list of potentially identified proteins (AC numbers) with scores and other evidences
- A detailed list of potentially identified peptides (associated or not to the potentially identified proteins) with scores
- Possibilities to validate/invalidate the provided results (info on the data processing, on the statistics, links to external resources, etc.)
- Possibilities to export the (validated) data in various formats
Hints to know when the identification is correct

With MS
- Good sequence coverage: the larger the sub-sequences and the higher the sequence coverage value, the better
- Consider the length of the protein versus the number of matched theoretical peptides
- Better when high intensity peaks have been used in the identification
- Scores: the higher, the better. The furthest from the 2nd hit the better
- Filter on the correct species if you know it (reduces the search space, time, and errors)
- Better when the errors are more or less constants among all peptides found.
- If you have time, try many tools and compare the results
Protein characterization with PMF data

1 protein entry does not represent 1 unique molecule

- Exact primary structure
- Splicing variants
- Sequence conflicts
- PTMs

Characterization tools at ExPASy using peptide mass fingerprinting data

Prediction tools
- PTMs and AA substitutions
- Oligosaccharide structures
- Unspecific cleavages
SWISS-PROT feature table: active protein is more than just translation of gene sequence (example: P20366)

<table>
<thead>
<tr>
<th>Key</th>
<th>From</th>
<th>To</th>
<th>Length</th>
<th>Description</th>
<th>FTId</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGNAL</td>
<td>1</td>
<td>19</td>
<td>19</td>
<td>Potential.</td>
<td></td>
</tr>
<tr>
<td>PROPEP</td>
<td>20</td>
<td>56</td>
<td>37</td>
<td>Potential.</td>
<td>PRO_000033529</td>
</tr>
<tr>
<td>PEPTIDE</td>
<td>58</td>
<td>68</td>
<td>11</td>
<td>Substance P.</td>
<td>PRO_000033530</td>
</tr>
<tr>
<td>PEPTIDE</td>
<td>72</td>
<td>107</td>
<td>36</td>
<td>Neuropeptide K.</td>
<td>PRO_000033531</td>
</tr>
<tr>
<td>PEPTIDE</td>
<td>72</td>
<td>73</td>
<td>2</td>
<td>Neuropeptide gamma, 1st part.</td>
<td>PRO_000033532</td>
</tr>
<tr>
<td>PEPTIDE</td>
<td>89</td>
<td>107</td>
<td>19</td>
<td>Neuropeptide gamma, 2nd part.</td>
<td>PRO_000033533</td>
</tr>
<tr>
<td>PEPTIDE</td>
<td>98</td>
<td>107</td>
<td>10</td>
<td>Neurokinin A.</td>
<td>PRO_000033534</td>
</tr>
<tr>
<td>PEPTIDE</td>
<td>111</td>
<td>126</td>
<td>16</td>
<td>C-terminal-flanking peptide.</td>
<td>PRO_000033535</td>
</tr>
<tr>
<td>MOD_RES</td>
<td>68</td>
<td>68</td>
<td>0</td>
<td>Methionine amide (G-69 provides amide group).</td>
<td></td>
</tr>
<tr>
<td>MOD_RES</td>
<td>107</td>
<td>107</td>
<td>0</td>
<td>Methionine amide (G-108 provides amide group).</td>
<td></td>
</tr>
<tr>
<td>VAR_SEQ</td>
<td>74</td>
<td>88</td>
<td>0</td>
<td>Missing (in isoform Gamma and isoform Delta).</td>
<td>VSP_006375</td>
</tr>
<tr>
<td>VAR_SEQ</td>
<td>97</td>
<td>114</td>
<td>0</td>
<td>Missing (in isoform Alpha and isoform Delta).</td>
<td>VSP_006376</td>
</tr>
<tr>
<td>VAR_SEQ</td>
<td>115</td>
<td>115</td>
<td>0</td>
<td>V -&gt; M (in isoform Alpha and isoform Delta).</td>
<td>VSP_006377</td>
</tr>
<tr>
<td>CONFLICT</td>
<td>32</td>
<td>32</td>
<td>0</td>
<td>S -&gt; P (in Ref. 3; BAD96677).</td>
<td></td>
</tr>
</tbody>
</table>
Detection of PTMs in MS

Unmodified tryptic masses

Tryptic masses of a modified protein

Δ m/z => PTM
FindMod


**FindMod tool**

FindMod is a tool that can predict potential protein post-translational modifications (PTM) and find potential single amino acid substitutions in peptides.

The experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot/TrEMBL entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest. [Documentation / Mass values and considered PTMs / Reference].

- **Swiss-Prot/TrEMBL ID or AC or user-entered sequence:**
- **DB entry**
- **AA modifications**
- **experim**
- **ental**
- **m**
- **as**
- **experim**
- **ental**
- **options**
- **(Supported formats):**
- **experimental**
- **masses**
- **PMP**

**Check:** all or unmatching peptides for the following:
- potential post-translational modifications (max. 1 or within one peptide).
- single amino acid substitutions.

If you wish to take into account other post-translational modifications than those already predictable by FindMod, you can specify them here.

**modification name**
**amino acids**
**this modification can be observed on**
**atom composition**

| 0 | example | ANF | H, 3, O, 1, C, 2, N, , S. |
| 1 |         |     | H, 3, O, 1, C, 2, N, , S. |
| 2 |         |     | H, 3, O, 1, C, 2, N, , S. |

All peptide masses are
- with cysteines treated with [nothing (in reduced form)]
- with acrylamide adducts on cysteines
- with methionines oxidized
- [M+H+], [M]+, [M+H2O]+, average or monoisotopic.

**Mass tolerance:** ± 0.5 Da

**Select an enzyme:** Trypsin

Allow for [0] missed cleavage sites [MC].

Display peptides sorted by [ ] peptide masses or in [ ] chronological order in the protein.

**Send the result by e-mail**

Your e-mail address:

name of the unknown protein: unknown

To run the search: Start FindMod
To clear all fields: Reset
### FindMod Output

<table>
<thead>
<tr>
<th>User mass</th>
<th>DB mass</th>
<th>Δ mass</th>
<th>#MC</th>
<th>peptide</th>
<th>position</th>
<th>known modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>965.51</td>
<td>965.506</td>
<td>-0.003</td>
<td>2</td>
<td>SKEKFER</td>
<td>1-7</td>
<td>(ACET: 1)</td>
</tr>
</tbody>
</table>

- **unmodified peptides**, modified peptides known in SWISS-PROT and chemically modified peptides

<table>
<thead>
<tr>
<th>User mass</th>
<th>DB mass</th>
<th>Δ mass</th>
<th>#MC</th>
<th>peptide</th>
<th>position</th>
<th>known modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1631.81</td>
<td>1603.772</td>
<td>28.038</td>
<td>28.031</td>
<td>DIMETH AFDQIDNAPEEKAR</td>
<td>45-58</td>
<td>(METH: 50)</td>
</tr>
</tbody>
</table>

- **putatively modified peptides predicted by mass differences**

<table>
<thead>
<tr>
<th>User mass</th>
<th>DB mass</th>
<th>Δ mass</th>
<th>#MC</th>
<th>peptide</th>
<th>position</th>
<th>known modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1053.35</td>
<td>1053.012</td>
<td>0.000</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note**: potential residues carrying modification.
Modification rules can be defined from SWISS-PROT, PROSITE and the literature

<table>
<thead>
<tr>
<th>modification</th>
<th>amino acid rule</th>
<th>exceptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>farnesylation</td>
<td>Cys</td>
<td>-</td>
</tr>
<tr>
<td>palmitoylation</td>
<td>Cys</td>
<td>Ser, Thr</td>
</tr>
<tr>
<td>O-GlcNAC</td>
<td>Ser, Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>amidation</td>
<td>Xaa (C-term)</td>
<td>where Gly</td>
</tr>
<tr>
<td>pyrrolidone carboxylic acid</td>
<td>Gln (N-term)</td>
<td>-</td>
</tr>
<tr>
<td>phosphorylation in eukaryotes:</td>
<td>Ser, Thr, Asp,</td>
<td>His, Tyr</td>
</tr>
<tr>
<td>in prokaryotes:</td>
<td>Ser, Thr, Asp,</td>
<td>His, Cys</td>
</tr>
<tr>
<td>sulfatation in eukaryotes</td>
<td>Tyr, PROSITE</td>
<td>PRODOC00003</td>
</tr>
</tbody>
</table>
FindMod Output - Application of Rules

- potentially modified peptides that agree with rules are listed
- amino acids that potentially carry modifications are shown

<table>
<thead>
<tr>
<th>User mass</th>
<th>DB mass</th>
<th>mass diff.</th>
<th>mod. diff.</th>
<th>Δmass</th>
<th>potential mod.</th>
<th>#MC</th>
<th>peptide</th>
<th>position</th>
<th>known modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1631.81</td>
<td>1603.771</td>
<td>28.039</td>
<td>28.031</td>
<td>-0.007</td>
<td>DIMETH</td>
<td>1</td>
<td>AFDQIDNAPEEKAR</td>
<td>45-58</td>
<td></td>
</tr>
<tr>
<td>1631.81</td>
<td>1617.787</td>
<td>14.023</td>
<td>14.016</td>
<td>-0.006</td>
<td>METH</td>
<td>1</td>
<td>AFDQIDNAPEEKAR</td>
<td>45-58</td>
<td>(METH: 56)</td>
</tr>
</tbody>
</table>

- peptides potentially modified only by mass difference

| Potential PTMs detected by mass differences, but not confirmed by rules: |
|-----------------------------|-------------------|-----------------|--------|----------------|-----|--------------------|----------|---------------------|
| 1631.81  | 1603.771  | 28.039  | 27.995  | -0.043| FORM           | 1   | AFDQIDNAPEEKAR     | 45-58    |                     |
| 1631.81  | 1632.736  | -0.925  | -0.983  | -0.057| AMID           | 1   | ETQKSTCTGVEMFR     | 249-262  | (1xMSO)             |

- predictions can be tested by MS-MS peptide fragmentation
MS/MS based identification tools

- **Tag search** - Tools that search peptides based on a MS/MS Sequence Tag
  - MS-Tag and MS-Seq, PeptideSearch

- **Ion search or PFF** - Tools that match MS/MS experimental spectra with “theoretical spectra” obtained via in-silico fragmentation of peptides generated from a sequence database
  - Phenyx, Mascot, Sequest, X!Tandem, OMSSA, ProID, ...

- **de novo sequencing** - Tools that directly interpret MS/MS spectra and try to deduce a sequence
  - Convolution/alignment (PEDENTA)
  - De-novo sequencing followed by sequence matching (Peaks, Lutefisk, Sherenga, PeptideSearch)
  - Guided Sequencing (Popitam)

In all cases, the output is a peptide sequence per MS/MS spectrum
Peptide fragmentation fingerprinting = PFF = ion search MS/MS database matching

Protein(s) -> Enzymatic digestion -> Peptides -> MS/MS spectra of peptides -> Ions peaklists

- MAILAGGHSGVR
- AFAGK
- VITNLNMNFNEIILK
- YPNVVIDDENNDK
...

In-silico digestion -> -MAILAG
- MAILA
- MAIL
- MAI
- M
- AIILAG

Sequence database entry

Theoretical proteolytic peptides

Theoretical fragmented peptides

Theoretical peaklist

Result: ranked list of peptide and protein candidates
Ion-types

<table>
<thead>
<tr>
<th>Ion Type</th>
<th>Ion Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>[N]+[M]-CO</td>
</tr>
<tr>
<td>a⁺</td>
<td>a-NH₃</td>
</tr>
<tr>
<td>a⁻</td>
<td>a-H₂O</td>
</tr>
<tr>
<td>a⁺⁺</td>
<td>(a+H)/2</td>
</tr>
<tr>
<td>b</td>
<td>[N]+[M]</td>
</tr>
<tr>
<td>b⁺</td>
<td>b-NH₃</td>
</tr>
<tr>
<td>b⁻</td>
<td>b-H₂O</td>
</tr>
<tr>
<td>b⁺⁺</td>
<td>(b+H)/2</td>
</tr>
<tr>
<td>c</td>
<td>[N]+[M]+NH₃</td>
</tr>
<tr>
<td>d</td>
<td>a-partial side chain</td>
</tr>
<tr>
<td>v</td>
<td>y-complete side chain</td>
</tr>
<tr>
<td>w</td>
<td>z-partial side chain</td>
</tr>
<tr>
<td>x</td>
<td>[C]+[M]+CO</td>
</tr>
<tr>
<td>y</td>
<td>[C]+[M]+H₂</td>
</tr>
<tr>
<td>y⁺</td>
<td>y-NH₃</td>
</tr>
<tr>
<td>y⁻</td>
<td>y-H₂O</td>
</tr>
<tr>
<td>y⁺⁺</td>
<td>(y+H)/2</td>
</tr>
<tr>
<td>z</td>
<td>[C]+[M]-NH</td>
</tr>
</tbody>
</table>

It is very important to know the ionic series produced by a spectrometer, otherwise potential matches will be missed.

Usually a, b and y

[N] is the mass of the N-term group
[M] is the mass of the sum of the neutral amino acid residue masses
Peptide fragmentation with MS/MS

MAPNCSCK

\[ [M+2H]^2+ \]

b7 MAPNCSC  Ky1
b6 MAPNCS  CKy2
b5 MAPNC  SCKy3
b4 MAPN  CSCKy4
...
...
...

MAPNCSCK

\[ [M+2H]^2+ \]

b7 MAPNCSC  Ky1
b6 MAPNCS  CKy2
b5 MAPNC  SCKy3
b4 MAPN  CSCKy4
...
...
...

There is a shift around the Serine (neutral loss – H3PO4), + other ions: a, c, x..., H2O and NH3 neutral losses, immoniums, internal fragments, ...missing ions
Instruments

Q-TOF
- easier to interpret since the spectra are less crowded (cleaner spectra)
- higher accuracy and resolution - which compensates for low frequency of a and b ions

3D Ion traps
- a, b and y ions - less easier to interpret but more significant since more ions to confirm the sequence (noisy spectra)
### Some PFF tools

Same principle of a PMF, but using MS/MS spectra

<table>
<thead>
<tr>
<th>Software</th>
<th>Source website</th>
</tr>
</thead>
<tbody>
<tr>
<td>InsPecT</td>
<td>peptide.ucsd.edu/inspect.py</td>
</tr>
<tr>
<td>Mascot</td>
<td><a href="http://www.matrixscience.com/search_form_select.html">www.matrixscience.com/search_form_select.html</a></td>
</tr>
<tr>
<td>MS-Tag and MS-Seq</td>
<td>prospector.ucsf.edu</td>
</tr>
<tr>
<td>PepFrag</td>
<td>prowl.rockefeller.edu/prowl/pepfragch.html</td>
</tr>
<tr>
<td>Phenyx</td>
<td>phenyx.vital-it.ch</td>
</tr>
<tr>
<td>Popitam</td>
<td><a href="http://www.expasy.org/tools/popitam">www.expasy.org/tools/popitam</a></td>
</tr>
<tr>
<td>ProID (download)</td>
<td>sashimi.sourceforge.net/software_mi.html</td>
</tr>
<tr>
<td>Sequest*</td>
<td>fields.scripps.edu/sequest/index.html</td>
</tr>
<tr>
<td>Sonar</td>
<td>65.219.84.5/service/prowl/sonar.html</td>
</tr>
<tr>
<td>SpectrumMill*</td>
<td><a href="http://www.home.agilent.com">www.home.agilent.com</a></td>
</tr>
<tr>
<td>VEMS</td>
<td><a href="http://www.bio.aau.dk/en/biotechnology/vems.htm">www.bio.aau.dk/en/biotechnology/vems.htm</a></td>
</tr>
<tr>
<td>X!Tandem (download)</td>
<td><a href="http://www.thegpm.org/TANDEM">www.thegpm.org/TANDEM</a></td>
</tr>
</tbody>
</table>

*Commercialized

Non exhaustive list!
## MASCOT MS/MS Ions Search

<table>
<thead>
<tr>
<th><strong>Your name</strong></th>
<th>Palagi</th>
<th><strong>Email</strong></th>
<th><a href="mailto:Patricia.Palagi@isb-sib.ch">Patricia.Palagi@isb-sib.ch</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Search title</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Database</strong></td>
<td>MSDB</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Taxonomy</strong></td>
<td>All entries</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Enzyme</strong></td>
<td>Trypsin</td>
<td><strong>Allow up to</strong></td>
<td>1 missed cleavages</td>
</tr>
<tr>
<td><strong>Fixed modifications</strong></td>
<td>Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term)</td>
<td><strong>Variable modifications</strong></td>
<td>Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term)</td>
</tr>
<tr>
<td><strong>Quantitation</strong></td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Peptide tol. ±</strong></td>
<td>1.2 Da #^{13}C 0</td>
<td><strong>MS/MS tol. ±</strong></td>
<td>0.6 Da</td>
</tr>
<tr>
<td><strong>Peptide charge</strong></td>
<td>2+</td>
<td><strong>Monoisotopic</strong></td>
<td>Average</td>
</tr>
<tr>
<td><strong>Data file</strong></td>
<td></td>
<td><strong>Precursor</strong></td>
<td>m/z</td>
</tr>
<tr>
<td><strong>Data format</strong></td>
<td>Mascot generic</td>
<td><strong>Error tolerant</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Instrument</strong></td>
<td>Default</td>
<td><strong>Report top</strong></td>
<td>AUTO hits</td>
</tr>
<tr>
<td><strong>Decoy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Start Search ...](#)  [Reset Form](#)
Sequest/Turbosequest output

```plaintext
al941203#2.0402.0405.2.out
SEQUEST v.22, Copyright 1993-95
Molecular Biotechnology, Univ. of Washington, J.Eng/J.Yates
Licensed to John Yates' Lab @ Univ. of Washington
11/15/95, 08:53 AM, 2 min. 32 sec. on thompson
mass=1472.0(±2), fragment tol.=0.00, mass tol.=3.00, ACG
# amino acids = 2904160, # proteins = 6254, # matched peptides = 151688
immonium (HPYWM) = (00000), total inten. = 6927.9, lowest Sp = 170.3
ion series nA nB nY ABCDUXWYZ: 0 1 1 0.5 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
rho=0.200, beta=0.075, top 10, /wfs/dbase/OWL/yeast

<table>
<thead>
<tr>
<th>Rank/Sp</th>
<th>(M+H)+</th>
<th>Cn</th>
<th>deltCn</th>
<th>C*10^4</th>
<th>Sp</th>
<th>Ions</th>
<th>Reference</th>
<th>Peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1471.7</td>
<td>1.0000</td>
<td>0.0000</td>
<td>3.8603</td>
<td>851.3</td>
<td>22/39</td>
<td>G3P1_YEAS+4 (R)UPTUDVSUVDLTVK</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1469.7</td>
<td>0.6042</td>
<td>0.3958</td>
<td>2.3323</td>
<td>381.5</td>
<td>15/39</td>
<td>S52527 (L)QAPPFPFSSTKSKF</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1472.9</td>
<td>0.5877</td>
<td>0.4123</td>
<td>2.2688</td>
<td>448.7</td>
<td>17/39</td>
<td>KEX1_YEAS (A)UUOTUULFIVULG</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1469.5</td>
<td>0.5572</td>
<td>0.4427</td>
<td>2.1515</td>
<td>378.5</td>
<td>17/39</td>
<td>CB31_YEAS (R)UPMTGDLSTGNTFE</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1471.8</td>
<td>0.5388</td>
<td>0.4644</td>
<td>2.0677</td>
<td>368.2</td>
<td>17/39</td>
<td>0DPA_YEAS (S)UKAULAEMLGRRAG</td>
</tr>
</tbody>
</table>

1. G3P1_YEAST GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE 1 (EC 1.2.1.12). - SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).
2. S52527 hypothetical protein - yeast (Saccharomyces cerevisiae)
3. KEX1_YEAST CARBOXYPEPTIDASE KEX1 PRECURSOR (EC 3.4.16.5) (CARBOXYPEPTIDASE D). - SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).
```
Phenyx Submission Page (TOFTOP_iTRAQ_std_Tryp)

**Profile**: TOFTOP_iTRAQ_std_Tryp (default, readonly)

**Title**

**Database(s)**
- NCBInr (version: 100510), type: AA, tax: Yes
- UniProt (version: 511), type: AA, tax: Yes
- Mascot (version: 2.1.15), type: AA, tax: NA
- PHD (version: 2.5), type: AA, tax: NA
- ZPE (version: 2.2.15), type: NA, tax: NA
- Mascot (version: 2.1.15), type: NA, tax: NA
- PHD (version: 2.5), type: NA, tax: NA
- ZPE (version: 2.2.15), type: NA, tax: NA

**AC list**

**Search Engine**
- MS/MS

**Instrument Type**:
- MALDI TOF-TOF

**Scoring Model**
- default

**Default Parent Charge**
- 1

**Tolerance**
- medium

**AA Modif. (def.)**
- TRYPSIN_19

**AA Modif. Details**
- Trypsin

**Enzyme**
- Trypsin

**Parent Error Tolerance**
- 0.1

**Collision Energy**
- 30

**Acceptance Parameters**

**Peak Lists**

**File Format**: .mgf

**Compounds Overview**

**Peak List(s)**

<table>
<thead>
<tr>
<th>#</th>
<th>AC</th>
<th>IE</th>
<th>Score</th>
<th>% Cov</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P00348</td>
<td>HMR_HUMAN</td>
<td>61.1</td>
<td>15/11</td>
<td>Nucleophosmin (NPM) (Nucleolar...)</td>
</tr>
<tr>
<td>2</td>
<td>P07919</td>
<td>HUMAN</td>
<td>48.4</td>
<td>1/1</td>
<td>Heterogeneous nuclear ribonucleoprotein...</td>
</tr>
</tbody>
</table>
The Proteins overview

<table>
<thead>
<tr>
<th>Protein group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum albumin precursor</td>
<td>(Allergen Bos d 6) (BSA)</td>
</tr>
<tr>
<td>Glycogen phosphorylase, muscle</td>
<td></td>
</tr>
<tr>
<td>Lactoperoxidase precursor</td>
<td>(EC 1.11.1.6)</td>
</tr>
<tr>
<td>Trypsin precursor</td>
<td>(EC 3.4.21.4)</td>
</tr>
<tr>
<td>Catalase</td>
<td>(EC 1.11.1.6)</td>
</tr>
<tr>
<td>Glycogen phosphorylase, brain</td>
<td></td>
</tr>
<tr>
<td>Catalase</td>
<td>(EC 1.11.1.6)</td>
</tr>
<tr>
<td>Leucylaminopeptidase</td>
<td></td>
</tr>
<tr>
<td>Catalase</td>
<td>(EC 1.11.1.6)</td>
</tr>
</tbody>
</table>

**List of identified proteins**

**Corresponding list of identified peptides**
The Proteins overview

Hints about the significance of the score
Better when high intensity peaks are matched and ion series are extended, without too many and too big holes
The scoring system in Phenyx

- The score is the sum of up to 12 *basic scores* such as:
  - presence of a, b, y, y++, B-H$_2$O...; co-occurrence of ion series (using HMMs), peak intensities, residue modifications (PTM or chemical), ...

- True probabilistic approach for each peptide match
  
  \[
  \frac{\text{likelihood of being correct}}{\text{likelihood of being random}} = \log \left( \frac{P\text{(correct)}}{P\text{(random)}} \right)
  \]

- Function of instruments and molecular types
  - Esquire 3000+, LCQ; iTRAQ vs. unmodified peptides

- Scores are normalised into z-scores
X!Tandem

The Global Proteome Machine, Simple search page

1. **spectra**
   - DTA, PKL or Matrix Science format only
   - [Browse...](#)

2. **taxon**
   - Select one or more.
   - Eukaryotes:
     - NCBI 36 (ENSEMBL)
     - Human (SwissProt)
     - H. sapiens (IPI)
     - Human Invitational DB
     - H. sapiens (UNIGENE)
   - Prokaryotes:
     - none
     - Acidiphilium cryptum JF-5
     - Acidobacteria bacterium Ellin345
     - Acidothermus cellulolyticus 11B
     - Acidovorax avenae citrulli AAC00-1

3. **measurement errors**
   - Fragment mass error: 0.4 [Da](#)
   - Feeling lucky

4. **residue modifications**
   - Complete modifications: Carbamidomethyl (C)
     - Specify your own
   - Potential modifications:
     - Oxidation (N)
     - Oxidation (W)
     - Deamidation (N)
     - Specify your own

5. **refinement specification**
   - Potential modifications (unimod):
     - Round 1
       - Oxidation (N)
       - Oxidation (W)
       - Deamidation (N)
       - ICAT-D2H(S) (C)
     - Round 2
       - Oxidation (M)
       - OxiCat (M)
       - Oxidation (W)
       - Deamidation (N)
   - Point mutations: yes
   - Semi-style cleavage: yes

---

PMP

[www.thegpm.org](http://www.thegpm.org)
The two-rounds search
Mascot, Phenyx and X!Tandem

The identification process may be launched in 2-rounds

- Each round is defined with a set of search criteria
  - First round searches the selected database(s) with stringent parameters,
  - Second round searches the proteins that have passed the first round (relaxed parameters):

  ⇒ Accelerate the job when looking for many variable modifications, or unspecific cleavages

  ⇒ Appropriate when the first round defines stringent criteria to capture a protein ID, and the second round looks for looser peptide identifications
Example 2\textsuperscript{nd} round

1\textsuperscript{rd} round,
Only 3 fixed mods
131 valid,
75% cov.

2\textsuperscript{nd} round,
Add variable mods
205 valid,
84% cov.

2\textsuperscript{nd} round,
With all mods
And half cleaved
348 valid,
90% cov.
Source of errors in assigning peptides

- Scores not adapted
- Parameters are too stringent or too loose
- Low MS/MS spectrum quality (many noise peaks, low signal to noise ratio, missing fragment ions, contaminants)
- Homologous proteins
- Incorrectly assigned charge state
- Pre-selection of the 2nd isotope (the parent mass is shifted of 1 Da. A solution is to take the parent mass tol. larger, but may drawn the good peptide too)...
- Novel peptide or variant
Hints to know when the identification is correct

With MS/MS

- The higher the number of peptides identified per protein, the better
- Sequence coverage: the larger the sub-sequences and the higher the sequence coverage value, the better
- Depends on the sample complexity and experiment workflow
- Scores: the higher, the better.
- Filter on the correct species if you know it (reduces the search space, time, and errors)
- Better when high intensity peaks are matched and ion series are extended, without too many and too big holes.
- Better when the errors are more or less constants among all ions.
- If you have time, try many tools and compare the results
Popitam, GutenTag, InsPect, OpenSea

- identify and characterize peptides with mutations or unexpected post-translational modifications
- "open-modification search": it takes into account any type and number of differences between an MS/MS spectrum and theoretical peptides

<table>
<thead>
<tr>
<th>#</th>
<th>Scenario score</th>
<th>Delta score</th>
<th>p-value</th>
<th>Mass</th>
<th>Delta mass</th>
<th>Peptide / scenario (shifts)</th>
<th>Found in ID(AC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>659853.40</td>
<td>0.00</td>
<td>-</td>
<td>1372.65</td>
<td>114.12</td>
<td>VFNCISYSPLCK</td>
<td>WDR12_HUMAN_C1 (Q9GZL7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>vFN*ISYSp1**</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>-</td>
<td>-</td>
<td>1300.59</td>
<td>186.18</td>
<td>FCSGSMKMLK</td>
<td>WDR12_HUMAN_C1 (Q9GZL7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>Csg</em>*******</td>
<td></td>
</tr>
</tbody>
</table>
De novo sequencing

- Sequencing = « read » the full peptide sequence out of the spectrum (from scratch)
- Then, eventually search database for sequence (not necessarily)
Why *de novo* sequencing is difficult

1. Leucine and isoleucine have the same mass
2. Glutamine and lysine differ in mass by 0.036Da
3. Phenylalanine and oxidized methionine differ in mass by 0.033Da
4. Cleavages do not occur at every peptide bond
   (or cannot be observed on the MS-MS)
   - Poor quality spectrum (some fragment ions are below noise level)
   - The C-terminal side of proline is often resistant to cleavage
   - Absence of mobile protons
   - Peptides with free N-termini often lack fragmentation between the first and second amino acids
Why de novo sequencing is difficult (II)

5. Certain amino acids have the same mass as pairs of other amino acids
   - Gly + Gly (114.0429) Asn (114.0429)
   - Ala + Gly (128.0586) Gln (128.0586)
   - Ala + Gly (128.0586) Lys (128.0950)
   - Gly + Val (156.0742) Arg (156.1011)
   - Ala + Asp (186.0641) Trp (186.0793)
   - Ser + Val (186.1005) Trp (186.0793)

6. Directionality of an ion series is not always known (are they b- or y-ions?)
de novo sequencing methods

- Manual inference by an expert

- Creation of all combinations of possible sequences and comparison with the spectrum \(\Rightarrow\) exponential with the length of the peptide

- Same thing but with prefixed filtering

- Genetic algorithm approach

- Spectrum transformed into a transition acyclic graph then search of the best path through the graph: SeqMS, SHERENGA, LUTEFISK97
## Summary of de novo sequencing tools

<table>
<thead>
<tr>
<th>Software</th>
<th>Source website</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEAKS*</td>
<td><a href="http://www.bioinformaticssolutions.com">www.bioinformaticssolutions.com</a></td>
</tr>
<tr>
<td>SeqMS (download)</td>
<td><a href="http://www.protein.osaka-u.ac.jp/rcsfp/profiling/SeqMS.html">www.protein.osaka-u.ac.jp/rcsfp/profiling/SeqMS.html</a></td>
</tr>
<tr>
<td>Sherenga (included in SpectrumMill)</td>
<td>N/A</td>
</tr>
<tr>
<td>Lutefisk (download)</td>
<td><a href="http://www.hairyfatguy.com/Lutefisk">www.hairyfatguy.com/Lutefisk</a></td>
</tr>
<tr>
<td>DeNovoX*</td>
<td><a href="http://www.thermo.com">www.thermo.com</a></td>
</tr>
<tr>
<td>PepNovo (download)</td>
<td>peptide.ucsd.edu/pepnovo.py</td>
</tr>
<tr>
<td>SpectrumMill*</td>
<td><a href="http://www.home.agilent.com">www.home.agilent.com</a></td>
</tr>
</tbody>
</table>

*Commercialized
PepNovo

- scoring method uses a probabilistic network whose structure reflects the chemical and physical rules that govern the peptide fragmentation
- specific for Ion Trap data

PepNovo Search Results

Spectra Filename: example_yxz02.dta
Parent Mass As: M+H
Charge: 2
Instrument: ESI-ION-TRAP
Protease: None
Minimum Tag Length: 3
Maximum Tag Length: 3
Limit: tags
Maximum Tags: 10

Save summary for all spectra

denovo Result

Score: 104.75
Start: [207.6]
Sequence: SLVGQTSPQ

<table>
<thead>
<tr>
<th>Mass</th>
<th>AA</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>207.6</td>
<td>S</td>
<td>0.6588</td>
</tr>
<tr>
<td>294.7</td>
<td>L</td>
<td>0.8655</td>
</tr>
<tr>
<td>407.9</td>
<td>V</td>
<td>0.6710</td>
</tr>
<tr>
<td>507.0</td>
<td>G</td>
<td>0.3332</td>
</tr>
<tr>
<td>564.1</td>
<td>Q</td>
<td>0.0868</td>
</tr>
<tr>
<td>692.2</td>
<td>T</td>
<td>0.2540</td>
</tr>
<tr>
<td>793.3</td>
<td>S</td>
<td>0.1955</td>
</tr>
<tr>
<td>880.4</td>
<td>P</td>
<td>0.2454</td>
</tr>
<tr>
<td>977.5</td>
<td>Q</td>
<td>0.0105</td>
</tr>
</tbody>
</table>

Tags

<table>
<thead>
<tr>
<th>Position</th>
<th>Start</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVG</td>
<td>295.2</td>
<td>0.692</td>
</tr>
<tr>
<td>SLV</td>
<td>207.6</td>
<td>0.681</td>
</tr>
<tr>
<td>FVG</td>
<td>261.2</td>
<td>0.641</td>
</tr>
<tr>
<td>SLY</td>
<td>207.6</td>
<td>0.461</td>
</tr>
</tbody>
</table>
Decoy database

- Is used to repeat the search, using identical search parameters, against a database in which the sequences have been reversed or randomised.

- Do not expect to get any real matches from the "decoy" database.

- Helps estimate the number of false positives that are present in the results from the real database.

- It is a good validation method for MS/MS searches of large data sets, it is not as useful for a search of a small number of spectra, because the number of matches is too small to give an accurate estimate.
Possible Decoy databases

Original sequence

- **Reverse**: each sequence of the real database is reversed (back to forth). Not used for MS search. When dealing MS/MS data, the b and y ions could be reversed (not appropriate)

- **Shuffle**: each sequence of the real database is shuffled with the same average AA composition

- **Random**: each sequence of the real database is a new randomised sequence based on the AA composition of the database
Help validating in Mascot, Phenyx and X!Tandem

• In Mascot - decoy shuffled with the same average amino acid composition (Decoy parameter). Otherwise, you may create your own.
• X!Tandem – reversed sequences
• Phenyx – reversed sequences available, otherwise you choose to create your own
E-values

• For a given score S, it indicates the number of matches that are expected to occur by chance in a database with a score at least equal to S.

• The e-value takes into account the size of the database that was searched. As a consequence it has a maximum of the number of sequences in the database.

• The lower the e-value, the more significant the score is.

• An e-value depends on the calculation of the p-value.
p-value

• A p-value describes the probability, which assesses the chance of validly rejecting the null hypothesis. If the p-value is $10^{-5}$ then the rejection of the null hypothesis is due to chance with a probability of $10^{-5}$.

• A p-value ranges between 0 and 1.0

• The larger the search space, the higher the p-value since the chance of a peptide being a random match increases.

• The lower the p-value, the more significant is the match.

Source: Lisacek, Practical Proteomics, 2006 Sep;6 Suppl 2:22-32
Z-score

- Z-score is a dimensionless quantity derived by subtracting the population mean from an individual (raw) score and then dividing the difference by the population standard deviation. \( z - \text{score} = \frac{x - \mu}{\sigma} \)

- The z score reveals how many units of the standard deviation a case is above or below the mean.

Source: wikipedia
So what?

• For small (significant) p-values, p and e are approximately equal, so the choice of one or the other is often equivalent. It is therefore reasonable to assimilate low p-values in Phenyx to e-values. X!Tandem simply switches e-values to log values to remove the powers of 10.

• For a single search (or set of sampled peptides), you can compare z-scores. However, when two or more searches are performed on different size spaces, you first need to look at the p-values before comparing z-scores.

Source: Lisacek, Practical Proteomics, 2006 Sep;6 Suppl 2:22-32
Thank you for your attention.