Biological Mass spectrometry in Protein Chemistry

Tuula Nyman
Institute of Biotechnology
tuula.nyman@helsinki.fi

MASS SPECTROMETRY

is an analytical technique that identifies the chemical composition of a compound or sample based on the mass-to-charge ratio of charged particles

ION SOURCE: molecules of interest are ionized

↓

MASS ANALYZER:
ions are separated according to their m/z-ratios

↓

DETECTOR: separated ions are detected

MS of small molecules

-In 1918, Arthur Jeffrey Dempster developed the first modern mass spectrometer, and established the basic theory and design of mass spectrometers that is still used to this day

-1919 Francis Aston constructs the first velocity focusing mass spectrograph with mass resolving power of 130 (1922 Nobel Prize in chemistry)

-The use of a mass spectrometer as the detector in gas chromatography was developed during the 1950s by Roland Gohlke and Fred McLafferty

-Ionization modes: chemical ionization (CI) and electron ionization (EI), not suitable for labile biomolecules

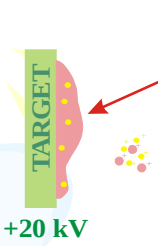
Biological mass spectrometry

-two modes of ionization:
MALDI (matrix assisted laser desorption ionization) and ESI (electrospray ionization)

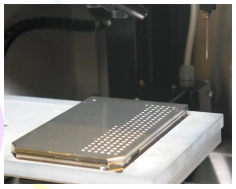
-developed in 1980's by Michael Karas & Franz Hillenkamp, Koichi Tanaka, John Fenn & Matthias Mann, Peter Roespstorff...

Nobel-price in Chemistry 2002 to Tanaka and Fenn 'For their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules'

MALDI= matrix assisted laser desorption ionization

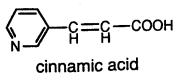


The analyte substance is embedded in a crystallized matrix, which is irradiated by a laser. The power of the laser beam is usually adjusted in a way that it has enough energy to ionize the biomolecules and matrix molecules but does not split the large analyte molecule.

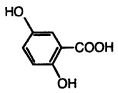


A 384 position MALDI-TOF sample target plate.

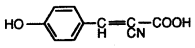
nicotinic acid



cinnamic acid



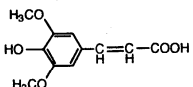
2,5-dihydroxy benzoic acid
(DHB)
peptides, small proteins and
oligonucleotides (< 10 bases)



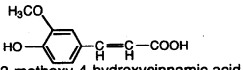
α-cyano-4-hydroxycinnamic acid
(CCA)

peptides and glycopeptides

==========
peptides



3,5-dimethoxy-4-hydroxycinnamic acid
(sinpinic acid)
peptides and proteins



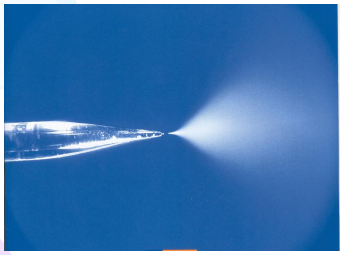
3-methoxy-4-hydroxycinnamic acid
(ferulic acid)
peptides, proteins, and some oligos

==========
proteins

Commonly used matrixes for MALDI-TOF mass spectrometry

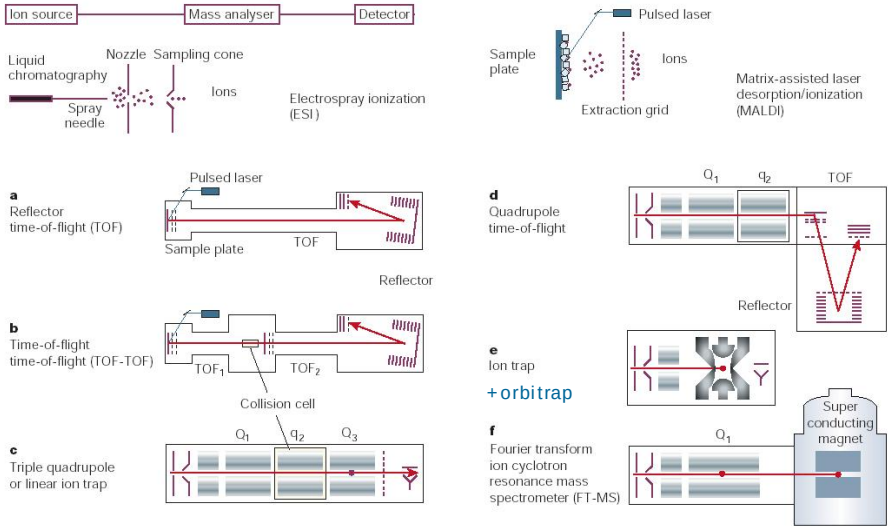
Electrospray ionization (ESI):

Creation of ions by spraying a solution into an electrical field. This process, which belongs to the "soft" ionization techniques, enables the analysis of intact biomolecules, such as e.g. proteins and peptides by mass spectrometry.



Electrospray of peptides in 0.1% FA /ACN from a 15 um I.D. fused silica glass needle. The liquid flow is 200 nl/min and the needle has a potential of 2000V as compared to the cone inlet of the mass spectrometer.

Mass analyser types



All mass analyzers work in high vacuum

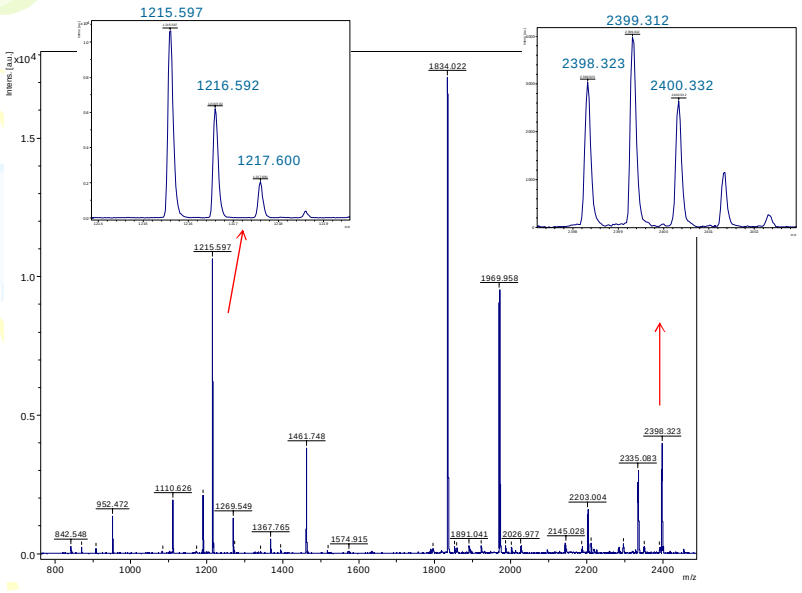
A tandem mass spectrometer is a mass spectrometer that has more than one analyser, in practice usually two.

Tandem mass spectrometry (MS/MS) is used to produce structural information about a compound by fragmenting specific sample ions inside the mass spectrometer and identifying the resulting fragment ions

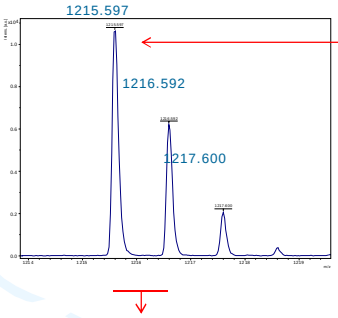
Common MS-instrument types:

-MALDI-TOF and MALDI-TOF/TOF
-ESI-triple quadrupole
-ESI-ion trap
-ESI-hybrid quadrupole TOF
-ESI-ion trap
-ESI-orbitrap

MALDI TOF spectrum of a peptide mixture



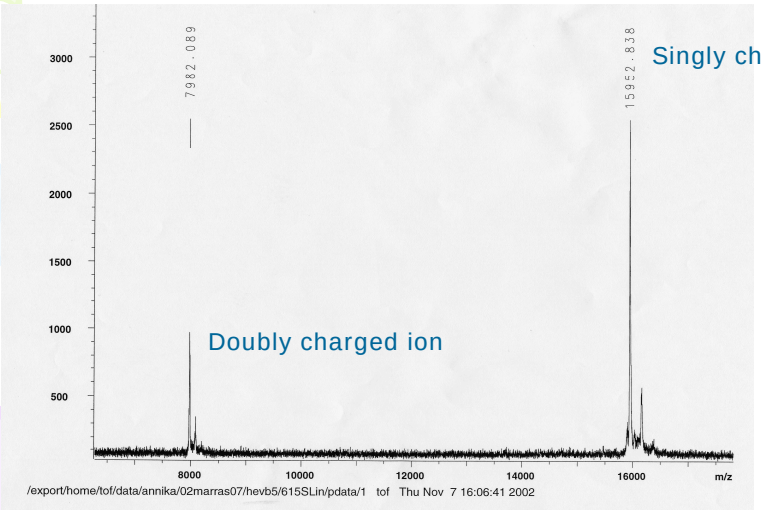
Monoisotopic mass is calculated using the mass of the most abundant natural isotope of each constituent element



^{12}C and ^{14}N

the mass difference between the isotopic peaks is 1 amu (1 Da) as isotopes differ in mass by the addition of 1 neutron which weighs 1 amu

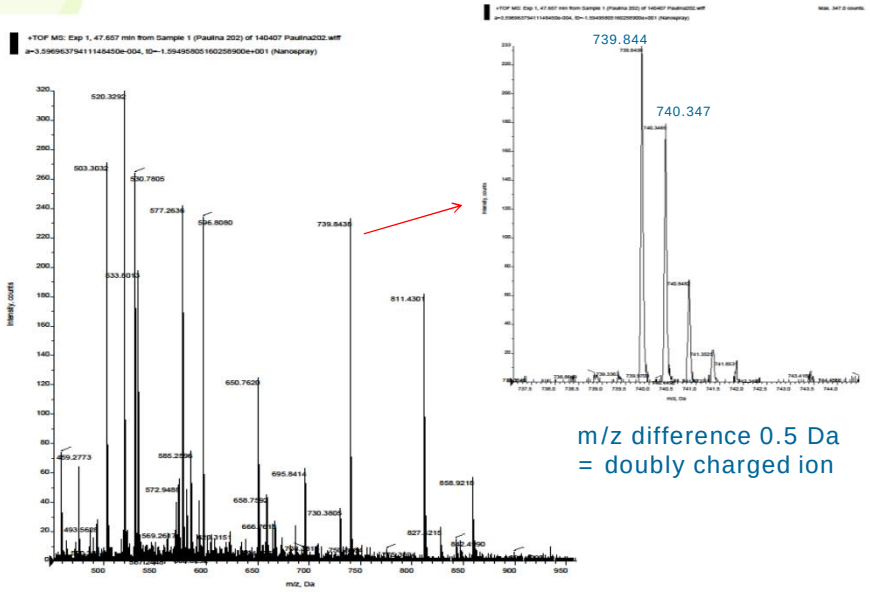
MALDI-TOF spectrum of a protein



Singly charged ion

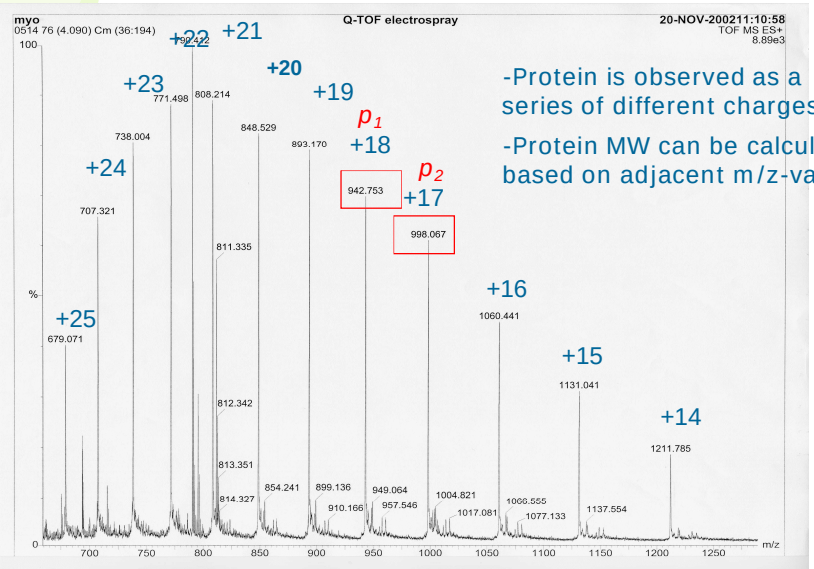
Doubly charged ion

ESI MS spectrum of a peptide mixture



m/z difference 0.5 Da
= doubly charged ion

ESI MS spectrum of a protein



-Protein is observed as a series of different charges

-Protein MW can be calculated based on adjacent m/z-values

MALDI

- sample is crystallized
- produces mainly singly charged ions
- simple, easy-to-use
- more tolerant to salts+other contaminants in the sample than ESI

ESI

- liquid sample
- produces multiply charged ions
- easy to couple with HPLC

http://www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm
http://www.ionsource.com/
http://planetorbitrap.com/

MS in protein chemistry/ proteomics/structural biology

-protein identification

-protein MW determination

-characterisation of post-translational modifications

-relative quantification of proteins between samples

-analysis of protein complexes

-essential role in proteomics

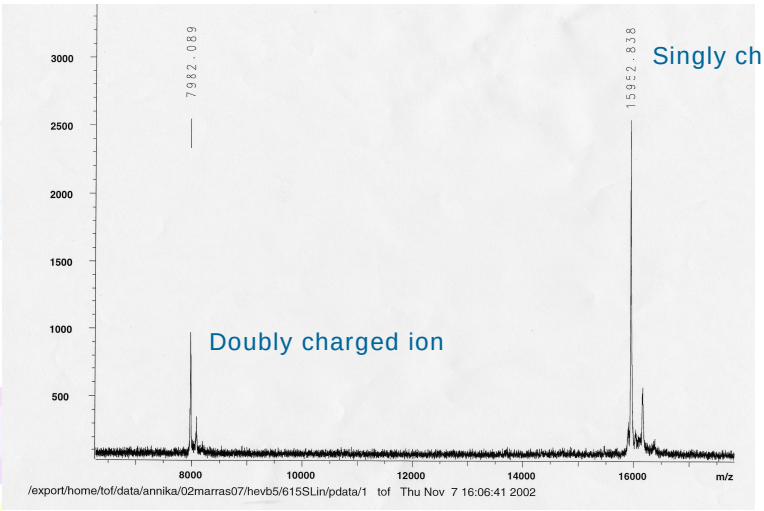
-MS imaging

Protein MW determination by MS (NOT= identification)

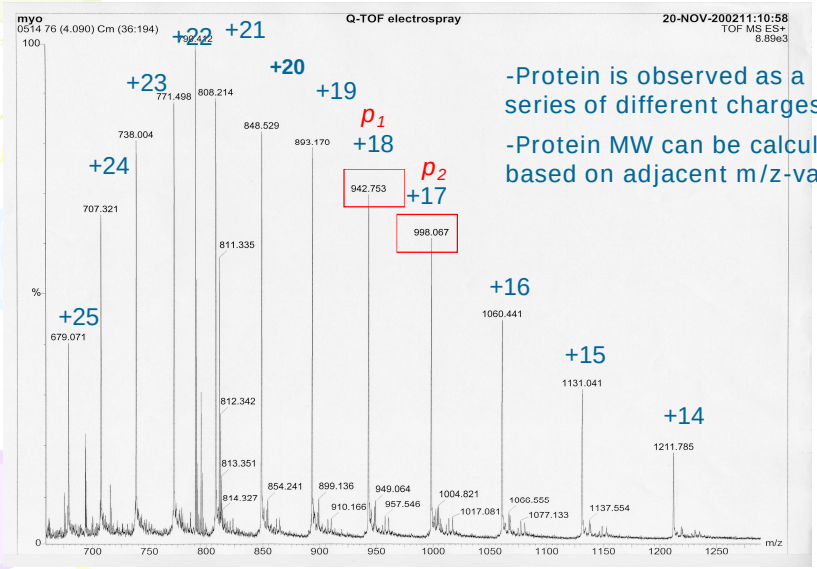
-for MW determination the protein needs to be in solution without salts and detergents

-usually proteins are first purified with RP chromatography before MW measurement

-MALDI TOF MS, linear mode
-accuracy is not as good with ESI MS



Protein MW determination, ESI MS



-Protein is observed as a series of different charges

-Protein MW can be calculated based on adjacent m/z-values

Protein MW calculation from ESI spectra

$$p = m/z$$
$$p_1 = (M_r + z_1)/z_1$$
$$p_2 = [M_r + (z_1 - 1)]/(z_1 - 1)$$

p = a peak in the mass spectrum
m = total mass of an ion
z = total charge
M_r = average mass of the protein

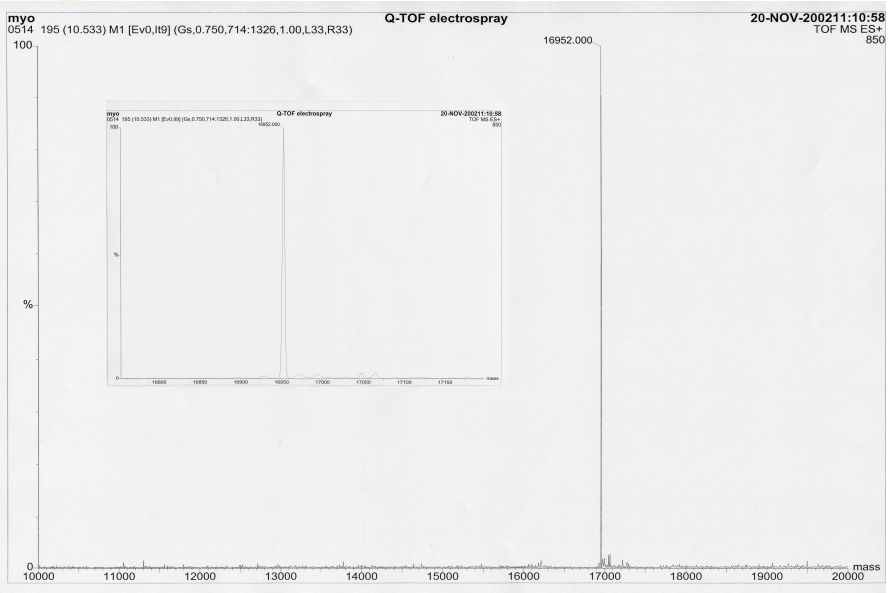
$p = m/z$
$p_1 = (M_r + z_1)/z_1$
$p_2 = [M_r + (z_1 - 1)]/(z_1 - 1)$

If $p_1 = 942.753$ and $p_2 = 998.067$

$942.753 = (M_r + z_1)/z_1$
$942.753z_1 = Mr + z_1$
$941.753z_1 = M_r$

$998.067 = (941.753z_1 + z_1 - 1)/z_1 - 1$
$998.067z_1 - 997.067 = 942.753z_1 - 1$
$(998.067 - 942.753)z_1 = 996.067$
$55.314z_1 = 996.067 \Rightarrow z_1 = 18.0075$

$M_r = 941.753z_1 = 16\ 951.6$ Da



Deconvoluted electrospray mass spectrum of myoglobin

Protein identification by mass spectrometry

- protein of interest is cleaved into peptides with a specific enzyme
- peptides are analyzed by MS (and MS/MS) followed by a database search with the acquired MS (and MS/MS) data

Protein identification methods:

- Peptide mass fingerprinting (PMF)
- Identification based on MS/MS data from one or more peptides

Peptide mass fingerprint (PMF)

A mass spectrum of the peptide mixture resulting from the digestion of a protein by an enzyme, usually measured by MALDI-TOF

Identification based on peptide MW information only

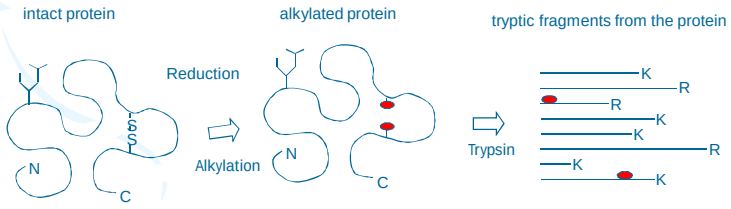
Database search engines create theoretical PMFs for all the proteins in the database and compare these to the measured PMF from protein X

Protein identification based on MS/MS data from one or more peptides

Database search engines create theoretical peptide fragmentation patterns for all the proteins in the database and compare these to the measured MS/MS data

Protein identification by MS

- Protein has to be digested into peptides
- Disulphide bridges need to be reduced and alkylated before digestion



Protein digestion into peptides before MS analysis

Digestion can be done *in-solution* or *in-gel*

The enzyme has to be as specific as possible

Trypsin is most commonly used enzyme:

-very specific, quite cheap
-cleaves peptide bond after lysines and arginines
-tryptic peptides are 'good' for MS analysis because they end up with basic amino acid
-works for both *in-solution* and *in-gel* digestion

Other enzymes:

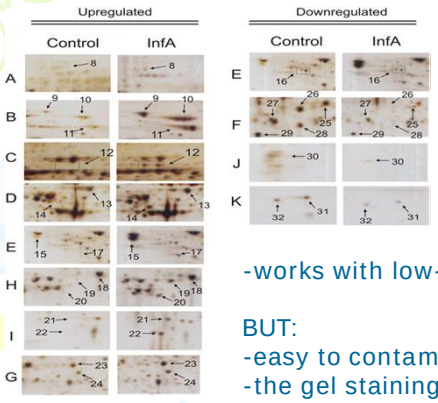
-LysC, cleaves after lysines
-LysN, cleaves before lysines
-AspN, cleaves before aspartic acid residues
-V8 protease, cleaves peptide bonds exclusively on the carbonyl side of aspartate and glutamate residues

-possible to do double-digestions

Chemical cleavage:

Cyanogen bromide, cuts after methionines

In-gel digestion

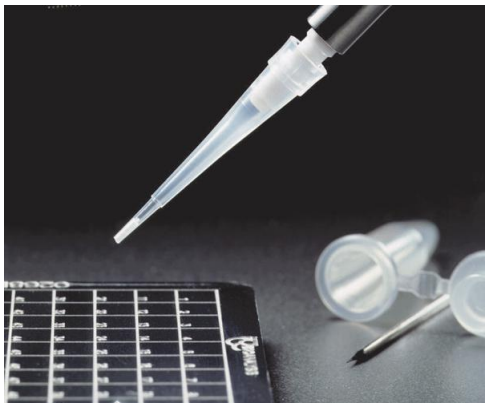


-works with low-femtomolar amounts of proteins

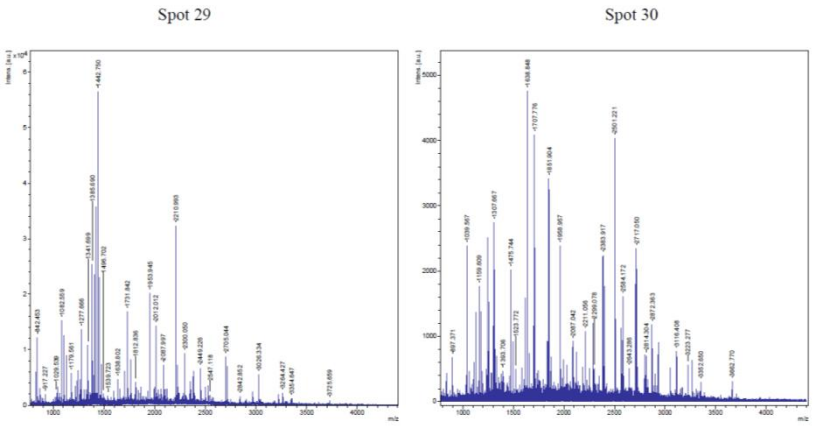
BUT:
-easy to contaminate samples with keratin
-the gel staining protocol needs to be compatible with MS
-silver-staining: fixing with glutaraldehy cross-links protein into gel matrix after which identification is not possible

Peptide mass fingerprinting

-usually peptides need to be desalted and concentrated before MALDI analysis
-> ZipTips, peptide elution directly onto MALDI target plate
-MALDI matrix included in the elution solution

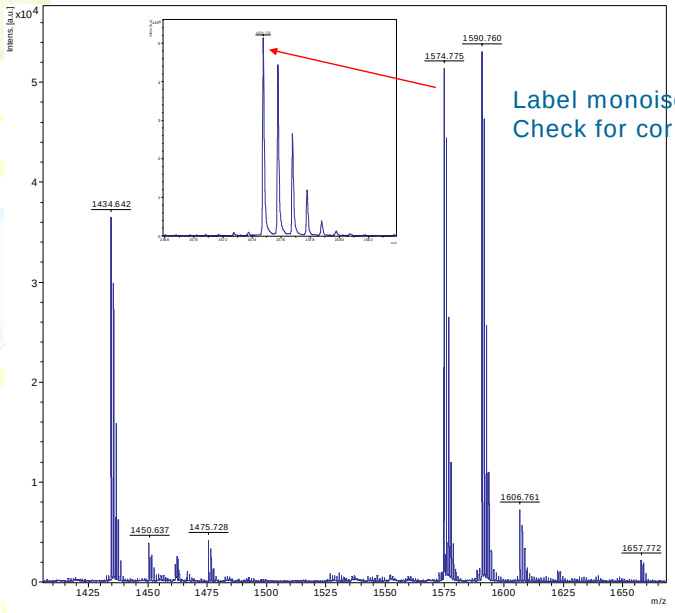


Peptide mass fingerprinting
MALDI, Positive ion reflector mode



-silver-stained spots from 2-DE gels, in-gel digestion+ZipTipping

Peak picking parameters important in PMF



Label monoisotopic masses
Check for correct isotope!

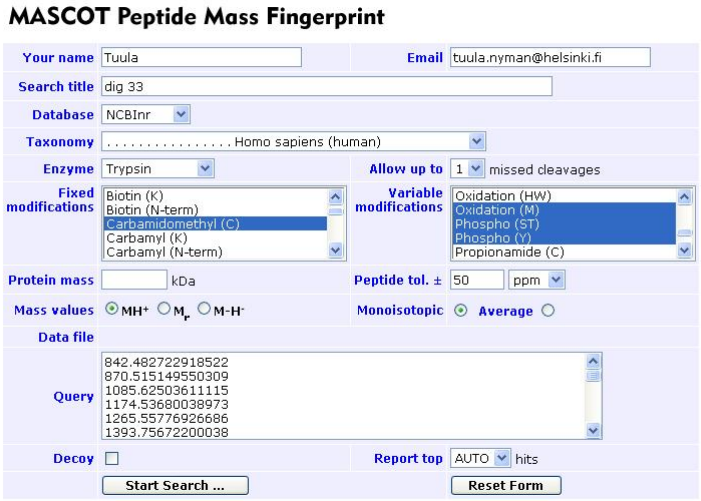
Publicly available search engines for PMF

- Mascot
- ProteinProspector/ MS-Fit
- PROWL/ ProFound
- Aldente

Databases

Database	Comment
EST	EST divisions of Genbank, (currently EST_human, EST_mouse, EST_others)
MSDB	Comprehensive, non-identical protein database
NCBInr	Comprehensive, non-identical protein database
SwissProt	High quality, curated protein database

PMF/ database search



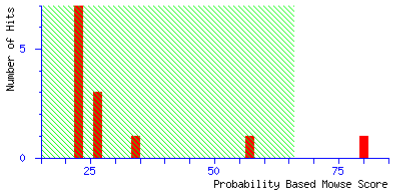
www.matrixscience.com

Mascot Search Results

User : Tuula
Email : tuula.nyman@helsinki.fi
Search title : dig 33
Database : NCBInr 20090912 (9680073 sequences; 3307708198 residues)
Taxonomy : Homo sapiens (human) (223942 sequences)
Timestamp : 17 Sep 2009 at 07:28:13 GMT
Top Score : 80 for gi|109096484, PREDICTED: tubulin, alpha, ubiquitous isoform 10 [Macaca mulatta]

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 66 are significant (p<0.05).



Concise Protein Summary Report

Format As: Concise Protein Summary — Help
Significance threshold p< 0.05 — Max. number of hits: AUTO

Re-Search All | Search Unmatched

1. gi|109096484 **Mass:** 46797 **Score:** 80 **Expect:** 0.0022 **Queries matched:** 12
PREDICTED: tubulin, alpha, ubiquitous isoform 10 [Macaca mulatta]
gi|193786502 **Mass:** 46725 **Score:** 80 **Expect:** 0.0022 **Queries matched:** 12
unnamed protein product [Homo sapiens]
gi|193787715 **Mass:** 46825 **Score:** 80 **Expect:** 0.0022 **Queries matched:** 12
unnamed protein product [Homo sapiens]
gi|34740335 **Mass:** 50804 **Score:** 76 **Expect:** 0.0052 **Queries matched:** 12
tubulin, alpha 1B [Mus musculus]
gi|73996547 **Mass:** 46781 **Score:** 69 **Expect:** 0.029 **Queries matched:** 11
PREDICTED: similar to tubulin, alpha 1 isoform 9 [Canis familiaris]
gi|109096516 **Mass:** 37707 **Score:** 66 **Expect:** 0.054 **Queries matched:** 10
PREDICTED: alpha tubulin isoform 2 [Macaca mulatta]
gi|158259731 **Mass:** 50788 **Score:** 66 **Expect:** 0.058 **Queries matched:** 11
unnamed protein product [Homo sapiens]

Mascot Search Results

Protein View

Match to: **gi|193786502** Score: **80** Expect: **0.0022**
unnamed protein product [Homo sapiens]

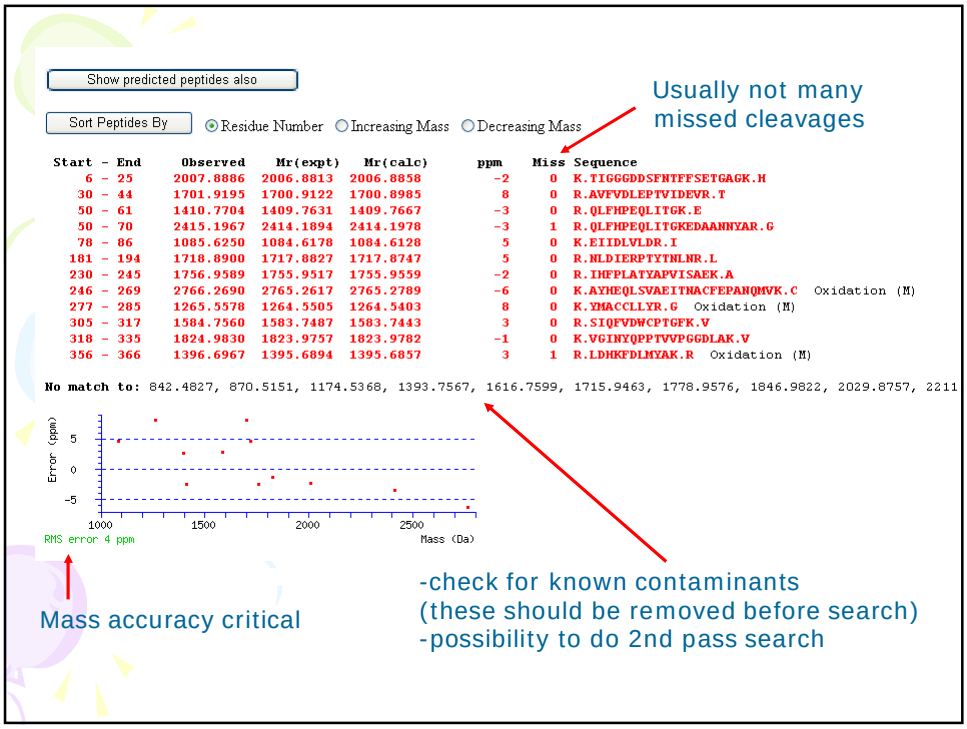
Nominal mass (M_r): **46725**; Calculated pI value: **4.99**
NCBI BLAST search of gi|193786502 against nr
Unformatted sequence string for pasting into other applications

Taxonomy: Homo sapiens

Fixed modifications: Carbamidomethyl (C)
Variable modifications: Oxidation (M),Phospho (ST),Phospho (Y)
Cleavage by Trypsin: cuts C-term side of KR unless next residue is P
Number of mass values searched: **24**
Number of mass values matched: **12**
Sequence Coverage: **40%**

Matched peptides shown in **Bold Red**

```
      1 MPSDKTIGGG DDSFNTFFSE TGAGKHVPRA VFVDLEPTVI DEVRTGTYRQ
     51 LFHPEQLITG KEDAANNYAR GHYTIGKEII DLVLDRIRKL ADQCTGLQGF
    101 LVFHSFGGGT GSGFTSLLME RLSVDYGKKS KLEFSIYPAP QVSTAVVEPY
    151 NSILTTHTTL EHSDCAFMVD NEAIYDICRR NLDIERPTYT NLNRLISQIV
    201 SSITASLRFD GALNVDLTEF QTNLVPYPRI HFPLATYAPV ISAEKAYHEQ
    251 LSVAEITNAC FEPANQMVKC DPRHGKYMAC CLLYRGDVVP KDVNAAIATI
    301 KTKRSIQFVD WCPTGFKVGI NYQPPTVVPG GDLAKVQRAV CMLSNTTAIA
    351 EAWARLDHKF DLMYAKRAFV HWYVGEGMEE GEFSEAREDM AALEKDYEEV
    401 GVDSVEGEGG EEGEEY
```



PEPTIDE MASS FINGERPRINTING

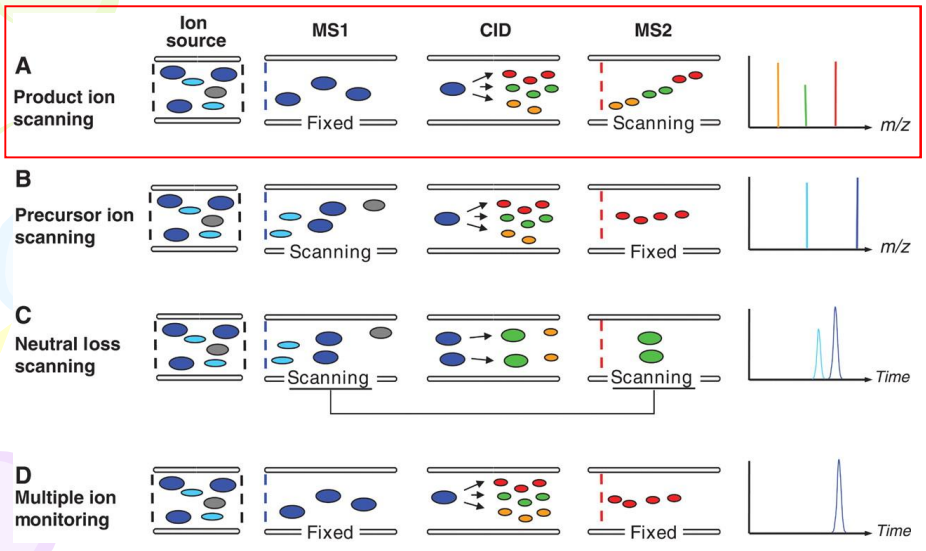
- 'quick and easy'
- requires
  - a very spesific enzyme
  - optimized digestion+ desalting protocols
  - internal/close external calibration of MALDI spectra
- works only for proteins which are already in the databases as protein sequences
- not suitable for complex protein mixtures

Protein identification based on MS/MS data from one or more peptides

-MALDI-TOF/TOF or nanoLC-ESI-MS/MS analysis

-suitable for (complex) protein mixtures, especially when combined with LC separation of peptides before MS/MS

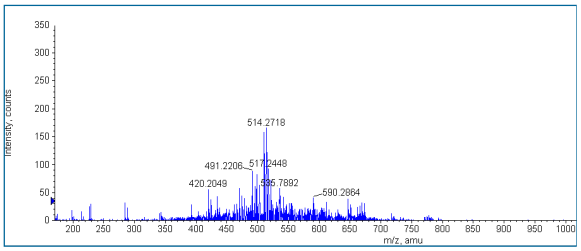
Tandem mass spectrometry scan types



Product ion scanning

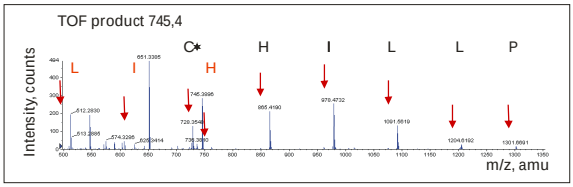
MS scan:

The ions are first separated according to their m/z ratios

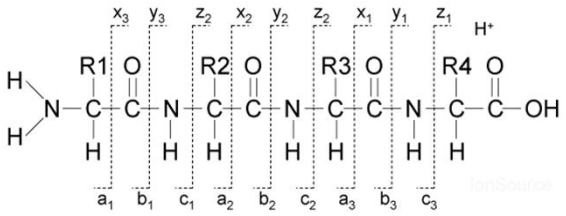


MS/MS scan:

Peptides with certain m/z-ratio are selected and fragmented inside the mass analyzer, and the m/z-ratios of the fragment ions are measured



Peptide Fragmentation Nomenclature



Peptides do not fragment sequentially, the fragmentation events are somewhat random.

The most common peptide fragments observed in low energy collisions are a, b and y ions. The b ions appear to extend from the amino terminus (N-terminus), and y ions appear to extend from the carboxyl terminus (C-terminus).

Protein identification with MALDI-TOF/TOF

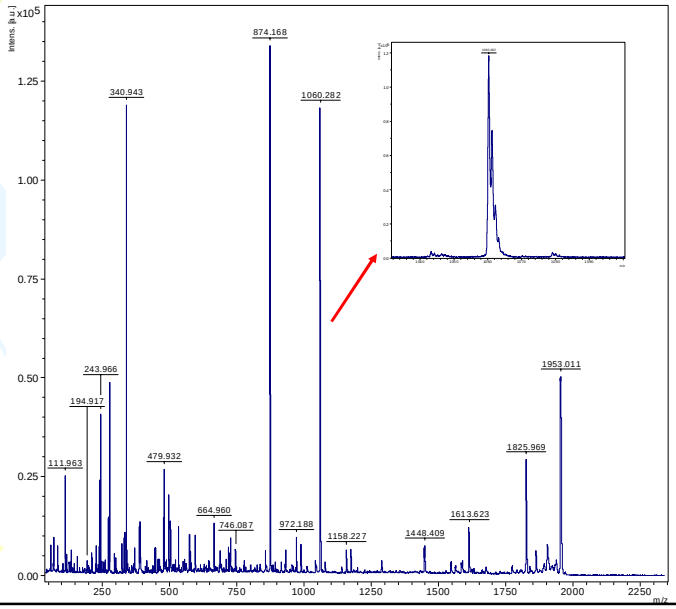
-first, PMF with MALDI-TOF

-next, selected precursor ions can be fragmented (TOF/TOF analysis)

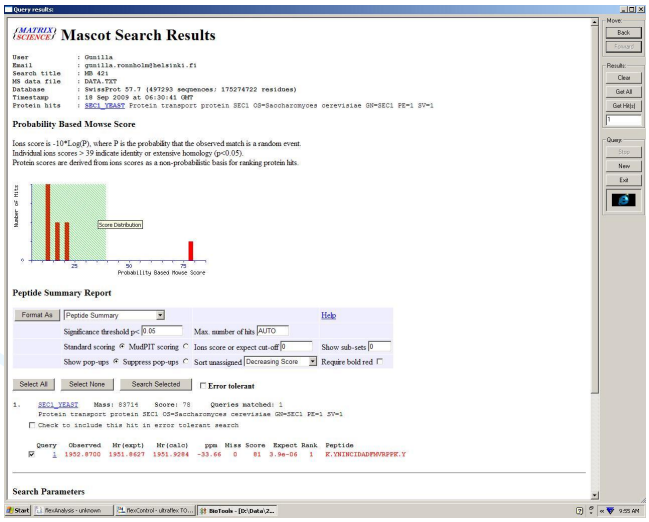
-MALDI produces singly charged parent ions
→ product ion spectra not as easy to interpret as in ESI-MS/MS

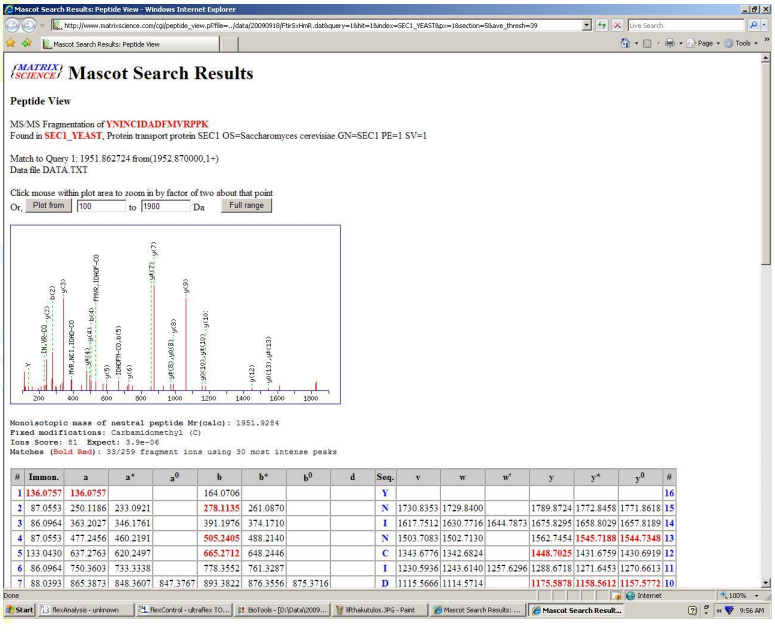
-database search with both PMF and MS/MS information

MALDI-TOF/TOF fragment ion spectra from parent ion m/z 1952



Sequence Query: One or more peptide mass values associated with information such as partial or ambiguous sequence strings, amino acid composition information, MS/MS fragment ion masses, etc.





#	Immon.	a	a*	a0	b	b*	b0	d	Seq.	v	w	w'	y	y*	y0	#
1	136.0757	136.0757			164.0706				Y							16
2	87.0553	250.1186	233.0921		278.1135	261.0870			N	1730.8353	1729.8400		1789.8724	1772.8458	1771.8618	15
3	86.0964	363.2027	346.1761		391.1976	374.1710			I	1617.7512	1630.7716	1644.7873	1675.8295	1658.8029	1657.8189	14
4	87.0553	477.2456	460.2191		505.2405	488.2140			N	1503.7083	1502.7130		1562.7454	1545.7188	1544.7348	13
5	133.0430	637.2763	620.2497		665.2712	648.2446			C	1343.6776	1342.6824		1448.7025	1431.6759	1430.6919	12
6	86.0964	750.3603	733.3338		778.3552	761.3287			I	1230.5936	1243.6140	1257.6296	1288.6718	1271.6453	1270.6613	11
7	88.0393	865.3873	848.3607	847.3767	893.3822	876.3556	875.3716		D	1115.5666	1114.5714		1175.5878	1158.5612	1157.5772	10

Protein identification using nanoLC-MS/MS

-50-75 um i.d. RP-columns, 200 nl/min
⟶ no need to split the effluent before MS

-DDA= data dependent analysis, can be fully automated

-suitable for complex protein mixtures, possibility to identify hundreds of proteins in one run

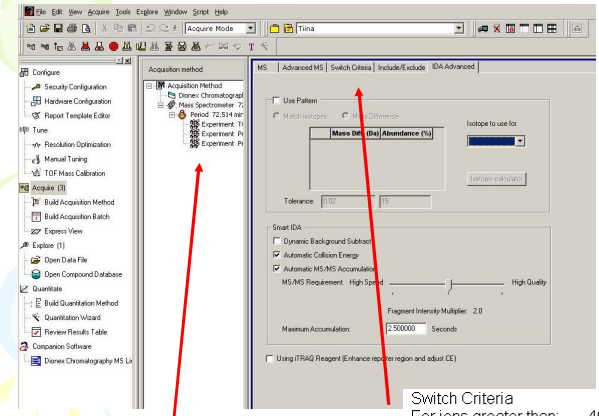
DDA =Data Dependent Acquisition
(IDA =Information Dependent Acquisition)

-fully automated experiment, first MS scan followed by two or more product ion scans

-the acquisition software is set to choose certain types of ions for fragmentation and to use 'suitable' collision energy for this ion

-in ESI tryptic peptides have usually 2-4 charges

DDA experiment

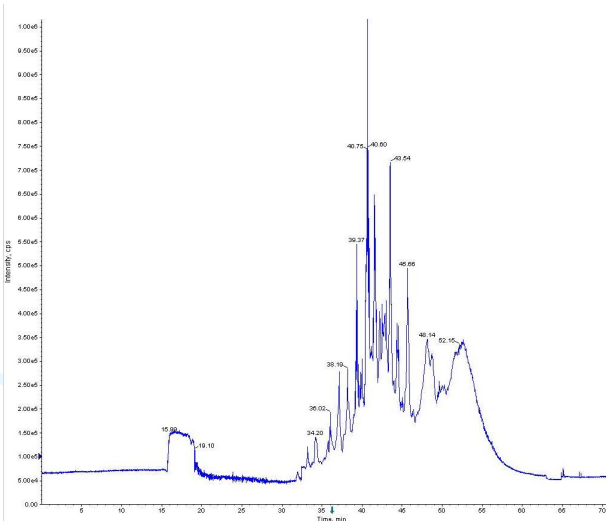


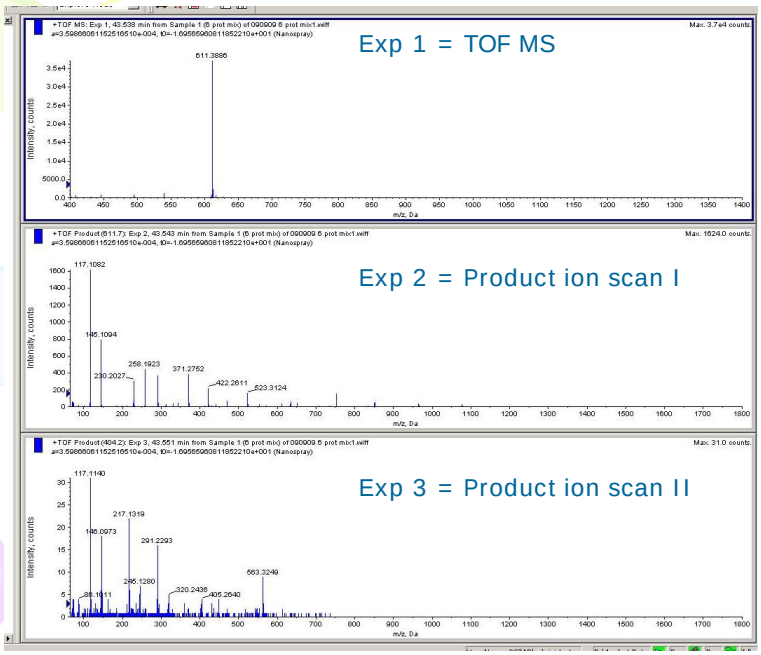
Switch Criteria

For ions greater than:	400.000 m/z	
For ions smaller than:	1400.000 m/z	
With charge state:	2 to 4	
Which exceeds:		10 counts
Exclude former target ions:		For: 60 seconds
Switch after:		1 spectra
Mass Defect Filter:	No	
Ions Tolerance:		50.000 mDa

Exp 1 = TOF MS
Exp 2 = Product ion scan from parent I
Exp 3 = Product ion scan from parent II

TIC = total ion current



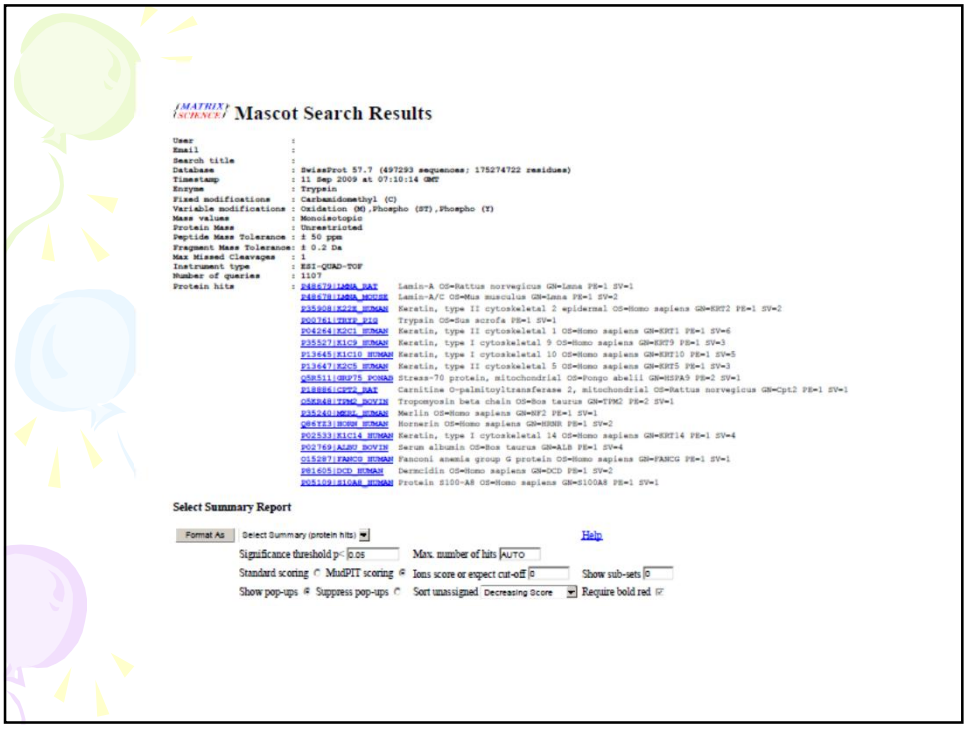


Search engines for (LC-)MS/MS data

-Mascot, Sequest, OMSSA etc
-the programs take the fragment ion spectrum of a peptide as input and score it against theoretical fragmentation patterns constructed for peptides from the searched database.

-in practise the user is often limited to use those search engines which accept the data format from the mass spec used

-mzXML is a open data format for storage and exchange of mass spec data
-raw, proprietary file formats from most vendors can be converted to the open mzXML format



Mascot Search Results

```
User                    :
Email                   :
Search title            :
Database                : SwissProt 57.7 (497293 sequences; 175274722 residues)
Timestamp               : 11 Sep 2009 at 07:10:14 GMT
Enzyme                  : Trypsin
Fixed modifications     : Carbamidomethyl (C)
Variable modifications  : Oxidation (M),Phospho (ST),Phospho (Y)
Mass values             : Monoisotopic
Protein Mass            : Unrestricted
Peptide Mass Tolerance  : ± 50 ppm
Fragment Mass Tolerance : ± 0.2 Da
Max Missed Cleavages    : 1
Instrument type         : ESI-QUAD-TOF
Number of queries       : 1107
Protein hits            : P48679|LMNA_RAT     Lamin-A OS=Rattus norvegicus GN=Lmna PE=1 SV=1
                          P48678|LMNA_MOUSE   Lamin-A/C OS=Mus musculus GN=Lmna PE=1 SV=2
                          P35908|K22E_HUMAN   Keratin, type II cytoskeletal 2 epidermal OS=Homo sapiens GN=KRT2 PE=1 SV=2
                          P00761|TRYP_PIG     Trypsin OS=Sus scrofa PE=1 SV=1
                          P04264|K2C1_HUMAN   Keratin, type II cytoskeletal 1 OS=Homo sapiens GN=KRT1 PE=1 SV=6
                          P35527|K1C9_HUMAN   Keratin, type I cytoskeletal 9 OS=Homo sapiens GN=KRT9 PE=1 SV=3
                          P13645|K1C10_HUMAN  Keratin, type I cytoskeletal 10 OS=Homo sapiens GN=KRT10 PE=1 SV=5
                          P13647|K2C5_HUMAN   Keratin, type II cytoskeletal 5 OS=Homo sapiens GN=KRT5 PE=1 SV=3
                          Q58511|GRP75_PONAB  Stress-70 protein, mitochondrial OS=Pongo abelii GN=HSPA9 PE=2 SV=1
                          P18886|CPT2_RAT     Carnitine O-palmitoyltransferase 2, mitochondrial OS=Rattus norvegicus GN=Cpt2 PE=1 SV=1
                          O5E848|TPM2_BOVIN   Tropomyosin beta chain OS=Bos taurus GN=TPM2 PE=2 SV=1
                          P35240|MERL_HUMAN   Merlin OS=Homo sapiens GN=NF2 PE=1 SV=1
                          Q86YZ3|HORN_HUMAN   Hornerin OS=Homo sapiens GN=HRNR PE=1 SV=2
                          P02533|K1C14_HUMAN  Keratin, type I cytoskeletal 14 OS=Homo sapiens GN=KRT14 PE=1 SV=4
                          P02769|ALBU_BOVIN   Serum albumin OS=Bos taurus GN=ALB PE=1 SV=4
                          O15287|FANCG_HUMAN  Fanconi anemia group G protein OS=Homo sapiens GN=FANCG PE=1 SV=1
                          P81605|DCD_HUMAN    Dermcidin OS=Homo sapiens GN=DCD PE=1 SV=2
                          P05109|S10A8_HUMAN  Protein S100-A8 OS=Homo sapiens GN=S100A8 PE=1 SV=1
```

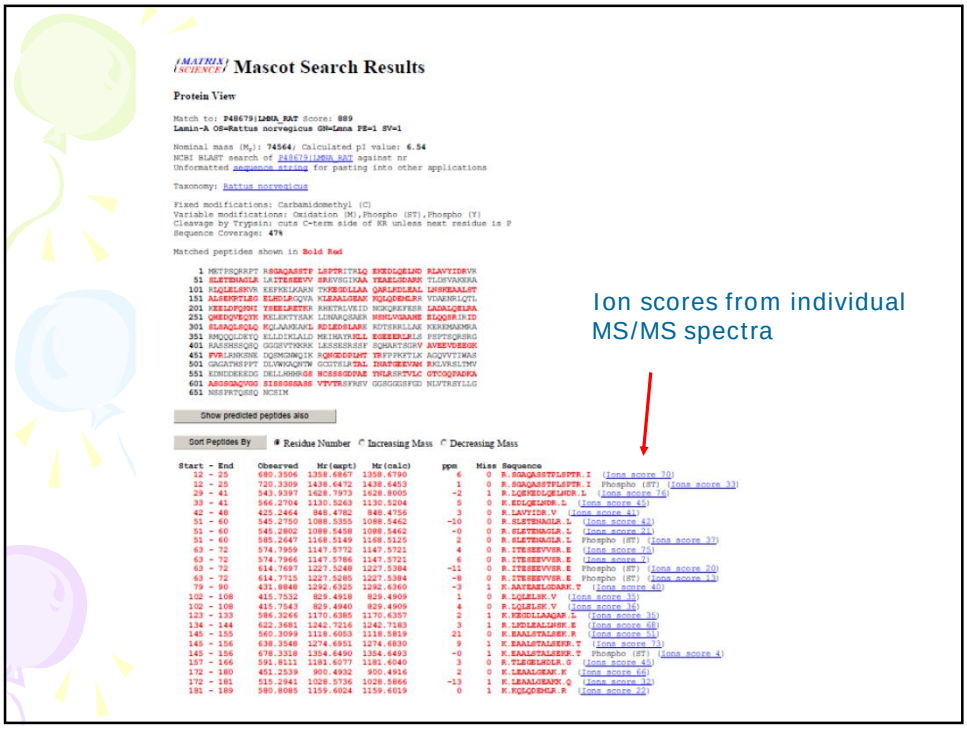
Select Summary Report

Format As | Select Summary (protein hits) | Help

Significance threshold p< 0.05 | Max. number of hits AUTO

Standard scoring ○ MudPIT scoring ● Ions score or expect cut-off 0 | Show sub-sets 0

Show pop-ups ● Suppress pop-ups ○ | Sort unassigned Decreasing Score | Require bold red ☑



Mascot Search Results

Protein View

```
Match to: P48679|LMNA_RAT Score: 889
Lamin-A OS=Rattus norvegicus GN=Lmna PE=1 SV=1

Nominal mass (Mr): 74564; Calculated pI value: 6.54
NCBI BLAST search of P48679|LMNA_RAT against nr
Unformatted sequence string for pasting into other applications

Taxonomy: Rattus norvegicus

Fixed modifications: Carbamidomethyl (C)
Variable modifications: Oxidation (M),Phospho (ST),Phospho (Y)
Cleavage by Trypsin: cuts C-term side of KR unless next residue is P
Sequence Coverage: 47%

Matched peptides shown in Bold Red

    1 METPSQRRFT RSGAQASSTP LSPTRITRLQ EKEDLQELND RLAVYIDRVR
   51 SLETENAGLR LRITESEEVV SREVSGIKAA YEAELGDARK TLDSVAKERA
  101 RLQLELSKVR EEFKELKARN TKKEGDLIAA QARLKDLEAL LNSKEAALST
  151 ALSEVFTLRS RLEDLSQQVA VLEAALGEAK KQLQDERSRL VDAQNRLQTL
  201 KEELDFQTNI YSEELRETKR RHETRLVEID NGKQREFESR LADALQELRA
  251 QHEDQVEQYK KELEKTYSAK LDNARQSAER NSNLVGAAHE ELQQSRIRID
  301 SLAAQLSQLQ KQLAAKEAKL KDLEDSLARE KDTSNKLLAS KERSMSDNVA
  351 RMQQQLDEYQ KLLDIKLALD MEIHAYRKLL EGEESRLSLS FSPTSQRSRG
  401 RASSHSSQSQ GGGSVTKKRK LESSESRSSF SQHARTSGRV AVEEVDEEGK
  451 FVRLRNKSNE DQSMGNWQIK RQNGDDPLMT YKYPKVTLKN ACQVVTVWAS
  501 GNGATHSPPT DLVWKAQNTW GCGTSLYTAL INATGEEVAM RKLVRSLTMV
  551 KDNDDEKEDG DELLHHHRGS NCSSSGDPAE YNLRSRTVLC GTCGQPADKA
  601 ASSSQAQVQG SISSGSSASS VTVTRSFRSV GSSGGGSFGD NLVTRSYLLG
  651 NSSPRTQRSQ NCSIM
```

Show predicted peptides also

Sort Peptides By | ● Residue Number ○ Increasing Mass ○ Decreasing Mass

Start - End	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Sequence
12 - 25	680.3506	1358.6867	1358.6790	6	0	R.SGAQASSTPLSPTR.I (Ions score 70)
12 - 25	720.3309	1438.6472	1438.6453	1	0	R.SGAQASSTPLSPTR.I Phospho (ST) (Ions score 33)
29 - 41	543.9397	1628.7973	1628.8005	-2	1	R.LQKEKDLQELNDR.L (Ions score 76)
33 - 41	566.2704	1130.5263	1130.5204	5	0	K.EDLQELNDR.L (Ions score 45)
42 - 48	425.2464	848.4782	848.4756	3	0	R.LAVYIDR.V (Ions score 41)
51 - 60	545.2750	1088.5355	1088.5462	-10	0	R.SLETENAGLR.L (Ions score 42)
51 - 60	545.2802	1088.5458	1088.5462	-0	0	R.SLETENAGLR.L (Ions score 21)
51 - 60	585.2647	1168.5149	1168.5125	2	0	R.SLETENAGLR.L Phospho (ST) (Ions score 37)
63 - 72	574.7959	1147.5772	1147.5721	4	0	R.ITESEEVVSR.E (Ions score 75)
63 - 72	574.7966	1147.5786	1147.5721	6	0	R.ITESEEVVSR.E (Ions score 7)
63 - 72	614.7697	1227.5248	1227.5384	-11	0	R.ITESEEVVSR.E Phospho (ST) (Ions score 20)
63 - 72	614.7715	1227.5285	1227.5384	-8	0	R.ITESEEVVSR.E Phospho (ST) (Ions score 13)
79 - 90	431.8848	1292.6325	1292.6360	-3	1	K.AAYEAELGDARK.T (Ions score 40)
102 - 108	415.7532	829.4918	829.4909	1	0	R.LQLELSK.V (Ions score 35)
102 - 108	415.7543	829.4940	829.4909	4	0	R.LQLELSK.V (Ions score 36)
123 - 133	586.3266	1170.6385	1170.6357	2	1	K.EKDLLAAQAR.L (Ions score 35)
134 - 144	622.3681	1242.7216	1242.7183	3	1	R.LNDLEALLNSK.E (Ions score 68)
145 - 155	560.3099	1118.6053	1118.5819	21	0	K.EAALSTALSEK.R (Ions score 51)
145 - 156	638.3548	1274.6951	1274.6830	9	1	K.EAALSTALSEKR.T (Ions score 73)
145 - 156	678.3318	1354.6490	1354.6493	-0	1	K.EAALSTALSEKR.T Phospho (ST) (Ions score 4)
157 - 166	591.8111	1181.6077	1181.6040	3	0	R.TLEGELHDLR.G (Ions score 45)
172 - 180	451.2539	900.4932	900.4916	2	0	K.LEAALGEAK.K (Ions score 66)
172 - 181	515.2941	1028.5736	1028.5866	-13	1	K.LEAALGEAKK.Q (Ions score 32)
181 - 189	580.8085	1159.6024	1159.6019	0	1	K.KQLQDEHLR.R (Ions score 22)

Ion scores from individual MS/MS spectra

Protein identification from complex mixtures using nanoLC-MS/MS

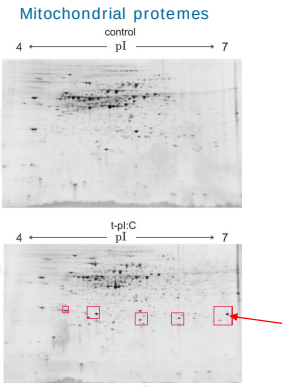
-produces huge amounts of raw data
-requires efficient data processing tools
-different database search programs can produce different results from the same raw data
-false discovery rate estimation
    -peptide level
    -protein level

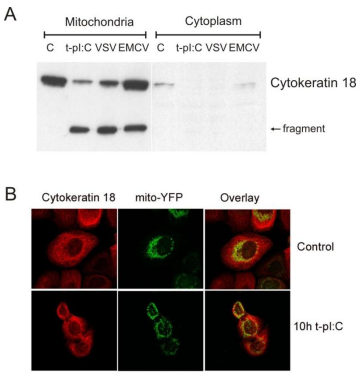
Now you're protein is identified
What else can we find in the data?

Protein fragmentation

-protein separation by SDS-PAGE/2-DE
followed by in-gel digestion and MS analysis

Cytokeratin-18 is cleaved during viral infection, and fragments localize onto mitochondria





Mascot Search Results

Protein View

Match to: **P05783|K1C18_HUMAN** Score: **927**
Keratin, type I cytoskeletal 18 OS=Homo sapiens GN=KRT18 PE=1 SV=2

Nominal mass (M_r): **48029**; Calculated pI value: **5.34**
NCBI BLAST search of P05783|K1C18_HUMAN against nr
Unformatted sequence string for pasting into other applications

Taxonomy: Homo sapiens

Fixed modifications: Carbamidomethyl (C)
Variable modifications: Oxidation (M)
Cleavage by Trypsin: cuts C-term side of KR unless next residue is P
Sequence Coverage: **40%**

Matched peptides shown in **Bold Red**

```
    1 MSFTTRSTFS TNYRSLGSVQ APSYGARPVS SAASVYAGAG GSGSRISVSR
   51 STSFRGGMGS GGLATGIAGG LAGMGGIQNE KETMQSLNDR LASYLDRVRS
  101 LETENRRLES KIREHLEKKG PQVRDWSHYF KIIEDLRAQI FANTVDNARI
  151 VLQIDNARLA ADDFRVKYET ELAMRQSVEN DIHGLRKVID DTNITRLQLE
  201 TEIEALKEEL LFMKKNHEEE VKGLQAQIAS SGLTVEVDAP KSQDLAKIMA
  251 DIRAQYDELA RKNREELDKY WSQQIEESTT VVTTQSAEVG AAETTLTELR
  301 RTVQSLEIDL DSMRNLKASL ENSLREVEAR YALQMEQLNG ILLHLESELA
  351 QTRAEGQRQA QEYEALLNIK VKLEAEIATY RRLLEDGEDF NLGDALDSSN
  401 SMQTIQKTTT RRIVDGKVVS ETNDTKVLRH
```

Show predicted peptides also

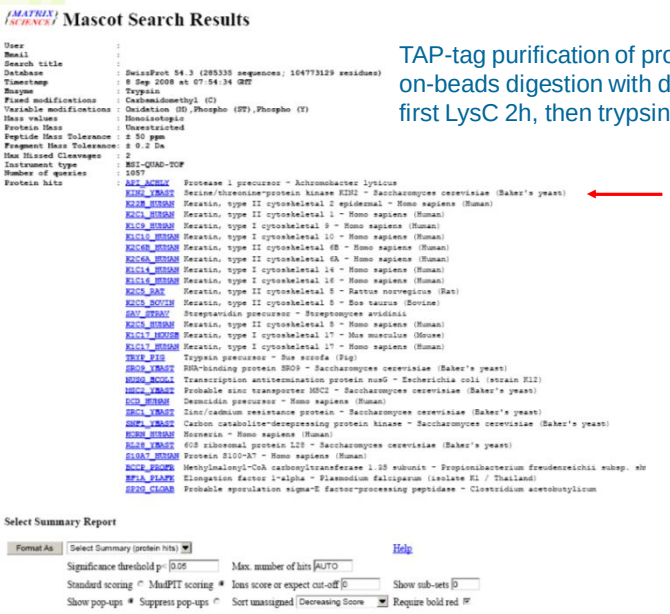
Sort Peptides By | Residue Number | Increasing Mass | Decreasing Mass

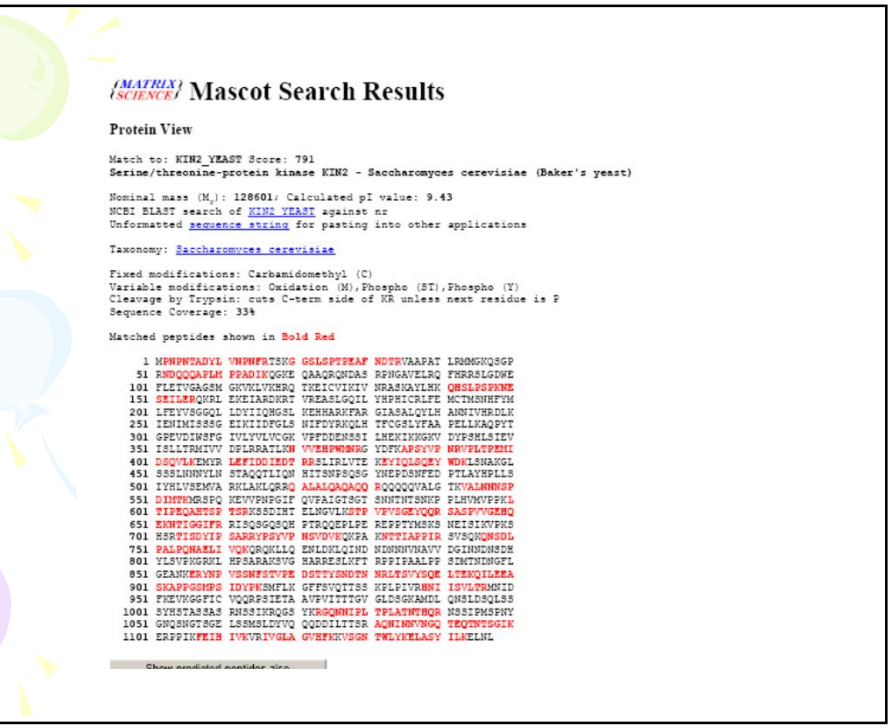
Start - End	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Sequence
7 - 14	488.2270	974.4394	974.4458	-0.0063	0	R.STFSTNYR.S (Ions score 52)
7 - 14	488.2307	974.4468	974.4458	0.0010	0	R.STFSTNYR.S (Ions score 30)
7 - 14	488.2311	974.4476	974.4458	0.0018	0	R.STFSTNYR.S (Ions score 3)
7 - 14	488.2316	974.4486	974.4458	0.0028	0	R.STFSTNYR.S (Ions score 22)
7 - 14	488.2481	974.4816	974.4458	0.0358	0	R.STFSTNYR.S (Ions score 41)
15 - 45	952.1442	2853.4109	2853.4005	0.0103	0	R.SLGSVQAPSYGARPVSSAASVYAGAGGSGSR.I (Ions score 86)
56 - 81	765.0267	2292.0581	2292.0838	-0.0257	0	R.GGMGSGGLATGIAGGLAGMGGIQNEK.E 2 Oxidation (M) (Ions score 136)
56 - 81	765.0359	2292.0858	2292.0838	0.0020	0	R.GGMGSGGLATGIAGGLAGMGGIQNEK.E 2 Oxidation (M) (Ions score 75)
56 - 90	846.6393	3382.5279	3382.5552	-0.0272	1	R.GGMGSGGLATGIAGGLAGMGGIQNEKETMQSLNDR.L 3 Oxidation (M) (Ions score
56 - 90	1128.5232	3382.5479	3382.5552	-0.0073	1	R.GGMGSGGLATGIAGGLAGMGGIQNEKETMQSLNDR.L 3 Oxidation (M) (Ions score
91 - 97	419.1995	836.3845	836.4392	-0.0547	0	R.LASYLDR.V (Ions score 5)
91 - 97	419.2254	836.4362	836.4392	-0.0030	0	R.LASYLDR.V (Ions score 14)
91 - 97	419.2257	836.4368	836.4392	-0.0024	0	R.LASYLDR.V (Ions score 57)
91 - 97	419.2261	836.4377	836.4392	-0.0015	0	R.LASYLDR.V (Ions score 15)
100 - 107	502.7557	1003.4969	1003.5046	-0.0077	1	R.SLETENRR.L (Ions score 1)

Based on 2-DE and MS data: N-terminal fragment of the identified protein

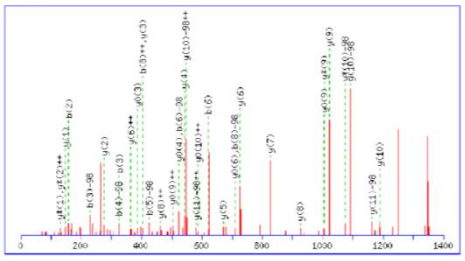
Post-translational modifications

TAP-tag purification of protein complexes, on-beads digestion with double enzyme, first LysC 2h, then trypsin o/n





Start - End	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Sequence
2 - 16	866.4372	1730.8598	1730.8376	13	0	M.PNPNTADYLVNPNFR.T (Ions score 82)
20 - 34	774.8776	1547.7406	1547.7216	12	0	K.GGSLSPPPEAFNDTR.V (Ions score 66)
52 - 66	833.4293	1664.8441	1664.8192	15	0	R.NDQQQAPLMPPADIK.Q (Ions score 57)
141 - 148	447.2544	892.4942	892.4767	20	0	K.QHSLPSPK.N (Ions score 10)
141 - 148	487.2387	972.4629	972.4430	21	0	K.QHSLPSPK.N Phospho (ST) (Ions score 5)
149 - 156	495.2514	988.4883	988.4825	6	0	K.NESEILER.Q (Ions score 65)
370 - 379	427.8884	1280.6435	1280.6084	27	0	K.NVVEHPWHNR.G (Ions score 22)
385 - 392	452.2459	902.4773	902.4610	18	0	K.APSYVPNR.V (Ions score 52)
393 - 406	785.4494	1568.8843	1568.8484	23	0	R.VPLTPEMIDSQVLK.E (Ions score 48)
411 - 422	507.9303	1520.7691	1520.7471	15	1	R.LEFIDDIEDTRR.S (Ions score 51)
432 - 443	801.3818	1600.7491	1600.7409	5	0	K.EYIQLSQEYWDK.L (Ions score 36)
520 - 531	663.3760	1324.7375	1324.7211	12	0	R.QALALQAQAQQR.Q (Ions score 96)
543 - 555	708.8656	1415.7166	1415.7078	6	0	K.VALNNNSPDIMTK.M (Ions score 91)
543 - 555	716.8680	1431.7215	1431.7028	13	0	K.VALNNNSPDIMTK.M Oxidation (M) (Ions score 106)
600 - 613	513.2744	1536.8014	1536.7896	8	0	K.LTIPEQAHTSPTSR.K (Ions score 29)
628 - 640	724.3667	1446.7189	1446.7103	6	0	K.STPVPVSGEYQQR.S (Ions score 32)
641 - 652	634.3262	1266.6378	1266.6204	14	0	R.SASPVVGEHQEK.N (Ions score 44)
641 - 652	423.2242	1266.6507	1266.6204	24	0	R.SASPVVGEHQEK.N (Ions score 58)
641 - 652	674.3103	1346.6061	1346.5867	14	0	R.SASPVVGEHQEK.N Phospho (ST) (Ions score 81)
653 - 660	439.2546	876.4946	876.4818	15	0	K.NTIGGIFR.R (Ions score 31)
704 - 713	561.7996	1121.5847	1121.5717	12	0	R.TISDYIPSAR.R (Ions score 47)
714 - 726	508.6109	1522.8110	1522.7780	22	1	R.RYPSVPNSVDVK.Q (Ions score 10)
732 - 740	491.7939	981.5732	981.5607	13	0	K.NTTIAPPIR.S (Ions score 24)
746 - 763	989.5486	1977.0826	1977.0531	15	0	K.QNSDLPALPQNAELIVQK.Q (Ions score 50)
856 - 882	1032.1285	3093.3637	3093.3548	3	1	K.ERYNPVSSNFSTVPEDSTTYSNDTNNR.L (Ions score 62)
883 - 894	699.3609	1396.7073	1396.7086	-1	0	R.LTSVYSQELTEK.Q (Ions score 68)
895 - 902	459.2582	916.5018	916.4865	17	0	K.QILEEASK.A (Ions score 42)
903 - 915	680.3405	1358.6664	1358.6540	9	0	K.APPGSMPSIDYPK.S (Ions score 51)
903 - 915	688.3393	1374.6640	1374.6489	11	0	K.APPGSMPSIDYPK.S Oxidation (M) (Ions score 27)
938 - 946	526.8238	1051.6330	1051.6138	18	0	R.EHIISVLTR.M (Ions score 52)
1023 - 1040	677.7172	2030.1299	2030.0769	26	1	K.RGQNNIPLTPLATNTHQR.N (Ions score 69)
1023 - 1040	508.5410	2030.1347	2030.0769	28	1	K.RGQNNIPLTPLATNTHQR.N (Ions score 14)
1081 - 1100	1066.0333	2130.0521	2130.0301	10	0	R.AQNINVVNGQTEQTNTPGIK.E (Ions score 127)
1107 - 1113	443.2667	884.5188	884.5120	8	0	K.FEIHIVK.V (Ions score 27)
1116 - 1125	520.8239	1039.6332	1039.6179	15	0	R.IVGLAGVHFK.K (Ions score 39)
1127 - 1135	534.2873	1066.5600	1066.5447	14	0	K.VSGNTWLYK.E (Ions score 62)
1136 - 1143	468.7756	935.5367	935.5327	4	0	K.ELASYILK.E (Ions score 38)



KIN2 phosphopeptide

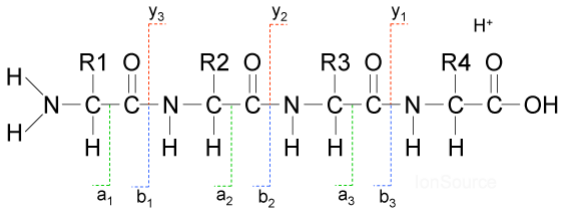
Monoisotopic mass of neutral peptide Mr(calc): 1346.5867
Fixed modifications: Carbamidomethyl (C)
Variable modifications:
S3 : Phospho (ST), with neutral losses 97.9769(shown in table), 0.0000
Ions Score: 81 **Expect:** 1.8e-005
Matches (Bold Red): 35/164 fragment ions using 57 most intense peaks

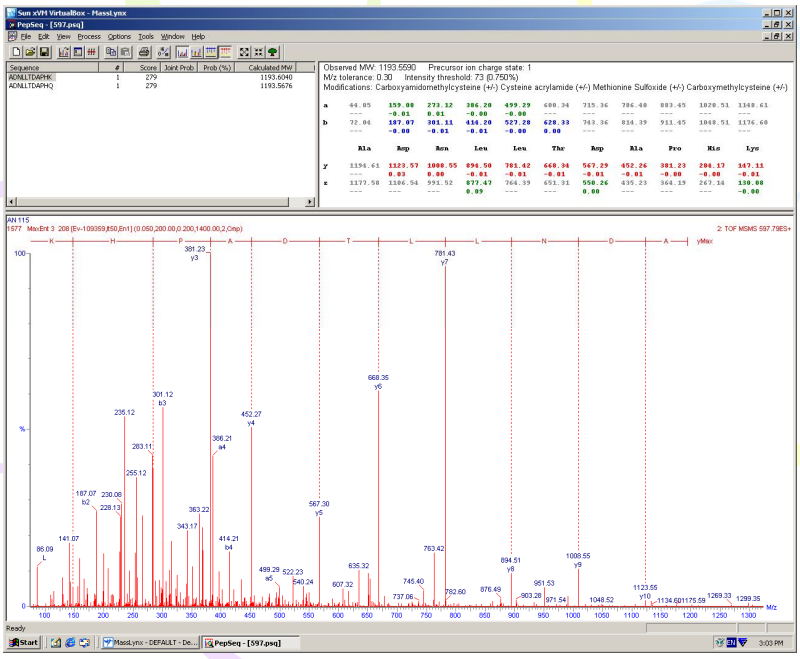
#	b	b^{++}	b^{*}	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	88.0393	44.5233			70.0287	35.5180	S							12
2	**159.0764**	80.0418			141.0659	71.0366	A	**1162.5851**	**581.7962**	1145.5586	573.2829	1144.5745	572.7909	11
3	**228.0979**	114.5526			210.0873	105.5473	S	**1091.5480**	**546.2776**	**1074.5214**	537.7644	1073.5374	537.2724	10
4	**325.1506**	163.0790			307.1401	154.0737	P	**1022.5265**	511.7669	**1005.5000**	503.2536	**1004.5160**	**502.7616**	9
5	**424.2191**	212.6132			406.2085	203.6079	V	**925.4738**	**463.2405**	908.4472	454.7272	907.4632	454.2352	8
6	**523.2875**	262.1474			505.2769	253.1421	V	**826.4054**	413.7063	809.3788	405.1930	808.3948	404.7010	7
7	580.3089	290.6581			562.2984	281.6528	G	**727.3369**	**364.1721**	710.3104	355.6588	**709.3264**	355.1668	6
8	**709.3515**	355.1794			691.3410	346.1741	E	**670.3155**	335.6614	653.2889	327.1481	652.3049	326.6561	5
9	846.4104	423.7089			828.3999	414.7036	H	**541.2729**	271.1401	524.2463	262.6268	**523.2623**	262.1348	4
10	974.4690	487.7381	957.4425	479.2249	956.4584	478.7329	Q	**404.2140**	202.6106	387.1874	194.0974	**386.2034**	193.6053	3
11	1103.5116	552.2594	1086.4851	543.7462	1085.5010	543.2542	E	**276.1554**	138.5813	259.1288	**130.0681**	258.1448	129.5761	2
12							K	**147.1128**	74.0600	**130.0863**	65.5468			1

De novo sequencing

De novo = peptide sequencing performed without prior knowledge of the amino acid sequence

-if enough material is available classical Edman degradation is still a vey good method
-partial peptide sequencing is possible based on MS/MS data





Publishing protein ID data

-most biological journals do not have any guidelines yet

Mat+Met:

- -MS instrument, identification type
- -data acquisition+processing software
- -database search engine

Results: Table from ID results

Protein name	Fold difference	Protein number	Access no. SwissProt	Access no. NCBInr	Identification type	No. of peptide matched	Sequence coverage	Mowse score	Theoretical MW	Theoretical pI	Biological process
Upregulated proteins											
14-3-3 protein sigma (epithelial cellmarker 1 / stratifin)	x	73[a]	P31947	187302	LC-MS/MS	10	32 %	216	27873	4,72	Signaling
cytokeratin 13 (frag.)			P13646	34033	LC-MS/MS	11	22 %	142	49956	4,87	Structural
14-3-3 protein sigma (epithelial cellmarker 1 / stratifin)	x	74[a]	P31947	187302	LC-MS/MS	12	39 %	444	27873	4,72	Signaling
cytokeratin 13 (frag.)			P13646	30377	LC-MS/MS	11	22 %	372	46179	4,83	Structural

Tutorial for Proteomics Data Submission

Katalin F. Medzihradszky
Robert J. Chalkley
UCSF

Mol Cell Proteomics
www.mcponline.org

Why Have Guidelines?

- Large-scale proteomics studies create huge amounts of data.
- It is impossible/impractical to present all the results.
- The function of the guidelines is to:
  - Provide enough information to be able to explain the experiment.
  - Provide an assessment of the reliability of the results.
  - Provide the data that supports the results, particularly those that have the greatest potential for mis-interpretation, so the readers can manually assess the results that are important to them.

The Problem

- Proteomics experiments are carried out by many different methods, using a variety of instrument types and employing different analysis tools. Hence, many experimental and analysis parameters need to be reported and the parameters required will differ depending on how the experiment was performed and analyzed.

In the experimental section:

Peak Picking Software

In the Experimental Section

- Name of peaklist-generating software and release version (number or date)
- The raw data acquired by the mass spectrometer is converted into a centroided peaklist file for database searching. The program that performs this can do many things, such as:
  - Remove peaks that are below a certain intensity or signal to noise ratio;
  - Require peaks to have a certain resolution in order to make the list;
  - Set a threshold for the maximum number of peaks within a mass range;
  - Assign a charge state to ions;
  - Merge together MS/MS spectra that have the same precursor mass.

Peak Picking (cont'd)

In the Experimental Section

- Parameters used – default vs altered

- Using different peak picking software / parameters will change the subsequent database searching results, so it is necessary to report the software used for peaklist generation and parameters used.

Examples of software

- Extract_msn in Bioworks 3.0 (Thermo)
- Mascot.dll v1.6b19 (Applied Biosystems)
- DataAnalysis 3.2 (Bruker)
- Mascot Distiller v2.1

- As many people use default parameters for whatever software they use, by specifying the version of the software it can be sufficient to state that the default parameters were used.

Search Engine

In the Experimental Section

- Name of the search engine and release version (number or date)

- Search engines change over time:
  - An improved scoring system may be implemented.
  - New ways to filter the results may be included.

Hence, the version number of the search engine is important.

Database Search Parameters

In the Experimental Section

- Enzyme specificity considered
- # of missed cleavages permitted
- Fixed modification(s) (including residue specificity)
- Variable modification(s) (including residue specificity)
- Mass tolerance for precursor ions
- Mass tolerance for fragment ions

Search engines work by:

1. Determining a list of potential peptides from a database that can be formed by the specified parameters and have the correct precursor mass.
2. Determining scores for the matches of the fragment peaks to each of these peptides.

- Changing search parameters will change the number of the potential hits.
- For software using probabilities or Expectation values (E-values) for scoring, it will change the scores.

Single-Peptide-Based Protein IDs and PTMs

In the Results Section

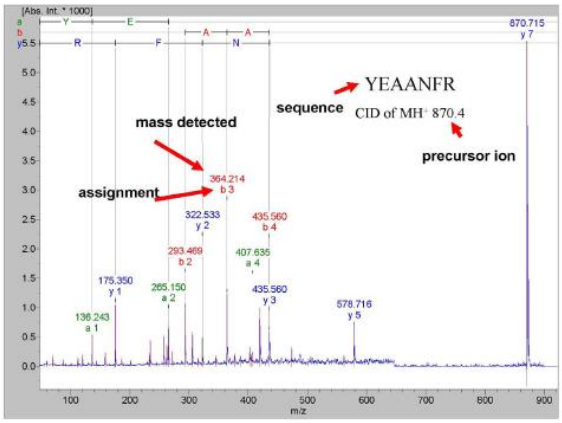
- For each protein or site, a MS/MS spectrum appropriately labeled should be included, with masses detected as well as fragment assignments

- It is recognized that properly labeled spectra are not always readily produced.
- Other acceptable options (in order of preference):
  - Two copies of the same spectrum/page – one labeled with the masses, the other with fragment assignments;
  - Spectrum labeled with masses, accompanied with a Table of the fragment assignments;
  - Fragment assignments provided by the search engine in the spectrum and corresponding masses highlighted in a Table of theoretical fragments (e.g. Mascot results output).

These files can be quite large. If there are large numbers of spectra it is often easier to split them between multiple files. We suggest to not include data on more than 50 proteins in a single file.

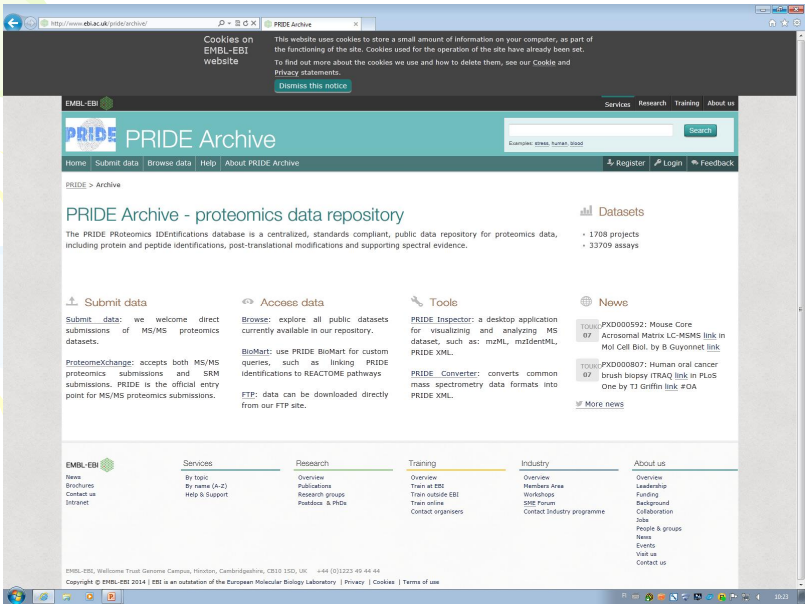
In the Results Section

Example Presentation of Spectrum of a Single-Peptide-Based Protein ID



The Future

- These guidelines will be continuously adjusted to the ever changing needs of the proteomics field.
- As software improves to become easier to output results into compliant formats, it will become easier to properly present proteomics data in any forum.
- As common formats e.g. mzML, AnalysisXML, become universally available, tools will be produced that will be able to automatically extract the relevant information from raw data and search results.
- Once these formatting problems are solved, journal submission guidelines for proteomics data are expected to become similar for all journals.
- Making raw data publicly available is going to become more common.



http://www.ebi.ac.uk/pride/archive/