Protein MW determination and protein identification by mass spectrometry

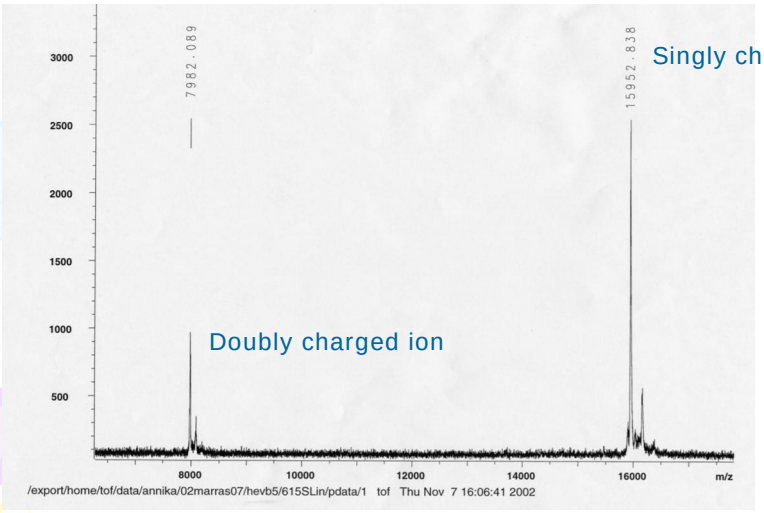
Tuula Nyman
Protein Chemistry Research Group
Institute of Biotechnology
tuula.nyman@helsinki.fi

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

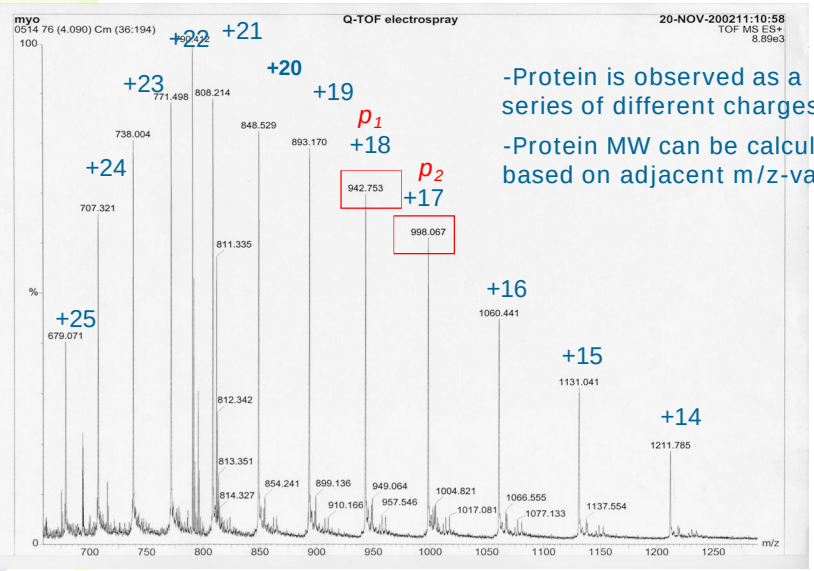
-MALDI TOF MS, linear mode
-accuracy is not as good with ESI MS



Singly charged ion

Doubly charged ion

Protein MW determination, ESI MS



p_1

p_2

942.753

998.067

-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

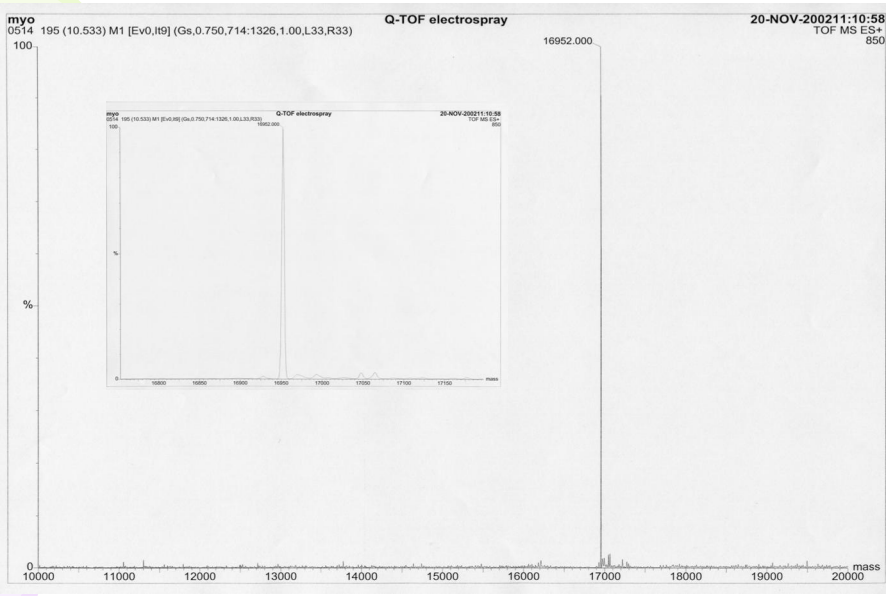
$p = m/z$
$p_1 = (M_r + z_1)/z_1$
$p_2 = [M_r + (z_1 - 1)]/(z_1 - 1)$

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

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

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

$$M_r = 941.753 z_1 = 16\,951.6 \text{ Da}$$



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)

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

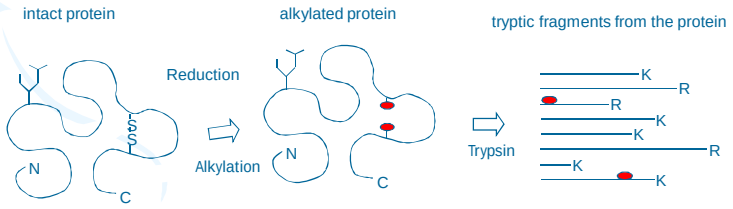
Database search engines create theoretical PMFs for all the proteins in the database and compare these to the measured PMF from protein X

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:

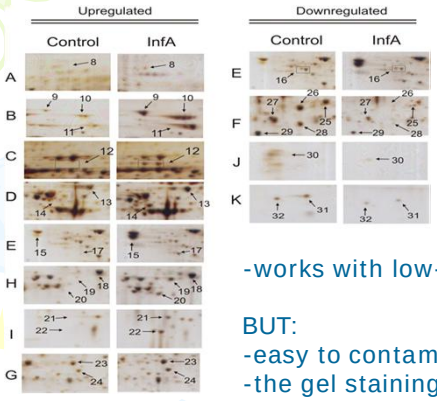
-LysC, cleaves after lysines
-LysN, cleaves before lysines
-AspN, cleaves before aspartic acid residues
-V8 protease, cleaves peptide bonds exclusively on the carbonyl side of aspartate and glutamate residues

-possible to do double-digestions

Chemical cleavage:

Cyanogen bromide, cuts after methionines

In-gel digestion

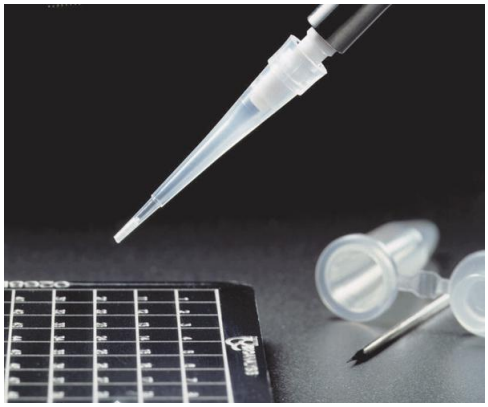


-works with low-femtomolar amounts of proteins

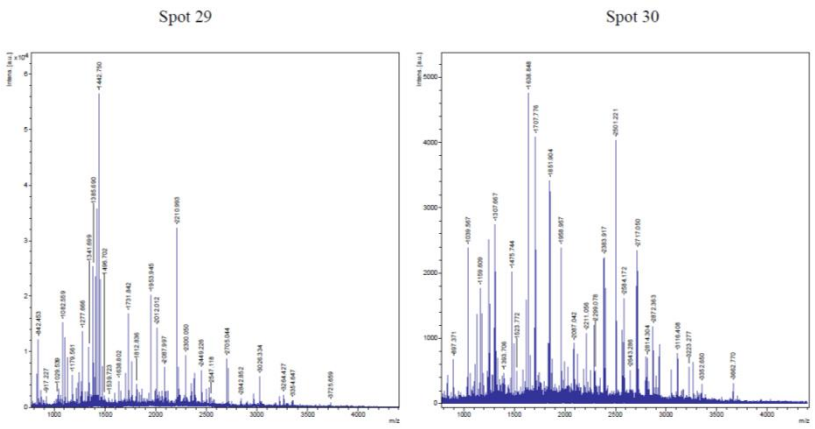
BUT:
-easy to contaminate samples with keratin
-the gel staining protocol needs to be compatible with MS
-silver-staining: fixing with glutaraldehy cross-links protein into gel matrix after which identification is not possible

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

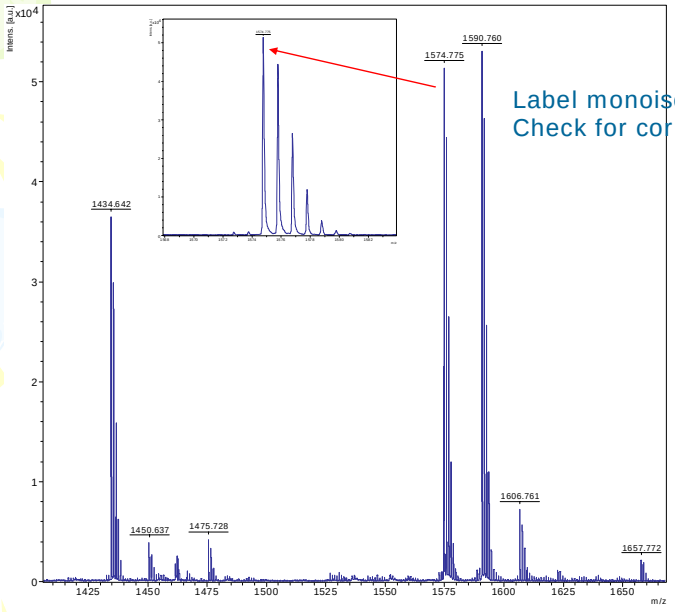


Peptide mass fingerprinting
MALDI, Positive ion reflector mode



-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

MASCOT Peptide Mass Fingerprint

Your name	Tuula	Email	tuula.nyman@helsinki.fi
Search title	dig 33		
Database	NCBInr		
Taxonomy	. . . . . Homo sapiens (human)		
Enzyme	Trypsin	Allow up to	1 missed cleavages
Fixed modifications	Biotin (K) Biotin (N-term) Carbamidomethyl (C) Carbamyl (K) Carbamyl (N-term)	Variable modifications	Oxidation (HW) Oxidation (M) Phospho (ST) Phospho (Y) Propionamide (C)
Protein mass	kDa	Peptide tol. ±	50 ppm
Mass values	MH+ M_r M-H+	Monoisotopic	Average
Data file			
Query	842.482722918522 870.515149550309 1085.62503611115 1174.53680038973 1265.55776926686 1393.75672200038		
Decoy		Report top	AUTO hits
	Start Search ...		Reset Form

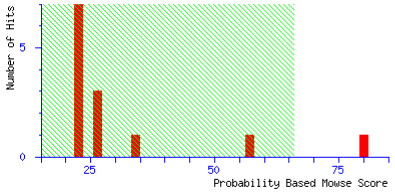
www.matrixscience.com

{MATRIX SCIENCE} 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]

MATRIX SCIENCE 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
```

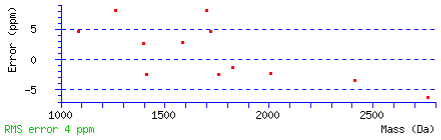
Show predicted peptides also

Sort Peptides By ◉ Residue Number ○ Increasing Mass ○ Decreasing Mass

Usually not many missed cleavages

Start - End	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Sequence
6 - 25	2007.8886	2006.8813	2006.8858	-2	0	K.TIGGGDDSFNTFFSETGAGK.H
30 - 44	1701.9195	1700.9122	1700.8985	8	0	R.AVFVDLEPTVIDEVR.T
50 - 61	1410.7704	1409.7631	1409.7667	-3	0	R.QLFHPEQLITGK.E
50 - 70	2415.1967	2414.1894	2414.1978	-3	1	R.QLFHPEQLITGKEDAANNYAR.G
78 - 86	1085.6250	1084.6178	1084.6128	5	0	K.EIIDLVLDR.I
181 - 194	1718.8900	1717.8827	1717.8747	5	0	R.NLDIERPTYTNLNR.L
230 - 245	1756.9589	1755.9517	1755.9559	-2	0	R.IHFPLATYAPVISAEK.A
246 - 269	2766.2690	2765.2617	2765.2789	-6	0	K.AYHEQLSVAEITNACFEPANQMVK.C Oxidation (M)
277 - 285	1265.5578	1264.5505	1264.5403	8	0	K.YMACCLLYR.G Oxidation (M)
305 - 317	1584.7560	1583.7487	1583.7443	3	0	R.SIQFVDWCPTGFK.V
318 - 335	1824.9830	1823.9757	1823.9782	-1	0	K.VGINYQPPTVVPGGDLAK.V
356 - 366	1396.6967	1395.6894	1395.6857	3	1	R.LDHKFDLMYAK.R Oxidation (M)

No match to: 842.4827, 870.5151, 1174.5368, 1393.7567, 1616.7599, 1715.9463, 1778.9576, 1846.9822, 2029.8757, 2211



Mass accuracy critical

-check for known contaminants
(these should be removed before search)
-possibility to do 2nd pass search

'Normal' search,
max 1missed cleavage
allowed

Match to: TPM2_HUMAN Score: 61 Expect: 0.43
Tropomyosin beta chain OS=Homo sapiens GN=TPM2 PE=1 SV=1

Nominal mass (M_r): 32945; Calculated pI value: 4.66
NCBI BLAST search of TPM2_HUMAN 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: 41
Number of mass values matched: 13
Sequence Coverage: 33%

Matched peptides shown in Bold Red

1 MDAIKKHKQM LKLDKENAID RAEQAEADKK QAEDRCKQLE KEQQALQKKL
51 KGTEDEVEKY SESVKEAQEK LEQAEKKATD AEADVASLNR RIQLVEEELD
101 RAQERLATAL QKLEEAEKAA DESERGMKVI ENRAMKDEEK MELQEMQLKE
151 AKHIAEDSDR KYEEVARKLV ILEGELERSE ERAEVAESKC GDLEEELKIV
201 TNNLKSLEAQ ADKYSTKEDK YEEEIKLLEE KLKEAETRAE FAERSVAKLE
251 KTIDDLEDEV YAQKMKYKAI SEELDNALND ITSL

Max 4 missed cleavages allowed

Protein View

Match to: TPM2_BOVIN Score: 120 Expect: 5e-07
Tropomyosin beta chain OS=Bos taurus GN=TPM2 PE=2 SV=1

Nominal mass (M_r): 32931; Calculated pI value: 4.66
NCBI BLAST search of TPM2_BOVIN against nr
Unformatted sequence string for pasting into other applications

Taxonomy: Bos taurus

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: 41
Number of mass values matched: 31
Sequence Coverage: 75%

Matched peptides shown in Bold Red

1 MDAIKKHKQM LKLDKENAID RAEQAEADKK QAEDRCKQLE KEQQALQKKL
51 KGTEDEVEKY SESVKDAQEK LEQAEKKATD AEADVASLNR RIQLVEEELD
101 RAQERLATAL QKLEEAEKAA DESERGMKVI ENRAMKDEEK MELQEMQLKE
151 AKHIAEDSDR KYEEVARKLV ILEGELERSE ERAEVAESKC GDLEEELKIV
201 TNNLKSLEAQ ADKYSTKEDK YEEEIKLLEE KLKEAETRAE FAERSVAKLE
251 KTIDDLEDEV YAQRMKYKAI SEELDNALND ITSL

Start - End	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Sequence
13 - 36	2043.7670	2042.7597	2043.0232	-129	3	K.LDKENAIDRAEQAEADKK.Q
36 - 48	1631.5830	1630.5757	1630.7984	-137	1	R.CKQLEEKQQALQK.K
38 - 48	1343.4900	1342.4827	1342.6728	-142	0	K.QLEEKQQALQK.K
49 - 59	1355.6060	1354.5987	1354.6381	-29	2	K.KLKGTEDEVEK.Y Phospho (ST)
52 - 76	2868.9540	2867.9467	2868.3512	-127	3	K.GTEDEVEKYSESVKDAQEKLEQAEK.K
60 - 76	1443.6670	1442.5597	1442.5367	44	1	K.YSESVKDAQEK.L Phospho (ST); Pho
77 - 90	1460.5220	1459.5247	1459.7267	-138	1	K.KATDAEADVASLNR.R
77 - 91	1616.6670	1615.5997	1615.8278	-141	2	K.KATDAEADVASLNRR.I
78 - 90	1332.4600	1331.4527	1331.6317	-134	0	K.ATDAEADVASLNR.R
78 - 91	1488.5360	1487.5287	1487.7328	-137	1	K.ATDAEADVASLNRR.I
91 - 101	1399.5710	1398.5637	1398.7467	-131	1	R.RIQLVEEELDR.A
91 - 105	1883.7500	1882.7427	1882.9860	-129	2	R.RIQLVEEELDRAQER.L
92 - 101	1243.4970	1242.4897	1242.6456	-125	0	R.IQLVEEELDR.A
92 - 105	1727.6700	1726.6627	1726.8848	-129	1	R.IQLVEEELDRAQER.L
106 - 125	2201.8610	2200.8537	2201.1175	-120	2	R.LATALQRLEEAERAADESER.G
113 - 128	1888.7230	1887.7157	1887.7921	-40	2	K.LEEAERAADESERGMK.V Oxidation (M
134 - 149	1980.6980	1979.6907	1979.9266	-125	2	K.AMAEEKMKELQKQLK.K
134 - 149	1996.6830	1995.6757	1995.9215	-129	2	K.AMAEEKMKELQKQLK.K Oxidation (M
137 - 161	3110.1290	3109.1217	3109.4596	-89	4	K.EEKMKELQKQLKKAKHLAEDSDKK.Y Pho
153 - 167	1817.8270	1816.8197	1816.8702	-138	2	K.HLAEDSDKKYEEVAK.K
168 - 178	1298.6620	1297.6547	1297.7805	-127	1	K.KLVILEGELEK.S
168 - 182	1759.7480	1758.7407	1758.9788	-132	2	K.KLVILEGELEKSKER.A
169 - 178	1170.5200	1169.5127	1169.6656	-122	0	K.LVILEGELEK.S
169 - 189	2385.9520	2384.9447	2385.2297	-123	2	K.LVILEGELEKSKERAEVAESK.C
206 - 226	2563.8880	2562.8807	2563.1562	-126	3	K.SLEAQADKYSTKEDKYEEEIK.L
206 - 231	3116.1560	3115.1487	3115.5448	-127	4	K.SLEAQADKYSTKEDKYEEEIKLLEEK.L
214 - 231	2014.0290	2013.0217	2013.0264	-2	2	K.YSTKEDKYEEEIKLLEEK.L +Phospho (S
218 - 231	1794.6600	1793.6527	1793.8934	-134	2	K.EDKYEEEIKLLEEK.L
218 - 231	1874.7160	1873.6987	1873.8597	-86	2	K.EDKYEEEIKLLEEK.L +Phospho (Y)
218 - 238	2622.1090	2621.1017	2621.2455	-122	4	K.EDKYEEEIKLLEEKLKKAETR.A
222 - 244	1538.5210	1537.5137	1537.7148	-131	0	K.YIDDLEDEVYAQR.M

PEPTIDE MASS FINGERPRINTING

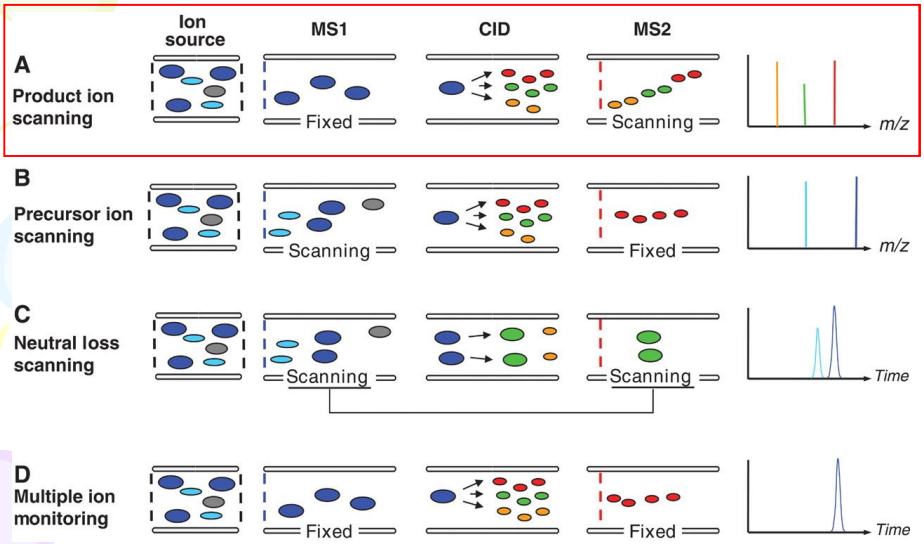
- 'quick and easy'
- requires
  - a very spesific enzyme
  - optimized digestion+ desalting protocols
  - internal/close external calibration of MALDI spectra
- works only for proteins which are already in the databases as protein sequences
- not suitable for complex protein mixtures

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

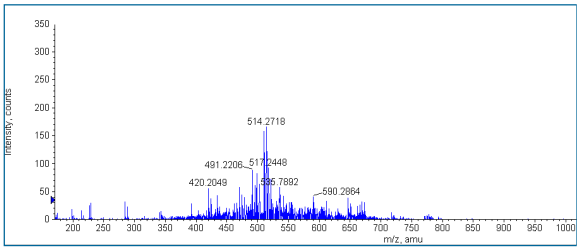
Tandem mass spectrometry scan types



Product ion scanning

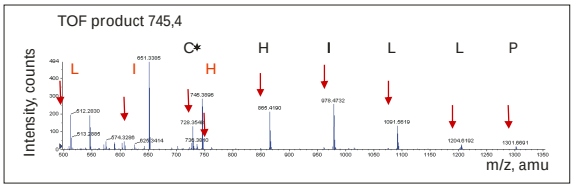
MS scan:

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

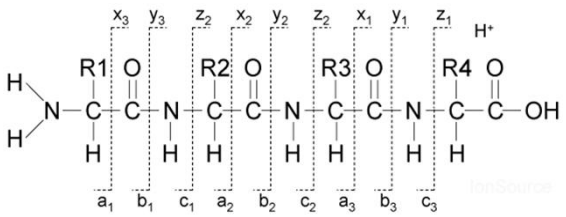


MS/MS scan:

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

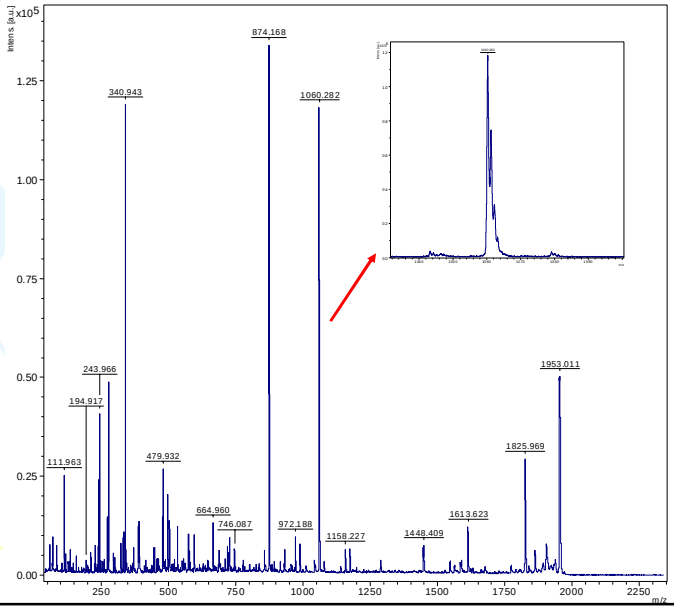
-first, PMF with MALDI-TOF

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

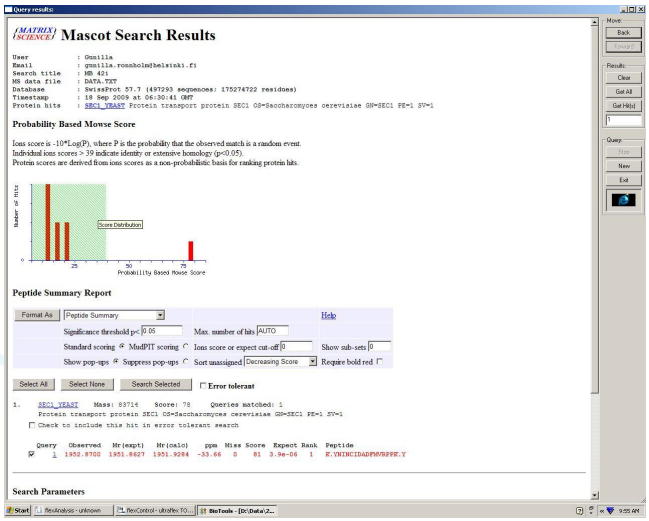
-MALDI produces singly charged parent ions
→ product ion spectra not as easy to interpret as in ESI-MS/MS

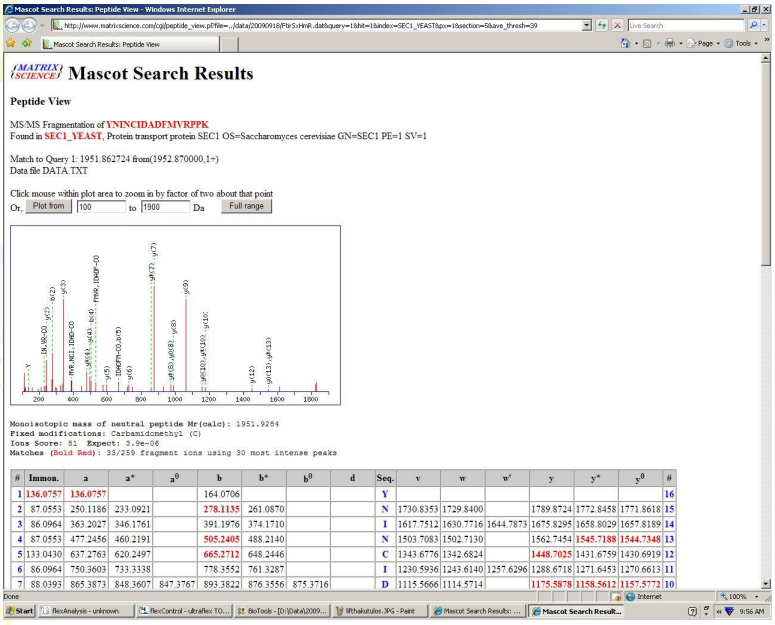
-database search with both PMF and MS/MS information

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.





Mascot Search Results: Peptide View - Windows Internet Explorer

http://www.matrixscience.com/cgi/peptide_view.pl?file=../data/20090918/F9r5cHnR.dat&query=1&hit=1&index=SEC1_YEAST&px=1&section=5&save_thresh=39

Mascot Search Results: Peptide View

Mascot Search Results

Peptide View

MS/MS Fragmentation of **YNINCIDADFMVRPPK**
Found in **SEC1_YEAST**, Protein transport protein SEC1 OS=Saccharomyces cerevisiae GN=SEC1 PE=1 SV=1

Match to Query 1: 1951.862724 from(1952.870000,1+)
Data file DATA.TXT

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 100 to 1900 Da Full range

Monoisotopic mass of neutral peptide Mr(calc): 1951.9284
Fixed modifications: Carbamidomethyl (C)
Ions Score: 91 Expect: 3.9e-06
Matches (Bold Red): 33/259 fragment ions using 30 most intense peaks

#	Immon.	a	a*	a^0	b	b*	b^0	d	Seq.	v	w	w'	y	y*	y^0	#
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

Done

Start | FlexAnalysis - unknown | FlexControl - ultraflex TO... | BioTools - [D:\Data\2009... | Whakututos.JPG - Paint | Mascot Search Results: ... | Mascot Search Result... | 9:56 AM

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

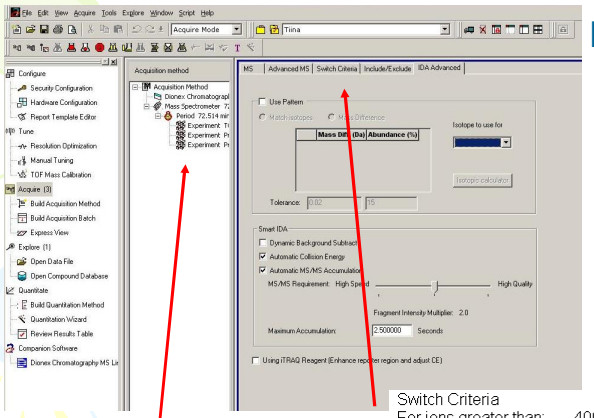
DDA =Data Dependent Acquisition
(IDA =Information Dependent Acquisition)

-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

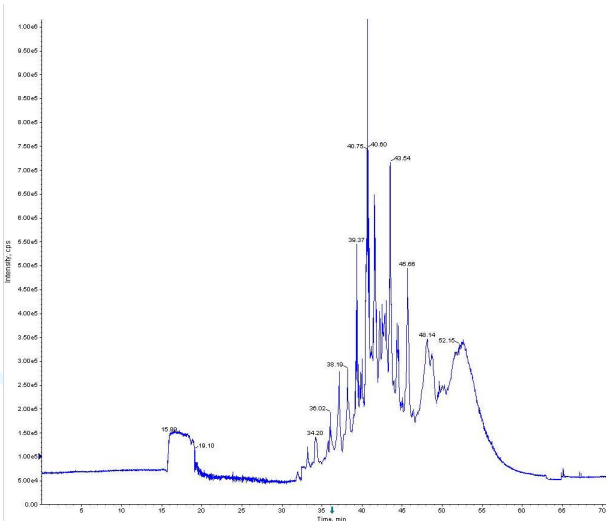


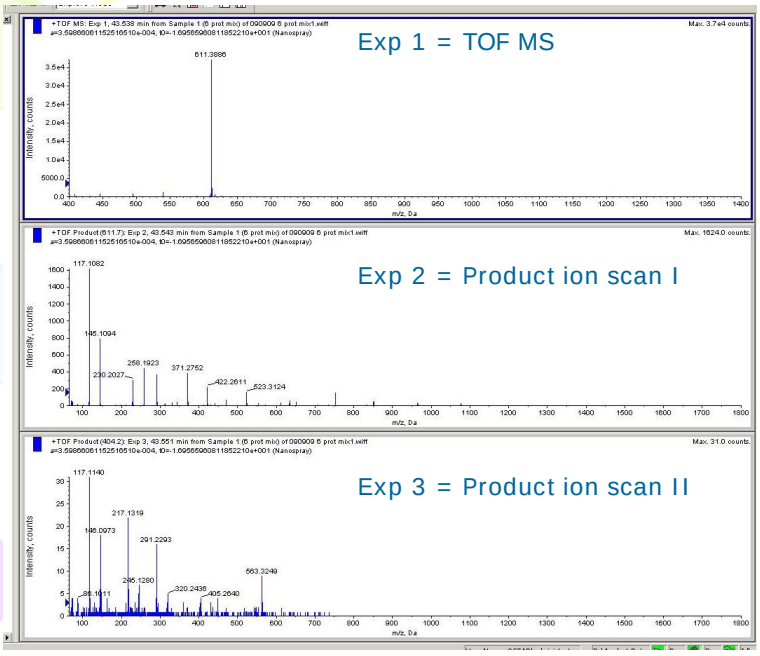
Switch Criteria

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



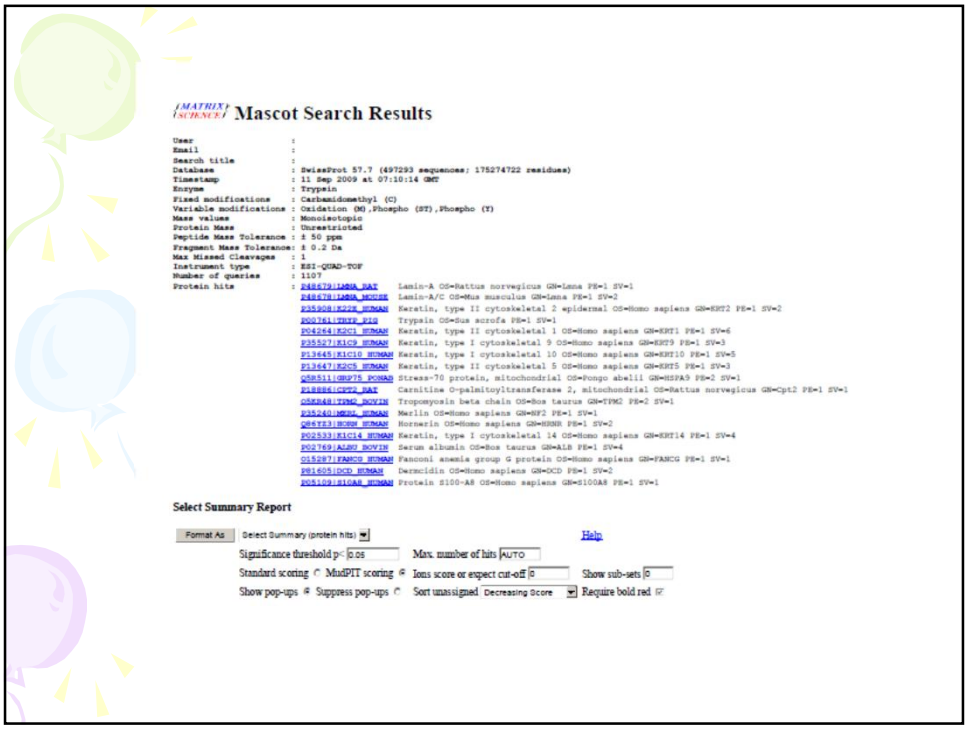


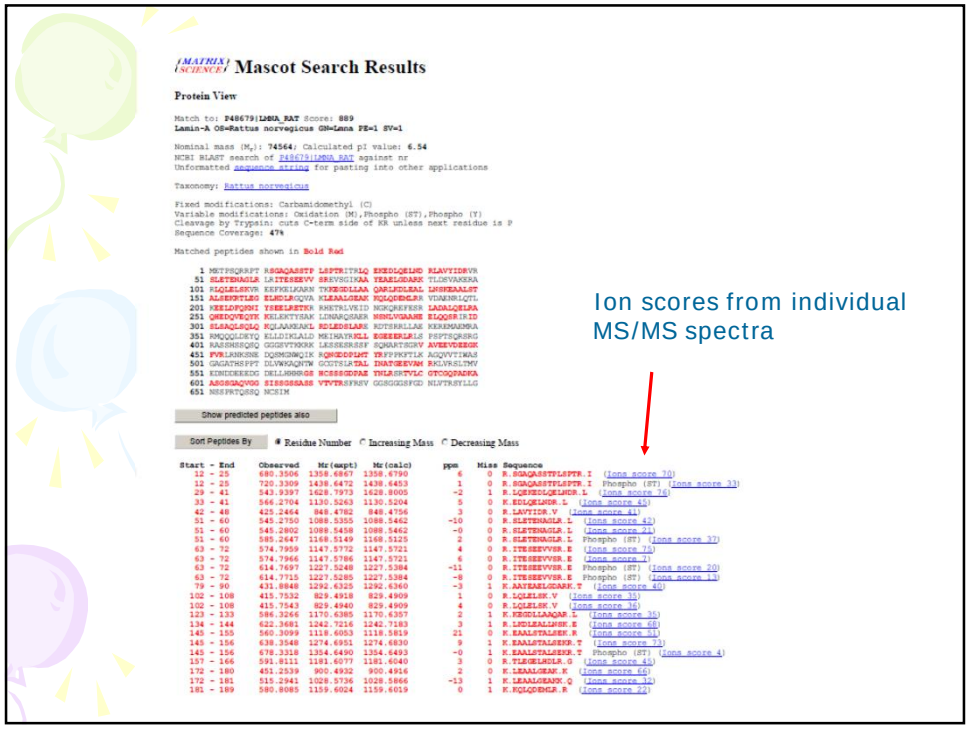
Search engines for (LC-)MS/MS data

-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



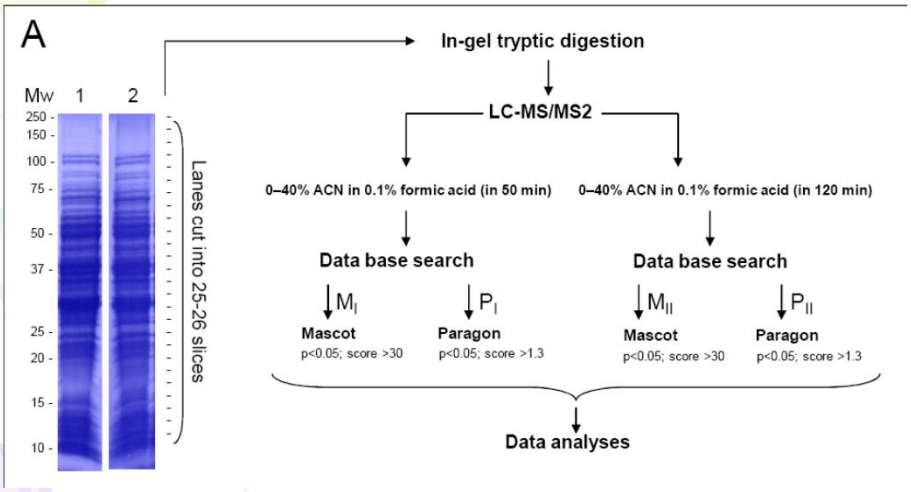


Protein identification from complex mixtures using nanoLC-MS/MS

- -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

Comparative proteome cataloging of Lactobacillus rhamnosus strains GG and Lc705

-in-depth proteome analysis of two Lactobacillus rhamnosus strains, the well-known probiotic strain GG and the dairy strain Lc705
-GeLC-MS/MS: proteins are separated using SDS_PAGE and identified using nanoLC-MS/MS
-to maximize the number of identifications, all data sets were searched against the target databases using two search engines, Mascot and Paragon



Savijoki et al, JPR 2011

Data analyses

Database Searches

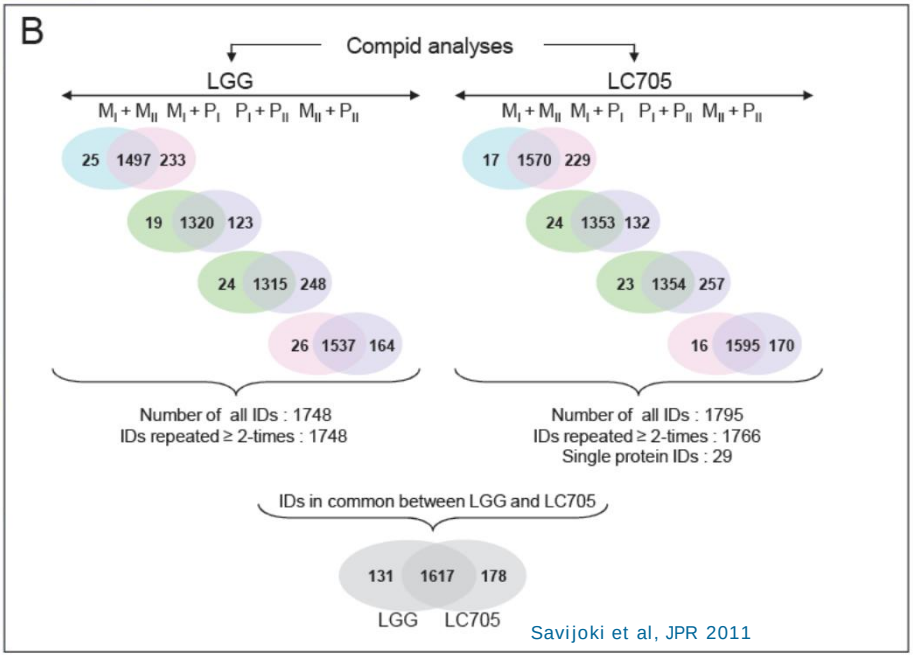
Protein Pilot (Mascot & Paragon)

Compid analyses

Compilation of Mascot and Paragon results

Bioinformatics

Cellular location of the identified proteins



Savijoki et al, JPR 2011

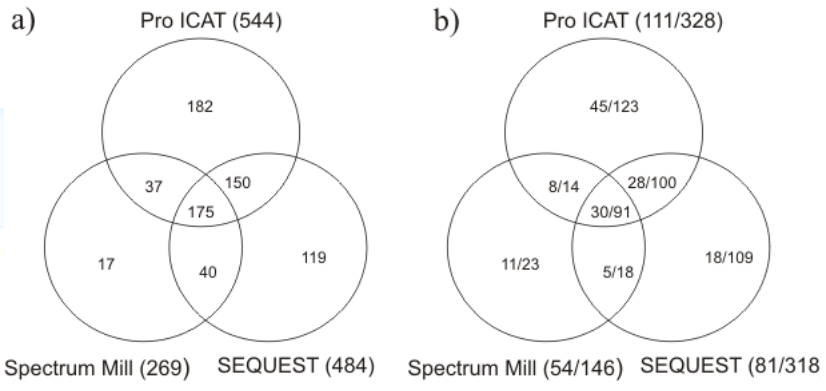
Natri et al, JPR 2010

Comparison of protein ID results from the same raw data with different search engines

- ProICAT SP2
- Spectrum Mill
- Sequest

Moulder *et al*, Proteomics. 2005;5:2748-60.

Comparison of the ID results with different search engines



a)Peptide identifications, b) Proteins identifications with two or more peptides given as a proportion of all protein identifications.

False discovery rate (FDR) estimation

$$FDR = FP/(TP+FP)$$

TP= true positive matches
FP =false positive matches

Previously: search the raw data first against the 'normal' database and then against the decoy database
-> calculate FDR based on these

NOW the preferred method:
One search against a database in which the target and decoy sequences have been concatenated. This means that you will only record a false positive when a match from the decoy sequences is better than any match from the target sequences.

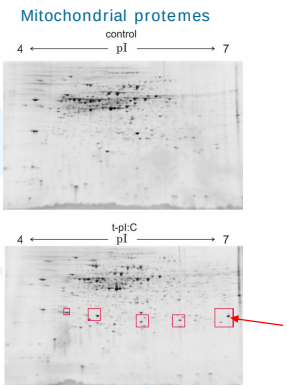
Elias, J. E., *et al*, Nature Methods 2 667-675 (2005).

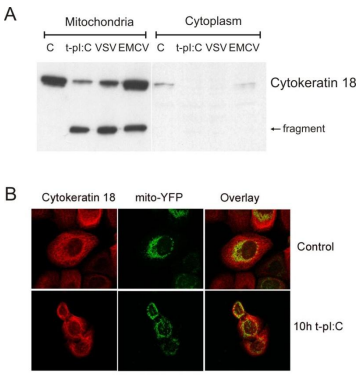
Now you're protein is identified
What else can we find in the data?

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 Page 1 of 7

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

Mascot Search Results

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

Mascot Search Results

Protein View

Match to: KIN2_YEAST Score: 791
Serine/threonine-protein kinase KIN2 - Saccharomyces cerevisiae (Baker's yeast)

Nominal mass (M_r): 128601; Calculated pI value: 9.43
NCBI BLAST search of KIN2_YEAST against nr
Unformatted sequence string for pasting into other applications

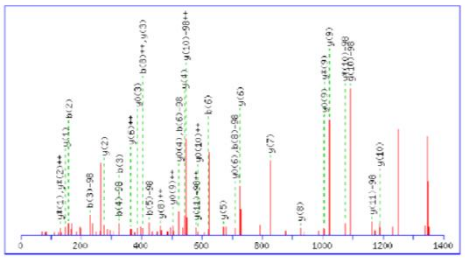
Taxonomy: Saccharomyces cerevisiae

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: 33%

Matched peptides shown in Bold Red

```
    1 MPNPNTADYL VNPNFRTSKG GSLSPTPEAF NDTRVAAPAT LRMMGMQSGP
   51 RNDQQQAPLM PPADIKQGKE QAAQRQNDAS RFNGAVELRQ FHRRSLGDWE
  101 FLETVGAGSM GKVKLVKHRQ TKEICVIKIV NRASKAYLMK QHSLPSPKNE
  151 SEILERQKRL EKEIARDKRT VREASLGQIL YMPHICRLFE MCTMSNHFYM
  201 LFEYVSGGQL LDYIIQHGSL KEHHARKFAR GIASALQYLH ANNIVHRDLK
  251 IENIMISSSG EIKIIDFGLS NIFDYRKQLK TFCGSLVFAA PELLKAQPYT
  301 GPEVDIWSFG IVLYVLVCGK VPFDDENSSI LHEKIKKGKV DYPSHLSIEV
  351 ISLLTRMIVV DPLRRATLKN VVEHPWMNRG YDFKAPSYVP NRVPLTPEMI
  401 DSQVLKEMYR LEFIDDIEDT RRSLIRLVTE KEYIQLSQEY WDKLSNAKGL
  451 SSSLNNNYLN STAQQTLIQM HITSNFSQSG YNEPDSNFED PTLAYHPLLS
  501 IYHLVSEMVA RKLAKLQRRQ ALALQAQAQQ RQQQQQVALG TKVALNNNSP
  551 DINTKMRSPQ KEVVPNPGIF QVPAIGTSGT SNNTNTSNKP PLHVMVPPKL
  601 TIPEQAMTSP TSRKSSDINT ELNGVLKSTP VPVSGEYQQR SASFVVGEHQ
  651 EKNTIGGIFR RIPQSGQSQH PTRQQEPLPE REPPTYMSKS NEISIKVPKS
  701 RSRTISDVIP SARKYPSYVP NSVDVKQKPA KNTSIAPPIR SVSQKQNSDL
  751 PALPQNAKLI VQNQRQKLLQ ENLDKLQIND NDNNNVNAVV DGINNDNSDH
  801 YLSVPKGRKL HPSARAKSVG HARRESLKFT RPPIPAALPF SIMTNDNGFL
  851 GEANKERYNP VSSNFSTVPE DSTTYSNDTN NRLTSVYSQE LTEKQILEEA
  901 SKAPPGSMPS IDYPKSMFLK GFFSVQTTSS KPLPIVRHNI ISVLTRMNID
  951 FKEVKGGFIC VQQRPSIETA AVPVITTTGV GLDSGKAMDL QNSLDSQLSS
 1001 SYHSTASSAS RNSSIKRQGS YKRGQNNIPL TPLATNTHQR NSSIPMSPNY
 1051 GMQSNGTSGE LSSMSLDYVQ QQDDILTTSR AQNINNVNGQ TEQTNTSGIK
 1101 ERPPIKFEIH IVKVRIVGLA GVEFKKVSGN TWLYKELASY ILKELNL
```

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.APSVVPNR.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.RYPSVVPNSVDVK.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