Peptide Mass Fingerprinting

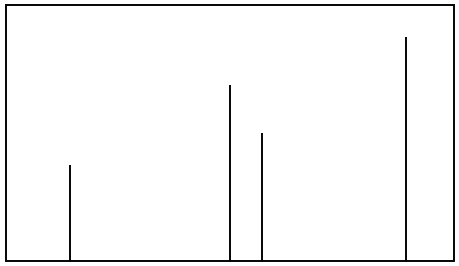
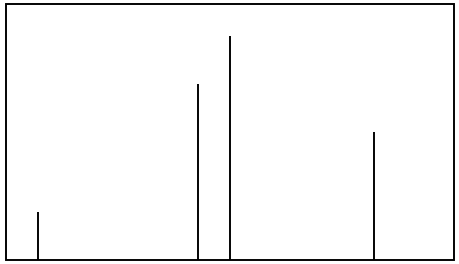
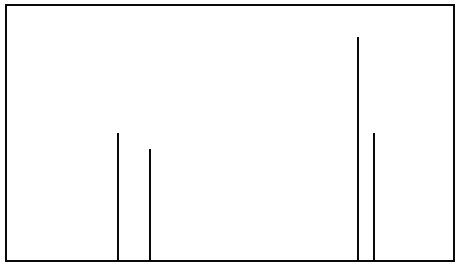
- Used to identify protein spots on gels or protein peaks from an HPLC run
- Depends of the fact that if a peptide is cut up or fragmented in a known way, the resulting fragments (and resulting masses) are unique enough to identify the protein*
- Requires a database of known sequences
- Uses software to compare observed masses with masses calculated from database

*Considering all aa sequence combinations that are theoretically possible, only a very minor portion of protein sequences is realized in nature, and therefore a short peptide sequence is already highly protein-specific,

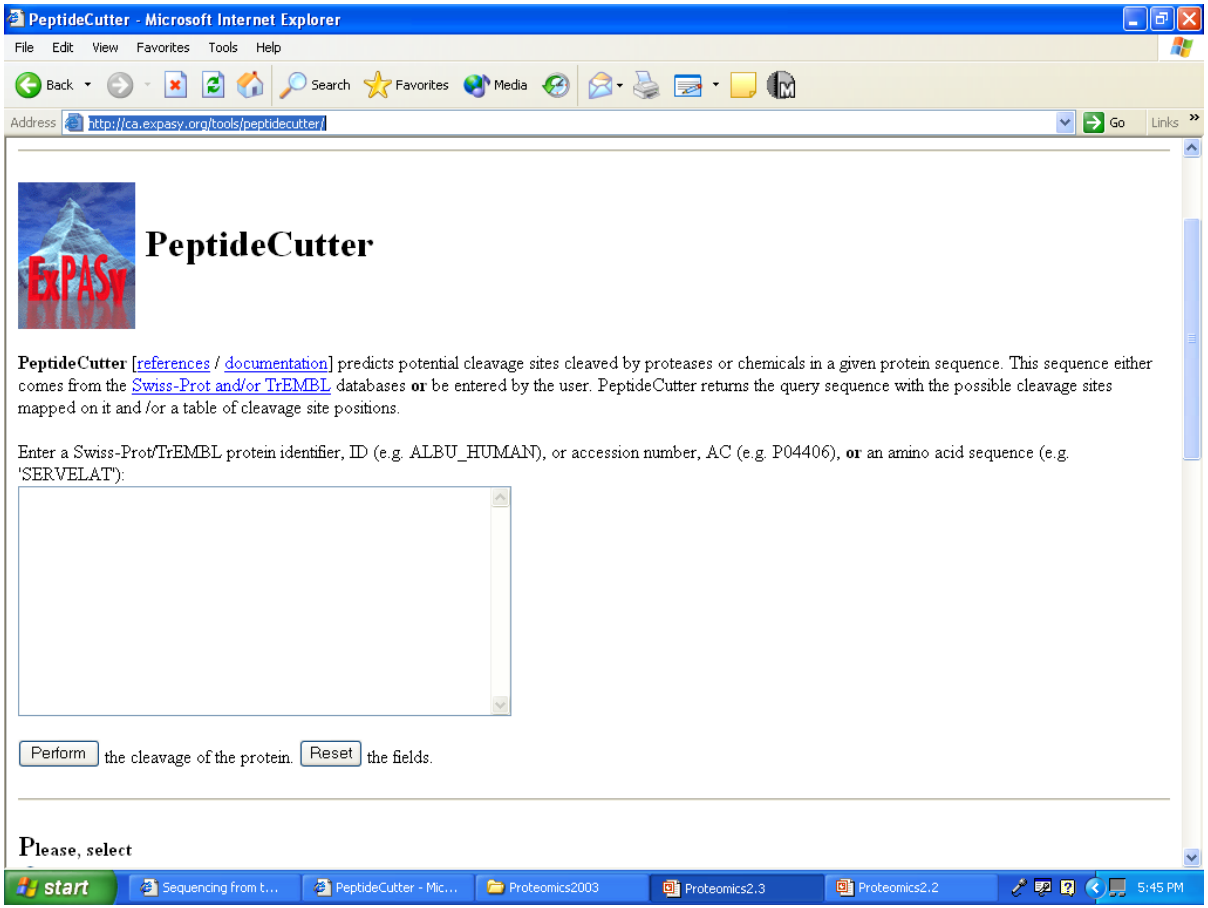
Principles of Fingerprinting

Sequence	Mass (M+H)	Tryptic Fragments
>Protein 1 acedfhsakdfqea sdfpkivtmeeewe ndadnfekqwfe	4842.05	acedfhsak dfgeasdfpk ivtmeeewendadnfek gwfe
>Protein 2 acekdfhsadfqea sdfpkivtmeeewe nkdadnfeqwfe	4842.05	acek dfhsadfgeasdfpk ivtmeeewenk dadnfeqwfe
>Protein 3 acedfhsadfqeka sdfpkivtmeeewe ndakdnfeqwfe	4842.05	acedfhsadfgek asdfpk ivtmeeewendak dnfegwfe

Principles of Fingerprinting

Sequence	Mass (M+H)	Mass Spectrum
>Protein 1 acedfhsakdfqea sdfpkivtmeeewe ndadnfekqwfe	4842.05	
>Protein 2 acekdfhsadfqea sdfpkivtmeeewe nkdadnfeqwfe	4842.05	
>Protein 3 acedfhsadfqeka sdfpkivtmeeewe ndakdnfeqwfe	4842.05	

Predicting Peptide Cleavages



http://ca.expasy.org/tools/peptidecutter/

http://ca.expasy.org/tools/peptidecutter/peptidecutter_enzymes.html#Tryps

PeptideCutter

The cleavage specificities of selected enzymes and chemicals:

A general model of enzymatic cleavage:

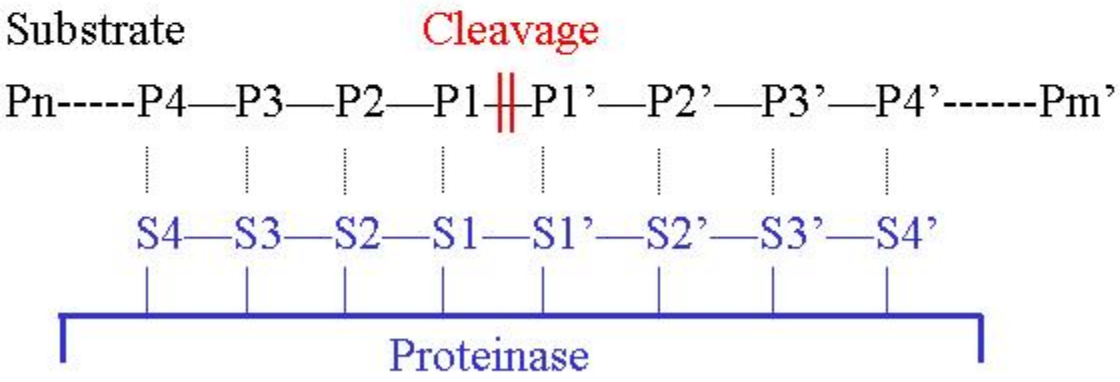


Fig.1 Schematic representation of enzyme-substrate complex with eight binding sites. Positions Pn to Pm' in the substrate are counted from the bond between P1 and P1', where the cleavage occurs.

Protease Cleavage Rules

Sometimes inhibition occurs

Trypsin	**XXX[KR]--[!P]XXX**
Chymotrypsin	**XX[FYW]--[!P]XXX**
Lys C	**XXXXXK-- XXXXX**
Asp N endo	**XXXXXD-- XXXXX**
CNBr	**XXXXXM--XXXXX**

K-Lysine, R-Arginine, F-Phenylalanine, Y-Tyrosine,
W-Tryptophan,D-Aspartic Acid, M-Methionine, P-Proline

Digest with specific protease

546 aa **60 kDa; 57 461 Da** **pI = 4.75**

>RBME00320 Contig0311_1089618_1091255 EC-mopA 60 KDa chaperonin GroEL
MAAKDVKFGR TAREKMLRGV DILADAVKVT LGPKGRNVVI EKSFGAPRIT KDGVSVAKEV
ELEDKFENMG AQMLREVASK TNDTAGDGTT TATVLGQAIV QEGAKAVAAG MNPMDLKRGI
DLAVNEVVAE LLKKAKKINT SEEVAQVGTI SANGEAEIGK MIAEAMQKVG NEGVITVEEA
KTAETELEVV EGMQFDRGYL SPYFVTNPEK MVADLEDAYI LLHEKKLSNL QALLPVLEAV
VQTSKPLLII AEDVEGEALA TLVVNKLRGG LKIAAVKAPG FGDCRKAMLE DIAILTGGQV
ISEDLGIKLE SVTLDMLGRA KKVSISKENT TIVDGAGQKA EIDARVGQIK QQIEETTSDY
DREKLQERLA KLAGGVAVIR VGGATEVEVK EKKDRVDDAL NATRAAVEEG IVAGGGTALL
RASTKITAKG VNADQEAGIN IVRRAIQAPA RQITTNAGEE ASVIVGKILE NTSETFGYNT
ANGEYGDLIS LGIVDPVKVV RTALQNAASV AGLLITTEAM IAELPKKDAA PAGMPGGMGG
MGGMDF

Digest with specific protease

Trypsin yields 47 peptides (theoretically)

Peptide masses in Da:

501.3	533.3	544.3	545.3	614.4	634.3
674.3	675.4	701.4	726.4	822.4	855.5
861.4	879.4	921.5	953.4	974.5	988.5
1000.6	1196.6	1217.6	1228.5	1232.6	1233.7
1249.6	1249.6	1344.7	1455.8	1484.6	1514.8
1582.9	1583.9	1616.8	1726.7	1759.9	1775.9
1790.6	1853.9	1869.9	2286.2	2302.2	2317.2
2419.2	2526.4	2542.4	3329.6	4211.4	

http://us.expasy.org/tools/peptide-mass.html

Digest with trypsin

In practice.....see far fewer by mass spec

- **possibly incomplete digest (we allow 1 miss)**
- **lose peptides during each manipulation**

  washes during digestion
  washes during cleanup step
  some peptides will not ionize well
  some signals (peaks) are poor
  low intensity; lack resolution

What Are Missed Cleavages?

Sequence

>Protein 1
acedfhsakdfqea
sdfpkivtmeeewe
ndadnfekqwfe

Tryptic Fragments (no missed cleavage)

acedfhsak (1007.4251)

dfgeasdfpk (1183.5266)

ivtmeeewendadnfek (2098.8909)
gwfe (609.2667)

Tryptic Fragments (1 missed cleavage)

acedfhsak (1007.4251)

dfgeasdfpk (1183.5266)

ivtmeeewendadnfek 2098.8909)
gwfe (609.2667)
acedfhsakdfgeasdfpk (2171.9338)
ivtmeeewendadnfekgwfe (2689.1398)
dfgeasdfpkivtmeeewendadnfek (3263.2997)

Calculating Peptide Masses

- **Sum the monoisotopic residue masses**

Monoisotopic Mass: the sum of the exact or accurate masses of the lightest stable isotope of the atoms in a molecule

- **Add mass of H_2O (18.01056)**
- **Add mass of H^+ (1.00785 to get M+H)**
- **If Met is oxidized add 15.99491**
- **If Cys has acrylamide adduct add 71.0371**
- **If Cys is iodoacetylated add 58.0071**

^{1}H-1.007828503 amu **^{12}C-12**

^{2}H-2.014017780 amu **^{13}C-13.00335, ^{14}C-14.00324**

Masses in MS

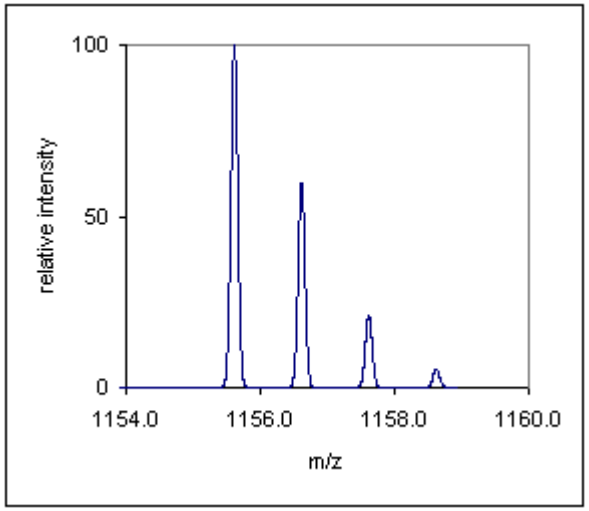
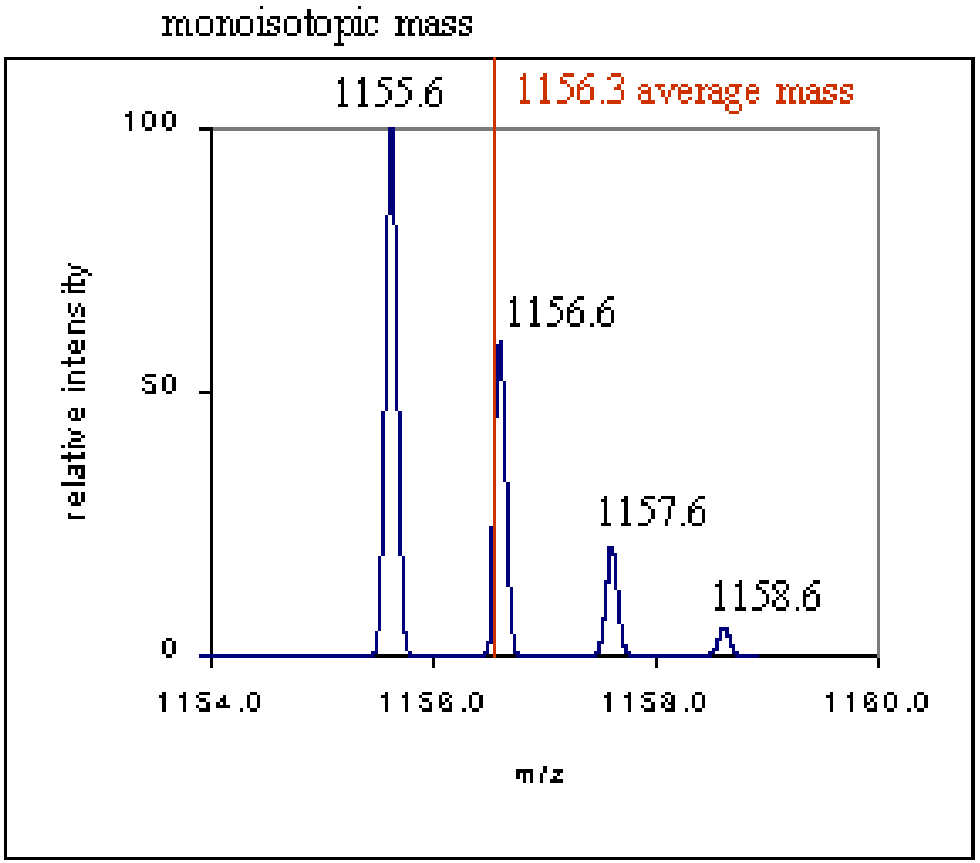


Figure shows a simulated isotopic distribution of the [M+H]+ ion of a compound with the following elemental composition, C_{48} H_{82} N_{16} O_{17} (Poly-Alanine)

Masses in MS



- **Monoisotopic mass is the mass determined using the masses of the most abundant isotopes**
- **Average mass is the abundance weighted mass of all isotopic components**

Mass Calculation (Glycine)

$NH_2—CH_2—COOH$ Amino acid

$R_1—NH—CH_2—CO—R_3$ Residue

Monoisotopic Mass

^{1}H = 1.007825

^{12}C = 12.00000

^{14}N = 14.00307

^{16}O = 15.99491

Glycine Amino Acid Mass

5xH + 2xC + 2xO + 1xN

= 75.032015 amu

Glycine Residue Mass

3xH + 2xC + 1xO + 1xN

=57.021455 amu

Amino Acid Residue Masses

Monoisotopic Mass

Glycine	**57.02147**	**Aspartic acid**	**115.02695**
Alanine	**71.03712**	**Glutamine**	**128.05858**
Serine	**87.03203**	**Lysine**	**128.09497**
Proline	**97.05277**	**Glutamic acid**	**129.0426**
Valine	**99.06842**	**Methionine**	**131.04049**
Threonine	**101.04768**	**Histidine**	**137.05891**
Cysteine	**103.00919**	**Phenylalanine**	**147.06842**
Isoleucine	**113.08407**	**Arginine**	**156.10112**
Leucine	**113.08407**	**Tyrosine**	**163.06333**
Asparagine	**114.04293**	**Tryptophan**	**186.07932**

Preparing a Peptide Mass Fingerprint Database

- Take a protein sequence database (Swiss-Prot or nr-GenBank)
- Determine cleavage sites and identify resulting peptides for each protein entry
- Calculate the mass (M+H) for each peptide
- Sort the masses from lowest to highest
- Have a pointer for each calculated mass to each protein accession number in databank

Building A PMF Database

Sequence DB

>P12345
acedfhsakdfqea
sdfpkivtmeeewe
ndadnfekqwfe

>P21234
acekdfhsadfqea
sdfpkivtmeeewe
nkdadnfeqwfe

>P89212
acedfhsadfqeka
sdfpkivtmeeewe
ndakdnfeqwfe

Calc. Tryptic Frags

acedfhsak
dfgeasdfpk
ivtmeeewendadnfek
gwfe

acek
dfhsadfgeasdfpk
ivtmeeewenk
dadnfeqwfe

acedfhsadfgek
asdfpk
ivtmeeewendak
dnfegwfe

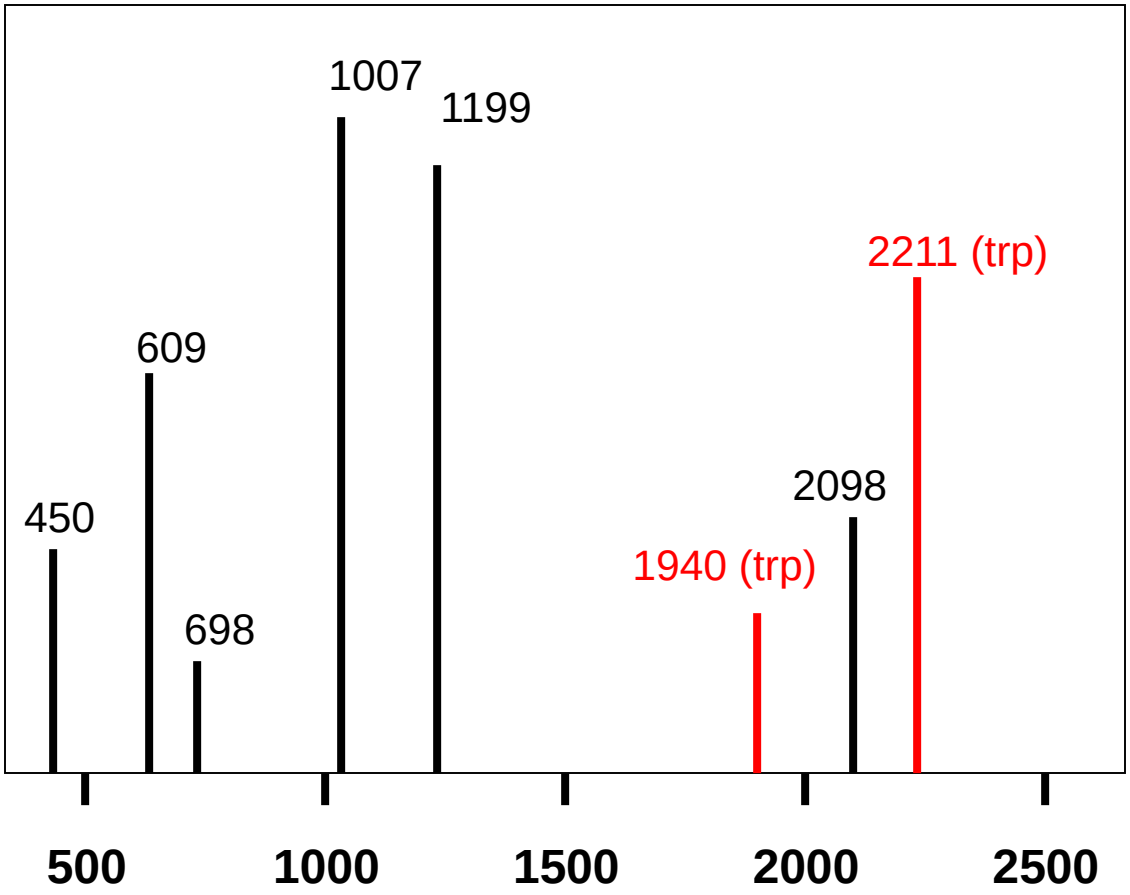
Mass List

450.2017 (P21234)
609.2667 (P12345)
664.3300 (P89212)
1007.4251 (P12345)
1114.4416 (P89212)
1183.5266 (P12345)
1300.5116 (P21234)
1407.6462 (P21234)
1526.6211 (P89212)
1593.7101 (P89212)
1740.7501 (P21234)
2098.8909 (P12345)

The Fingerprint (PMF) Algorithm

- **Take a mass spectrum of a trypsin-cleaved protein (from gel or HPLC peak)**
- **Identify as many masses as possible in spectrum (avoid autolysis peaks of trypsin)**
- **Compare query masses with database masses and calculate # of matches or matching score (based on length and mass difference)**
- **Rank hits and return top scoring entry – this is the protein of interest**

Query (MALDI) Spectrum



Query vs. Database

Query Masses

450.2201
609.3667
698.3100
1007.5391
1199.4916
2098.9909

Database Mass List

450.2017 (P21234)
609.2667 (P12345)
664.3300 (P89212)
1007.4251 (P12345)
1114.4416 (P89212)
1183.5266 (P12345)
1300.5116 (P21234)
1407.6462 (P21234)
1526.6211 (P89212)
1593.7101 (P89212)
1740.7501 (P21234)
2098.8909 (P12345)

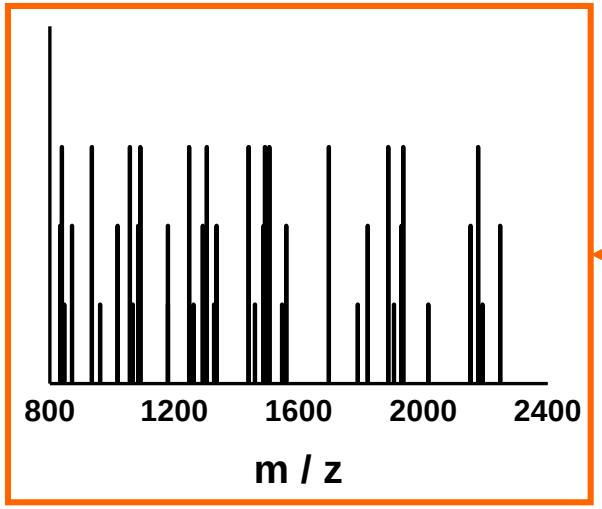
Results

2 Unknown masses
1 hit on P21234
3 hits on P12345

Conclude the query protein is P12345

Database search

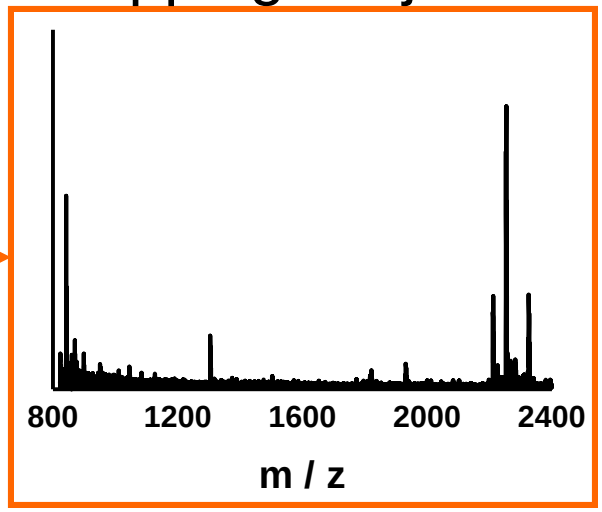
PeptIdent (ExPasy)
→ Mascot (Matrix Science)
MS-Fit (Prospector; UCSF)
ProFound (Proteometrics)
MOWSE (HGMP)
Human Genome Mapping Project



theoretical

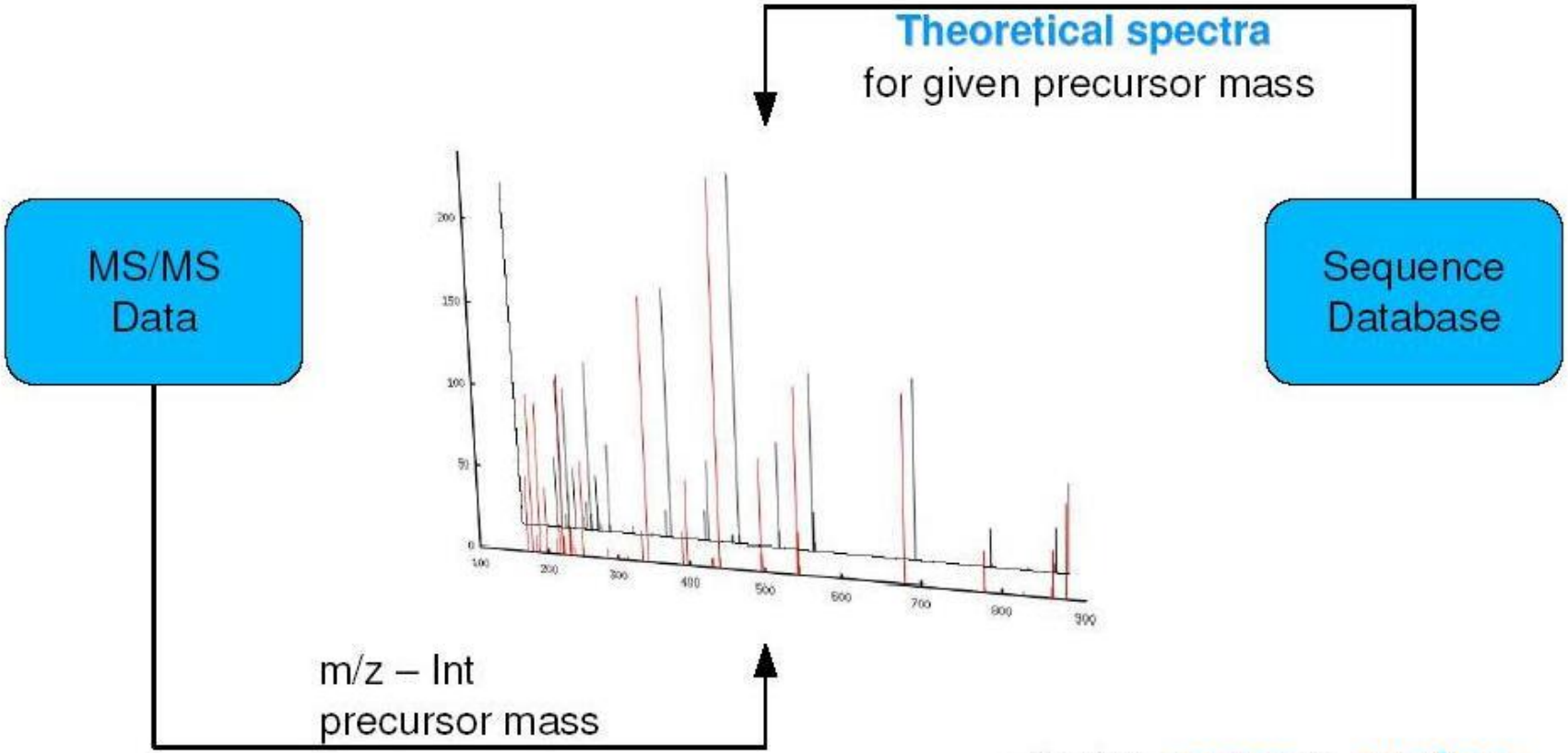
Mascot

Protein ID



experimental

Uninterpreted MS/MS Database Search



- Assign **scores** to **overlaps**
- (Normalize scores)
- Keep best match

What You Need To Do PMF

- **A list of query masses (as many as possible)**
- **Protease(s) used or cleavage reagents**
- **Databases to search (SWProt, Organism)**
- **Estimated mass and pI of protein spot**
- **Cysteine (or other) modifications**
- **Minimum number of hits for significance**
- **Mass tolerance (100 ppm = 1000.0 ± 0.1 Da)**
- ***A PMF website (ProFound, Mascot, etc.)***

ProFound

ProFound - Peptide Mapping [Short Form]

Version 4.10.5
The Rockefeller University Edition

General

Sample ID

Database: NCBInr (2002/11/27)

Taxonomic Category: All taxa

Search for: single protein only

Protein Mass: 0 - 3000 kDa

Protein pI: 0 - 14

Report Top: 10 Candidates

Questions? Please write to ProFound

What's new about ProFound?

Digestion

Allow maximum: 1 missed cleavages

Enzyme: Trypsin

For user-defined cleavage, please click here.

Modifications

Complete Modification(s):
- Unmodified
- 4-vinyl-pyridine (Cys)
- Acrylamide (Cys)
- Iodoacetamide (Cys)
- Iodoacetic acid (Cys)

Partial Modification: ☐ Methionine oxidation

For more partial modifications, please click here.

Masses

Average Masses:

Monoisotopic Masses:

Mass tolerance for average data: +/- 1

Mass tolerance for monoisotopic data: +/- 0.1

Tolerance unit: ◉ Da ○ % ○ ppm

Charge state: ◉ M ○ MH+

Identify Protein | Extra Settings | Example | Reset Form

ProFound Results

ProFound - Search Result Summary

Version 4.10.5
The Rockefeller University Edition

Protein Candidates for search B9403AFB-07C0-76BF87E5 [1209637 sequences searched]

Rank	Probability	Est'd Z	Protein Information and Sequence Analyse Tools (T)	%	pI	kDa	®
1	2.2e-001	0.12	T gi|15222204|ref|NP_172776.1| putative oxysterol-binding protein; protein id: At1g13170.1 [Arabidopsis thaliana]	8	6.1	92.31	®
2	2.2e-001	0.12	T gi|17547403|ref|NP_520805.1| PROBABLE OXIDOREDUCTASE PYRROLINE-5-CARBOXYLATE REDUCTASE SIGNAL PEPTIDE PROTEIN [Ralstonia solanacearum]	11	5.8	28.10	®
3	7.6e-002	-	T gi|23054472|gb|ZP_00080629.1| hypothetical protein [Geobacter metallireducens]	11	6.1	51.76	®
4	7.6e-002	-	T gi|19920902|ref|NP_609168.1| CG7228-PA [Drosophila melanogaster]	7	8.6	66.18	®
5	2.6e-002	-	T gi|19572314|emb|CAD19081.1| potassium channel beta chain [Stigmatella aurantiaca]	10	9.6	41.10	®
+6	2.5e-002	-	T gi|2133779|pir||S63985 collagen alpha 2 chain precursor - sea urchin (Strongylocentrotus purpuratus) (fragment)	3	4.4	200.03	®
7	2.3e-002	-	T gi|15450423|gb|AAK96505.1| AT4g20760/F21C20_110 [Arabidopsis thaliana]	13	9.8	32.46	®
+8	2.0e-002	-	T gi|7495844|pir||T25534 hypothetical protein C10H11.6 - Caenorhabditis elegans	8	6.7	58.38	®
9	1.9e-002	-	T gi|21293583|gb|EAA05728.1| agCP10259 [Anopheles gambiae str. PEST]	4	6.3	66.10	®
10	1.6e-002	-	T gi|16121031|ref|NP_404344.1| sigma-54 transcriptional regulatory protein [Yersinia pestis]	10	6.1	37.74	®

Advantages of PMF

- Uses a “robust” & inexpensive form of MS (MALDI)
- Doesn’t require too much sample optimization
- Can be done by a moderately skilled operator (don’t need to be an MS expert)
- Widely supported by web servers
- Improves as DB’s get larger & instrumentation gets better

Limitations With PMF

- Requires that the protein of interest already be in a sequence database
- Spurious or missing critical mass peaks always lead to problems
- Mass resolution/accuracy is critical
- Generally found to only be about 40% effective in positively identifying gel spots