Protein Identification using Mass Spectrometry

PC219

Mike Baldwin

mbaldwin@cgl.ucsf.edu

Mass Spectrometers measure mass.

Why can't we identify a protein from its molecular mass?

- It is challenging to measure the molecular mass of a large species such as a protein with high accuracy and needs sophisticated and expensive equipment.
- Too many proteins have the same mass.
- Due to posttranslational modifications or chemical changes the measured mass might be "wrong", i.e. not conform to the gene sequence.
- Modifications are often heterogeneous or not present on all molecules, resulting in multiple molecular masses for a single protein.

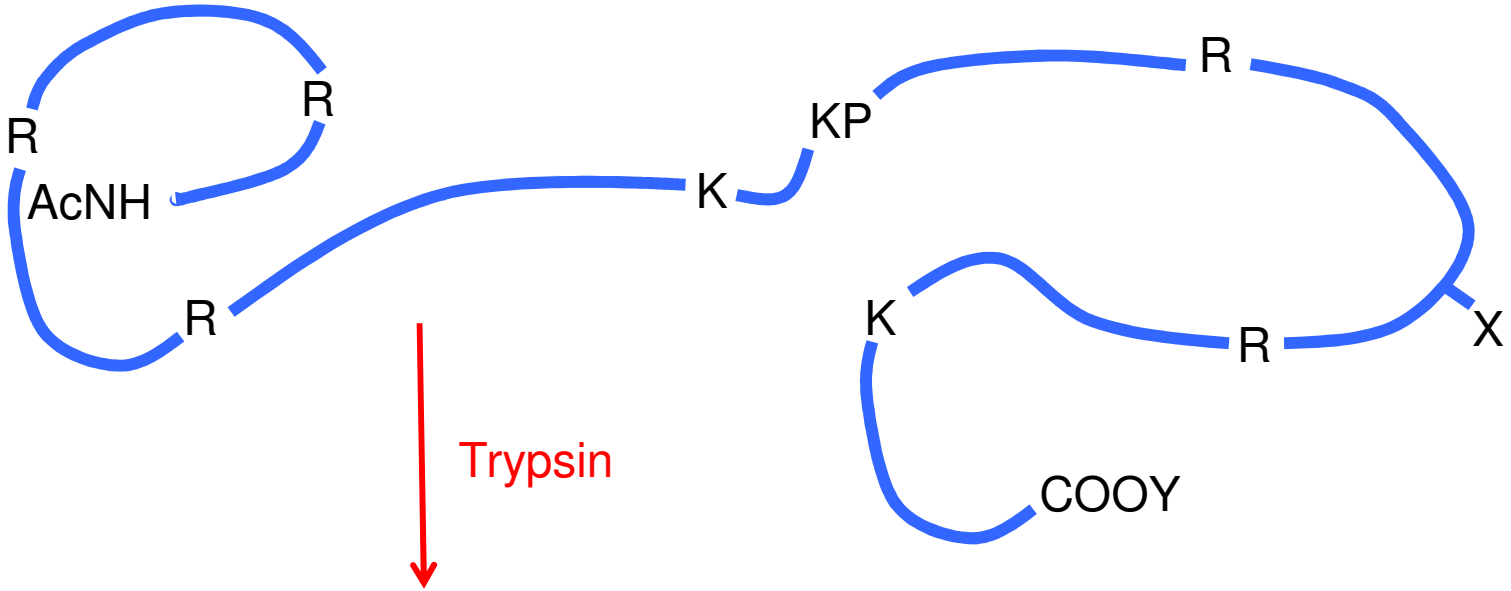
What can we do?

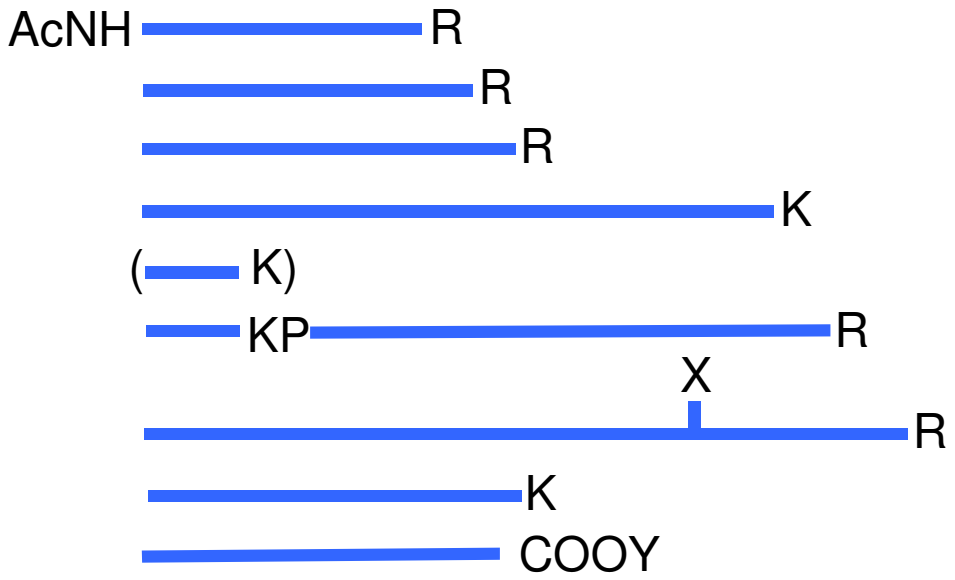
- Digest the protein with a specific protease (most often trypsin) for Peptide Mass Fingerprinting.
- Use proteolysis, CID and peptide fragmentation to identify sections of protein sequence. (Bottom up).
- Fragmentation analysis of the intact protein ‘Protein Sequencing’ (Top down). Not normally used for protein ID’s, more for PTM analysis.
- All these approaches rely on database searching and scoring

Protocol for Peptide Mass Fingerprinting (PMF)

- Reduce, alkylate and digest protein.
- Acquire mass spectrum of peptide mixture, usually by MALDI (or LC-MS).
- Process the raw data and input the list of observed masses into a database search program.
- Use a search program that creates a theoretical enzyme digest of all proteins in database, and compares the mass list observed to theoretical mass lists for all proteins, and returns ‘best matches’.
- Assess scores for the best hit.

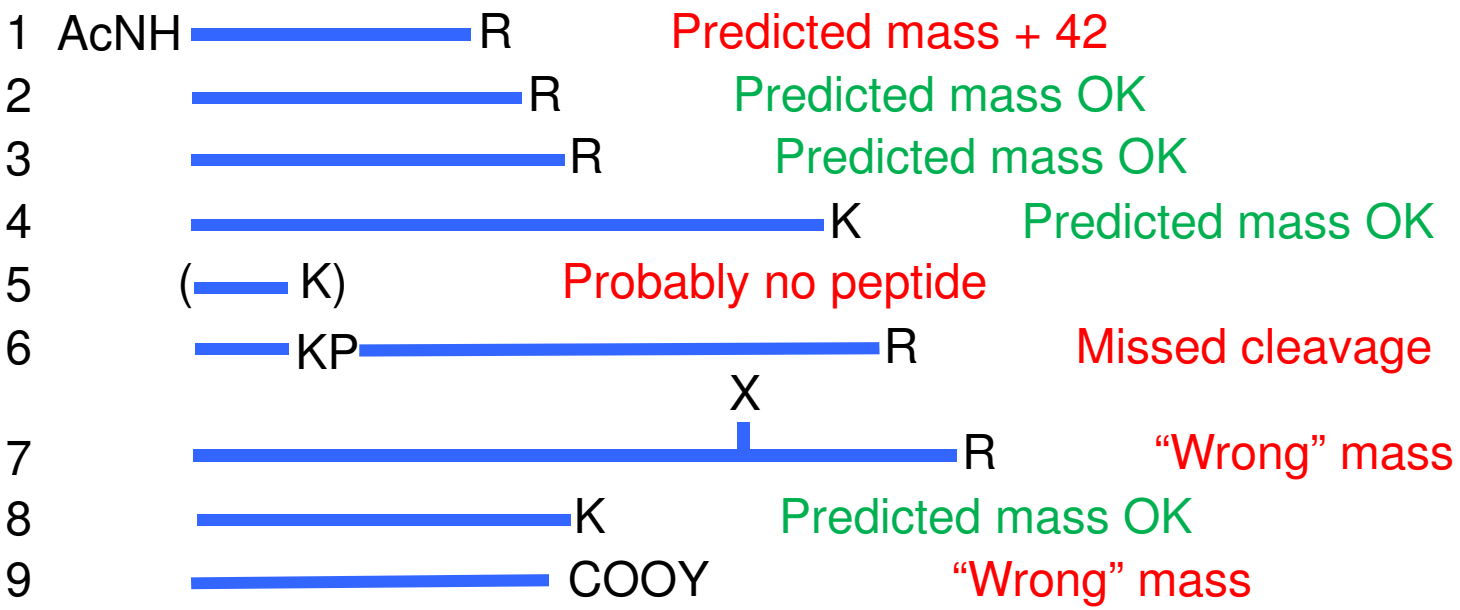
PMF has Limitations





Digestion of this protein with trypsin may yield 9 peptides.

How many of these can we use to identify the protein by mass measurements alone?



Conclusion: Maybe 4 of 9 predicted peptides give the "correct" mass

Improved strategy: Allow for missed cleavages; allow for N-terminal Ac.

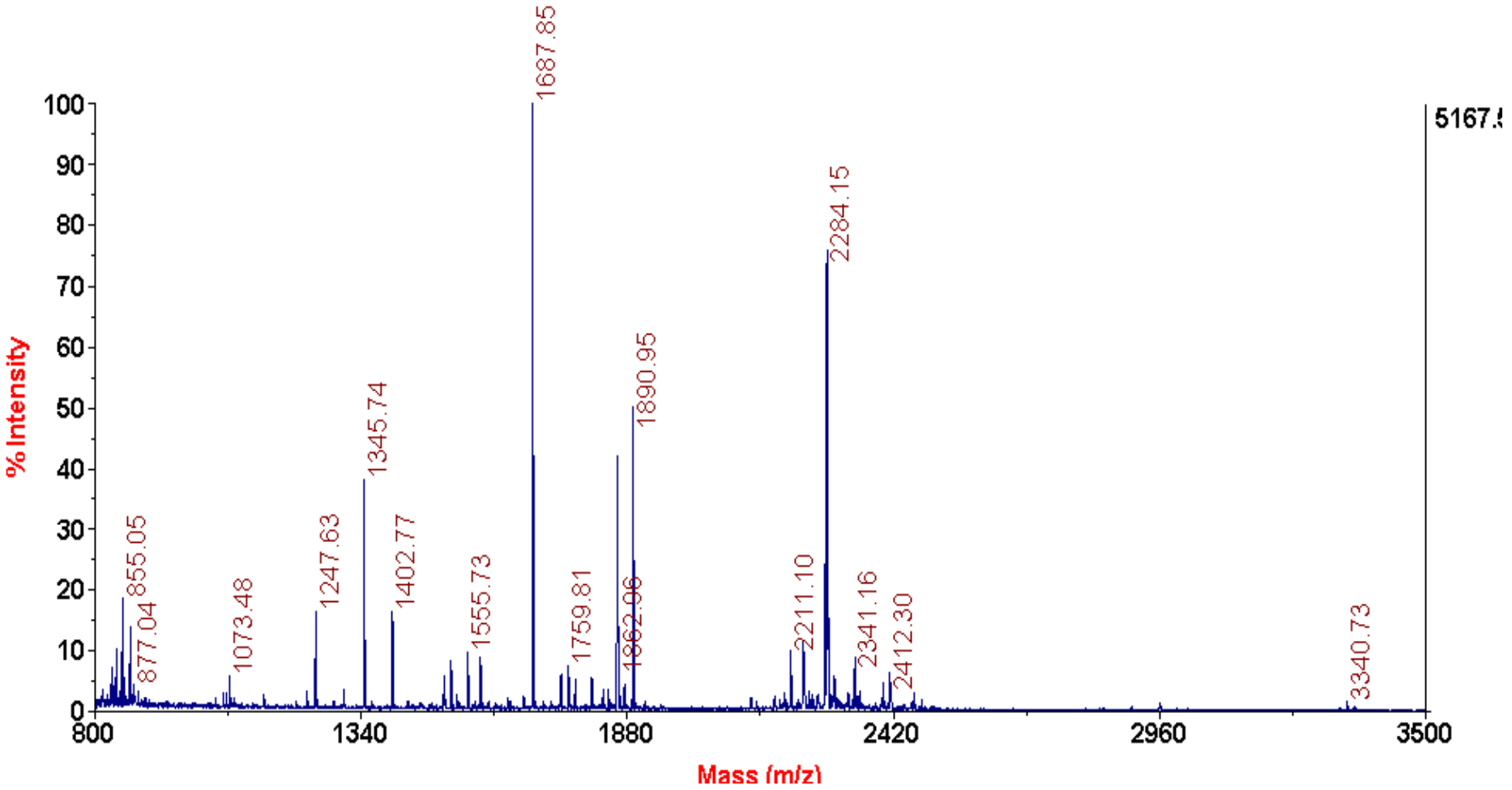
But: Not all peptides are observed by mass spectrometry, and some unexpected peptides are formed by non-specific cleavages.

MALDI of mixtures favors some peptides and suppresses others.

HPLC-ESI can fail to retain small hydrophilic peptides and large hydrophobic peptides may not elute from the column.

MALDI Mass Spectrum of a Tryptic Digest

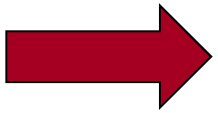
We assume there are no fragment ions and each peak represents a peptide formed by digestion of the protein.



Peaks are “deisotoped” and a peak list is generated and searched against a theoretical digest of a database

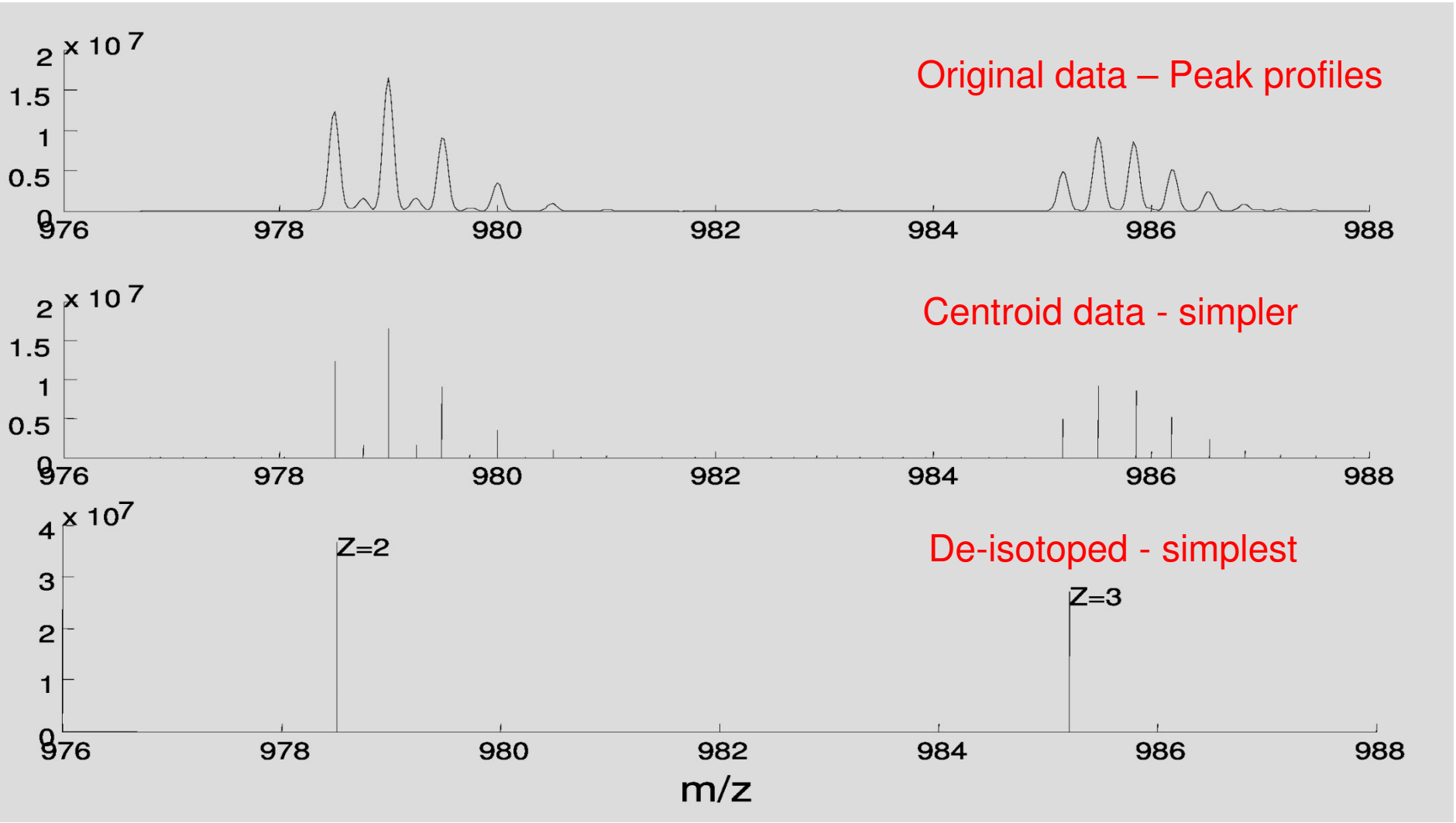
Monoisotopic Masses

771.478027	1773.904297
833.069885	3340.733154
842.510010	1807.806519
855.051453	1830.928589
861.066223	1841.965332
871.022034	1858.968140
877.037292	1862.963867
1073.484009	1873.940552
1247.629395	1890.951782
1304.651123	2211.104736
1345.741699	2236.117920
1402.770264	2250.125977
1507.718140	2281.173584
1522.794678	2284.152100
1555.726318	2300.157227
1581.734375	2341.159668
1687.847168	2398.184326
1744.864990	2412.297363
1759.811157	2461.312012



Database Search Program

Data processing before database searching



- This simplifies and reduces the amount of data and speeds up searches.
- But due to noise, overlapping isotope patterns, etc., it can introduce errors by wrong identification of the first isotope peak.

PMF Database Search Engines

Software is required to search the observed peptides against predictions from a theoretical digestion of all proteins in a database.

Protein Prospector

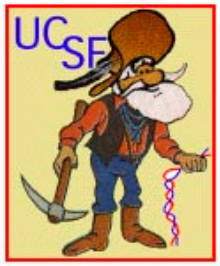
Developed at UCSF. Provides a suite of tools for all kinds of proteomic analysis, including protein mass fingerprinting, MSMS analysis, theoretical protein digestion, peptide fragmentation tools, etc.

Mascot

Search engine for analyzing protein mass fingerprinting data and LC-MSMS data.

- Data is input and searched in a similar fashion for both, but they have different 'scoring systems' for deciding which matches are correct.
- Both are publicly available for on-line searching or users can purchase licenses for dedicated in-house versions.

orts quantitation for Analyst and Xcalibur data files! Instructions available!



ProteinProspector

v 5.3.2

Proteomics tools for mining sequence databases in conjunction with Mass Spectrometry experiments.

New version of ProteinProspector with batch MSMS searching!

ProteinProspector Asia Pacific

ProteinProspector Tools

Batch MSMS Database Searching Instructions

Search Compare | Batch-Tag Web | Batch-Tag

Results Management | Search Table

Database Search Programs

MS-Fit | MS-Tag | MS-Homology | MS-Bridge

MS-Fit Upload | MS-Seq | MS-Pattern | MS-NonSpecific

Peptide / Protein MS Utility Programs

MS-Digest | MS-Product

MS-Isotope | MS-Comp

Database Management

DB-Stat

Administration/Help

Administering ProteinProspector
User's Manual
FAQ
Bug Listing
ProteinProspector Revision History
ProteinProspector Automation Guidance

Useful Tables

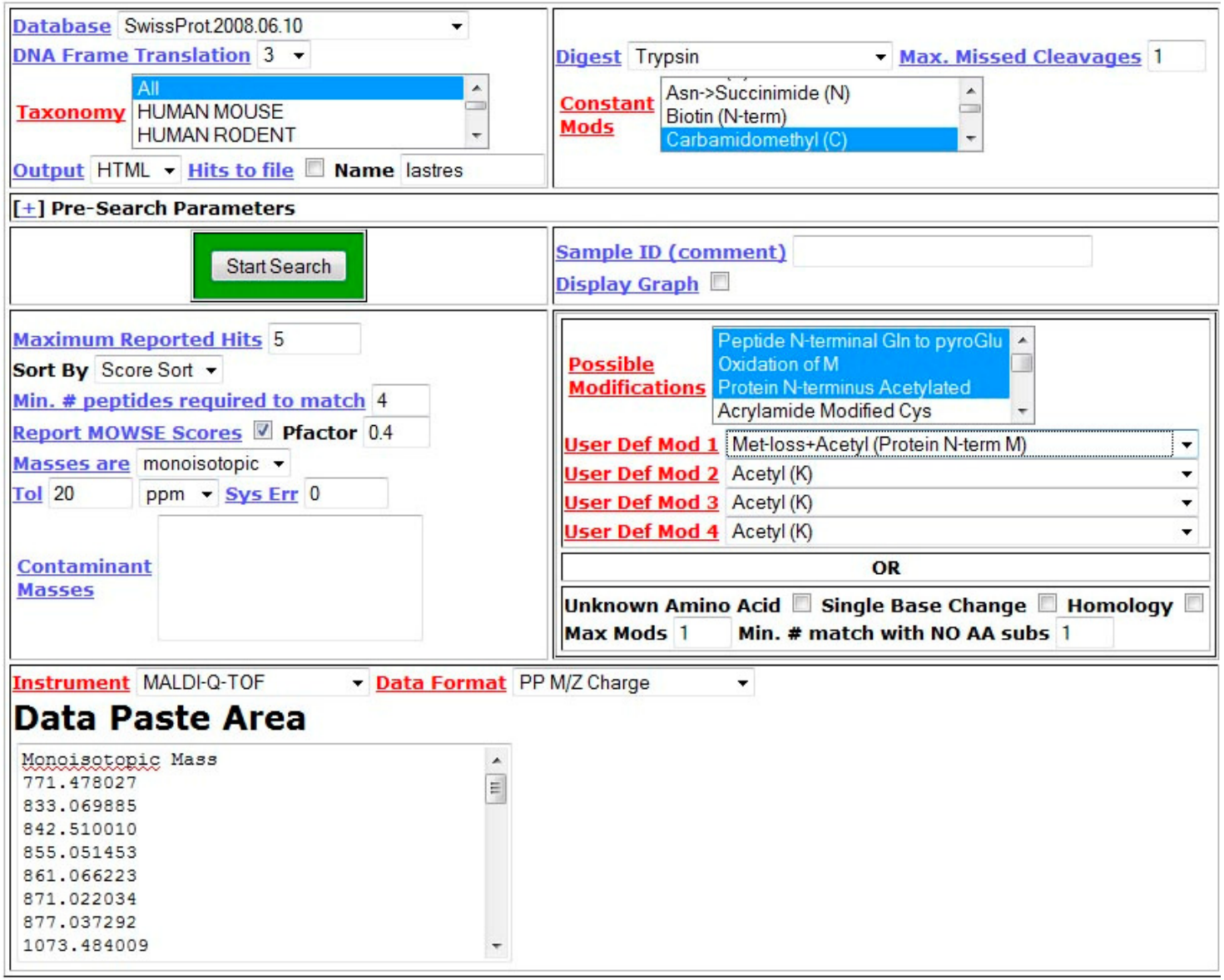
- Mutation Mass Shifts
- Dipeptide Masses
- Trypsin Autolysis Products

Publications
Useful Links

Questions/comments email: ppadmin@cgl.ucsf.edu

These programs were developed in the UCSF Mass Spectrometry Facility, which is directed by Dr. Alma Burlingame, Professor of Chemistry and Pharmaceutical Chemistry at UCSF and funded by the NIH National Center for Research Resources.

MS-Fit



In Protein Prospector the PMF program is MS-FIT.

The program allows the user to define certain parameters before carrying out a search against their specified database.

MS-Fit Search Results

[+] Parameters

[+] Pre Search Results

Fraction-Spot-Run ID: **1-1-1**
MS-Fit search selects **238** entries (results displayed for top **5** matches).

[−] Results Summary

Protein Hit Number	MOWSE Score	# pep # mat % mat 38 pks	% Cov	% TIC	Mean Err ppm	Data Tol ppm	# Hom Prot	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
1	1.77e+7	13/12/32	44.3	31.6	2.02	8.34	No	197745	42882/5.2	P01012	CHICK	Ovalbumin
2	10587	7/7/18	7.3	18.4	-1.54	21.6	No	301067	118589/5.4	Q7X923	ORYSJ	FACT complex subunit SPT16
3	2238	4/4/11	11.4	10.5	-5.21	29.2	No	42372	50598/6.4	Q5JLD8	ORYSJ	CBL-interacting protein kinase 8
4	1474	4/4/11	7.0	10.5	6.32	18.9	No	31318	66766/8.6	Q5XI58	RAT	Calicin
5	1133	5/5/13	12.5	13.2	-4.31	24.0	No	52140	65862/9.7	Q5QNV8	MOUSE	Uncharacterized protein C17orf66 homolog

197745	.	.	.	.	.	.	.	.	x	.	x	.	.	x	x	x	x	.	.	x	x	.	.	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.
301067	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	.
42372	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	x	.	.	x	.	x	.	.	.	.	.	.	.	.	.	.	.	.
31318	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.	.	x	.	.
52140	.	.	.	.	.	.	.	.	.	.	.	x	.	x	.	.	x	.	.	x	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

[+] Detailed Results

MS-Fit in ProteinProspector 5.3.2

Detailed Results

1. 12/38 matches (31%).
Acc. #: P01012 **Species:** CHICK **Name:** Ovalbumin
Index: 197745 **MW:** 42882 Da **pI:** 5.2

m/z Submitted	MH^+ Matched	Intensity	Delta ppm	Modifications	Start	End	Missed Cleavages	Sequence
1247.6294	1247.6241	100.0	4.21		361	370	0	(R)ADHPFLFC(Carbamidomethyl)IK(H)
1345.7417	1345.7375	100.0	3.10		371	382	0	(K)HIATNAVLFFGR(C)
1522.7947	1522.7974	100.0	-1.80		112	123	0	(R)YPILPEYLQC(Carbamidomethyl)VK(E)
1555.7263	1555.7210	100.0	3.45		188	200	1	(K)AFKDEDTQAMPFR(V)
1581.7344	1581.7213	100.0	8.24		265	277	0	(K)LTEWTSSNVMEER(K)
1687.8472	1687.8398	100.0	4.35		128	143	0	(R)GGLEPINFQTAADQAR(E)
1773.9043	1773.8991	100.0	2.95		324	340	0	(K)ISQAVHAAHAEINEAGR(E)
1807.8065	1807.8030	100.0	1.96	1Met-loss+Acetyl	1	17	0	(-)MGSIGAASMEFC(Carbamidomethyl)FDVFK(E)
1858.9681	1858.9658	100.0	1.28		144	159	0	(R)ELINSWVESQTNGIIR(N)
2281.1736	2281.1823	100.0	-3.82		86	105	0	(R)DILNQITKPNDVYSFSLASR(L)
2284.1521	2284.1464	100.0	2.48		201	219	0	(R)VTEQESKPVQMMYQIGLFR(V)
2284.1521	2284.1682	100.0	-7.06		106	123	1	(R)LYAEERYPILPEYLQC(Carbamidomethyl)VK(E)
2300.1572	2300.1414	100.0	6.90	1Oxidation	201	219	0	(R)VTEQESKPVQMMYQIGLFR(V)

Num Unmatched Masses: **26**
Search for disulfide linked peptides.
Do a non-specific cleavage search.
Search for another component.

The matched peptides cover **44.3%** (171/386AA's) of the protein.
Coverage Map for This Hit (MS-Digest index #): 197745

2. 7/38 matches (18%).
Acc. #: Q7X923 **Species:** ORYSJ **Name:** FACT complex subunit SPT16
Index: 301067 **MW:** 118589 Da **pI:** 5.4

m/z Submitted	MH^+ Matched	Intensity	Delta ppm	Modifications	Start	End	Missed Cleavages	Sequence
1522.7947	1522.7972	100.0	-1.68		616	628	1	(K)DPRHSSEVVQQIK(T)
2284.1521	2284.1842	100.0	-14.0		721	739	0	(K)EMITLLHFHLHNHIMVGNK(K)
2300.1572	2300.1791	100.0	-9.50	1Oxidation	721	739	0	(K)EMITLLHFHLHNHIMVGNK(K)
2341.1597	2341.1605	100.0	-0.358	1Oxidation	743	762	1	(K)DVQFYVEVMDVVQTLGGNRR(S)
2398.1843	2398.2071	100.0	-9.51		741	761	1	(K)TKDVQFYVEVMDVVQTLGGNR(R)
2412.2974	2412.2791	100.0	7.56		721	740	1	(K)EMITLLHFHLHNHIMVGNKK(T)
2461.3120	2461.2708	100.0	16.7		884	905	1	(R)IDSIPSTSLDAIKEWLDTTDLK(Y)

MASCOT Peptide Mass Fingerprint

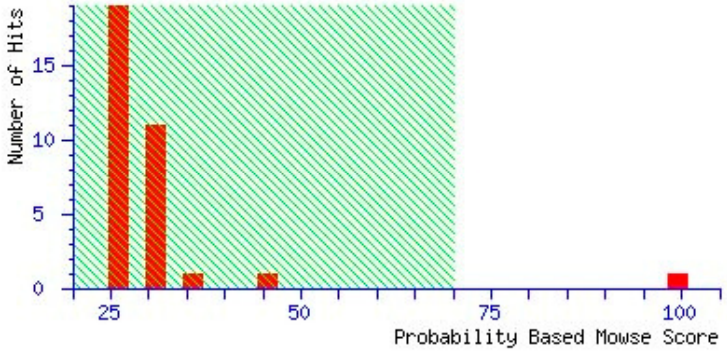
Your name	Mike	**Email**	mbaldwin@cgl.ucsf.edu
Search title			
Database	SwissProt		
Taxonomy	All entries		
Enzyme	Trypsin	**Allow up to**	1 missed cleavages
Fixed modifications	Biotin (N-term) Carbamidomethyl (C) Carbamyl (K) Carbamyl (N-term) Carboxymethyl (C)	**Variable modifications**	mTRAQ:13C(3)15N(1) (Y) NIPCAM (C) Oxidation (HW) Oxidation (M) Phospho (ST)
Protein mass	kDa	**Peptide tol. ±**	20 ppm
Mass values	MH^+ M_r $M\text{-}H^-$	**Monoisotopic**	**Average**
Data file	Browse...		
Query NB Contents of this field are ignored if a data file is specified.	771.478027 833.069885 842.510010 855.051453 861.066223 871.022034 877.037292		
Decoy		**Report top**	AUTO hits
	Start Search ...		**Reset Form**

Mascot Search Results

User : Mike
Email : mbaldwin@cgl.ucsf.edu
Search title :
Database : SwissProt 57.15 (515203 sequences; 181334896 residues)
Timestamp : 10 Mar 2010 at 02:33:42 GMT
Top Score : 99 for OVAL_CHICK, Ovalbumin OS=Gallus gallus GN=SERPINB14 PE=1 SV=2

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 70 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. OVAL_CHICK Mass: 43196 Score: 99 Expect: 5.9e-05 Queries matched: 11
Ovalbumin OS=Gallus gallus GN=SERPINB14 PE=1 SV=2

2. NUSB_GEOSF Mass: 16001 Score: 46 Expect: 12 Queries matched: 4
N utilization substance protein B homolog OS=Geobacter sp. (strain FRC-32) GN=nusB PE=3 SV=1

3. T1SD ECOLX Mass: 50034 Score: 37 Expect: 1.1e+02 Queries matched: 5
Type-1 restriction enzyme EcoDI specificity protein OS=Escherichia coli GN=hsdS PE=3 SV=1

4. UCP3 CANFA Mass: 34572 Score: 32 Expect: 3e+02 Queries matched: 4
Mitochondrial uncoupling protein 3 OS=Canis familiaris GN=UCP3 PE=2 SV=1

5. GSA AERS4 Mass: 46305 Score: 32 Expect: 3.1e+02 Queries matched: 4
Glutamate-1-semialdehyde 2,1-aminomutase OS=Aeromonas salmonicida (strain A449) GN=hemL PE=3 SV=1

Search Parameters

```
Type of search         : Peptide Mass Fingerprint
Enzyme                 : Trypsin
Fixed modifications    : Carbamidomethyl (C)
Variable modifications : Oxidation (M)
Mass values            : Monoisotopic
Protein Mass           : Unrestricted
Peptide Mass Tolerance : ± 20 ppm
Peptide Charge State   : 1+
Max Missed Cleavages   : 1
Number of queries      : 38
```

Databases

- *SwissProt – well curated, manually annotated with detailed protein descriptions and some known PTMs.
- *Uniprot – Combination of SwissProt and TrEMBL. Much larger than SwissProt. All entries annotated, but TrEMBL annotated automatically.
- NCBI – combination of GeneProt, SwissProt, Refseq, PIR, PRF, PDB… Very large, but many entries per protein and some with no annotation. Lot of redundancy.
- dbEST – translation of Genbank cDNA sequences – i.e. predicted coding sequences. Very large!
- Species specific databases: Yeast, Human, Fruit Fly… Small, but generally well annotated.

Dilema: Small databases give better results, i.e. small is better – as long as the chosen database includes the protein of interest.

Mass accuracy affects the number of peaks required for a correct match

Table 3. MS-Fit Searches[1] at Various Mass Tolerances Using 23 Masses Measured in Figure 2 (Dashed Lines Show Levels Below Which Only the Correct Proteins Are Matched)

Minimum # Peptides Matched	Number of Proteins Matched — Mass Tolerance supplied to MS-Fit						
	±2.0 Da	±1.0 Da	±0.5 Da	±0.3 Da	±0.1 Da	±50 ppm	±10 ppm
1	156,793	117,419	77,906	77,374	63,730	47,461	11,703
2	104,022	58,188	24,997	24,708	16,842	9,344	723
3	67,400	26,460	7,455	7,297	4,087	1,766	36
4	42,295	11,623	2,048	1,991	923	323	7
5	25,638	4,846	509	496	190	44	3
6	14,987	1,882	145	135	51	8	3
7	8,192	687	36	33	10	3	3
8	4,378	248	12	9	3	3	3
9	2,208	88	3	3	3	3	3
10	1,062	35	3	3	3	3	3
11	466	9	3	3	3	3	3
12	200	3	3	3	3	3	3
13	72	3	3	3	3	3	3
14	34	3	3	3	3	3	2
15	12	3	3	3	3	3	2
16	3	3	3	3	2	2	0

Clauser, K. et al. *Anal Chem* (1999) **71** 2871-2882

A high performance instrument can achieve +/- 10 ppm or better.

Peptide Modifications Commonly Observed

Chemically induced, either deliberately or unintentionally:

- Carbamidomethylation of Cys +57 Da
- Oxidation of Met +16 Da
- Pyroglutamate formation -17 Da
- Deamidation of Asn (or Gln) +1 Da

PTM's etc

- Acetylation +42 Da
- Phosphorylation +80 Da
- Sulfation +80 Da
- Methylation +14 Da
- Glycosylation +Various

Analytical adducts

- Sodium ion instead of proton +22 Da
- Potassium ion instead of proton +38 Da
- Detergents, phosphate, etc. +Various

Other "artifact" peaks may be seen for enzyme self-digestion, impurities, etc.

What Modifications Should You Search For?

Only search for modifications that are common or you have reason to expect, such as:

- Fixed: carbamidomethyl cysteine. We assume every Cys is modified so this does not alter the number of potential peptides or the size of the database.
- Variable: N-Acetyl (protein); oxidised Met; pyroGlu (from Q). We assume these MAY occur so we test for both unmodified and modified versions.

Variable modifications increase the number of potential peptides. e.g. A single peptide containing 2 serine residues. Allowing for serine phosphorylation this results in 4 possible versions:

GSGASMER G**S**IGASMER GSIGA**S**MER G**S**IGA**S**MER

Consequently variable modifications cause databases to become substantially larger, slow down searches and increase the chance of false positive matches.

How are PMF Results Scored / Results Ranked?

- Which protein matches the highest fraction of the peptide masses observed?
- What is the probability that ‘x’ peaks match to a given protein at random?

What will affect this probability?

- How many peaks are submitted for the search?
- What mass accuracy are you allowing for the peaks?
- Size of protein: bigger protein will form more tryptic peptides, so is likely to match more peptides at random.
- Number of proteins in the database.
- What modifications you allow for.

- The scoring algorithm most commonly used is the “molecular weight search” (MOWSE) developed by Pappin et al, 1993.

PMF Conclusions

PMF has advantages:

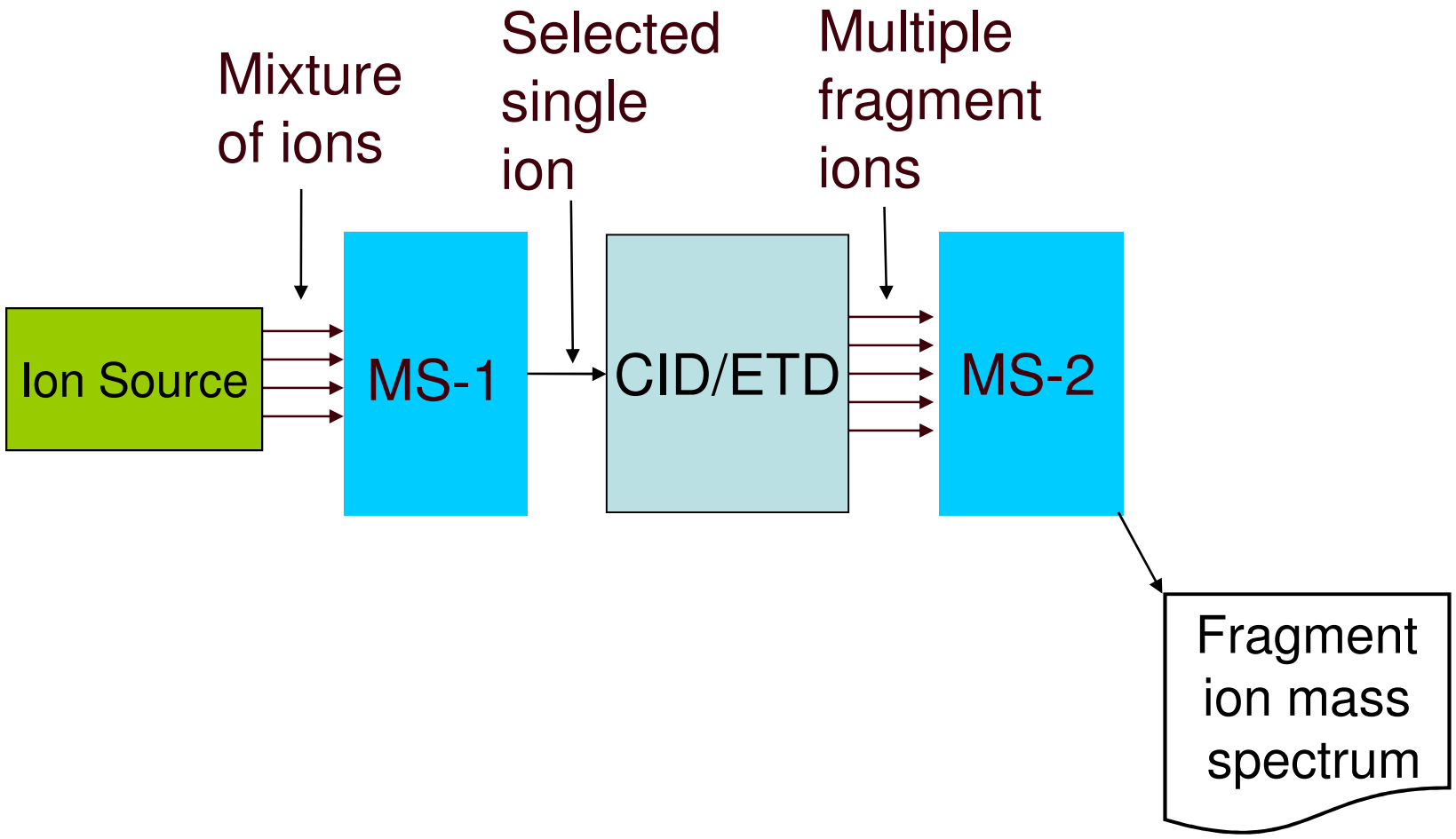
- Quick and simple to acquire data.
- Sensitive.
- Data can be obtained on a relatively simple mass spectrometer as MSMS is not required.

And disadvantages:

- Not good for protein mixture analysis (even a simple mixture).
- Confidence of many search result assignments is low.

Enhanced alternatives involve collision induced dissociation (CID) and/or electron transfer dissociation (ETD) and sequence analysis, usually of peptides within the same digest as PMF, but sometimes of intact proteins.

MS/MS (Tandem Mass Spectrometry)



Advantages of MS/MS Analysis

- More specific and reliable than peptide mass fingerprinting
- Searches employ the intact peptide mass as well as the masses of fragment ions.
- All fragment ions should be derived from the selected precursor ion.
- Protein identifications can be made on the basis of as few as one or two peptides.
- MS/MS allows the identification of proteins in complex mixtures.

Note: MS/MS can also be used for *de novo* sequencing; i.e. when the protein sequence is not previously known or in the database.

Why Trypsin?

There are very many specific proteases so why is trypsin widely favored for PMF and MS/MS?

- It is highly specific and digests at basic residues (Arg and Lys) that are common and widely distributed throughout most proteins. Consequently it produces peptides of a size generally amenable to MS analysis.
- Except for the peptide from the protein C-terminus, all other tryptic peptides have a basic residue at their C-terminus (Arg or Lys) which is a natural site for a positive charge. Such peptides are favored to give strong singly charged ions in MALDI or doubly charged ions in ESI, the 2^{nd} charge being at the N-terminal amino group.
- In ESI-MS/MS the basic residue at the C-terminus favors the formation of strong y-ion series.

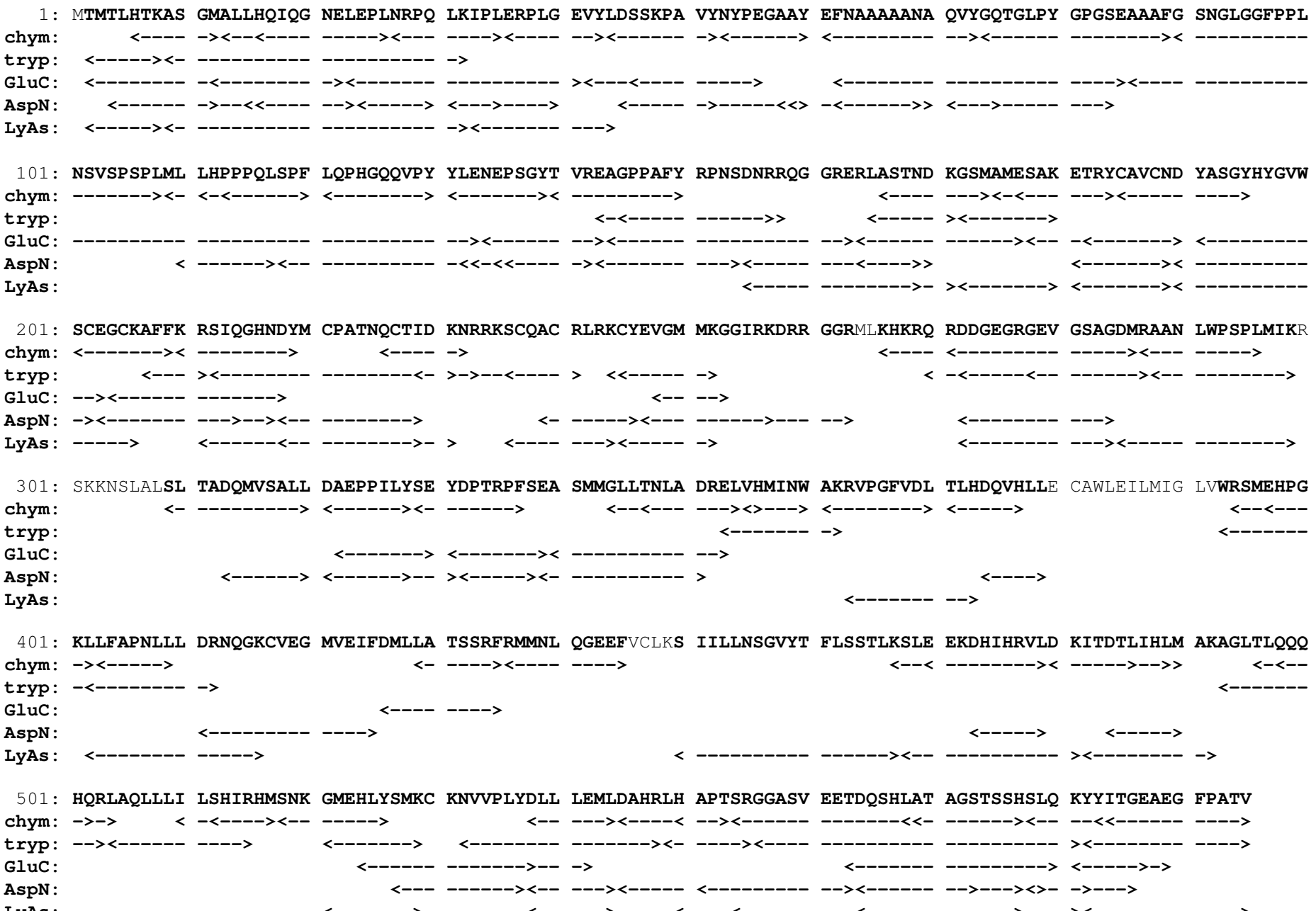
Note: Trypsin also digests itself, giving known autolysis products that can serve as useful mass markers.

Why NOT Trypsin?

- In some protein regions Arg and Lys residues may come very close together giving small peptides (di- and tri-peptides) that are too small for most MALDI experiments and are not retained on HPLC columns. In such cases Lys-C may be better as it digests only at Lys residues.
- Conversely, some proteins have regions that are devoid of basic residues, giving rise to very large peptides outside the range of routine MS or MS/MS experiments
- In such cases other proteases or combinations of proteases may be favored, e.g. Asp-N, Glu-C, chymotrypsin.

Note: For complete *de novo* sequence analysis of a protein as distinct from protein ID, it is usually necessary to carry out multiple different digestions, each of which can reveal different and overlapping regions of the sequence.

Estrogen Receptor: 94% sequence coverage by MSMS



Amino Acid Residue Masses (Molecular mass minus H_2O)

Amino acid residue		Monoisotopic mass
Ala	A	71.03711
Cys	C	103.00919
Asp	D	115.02694
Glu	E	129.04259
Phe	F	147.06841
Gly	G	57.02146
His	H	137.05891
Ile	I	113.08406
Lys	K	128.09496
Leu	L	113.08406
Met	M	131.04049
Asn	N	114.04293
Pro	P	97.05276
Gln	Q	128.05858
Arg	R	156.10111
Ser	S	87.03203
Thr	T	101.04768
Val	V	99.06841
Trp	W	186.07931
Tyr	Y	163.06333

Modified Amino Acid Residue	Monoisotopic mass
Homoserine Lactone	83.03712
Pyroglutamic acid	111.03203
Hydroxyproline	113.04768
Oxidised Methionine	147.03541
Carbamidomethylcysteine	160.03065

The mass of a peptide is equal to the sum of the masses of the residues plus the mass of H_2O (18.01528).

The mass of a singly charged peptide ion is greater by an H atom (1.007825) minus the mass of an electron (0.000547).

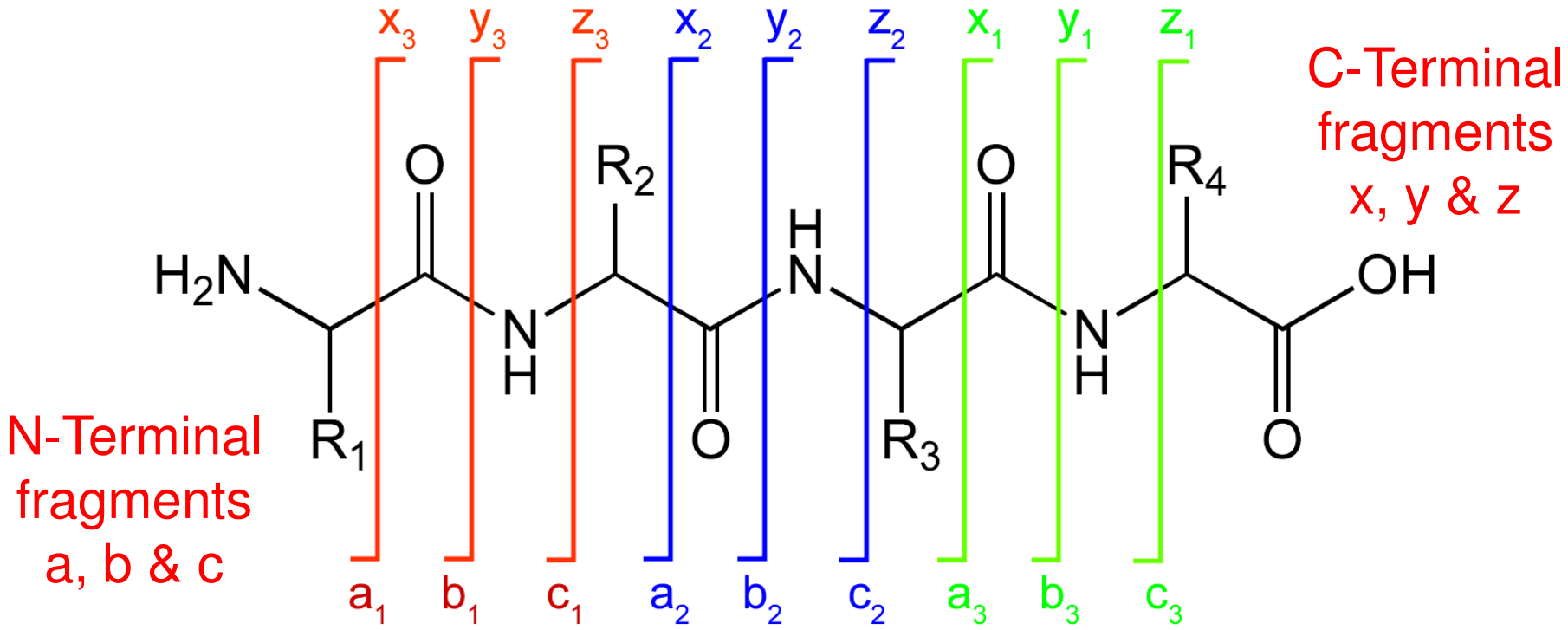
It is useful to learn the integer mass of each amino acid so the you can calculate the nominal mass of a peptide and predict simple fragment masses.

Ion Fragmentation Methods in MS/MS

- Thermal / energy based fragmentation
  - Introduces vibronic energy into molecule and breaks the weakest bonds.
    - Collision-Induced Dissociation (CID) (common)
    - Surface-Induced Dissociation (SID) (uncommon)
    - Infra-Red MultiPhoton Dissociation (IRMPD) (uncommon)

- Radical-based fragmentation
  - Introduces an electron to create an unstable radical ion, which spontaneously fragments at sites related to the location of electron capture.
    - Electron Capture Dissociation (ECD)
    - Electron Transfer Dissociation (ETD)

Peptide Fragmentation. Roepstorff and Fohlman (1984). Biemann, (1990)



- As shown the fragmentation is of a neutral molecule whereas the peptide ion is actually protonated and is an even-electron species. Backbone cleavage is usually accompanied by a hydrogen rearrangement to retain this favored state.

- The numbering of the N-terminal residues 1, 2, 3, etc. is independent of the numbering of the C-terminal residues as both termini start at 1. This has the advantage that it is not necessary to know the total number of residues in a peptide to assign ion labels.

High and Low Energy CID

Tandem mass spectrometers either impart high kinetic energy to ions (TOF/TOF) or low kinetic energy (QIT, QQQ, QTOF). This has some effect on CID and the type of fragment ions formed. The instrument type also affects the ability to monitor some ions, particularly low mass ions.

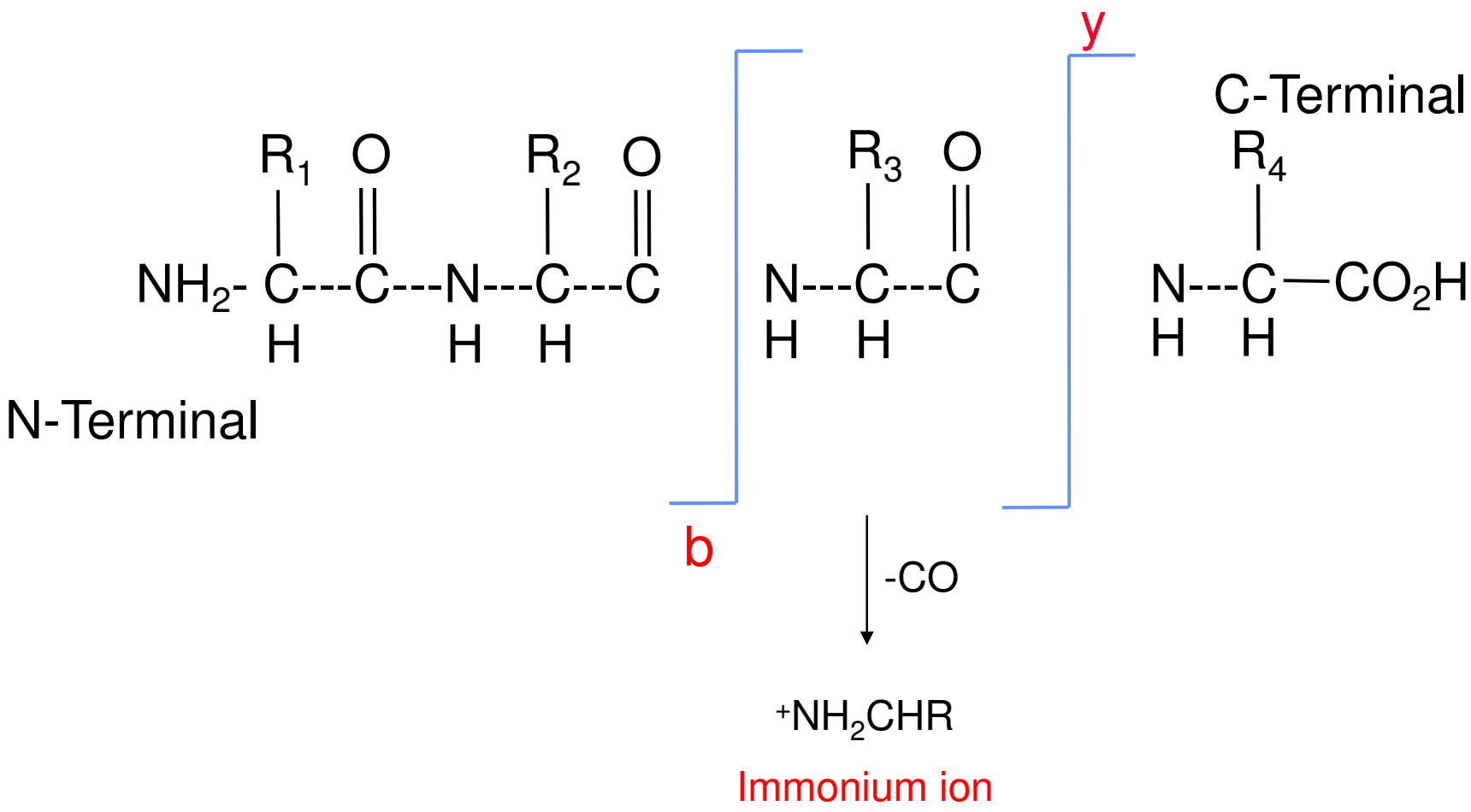
- TOF/TOF: Generally get single bond cleavages with minimal rearrangements. Multiple higher energy backbone cleavages occur in addition to the lower energy b- and y-ion hydrogen rearrangements. Small fragments characteristic of specific amino acids (immonium ions) are also seen.

- Ion trap: Excitation (for CID) is m/z dependent. Once an ion has fragmented its m/z changes so it is no longer excited. QIT generally gives a single fragmentation event and multiple fragmentation events are rare.

- QQQ or QTOF: Fragment ions retain vibronic energy and may give multiple fragmentation events. QTOF (QSTAR) gives higher selectivity, resolution, mass accuracy and the spectra show the low mass fragment ions.

Immonium Ions

A special type of a ion characteristic of a given amino acid



Immonium Ion Masses

IMMONIUM AND RELATED IONS CHARACTERISTIC OF THE 20 STANDARD AMINO ACIDS[a]

Amino acid	Immonium and related ion(s) masses		Comments
Ala	44		
Arg	129	59, 70, 73, 87, 100, 112	129, 73 usually weak
Asn	87	70	87 often weak, 70 weak
Asp	88		Usually weak
Cys	76		Usually weak
Gly	30		
Gln	101	84, 129	129 weak
Glu	102		Often weak if C-terminal
His	110	82, 121, 123, 138, 166	110 very strong 82, 121, 123, 138 weak
Ile/Leu	86		
Lys	101	84, 112, 129	101 can be weak
Met	104	61	104 often weak
Phe	120	91	120 strong, 91 weak
Pro	70		Strong
Ser	60		
Thr	74		
Trp	159	130, 170, 171	Strong
Tyr	136	91, 107	136 strong, 107, 91 weak
Val	72		Fairly strong

The mass of a true immonium ion is the amino acid residue mass minus 27 Da

Fragment ions may lose water or ammonia

Losses can occur from a, b or y ions.

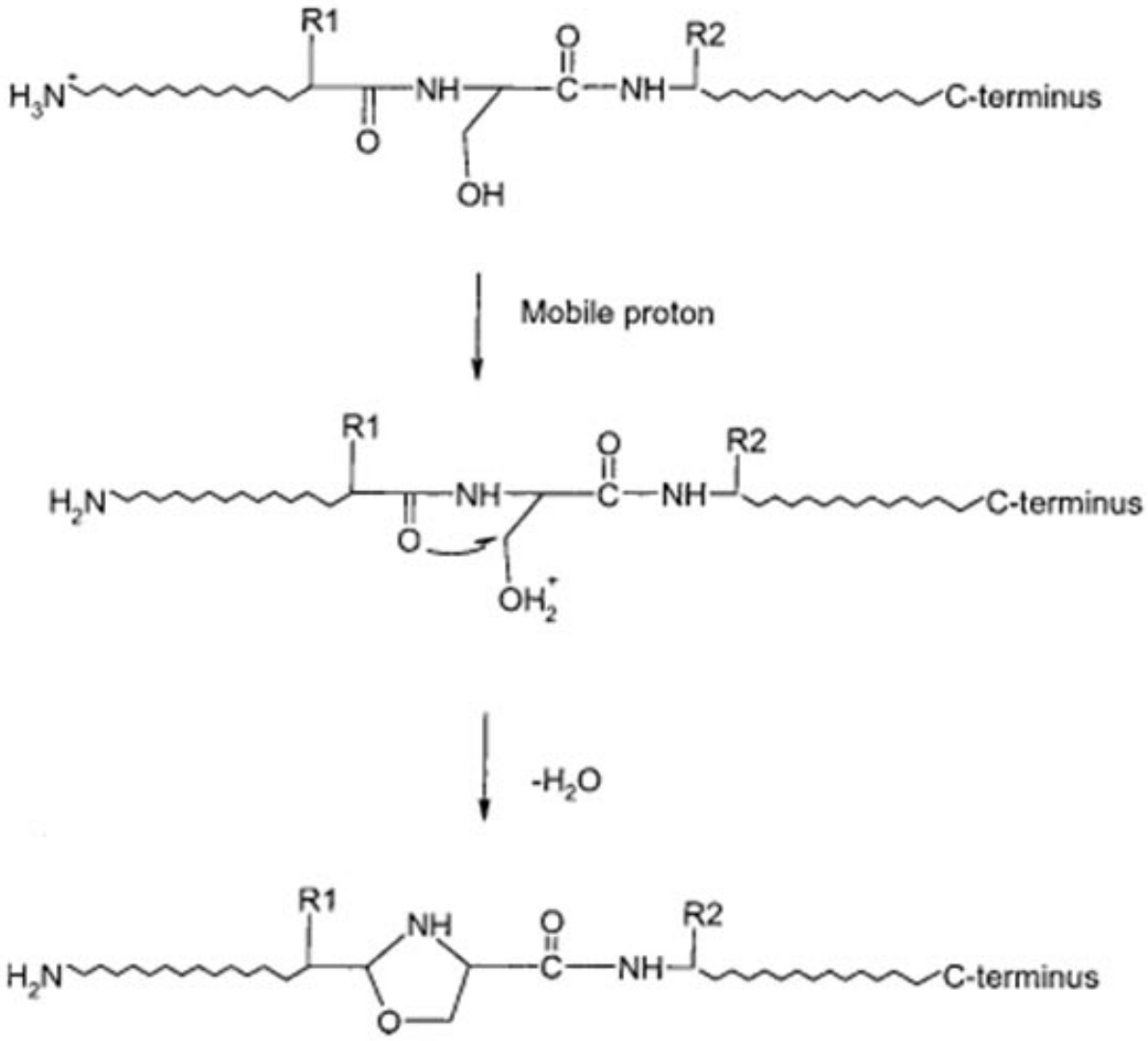
Water loss: e.g. y_n-18

S, T, E and D.

The figure shows a possible mechanism for water loss from S

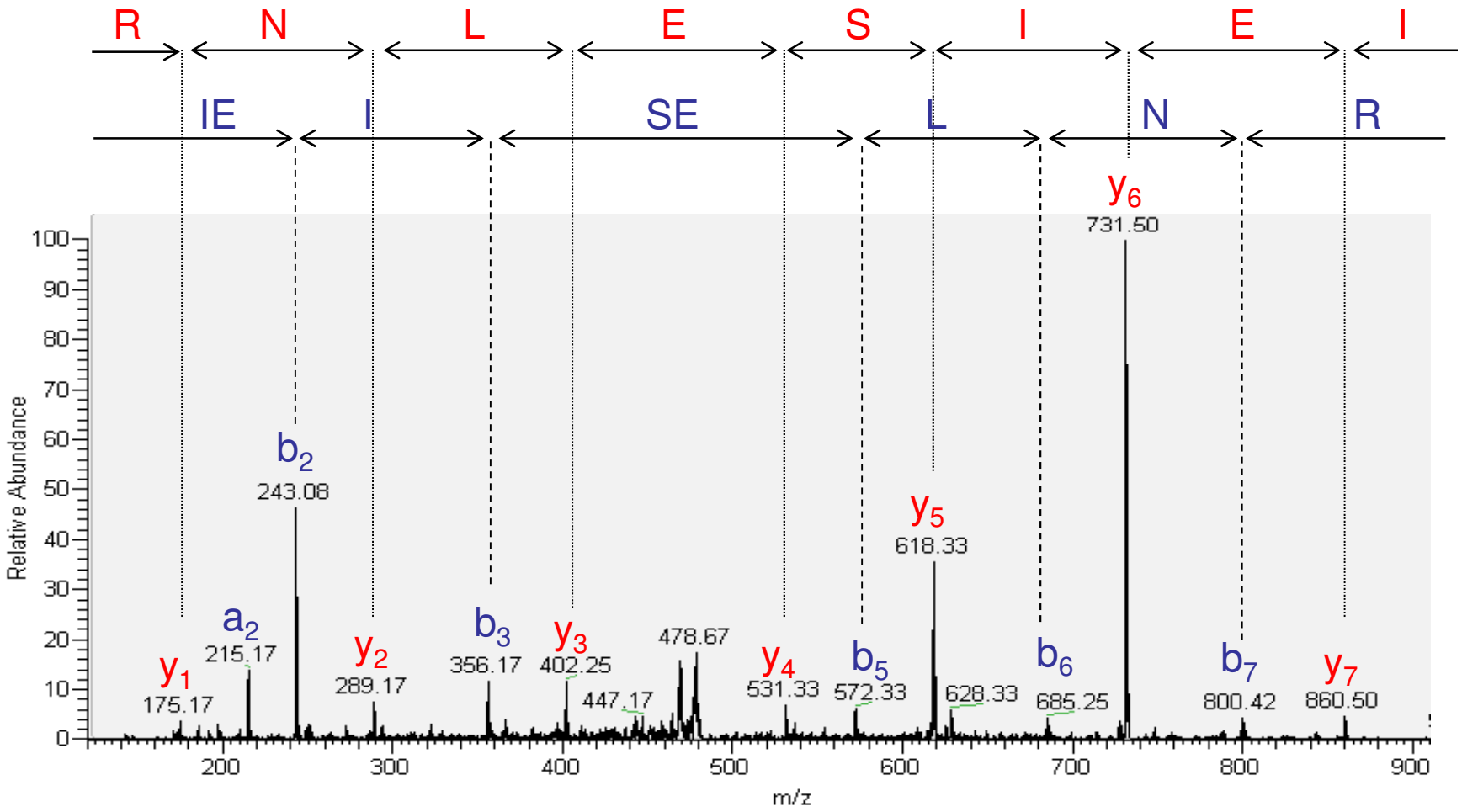
Ammonia loss: e.g. y_n-17

R, K, N and Q



ESI-MSMS 487.27^{2+} IEISELNR in ion trap

- The precursor is doubly charged but the fragment ions are singly charged.
- The y-ion series is stronger than the b-ion series.
- Low energy MSMS cannot distinguish between the isomers Leu and Ile.



Why are some fragment ions more intense than others? (and some aren't even detectable!)

- Amino acids are chemical structures, not homogeneous 'building blocks', and the cleavage reactions of protonated peptide ions are subject to the normal rules of kinetics and thermodynamics.
- Consequently certain fragment ions are favored over others.
- Statistical analysis on large amounts of CID data allow some predictions of fragment ion intensities[1,2].

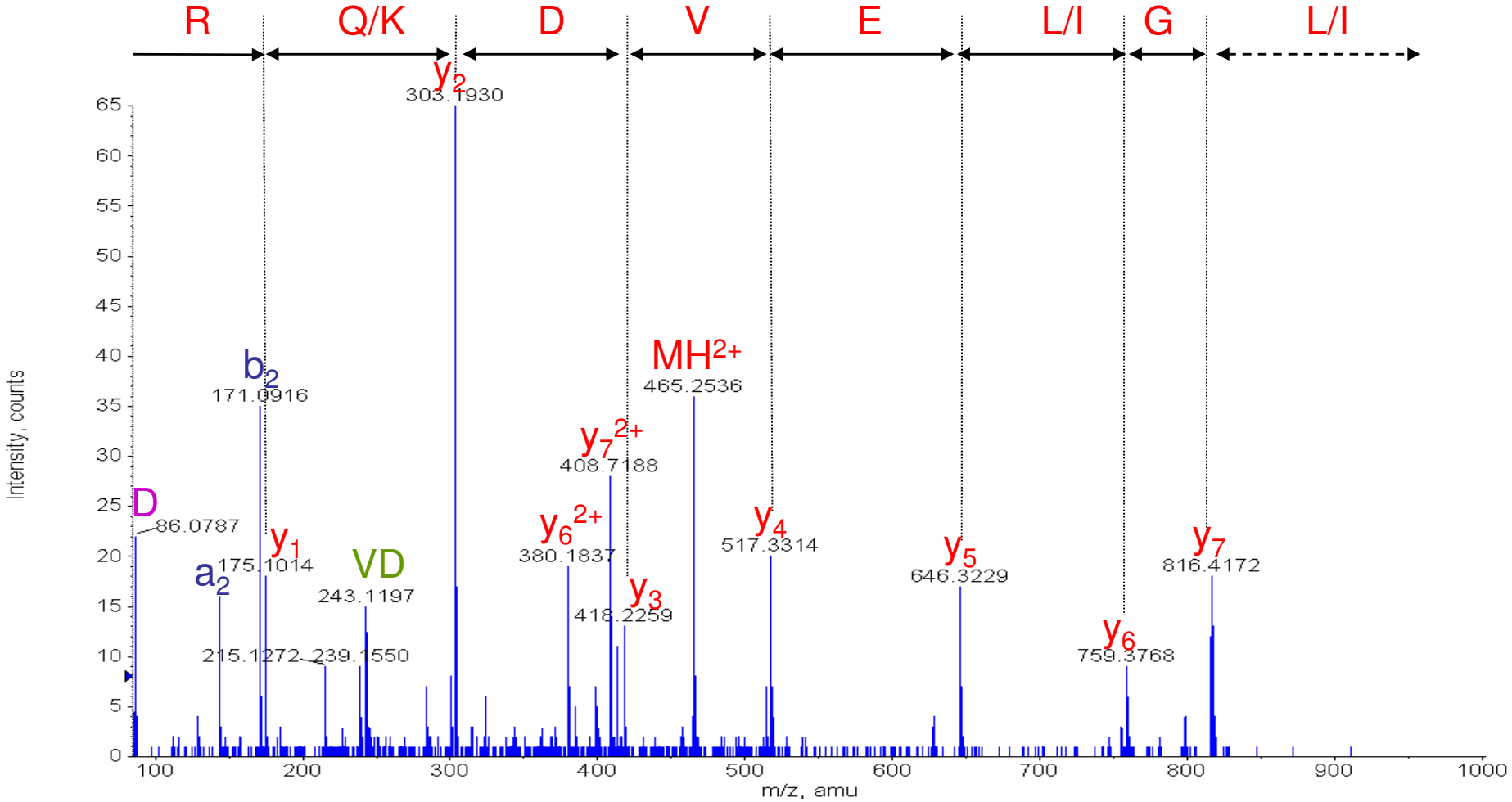
[1]Kapp, E.A. et al *Anal Chem* (2003) **75** 22: 6251-6264
[2]Huang, Y. et al. *Anal Chem* (2005) **77** 18: 5800-5813

Examples:

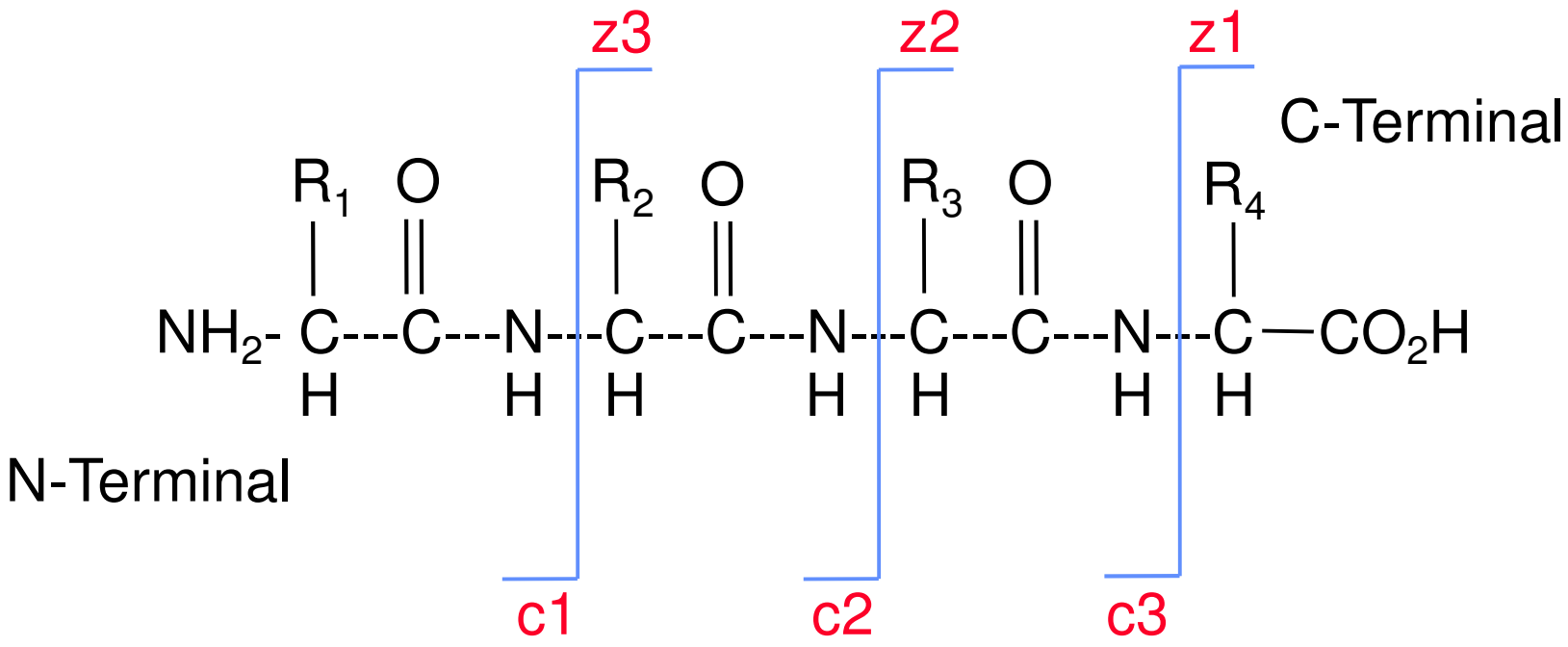
- Cleavage N-terminal to proline gives intense fragment ions.
- Cleavage C-terminal to proline generally is not seen.
- Cleavage C-terminal to aspartic acid is favored.

ESI-MSMS 465.25^{2+} IGLEVDKR in quadrupole

- Note that the strongest ion y_2 results from cleavage C-terminal to D.
- The precursor is doubly charged and so are some fragment ions. We could establish this by looking at the isotope peak spacing.



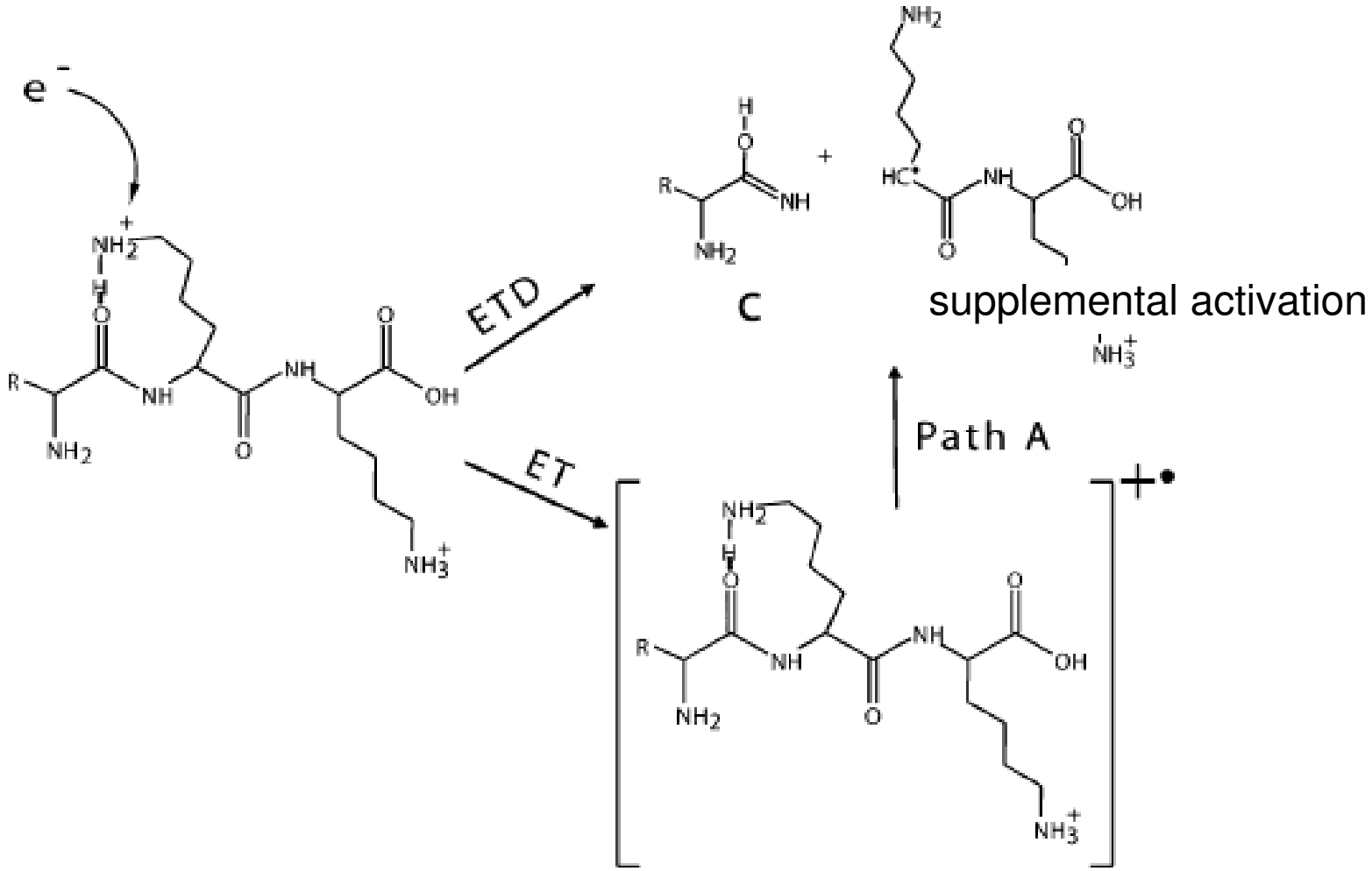
ETD / ECD favors c/z ion formation rather than b/y



Amide bond cleavages are not favored, unlike CID

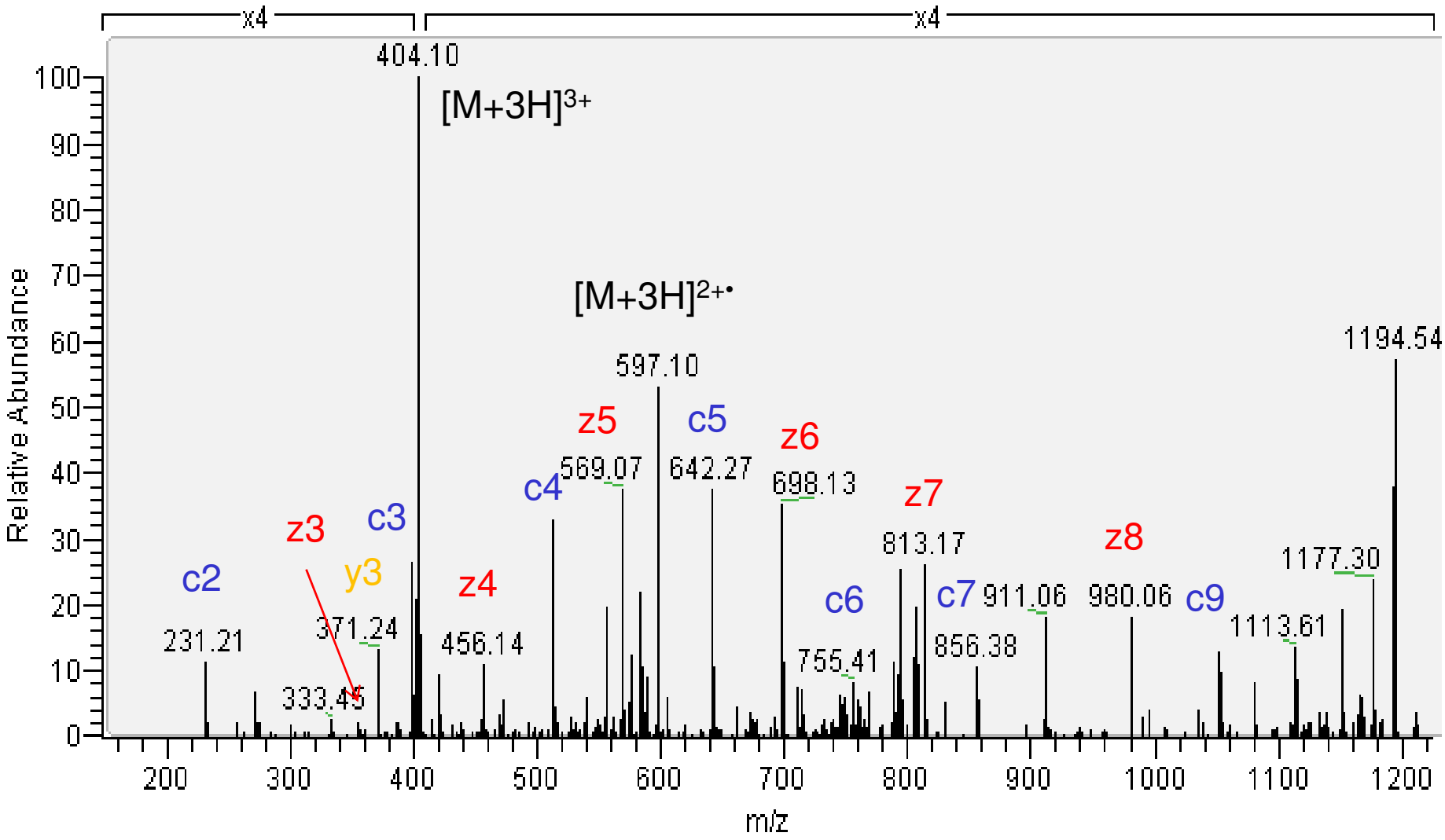
ETD / ECD reactions

- Radical cations are unstable and fragment rapidly.
- These reactions are promoted by unpaired electrons, not by protons.



ETD Spectrum of 3^+ Precursor

R G S(Phospho) D E L T V P R^{3+} : All identified fragments are c or z ions

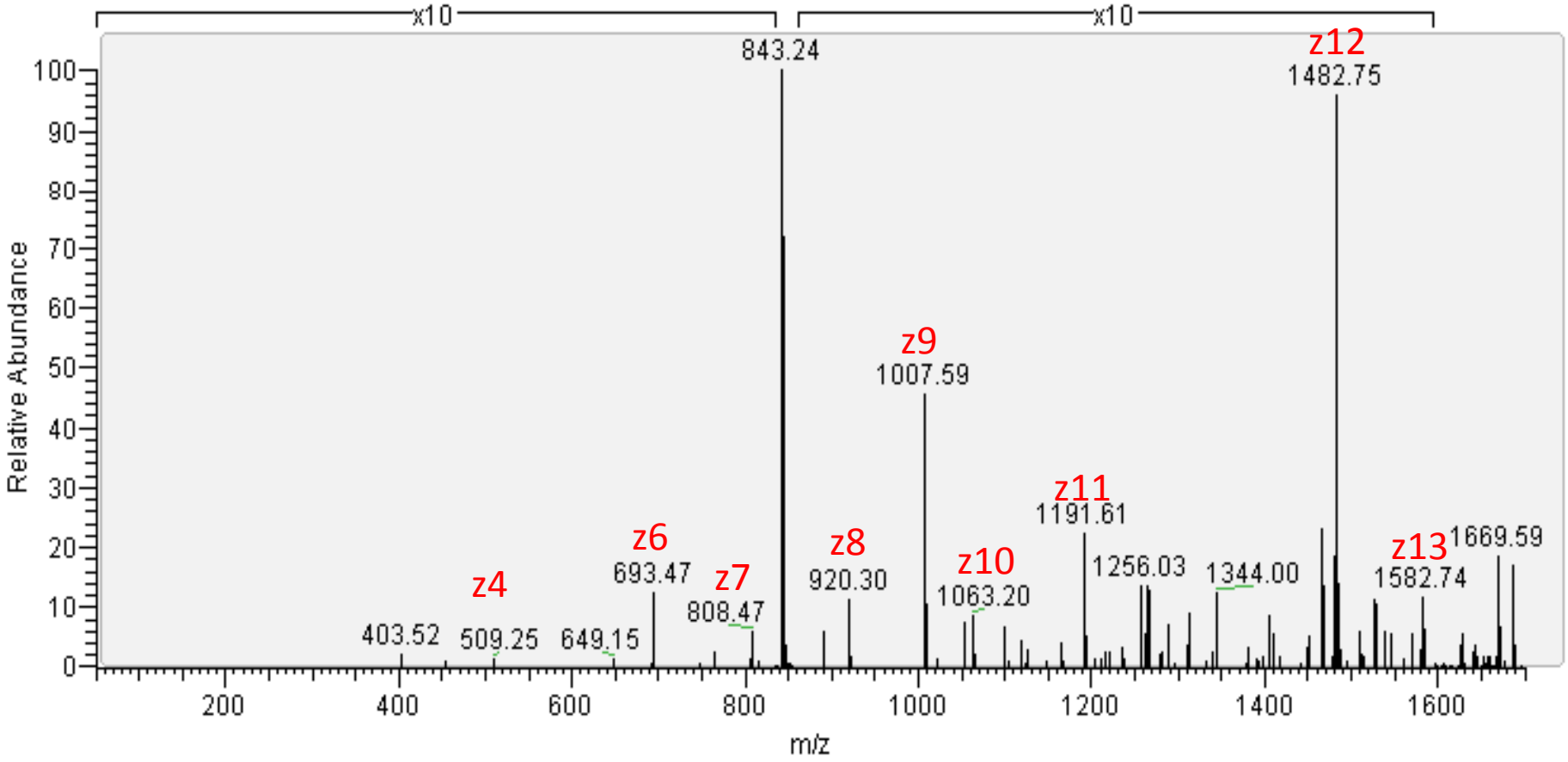


ETD Spectrum of 2^+ Precursor

m/z 843.402^{2+}

STS(HexNAc)QGSINSPVYSR - Actin-binding LIM protein 1

Mass difference between z_{11}-z_{12} identifies modification site as residue 496



Protein Prospector Predicts all Fragment Ions

The output from MS-Product is based only on arithmetic, not chemistry, and makes no predictions of ion intensities.

MH-H_3PO_4 ions							
MH-$SOCH_4$ ions							
MH-H_2O ions	782.3567						
MH-NH_3 ions							
MH ions	800.3672						
Immonium and Related Ions	70.0651 126.0550	102.0550	70.0651 126.0550	74.0600	86.0964	88.0393	102.0550

N-terminal ions

a-H_2O ions	---	181.0972	278.1499	379.1976	492.2817	607.3086	---
a ions	---	199.1077	296.1605	397.2082	510.2922	625.3192	---
b-H_2O ions	---	209.0921	306.1448	407.1925	520.2766	635.3035	---
b ions	---	227.1026	324.1554	425.2031	538.2871	653.3141	---

			1		2		3		4		5		6		7	
-	P		E		P		T		I		D		E	-
	7		6		5		4		3		2		1	

C-terminal ions

y ions	---	703.3145	574.2719	477.2191	376.1714	263.0874	148.0604
y-H_2O ions	---	685.3039	556.2613	459.2086	358.1609	245.0768	130.0499

[+] Internal Ions
[+] Theoretical Peak Table

Matching observed masses to ion types

Internal Ions

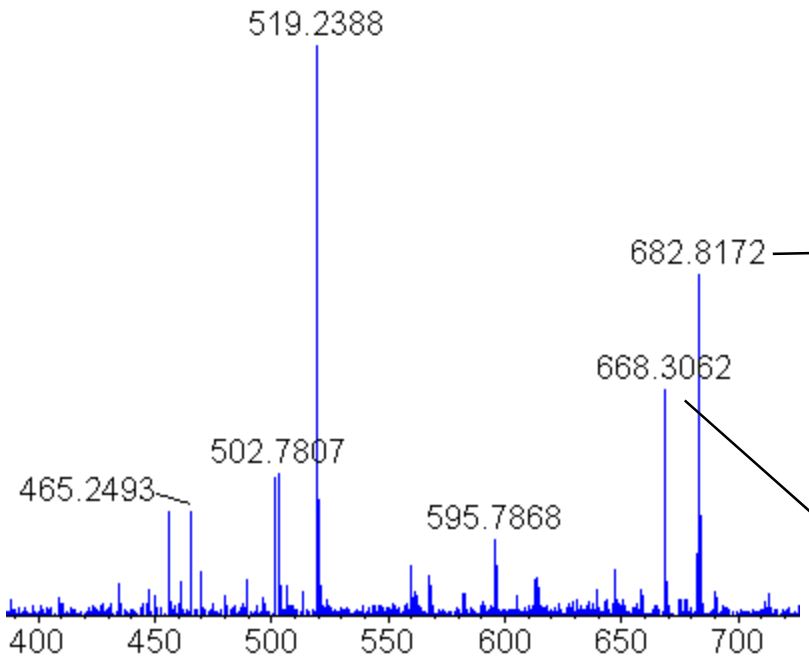
Internal Sequence	Internal ions	Internal-28 ions	Internal-NH_3 ions	Internal-H_2O ions
PT	199.1077	171.1128	---	181.0972
TI	215.1390	187.1441	---	197.1285
EP	227.1026	199.1077	---	209.0921
ID	229.1183	201.1234	---	211.1077
PTI	312.1918	284.1969	---	294.1812
EPT	328.1503	300.1554	---	310.1397
TID	330.1660	302.1710	---	312.1554
PTID	427.2187	399.2238	---	409.2082
EPTI	441.2344	413.2395	---	423.2238
EPTID	556.2613	528.2664	---	538.2508

Theoretical Peak Table

70.0651	P	199.1077	EP-28	294.1812	PTI-H_2O	397.2082	a_4	528.2664	EPTID-28
74.0600	T	199.1077	a_2	296.1605	a_3	399.2238	PTID-28	538.2508	EPTID-H_2O
86.0964	I	201.1234	ID-28	300.1554	EPT-28	407.1925	b_4-H_2O	538.2871	b_5
88.0393	D	209.0921	b_2-H_2O	302.1710	TID-28	409.2082	PTID-H_2O	556.2613	EPTID
102.0550	E	209.0921	EP-H_2O	306.1448	b_3-H_2O	413.2395	EPTI-28	556.2613	y_5-H_2O
126.0550	P	211.1077	ID-H_2O	310.1397	EPT-H_2O	423.2238	EPTI-H_2O	574.2719	y_5
130.0499	y_1-H_2O	215.1390	TI	312.1554	TID-H_2O	425.2031	b_4	607.3086	a_6-H_2O
148.0604	y_1	227.1026	EP	312.1918	PTI	427.2187	PTID	625.3192	a_6
171.1128	PT-28	227.1026	b_2	324.1554	b_3	441.2344	EPTI	635.3035	b_6-H_2O
181.0972	PT-H_2O	229.1183	ID	328.1503	EPT	459.2086	y_4-H_2O	653.3141	b_6
181.0972	a_2-H_2O	245.0768	y_2-H_2O	330.1660	TID	477.2191	y_4	685.3039	y_6-H_2O
187.1441	TI-28	263.0874	y_2	358.1609	y_3-H_2O	492.2817	a_5-H_2O	703.3145	y_6
197.1285	TI-H_2O	278.1499	a_3-H_2O	376.1714	y_3	510.2922	a_5	782.3567	MH-H_2O
199.1077	PT	284.1969	PTI-28	379.1976	a_4-H_2O	520.2766	b_5-H_2O	800.3672	MH

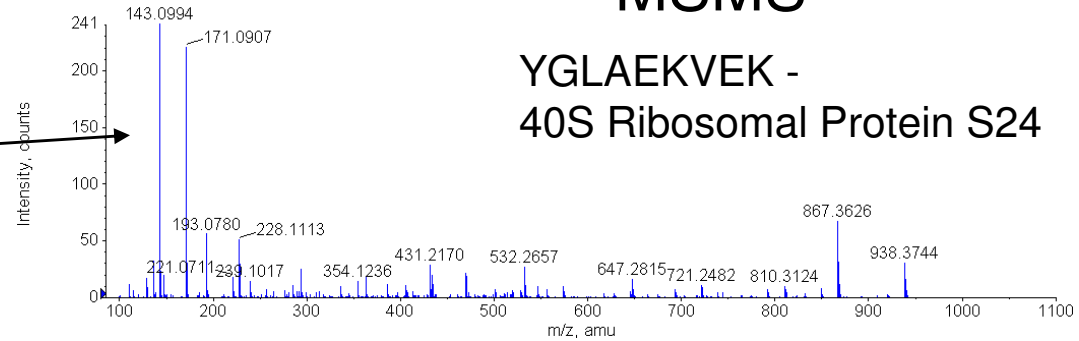
MSMS Allows Analysis of Complex Mixtures

MS

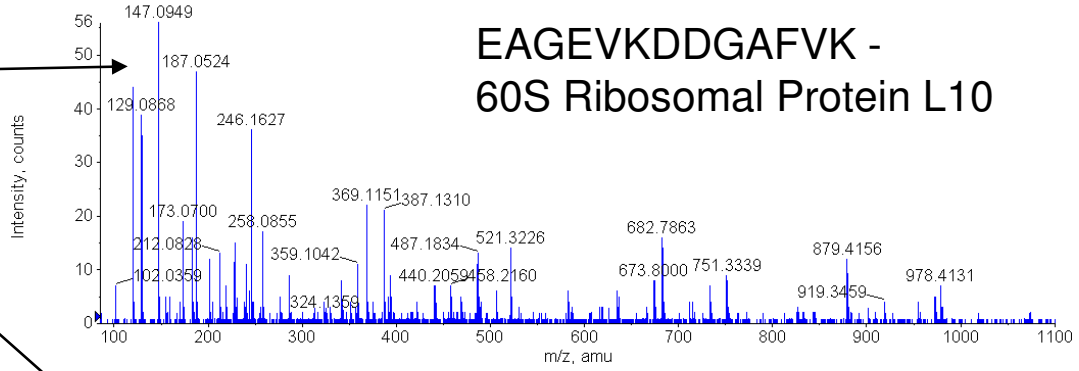


MSMS

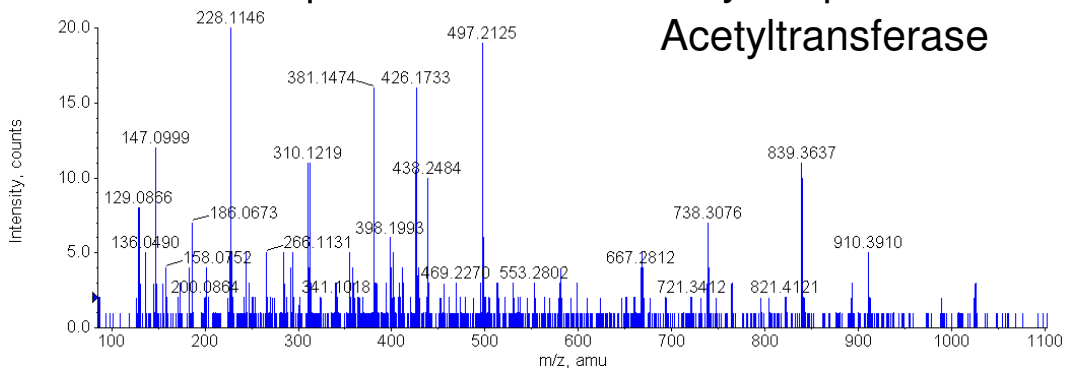
YGLAEKVEK -
40S Ribosomal Protein S24



EAGEVKDDGAFVK -
60S Ribosomal Protein L10



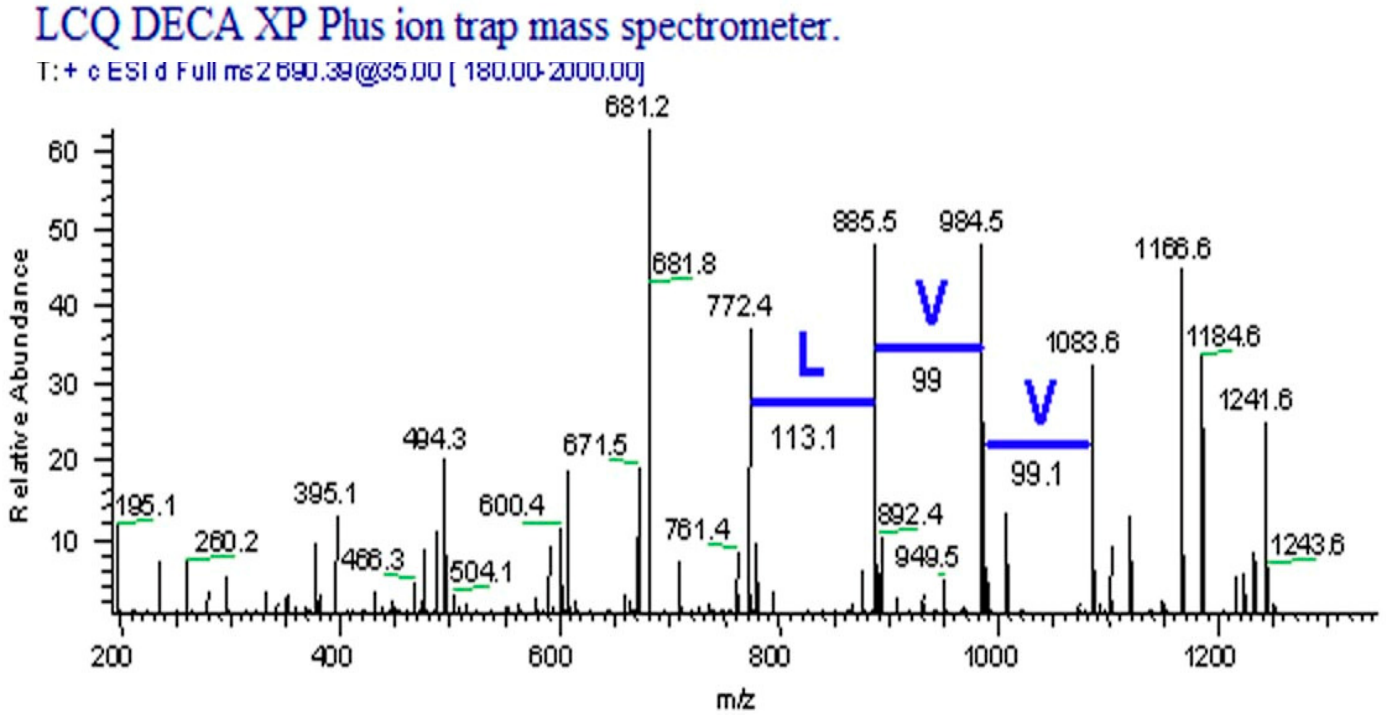
qSLNATANDKYK - Dihydrolipoamide Acetyltransferase



NOTE: It is highly improbable that PMF would correctly identify any of these proteins.

MS/MS Sequence Tags

- It is difficult to determine a complete and unambiguous peptide sequence from an MS/MS spectrum, but a series of peaks providing several adjacent residues can often be identified.
- This approach was pioneered by Mann and co-workers at EMBL [Mann, 1994].
- They defined a sequence tag derived from an MS/MS spectrum as the mass of the precursor peptide, the mass of the first peak of the identified sequence ladder, a stretch of interpreted sequence, and the mass of the final peak of the ladder.

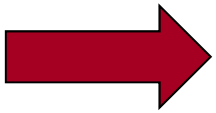


Note: Depending on whether the identified peaks are b- or y-ions, this sequence might be read in either direction, LVV or VVL.

Database Searching of MSMS Data

Input precursor ion m/z and charge, plus list of all fragment ions

```
PEPMASS=428.764297517301
CHARGE=2+
TITLE=Elution from: 41.95 to 42.23
59.038 6
60.041 13
61.034 9
63 4
70.059 10
71.074 24
72.075 59
72.153 2
73.028 2
74.056 8
75.045 6
85.018 4
85.088 2
86.092 110
86.153 6
87.098 11
89.078 2
92.009 8
93.061 2
95.053 8
96.077 3
97.069 11
98.088 15
98.979 42
99.044 11
99.111 2
99.176 2
100.061 7
100.995 5
101.082 13
101.993 4
102.085 6
103.045 6
```



Search engine de-isotopes mass list and filters out ‘n’ most intense peaks for searching

Compare peak list observed with theoretical fragmentation peak list produced for all peptides with the molecular weight observed for the parent ion

MSMS Database Search Engines

- There are many commercial and freely available search engines.
- Different instrument vendors promote their own tools.
- Some tools are open-source. In most cases access to an internet version is free. More advanced versions require a site license.
- In all cases the data is input and searched in a similar fashion.
- Different programs have different 'scoring systems' for deciding which matches are correct.

Available search programs: Protein Prospector (MS Tag); Mascot; Sequest; OMSSA; Xtandem; etc.

MSMS Search Parameters

As with PMF, efficient and accurate database searching of MS/MS data is best achieved if the operator makes intelligent use of all available knowledge.

- Protein Database.
- Enzyme used.
- Mass accuracy of precursor ion.
- Mass accuracy of fragment ions.
- Fragment ion types to look for – specify instrument type.
- What types of peptide modifications should be allowed for?

How do you determine a good peptide match? Scoring Systems

Count number of peaks matched? This is insensitive as:

- Certain ion types are more likely to be observed than others.
- In low energy CID ‘b’ and ‘y’ ions are going to be common.
- For tryptic peptides ‘y’ ions are more common (due to basic C-terminal residue).
- CID in quadrupole produces internal ions, in an ion-trap they are not formed.
- Certain ion types are more diagnostic than others.
- Immonium ions identify an amino acid but no sequence.
- ‘b’ and ‘y’ ions more specific than internal ions.

Practical approach:

- Depending on instrument type, look for different sets of ions.
- Give different scores for different ion types observed (more for ‘y’ ions, less for internal ions)

MS-Tag Search Result

Result Summary

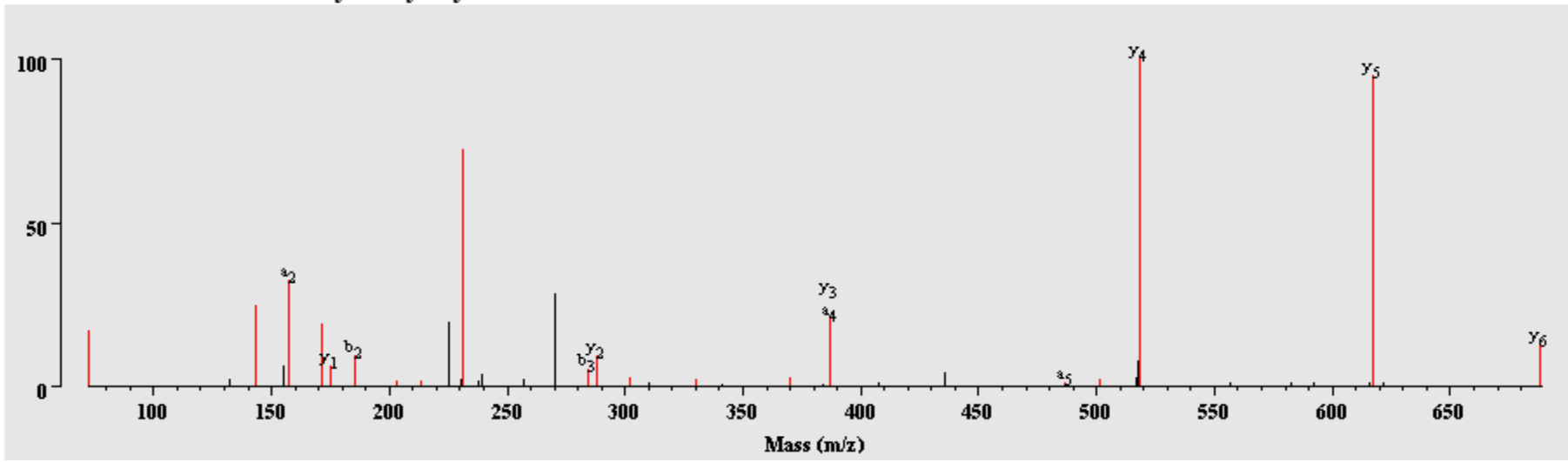
Rank	# Unmatched Ions	Sequence	Score	m/z Submitted	MH^+ Calculated (Da)	Error (ppm)	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
1	20	(R)IAVMVIR (W)	26.5	401.2744^{+2}	801.5015	50.0	16576	118069/5.2	P81650	PSEHA	Beta-galactosidase (EC 3.2.1.23) (Lactase) (Beta-D-galactoside galactohydrolase)
1	20	(R)LAVMVLR (W)	26.5	401.2744^{+2}	801.5015	50.0	16559	116353/5.3	P00722	ECOLI	Beta-galactosidase (EC 3.2.1.23) (Lactase)
1	20	(R)LAVMVLR (W)	26.5	401.2744^{+2}	801.5015	50.0	16560	116679/5.8	Q47077	ENTCL	Beta-galactosidase (EC 3.2.1.23) (Lactase)
2	21	(R)LAVTELR (G)	21.6	401.2744^{+2}	801.4829	73.2	123666	104062/6.0	Q13608	HUMAN	Peroxisome assembly factor 2 (PAF-2) (Peroxisomal-type ATPase 1) (Peroxin-6) (Peroxisomal biogenesis factor 6)
2	21	(R)LAVTELR (G)	21.6	401.2744^{+2}	801.4829	73.2	123668	104549/7.0	Q99LC9	MOUSE	Peroxisome assembly factor 2 (PAF-2) (Peroxisomal-type ATPase 1) (Peroxin-6) (Peroxisomal biogenesis factor 6)

Note: MV (131+99 = 230) and TE (101+129 = 230) can only be distinguished if fragmentation occurs between them, i.e. look for y_3 or b_4.

Is the top match significantly better than random?

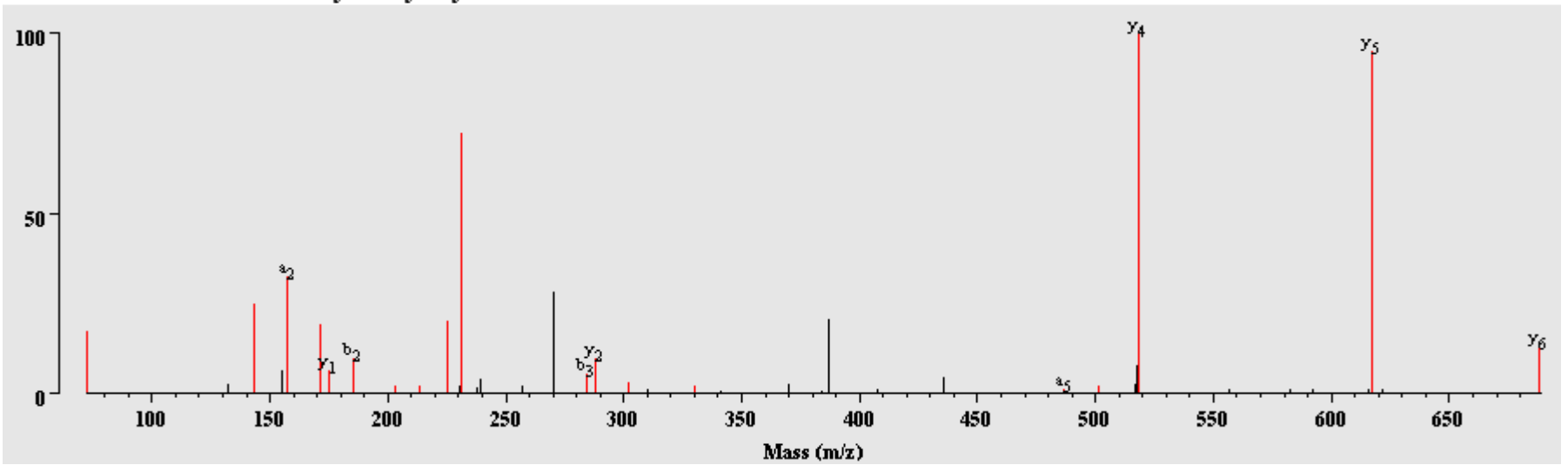
LAVMVLR^{+2}

Constant Modification: **Carboxymethyl Cysteine**



LAVTELR^{+2}

Constant Modification: **Carboxymethyl Cysteine**



How do you determine a good peptide match? Is the top match correct?

- You have a score for all peptides in the database that have the same precursor mass as your spectrum.
- You have a top scoring match.

How do you decide whether this top scoring match is correct?

Calculate a probability that it is correct?

Very difficult to do.

Calculate a probability that it is incorrect?

Easier.

Most search engines now report an Expectation value.

Expectation Values

- The expectation value is a prediction of the number of times an event is expected to happen at random.

- For a peptide result the expectation value is the number of times the given score (or greater) will be achieved by random (incorrect) matches.

- Expectation value of a score = probability of score x number of peptides in the database having the same precursor mass

e.g. If the probability of a random match scoring ‘20’ is 1e-5, but there are 1000 peptides in the database with the same precursor mass, then the expectation value is (1e-5 x 1000 =) 1e-2; i.e. there is a 1% chance that the score of 20 is a random (incorrect) match.

Calculation of Expectation Values

Theoretical Calculation (Mascot): What is the probability of 10 out of 25 peaks matching a random (incorrect) assignment?

- Assumes theoretical model takes into account all variables that can change the number of peaks matching at random.
- Assumes sequences in database are random.

Calculation based on results (Protein Prospector): Model scores of the incorrect answers to a distribution and extrapolates the probability of a given score being part of this distribution.

- More flexible / applicable to more scoring systems
- Model incorporates non-random nature of protein sequences
- Reliant on having enough data points to accurately model the distribution

From Peptide ID's to Protein ID

- Other peptides from the same protein may be identified in the same experiment.
- If the identified protein is actually in the sample, it is more likely that other peptides from the same protein will be found.

1 **Acc. #:** P00722 **Gene:** BGAL_ECOLI **Species:** ECOLI **Name:** Beta-galactosidase (EC 3.2.1.23) (Lactase)

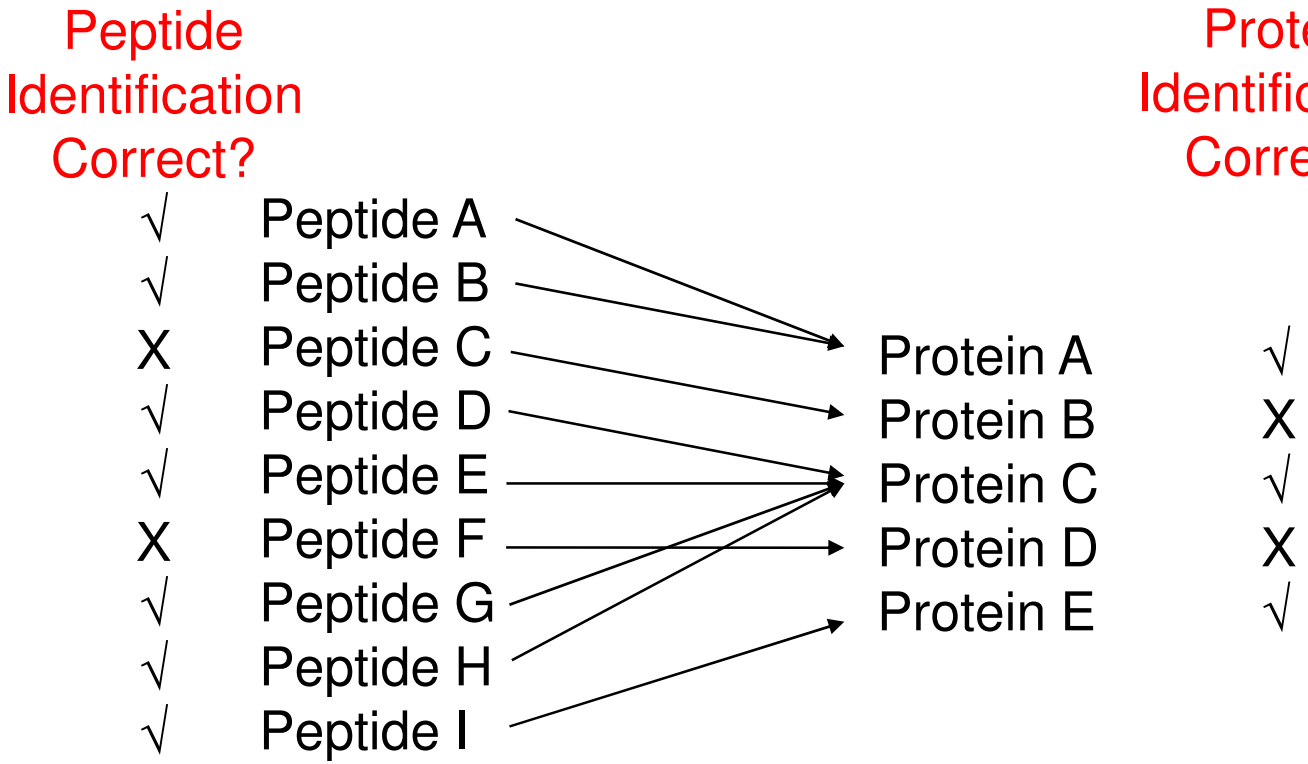
Protein MW: 116352.7 **Protein pI:** 5.3

Num Unique	% Cov	Best Disc Score	Best Expect Val
11	11.1	3.18	4.8e-7

m/z	z	ppm	Peptide	S	Score	Expect	# in DB
729.3964	2	43	APLDNDIGVSEATR	27.3	35.7	4.8e-7	1
567.0809	4	46	DVSLLHKPTTQISDFHVATR	35.51	26.8	4.6e-5	1
681.3903	2	38	LWSAEIPNLYR	35.28	28.1	1.7e-4	1
736.9074	2	37	IGLNCQLAQVAER	31.98	26.6	3.7e-4	1
503.2559	3	38	YSQQQLMETSHR	25.56	24.3	5.4e-4	1
542.2814	2	31	GDFQFNISR	31.3	27.3	0.0050	1
401.2744	2	50	LAVMVLR	31.04	26.5	0.010	3
450.7146	2	41	FNDDFSR	26.1	25.4	0.015	1
355.6959	2	27	MSGIFR	27.98	18.8	0.024	9
477.7467	2	54	LTAACFDR	27.1	19.1	0.029	1
407.2368	2	24	LNVENPK	22.04	22.6	0.032	1

Peptide Errors Are Amplified in Protein ID's

- Peptides correctly identified are more likely to be from proteins from which other peptides have been observed.
- Incorrect peptide identifications almost always represent the sole identification of a particular protein.



7 of 9 peptides correct (78%)

only 3 of 5 proteins correct (60%)

Best strategy

The conversion of peptide to protein information is also complicated by:

- multiple database entries for the same protein.
- sequence variants / isoforms.
- splice variants.

It is best to combine multiple parameters from a search result to create a new score that is better at discriminating between correct and incorrect answers than any one parameter from the search result.

This can be used to assign a new measure of reliability to a result.

- Protein Prospector reports a discriminant score.
- PeptideProphet / ProteinProphet (free open source software) can be used to re-analyze other search engine results.

Peptide to Protein - Mascot

- Combine peptide scores together to calculate a protein score.
- Only report matches to proteins above a certain score threshold.
- Report all peptide matches to these proteins.

4. BGAL_ECOLI **Mass:** 116278 **Score:** 316 **Peptides matched:** 12

P00722|BGAL_ECOLI Beta-galactosidase (EC 3.2.1.23) (Lactase) - Escherichia coli

☐ Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑ 44	355.70	709.38	709.36	0.02	0	22	2.5	1	MSGIFR
58	368.74	735.46	735.43	0.03	0	20	2.4	4	TLFISR
☑ 76	401.27	800.53	800.49	0.04	0	27	0.83	1	LAVMVLR
☑ 82	407.24	812.46	812.44	0.02	0	18	7.8	1	LNVENPK
☑ 107	450.71	899.41	899.38	0.04	0	27	0.63	1	FNDDFSR
151	477.75	953.48	953.43	0.05	0	19	6.3	2	LTAACFDR + Carboxymethyl (C)
☑ 279	542.28	1082.55	1082.51	0.03	0	36	0.12	1	GDFQFNISR
☑ 361	671.36	1340.71	1340.66	0.05	0	13	32	1	VDEDQPFPAVPK
☑ 368	681.39	1360.77	1360.71	0.05	0	37	0.13	1	LWSAEIPNLYR
☑ 403	729.40	1456.78	1456.72	0.06	0	42	0.042	1	APLDNDIGVSEATR
☑ 409	736.91	1471.80	1471.75	0.06	0	44	0.026	1	IGLNCQLAQVAER + Carboxymethyl (C)
☑ 480	567.08	2264.29	2264.19	0.10	0	13	24	1	DVSLLHKPTTQISDFHVATR

Real example: Why Many Spectra are not Identified

Careful analysis of 3269 spectra yielded 904 that could not be identified.

22 peptides too short to be confident of assignment (m/z <620)
43 from mixtures of precursor ions
24 spectra of methylated trypsin
24 Deamidation of N
4 peptides sequences not in the database
226 spectra not of a peptide (ICAT, PEG …)
48 peptides products of non-specific enzyme cleavages
312 spectra not good enough to assign
1 spectrum with a methylated lysine
82 assigned the wrong charge
1 wrong charge and mixture
2 wrong charge – not peptide
78 wrong isotope selected
14 wrong charge and monoisotopic peak
3 wrong isotope and mixture
11 MSMS of peptides that lost water in-source
8 peptides formed from in-source fragmentation of abundant co-eluting peak
1 peptide containing an internal disulfide bond

Chalkley, R. J. et al. *Mol Cell Proteomics* (2005) **4** (8) p.1189-1193

Homology-based searching – Brief introduction

- If your protein is not in the database, how do you identify it?
- It may be highly homologous either to another protein, or to the same protein from a different species
- *De Novo* Sequencing, then BLAST or MS-Homology
  - Searching allowing for amino acid substitutions

    [213]ENFAGVGV[I|L]DFES 6
    [217]GA[Q|K][242]DENTR 4

  - Scoring system based on likelihood of amino acid substitution
    - Ser to Thr: similar amino acids
    - Gly to Arg: very different amino acids

Summary of Protein Identification and Characterization

Peptide Mass Fingerprinting (PMF)

- Protein is digested into peptides; MWs are measured on MS.
- Peptide MWs are searched against a database.
- Works for simple mixtures and the whole experiment is simple and fast.

Protein Identification Based on Peptide MSMS

- One or two peptide ID's by MSMS can give protein ID.
- Works with complicated mixtures.
- Typically the data are acquired by LCMSMS.
- Desirable with HighRes on precursor ions or survey scans.
- HighRes on MSMS fragment ions is less critical.
- May provide PTM site assignment.

BUT: Search engines make mistakes

- Appropriate choice of search engine parameters is important.
- Use probability/expectation values to measure assignment reliability.
- Use of random/concatenated database searching can estimate false positive rates for the dataset as a whole.