PC235 2009 Lecture 3: Protein Identification using Mass Spectrometry

Robert Chalkley
E-mail: chalkley@cgl.ucsf.edu

Outline

- Peptide Mass Fingerprinting
- Peptide Fragmentation Mechanisms - CID
- Fragmentation Analysis: ‘Protein Sequencing’
- Database Searching and Scoring

Peptide Mass Fingerprinting (PMF)

- Acquire mass spectrum of protein digest (peptides)
- Input list of observed masses into database search program
- Search program creates theoretical enzyme digest of all proteins in database, and compares the mass list you observed to theoretical mass lists for all proteins, and returns ‘best matches’.

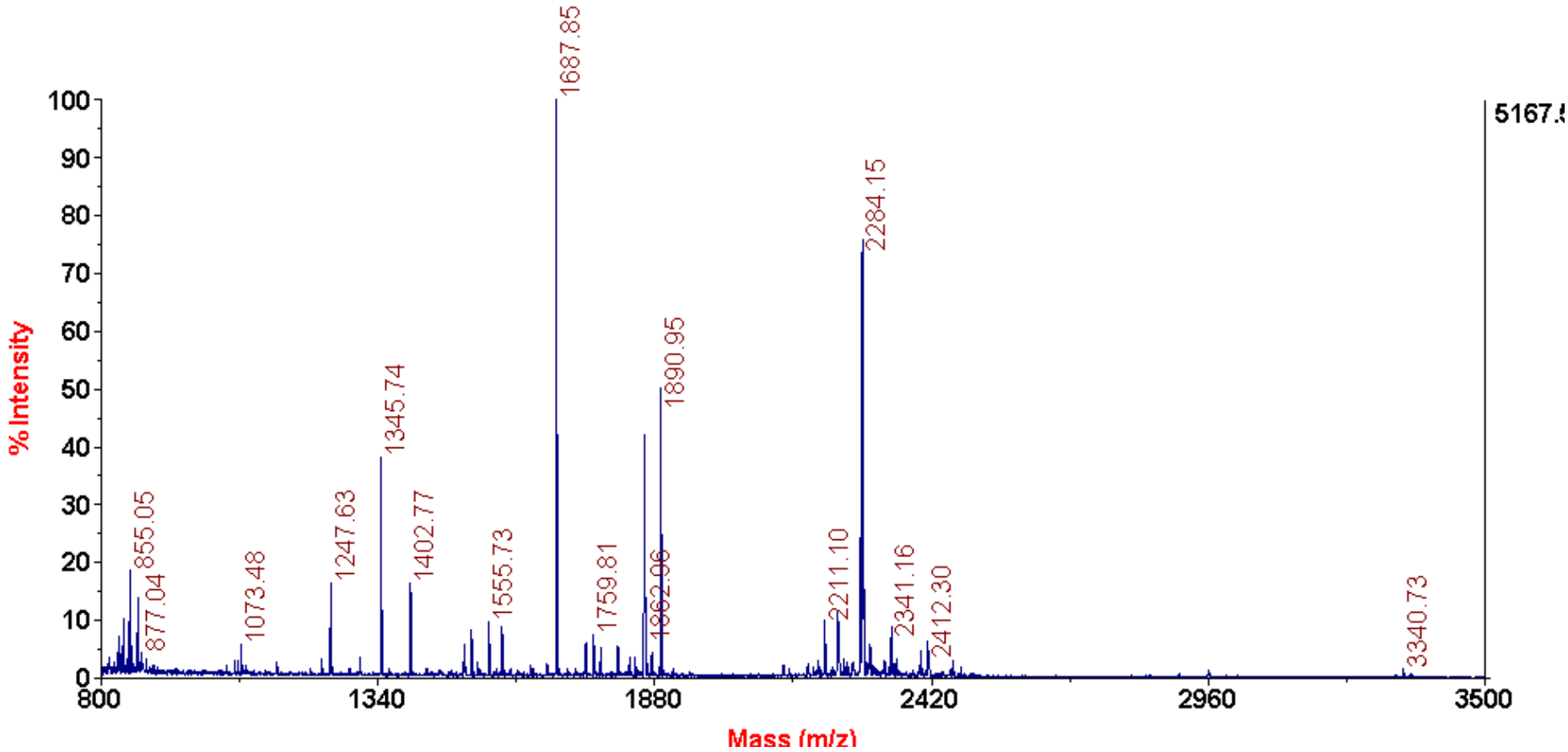
Advantages:

- Quick and simple to acquire data.
- Sensitive.

Disadvantages:

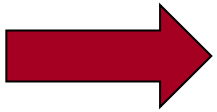
- Not good for protein mixture analysis (even a simple mixture).
- Confidence of many search result assignments is low.

MALDI Mass Spectrum of a Tryptic Digest



Peak List is Generated and Searched

Monoisotopic Mass
771.478027
833.069885
842.510010
855.051453
861.066223
871.022034
877.037292
1073.484009
1247.629395
1304.651123
1345.741699
1402.770264
1507.718140
1522.794678
1555.726318
1581.734375
1687.847168
1744.864990
1759.811157
1773.904297
1807.806519
1830.928589
1841.965332
1858.968140
1862.963867
1873.940552
1890.951782
2211.104736



Database Search Program

Mass Fingerprinting Database Search Engines

Protein Prospector

- •Developed at UCSF. Suite of tools for all kinds of proteomic analysis, including protein mass fingerprinting, MSMS analysis, theoretical protein digestion, peptide fragmentation tools…
  - •Mass Fingerprint analysis software is called ‘MS-Fit’.

Mascot

- •Search engine for analyzing protein mass fingerprinting data and LC-MSMS data. Limited version available for free over internet; more advanced version requires site license.

•For both, data is input and searched in a similar fashion, but they have different ‘scoring systems’ for deciding which matches are correct.

PMF Search Parameters

MASCOT Peptide Mass Fingerprint

Your name		**Email**	
Search title			
Database	MSDB		
Taxonomy	All entries		
Enzyme	Trypsin	**Allow up to**	1 missed cleavages
Fixed modifications	Acetyl (K) Acetyl (N-term) Amide (C-term) Biotin (K) Biotin (N-term)	**Variable modifications**	Acetyl (K) Acetyl (N-term) Amide (C-term) Biotin (K) Biotin (N-term)
Protein mass	kDa	**Peptide tol. ±**	1.0 Da
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.			
Overview		**Report top**	20 hits
	Start Search ...		**Reset Form**

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.

*Database accessions disappear!

Effect of Mass Accuracy and Number of Peaks Required to 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

Peptide Modifications Commonly Observed

Modification	ΔMass
Carbamidomethylation of Cys	+57Da
Oxidation of Met	+16Da
Pyroglutamate formation	-17Da
Deamidation of Asn (or Gln)	+1Da
Acetylation	+42Da
Phosphorylation	+80Da
Sulfation	+80Da
Methylation	+14Da
Glycosylation	+??
Sodium Adduct	+22Da
Potassium Adduct	+38Da

What Modifications Should You Search For?

- Only ones that are common / you are expecting.

- Allowing for modifications can increase dramatically the number of potential peptides masses in your database; e.g.
  - Single peptide: GSIGASMER
  - If you allow for phosphorylation becomes 4 potential peptides:
  - GSIGASMER; GS(phos)IGASMER; GSIGAS(phos)MER; GS(phos)IGAS(phos)MER
    - Database is now 4x bigger!

- Acetylation of N-terminus of protein: Adds one peptide per protein.
- Fixed modifications; e.g. carbamidomethyl cysteine, do not increase database size.
- Modifications I generally allow for:
  - Fixed modification: carbamidomethyl cysteine
  - Variable modifications: N-Acetyl (protein); oxidised Met; pyroGlu (from Q)

How are PMF Results Scored / Results Ranked?

- What protein matches the most 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.

MS-Fit Search Result

Allowing up to 100ppm mass accuracy

MS-Fit search selects **2809** entries (results displayed for top **5** matches).

Results Summary

Protein Hit Number	MOWSE Score	# pep # mat % mat 34 pks	% Cov	% TIC	Mean Err ppm	Data Tol ppm	# Hom Prot	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
☑ 1	1.285e+007	13/12/35	44.3	35.3	1.70	8.29	No	133728	42882/5.2	P01012	CHICK	Ovalbumin (Plakalbumin)
☑ 2	4.995e+004	8/8/24	11.8	23.5	-3.43	129	No	21875	113232/5.4	Q09349	CAEEL	Probable ubiquitin conjuga
☑ 3	2.951e+004	8/7/21	8.2	20.6	27.3	76.7	No	9965	149835/8.3	P11675	PRVIF	IMMEDIATE-EARLY P
☑ 4	2.106e+004	10/9/26	18.2	26.5	13.1	65.2	No	121421	61403/9.4	P12175	ORYSA	Maturase K (Intron matur
☑ 5	1.100e+004	5/5/15	15.5	14.7	25.4	58.2	No	80260	48165/5.5	P95178	MYCTU	NADH-quinone oxidoredu chain D)

Peptide Match Summary: Top Match

1. 12/34 matches (35%).
Acc. #: P01012 **Species:** CHICK **Name:** Ovalbumin (Plakalbumin) (Allergen Gal d 2) (Gal d II)
Index: 133728 **MW:** 42882 Da **pI:** 5.2

m/z Submitted	MH^+ Matched	Intensity	Delta ppm	Modifications	Start	End	Missed Cleavages	Sequence
1247.6294	1247.6247	100.0	3.8		361	370	0	(R)ADHPFLFCIK (H)
1345.7417	1345.7381	100.0	2.7		371	382	0	(K)HIATNAVLFFGR (C)
1522.7947	1522.7980	100.0	-2.2		112	123	0	(R)YPILPEYLQCVK (E)
1555.7263	1555.7215	100.0	3.1		188	200	1	(K)AFKDEDTQAMPFR (V)
1581.7344	1581.7219	100.0	7.9		265	277	0	(K)LTEWTSSNVMEER (K)
1687.8472	1687.8404	100.0	4.0		128	143	0	(R)GGLEPINFQTAADQAR (E)
1773.9043	1773.8996	100.0	2.6		324	340	0	(K)ISQAVHAAHAEINEAGR (E)
1807.8065	1807.8035	100.0	1.7	AcetN	1	17	0	(-)GSIGAASMEFCFDVFK(E)
1858.9681	1858.9663	100.0	0.99		144	159	0	(R)ELINSWVESQTNGIIR (N)
2281.1736	2281.1828	100.0	-4.1		86	105	0	(R)DILNQITKPNDVYSFSLASR (L)
2284.1521	2284.1470	100.0	2.2		201	219	0	(R)VTEQESKPVQMMYQIGLFR (V)
2284.1521	2284.1688	100.0	-7.3		106	123	1	(R)LYAEERYPILPEYLQCVK (E)
2300.1572	2300.1419	100.0	6.7	1Met-ox	201	219	0	(R)VTEQESKPVQMMYQIGLFR(V)

Click link below to search for cysteine linked fragments.
22 unmatched masses:

Click link below to do a non-specific cleavage search.
22 unmatched masses:

Click link below to search for another component.
22 unmatched masses:

The matched peptides cover **44.3%** (171/386AA's) of the protein.
Coverage Map for This Hit (MS-Digest index #): 133728

Peptide Match Summary: Second Match

2. 8/34 matches (23%).
Acc. #: Q09349 **Species:** CAEEL **Name:** Probable ubiquitin conjugation factor E4
Index: 21875 **MW:** 113232 Da **pI:** 5.4

m/z Submitted	MH^+ Matched	Intensity	Delta ppm	Modifications	Start	End	Missed Cleavages	Sequence
771.4780	771.4365	100.0	54		788	794	0	(R)TPVLGER (L)
1402.7703	1402.7694	100.0	0.59		351	362	0	(R)WIATIISTNDIR (T)
1759.8112	1759.9417	100.0	-74		454	468	0	(R)LVIPPLMNQISEYSR (H)
1841.9653	1841.7969	100.0	91		897	911	0	(K)AELEEEYDDVPEEFK (D)
2284.1521	2284.2746	100.0	-54		602	620	1	(R)LTVLFTQYHYIKSPFLVSK (L)
2300.1572	2300.2331	100.0	-33		821	839	1	(R)SYGWEPREFVSLLISIYLK (L)
2341.1597	2341.3245	100.0	-70	1Met-ox	828	847	1	(R)EFVSLLISIYLKLNMPAFVK(Y)
2412.2974	2412.1579	100.0	58		704	724	0	(R)FVNMVINDATWCIDESLSGLK (S)

Click link below to search for cysteine linked fragments.
26 unmatched masses:

Click link below to do a non-specific cleavage search.
26 unmatched masses:

Click link below to search for another component.
26 unmatched masses:

The matched peptides cover **11.8**% (116/980AA's) of the protein.
Coverage Map for This Hit (MS-Digest index #): 21875

MS-Fit Search Result

Allowing up to 100ppm mass accuracy

MS-Fit search selects **2809** entries (results displayed for top **5** matches).

Results Summary

Protein Hit Number	MOWSE Score	# pep # mat % mat 34 pks	% Cov	% TIC	Mean Err ppm	Data Tol ppm	# Hom Prot	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
☑ 1	1.285e+007	13/12/35	44.3	35.3	1.70	8.29	No	133728	42882/5.2	P01012	CHICK	Ovalbumin (Plakalbumin)
☑ 2	4.995e+004	8/8/24	11.8	23.5	-3.43	129	No	21875	113232/5.4	Q09349	CAEEL	Probable ubiquitin conjug
☑ 3	2.951e+004	8/7/21	8.2	20.6	27.3	76.7	No	9965	149835/8.3	P11675	PRVIF	IMMEDIATE-EARLY P
☑ 4	2.106e+004	10/9/26	18.2	26.5	13.1	65.2	No	121421	61403/9.4	P12175	ORYSA	Maturase K (Intron matu
☑ 5	1.100e+004	5/5/15	15.5	14.7	25.4	58.2	No	80260	48165/5.5	P95178	MYCTU	NADH-quinone oxidored chain D)

133728	.	.	.	.	x	.	x	.	.	x	x	x	x	.	.	x	x	.	.	x	.	.	.	.	.	.	x	x	x	.	.	.	.
21875	.	x	.	.	.	.	.	x	.	.	.	.	.	.	x	.	.	.	x	.	.	.	.	.	.	.	.	x	x	x	.	x	.	.
9965	.	.	.	.	x	.	.	.	x	.	.	.	.	x	.	.	.	x	x	.	.	x	.	.	.	.	.	.	.	.	.	.	x	.
121421	x	.	.	.	x	.	.	.	.	.	.	.	.	x	.	x	x	.	x	.	.	x	.	.	.	.	.	x	x	.	.	.	.	.
80260	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	x	x	.	.	.	.	.	.	.	.	.	x	x	.	.	.	.	.

MS-Fit Search Result

Allowing up to 20ppm mass accuracy

MS-Fit search selects **107** entries (results displayed for top **5** matches).

Results Summary

Protein Hit Number	MOWSE Score	# pep # mat % mat 34 pks	% Cov	% TIC	Mean Err ppm	Data Tol ppm	# Hom Prot	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
1	1.285e+007	13/12/35	44.3	35.3	1.70	8.29	No	133728	42882/5.2	P01012	CHICK	Ovalbumin (Plakalbumin) (Allergen
2	764	5/5/15	7.6	14.7	-4.96	4.51	No	72594	118004/8.3	P77839	BROTH	DNA-directed RNA polymerase be subunit)
3	503	4/4/12	11.0	11.8	-4.07	9.12	No	59765	54123/8.8	O00330	HUMAN	Pyruvate dehydrogenase protein X dehydrogenase-binding protein of p (proX)
4	315	4/4/12	6.8	11.8	8.70	19.7	No	14256	121883/8.8	Q9TM35	CYACA	DNA-directed RNA polymerase be
5	281	4/4/12	5.8	11.8	-0.476	27.4	No	113581	124885/6.5	Q93100	HUMAN	Phosphorylase B kinase beta regula

133728	.	.	.	.	x	.	x	.	.	x	x	x	x	.	.	x	x	.	.	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.
72594	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	.	x	.	x	x	.	.	x	.	.
59765	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	.	x	.	.	.	.	.	.	.	.	.	x	.
14256	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	.
113581	.	.	.	.	.	.	x	.	.	x	.	.	.	.	.	.	.	.	x	.	.	.	.	.	.	.	.	.	.	.	x	.	.	.

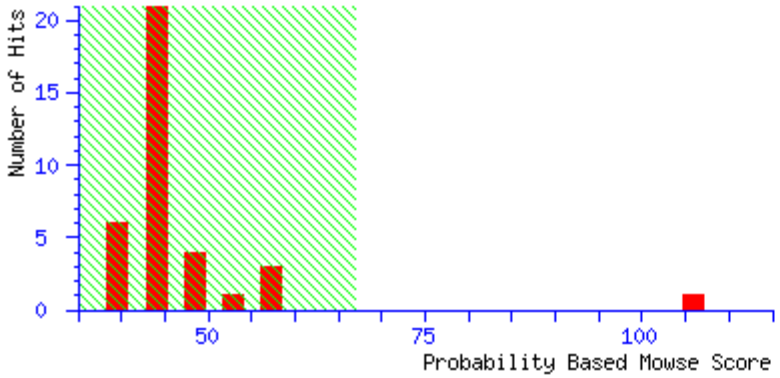
Mascot Search Results

100ppm

User : robert
Email : robertc@itsa.ucsf.edu
Search title :
Database : SwissProt 48.0 (250646 sequences; 113845621 residues)
Timestamp : 2 Oct 2005 at 06:37:17 GMT
Top Score : 106 for P01012-00-00-00, (OVAL_CHICK) Splice isoform Displayed; Variant Displayed; Conflic

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 67 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. P01012-00-00-00 **Mass:** 43065 **Score:** 106 **Expect:** 6.3e-06 **Queries matched:** 12
(OVAL_CHICK) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from P01012 Ovalbumin (P
P01012-00-01-00 **Mass:** 43066 **Score:** 106 **Expect:** 6.3e-06 **Queries matched:** 12

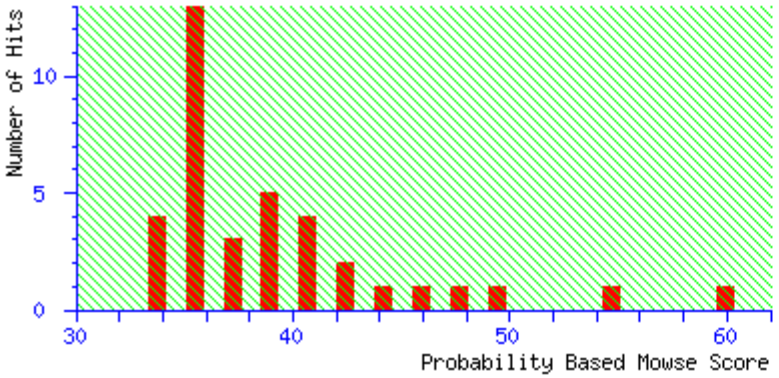
{MATRIX SCIENCE} Mascot Search Results

100ppm (allow for phosphorylation)

```
User            : robert
Email           : robertc@itsa.ucsf.edu
Search title    :
Database        : SwissProt 48.0 (250646 sequences; 113845621 residues)
Timestamp       : 2 Oct 2005 at 06:38:25 GMT
Top Score       : 60 for P01012-00-00-00, (OVAL_CHICK) Splice isoform Displayed; Variant Displayed; Conflict
```

Probability Based Mowse Score

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 67 are significant (p<0.05).



Concise Protein Summary Report

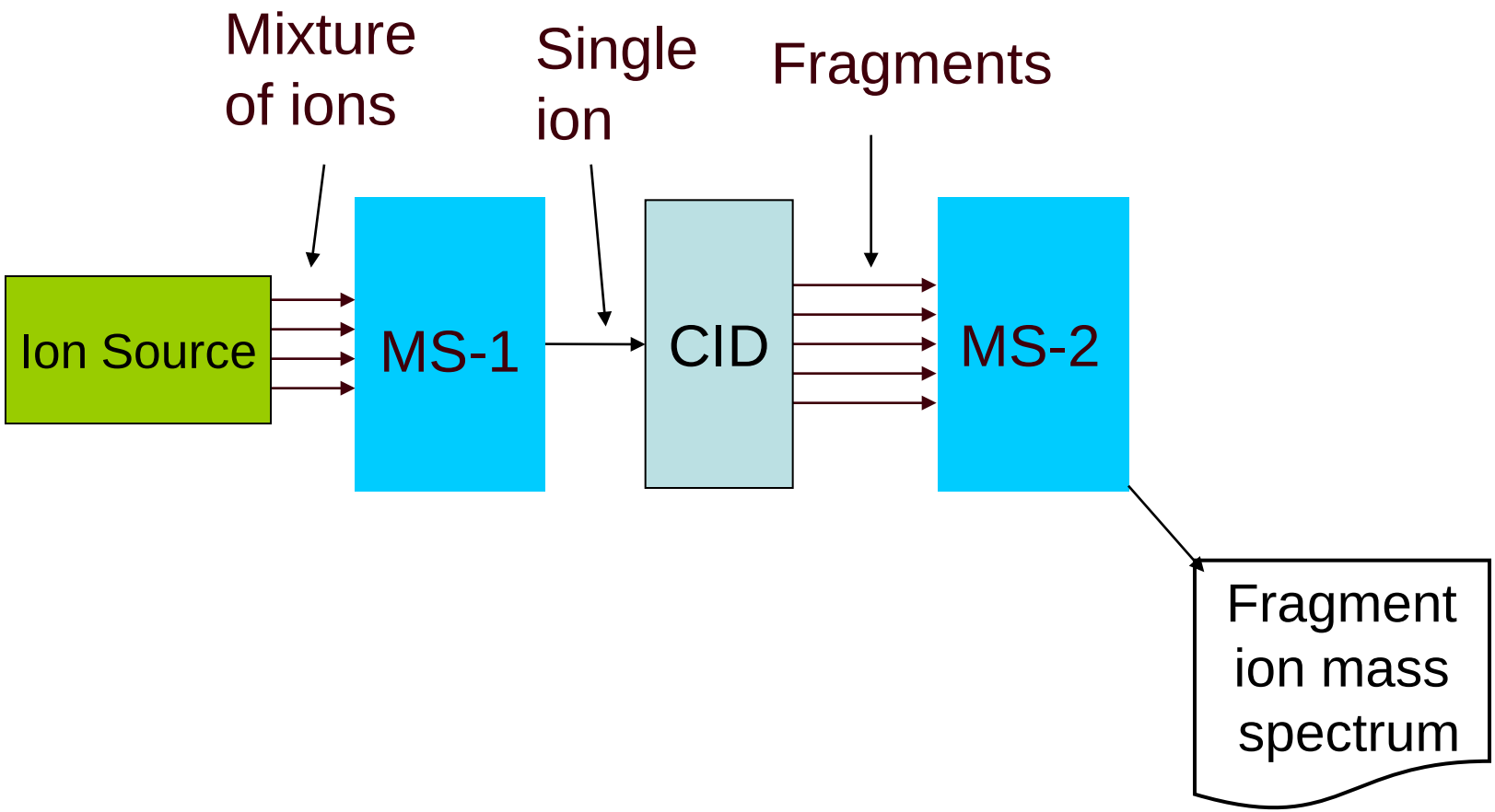
Format As: Concise Protein Summary — Help

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

Re-Search All | Search Unmatched

1. P01012-00-00-00 **Mass:** 43065 **Score:** 60 **Expect:** 0.26 **Queries matched:** 12
(OVAL_CHICK) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from P01012 Ovalbumin (P
P01012-00-01-00 **Mass:** 43066 **Score:** 60 **Expect:** 0.26 **Queries matched:** 12

MS/MS (Tandem Mass Spectrometry)



Advantages of MS/MS Analysis

- More specific and reliable than peptide mass fingerprinting
  - Search with intact peptide mass, and masses of fragment ions.
  - All fragment ions should be derived from precursor ion.
- Can be used for *de novo* sequencing; i.e. protein sequence does not need to be previously known.

- Allows identification of proteins on the basis of one or two peptides.
- Can identify proteins in complex mixtures.

Amino Acid Residue Masses

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

Different Methods of Fragmentation

- Thermal / energy based fragmentation
  - Put energy into molecule. Breaks weakest bonds.
    - Collision-Induced Dissociation (CID)
    - Infra-Red MultiPhoton Dissociation (IRMPD)

- Radical-based fragmentation
  - Introduce 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)

Different Flavors of CID

Vanilla, Strawberry, Mint Choc Chip…

- Low Energy CID
  - In an ion trap – generally get a single fragmentation event
  - In a quadrupole – may get multiple fragmentation events

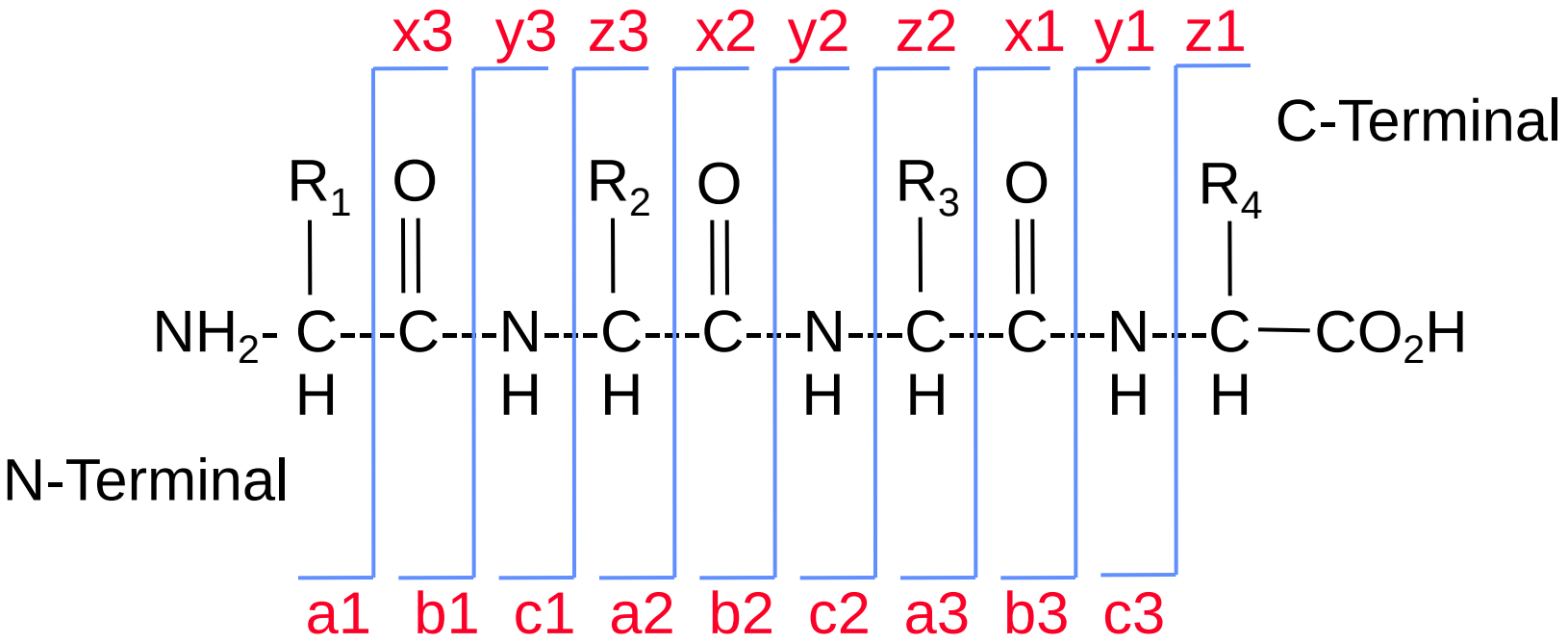
- High Energy CID
  - MALDI-CID / MALDI-TOFTOF – higher energy can allow formation of fragment types not observed in low energy CID. Can get multiple fragmentation events.

Ion Trap CID Fragmentation

- In ion trap, excitation (for CID) is m/z dependent.
- Once a molecule has fragmented (so its m/z has changed), it is no longer excited.
    - It is unusual to see fragments that are products of multiple fragmentation events.

- When exciting a component of a given m/z, fragments that are formed that are less than one third of the mass of the precursor ion cannot be trapped.
    - The bottom part of the CID mass spectrum from an ion trap will contain no fragment ions.

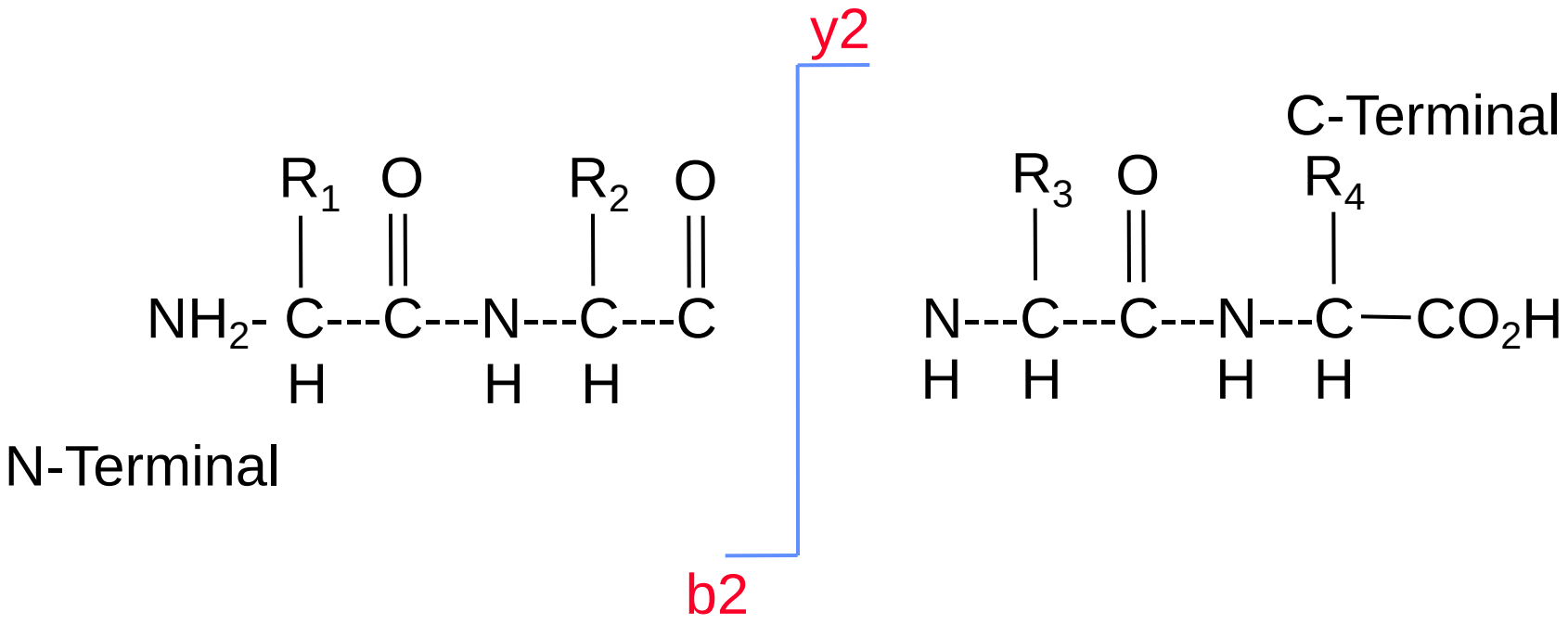
Peptide Fragment Ions



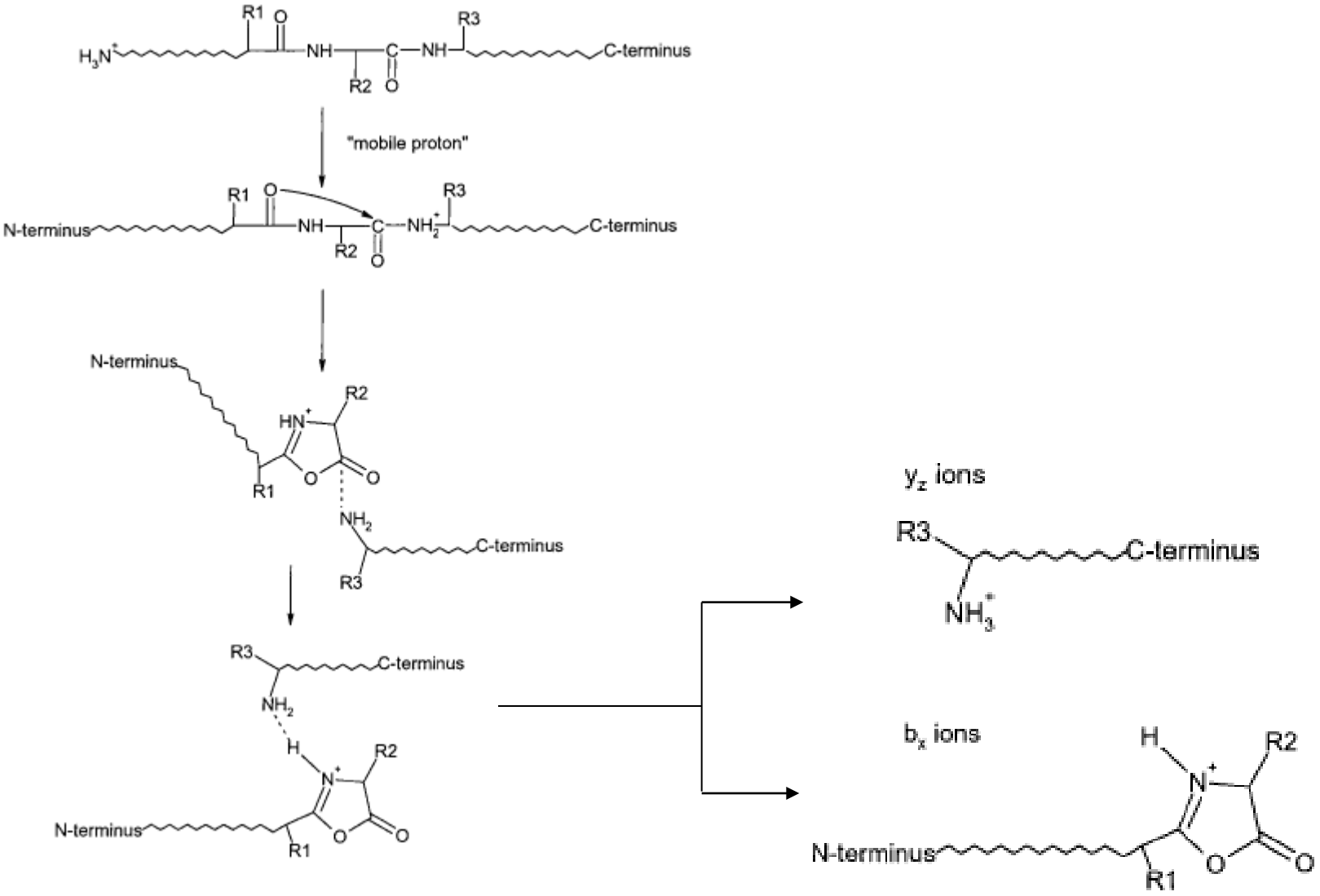
Biemann, K *Methods Enzymol* (1990) **193** 886-887

Peptide Fragment Ions

•Most common fragmentation in CID is peptide bond cleavage, forming b and y ions.



CID: b-y Fragmentation



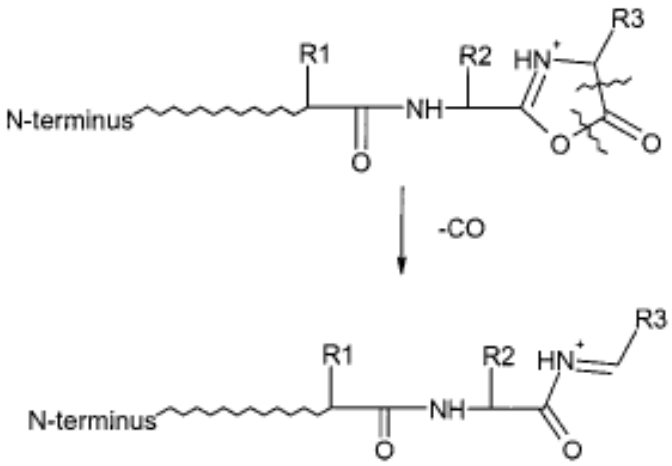
Paizs, B. and Suhai, S. **Mass Spectrom Rev.** (2005) **24** 4: 508-548

After b-y fragmentation…

•y ion is stable (identical to a normal peptide).

•b ion is relatively unstable, so readily undergoes further fragmentation:

- •$b_n \rightarrow a_n$
- •$b_n \rightarrow b_{n-1}$



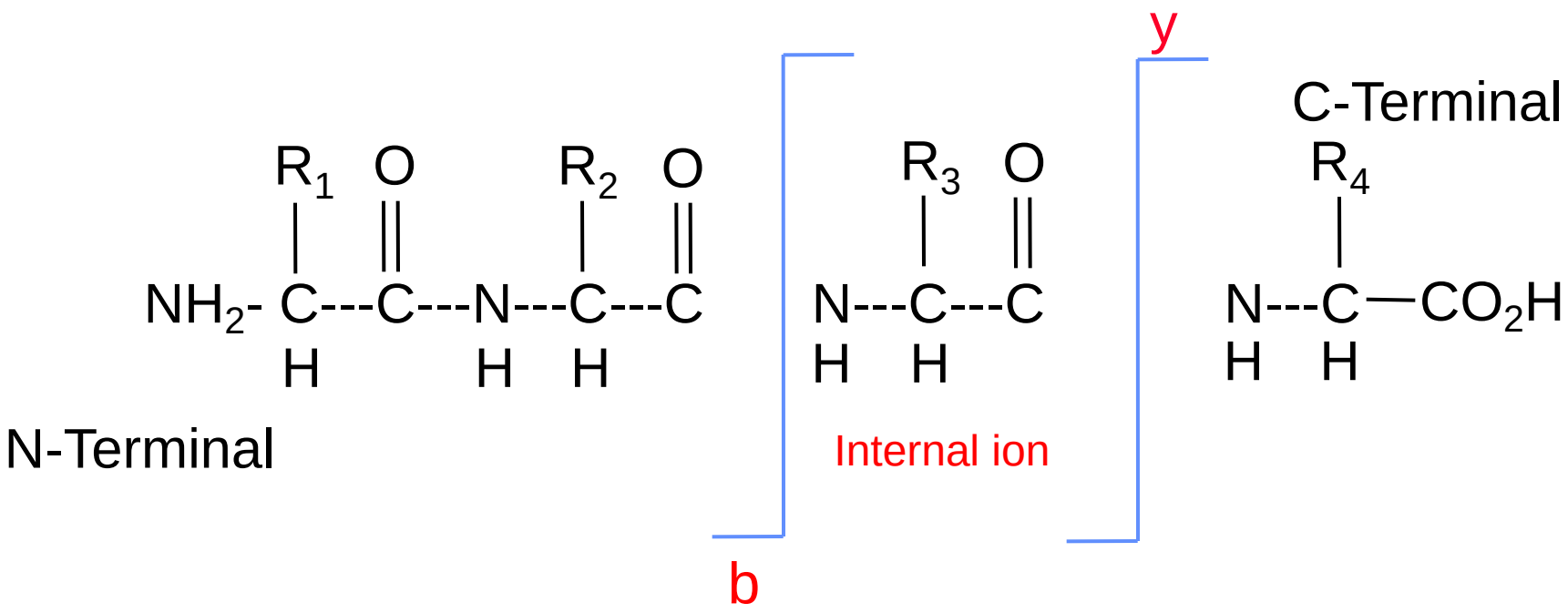
b ion

a ion

•Cannot form a b_1 ion. Therefore, b_2 is most stable b ion.

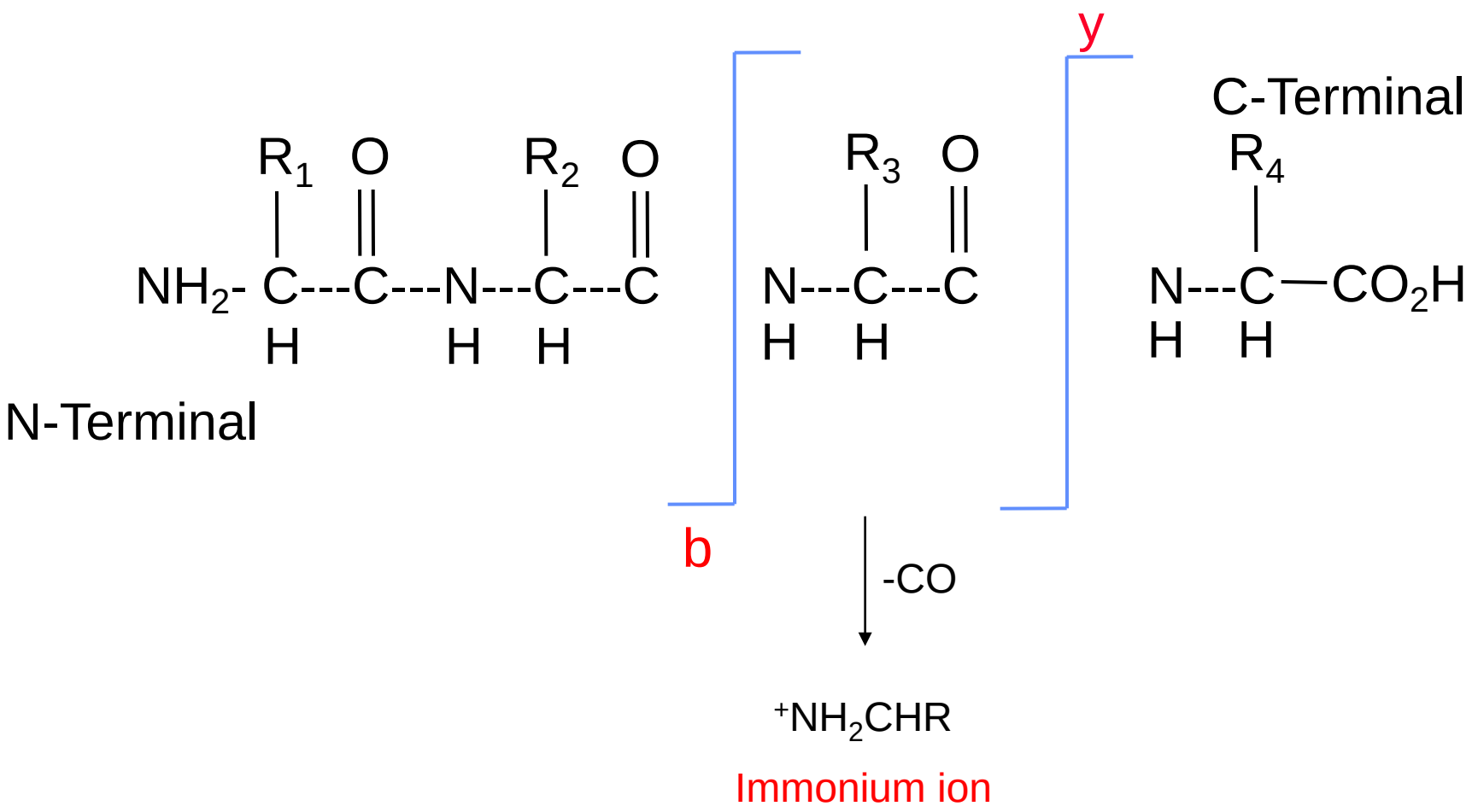
Internal Ions

•If a b or y ion undergoes a second b-y type fragmentation, internal ions can be formed.



Immonium Ions

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



Immonium Ion Masses

- Mass of amino acid residue –27 Da

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

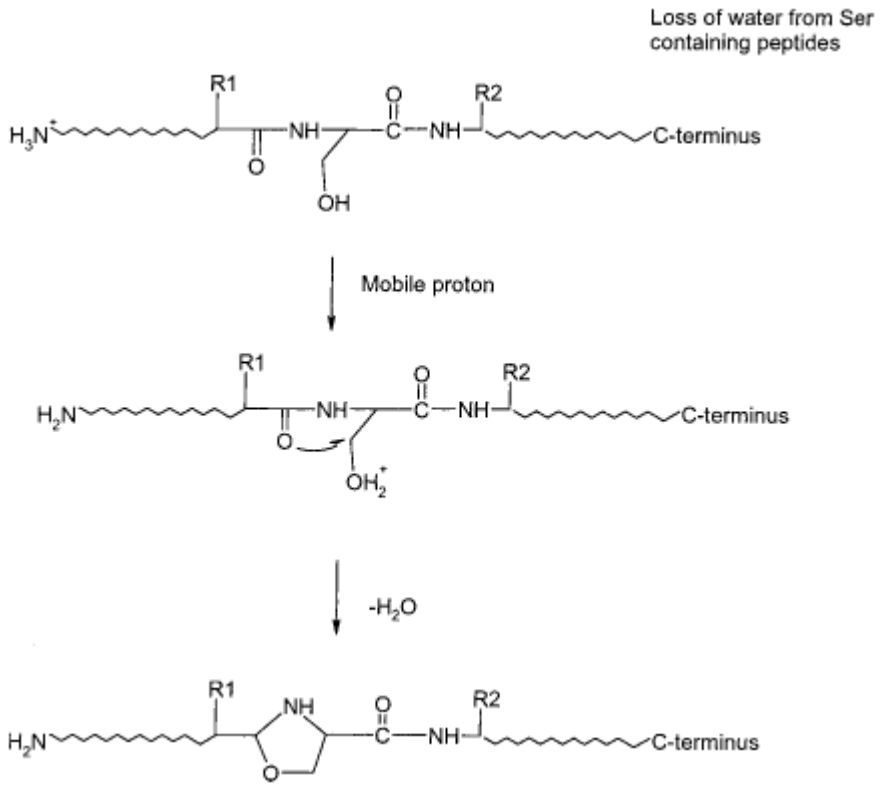
Amino acid	Immonium and related ion(s) masses		Comments
Ala	44		
Arg	129	59, 70, 73, 87, 100, 112	129, 73 usually weak
Asn	87	70	87 often weak, 70 weak
Asp	88		Usually weak
Cys	76		Usually weak
Gly	30		
Gln	101	84, 129	129 weak
Glu	102		Often weak if C-terminal
His	110	82, 121, 123, 138, 166	110 very 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

‘Loss’ Fragments

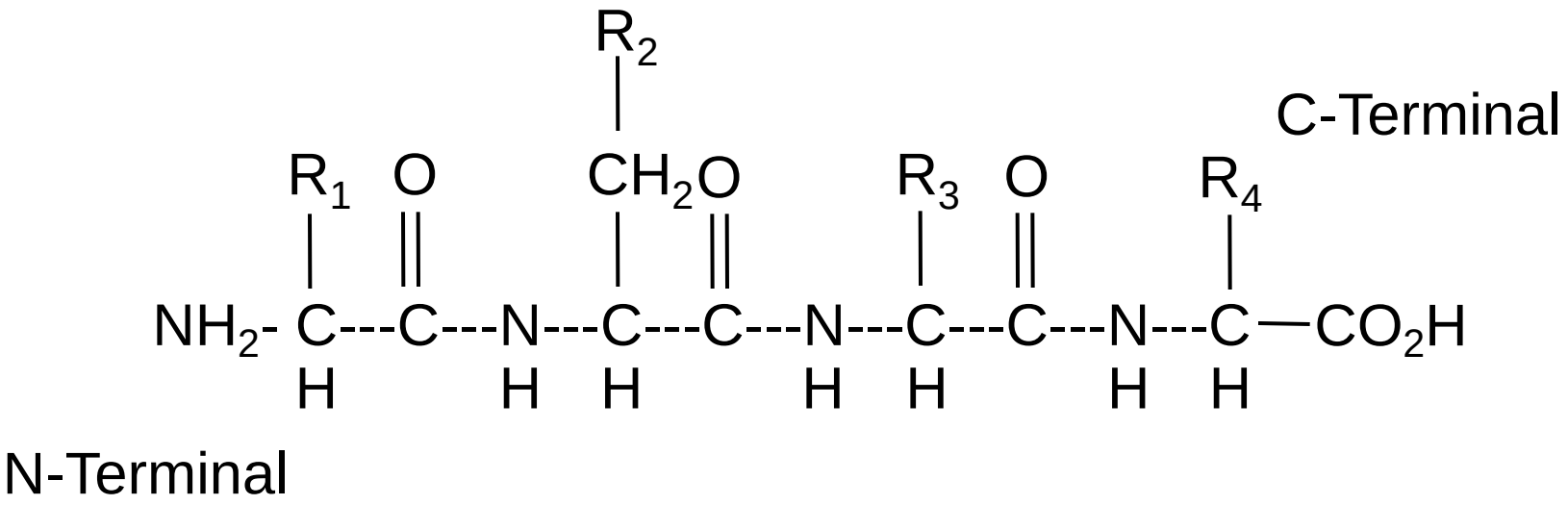
From a, b or y ions:

S, T, E and D can lose water.

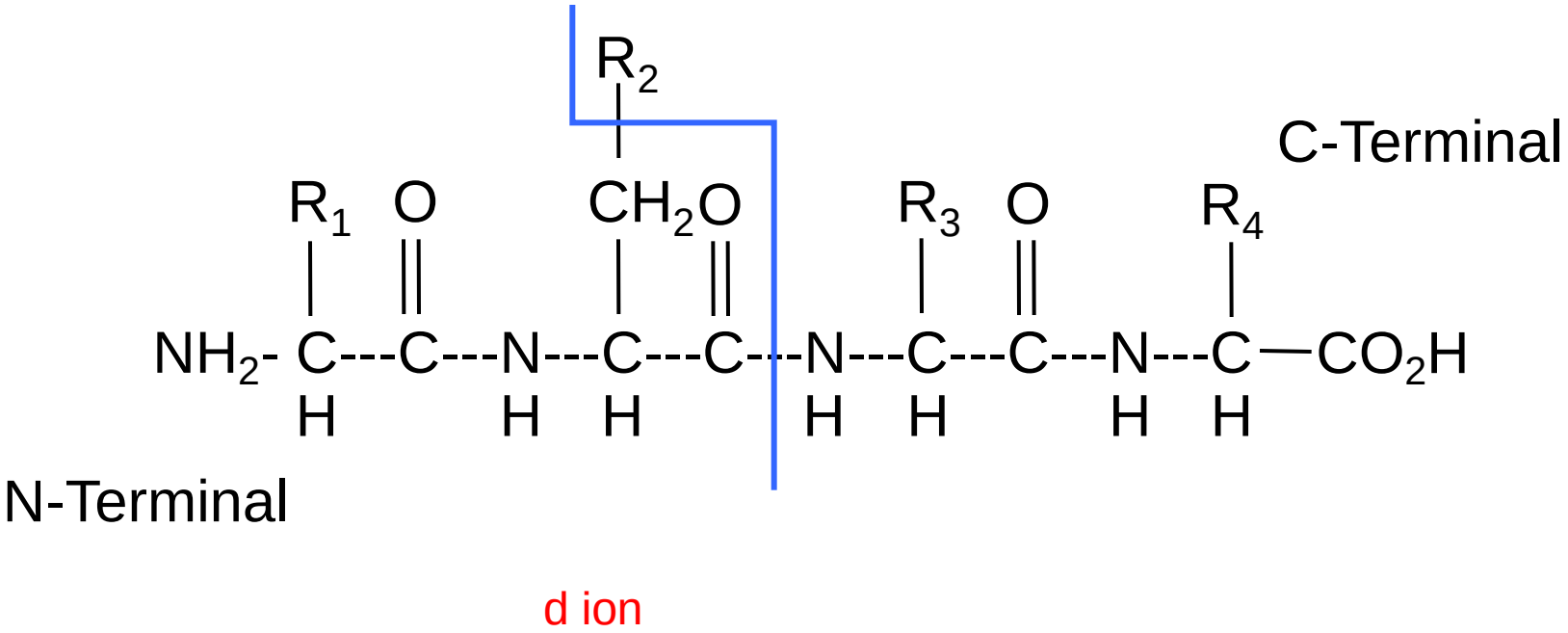
R, K, N and Q can lose ammonia.



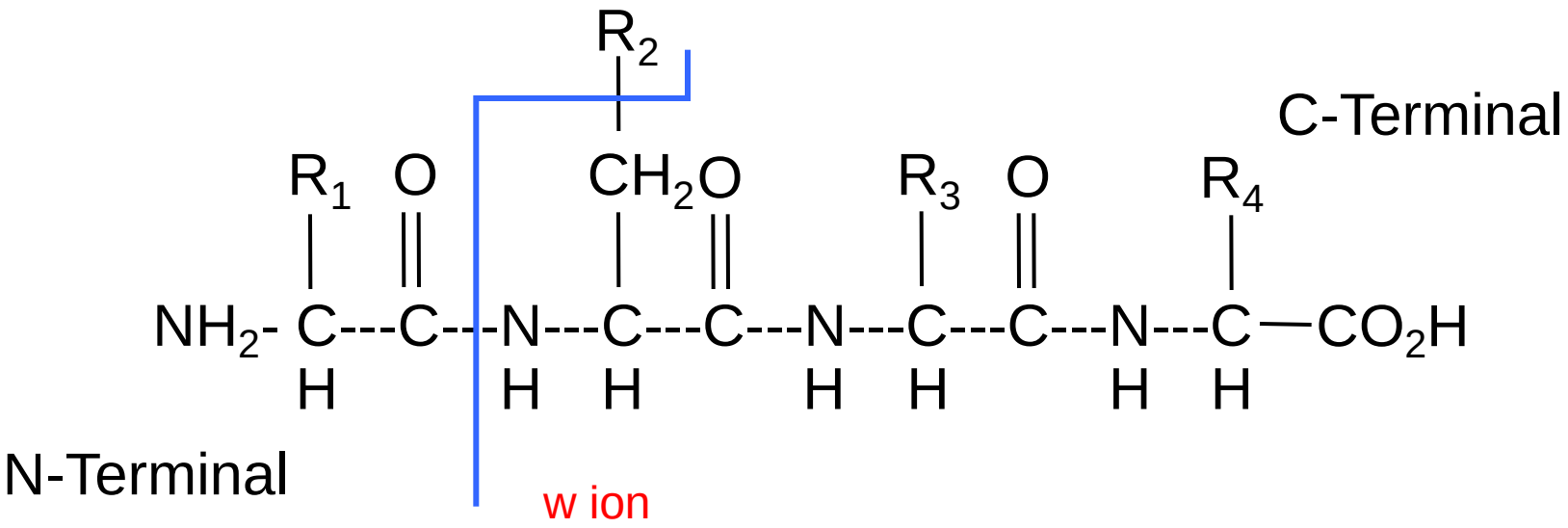
Side-Chain Fragments



Side-Chain Fragments



Side-Chain Fragments



Side-Chain Cleavage

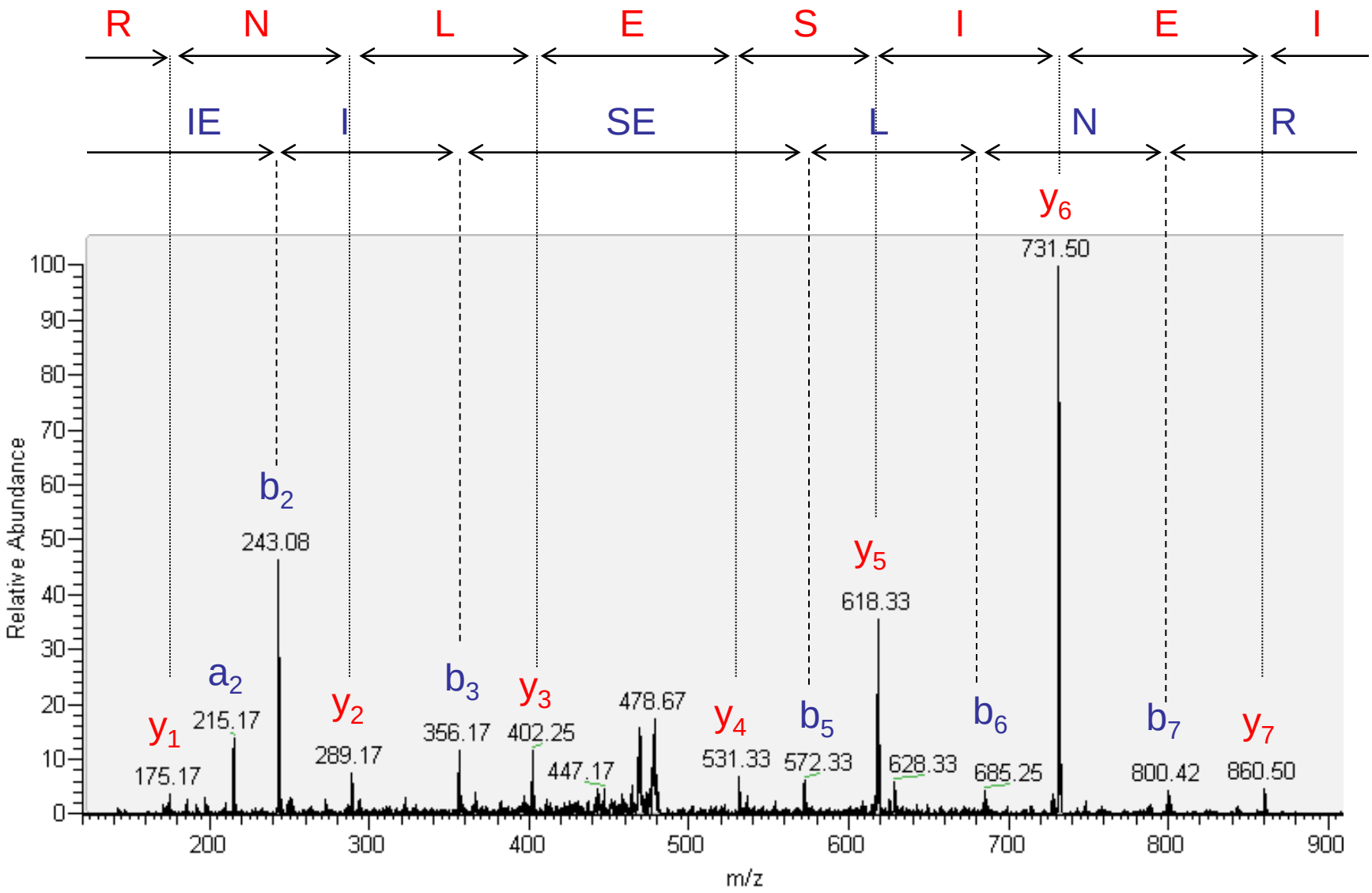
- Apart from a couple of special case exceptions, side-chains are not fragmented in low energy CID, but can be in high energy CID.
- Side-chain cleavage is the only way in a mass spectrometer to distinguish between Leu and Ile.

Special cases where side-chains are lost in low energy CID:

- Oxidized methionines can lose their side-chain ($SOCH_4$).
  - Unmodified methionine is stable
- Alkylated cysteines (e.g. carbamidomethylated) can lose their side-chain.
  - Unmodified cysteine is stable.

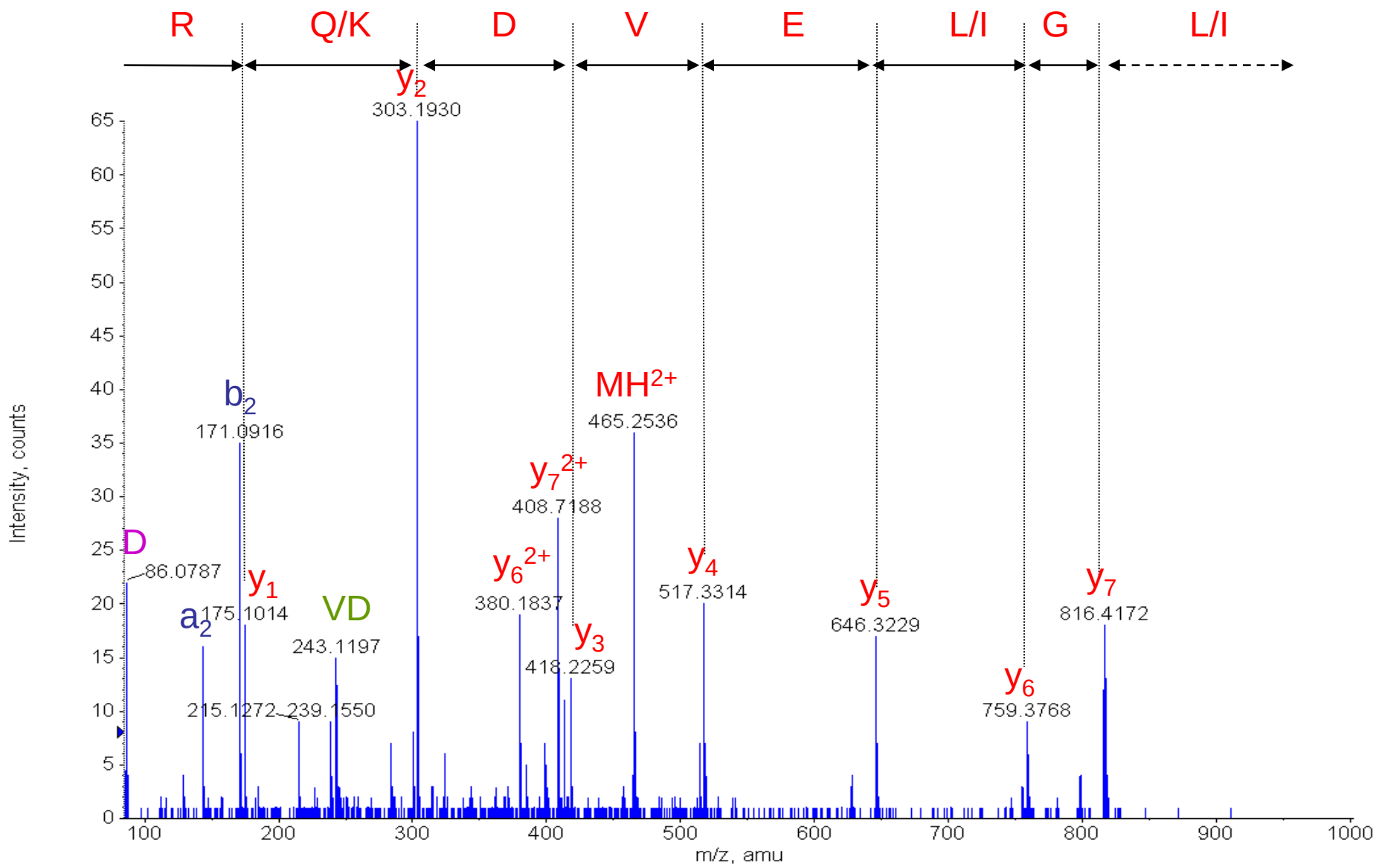
ESI-MSMS 487.27^{2+}: IEISELNR

- MSMS in iontrap



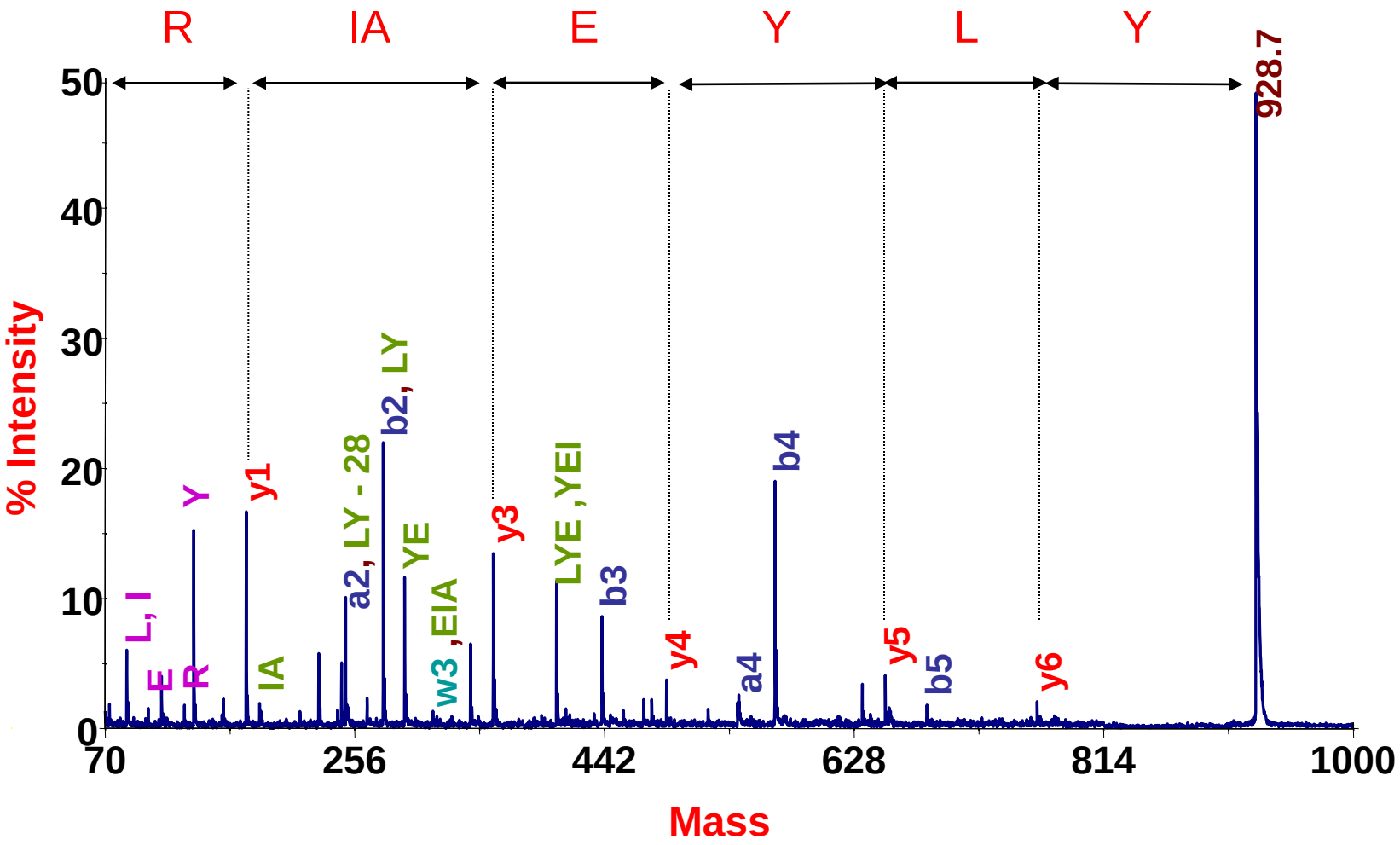
ESI-MSMS 465.25^{2+}: IGLEVDKR

- MSMS in quadrupole



MALDI-TOF-TOF MS/MS 928.7^{+}: YLYEIAR

•MSMS by PSD/CID



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

- Amino acids are chemicals, not homogeneous 'building blocks'.
- Certain cleavages are favored over others.
    - Can we predict which ones will be seen?
        - Answer: Sort of…
        - Statistical studies of large amounts of CID have been carried out[1,2]

- Cleavage N-terminal to proline gives intense fragment ions.
- Cleavage C-terminal to proline is often not seen.
- Different preferences for singly charged precursors
    - Preferential cleavage C-terminal to aspartic acid.

[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

What fragment masses might I see?

- MS-Product (part of Protein Prospector)

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

What fragment masses might I see?

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

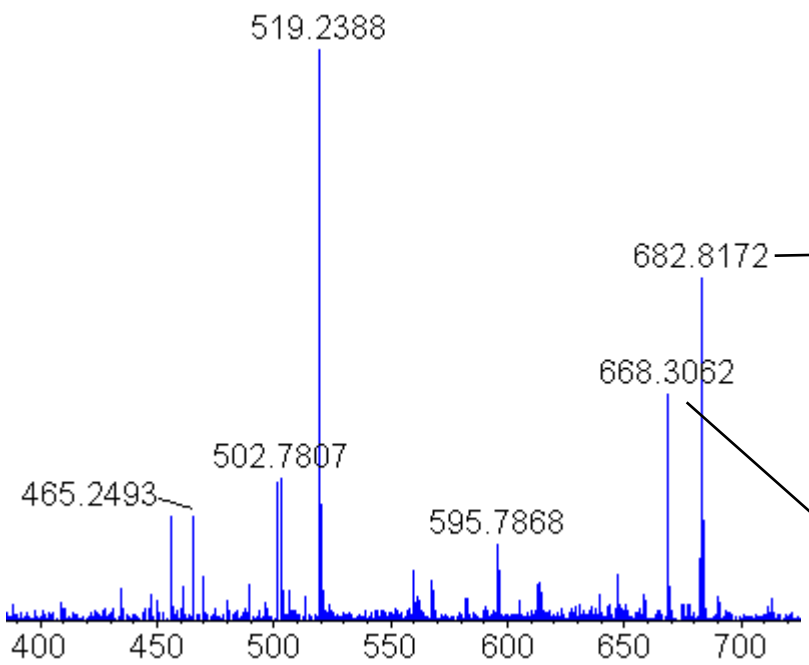
Theoretical Peak Table

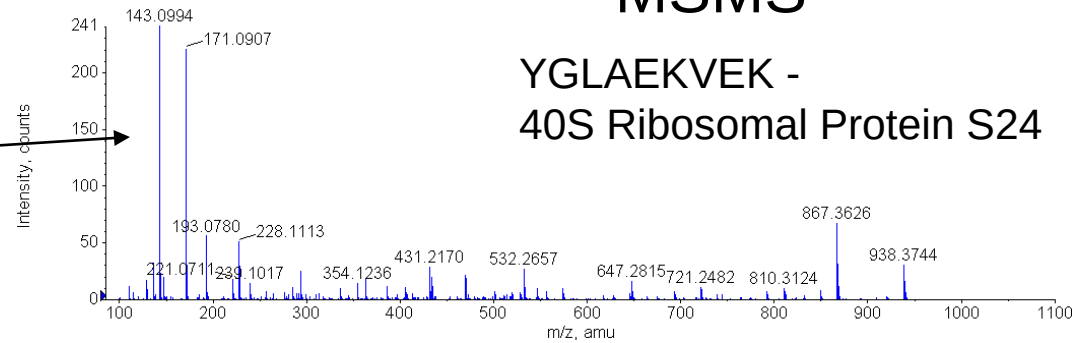
70.0651	P	199.1077	EP-28	294.1812	PTI-H_2O	397.2082	a_4	528.2664	EPTID-28
74.0600	T	199.1077	a_2	296.1605	a_3	399.2238	PTID-28	538.2508	EPTID-H_2O
86.0964	I	201.1234	ID-28	300.1554	EPT-28	407.1925	b_4-H_2O	538.2871	b_5
88.0393	D	209.0921	b_2-H_2O	302.1710	TID-28	409.2082	PTID-H_2O	556.2613	EPTID
102.0550	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

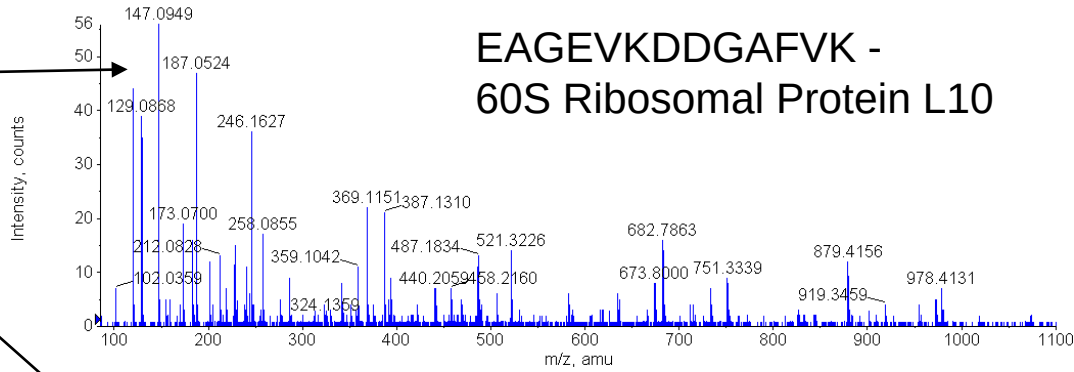
MS

MSMS



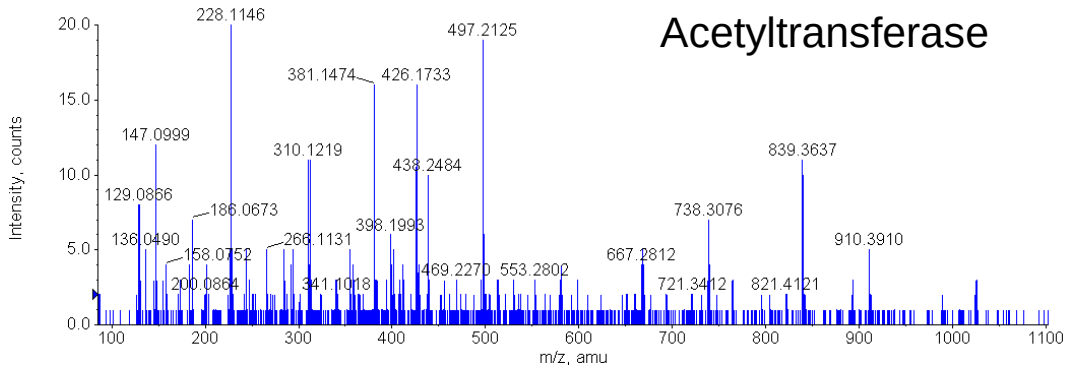


YGLAEKVEK -
40S Ribosomal Protein S24



EAGEVKDDGAFVK -
60S Ribosomal Protein L10

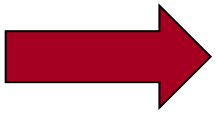
qSLNATANDKYK - Dihydrolipoamide Acetyltransferase



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
```



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

- Many commercial and freely available search engines.
  - Different instrument vendors promote their own tools.
  - Some tools are open-source.
  - Sequest; OMSSA; Xtandem!…
- Protein Prospector
  - MSMS searching ability has only recently been made freely available on web.
- Mascot
  - Limited version available for free over internet; more advanced version requires site license.

- For all, data is input and searched in a similar fashion, but they have different ‘scoring systems’ for deciding which matches are correct.

MSMS Search Parameters

- 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 do you allow for?

How do you determine a good match?
Scoring Systems

- Count number of peaks matched?

- 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

Solution:

- 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)

Scoring Systems – Protein Prospector

•Values based on significance and frequency of observation of ion type in correct vs incorrect answers.

Ion Type	Score
‘y’ ion	3
‘y’ loss ion*	1.5
‘b’ ion	1.5
‘b’ loss ion*	0.5
‘b’ + H_2O	1.0
‘a’ ion	0.5
‘a’ loss ion*	0
Internal ion	0.25
Immonium Ion	0.5
M-$SOCH_4$**	5.0

* loss of either H_2O or NH_3

** sidechain loss from parent ion of oxidized-methionine-containing peptide

Scoring Systems - Mascot

- Score is calculated by unpublished method/magic.
- Score is related to a probability of being incorrect.

Score = -10log (P)

i.e. Score of 10 = probability of being incorrect of 0.1.

Score of 20 = probability of being incorrect of 0.01.

MS-Tag Search Result

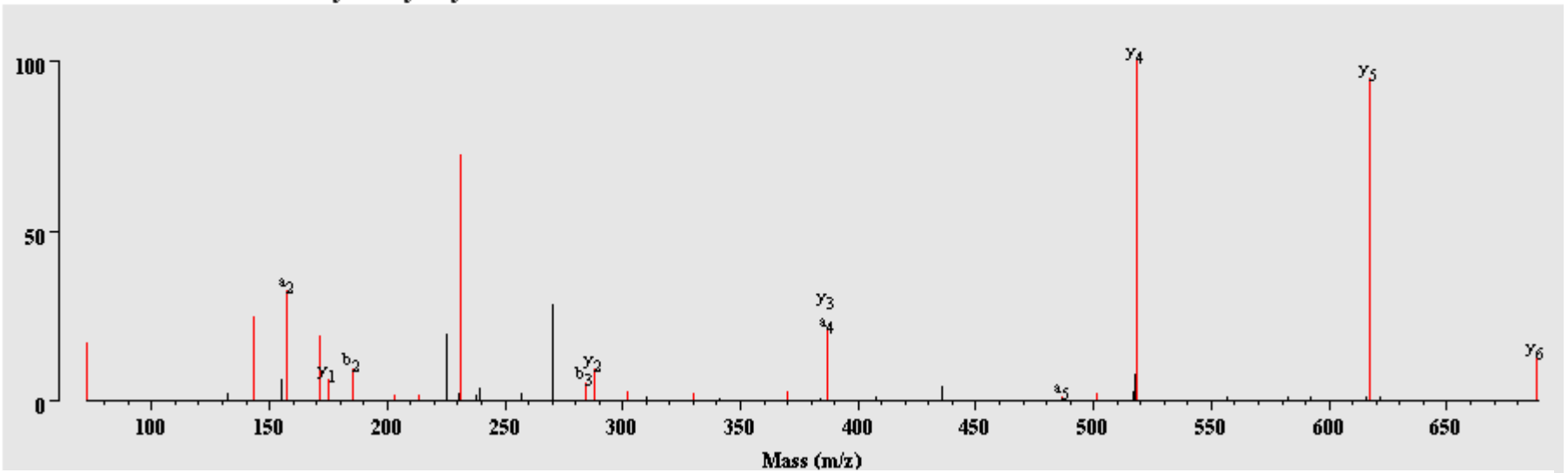
Result Summary

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)

Is the top match significantly better than random?

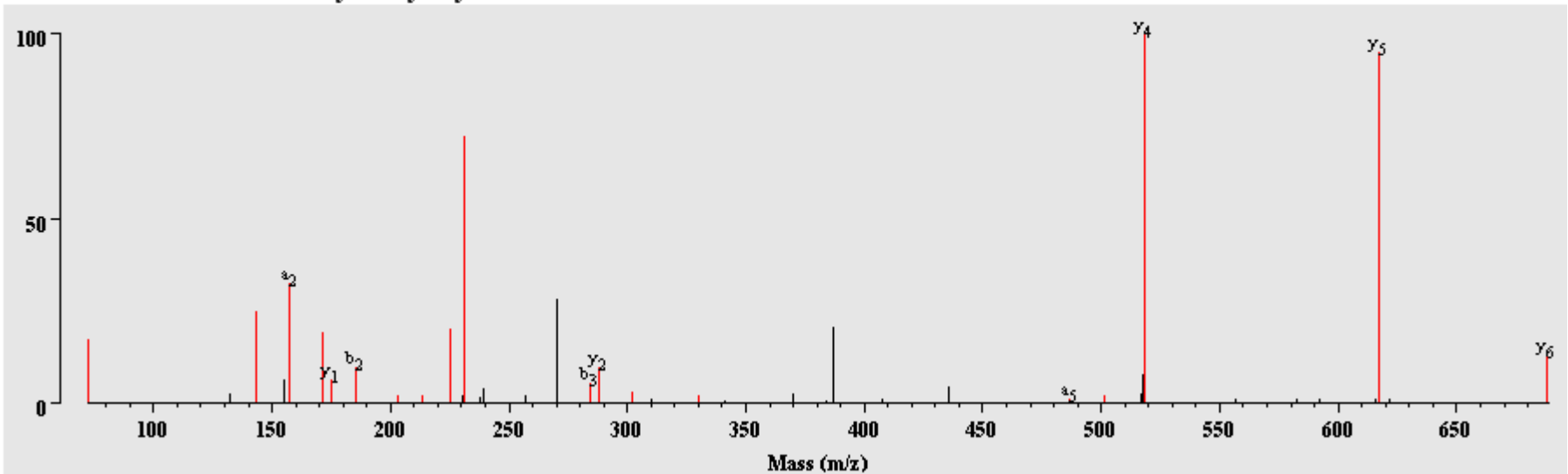
LAVMVLR^{+2}

Constant Modification: **Carboxymethyl Cysteine**



LAVTELR^{+2}

Constant Modification: **Carboxymethyl Cysteine**



How do you determine a good match?
Is my 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

- What is an expectation value?
  - Prediction of the number of times an event is expected to happen at random.

- For a peptide result an expectation value is:
  - Number of times a 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 that have 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.

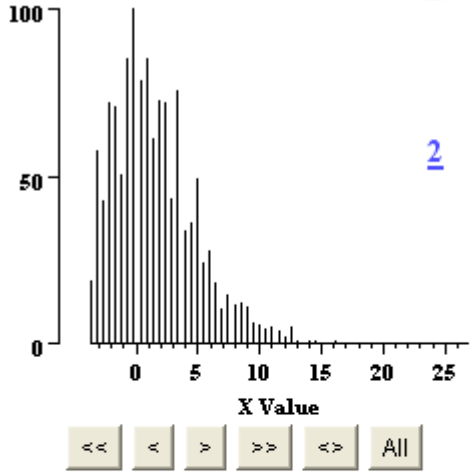
How Can We Calculate an Expectation Value?

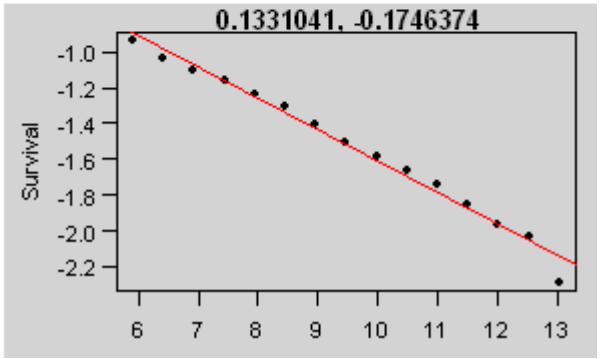
- Theoretical Calculation: What is the probability of 10 out of 25 peaks matching a random (incorrect) assignment?
  - Assumes theoretical model takes into account all variables that can change the number of peaks matching at random.
  - Assumes sequences in database are random.
- Approach used by Mascot

- Calculation based on results: Model scores of the incorrect answers to a distribution and extrapolate 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 datapoints to accurately model the distribution
- Approach used by Protein Prospector

Example of Prospector E-Value Calculation

1	(R)LAVMVIR (W)	26.5	401.2744^{+2}	8069/5.2	P81650	PSEHA	Beta-galactosidase (EC 3.2.1.23) (Lactase) (Beta-D-galactoside galactohydrolase)
1	(R)LAVMVLR (W)	26.5	401.2744^{+2}	6353/5.3	P00722	ECOLI	Beta-galactosidase (EC 3.2.1.23) (Lactase)
1	(R)LAVMVLR (W)	26.5	401.2744^{+2}	6679/5.8	Q47077	ENTCL	Beta-galactosidase (EC 3.2.1.23) (Lactase)
2	(R)LAVTELR (G)	21.6	401.2744^{+2}	4062/6.0	Q13608	HUMAN	Peroxisome assembly factor 2 (PAF-2) (Peroxisomal-type ATPase 1) (Peroxin-6) (Peroxisomal biogenesis factor 6)





E-value

401.2744	2	50	LAVMVLR	31.04	26.5	0.010	3

Can we use any more information to increase our confidence in an assignment?

- Other peptides from the same protein may be identified in the same experiment
  - If you are confident a protein is in the sample, you are more likely to find more peptides from the same protein.

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

From Peptide ID to Protein ID

- We are identifying peptides.
- We want to know what proteins were in the original sample.

- Conversion of peptide to protein information not completely straightforward
    - Multiple database entries for the same protein
        - Sequence variants / isoforms
        - Splice variants

- If peptides from a given protein have already been identified, then it is more likely we will find peptides from this protein.
    - How do we use this information?

Peptide to Protein - Mascot

- Combine peptide scores together to give 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

Peptide to Protein – Protein Prospector

Discriminant Score:

Combine multiple parameters about 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:

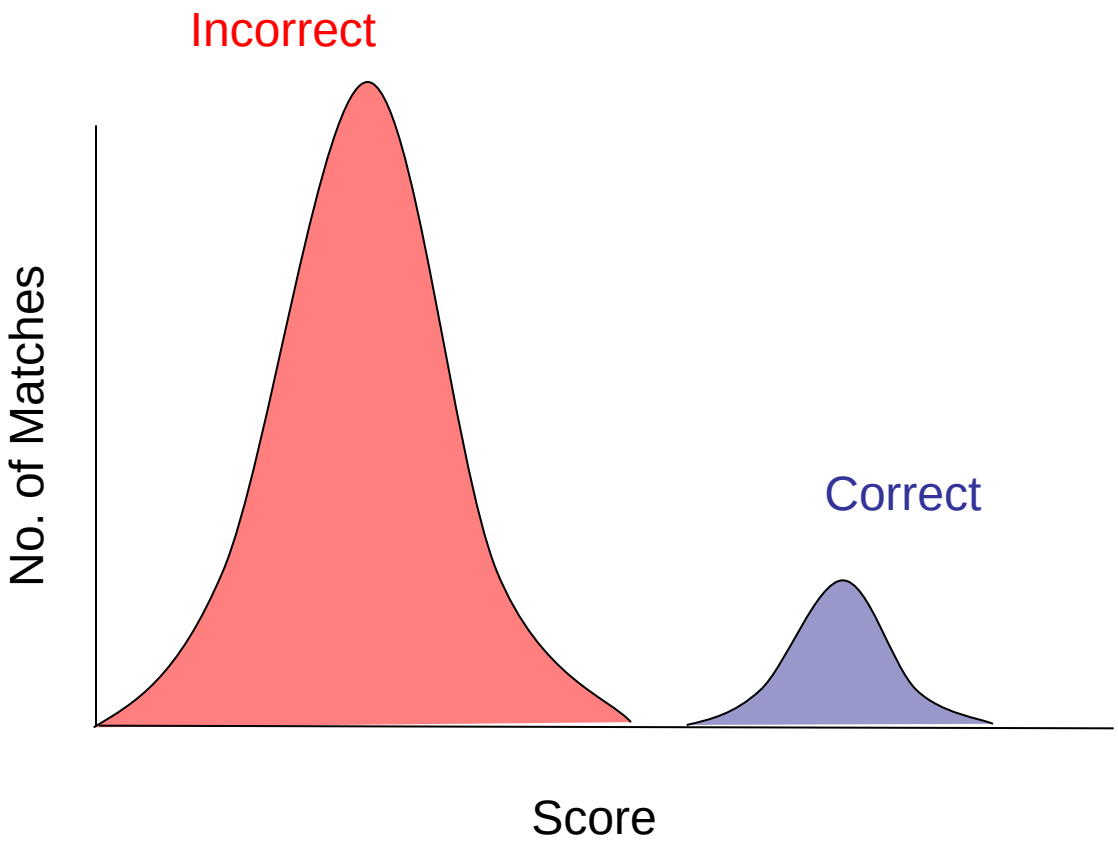
Protein Prospector Discriminant Score combines two parameters:

- Expectation value for peptide identification
- 'Best peptide score': Score of the highest scoring peptide matched to the particular protein database entry.

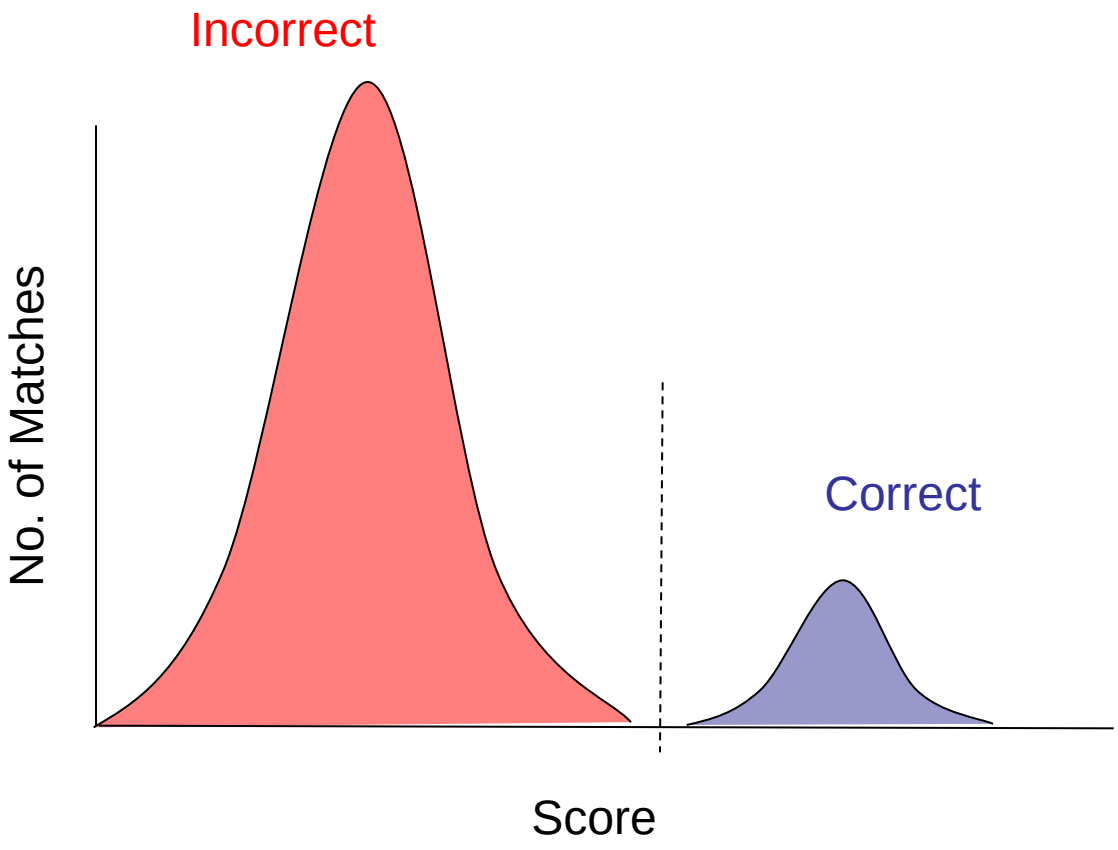
Disc Score = a*(-log E-value) + b*(best peptide score) + constant.

- Weightings are instrument specific.

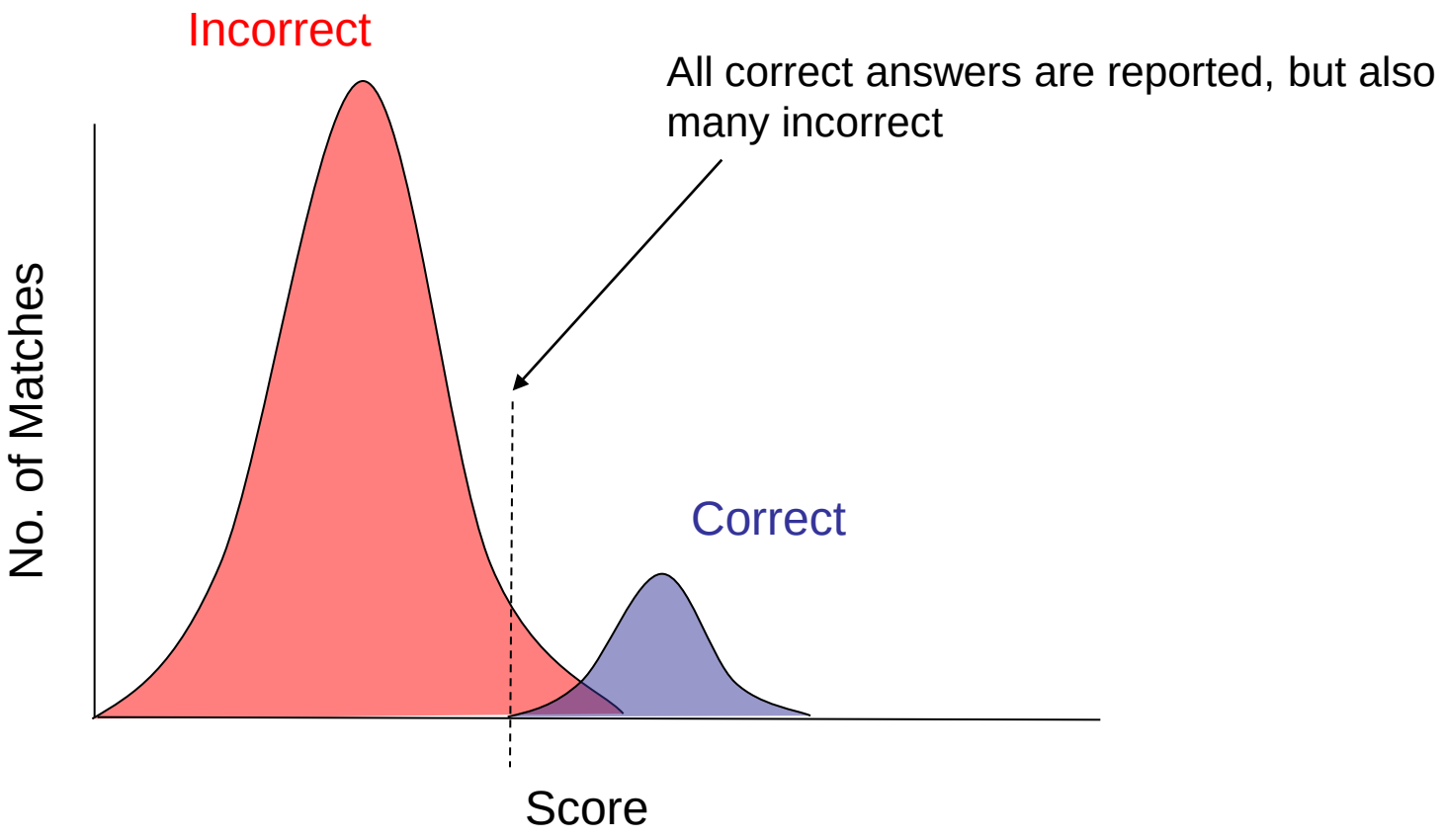
Search Result Scoring Systems



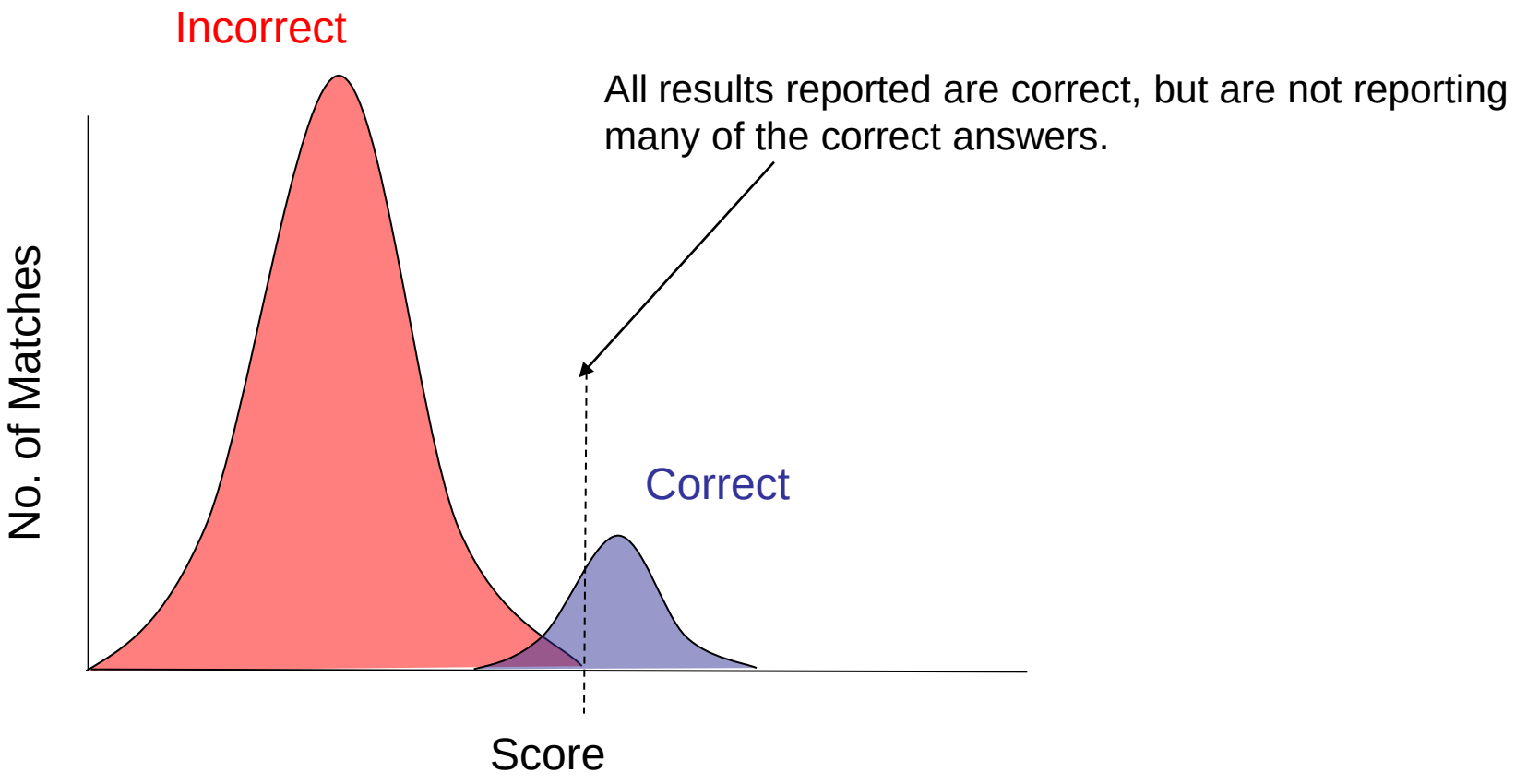
Search Result Scoring Systems



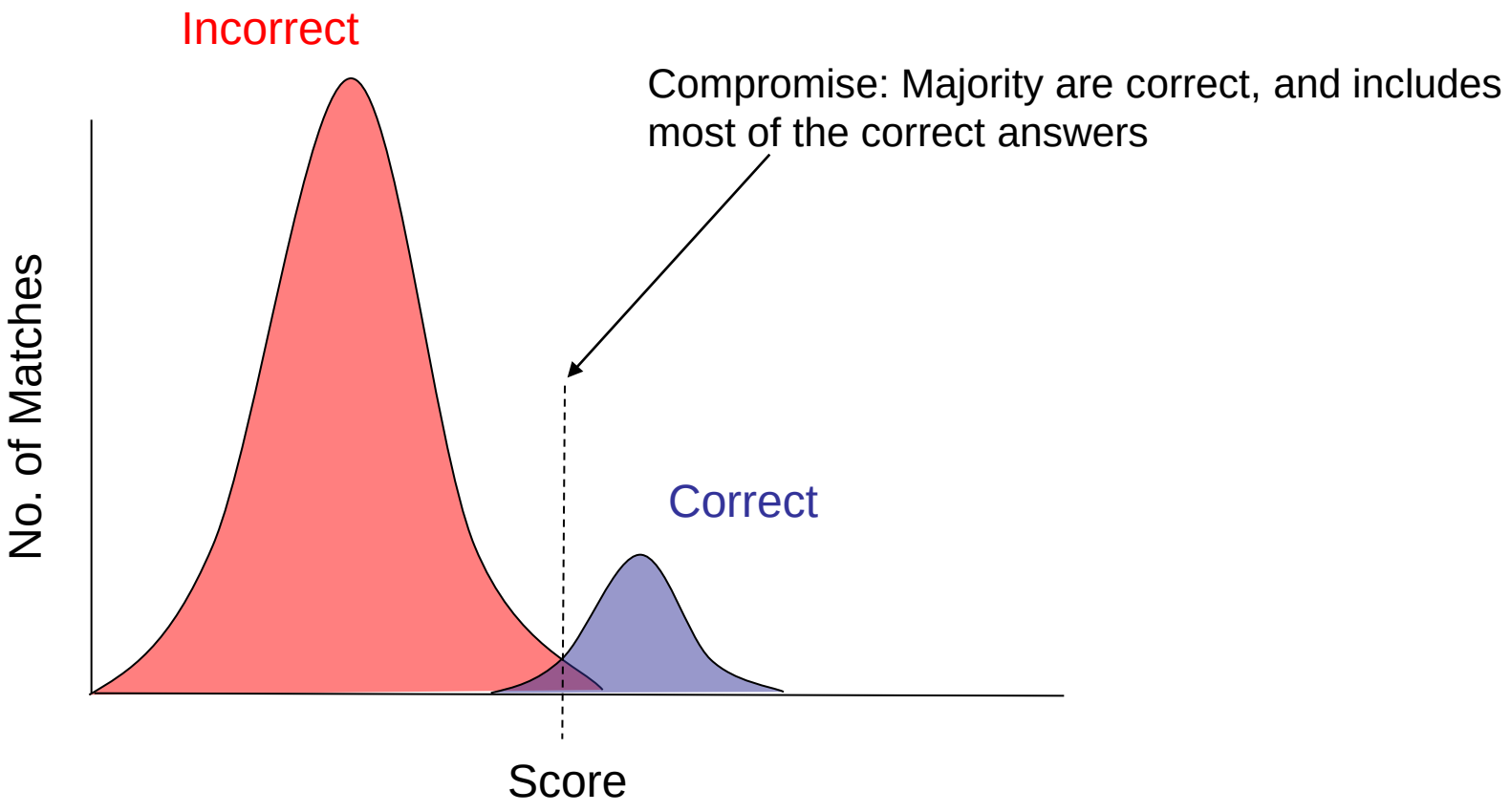
Search Result Scoring Systems



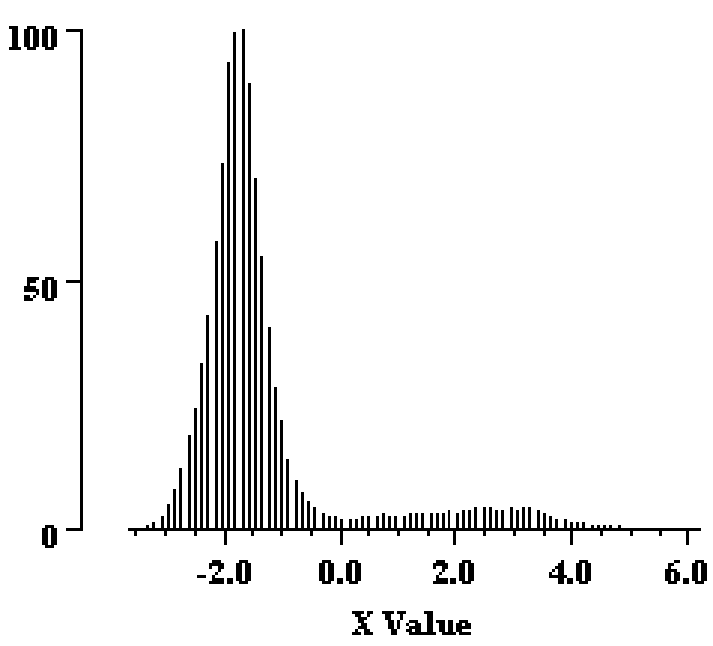
Search Result Scoring Systems

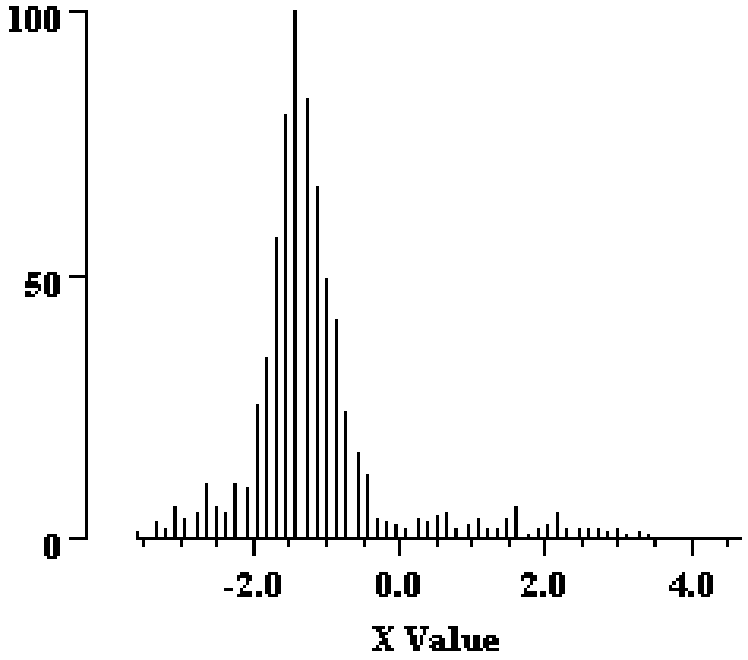


Search Result Scoring Systems



Discriminant Score Distributions – Real Results





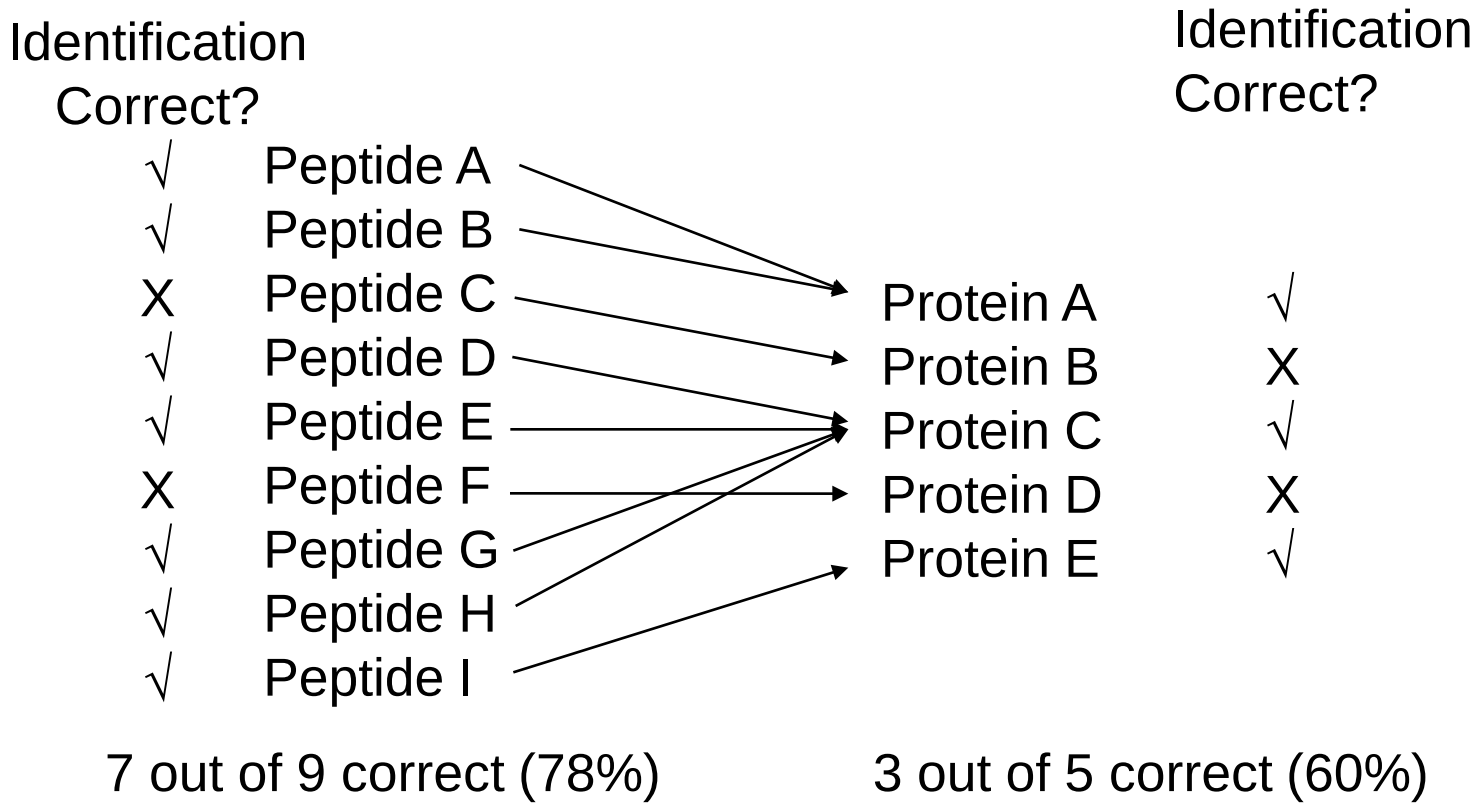
Search Engines give different results

m/z	z	ppm	Peptide	S	Score	Expect	Disc Score	# in DB
355.6959	2	27	MSGIFR	27.98	18.8	0.024	1.26	9
401.2744	2	50	LAVMVLR	31.04	26.5	0.010	1.41	3
407.2368	2	24	LNVENPK	22.04	22.6	0.032	1.21	1
450.7146	2	41	FNDDFSR	26.1	25.4	0.015	1.34	1
477.7467	2	54	LTAACFDR	27.1	19.1	0.029	1.23	1
542.2814	2	31	GDFQFNISR	31.3	27.3	0.0050	1.54	1
681.3903	2	38	LWSAEIPNLYR	35.28	28.1	1.7e-4	2.14	1
729.3964	2	43	APLDNDIGVSEATR	27.3	35.7	4.8e-7	3.18	1
736.9074	2	37	IGLNCQLAQVAER	31.98	26.6	3.7e-4	2.00	1
503.2559	3	38	YSQQQLMETSHR	25.56	24.3	5.4e-4	1.93	1
567.0809	4	46	DVSLLHKPTTQISDFHVATR	35.51	26.8	4.6e-5	2.37	1

	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

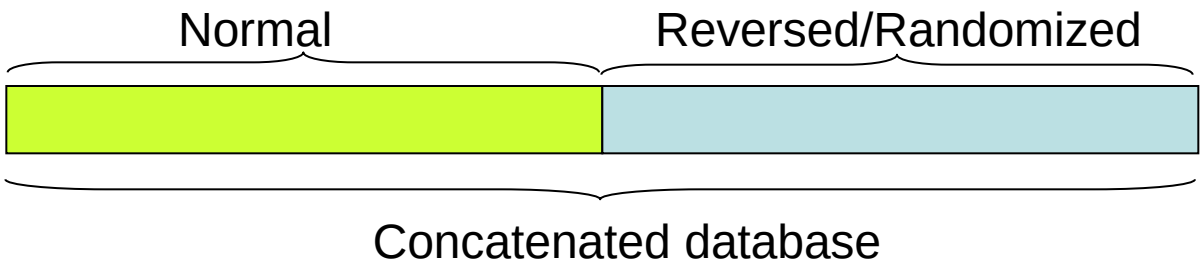
Peptide Errors Are Amplified in Protein Identifications

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



How Do You Assess the Reliability of Results?

- Create a randomized/reversed version of the database and concatenate it onto the end of the normal database. (target-decoy)



- Search data against a concatenated database.
- At a given threshold, for every match to the random part of the database you predict one incorrect match to the normal part of the database.

- E.g. if in your database search you match 10 spectra above your score / expectation value threshold to the random part of the database, then there are probably 10 incorrect matches among those assigned to peptides in the normal part of the database.

- This is calculating a reliability for the dataset as a whole; not individual matches.

Size of Database Affects Protein Identification Reliability

Search of Yeast Nuclear Pore Interacting Proteins

Database	Correct	Incorrect
SwissProt Yeast+	2214	1055
SwissProt All	2118	1151
NCBI	2045	1224

- Larger database gives fewer correct results with lower confidence, providing the majority of the proteins are in the smaller database.

Homology-based Searching

- If your protein is not in the database, how do you identify it?

- It may be highly homologous to another protein / same protein in different species

- *De Novo* Sequencing, then BLAST or MS-Homology
  - Searching allowing for amino acid substitutions

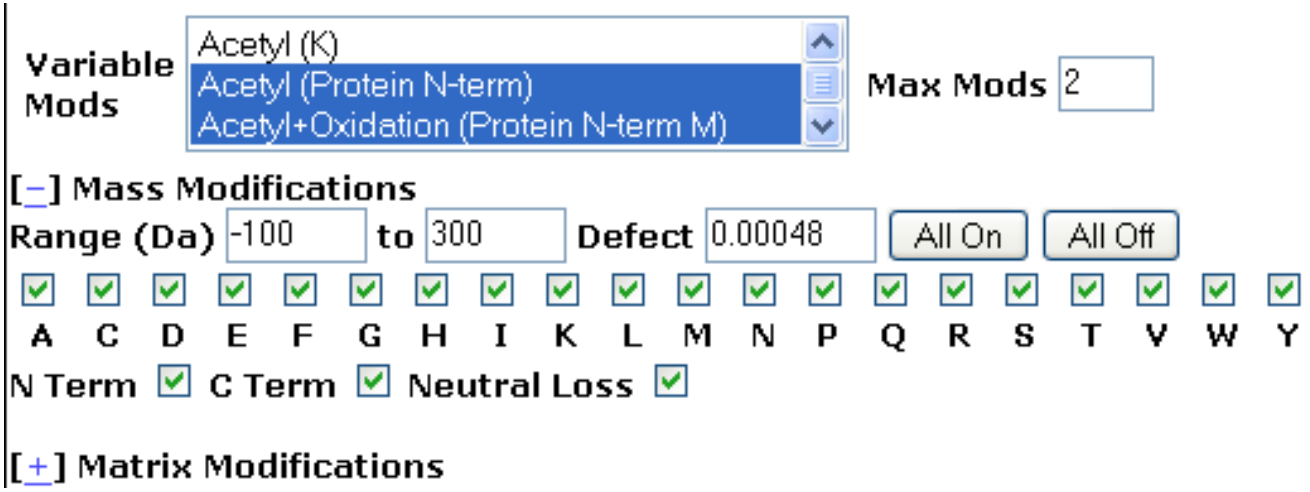
    [213]ENFAGVGV[I|L]DFES 6
    [217]GA[Q|K][242]DENTR 4

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

Unpredicted Mass Modifications

- User tells software a list of modifications to consider.
- What if there are modifications present that are not considered?
  - Peptides are not identified.
- How can I identify these peptides?
  - Considering more modifications makes it harder to get high confidence assignments.

- After initial search to identify proteins present in the sample, can perform second search only against proteins (accession numbers) identified in the initial search.
  - If modified peptides are present, there will probably be unmodified peptides from the same protein present.



Mass Modification Searches: How Reliable are these Results?

Expectation values are reported, so they have the same level of reliability – right?

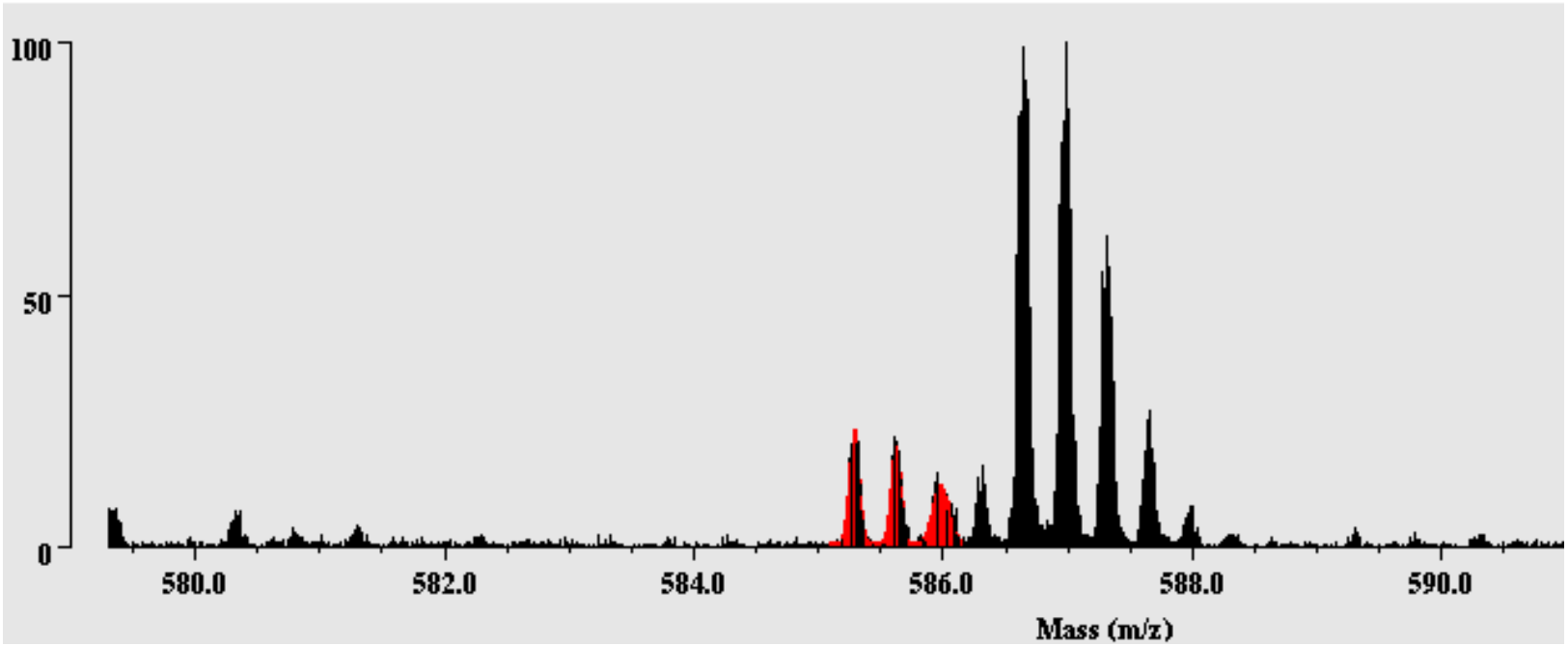
Two issues:

1. Expectation value is measure of likelihood that a match is random, assuming you are searching against a database representing all the options. You are using a very restricted protein database. Is this still representative?
2. Determining a match is non-random, does not mean it is correct. The match may be to the correct peptide, but with the wrong modification and/or in the wrong location; i.e. it is homologous to the correct answer.

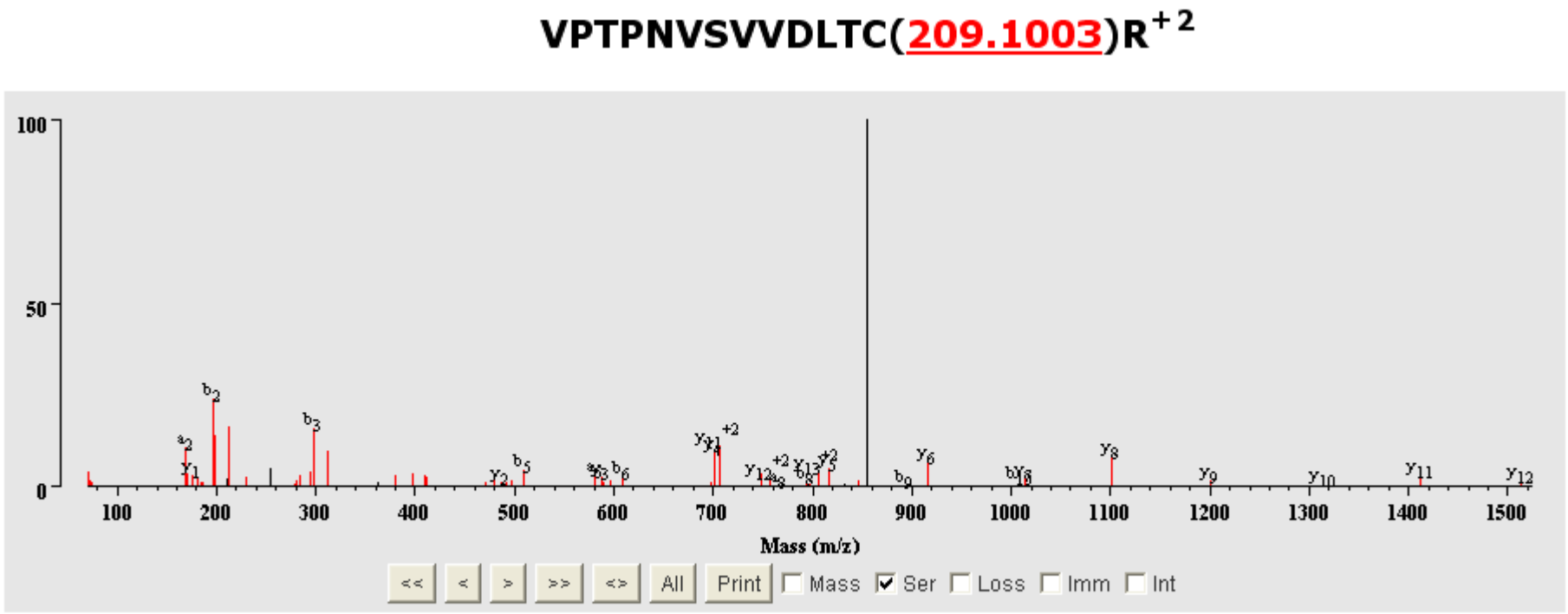
- How do you determine which ones are correct?
  - Manual verification
  - Software is being developed that, assuming the peptide ID is correct, can report a likelihood of a particular site assignment over different locations in the peptide.

What do you Find?

- Non-tryptic peptides.
- Incorrect precursor mass – selected 2nd isotope instead of monoistopic peak.
- Spectrum is of different precursor!



‘Real’ Unexpected Modifications



Max Intensity: 2701
Num Matched: 62/80 (22.5% unmatched)

- Carbamidomethylated DTT modification of cysteine.

Why Are All Spectra not Identified?

3269 Spectra:

22 peptides too short to be confident of assignment (m/z <620)
43 mixture of precursor ions
24 spectra of methylated trypsin
24 Deamidation of N
4 peptides sequences not in db
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 contains a methylated lysine
82 wrong charge
1 wrong charge and mixture
2 wrong charge – not peptide
78 wrong isotope
14 wrong charge and monoisotopic peak
3 wrong isotope and mixture
11 MSMS of peptide that has lost water in-source
8 peptides formed from in-source fragmentation of abundant co-eluting peak
1 peptide contains an internal disulfide bond

=904 Spectra

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

Conclusions

- Mass Spectrometry can be used for analysis of simple or complex mixtures.

- Bioinformatic tools available on web and in-house greatly facilitate data analysis.

- Search engines make mistakes.

- Need to be able to distinguish between real and incorrect search results.
  - Appropriate choice of search engine parameters important.
  - Use of probability/expectation value to measure assignment reliability
  - Use of random/concatenated database searching can estimate false positive rates for dataset as a whole.

Labs

- GH-S132 (Mass Spectrometry Lab)
- Tuesday 18th (Tomorrow) 1pm

  We will explain instruments, how data is acquired, give you your data and explain how to analyze it.