Protein Identification Using Tandem Mass Spectrometry

Nathan Edwards

Informatics Research

Applied Biosystems

Outline

- Proteomics context
- Tandem mass spectrometry
- Peptide fragmentation
- Peptide identification
  - *De novo*
  - Sequence database search
- Mascot screen shots
- Traps and pitfalls
- Summary

Proteomics Context

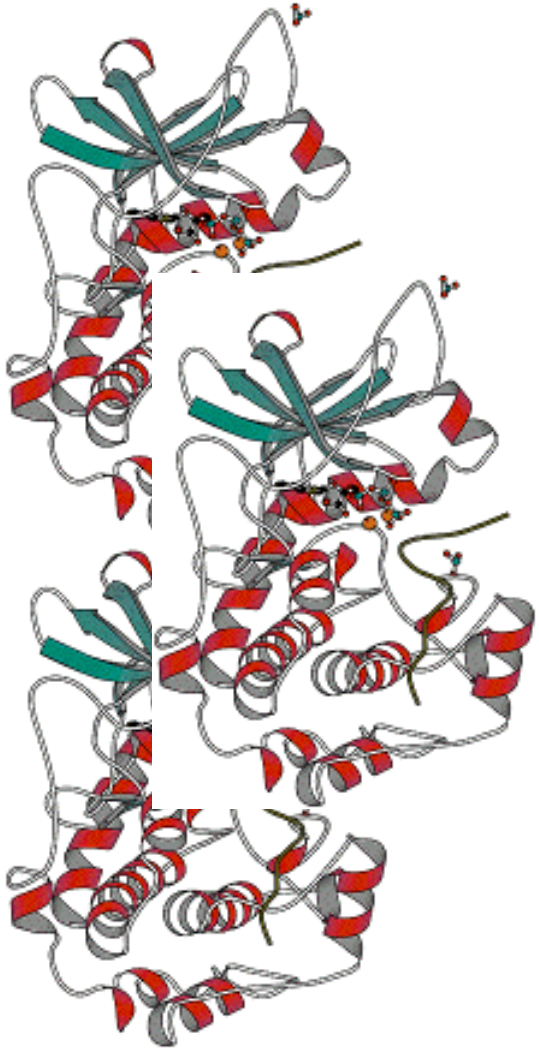
High-throughput proteomics focus

- (Differential) Quantitation
  - How much of each protein is there?
- Identification
  - What proteins are present?

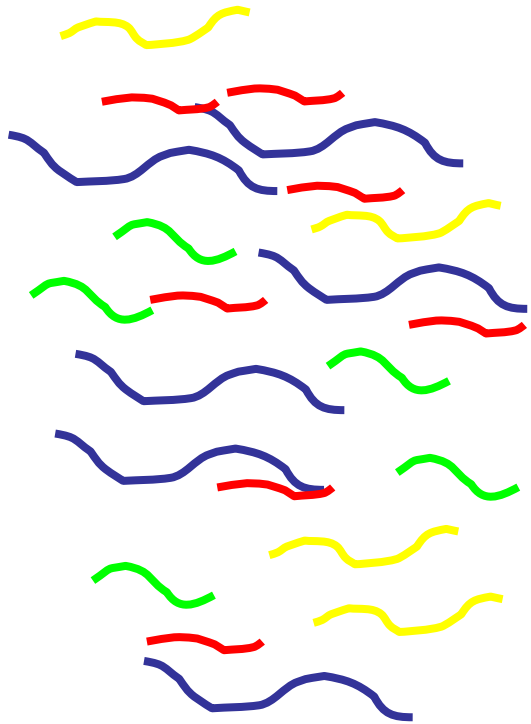
Two established workflows

- 2-D Gels
- LC-MS, LC-MALDI

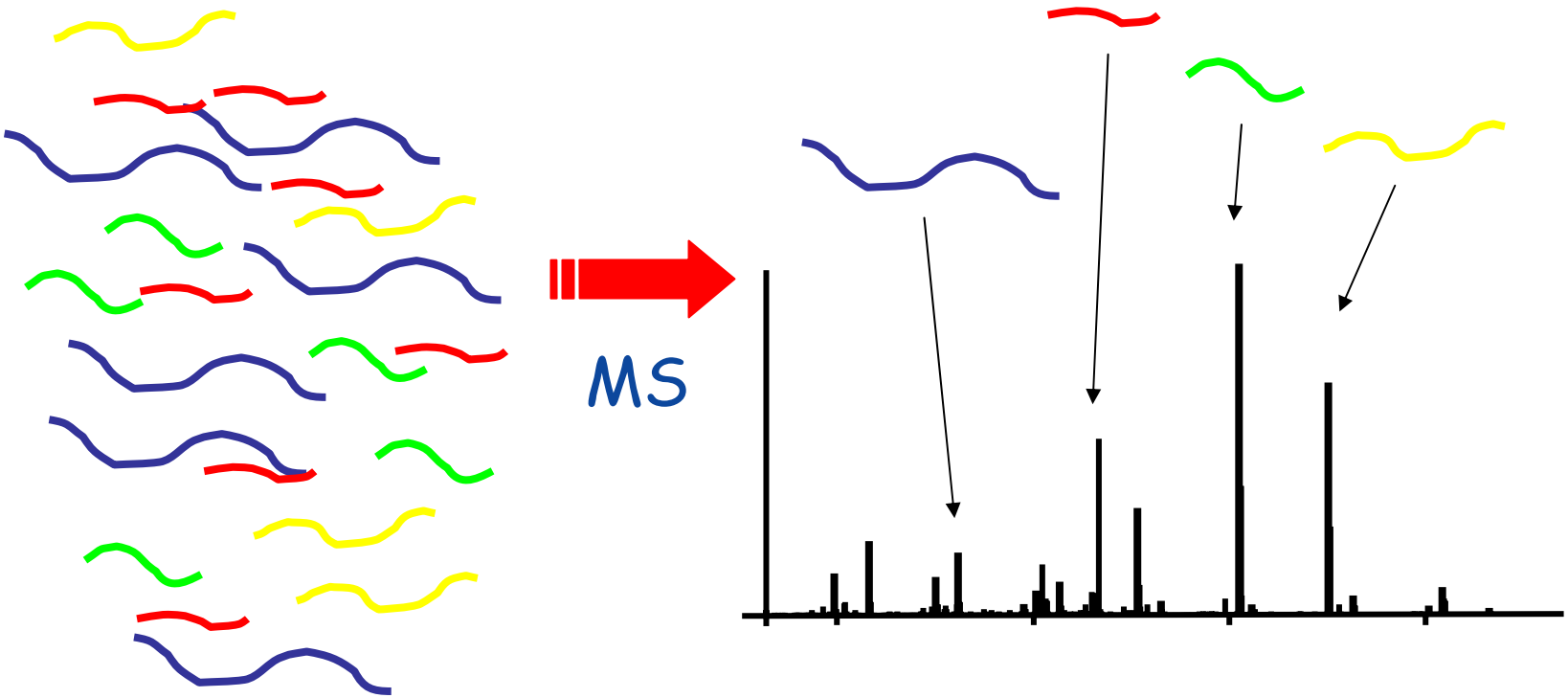
Sample Preparation for Tandem Mass Spectrometry



Enzymatic Digest
and
Fractionation



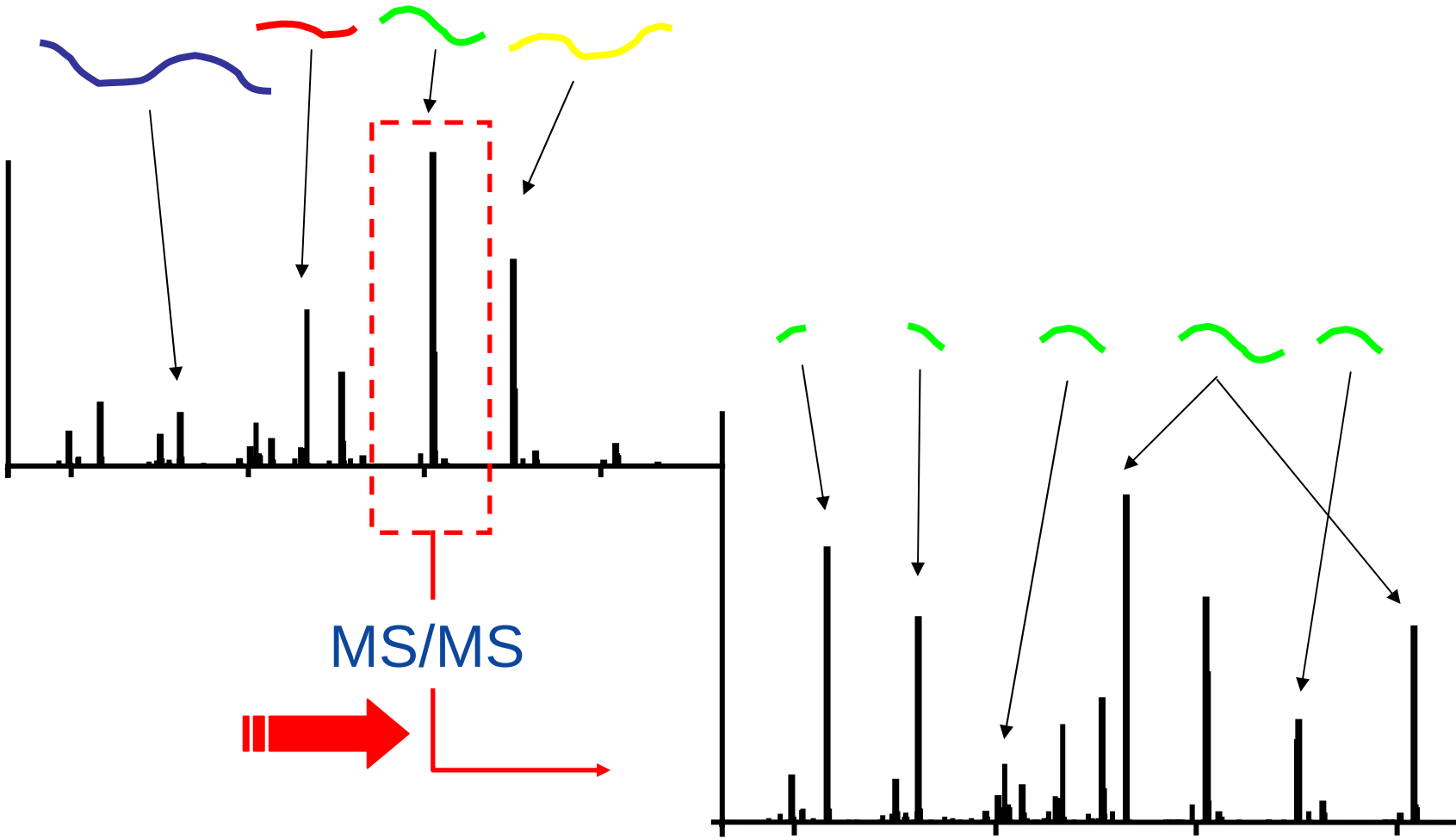
Single Stage MS



Tandem Mass Spectrometry (MS/MS)

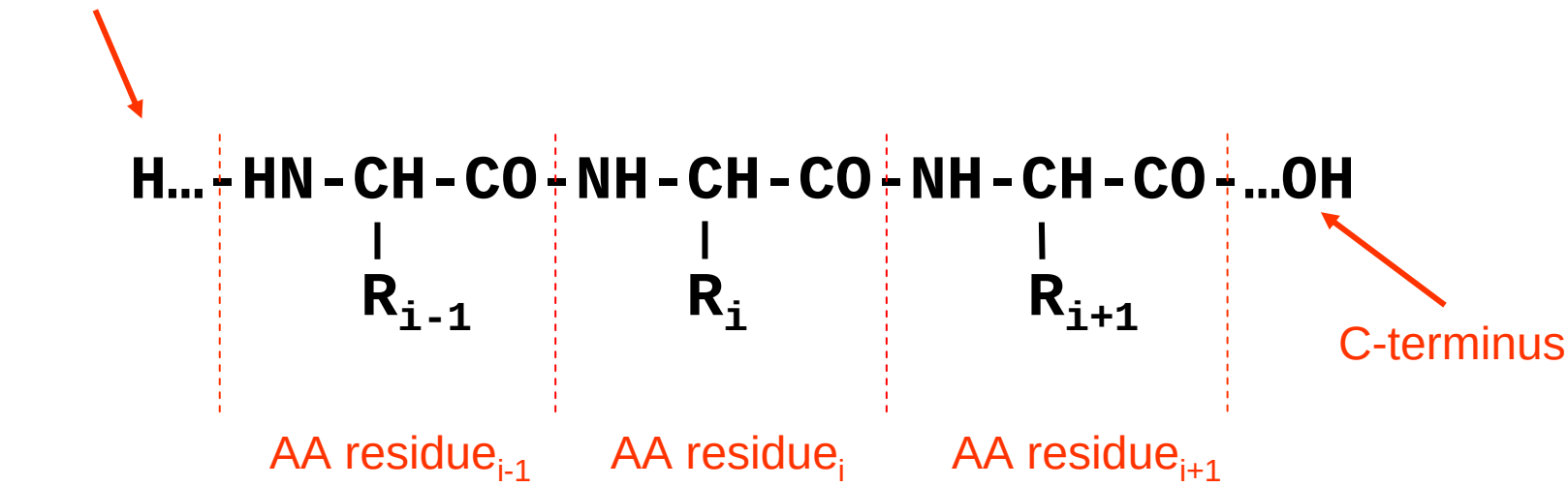
- Acquire mass spectrum of sample
- Select interesting ion by m/z value
- Fragment the selected “parent” ion
- Acquire mass spectrum of parent ion’s fragments

Tandem Mass Spectrometry (MS/MS)



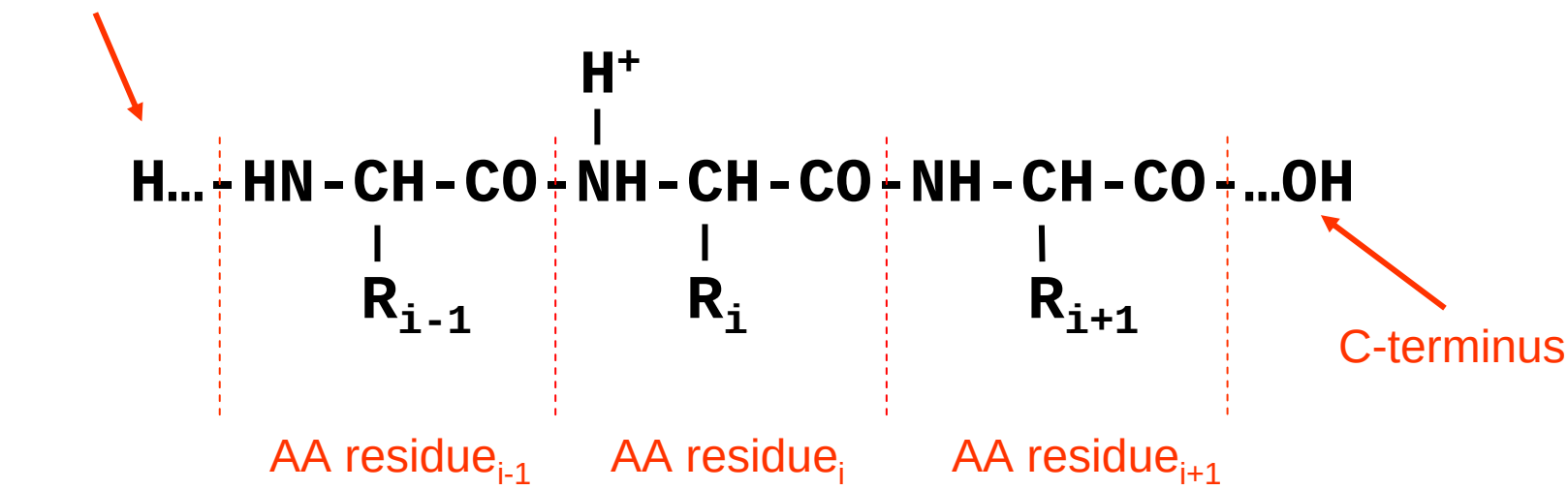
Peptide Fragmentation

Peptides consist of amino-acids arranged in a linear backbone.



Peptide Fragmentation

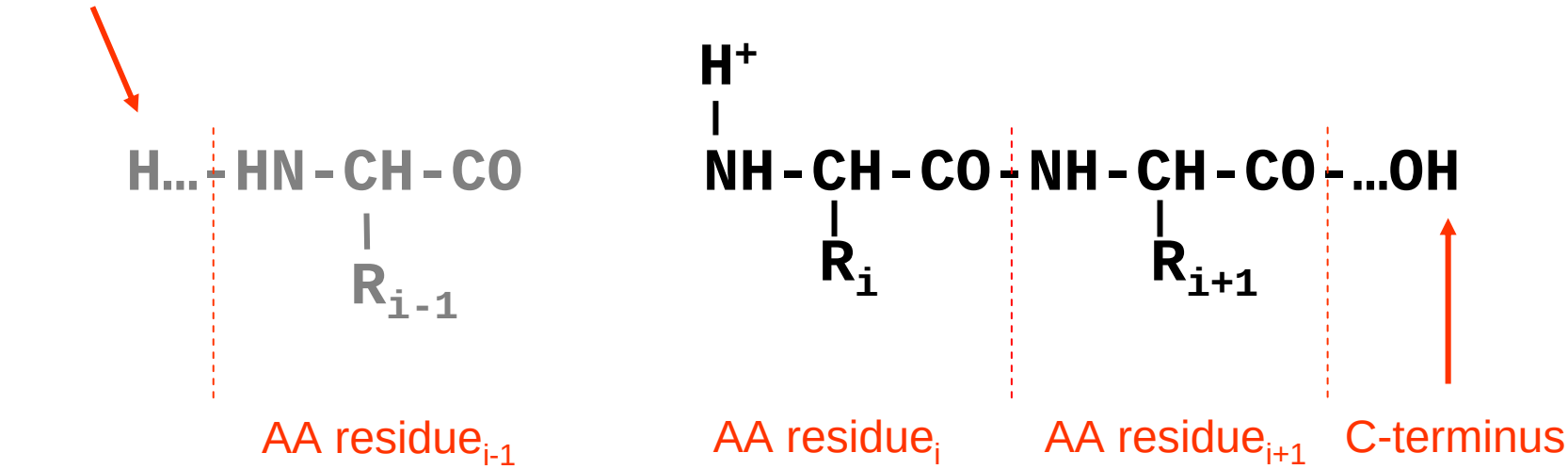
Peptides consist of amino-acids arranged in a linear backbone.



Ionized peptide (addition of a proton)

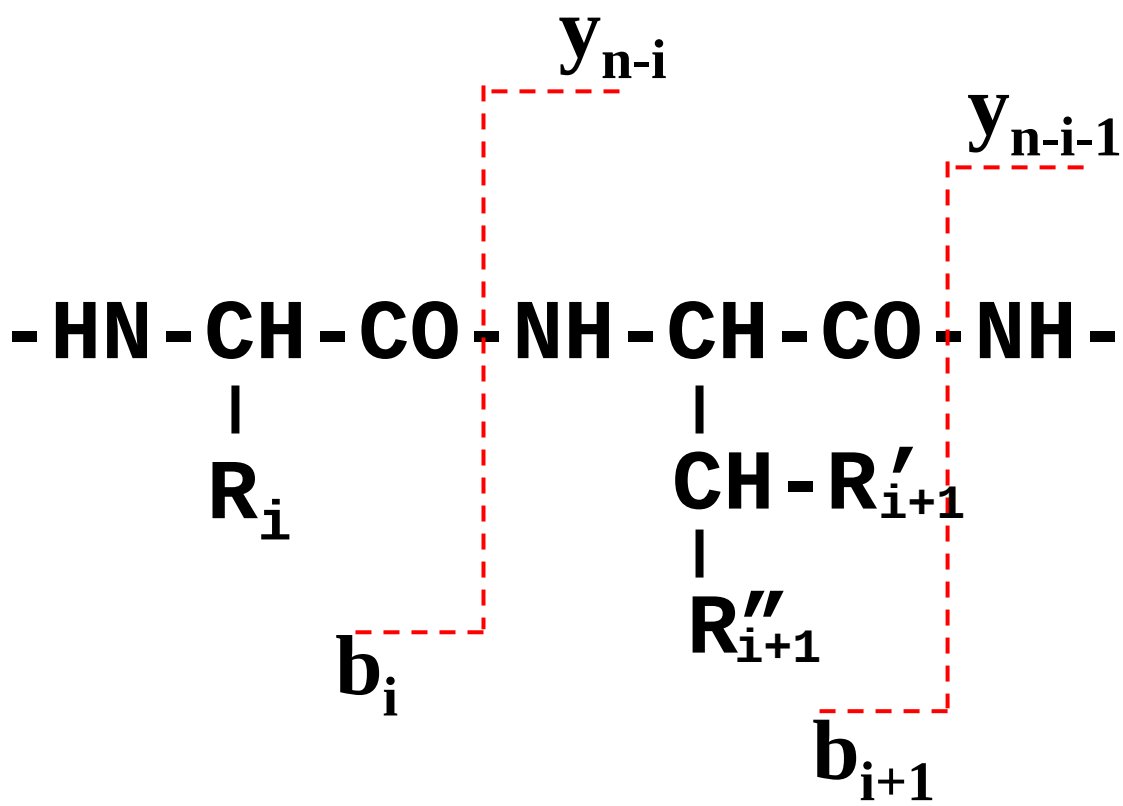
Peptide Fragmentation

Peptides consist of amino-acids arranged in a linear backbone.



Fragmented peptide
C-terminus fragment observed

Peptide Fragmentation



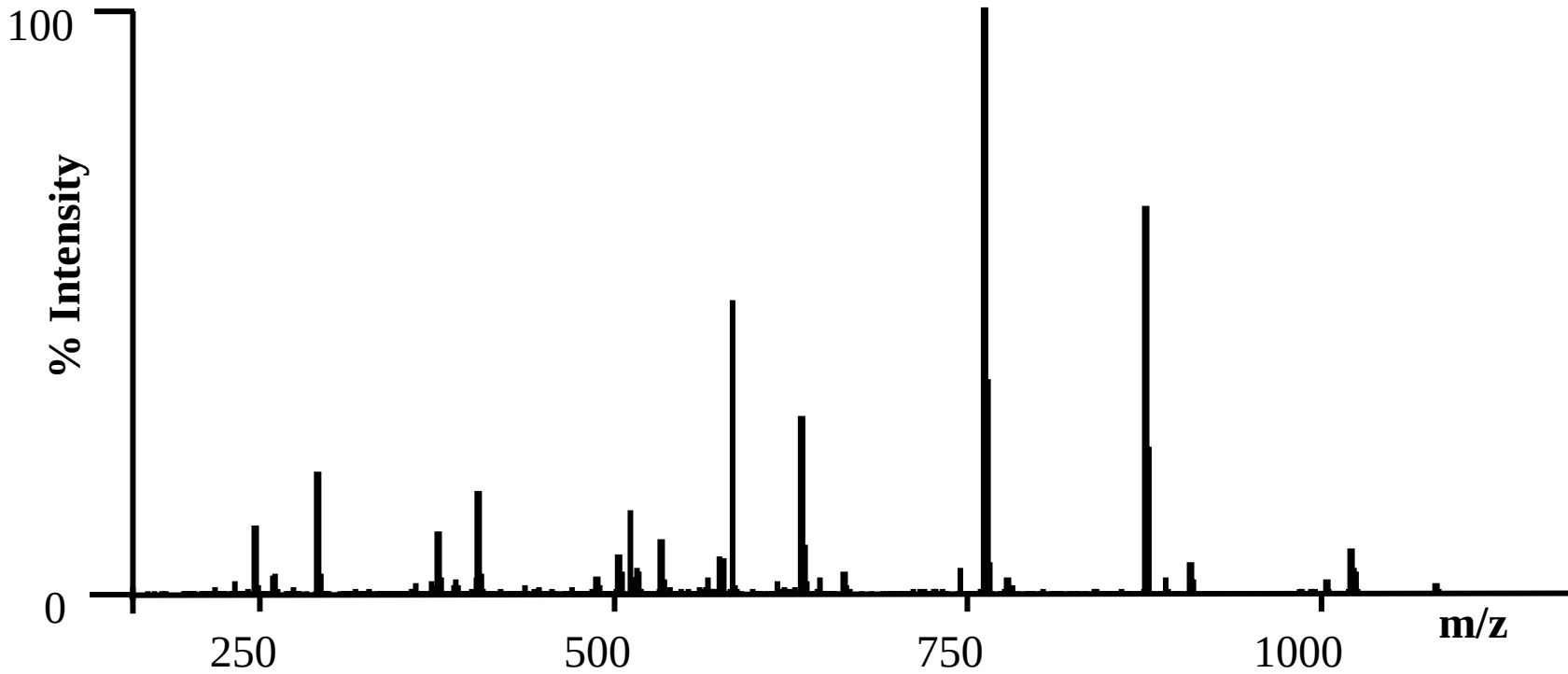
Peptide Fragmentation

Peptide: S-G-F-L-E-E-D-E-L-K

MW	ion			ion	MW
88	b_1	S	GFLEEDELK	y_9	1080
145	b_2	SG	FLEEDELK	y_8	1022
292	b_3	SGF	LEEDELK	y_7	875
405	b_4	SGFL	EEDELK	y_6	762
534	b_5	SGFLE	EDELK	y_5	633
663	b_6	SGFLEE	DELK	y_4	504
778	b_7	SGFLEED	ELK	y_3	389
907	b_8	SGFLEEDE	LK	y_2	260
1020	b_9	SGFLEEDEL	K	y_1	147

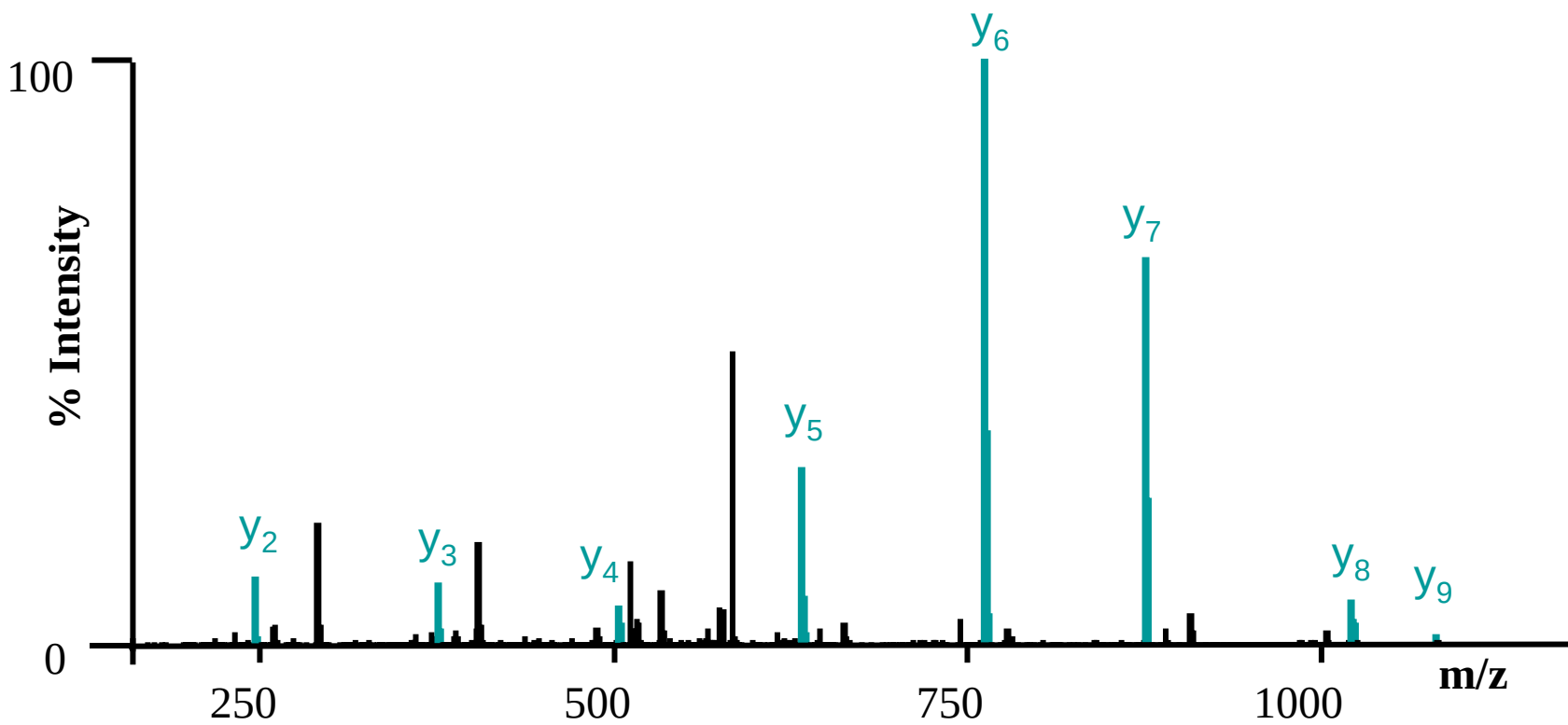
Peptide Fragmentation

88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



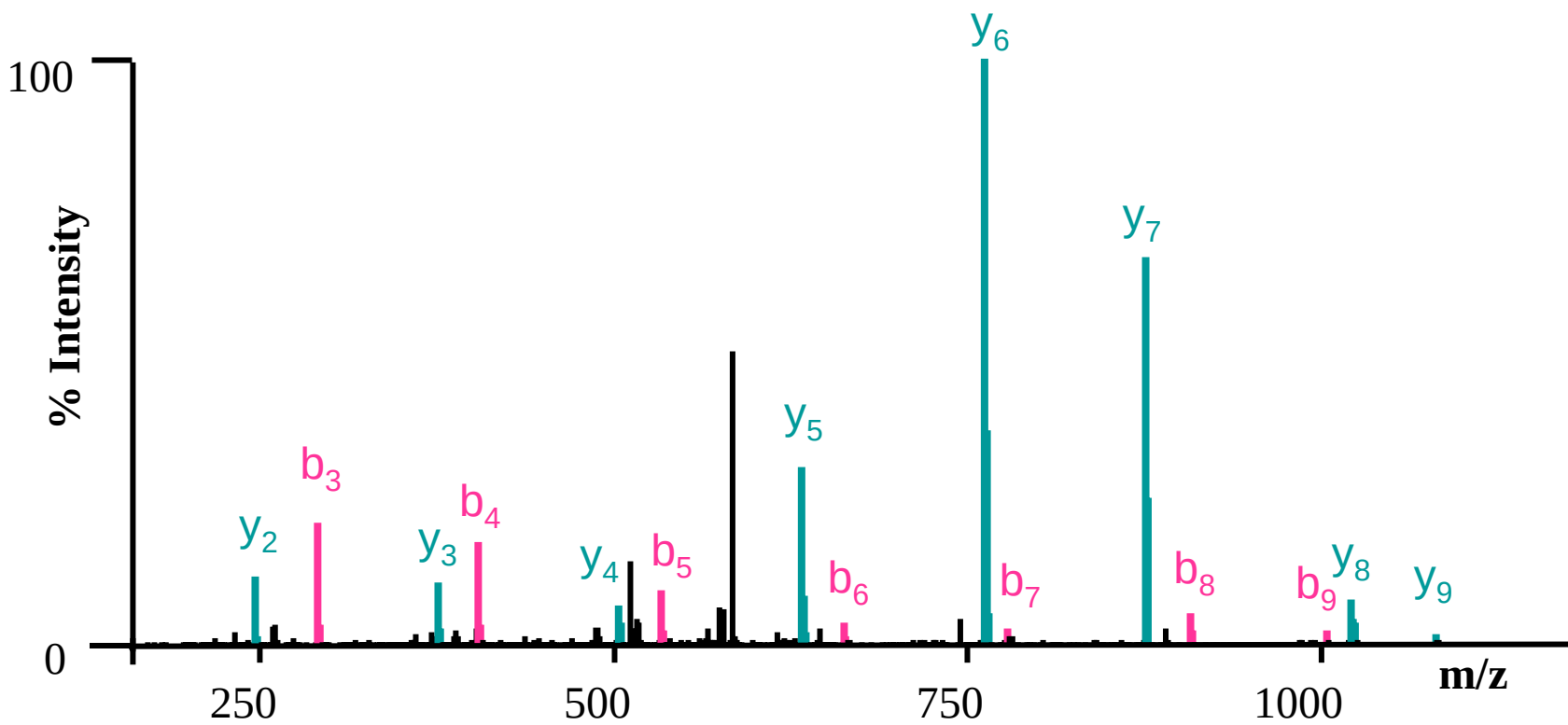
Peptide Fragmentation

88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



Peptide Fragmentation

88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



Peptide Identification

Given:

- The mass of the parent ion, and
- The MS/MS spectrum

Output:

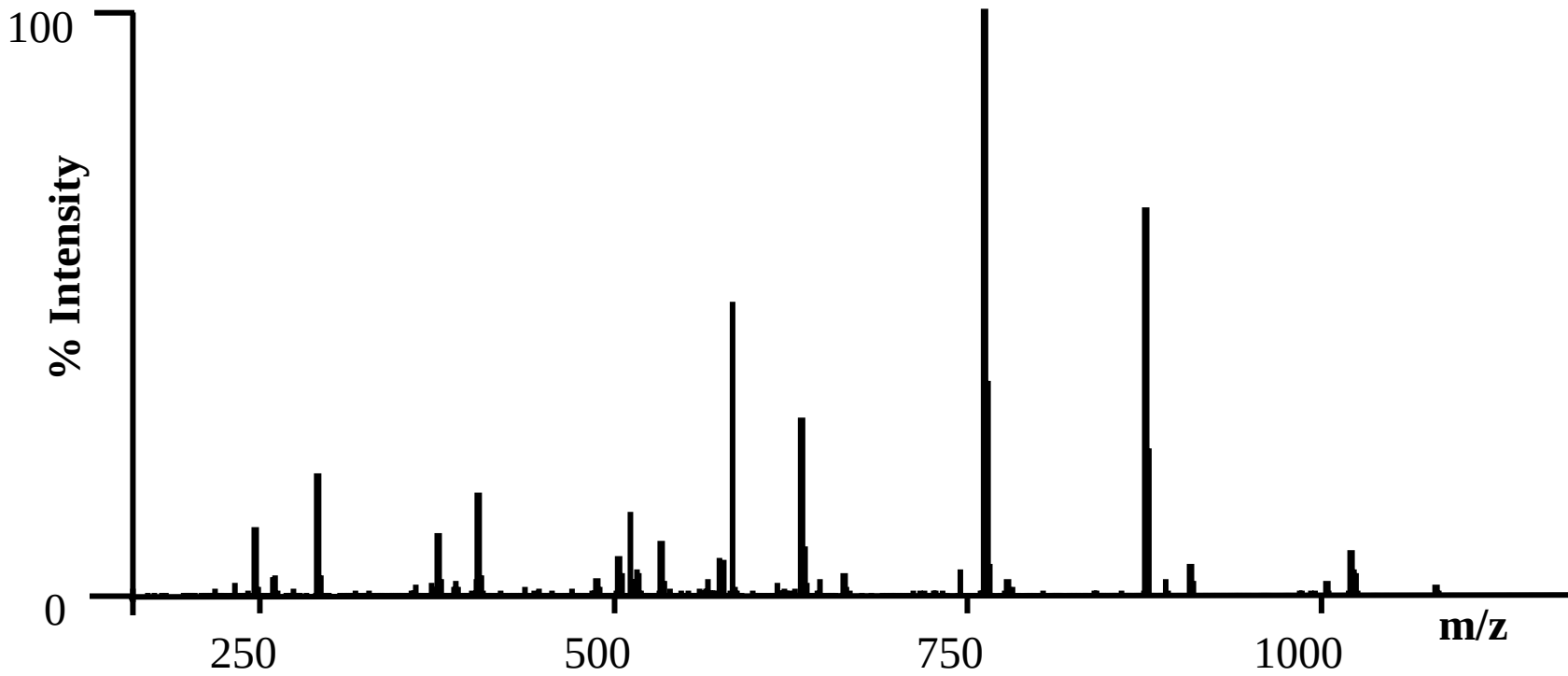
- The amino-acid sequence of the peptide

Peptide Identification

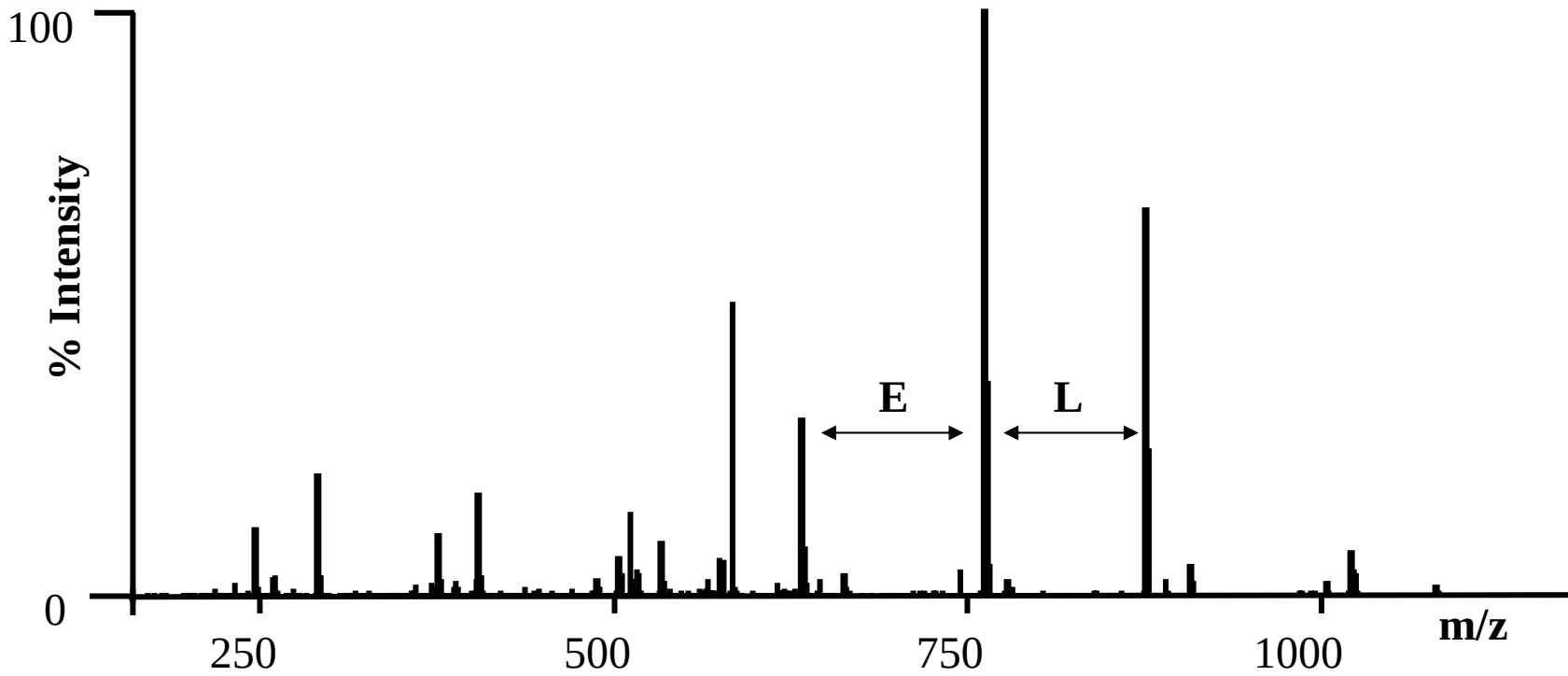
Two paradigms:

- *De novo* interpretation
- Sequence database search

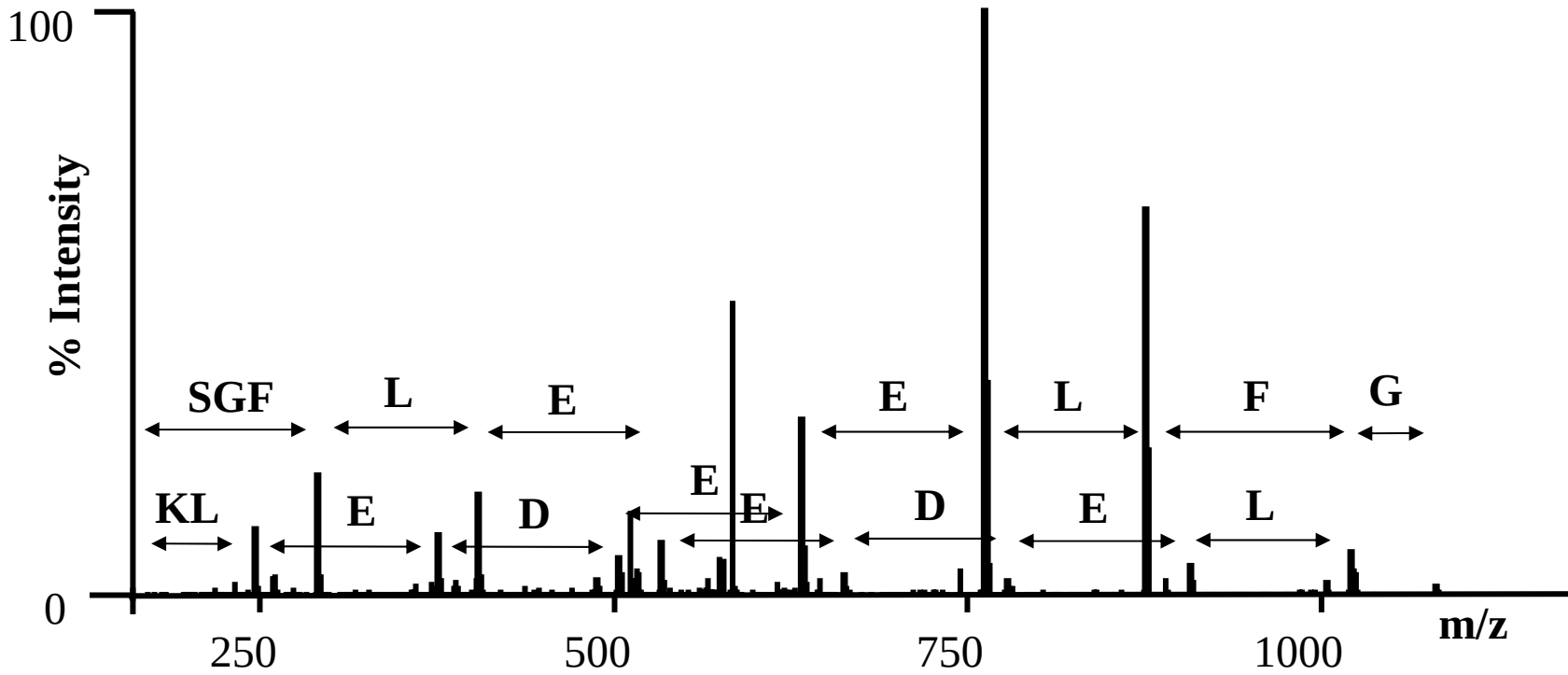
De Novo Interpretation



De Novo Interpretation



De Novo Interpretation



De Novo Interpretation

- Amino-acids have duplicate masses!
- Incomplete ladders create ambiguity.
- Noise peaks and unmodeled fragments create ambiguity
- “Best” *de novo* interpretation may have no biological relevance
- Current algorithms cannot model many aspects of peptide fragmentation
- Identifies relatively few peptides in high-throughput workflows

De Novo Interpretation

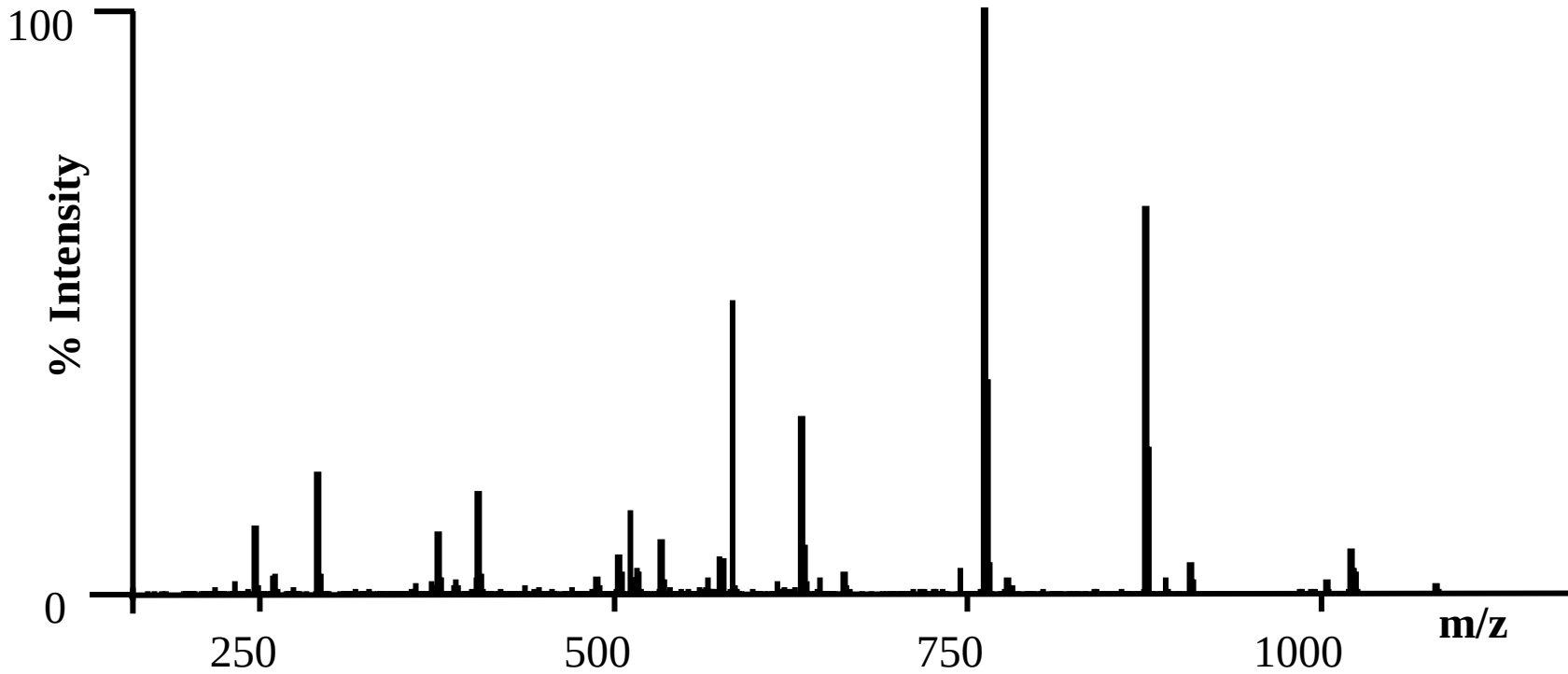
	Amino-Acid	Residual MW		Amino-Acid	Residual MW
A	Alanine	71.03712	M	Methionine	131.04049
C	Cysteine	103.00919	N	Asparagine	114.04293
D	Aspartic acid	115.02695	P	Proline	97.05277
E	Glutamic acid	129.04260	Q	Glutamine	128.05858
F	Phenylalanine	147.06842	R	Arginine	156.10112
G	Glycine	57.02147	S	Serine	87.03203
H	Histidine	137.05891	T	Threonine	101.04768
I	Isoleucine	113.08407	V	Valine	99.06842
K	Lysine	128.09497	W	Tryptophan	186.07932
L	Leucine	113.08407	Y	Tyrosine	163.06333

Sequence Database Search

- Compares peptides from a protein sequence database with spectra
- Filter peptide candidates by
  - Parent mass
  - Digest motif
- Score each peptide against spectrum
  - Generate all possible peptide fragments
  - Match putative fragments with peaks
  - Score and rank

Sequence Database Search

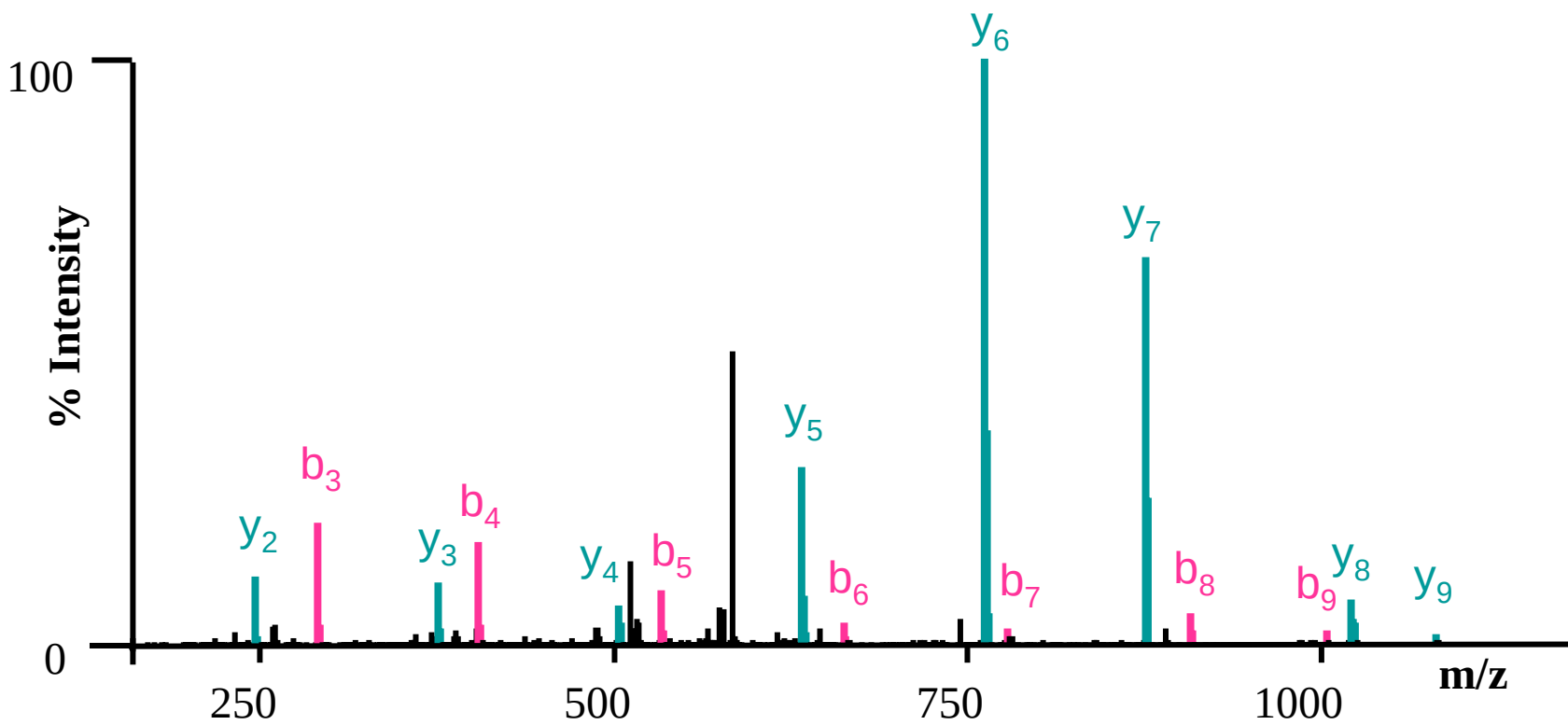
88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



Applied Biosystems

Sequence Database Search

88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



Sequence Database Search

- No need for complete ladders
- Possible to model all known peptide fragments
- Sequence permutations eliminated
- All candidates have *some* biological relevance
- Practical for high-throughput peptide identification
- Correct peptide might be missing from database!

Peptide Candidate Filtering

Digestion Enzyme: Trypsin

- Cuts just after K or R unless followed by a P.
- Basic residues (K & R) at C-terminal attract ionizing charge, leading to strong y-ions
- “Average” peptide length about 10-15 amino-acids
- Must allow for “missed” cleavage sites

Peptide Candidate Filtering

>ALBU_HUMAN
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKA
LVLIAFAQYLQQCPFEDHVKLVNEVTEFAK…

No missed cleavage sites

MKWVTFISLLFLFSSAYSRGVFR
R
DAHK
SEVAHR
FK
DLGEENFK
ALVLIAFAQYLQQCPFEDHVK
LVNEVTEFAK
…

Peptide Candidate Filtering

>ALBU_HUMAN
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKA
LVLIAFAQYLQQCPFEDHVKLVNEVTEFAK…

One missed cleavage site

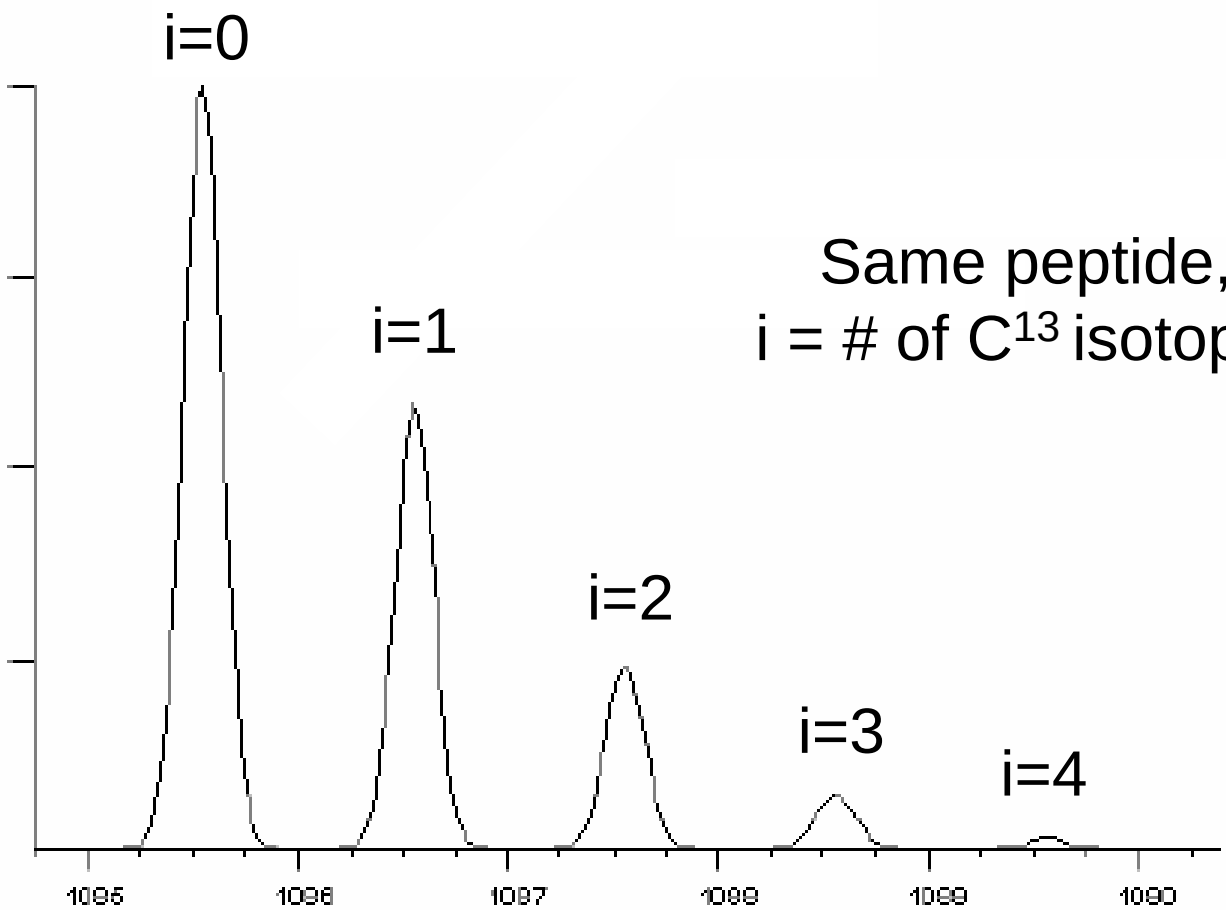
MKWVTFISLLFLFSSAYSRGVFRR
RDAHK
DAHKSEVAHR
SEVAHRFK
FKDLGEENFK
DLGEENFKALVLIAFAQYLQQCPFEDHVK
ALVLIAFAQYLQQCPFEDHVKLVNEVTEFAK
…

Peptide Candidate Filtering

Peptide molecular weight

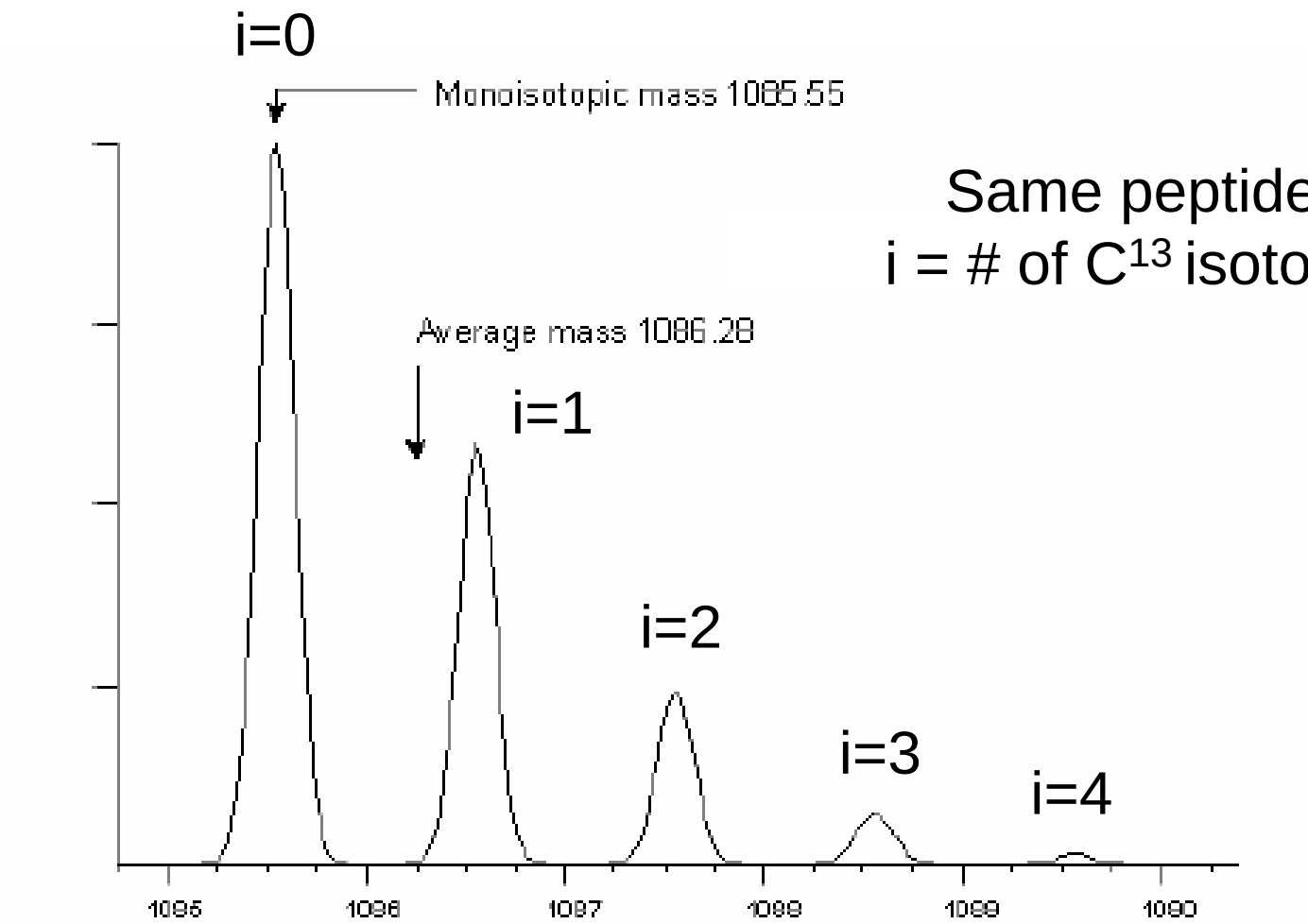
- Only have m/z value
  - Need to determine charge state
- Ion selection tolerance
- Mass for each amino-acid symbol?
  - Monoisotopic vs. Average
  - “Default” residual mass
  - Depends on sample preparation protocol
  - Cysteine almost always modified

Peptide Molecular Weight



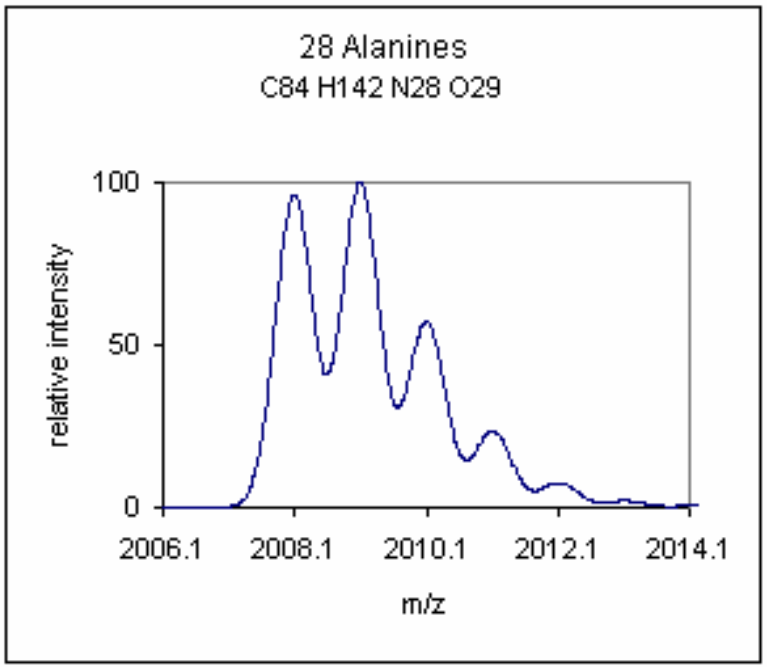
Same peptide,
i = # of C^{13} isotope

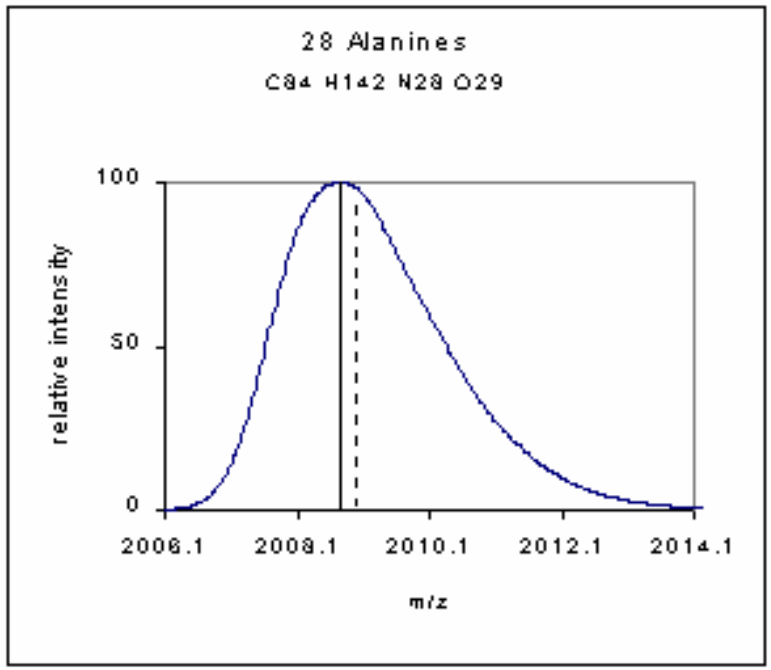
Peptide Molecular Weight



Same peptide,
i = # of C^{13} isotope

Peptide Molecular Weight



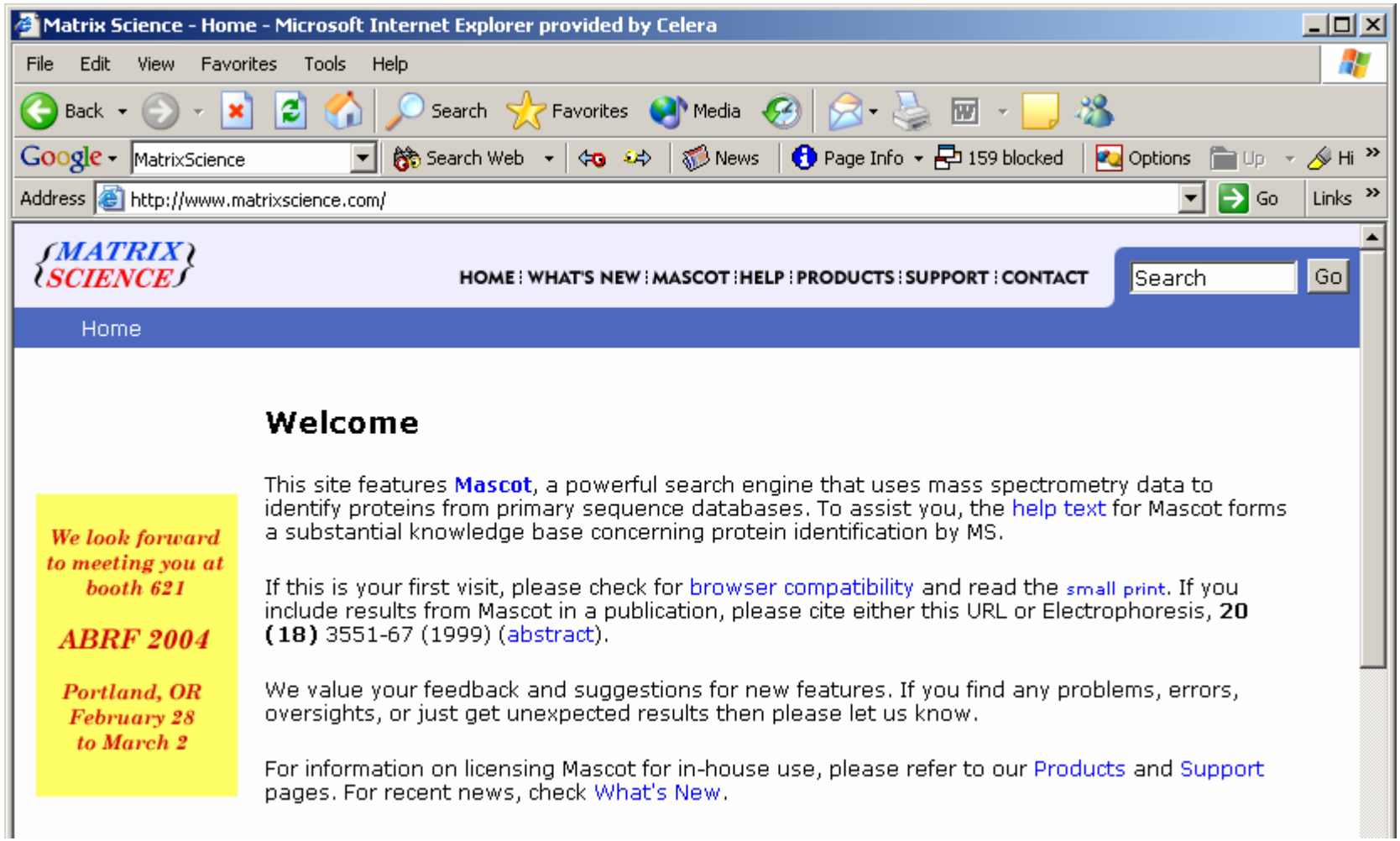


…from “Isotopes” – An IonSource.Com Tutorial

Peptide Scoring

- Peptide fragments vary based on
  - The instrument
  - The peptide's amino-acid sequence
  - The peptide's charge state
  - Etc…
- Search engines model peptide fragmentation to various degrees.
  - Speed vs. sensitivity tradeoff
  - y-ions & b-ions occur most frequently

Mascot Search Engine



Mascot MS/MS Ions Search

Your name		**Email**	
Search title			
Database	MSDB		
Taxonomy	All entries		
Enzyme	Trypsin	**Allow up to**	1 missed cleavages
Fixed modifications	AB_old_ICATd0 (C) AB_old_ICATd8 (C) Acetyl (K) Acetyl (N-term) Amide (C-term)	**Variable modifications**	AB_old_ICATd0 (C) AB_old_ICATd8 (C) Acetyl (K) Acetyl (N-term) Amide (C-term)
Protein mass	kDa	**ICAT**	☐
Peptide tol. ±	2.0 Da	**MS/MS tol. ±**	0.8 Da
Peptide charge	2+	**Monoisotopic**	◉ **Average** ○
Data file	Browse...		
Data format	Mascot generic	**Precursor**	m/z
Instrument	Default		
Overview	☐	**Report top**	20 hits
	Start Search ...		**Reset Form**

Mascot MS/MS Search Results

1. Q9XZJ2 **Mass:** 79480 **Score:** 286 **Peptides matched:** 4

HEAT SHOCK PROTEIN 70.- Crassostrea gigas (Pacific oyster).

☐ Check to include this hit in error tolerant search

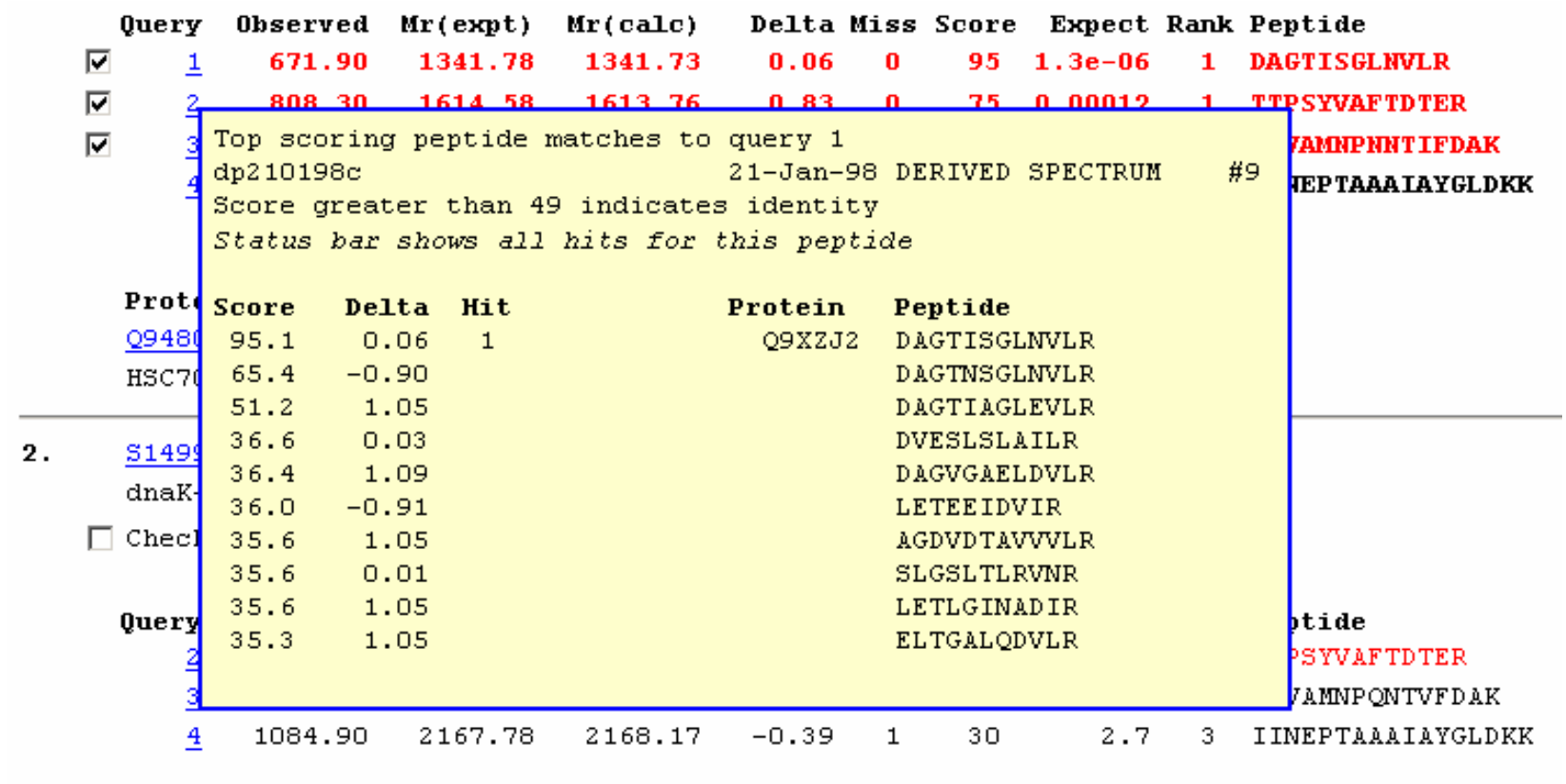
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑ 1	671.90	1341.78	1341.73	0.06	0	95	1.3e-06	1	DAGTISGLNVLR
☑ 2	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TTPSYVAFTDTER
☑ 3	973.90	1945.78	1945.92	-0.14	0	90	2.9e-06	1	NQVAMNPNNTIFDAK
4	1084.90	2167.78	2168.17	-0.39	1	30	2.7	3	IINEPTAAAIAYGLDKK

Proteins matching the same set of peptides:

Q94805 **Mass:** 79333 **Score:** 283 **Peptides matched:** 4

HSC70.- Trichoplusia ni (Cabbage looper).

Mascot MS/MS Search Results



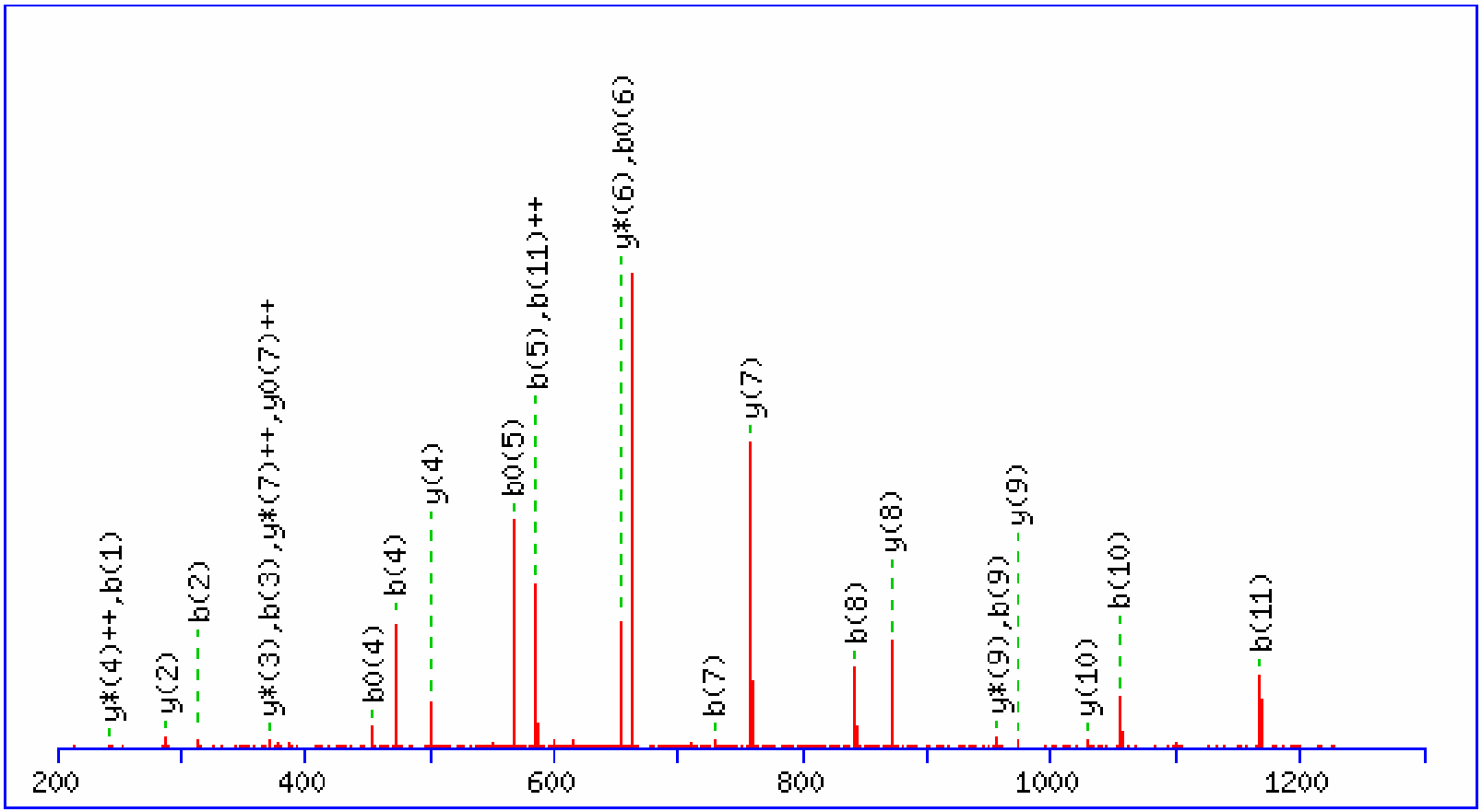
Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
1	671.90	1341.78	1341.73	0.06	0	95	1.3e-06	1	DAGTISGLNVLR
2	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TTPSYVAFTDTER
3									VAMNPNNTIFDAK
4									NEPTAAAIAYGLDKK

Top scoring peptide matches to query 1
dp210198c 21-Jan-98 DERIVED SPECTRUM #9
Score greater than 49 indicates identity
Status bar shows all hits for this peptide

Score	Delta	Hit	Protein	Peptide
95.1	0.06	1	Q9XZJ2	DAGTISGLNVLR
65.4	-0.90			DAGTNSGLNVLR
51.2	1.05			DAGTIAGLEVLR
36.6	0.03			DVESLSLAILR
36.4	1.09			DAGVGAELDVLR
36.0	-0.91			LETEEIDVIR
35.6	1.05			AGDVDTAVVVLR
35.6	0.01			SLGSLTLRVNR
35.6	1.05			LETLGINADIR
35.3	1.05			ELTGALQDVLR

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
2									PSYVAFTDTER
3									VAMNPQNTVFDAK
4	1084.90	2167.78	2168.17	-0.39	1	30	2.7	3	IINEPTAAAIAYGLDKK

Mascot MS/MS Search Results



Mascot MS/MS Search Results

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	243.10	122.05			225.09	113.05	D							12
2	314.13	157.57			296.12	148.57	A	1100.64	550.82	1083.62	542.31	1082.63	541.82	11
3	371.16	186.08			353.15	177.08	G	1029.61	515.31	1012.58	506.79	1011.59	506.30	10
4	472.20	236.61			454.19	227.60	T	972.58	486.80	955.56	478.28	954.57	477.79	9
5	585.29	293.15			567.28	284.14	I	871.54	436.27	854.51	427.76	853.53	427.27	8
6	672.32	336.66			654.31	327.66	S	758.45	379.73	741.43	371.22	740.44	370.72	7
7	729.34	365.17			711.33	356.17	G	671.42	336.21	654.39	327.70			6
8	842.43	421.72			824.41	412.71	L	614.40	307.70	597.37	299.19			5
9	956.47	478.74	939.44	470.22	938.46	469.73	N	501.31	251.16	484.29	242.65			4
10	1055.54	528.27	1038.51	519.76	1037.53	519.27	V	387.27	194.14	370.24	185.63			3
11	1168.62	584.81	1151.59	576.30	1150.61	575.81	L	288.20	144.61	271.18	136.09			2
12							R	175.12	88.06	158.09	79.55			1

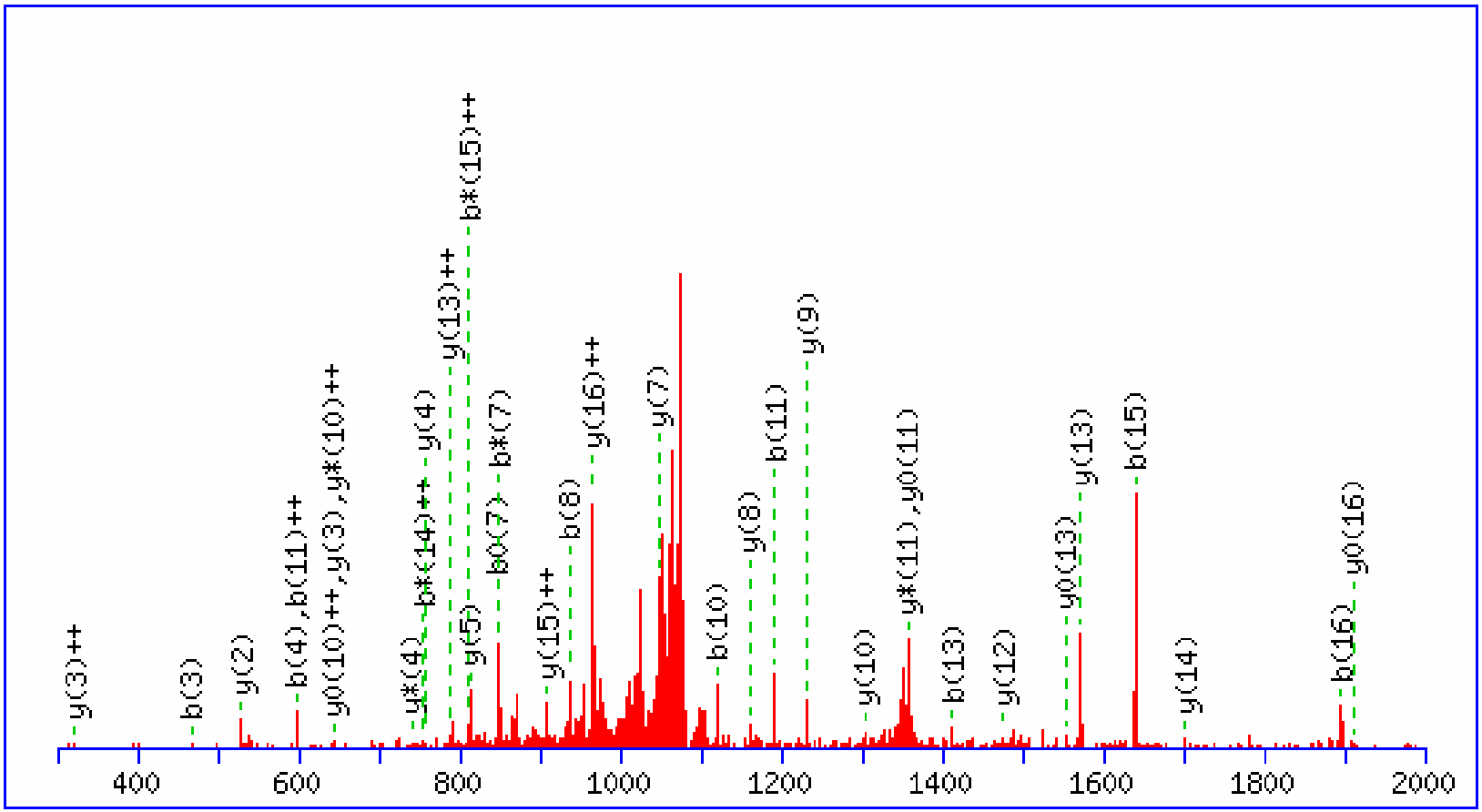
Mascot MS/MS Search Results

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
☑	1	671.90	1341.78	1341.73	0.06	0	95	1.3e-06	1	DAGTISGLNVLR
☑	2	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TTPSYVAFTDTER
☑	3	973.90	1945.78	1945.92	-0.14	0	90	2.9e-06	1	NQVAMNPNNTIFDAK
	4	1084.90	2167.78	2168.17	-0.39	1	30	2.7	3	IINEPTAAAIAYGLDKK

Top scoring peptide matches to query 4
dp210198 21-Jan-98 DERIVED SPECTRUM #9
Score greater than 47 indicates identity
Status bar shows all hits for this peptide

Score	Delta	Hit	Protein	Peptide
39.1	0.57	5	Q95PU3	IIIEPTAAAIAYGLDKK
35.3	0.60			IINQPTAAAIAYGLDKK
30.1	-0.39	1+	Q9XZJ2	IINEPTAAAIAYGLDKK
30.1	-0.39			IINEPTAAALAYGLDKK
30.1	-0.39			IINEPTAAAIAYGIDKK
30.1	-0.39			ILNEPTAAAIAYGLDKK
30.1	-0.39			ILNEPTAAALAYGLDKK
26.4	-0.39			IINEPTAAALSFGLDKK
20.5	-0.26			IPQQWTPPCGSKCTNK
13.3	-0.26			TNCPNVSSFFQSNSVRVR

Mascot MS/MS Search Results



Mascot MS/MS Search Results

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	241.15	121.08					I							17
2	354.24	177.62					I	1929.03	965.02	1912.01	956.51	1911.02	956.01	16
3	468.28	234.64	451.26	226.13			N	1815.95	908.48	1798.92	899.96	1797.94	899.47	15
4	597.32	299.17	580.30	290.65	579.31	290.16	E	1701.91	851.46	1684.88	842.94	1683.90	842.45	14
5	694.38	347.69	677.35	339.18	676.37	338.69	P	1572.86	786.94	1555.84	778.42	1554.85	777.93	13
6	795.42	398.22	778.40	389.70	777.41	389.21	T	1475.81	738.41	1458.78	729.90	1457.80	729.40	12
7	866.46	433.73	849.44	425.22	848.45	424.73	A	1374.76	687.88	1357.74	679.37	1356.75	678.88	11
8	937.50	469.25	920.47	460.74	919.49	460.25	A	1303.73	652.37	1286.70	643.85	1285.72	643.36	10
9	1008.54	504.77	991.51	496.26	990.53	495.77	A	1232.69	616.85	1215.66	608.33	1214.68	607.84	9
10	1121.62	561.31	1104.59	552.80	1103.61	552.31	I	1161.65	581.33	1144.62	572.82	1143.64	572.32	8
11	1192.66	596.83	1175.63	588.32	1174.65	587.83	A	1048.57	524.79	1031.54	516.27	1030.56	515.78	7
12	1355.72	678.36	1338.69	669.85	1337.71	669.36	Y	977.53	489.27	960.50	480.76	959.52	480.26	6
13	1412.74	706.87	1395.72	698.36	1394.73	697.87	G	814.47	407.74	797.44	399.22	796.46	398.73	5
14	1525.83	763.42	1508.80	754.90	1507.82	754.41	L	757.45	379.23	740.42	370.71	739.43	370.22	4
15	1640.85	820.93	1623.83	812.42	1622.84	811.92	D	644.36	322.68	627.33	314.17	626.35	313.68	3
16	1896.01	948.51	1878.98	940.00	1878.00	939.50	K	529.33	265.17	512.31	256.66			2
17							K	274.18	137.59	257.15	129.08			1

Mascot MS/MS Search Results

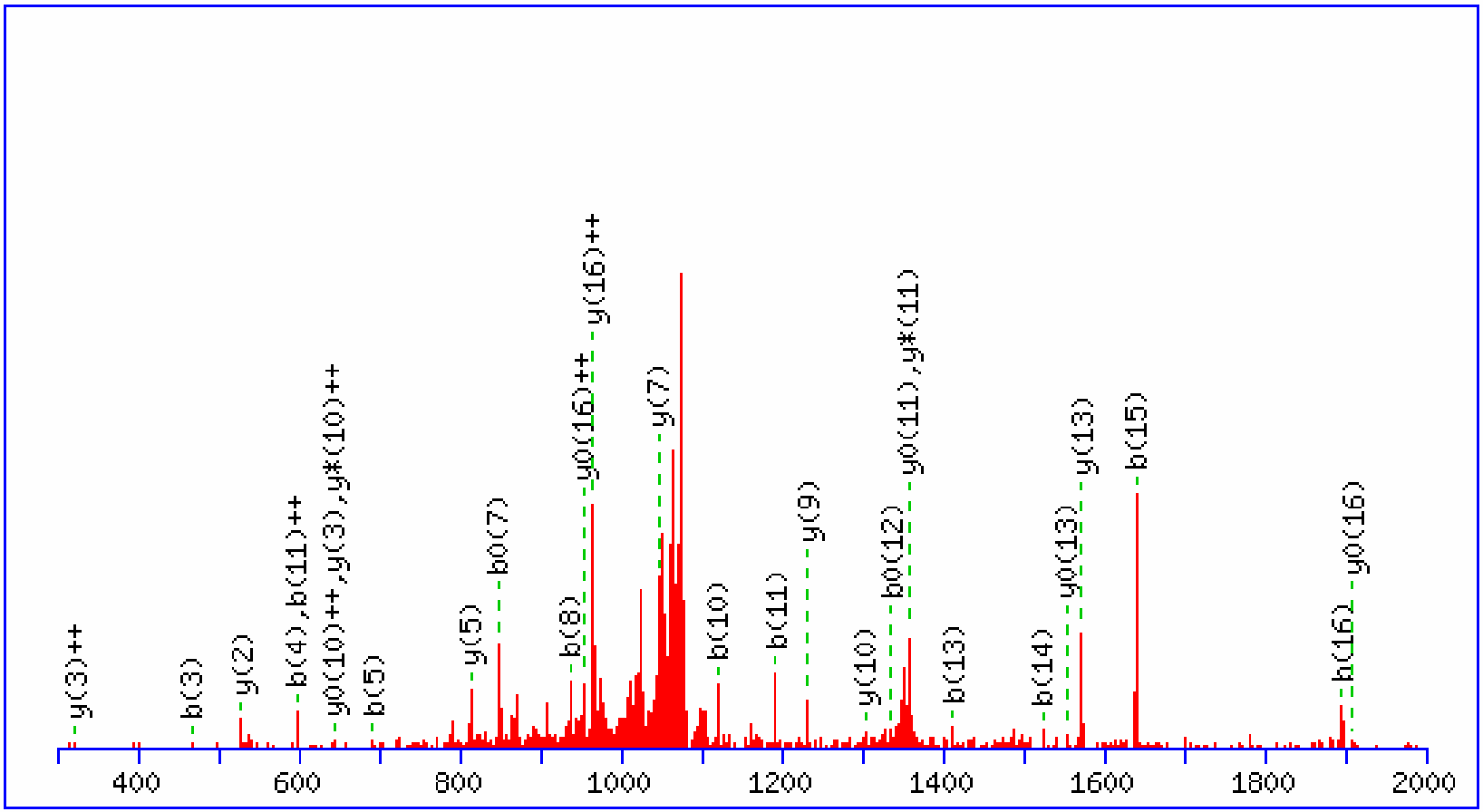
5. Q95PU3 **Mass:** 80430 **Score:** 114 **Peptides matched:** 2

HEAT SHOCK PROTEIN (HSP70).- Euplotes crassus.

☐ Check to include this hit in error tolerant search

	Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
	2	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TTPSYVAFTDTER
☑	4	1084.90	2167.78	2167.21	0.57	1	39	0.34	1	IIIEPTAAAIAYGLDKK

Mascot MS/MS Search Results



Mascot MS/MS Search Results

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	241.15	121.08					I							17
2	354.24	177.62					I	1928.07	964.54	1911.05	956.03	1910.06	955.54	16
3	467.32	234.17					I	1814.99	908.00	1797.96	899.49	1796.98	898.99	15
4	596.37	298.69			578.35	289.68	E	1701.91	851.46	1684.88	842.94	1683.90	842.45	14
5	693.42	347.21			675.41	338.21	P	1572.86	786.94	1555.84	778.42	1554.85	777.93	13
6	794.47	397.74			776.46	388.73	T	1475.81	738.41	1458.78	729.90	1457.80	729.40	12
7	865.50	433.26			847.49	424.25	A	1374.76	687.88	1357.74	679.37	1356.75	678.88	11
8	936.54	468.77			918.53	459.77	A	1303.73	652.37	1286.70	643.85	1285.72	643.36	10
9	1007.58	504.29			989.57	495.29	A	1232.69	616.85	1215.66	608.33	1214.68	607.84	9
10	1120.66	560.83			1102.65	551.83	I	1161.65	581.33	1144.62	572.82	1143.64	572.32	8
11	1191.70	596.35			1173.69	587.35	A	1048.57	524.79	1031.54	516.27	1030.56	515.78	7
12	1354.76	677.88			1336.75	668.88	Y	977.53	489.27	960.50	480.76	959.52	480.26	6
13	1411.78	706.40			1393.77	697.39	G	814.47	407.74	797.44	399.22	796.46	398.73	5
14	1524.87	762.94			1506.86	753.93	L	757.45	379.23	740.42	370.71	739.43	370.22	4
15	1639.89	820.45			1621.88	811.45	D	644.36	322.68	627.33	314.17	626.35	313.68	3
16	1895.05	948.03	1878.03	939.52	1877.04	939.02	K	529.33	265.17	512.31	256.66			2
17							K	274.18	137.59	257.15	129.08			1

Sequence Database Search Traps and Pitfalls

Search options may eliminate the correct peptide

- Parent mass tolerance too small
- Fragment m/z tolerance too small
- Incorrect parent ion charge state
- Non-tryptic or semi-tryptic peptide
- Incorrect or unexpected modification
- Sequence database too conservative
- Unreliable taxonomy annotation

Sequence Database Search Traps and Pitfalls

Search options can cause infinite search times

- Variable modifications increase search times exponentially
- Non-tryptic search increases search time by two orders of magnitude
- Large sequence databases contain many irrelevant peptide candidates

Sequence Database Search Traps and Pitfalls

Best available peptide isn't necessarily correct!

- Score statistics (e-values) are essential!
  - What is the chance a peptide could score this well by chance alone?
- The wrong peptide can *look* correct if the right peptide is missing!
- Need scores (or e-values) that are invariant to spectrum quality and peptide properties

Sequence Database Search Traps and Pitfalls

Search engines often make incorrect assumptions about sample prep

- Proteins with lots of identified peptides are not more likely to be present
- Peptide identifications do not represent independent observations
- All proteins are not equally interesting to report

Sequence Database Search Traps and Pitfalls

Good spectral processing can make a big difference

- Poorly calibrated spectra require large m/z tolerances
- Poorly baselined spectra make small peaks hard to believe
- Poorly de-isotoped spectra have extra peaks and misleading charge state assignments

Summary

- Protein identification from tandem mass spectra is a key proteomics technology.
- Protein identifications should be treated with healthy skepticism.
  - Look at *all* the evidence!
- Spectra remain unidentified for a variety of reasons.
- Lots of open algorithmic problems!

Further Reading

- Matrix Science (Mascot) Web Site
  - www.matrixscience.com
- Seattle Proteome Center (ISB)
  - www.proteomecenter.org
- *Proteomic Mass Spectrometry Lab* at The Scripps Research Institute
  - fields.scripps.edu
- UCSF ProteinProspector
  - prospector.ucsf.edu