



# 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

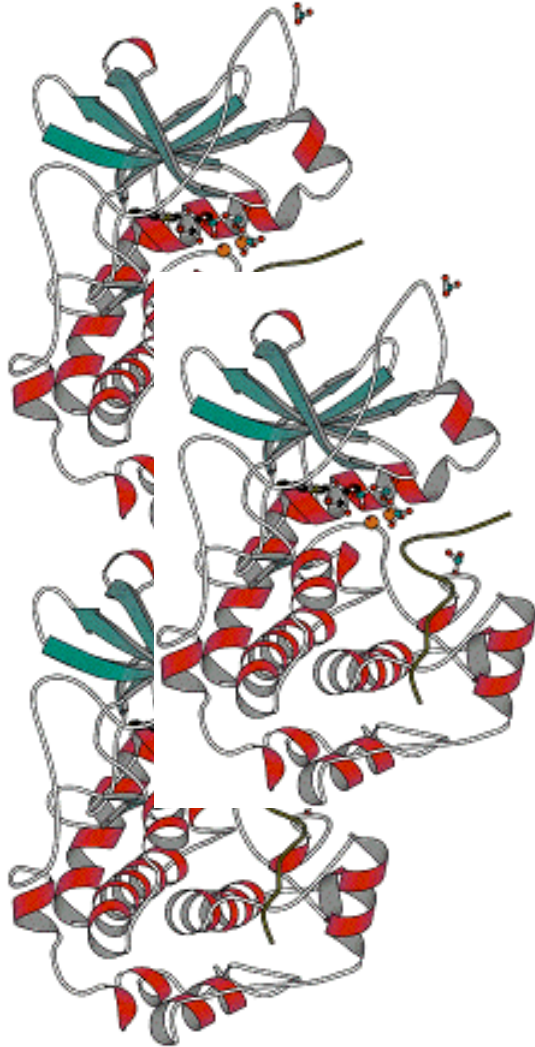
High-throughput proteomics focus

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

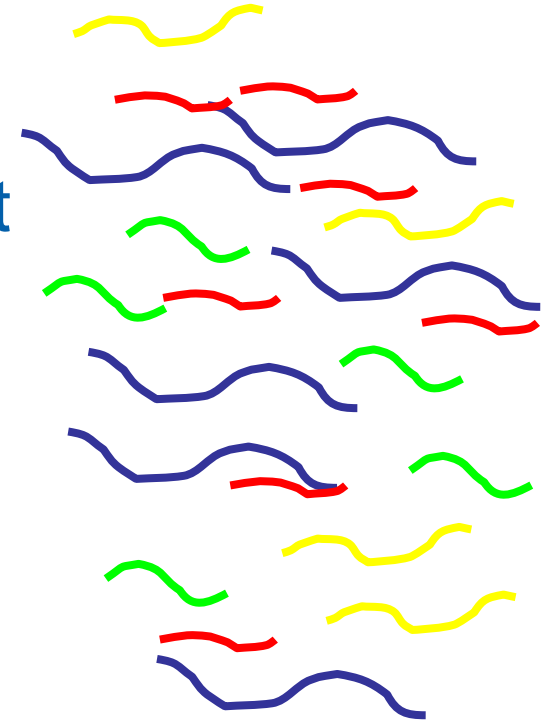
Two established workflows

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

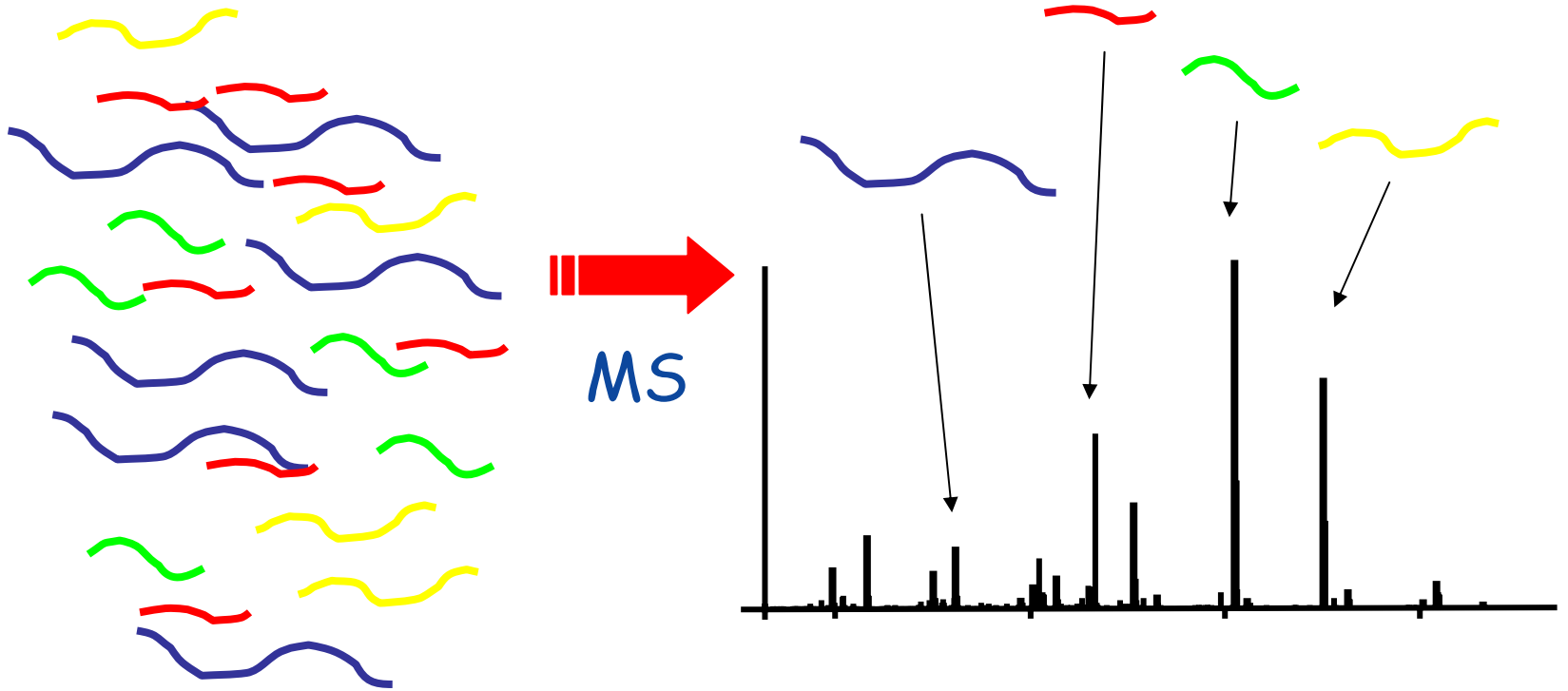
# Sample Preparation for Tandem Mass Spectrometry



Enzymatic Digest  
and  
Fractionation



# Single Stage MS

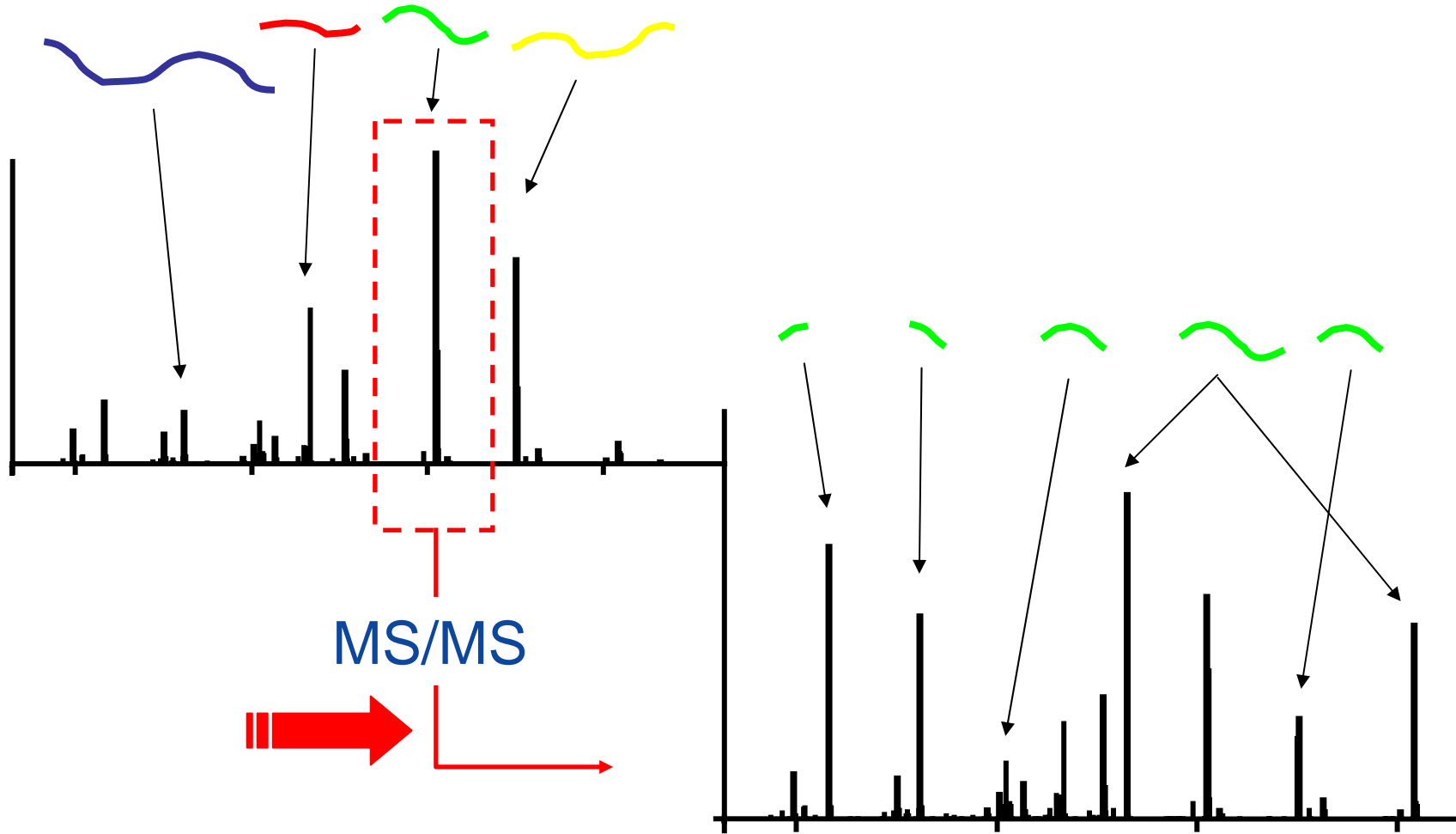




# Tandem Mass Spectrometry (MS/MS)

- 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

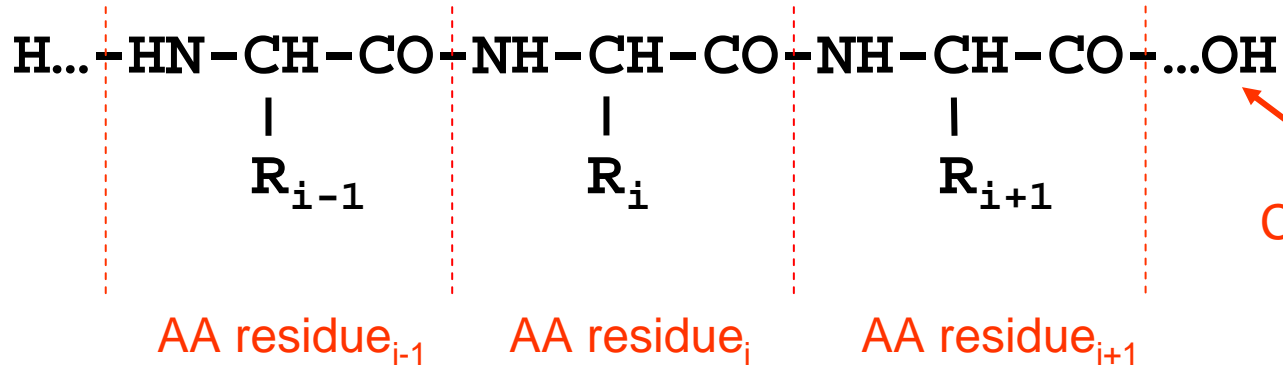
# Tandem Mass Spectrometry (MS/MS)



# Peptide Fragmentation

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

N-terminus



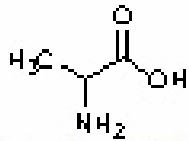
C-terminus



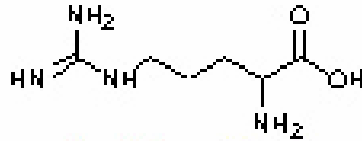
# Peptide Fragmentation

Schematic diagrams of the 20 amino acids

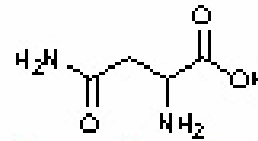
(picture taken from [www.chemistry.pomona.edu](http://www.chemistry.pomona.edu))



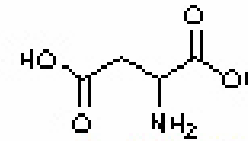
Alanine (Ala)



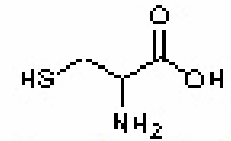
Arginine (Arg)



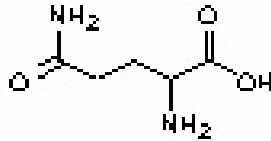
Asparagine (Asn)



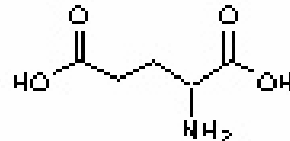
Aspartic Acid (Asp)



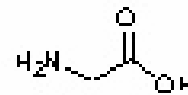
Cysteine (Cys)



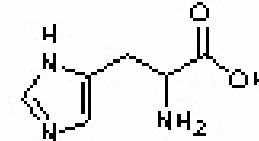
Glutamine (Gln)



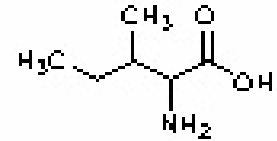
Glutamic Acid (Glu)



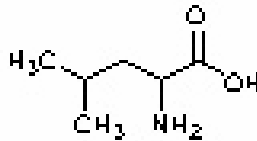
Glycine (Gly)



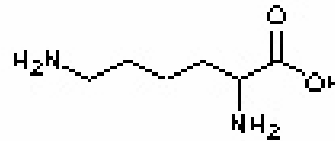
Histidine (His)



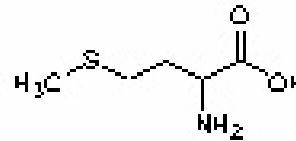
Isoleucine (Ile)



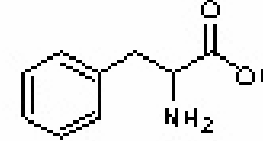
Leucine (Leu)



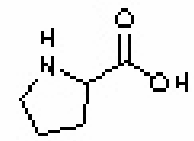
Lysine (Lys)



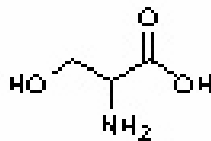
Methionine (Met)



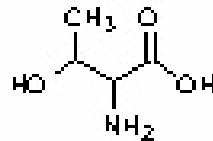
Phenylalanine (Phe)



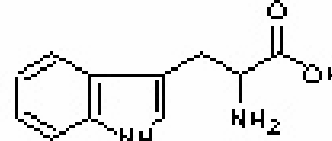
Proline (Pro)



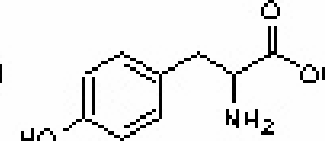
Serine (Ser)



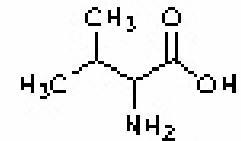
Threonine (Thr)



Tryptophan (Trp)



Tyrosine (Tyr)

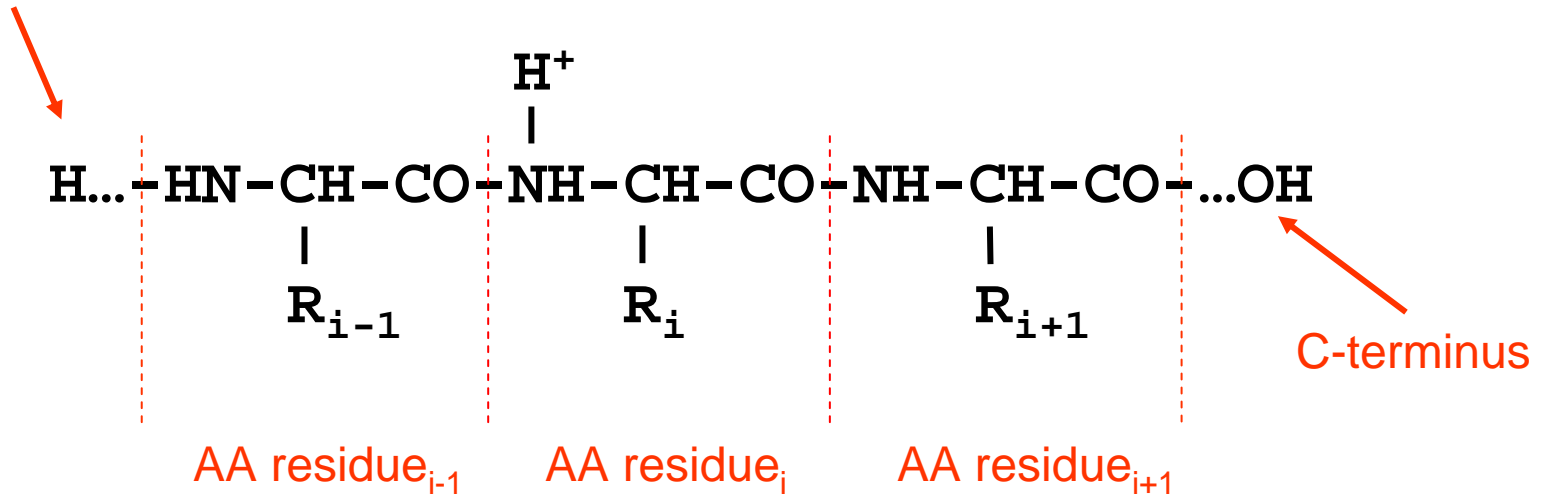


Valine (Val)

# Peptide Fragmentation

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

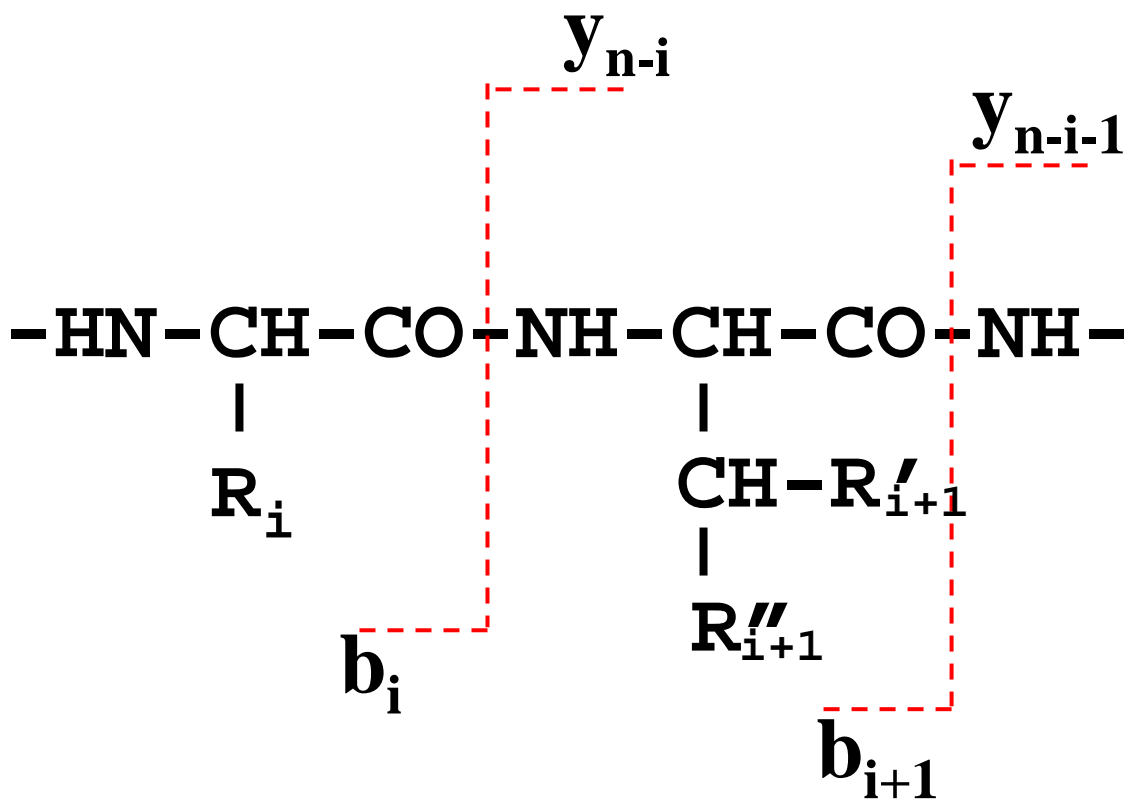
N-terminus



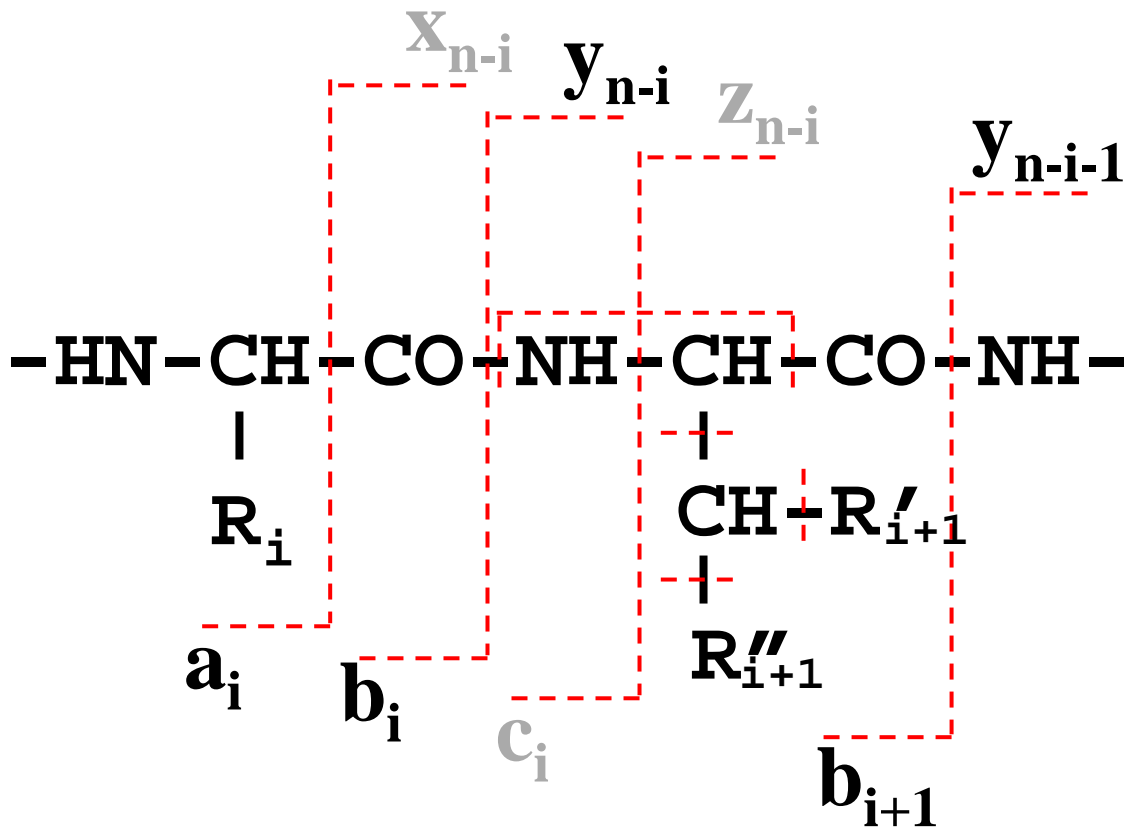
Ionized peptide (addition of a proton)



# Peptide Fragmentation



# Peptide Fragmentation



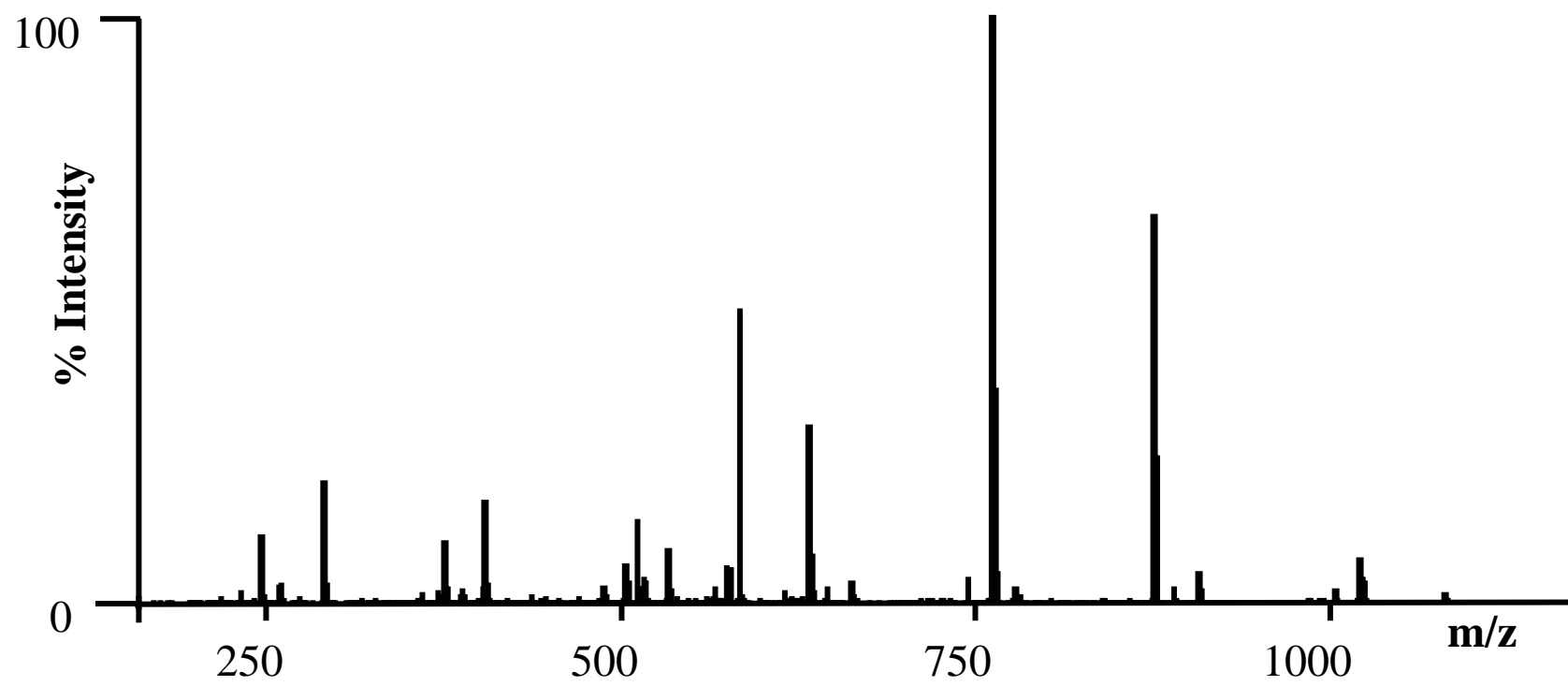
# Peptide Fragmentation

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

MW	ion			ion	MW
88	b <sub>1</sub>	S	GFLEEDELK	y <sub>9</sub>	1080
145	b <sub>2</sub>	SG	FLEEDELK	y <sub>8</sub>	1022
292	b <sub>3</sub>	SGF	LEEDELK	y <sub>7</sub>	875
405	b <sub>4</sub>	SGFL	EEDELK	y <sub>6</sub>	762
534	b <sub>5</sub>	SGFLE	EDELK	y <sub>5</sub>	633
663	b <sub>6</sub>	SGFLEE	DELK	y <sub>4</sub>	504
778	b <sub>7</sub>	SGFLEED	ELK	y <sub>3</sub>	389
907	b <sub>8</sub>	SGFLEEDE	LK	y <sub>2</sub>	260
1020	b <sub>9</sub>	SGFLEEDEL	K	y <sub>1</sub>	147

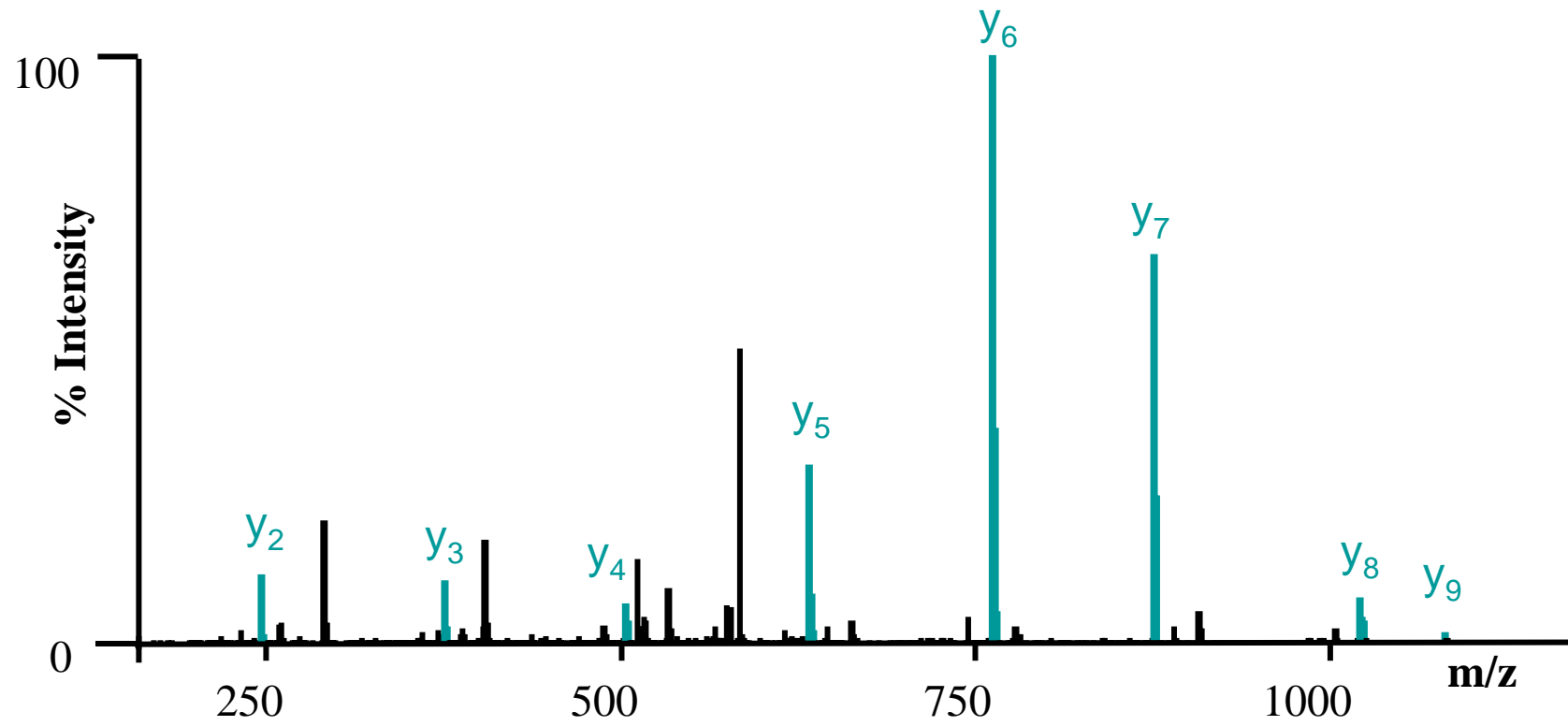
# Peptide Fragmentation

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



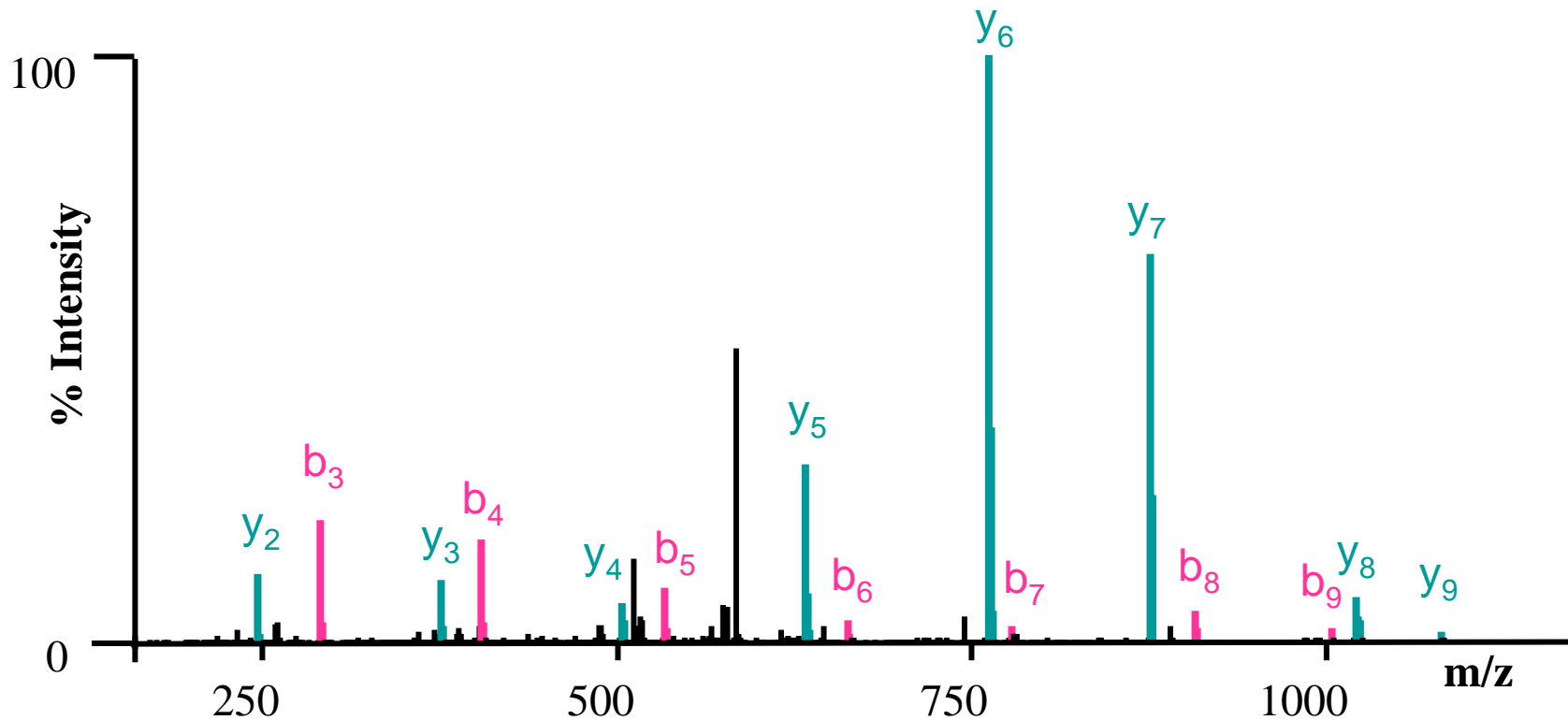
# Peptide Fragmentation

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



# Peptide Fragmentation

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



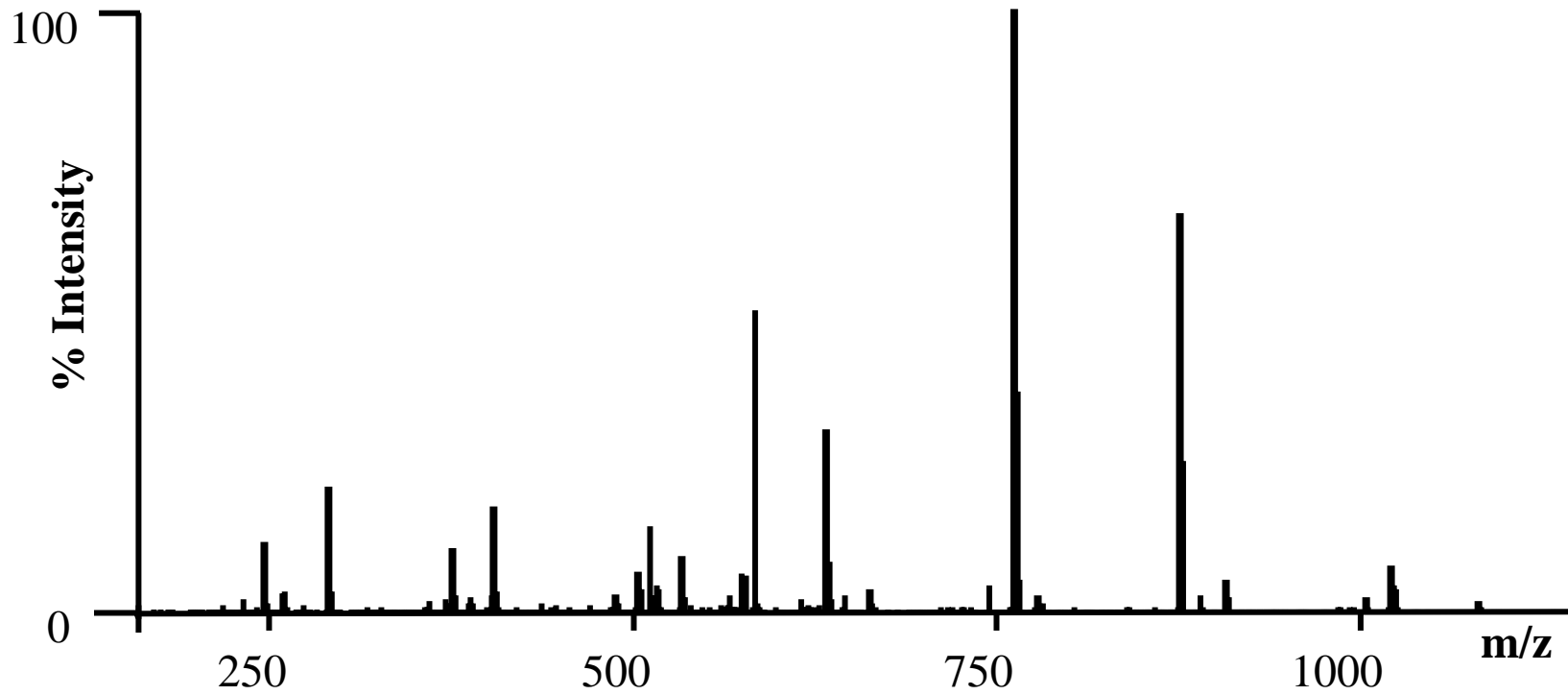
# Peptide Identification

Two paradigms:

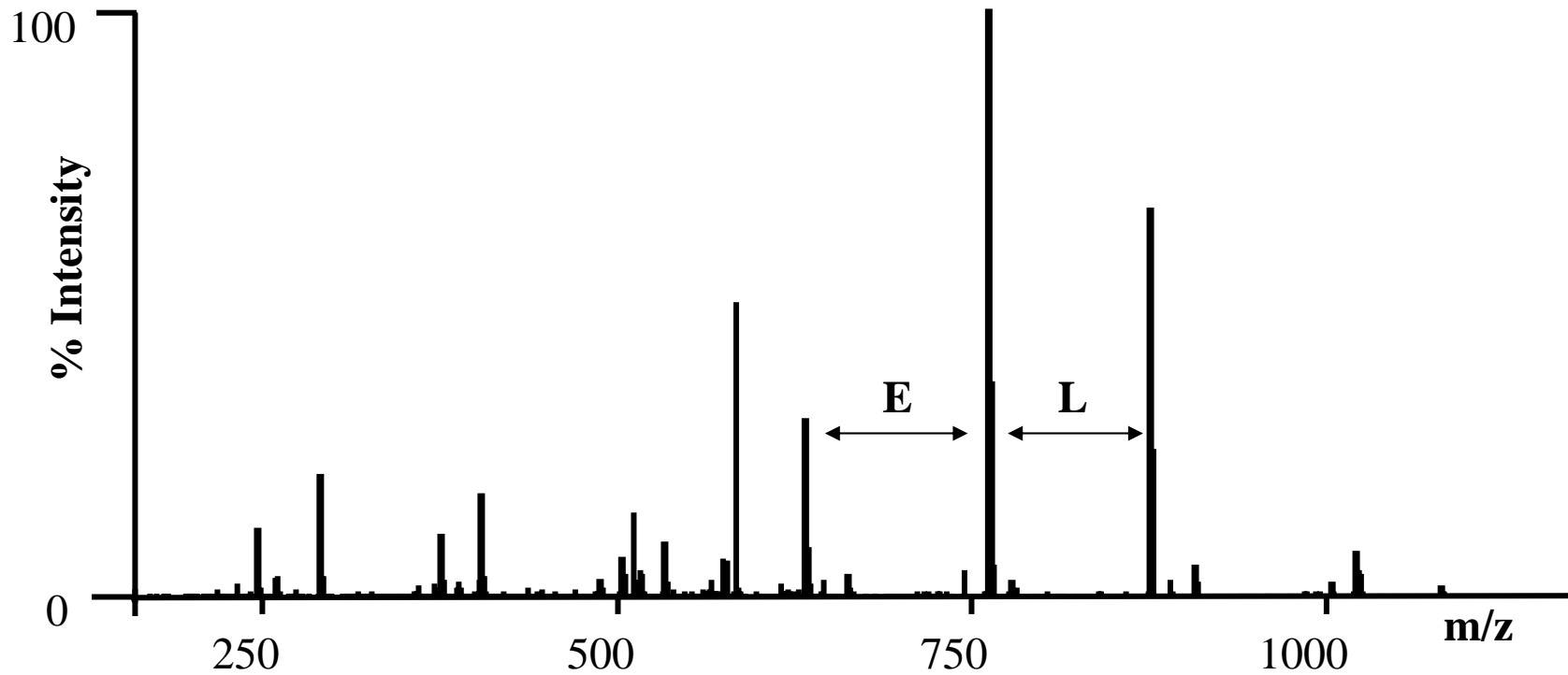
- *De novo* interpretation
- Sequence database search



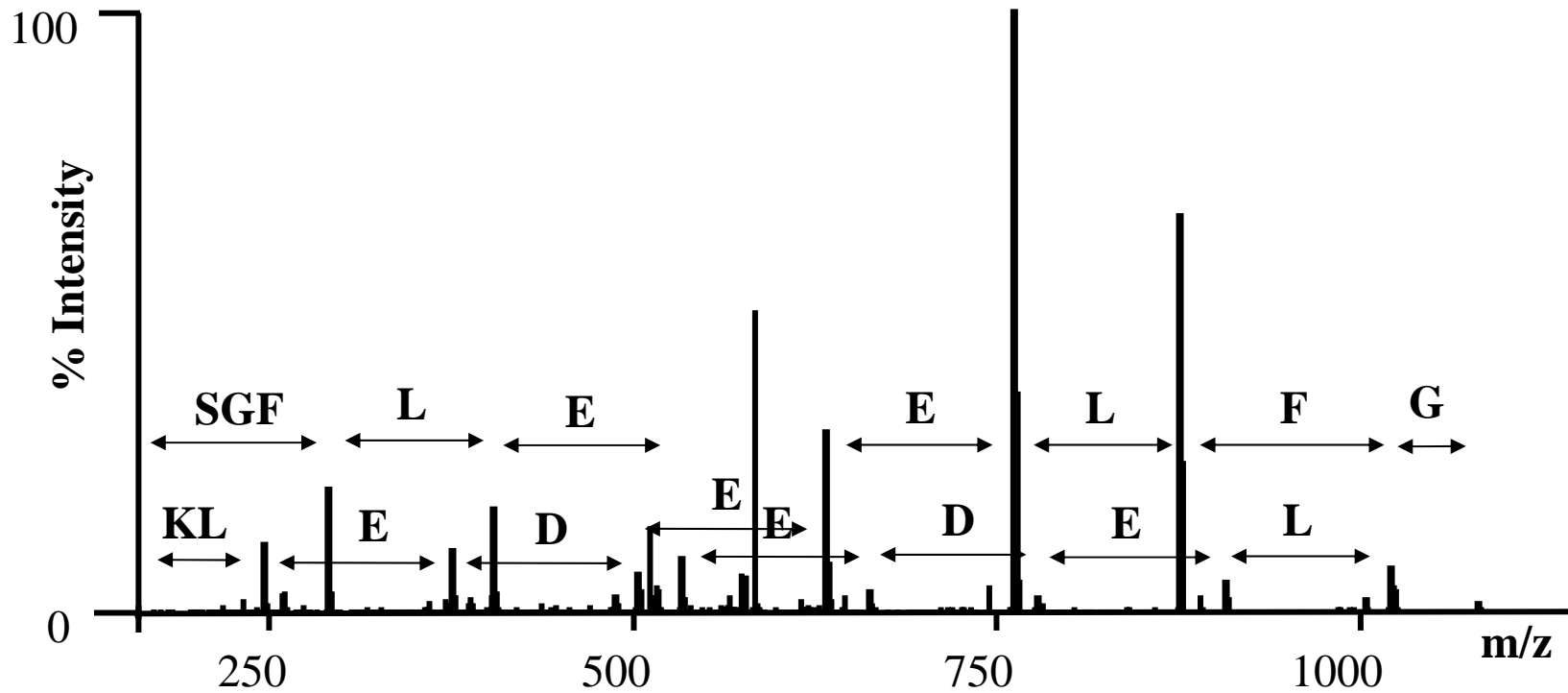
# De Novo Interpretation



# De Novo Interpretation



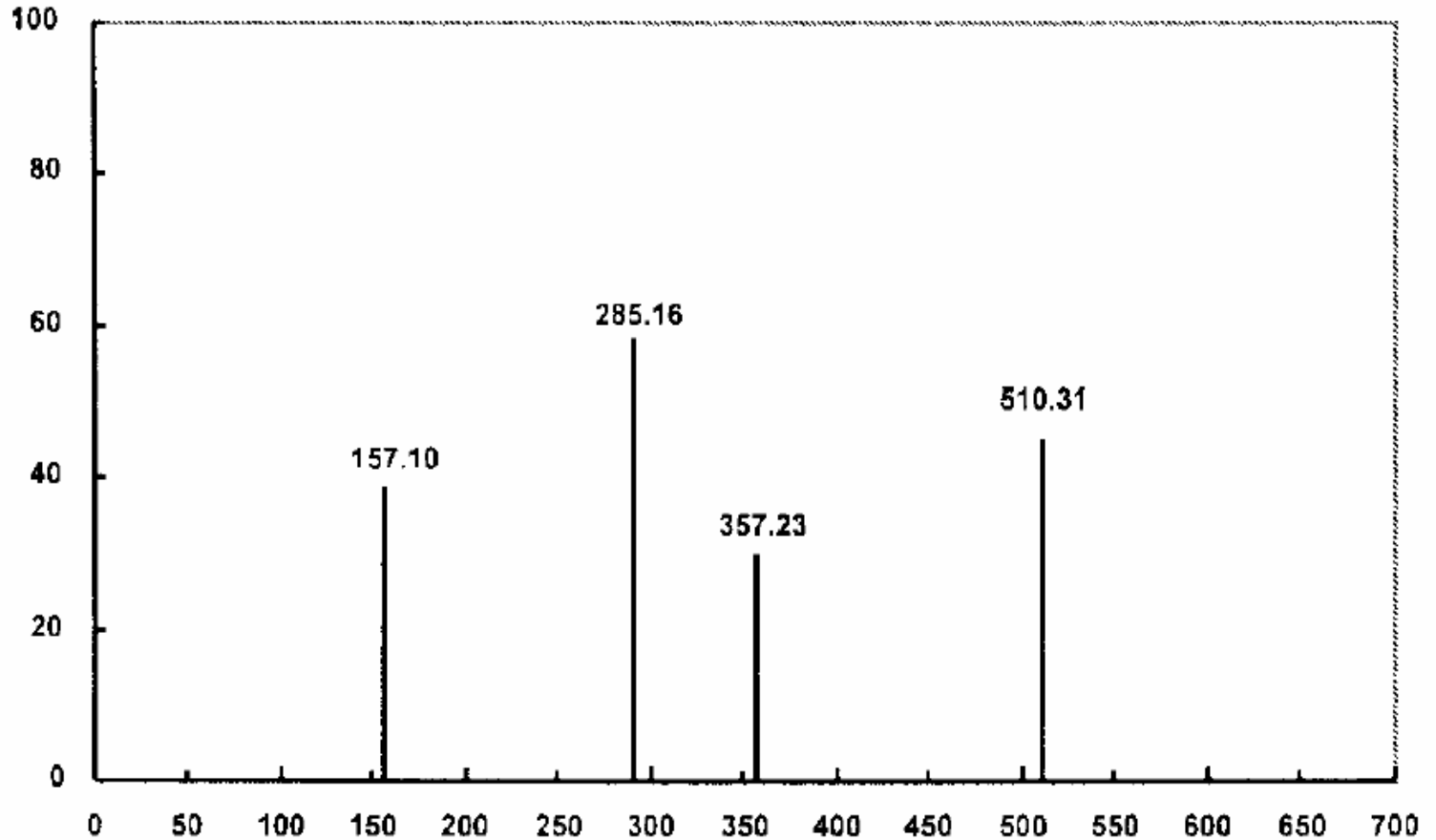
# De Novo Interpretation



# De Novo Interpretation

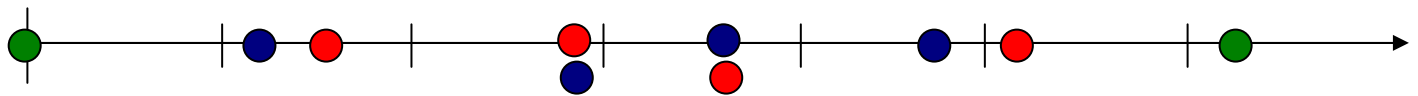
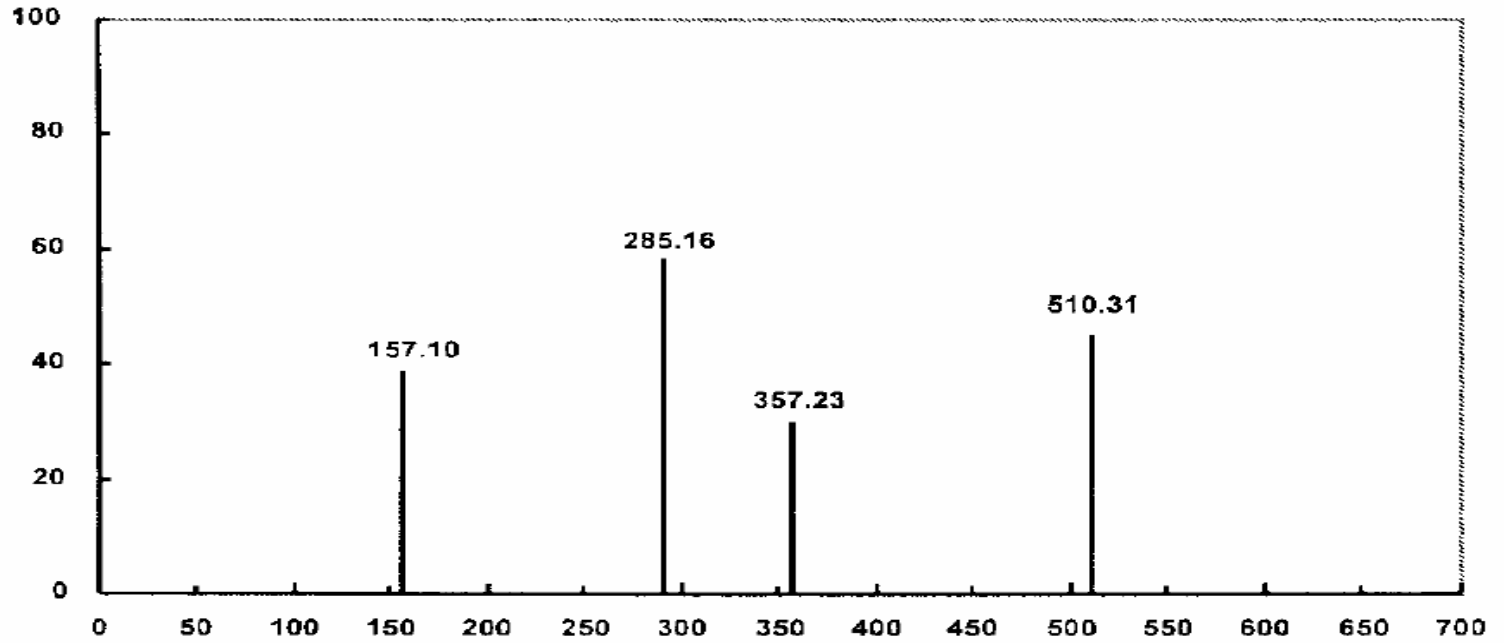
	<b>Amino-Acid</b>	<b>Residual MW</b>		<b>Amino-Acid</b>	<b>Residual MW</b>
A	Alanine	71.03712	M	Methionine	131.04049
C	Cysteine	103.00919	N	Asparagine	114.04293
D	Aspartic acid	115.02695	P	Proline	97.05277
E	Glutamic acid	129.04260	Q	Glutamine	128.05858
F	Phenylalanine	147.06842	R	Arginine	156.10112
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

# De Novo Interpretation

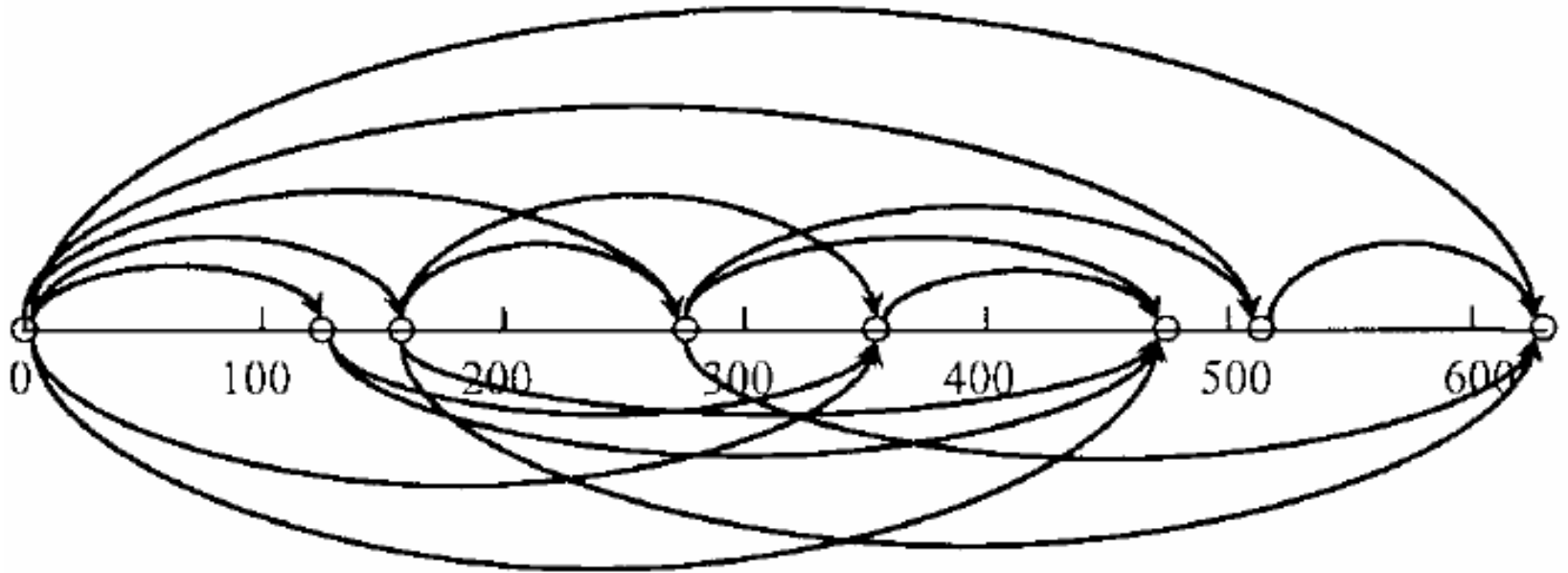


...from Lu and Chen (2003), JCB 10:1

# De Novo Interpretation



# De Novo Interpretation



...from Lu and Chen (2003), JCB 10:1



# De Novo Interpretation

- Find *good* paths in spectrum graph
- Can't use same peak twice
  - Forbidden pairs: NP-hard
  - “Nested” forbidden pairs: Dynamic Prog.
- Simple peptide fragmentation model
- Usually many apparently good solutions
- Needs better fragmentation model
- Needs better path scoring



# 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



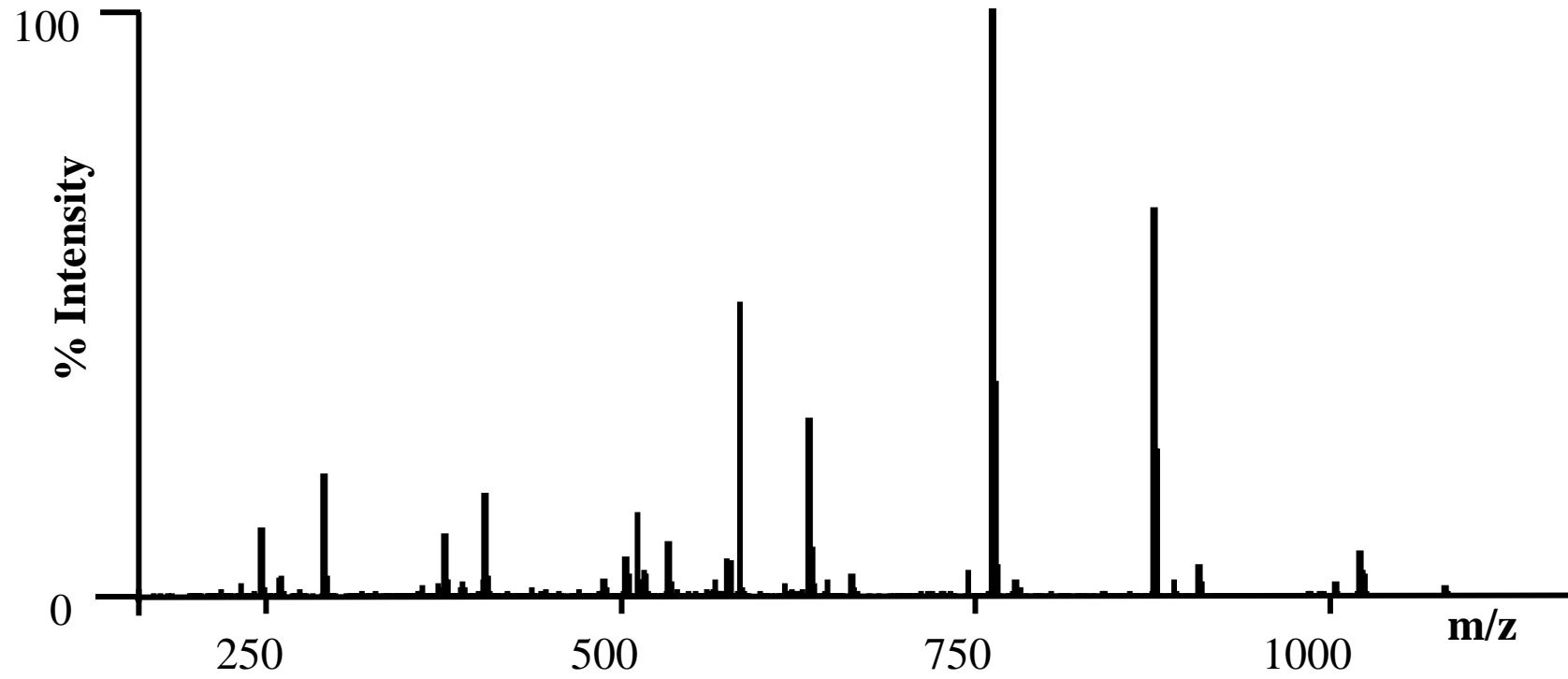
# Sequence Database Search

- 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

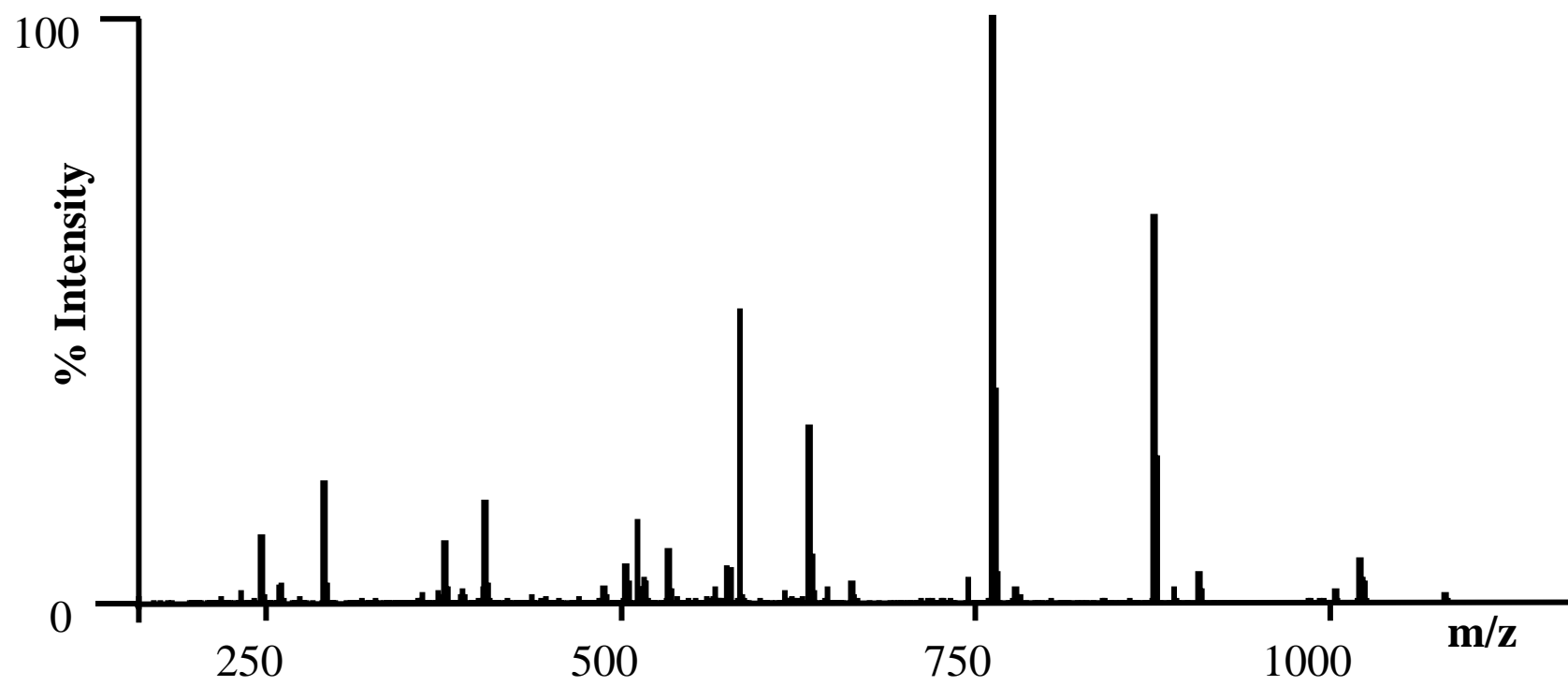
S G F L E E D E L K





# Sequence Database Search

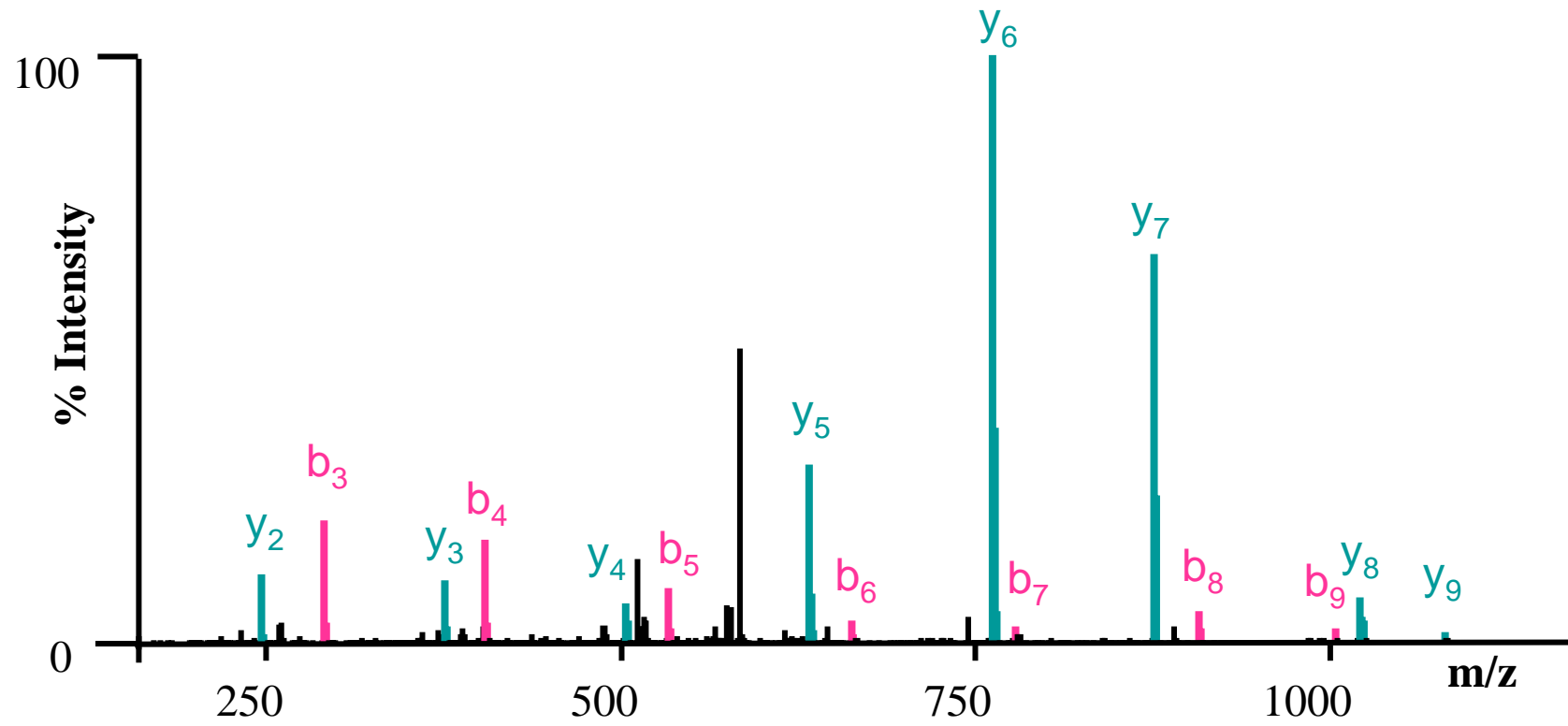
<u>88</u>	<u>145</u>	<u>292</u>	<u>405</u>	<u>534</u>	<u>663</u>	<u>778</u>	<u>907</u>	<u>1020</u>	<u>1166</u>	<u>b ions</u>
S	G	F	L	E	E	D	E	L	K	
1166	<u>1080</u>	<u>1022</u>	<u>875</u>	<u>762</u>	<u>633</u>	<u>504</u>	<u>389</u>	<u>260</u>	147	y ions





# Sequence Database Search

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





# 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

MKWVTFISLLFLFSSAYSRRDAHKSEVAHRFKDLGEENFKA  
LVLIAFAQYLQQCPFEDHVKLVNEVTEFAK...

No missed cleavage sites

MK

WVTFISLLFLFSSAYSR

GVFR

R

DAHK

SEVAHR

FK

DLGEENFK

ALVLIAFAQYLQQCPFEDHVK

LVNEVTEFAK

...



# Peptide Candidate Filtering

>ALBU\_HUMAN

MKWVTFISLLFLFSSAYS  
SRGVFRRDAHKSEVAHRFKDLGEENFKA  
LVLIAFAQYLQQCPFEDHV  
KLVNEVTEFAK...

## One missed cleavage site

MKWVTFISLLFLFSSAYS  
R  
WVTFISLLFLFSSAYS  
RGVFR  
GVFRR  
RDAHK  
DAHKSEVAHR  
SEVAHRFK  
FKDLGEENFK  
DLGEENFKALVLIAFAQYLQQCPFEDHV  
K  
ALVLIAFAQYLQQCPFEDHV  
KLVNEVTEFAK  
...

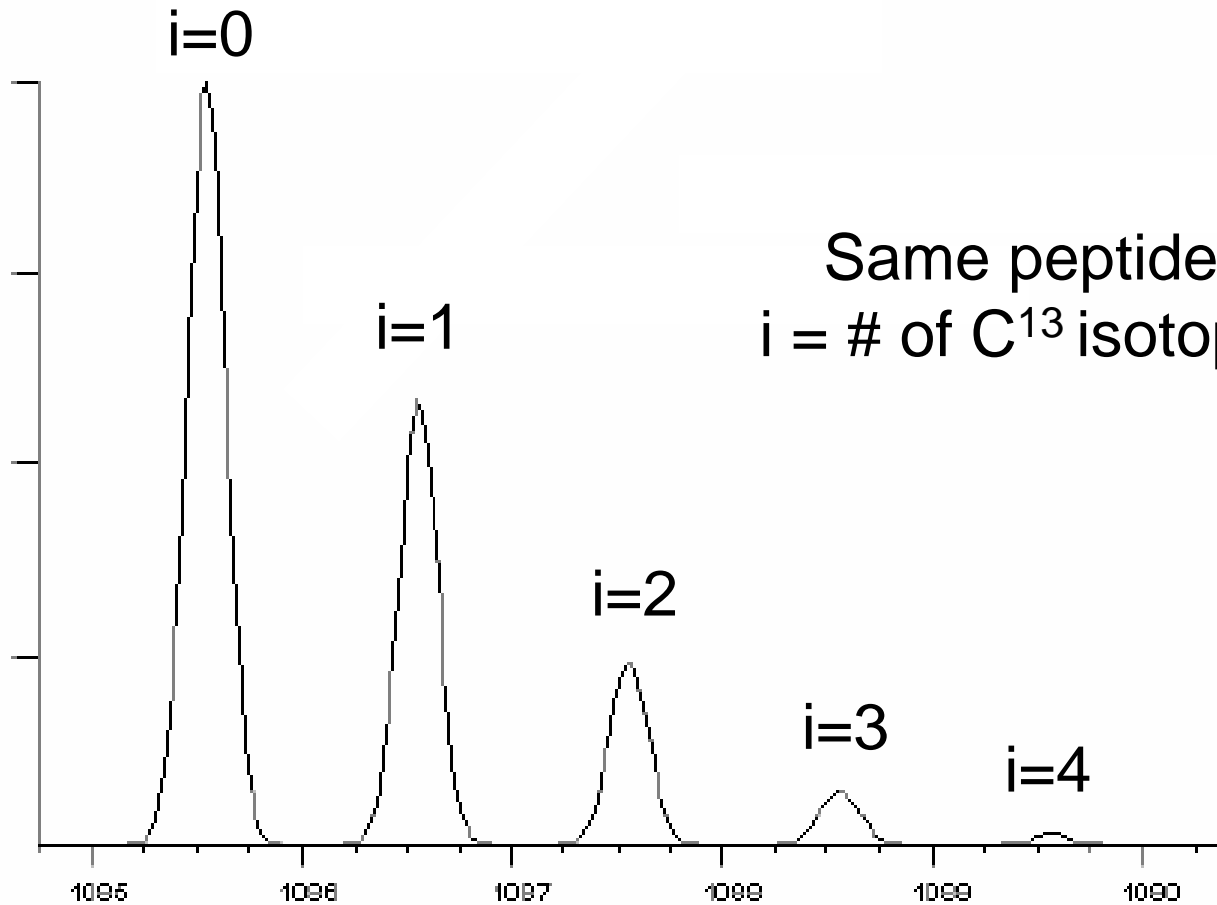


# Peptide Candidate Filtering

## Peptide molecular weight

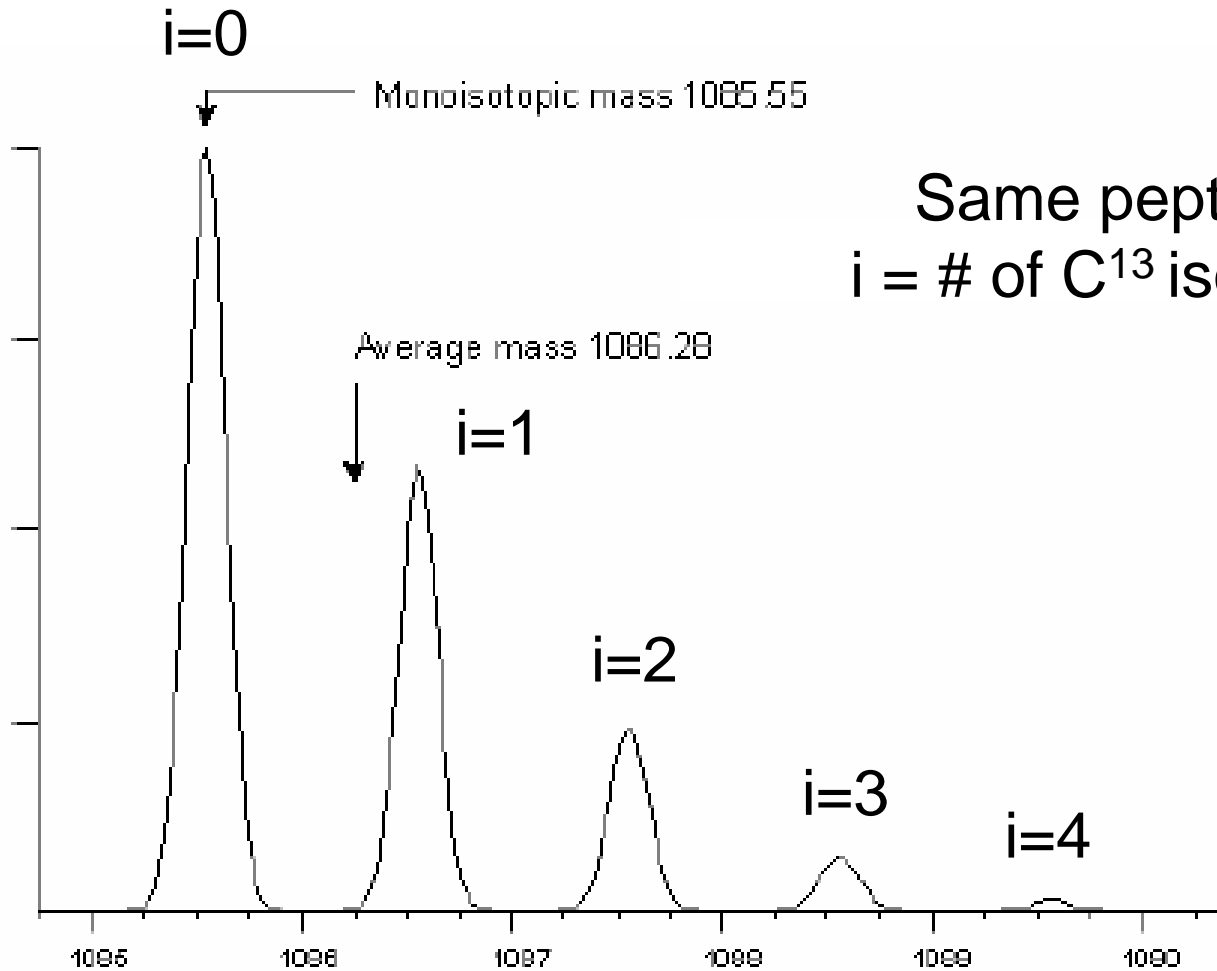
- Only have  $m/z$  value
  - Need to determine charge state
- Ion selection tolerance
- Mass for each amino-acid symbol?
  - Monoisotopic vs. Average
  - “Default” residual mass
  - Depends on sample preparation protocol
  - Cysteine almost always modified

# Peptide Molecular Weight



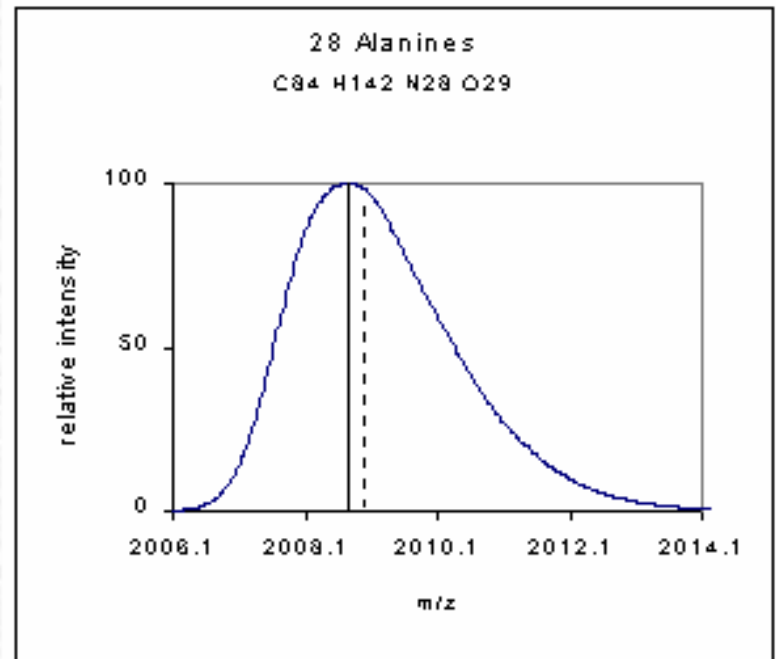
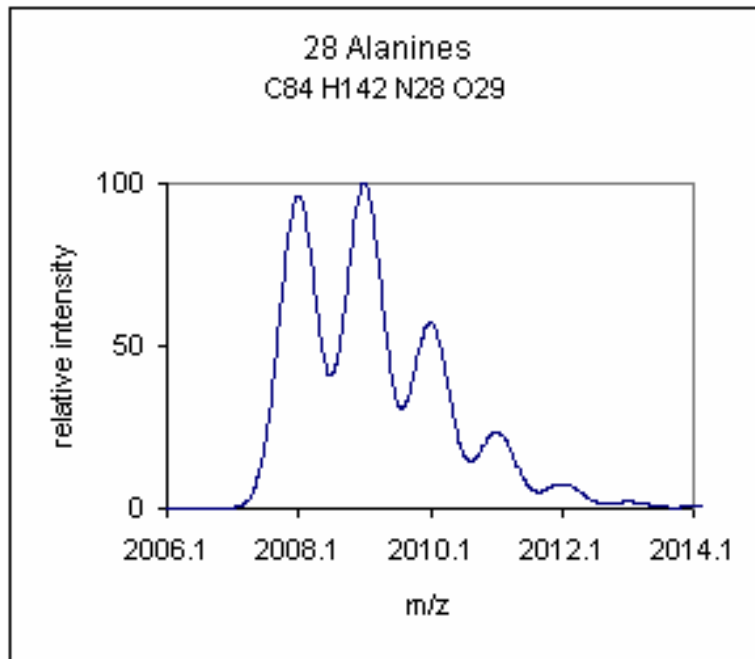
Same peptide,  
 $i = \#$  of  $C^{13}$  isotope

# Peptide Molecular Weight



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

# Peptide Molecular Weight



...from "Isotopes" – An IonSource.Com Tutorial

# Peptide Molecular Weight

- Peptide sequence WVTFISLLFLFSSAYS
- Potential phosphorylation?
  - S,T,Y + 80 Da

WVTFISLLFLFSSAYS	2018.06
WV <b>T</b> FISLLFLFSSAYS	2098.06
WVTFIS <b>S</b> LLFLFSSAYS	2098.06
WVTFISLLFLF <b>S</b> SAYS	2098.06
WVTFISLLFLFSS <b>S</b> AYS	2098.06
WVTFISLLFLFSSAY <b>S</b>	2098.06
WVTFISLLFLFSSAYS <b>R</b>	2098.06
WV <b>T</b> FIS <b>S</b> LLFLFSSAYS	2178.06
WV <b>T</b> FISLLFLF <b>S</b> SAYS	2178.06
...	...
WV <b>T</b> FIS <b>S</b> LLFLF <b>S</b> SAYS	2418.06

- 7 Molecular Weights  
- 64 “Peptides”



# 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

**MATRIX SCIENCE** HOME | WHAT'S NEW | MASCOT | HELP | PRODUCTS | SUPPORT | CONTACT

Home

## Welcome

This site features **Mascot**, a powerful search engine that uses mass spectrometry data to identify proteins from primary sequence databases. To assist you, the [help text](#) for Mascot forms a substantial knowledge base concerning protein identification by MS.

If this is your first visit, please check for [browser compatibility](#) and read the [small print](#). If you include results from Mascot in a publication, please cite either this URL or *Electrophoresis*, **20 (18)** 3551-67 (1999) ([abstract](#)).

We value your feedback and suggestions for new features. If you find any problems, errors, oversights, or just get unexpected results then please let us know.

For information on licensing Mascot for in-house use, please refer to our [Products](#) and [Support](#) pages. For recent news, check [What's New](#).

*We look forward to meeting you at booth 621*

**ABRF 2004**

*Portland, OR  
February 28  
to March 2*



# Mascot MS/MS Ions Search

<b>Your name</b>	<input type="text"/>	<b>Email</b>	<input type="text"/>
<b>Search title</b>	<input type="text"/>		
<b>Database</b>	MSDB <input type="button" value="v"/>		
<b>Taxonomy</b>	All entries <input type="button" value="v"/>		
<b>Enzyme</b>	Trypsin <input type="button" value="v"/>	<b>Allow up to</b>	1 <input type="button" value="v"/> missed cleavages
<b>Fixed modifications</b>	<input type="text" value="AB_old_ICATd0 (C)"/> <input type="text" value="AB_old_ICATd8 (C)"/> <input type="text" value="Acetyl (K)"/> <input type="text" value="Acetyl (N-term)"/> <input type="text" value="Amide (C-term)"/>	<b>Variable modifications</b>	<input type="text" value="AB_old_ICATd0 (C)"/> <input type="text" value="AB_old_ICATd8 (C)"/> <input type="text" value="Acetyl (K)"/> <input type="text" value="Acetyl (N-term)"/> <input type="text" value="Amide (C-term)"/>
<b>Protein mass</b>	<input type="text"/> kDa	<b>ICAT</b>	<input type="checkbox"/>
<b>Peptide tol. ±</b>	<input type="text" value="2.0"/> <input type="button" value="Da"/> <input type="button" value="v"/>	<b>MS/MS tol. ±</b>	<input type="text" value="0.8"/> <input type="button" value="Da"/> <input type="button" value="v"/>
<b>Peptide charge</b>	<input type="text" value="2+"/> <input type="button" value="v"/>	<b>Monoisotopic</b>	<input checked="" type="radio"/> <b>Average</b> <input type="radio"/>
<b>Data file</b>	<input type="text"/>	<input type="button" value="Browse..."/>	
<b>Data format</b>	Mascot generic <input type="button" value="v"/>	<b>Precursor</b>	<input type="text"/> m/z
<b>Instrument</b>	Default <input type="button" value="v"/>		
<b>Overview</b>	<input type="checkbox"/>	<b>Report top</b>	20 <input type="button" value="v"/> hits
<input type="button" value="Start Search ..."/>		<input type="button" value="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
<input checked="" type="checkbox"/> <a href="#">1</a>	671.90	1341.78	1341.73	0.06	0	95	1.3e-06	1	DAGTISGLNVLR
<input checked="" type="checkbox"/> <a href="#">2</a>	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TTPSYVAFDTER
<input checked="" type="checkbox"/> <a href="#">3</a>	973.90	1945.78	1945.92	-0.14	0	90	2.9e-06	1	NQVAMNPNTIFDAK
<a href="#">4</a>	1084.90	2167.78	2168.17	-0.39	1	30	2.7	3	IINEPTAAAIAAYGLDKK

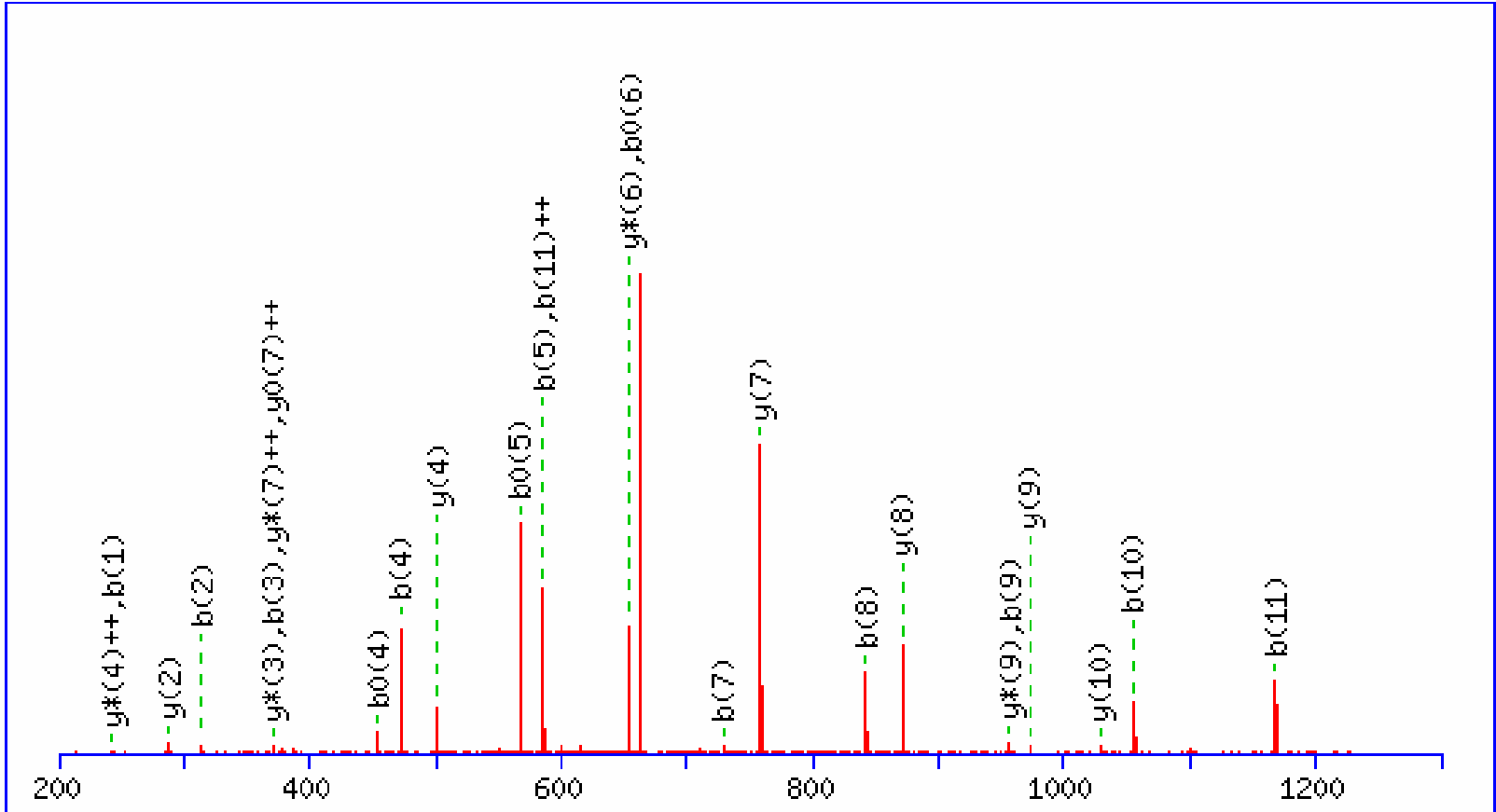
Proteins matching the same set of peptides:

[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
<input checked="" type="checkbox"/> 1	671.90	1341.78	1341.73	0.06	0	95	1.3e-06	1	DAGTISGLNVLR
<input checked="" type="checkbox"/> 2	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TTPSYVAFTDTER
<input checked="" type="checkbox"/> 3	Top scoring peptide matches to query 1								VAMNPNTIFDAK
4	dp210198c 21-Jan-98 DERIVED SPECTRUM #9								NEPTAAAIAYGLDKK
5	Score greater than 49 indicates identity								
6	Status bar shows all hits for this peptide								
Protein	Score	Delta	Hit	Protein	Peptide				
<a href="#">Q9480</a>	95.1	0.06	1	Q9XZJ2	DAGTISGLNVLR				
HSC70	65.4	-0.90			DAGTNSGLNVLR				
	51.2	1.05			DAGTIAGLEVLR				
2. <a href="#">S1499</a>	36.6	0.03			DVESLSLAILR				
dnaK	36.4	1.09			DAGVGAELDVLR				
<input type="checkbox"/> Check	36.0	-0.91			LETEEIDVIR				
	35.6	1.05			AGDVDTAVVVLR				
	35.6	0.01			SLGSLTLRVNR				
Query	35.6	1.05			LETGINADIR				
	35.3	1.05			ELTGALQDVLR				
2									PSYVAFTDTER
3									VAMNPQNTVFDK
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 <sup>++</sup>	b <sup>*</sup>	b <sup>*++</sup>	b <sup>0</sup>	b <sup>0++</sup>	Seq.	y	y <sup>++</sup>	y <sup>*</sup>	y <sup>*++</sup>	y <sup>0</sup>	y <sup>0++</sup>	#
1	<b>243.10</b>	122.05			225.09	113.05	D							12
2	<b>314.13</b>	157.57			296.12	148.57	A	1100.64	550.82	1083.62	542.31	1082.63	541.82	11
3	<b>371.16</b>	186.08			353.15	177.08	G	<b>1029.61</b>	515.31	1012.58	506.79	1011.59	506.30	10
4	<b>472.20</b>	236.61			<b>454.19</b>	227.60	T	<b>972.58</b>	486.80	<b>955.56</b>	478.28	954.57	477.79	9
5	<b>585.29</b>	293.15			<b>567.28</b>	284.14	I	<b>871.54</b>	436.27	854.51	427.76	853.53	427.27	8
6	672.32	336.66			<b>654.31</b>	327.66	S	<b>758.45</b>	379.73	741.43	<b>371.22</b>	740.44	<b>370.72</b>	7
7	<b>729.34</b>	365.17			711.33	356.17	G	671.42	336.21	<b>654.39</b>	327.70			6
8	<b>842.43</b>	421.72			824.41	412.71	L	614.40	307.70	597.37	299.19			5
9	<b>956.47</b>	478.74	939.44	470.22	938.46	469.73	N	<b>501.31</b>	251.16	484.29	<b>242.65</b>			4
10	<b>1055.54</b>	528.27	1038.51	519.76	1037.53	519.27	V	387.27	194.14	<b>370.24</b>	185.63			3
11	<b>1168.62</b>	<b>584.81</b>	1151.59	576.30	1150.61	575.81	L	<b>288.20</b>	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
<input checked="" type="checkbox"/> <a href="#">1</a>	671.90	1341.78	1341.73	0.06	0	95	1.3e-06	1	DAGTISGLNVLR
<input checked="" type="checkbox"/> <a href="#">2</a>	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TPSYVAFDTER
<input checked="" type="checkbox"/> <a href="#">3</a>	973.90	1945.78	1945.92	-0.14	0	90	2.9e-06	1	NQVAMNPNTIFDAK
<a href="#">4</a>	1084.90	2167.78	2168.17	-0.39	1	30	2.7	3	IINEPTAAAIAYGLDKK

Top scoring peptide matches to query 4

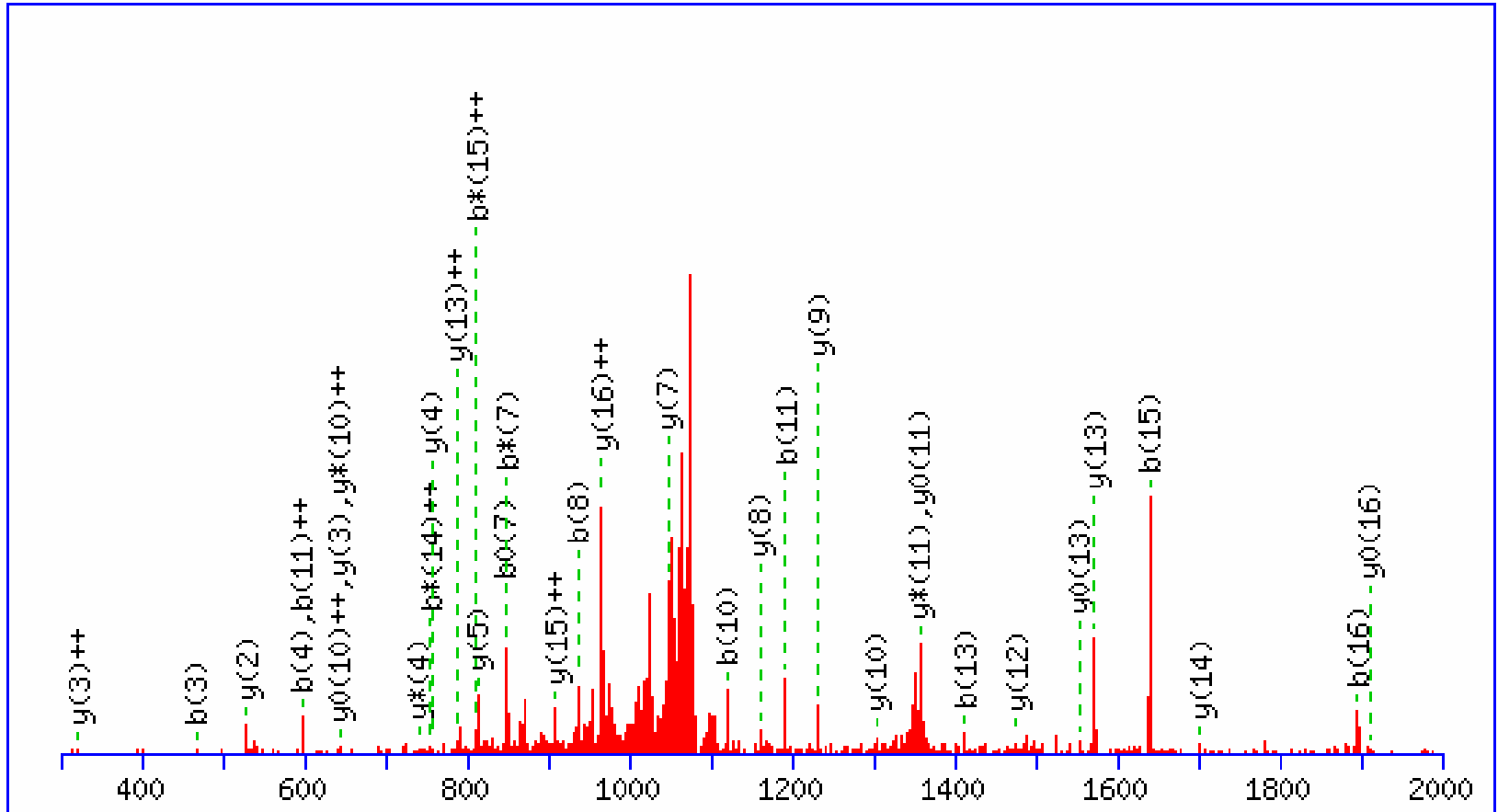
Protein: dp210198 21-Jan-98 DERIVED SPECTRUM #9

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	IIEPTAAAIAYGLDKK
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			IINEPTAAALSFGGLDKK
20.5	-0.26			IPQQWTPPCGSKCTNK
13.3	-0.26			TNCPNVSSFFQSNSVRVR

# Mascot MS/MS Search Results



# Mascot MS/MS Search Results

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



# Mascot MS/MS Search Results

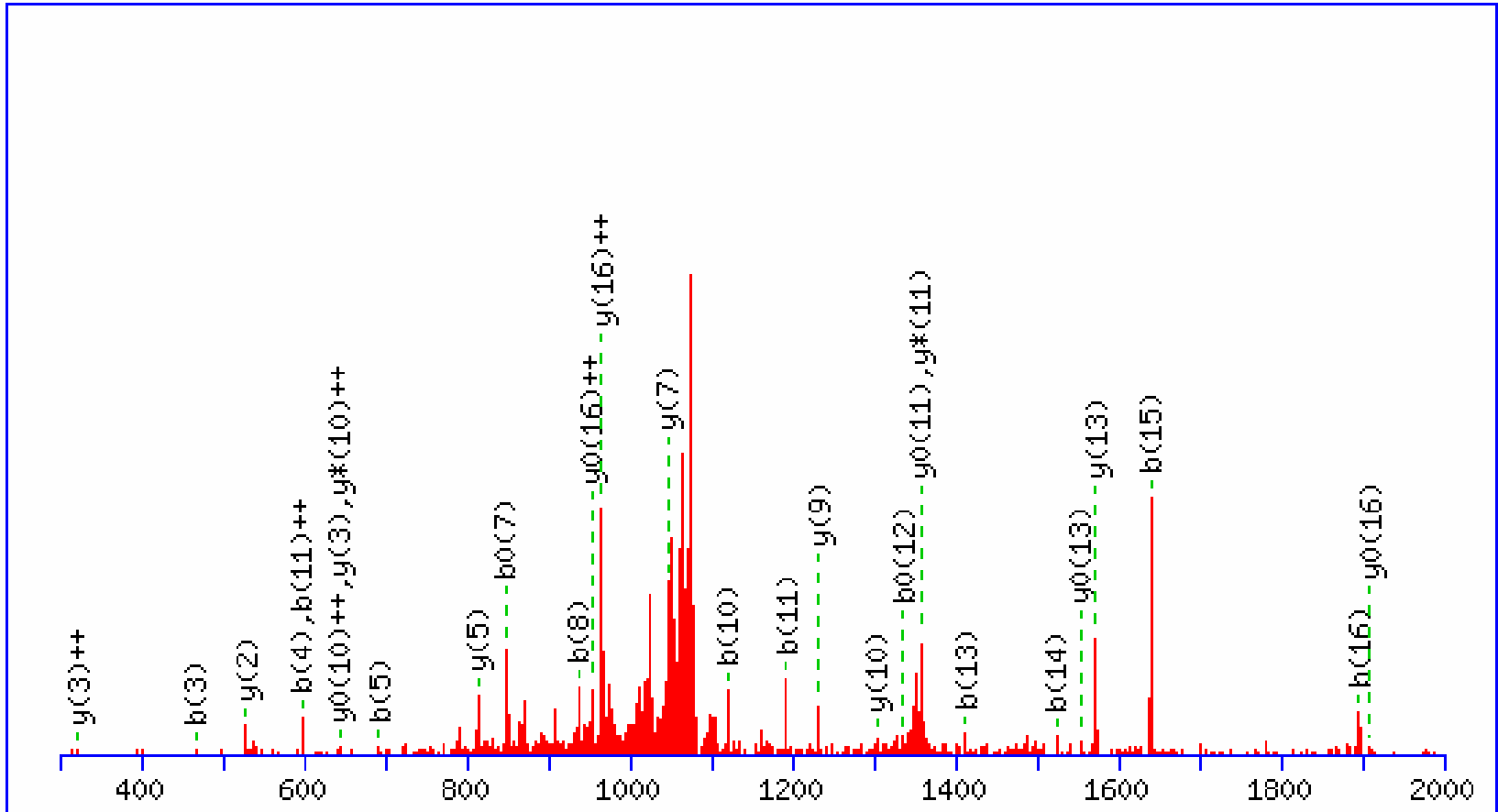
5. [Q95PU3](#)                      **Mass:** 80430      **Score:** 114      **Peptides matched:** 2

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
<a href="#">2</a>	808.30	1614.58	1613.76	0.83	0	75	0.00012	1	TTPSYVAFTDTER
<input checked="" type="checkbox"/> <a href="#">4</a>	1084.90	2167.78	2167.21	0.57	1	39	0.34	1	IIIIEPTAAAIAYGLDKK

# Mascot MS/MS Search Results



# Mascot MS/MS Search Results

#	b	b <sup>++</sup>	b <sup>*</sup>	b <sup>***</sup>	b <sup>0</sup>	b <sup>0++</sup>	Seq.	y	y <sup>++</sup>	y <sup>*</sup>	y <sup>***</sup>	y <sup>0</sup>	y <sup>0++</sup>	#
1	241.15	121.08					I							17
2	354.24	177.62					I	1928.07	<b>964.54</b>	1911.05	956.03	<b>1910.06</b>	<b>955.54</b>	16
3	<b>467.32</b>	234.17					I	1814.99	908.00	1797.96	899.49	1796.98	898.99	15
4	<b>596.37</b>	298.69			578.35	289.68	E	1701.91	851.46	1684.88	842.94	1683.90	842.45	14
5	<b>693.42</b>	347.21			675.41	338.21	P	<b>1572.86</b>	786.94	1555.84	778.42	<b>1554.85</b>	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			<b>847.49</b>	424.25	A	1374.76	687.88	<b>1357.74</b>	679.37	<b>1356.75</b>	678.88	11
8	<b>936.54</b>	468.77			918.53	459.77	A	<b>1303.73</b>	652.37	1286.70	<b>643.85</b>	1285.72	<b>643.36</b>	10
9	1007.58	504.29			989.57	495.29	A	<b>1232.69</b>	616.85	1215.66	608.33	1214.68	607.84	9
10	<b>1120.66</b>	560.83			1102.65	551.83	I	1161.65	581.33	1144.62	572.82	1143.64	572.32	8
11	<b>1191.70</b>	<b>596.35</b>			1173.69	587.35	A	<b>1048.57</b>	524.79	1031.54	516.27	1030.56	515.78	7
12	1354.76	677.88			<b>1336.75</b>	668.88	Y	977.53	489.27	960.50	480.76	959.52	480.26	6
13	<b>1411.78</b>	706.40			1393.77	697.39	G	<b>814.47</b>	407.74	797.44	399.22	796.46	398.73	5
14	<b>1524.87</b>	762.94			1506.86	753.93	L	757.45	379.23	740.42	370.71	739.43	370.22	4
15	<b>1639.89</b>	820.45			1621.88	811.45	D	<b>644.36</b>	<b>322.68</b>	627.33	314.17	626.35	313.68	3
16	<b>1895.05</b>	948.03	1878.03	939.52	1877.04	939.02	K	<b>529.33</b>	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](http://www.matrixscience.com)
- Seattle Proteome Center (ISB)
  - [www.proteomecenter.org](http://www.proteomecenter.org)
- *Proteomic Mass Spectrometry Lab* at The Scripps Research Institute
  - [fields.scripps.edu](http://fields.scripps.edu)
- UCSF ProteinProspector
  - [prospector.ucsf.edu](http://prospector.ucsf.edu)