De Novo Sequencing of Peptides

Sequenciação de Péptidos *de Novo*

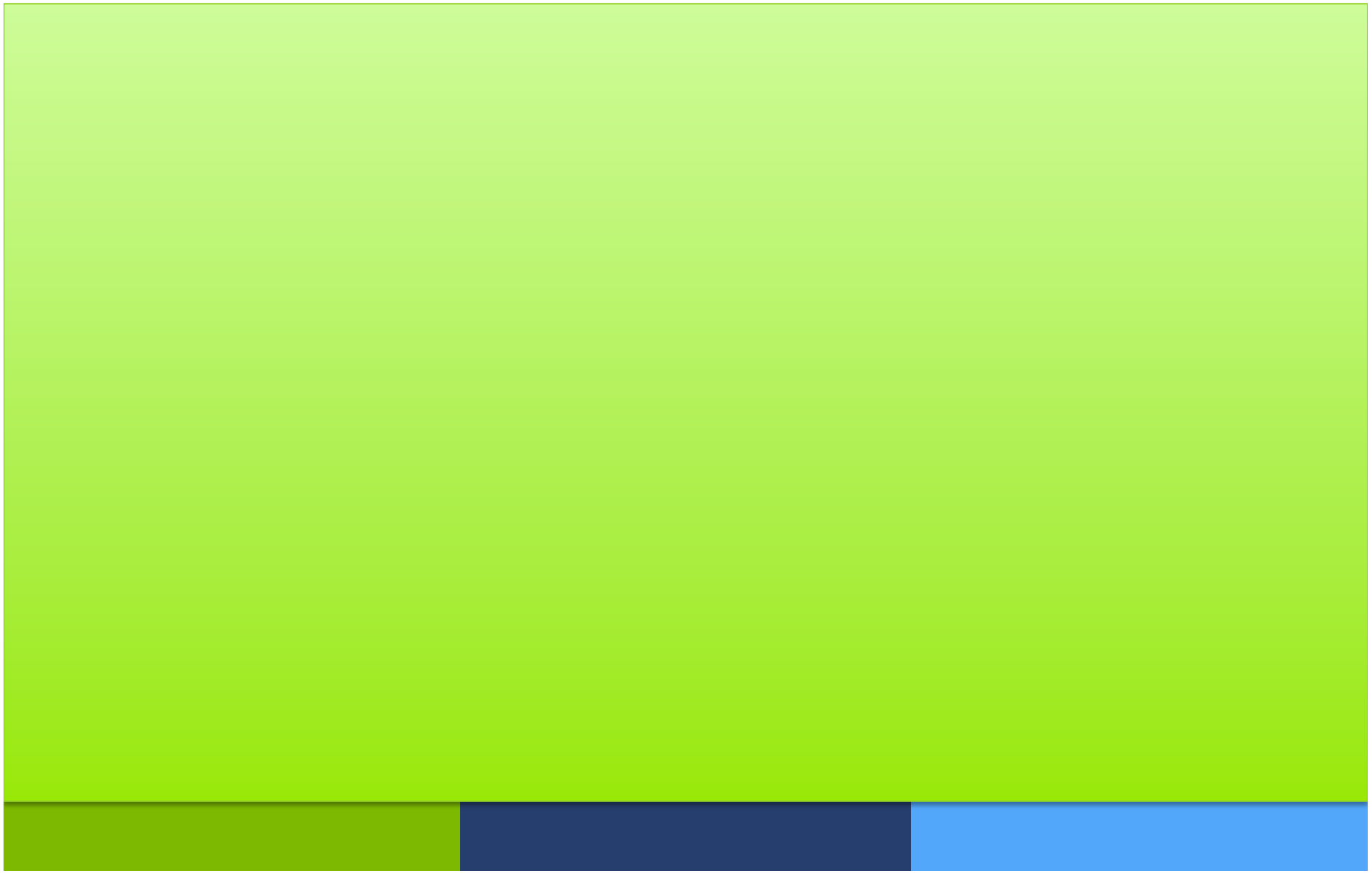
José Luis Capelo Martínez | M.Sc. | Ph.D. | FRSC

Professor Auxiliar | DQ FCT UNL

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

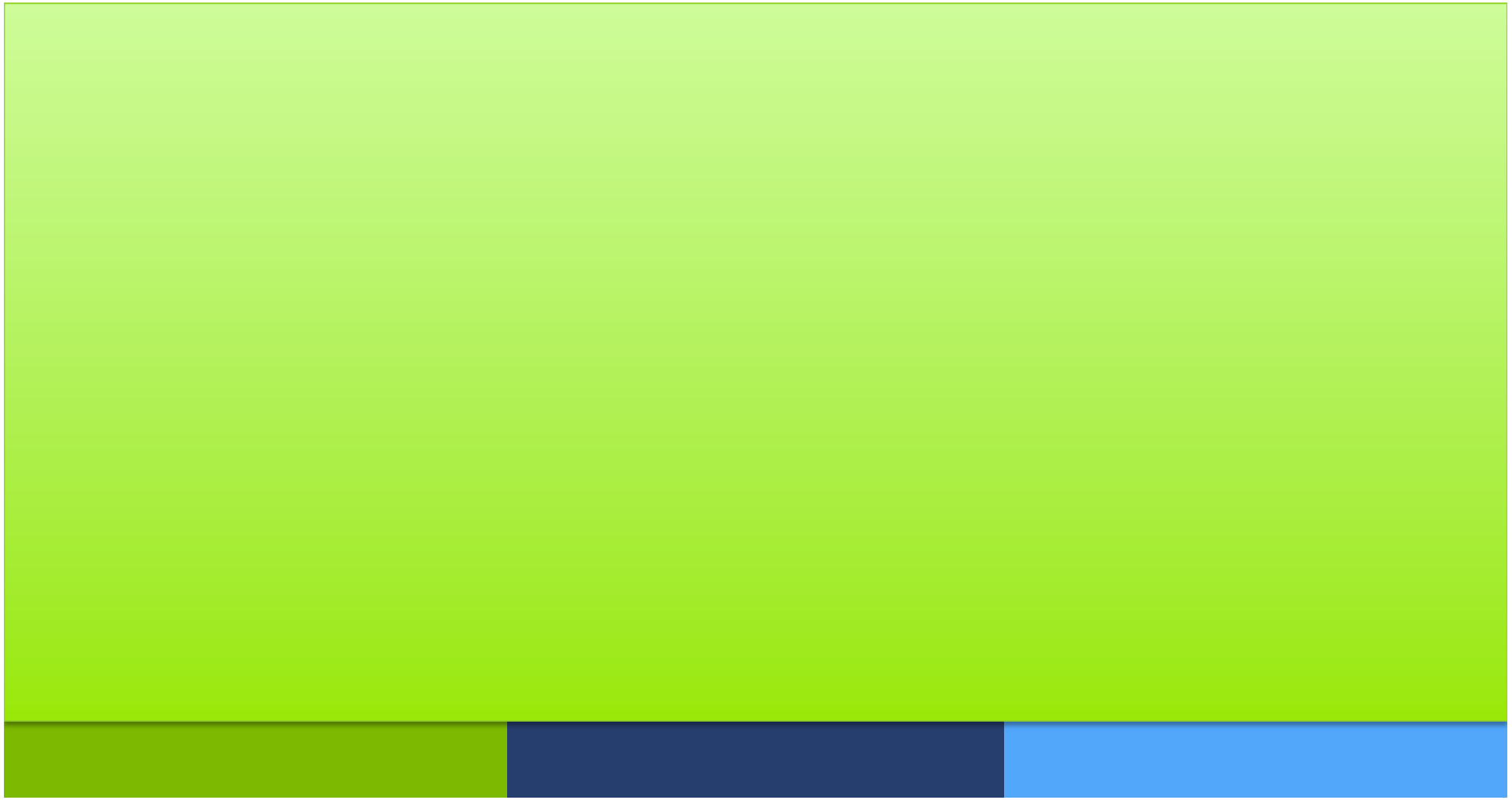
De Novo Sequencing of Peptides

1. Improves your memory



1. Improves your memory

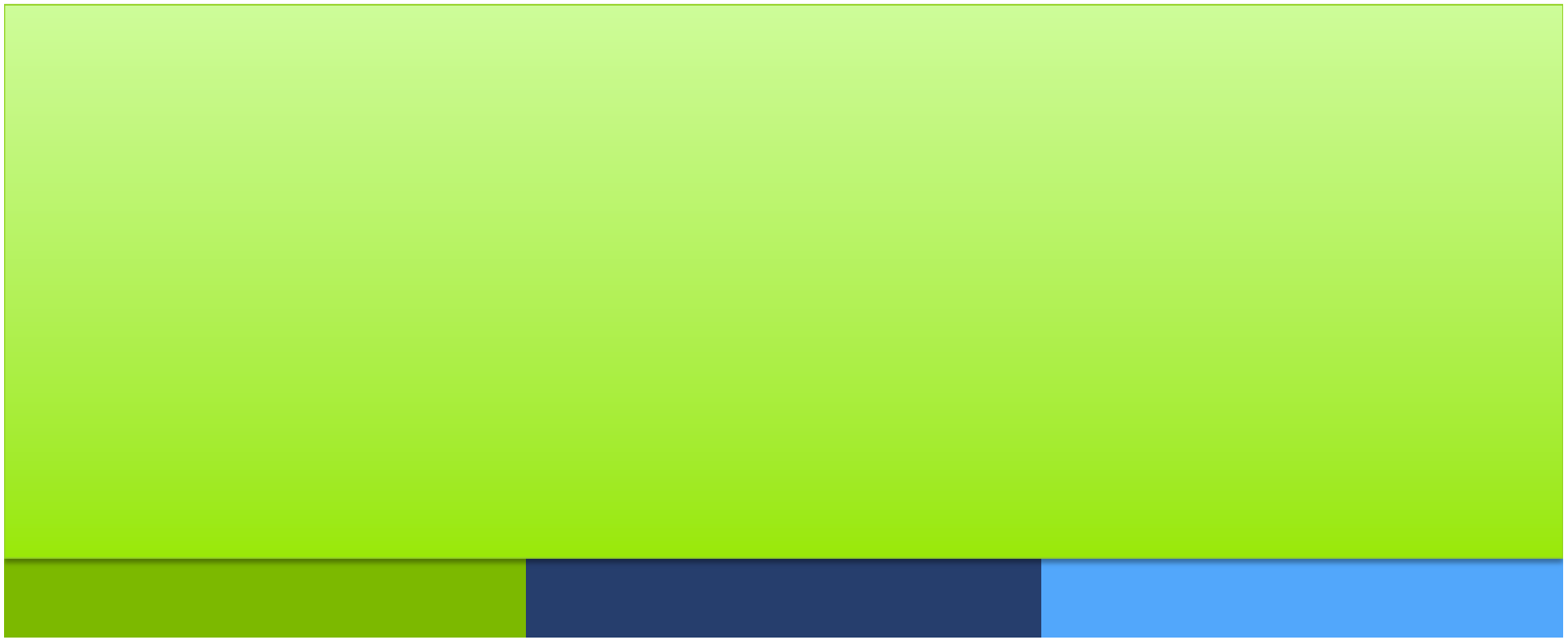
2. Stimulates your mind



1. Improves your memory

2. Stimulates your mind

3. Reduces the chances of developing Alzheimer



1. Improves your memory

2. Stimulates your mind

3. Reduces the chances of developing Alzheimer

4. Learns to do things quickly

1. Improves your memory

2. Stimulates your mind

3. Reduces the chances of developing Alzheimer

4. Learns to do things quickly

5. Increases your concentration power

1. Improves your memory

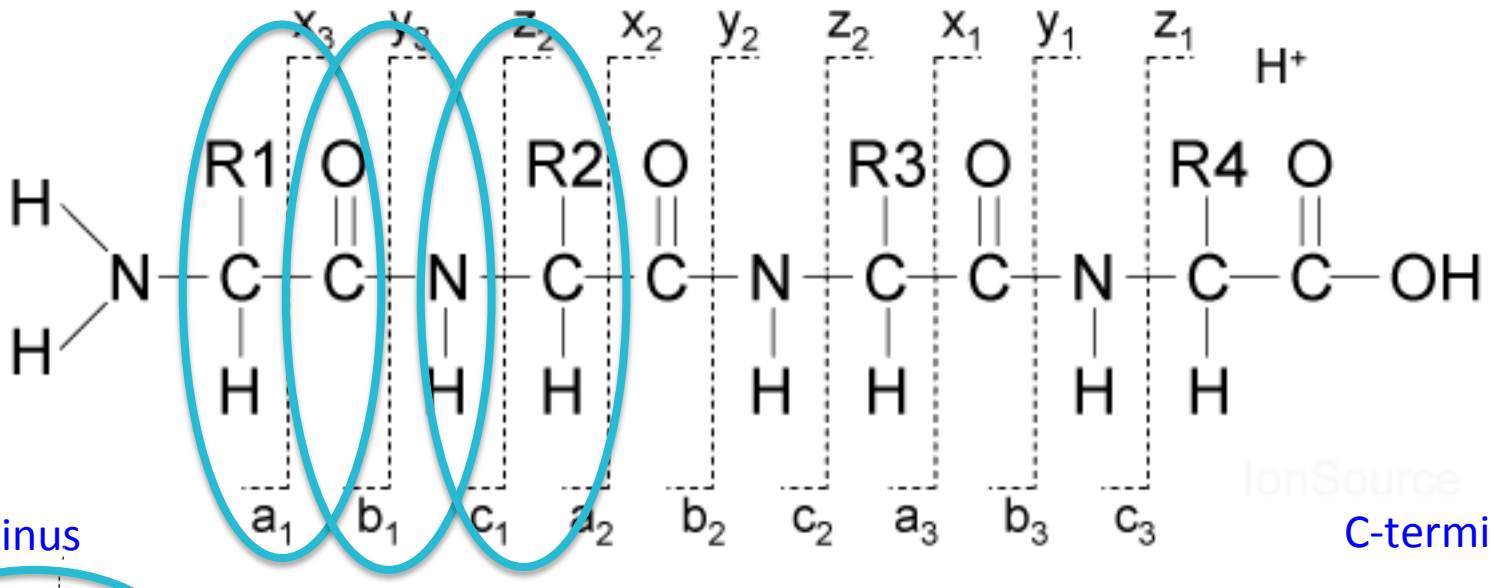
2. Stimulates your mind

3. Reduces the chances of developing Alzheimer

4. Learns to do things quickly

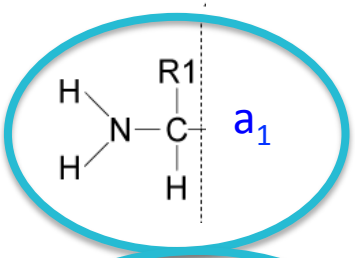
5. Increases your concentration power

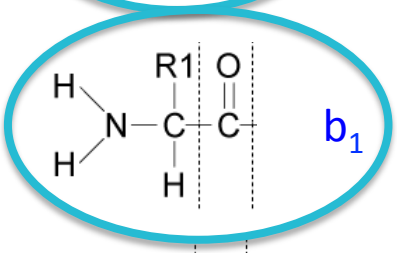
6. Feel Happy

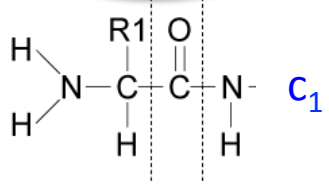


N-terminus

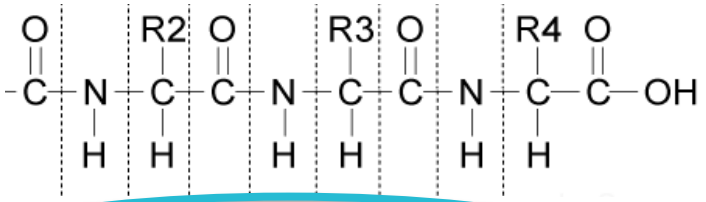
C-terminus



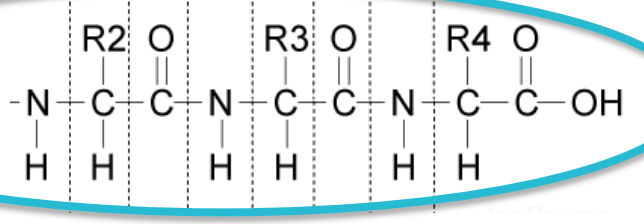




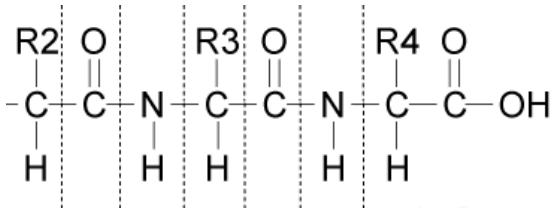
x_3



y_3



z_3



Key points

Peptides are fragmented in collision cells by colliding with molecules of He or N_2 or Ar

COLLISION-INDUCED DISSOCIATION - CID

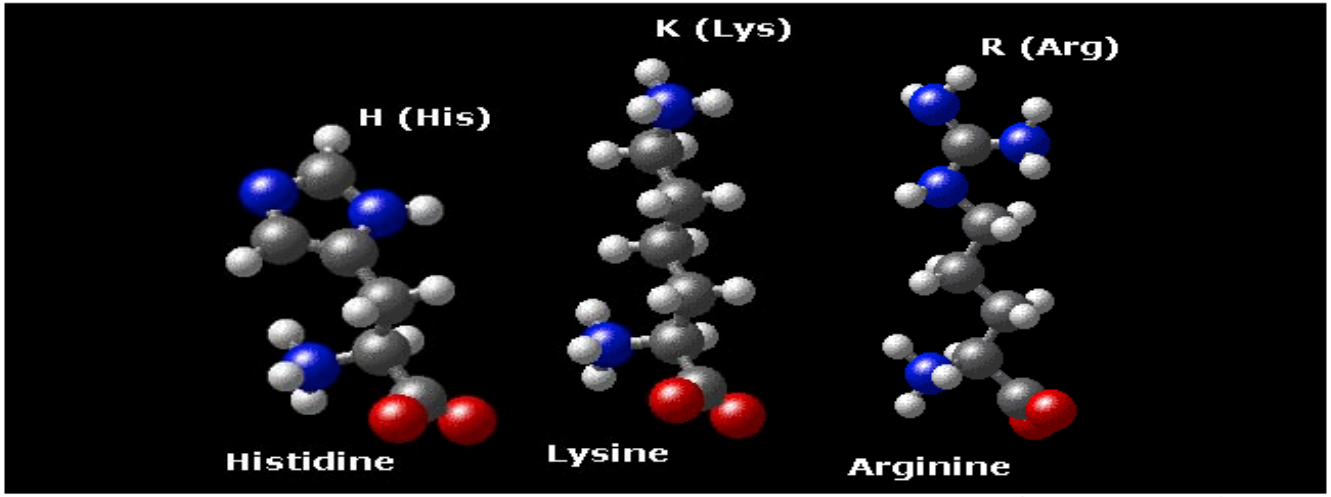
Under mild conditions **b** and **y** ions fragments are formed preferentially

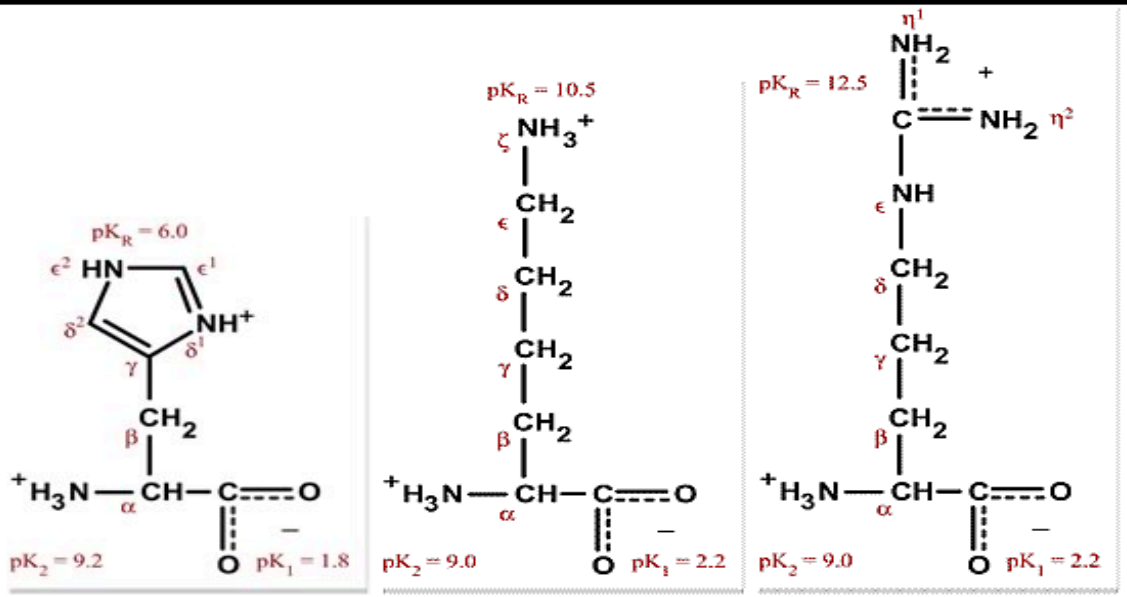
Key points

Basic residues in the N-terminus and digested with
LysN promotes spectrum
with short y-ion series and long b-ion series

Basic Amino Acids

Basic amino acids are polar and positively charged at pH values below their pK_a's, and are very hydrophilic. Even though the basic amino acids are almost always in contact with the solvent, the side chain of lysine has a marked hydrocarbon character, so it is often found NEAR the surface, with the amino group of the side chain in contact with solvent. Note that in the drawing, histidine is shown in the protonated form, while at pH 7.0, the imidazole would exist predominantly in the neutral form.





Key points

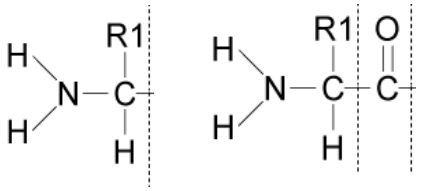
Basic residues in the N-terminus and in the C-terminus promotes spectra with y-ion series and b-ion series of comparable length

De NOVO Sequencing Rules
y or b ions

Loss of ammonia
- 17 u.a.

- **R** Arginine
- **K** Lysine
- **Q** Glutamine
- **N** Asparagine

C=O
- 28 u.a.



Loss of H_2O
- 18 u.a.

- **S** Serine
- **T** Threonine
- **E** Glutamic

De NOVO Sequencing Rules
Drop intensity

b ion

- **R** Arginine
- **K** Lysine
- **H** Histidine
- **P** Proline
- **G** Glycine

y ion

- **E** Glutamic

De NOVO Sequencing Rules
Complementary Masses

y ion

$y = (M+H)^{1+} - b + 1$

b ion

$b = (M+H)^{1+} - y + 1$

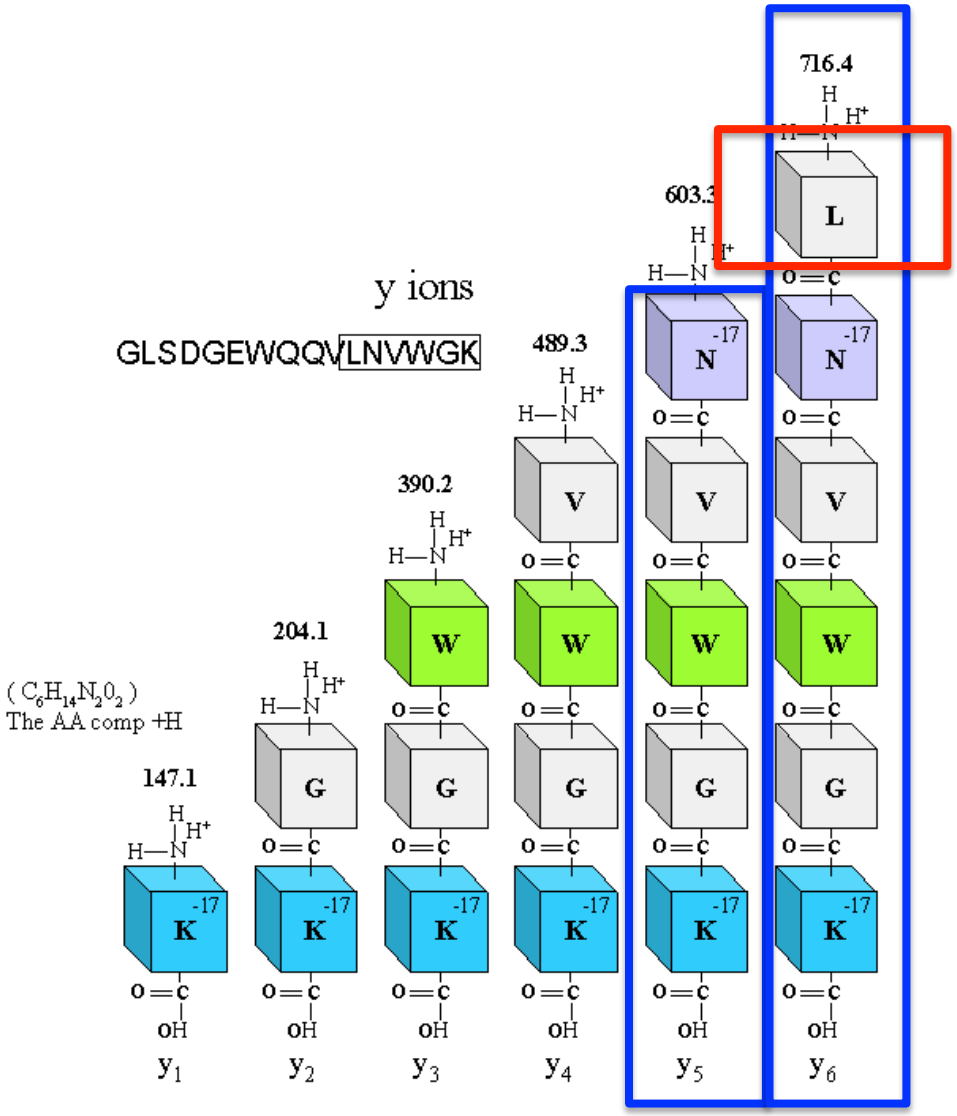
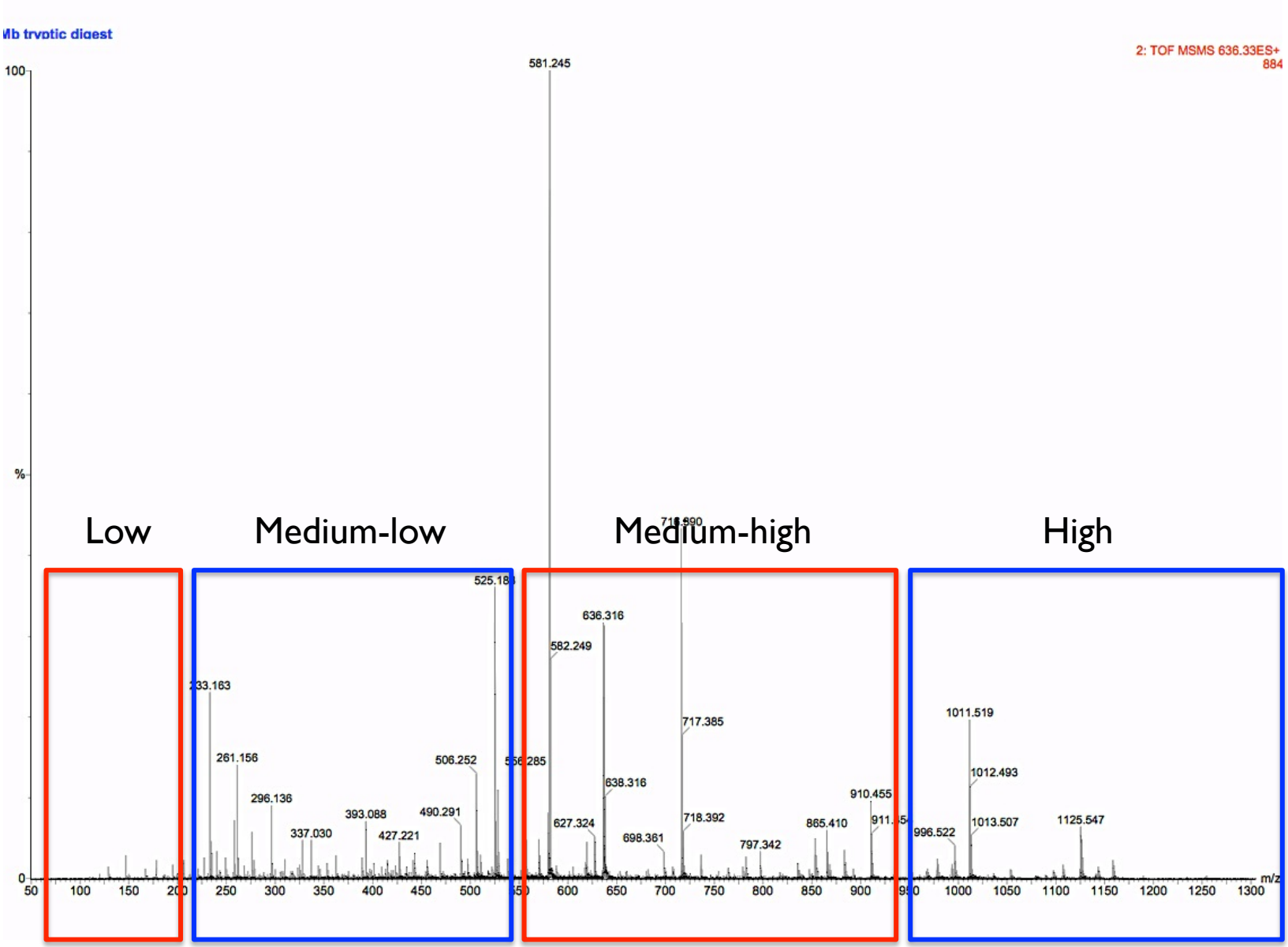


Figure 5. The first six **y** ions are illustrated. The calculated masses are shown above each **y** ion in bold numbers.

716.4-603.3 = 113.1

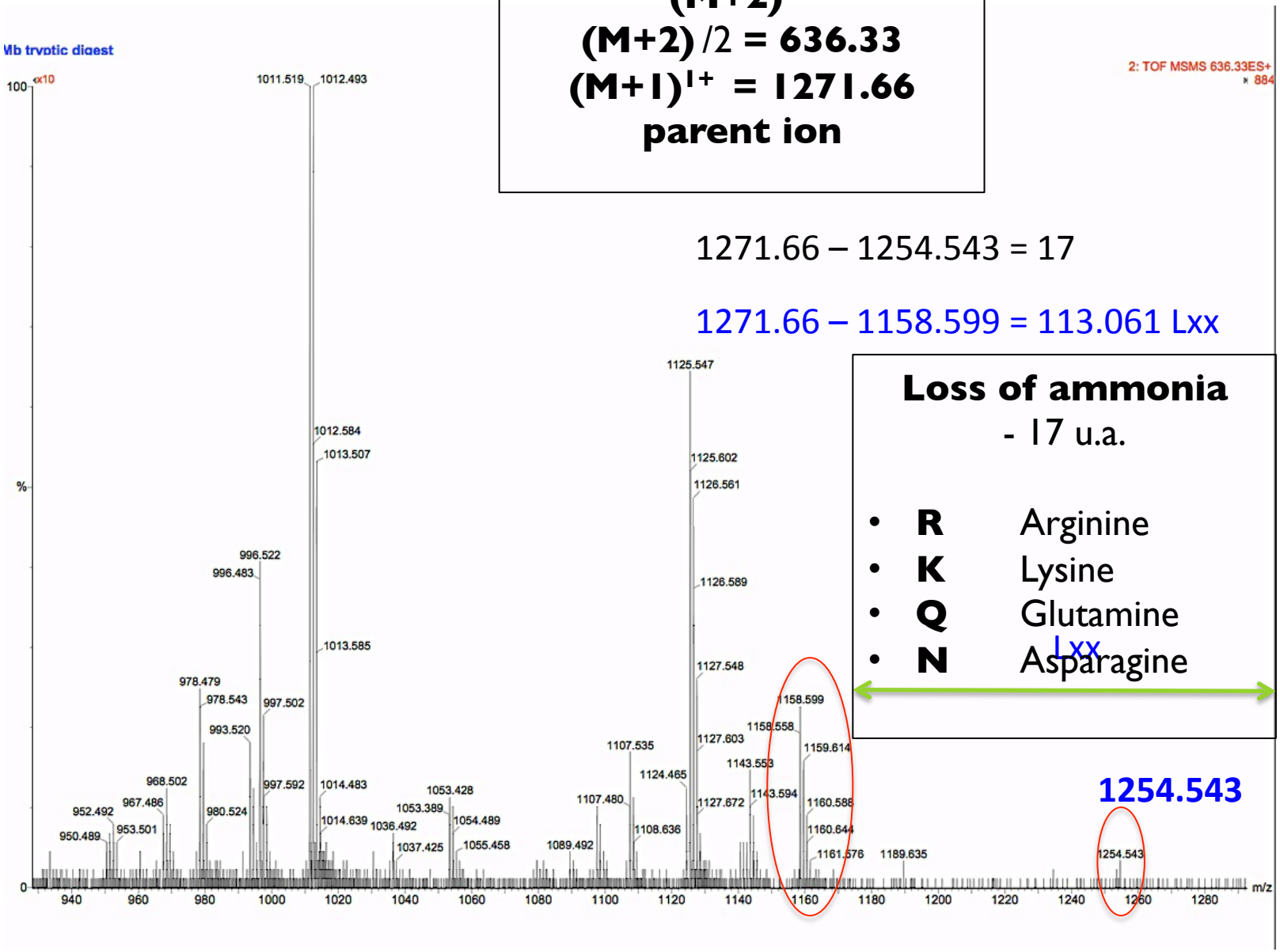
Leucine or Isoleucine

The difference between two consecutive ions of the same series correspond to one amino acid


Mb tryptic digest
2: TOF MSMS 636.33ES+
884
100
%
0
581.245
Low
Medium-low
Medium-high
High
525.18
636.316
582.249
233.163
717.385
1011.519
261.156
506.252
638.316
1012.493
296.136
910.455
393.088
490.291
718.392
627.324
865.410
1013.507
1125.547
337.030
427.221
698.361
797.342
996.522
50
100
150
200
250
300
350
400
450
500
600
650
700
750
800
850
900
1000
1050
1100
1150
1200
1250
1300
m/z

High

$(M+2)^{2+}$
(M+2) /2 = 636.33
$(M+1)^{1+}$ = 1271.66
parent ion



1271.66 – 1254.543 = 17

1271.66 – 1158.599 = 113.061 Lxx

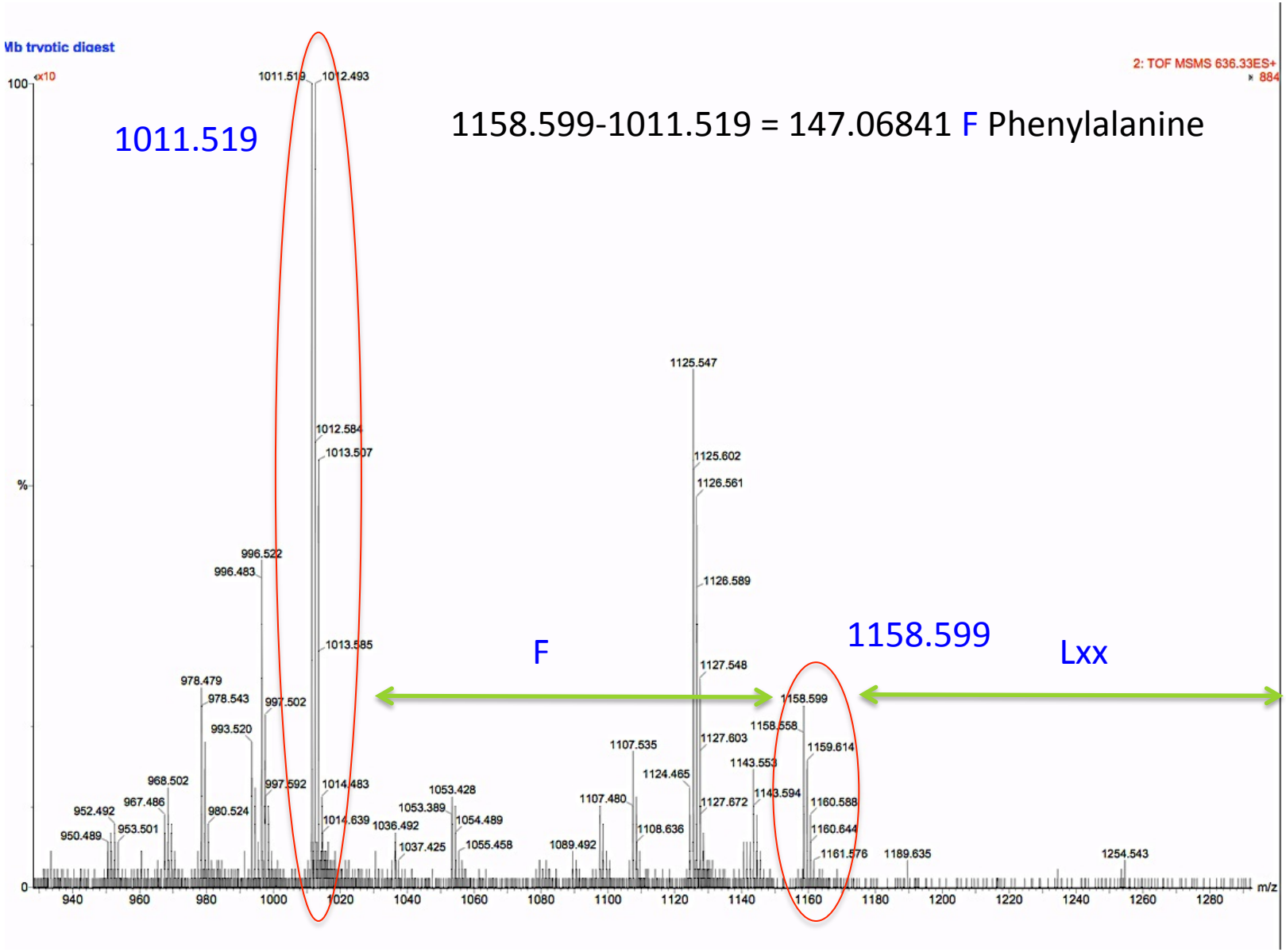
Loss of ammonia
- 17 u.a.

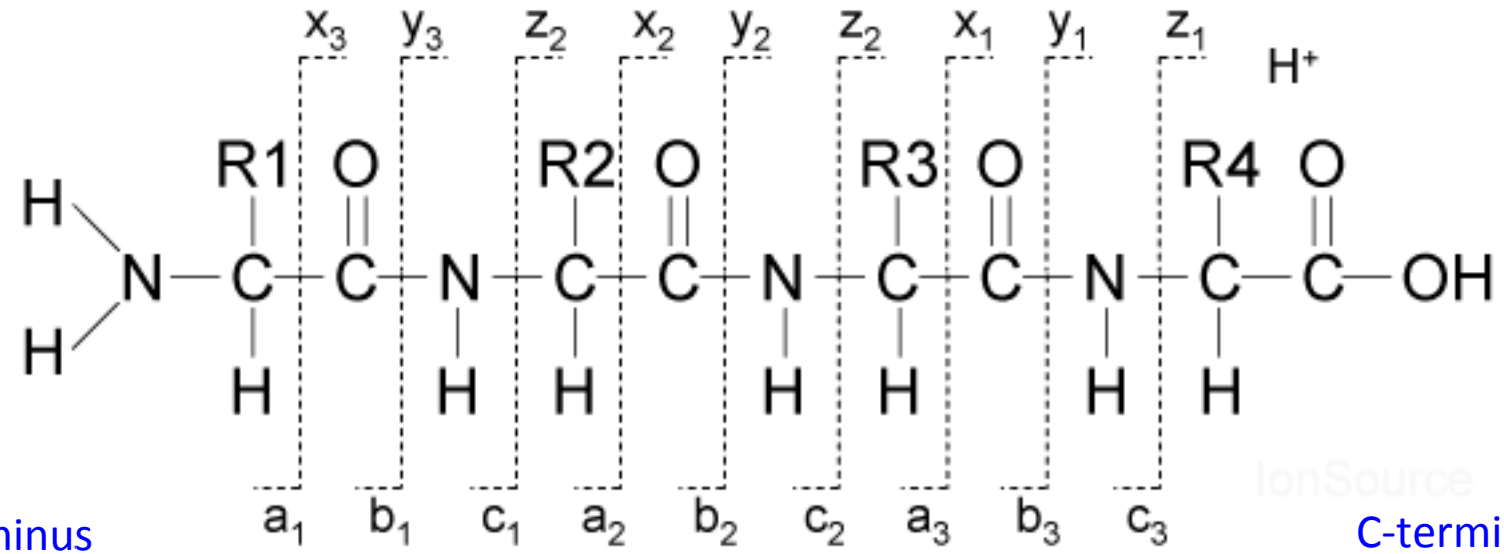
- **R** Arginine
- **K** Lysine
- **Q** Glutamine
- **N** Asparagine

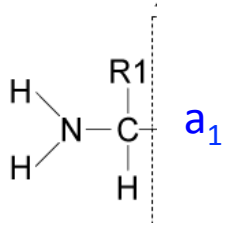
Lxx

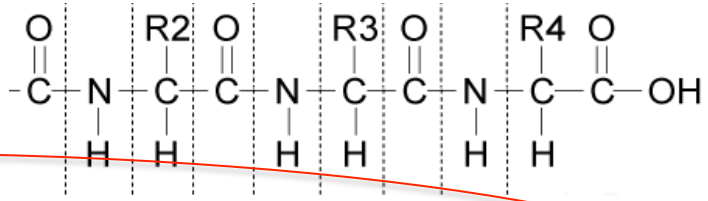
1254.543

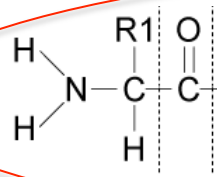
High

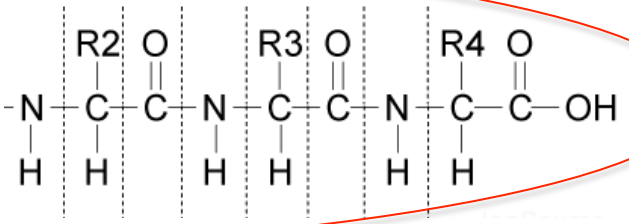

Mb tryptic digest
2: TOF MSMS 636.33ES+
884
100
%
0
x10
1011.519
1158.599-1011.519 = 147.06841 F Phenylalanine
F
1158.599
Lxx
m/z
940
960
980
1000
1020
1040
1060
1080
1100
1120
1140
1160
1180
1200
1220
1240
1260
1280

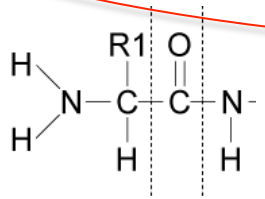

x3 y3 z2 x2 y2 z2 x1 y1 z1
H+
R1 O R2 O R3 O R4 O
H N C C N C C N C C N C C OH
H H H H H H H H
a1 b1 c1 a2 b2 c2 a3 b3 c3
N-terminus
C-terminus

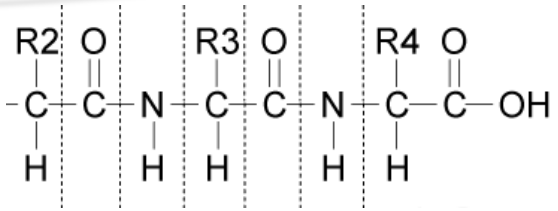

R1
H N C
H H
a1


x3
O R2 O R3 O R4 O
C N C C N C C N C C OH
H H H H H H


R1 O
H N C C
H H
b1


y3
R2 O R3 O R4 O
N C C N C C N C C OH
H H H H H H

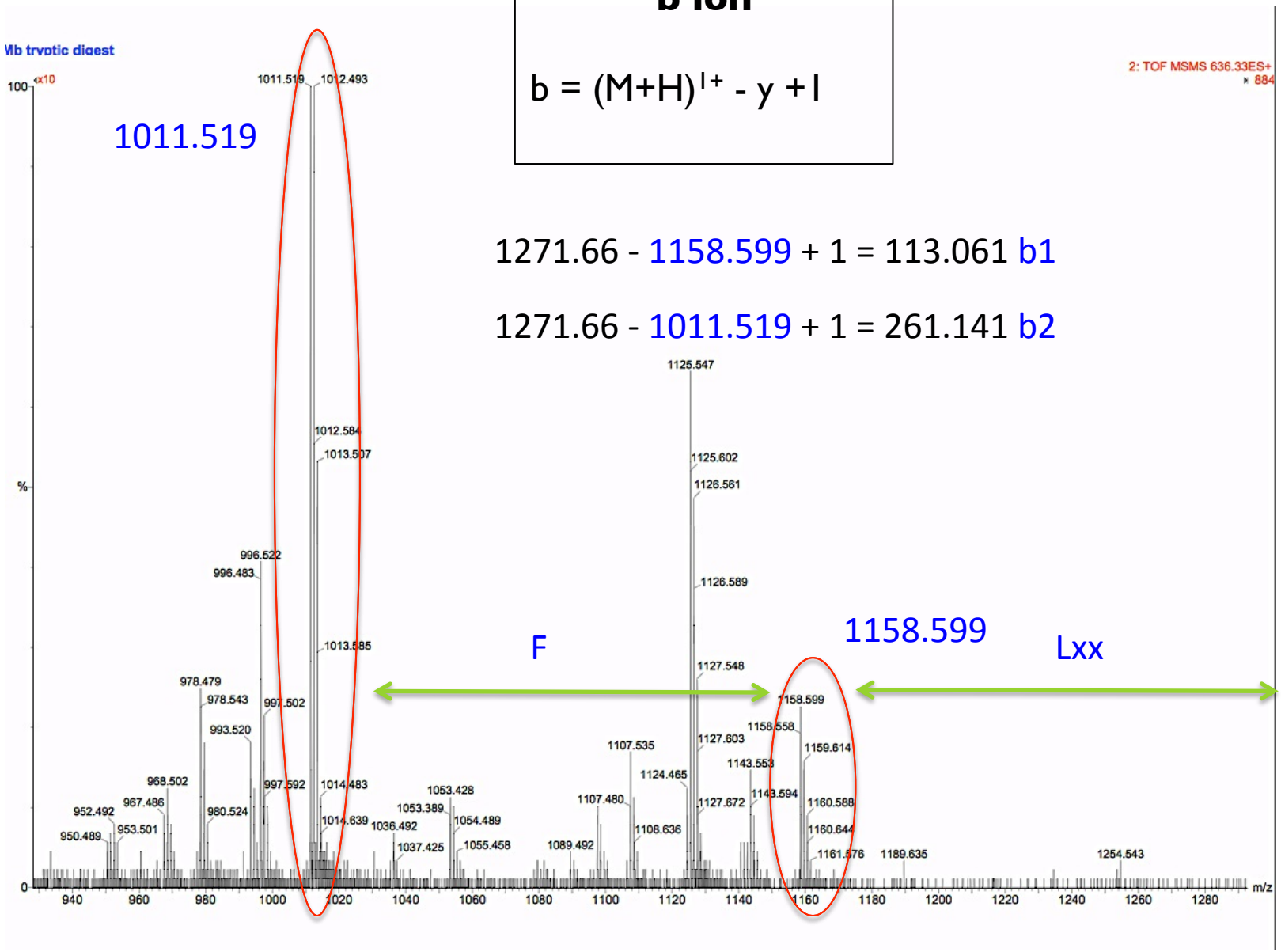

R1 O
H N C C N
H H H
c1


z3
R2 O R3 O R4 O
C C N C C N C C OH
H H H H H

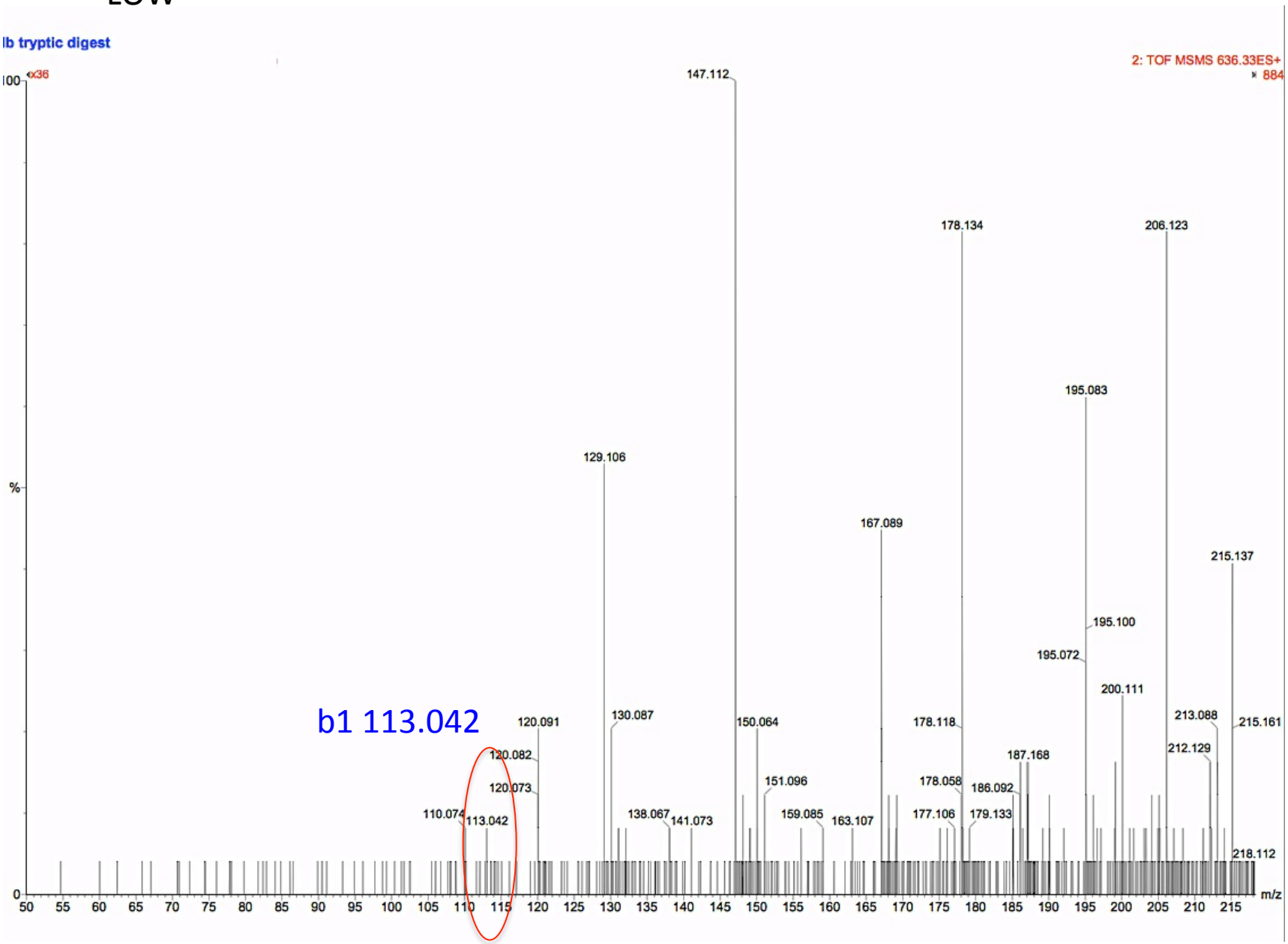
High

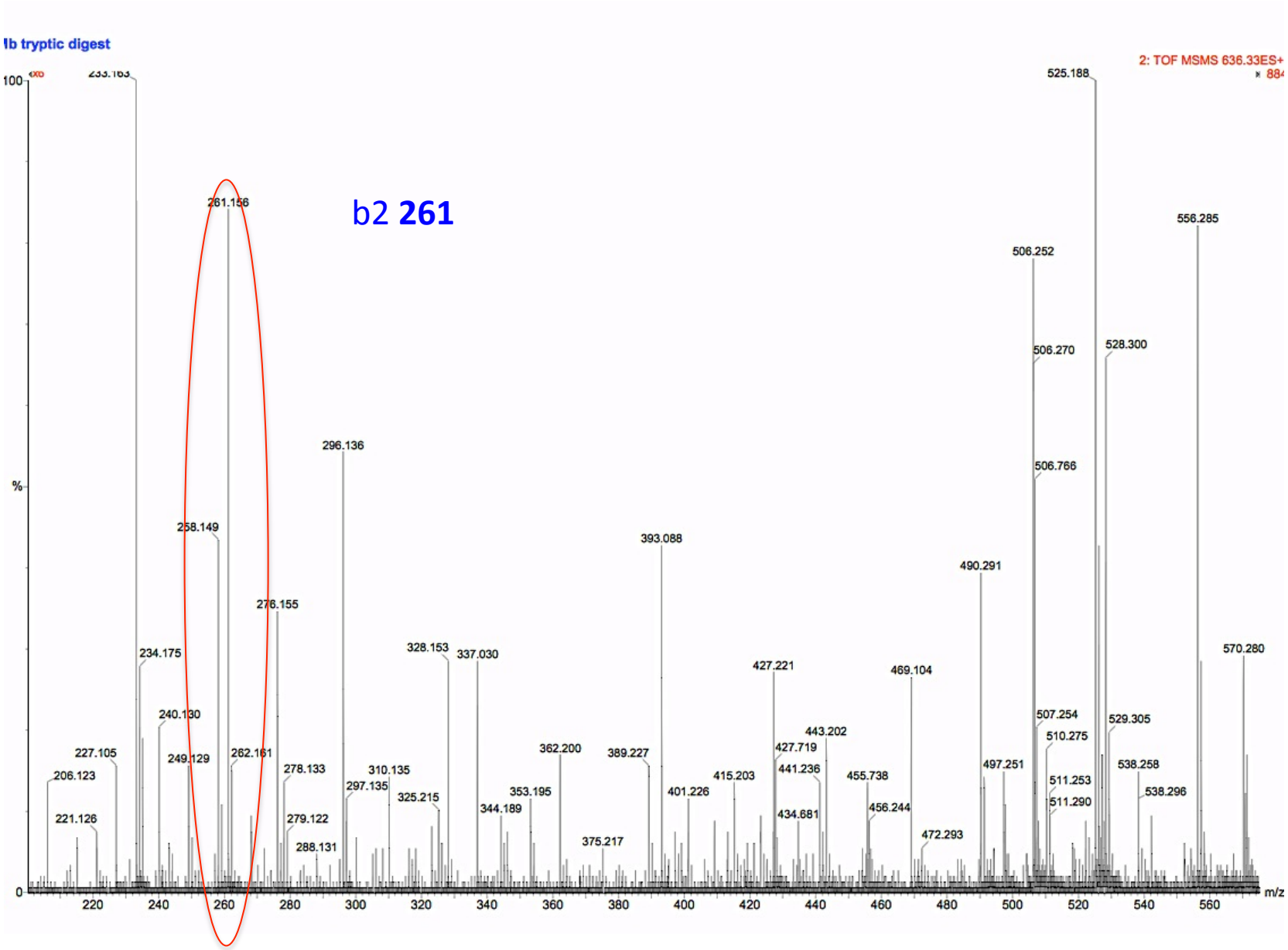
b ion

$b = (M+H)^{1+} - y + 1$



LOW


lb tryptic digest
2: TOF MSMS 636.33ES+
884
x36
100
%
0
147.112
178.134
206.123
195.083
129.106
167.089
215.137
195.100
195.072
200.111
213.088
215.161
212.129
187.168
178.118
150.064
130.087
120.091
120.082
120.073
151.096
178.058
186.092
110.074
113.042
138.067
141.073
159.085
163.107
177.106
179.133
218.112
b1 113.042
50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215
m/z

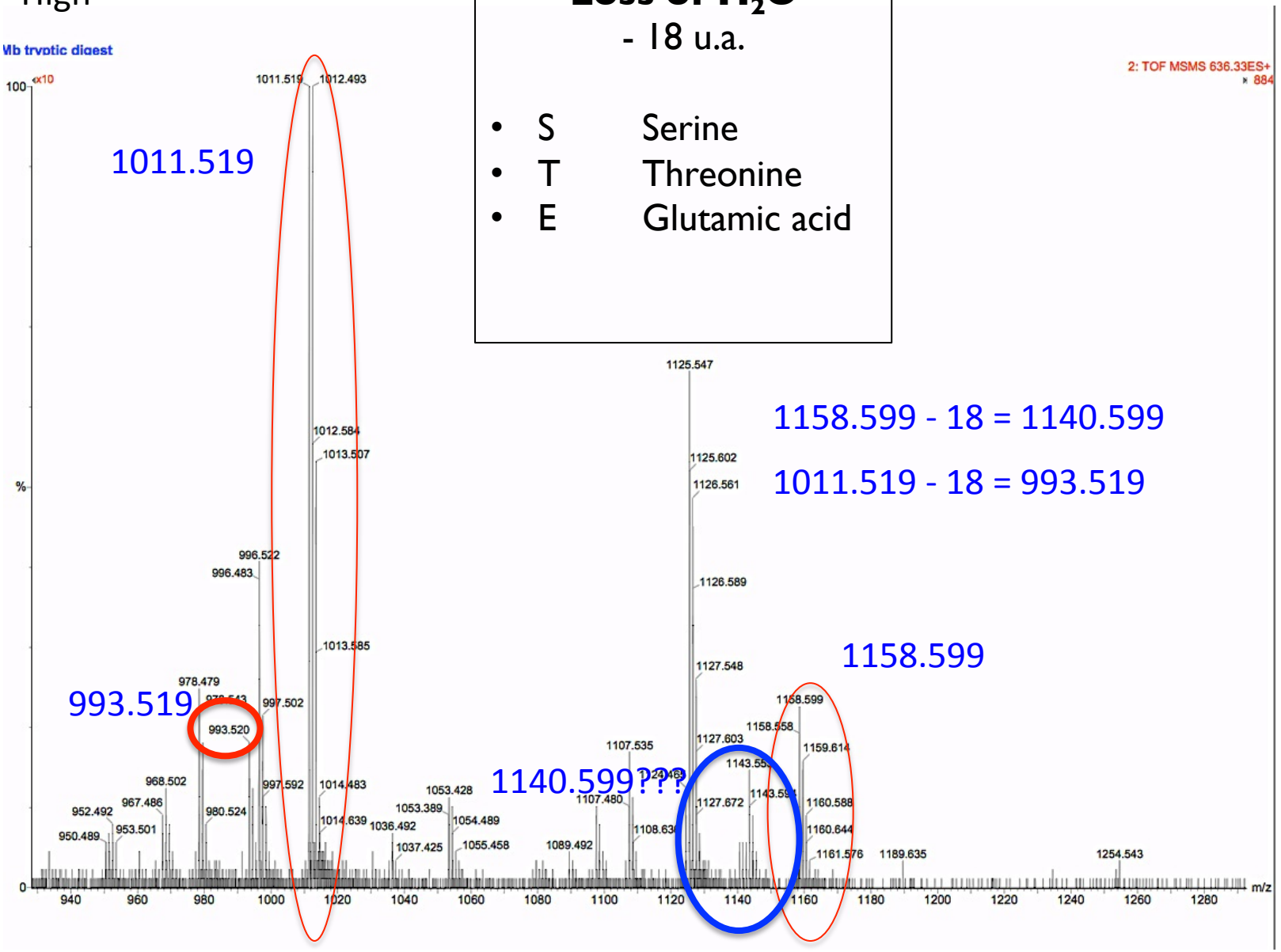

lb tryptic digest
2: TOF MSMS 636.33ES+
b2 261
100
%
0
m/z
261.156
296.136
556.285
506.252
506.270
528.300
506.766
258.149
393.088
490.291
276.155
234.175
328.153
337.030
427.221
469.104
570.280
240.130
507.254
529.305
443.202
427.719
510.275
227.105
249.129
262.161
362.200
389.227
441.236
455.738
497.251
538.258
206.123
278.133
310.135
297.135
415.203
511.253
538.296
325.215
353.195
401.226
511.290
221.126
279.122
344.189
434.681
456.244
288.131
375.217
472.293
220
240
260
280
300
320
340
360
380
400
420
440
460
480
500
520
540
560

High

Loss of H_2O

- 18 u.a.

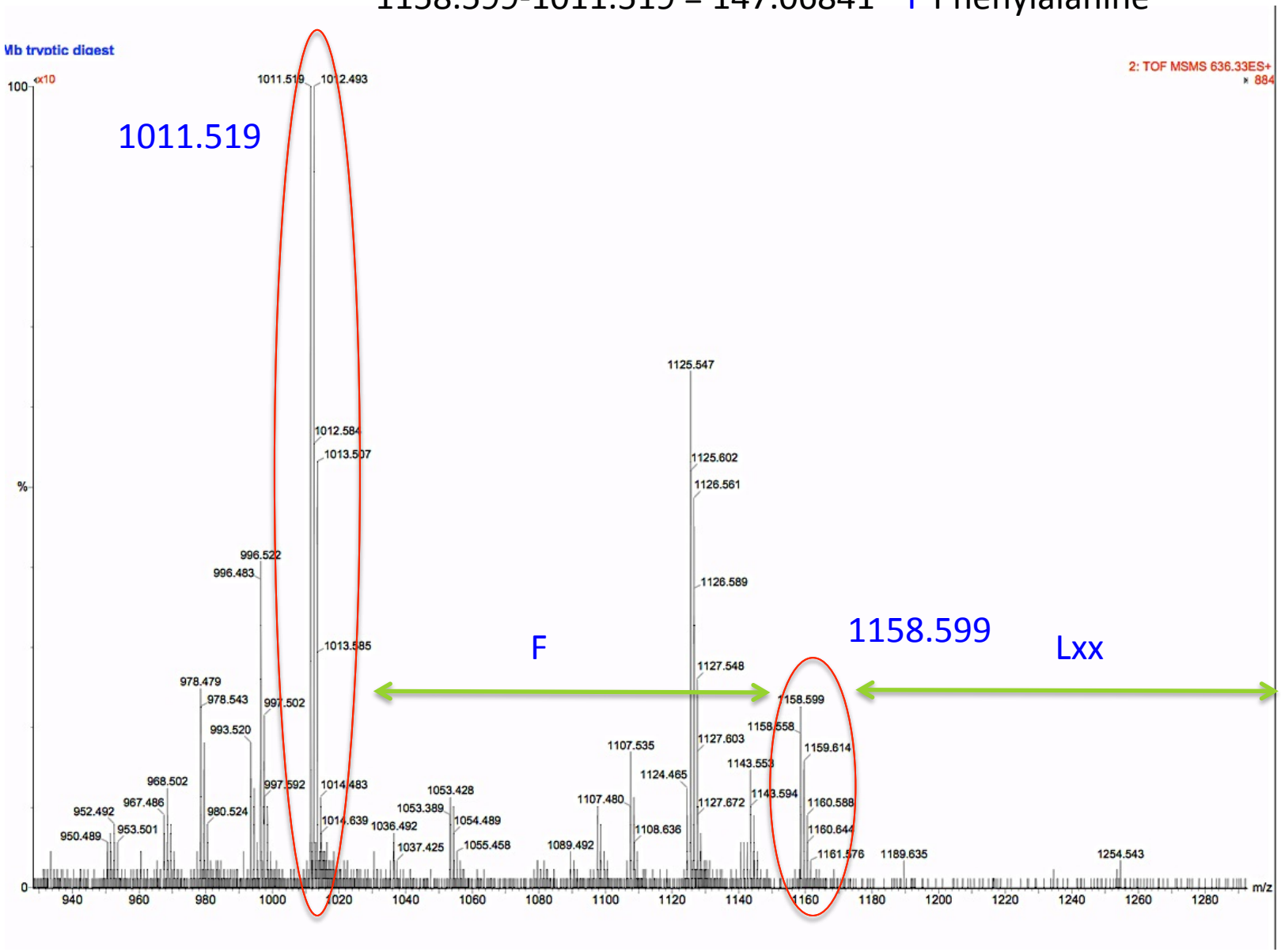
- S Serine
- T Threonine
- E Glutamic acid



High

1271.66 – 1158.599 = 113.061 X (leucine or isoleucine)

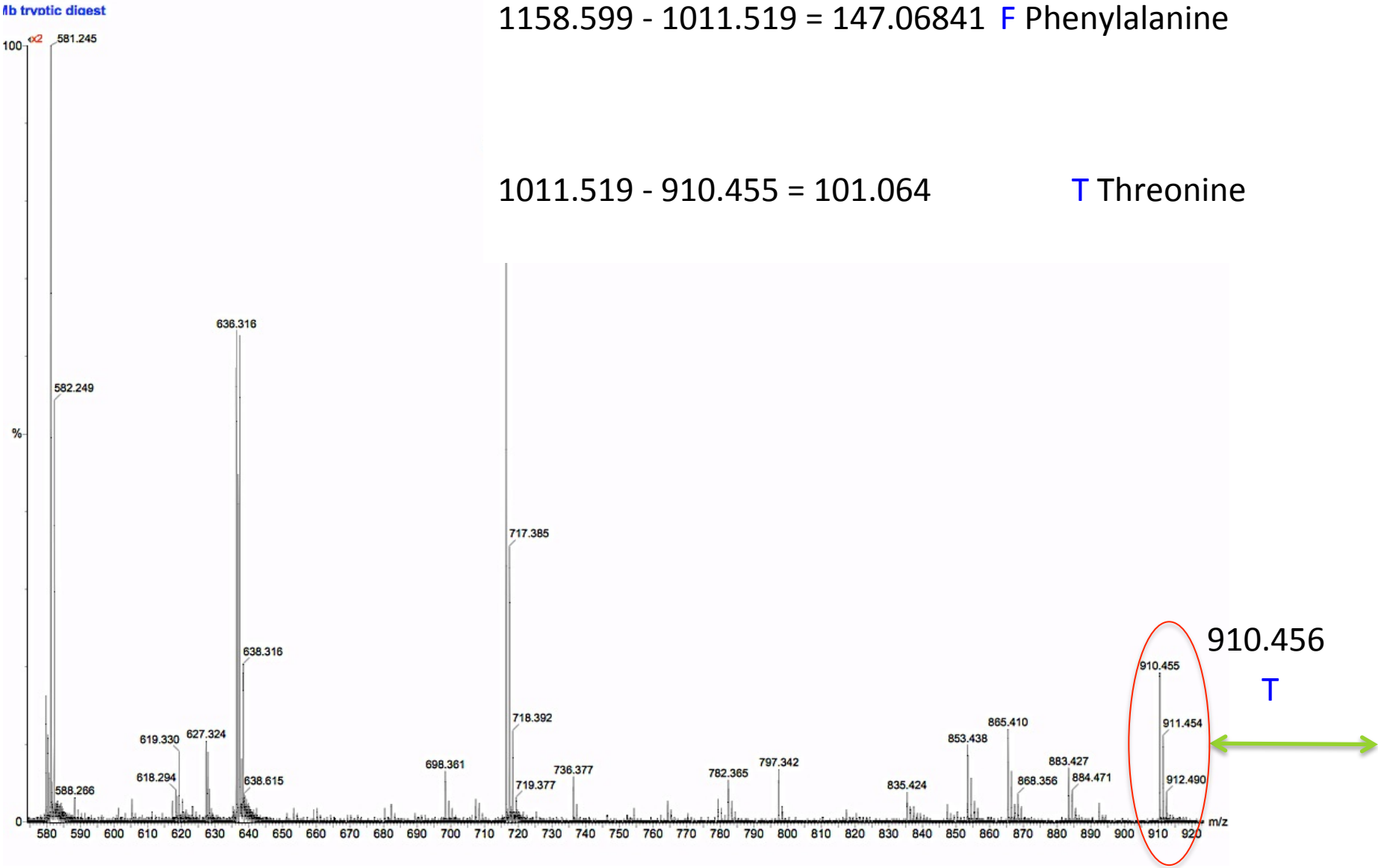
1158.599-1011.519 = 147.06841 F Phenylalanine



Medium High

1271.660 - 1254.543 = 17
1271.660 - 1158.599 = 113.061 X (leucine or isoleucine)
1158.599 - 1011.519 = 147.06841 F Phenylalanine

1011.519 - 910.455 = 101.064 T Threonine



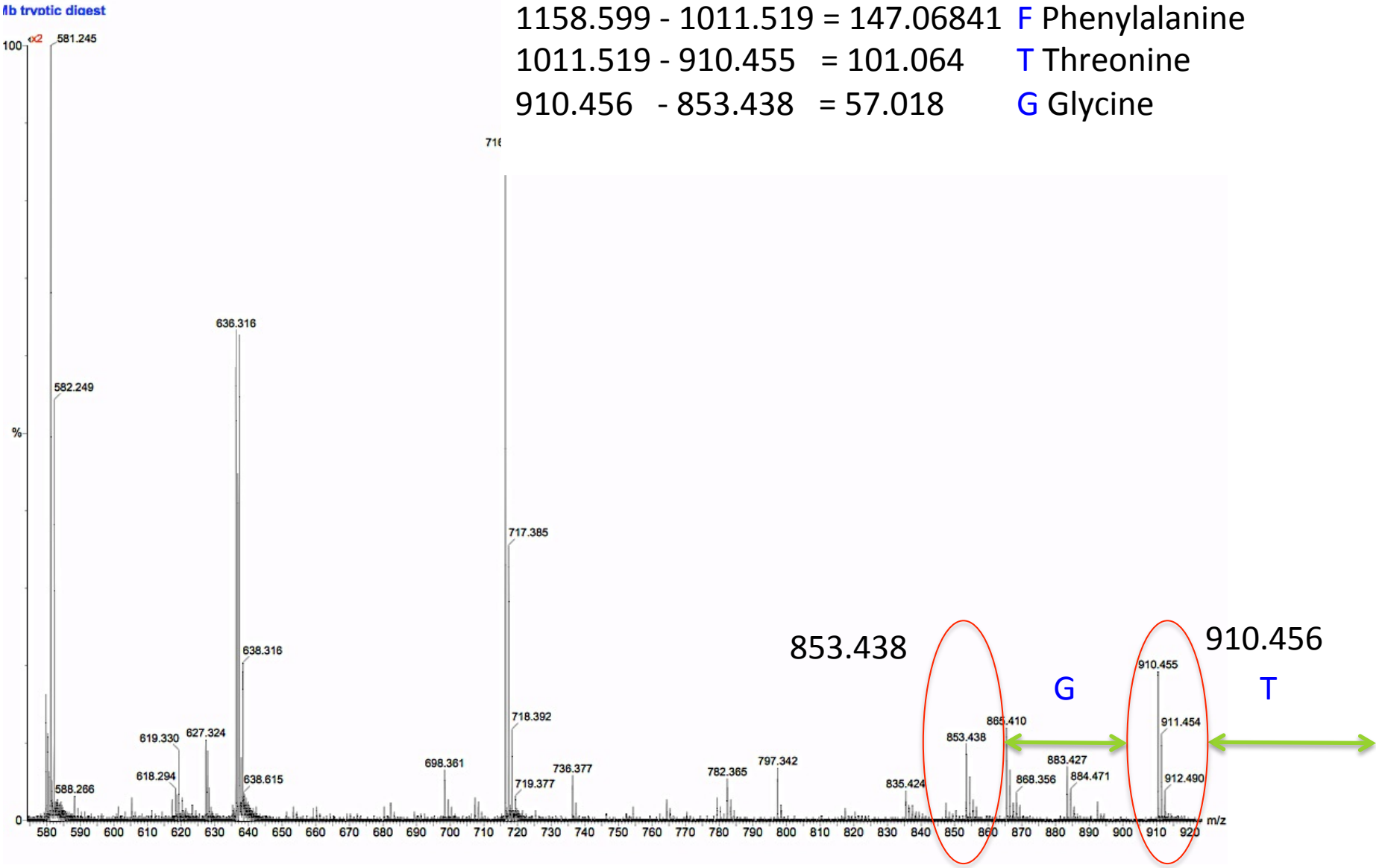
1271.660 - 1254.543 = 17

1271.660 - 1158.599 = 113.061 X (leucine or isoleucine)

1158.599 - 1011.519 = 147.06841 F Phenylalanine

1011.519 - 910.455 = 101.064 T Threonine

910.456 - 853.438 = 57.018 G Glycine



1271.660 - 1254.543 = 17

1271.660 - 1158.599 = 113.061 X (leucine or isoleucine

1158.599 - 1011.519 = 147.068 F Phenylalanine

1011.519 - 910.455 = 101.064 T Threonine

910.456 - 853.438 = 57.018 G Glycine

853.438 - 716.390 = 137.048 H Histidine

716.390

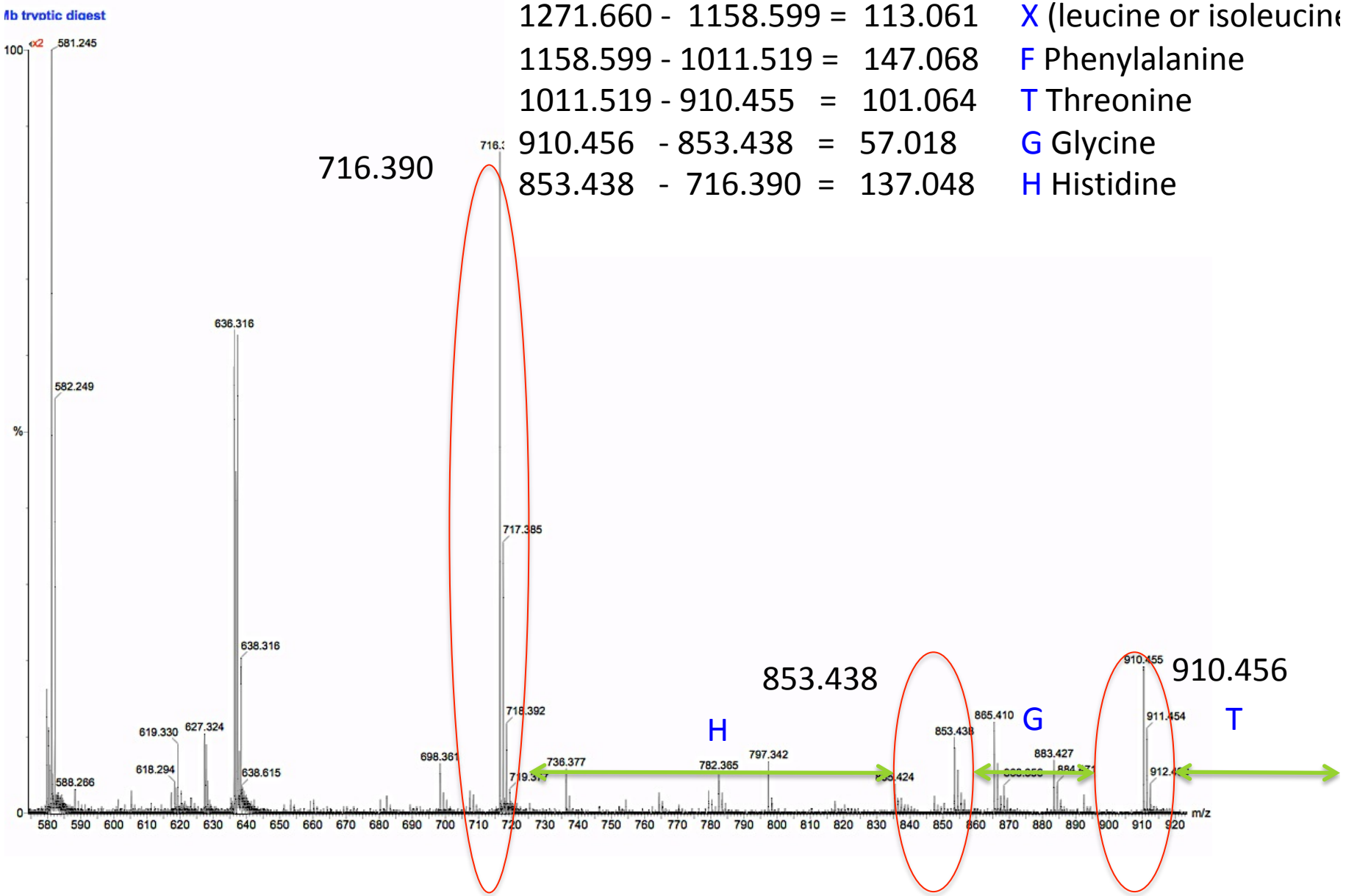
853.438

H

G

910.456

T



1271.660 - 1254.543 = 17

1271.660 - 1158.599 = 113.061 X (leucine or isoleucine

1158.599 - 1011.519 = 147.06841 F Phenylalanine

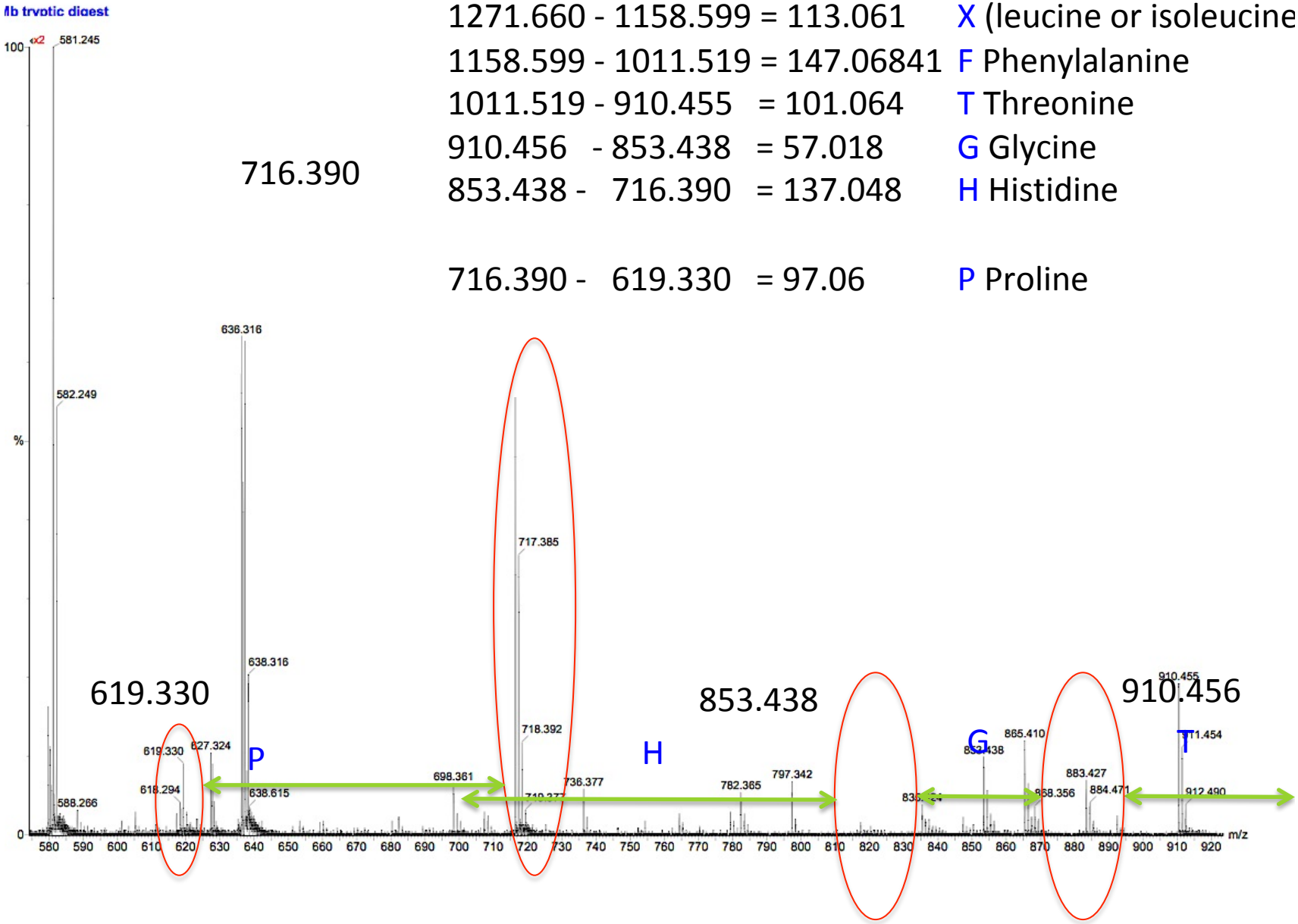
1011.519 - 910.455 = 101.064 T Threonine

910.456 - 853.438 = 57.018 G Glycine

853.438 - 716.390 = 137.048 H Histidine

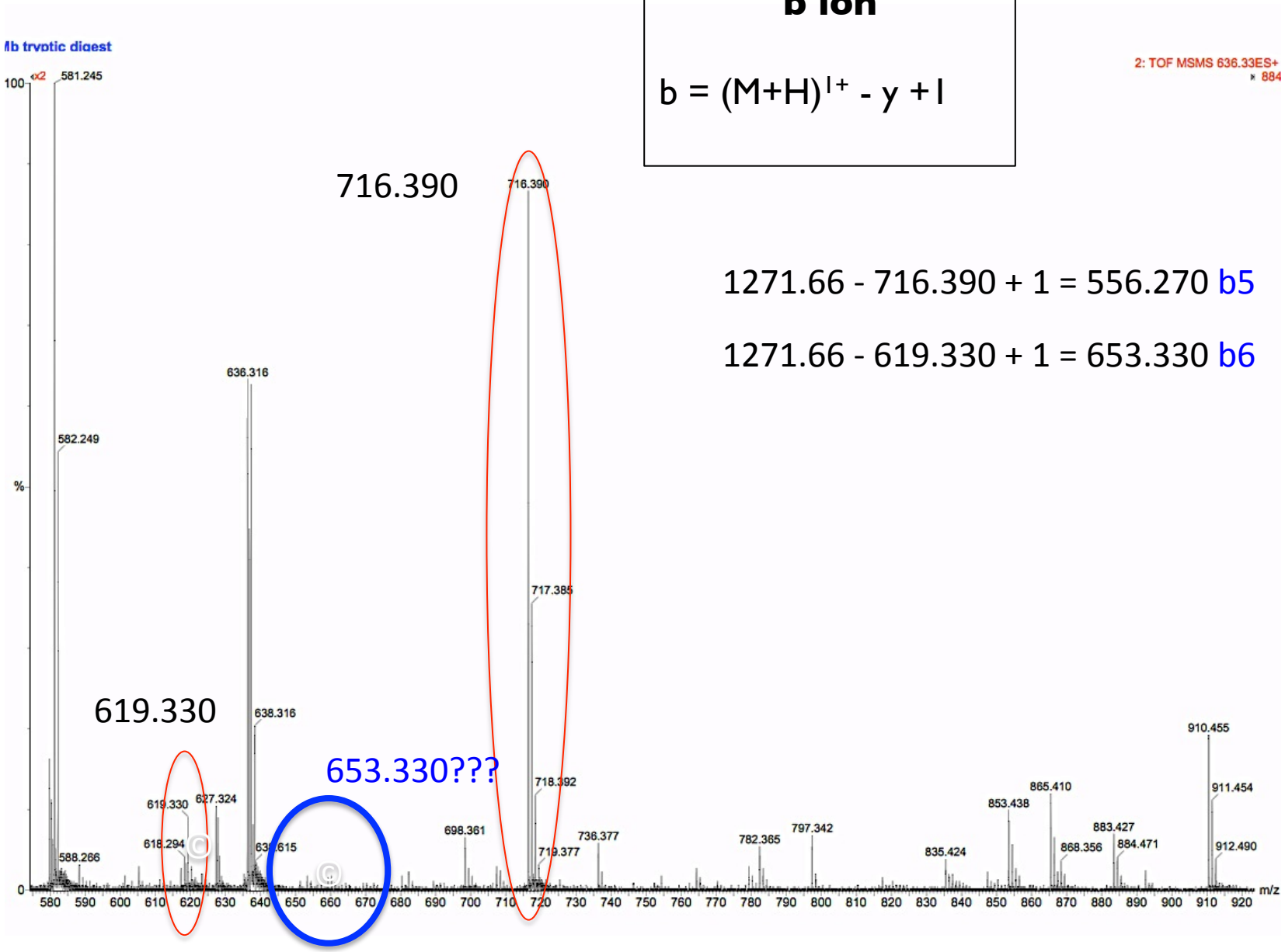
716.390

716.390 - 619.330 = 97.06 P Proline



b ion

$b = (M+H)^{1+} - y + 1$



716.390

619.330

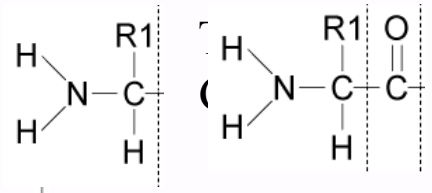
653.330???

1271.66 - 716.390 + 1 = 556.270 b5

1271.66 - 619.330 + 1 = 653.330 b6

Loss of H_2O
- 18 u.a.

- S Serine



b a

b ion

$b = (M+H)^{1+} - y + I$

556.27

528.27

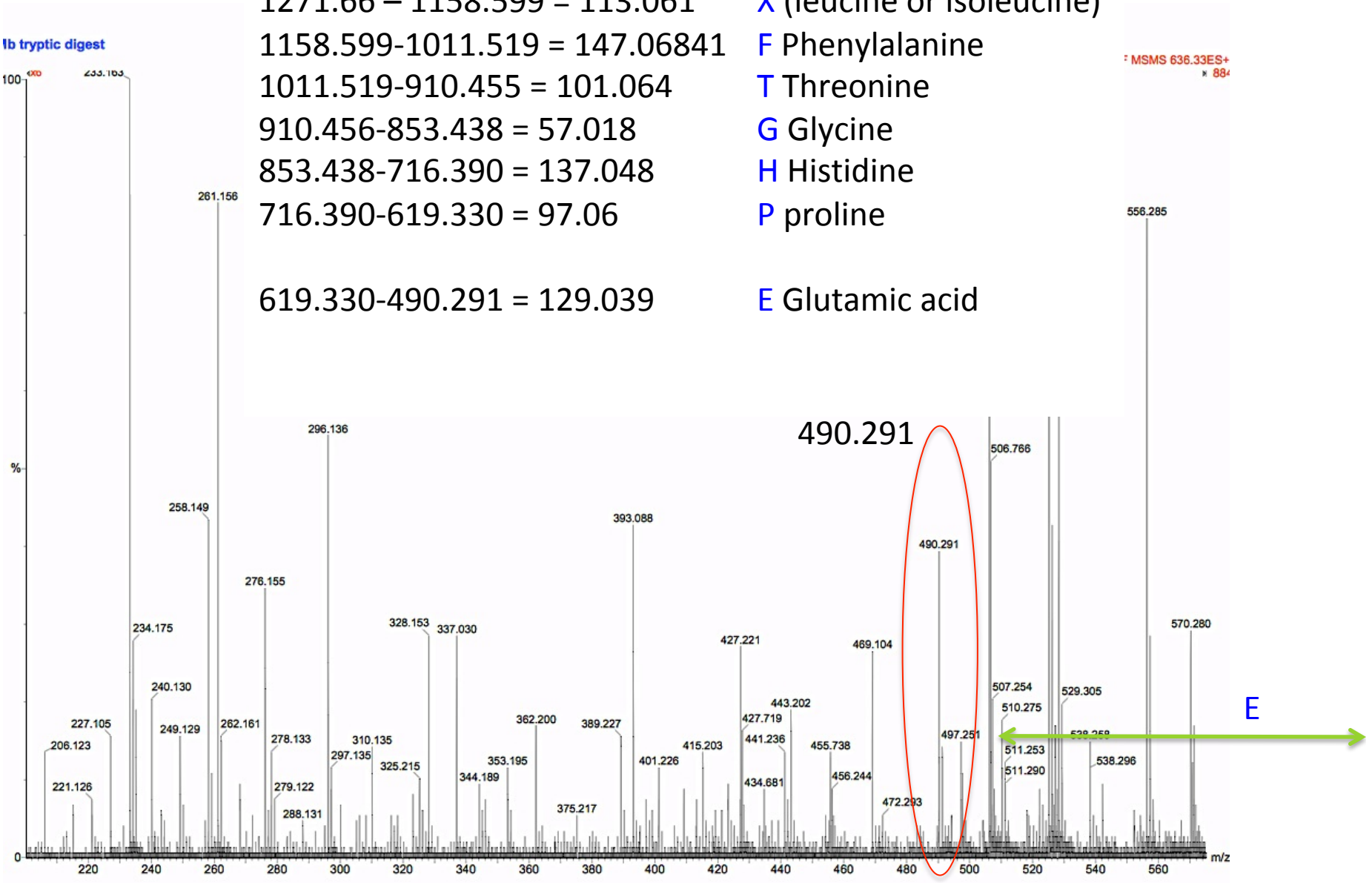
556.27-18 = 538.27

556.27-28 = 528.27

538.270

1271.66 – 1254.543 = 17

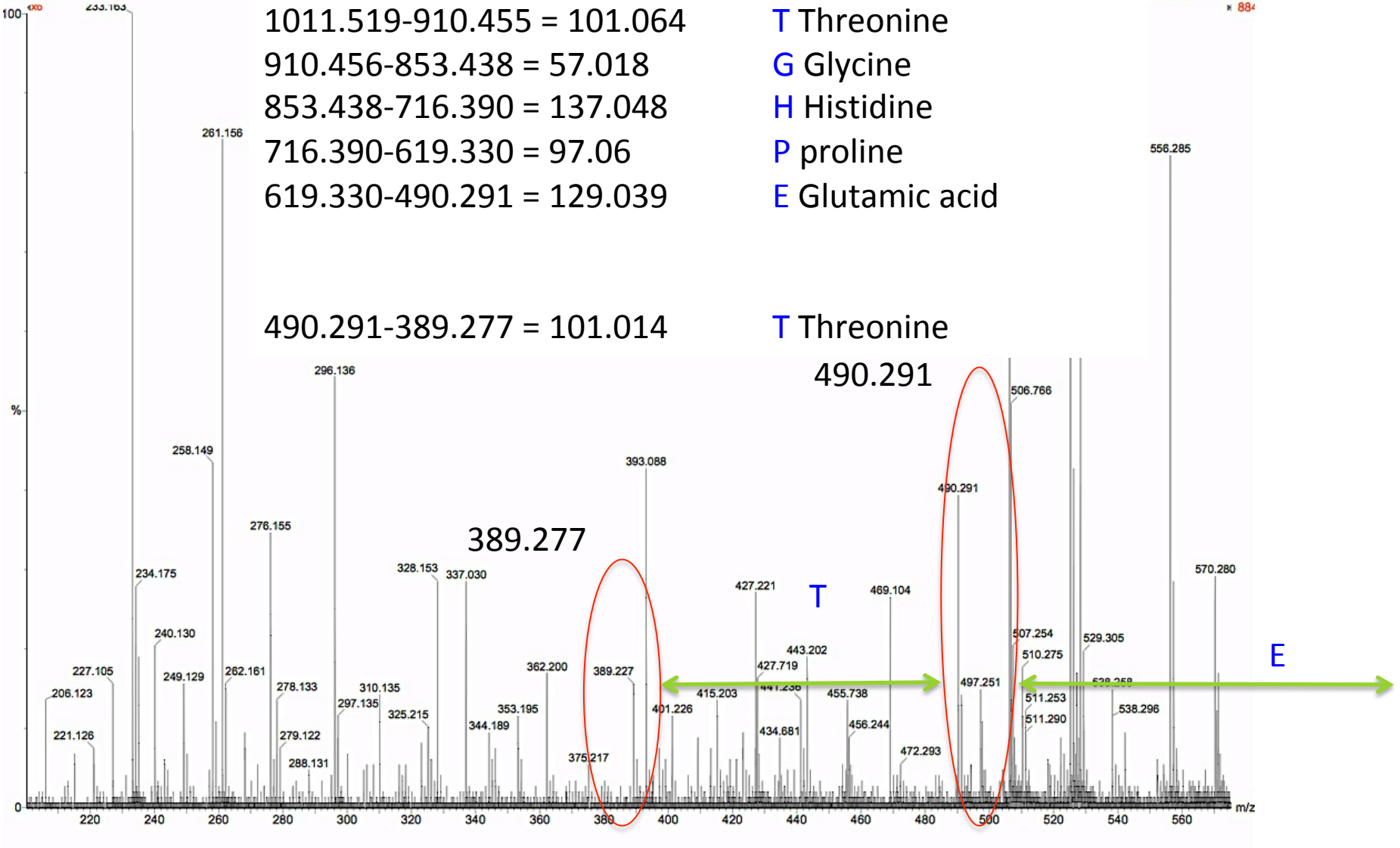
1271.66 – 1158.599 = 113.061	X (leucine or isoleucine)
1158.599-1011.519 = 147.06841	F Phenylalanine
1011.519-910.455 = 101.064	T Threonine
910.456-853.438 = 57.018	G Glycine
853.438-716.390 = 137.048	H Histidine
716.390-619.330 = 97.06	P proline
619.330-490.291 = 129.039	E Glutamic acid



1271.66 – 1254.543 = 17

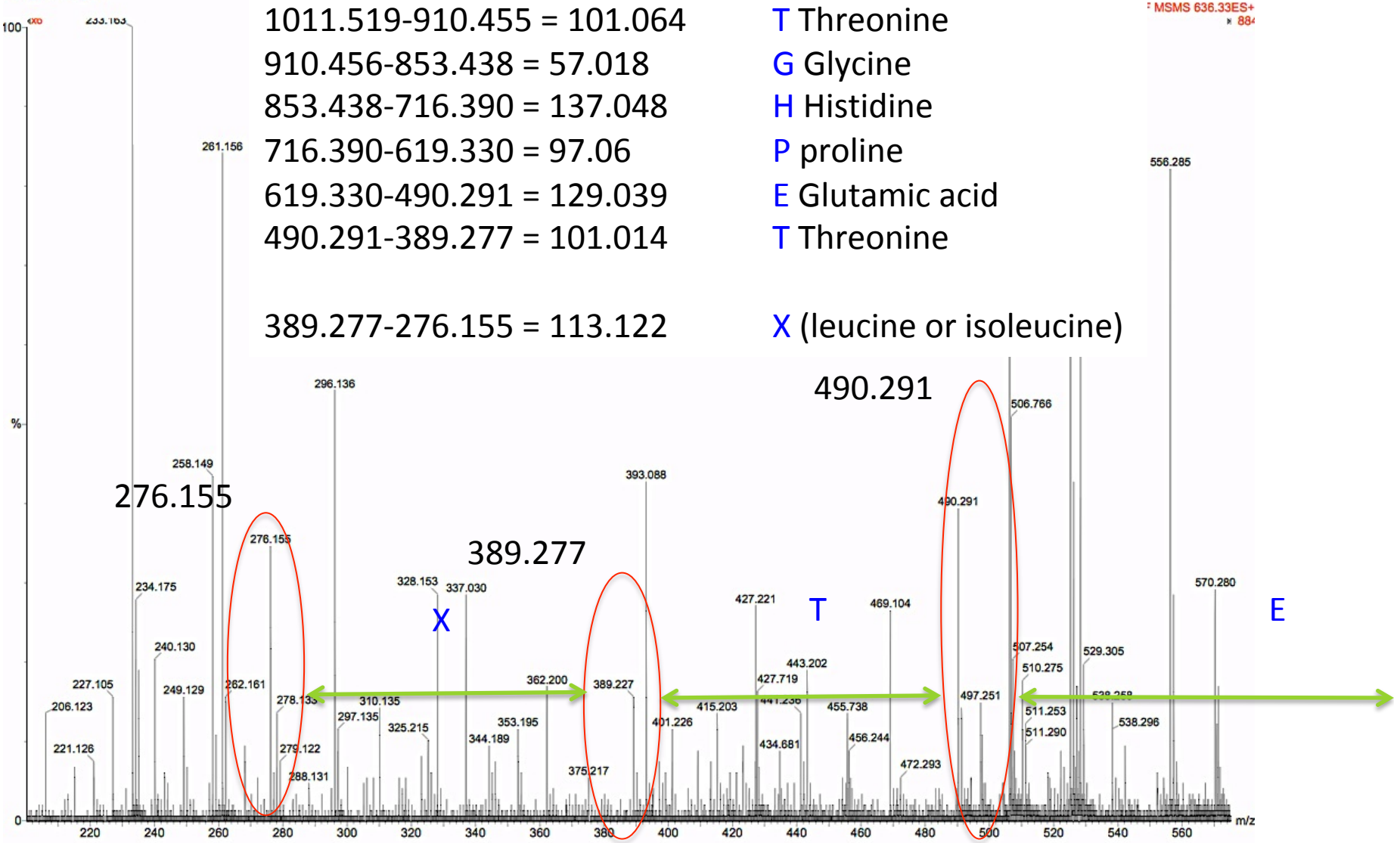
1271.66 – 1158.599 = 113.061	X (leucine or isoleucine)
1158.599-1011.519 = 147.06841	F Phenylalanine
1011.519-910.455 = 101.064	T Threonine
910.456-853.438 = 57.018	G Glycine
853.438-716.390 = 137.048	H Histidine
716.390-619.330 = 97.06	P proline
619.330-490.291 = 129.039	E Glutamic acid

490.291-389.277 = 101.014 T Threonine

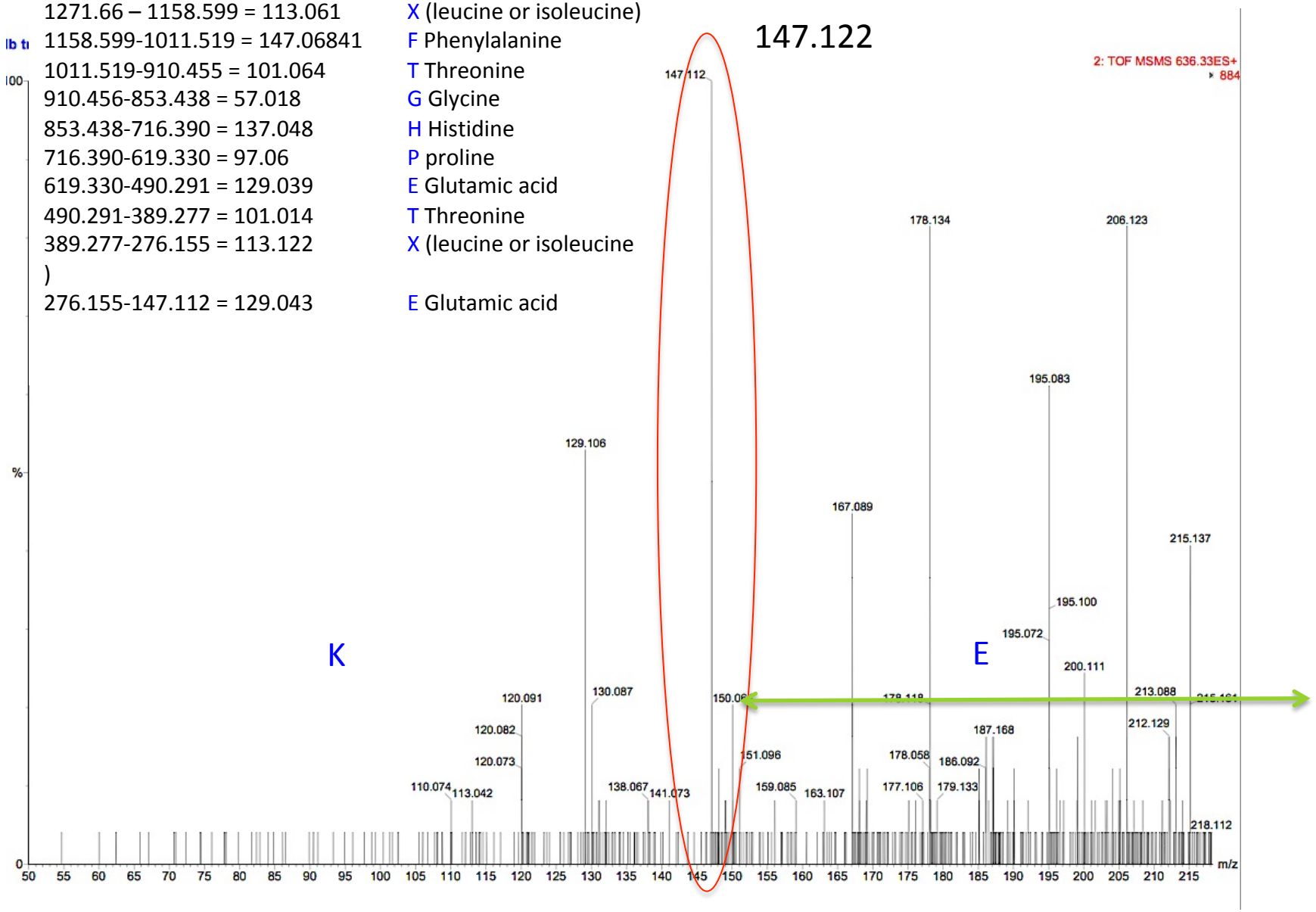


1271.66 – 1254.543 = 17

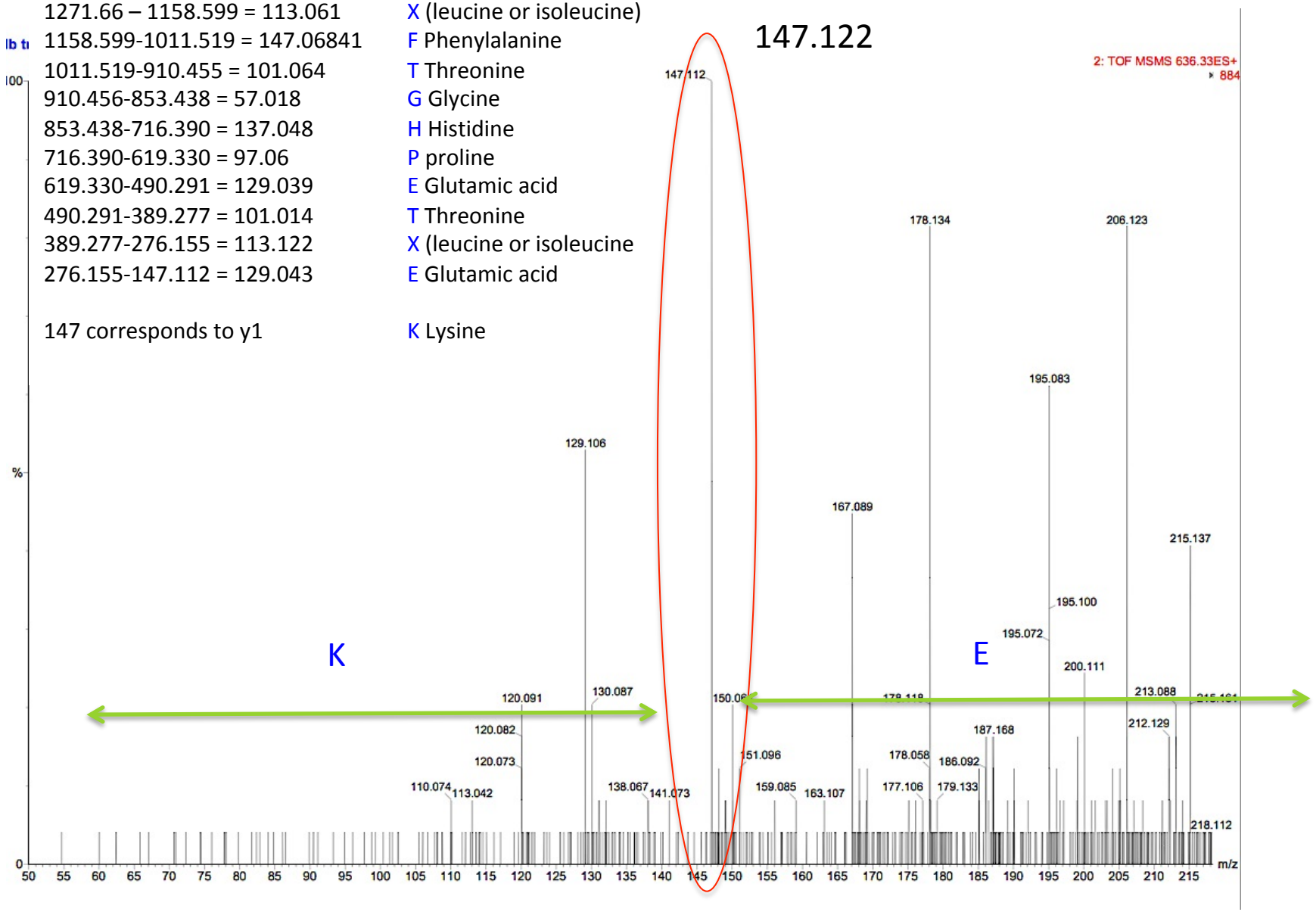
1271.66 – 1158.599 = 113.061	X (leucine or isoleucine)
1158.599-1011.519 = 147.06841	F Phenylalanine
1011.519-910.455 = 101.064	T Threonine
910.456-853.438 = 57.018	G Glycine
853.438-716.390 = 137.048	H Histidine
716.390-619.330 = 97.06	P proline
619.330-490.291 = 129.039	E Glutamic acid
490.291-389.277 = 101.014	T Threonine
389.277-276.155 = 113.122	X (leucine or isoleucine)



1271.66 – 1254.543 = 17	
1271.66 – 1158.599 = 113.061	X (leucine or isoleucine)
1158.599-1011.519 = 147.06841	F Phenylalanine
1011.519-910.455 = 101.064	T Threonine
910.456-853.438 = 57.018	G Glycine
853.438-716.390 = 137.048	H Histidine
716.390-619.330 = 97.06	P proline
619.330-490.291 = 129.039	E Glutamic acid
490.291-389.277 = 101.014	T Threonine
389.277-276.155 = 113.122	X (leucine or isoleucine
)	
276.155-147.112 = 129.043	E Glutamic acid



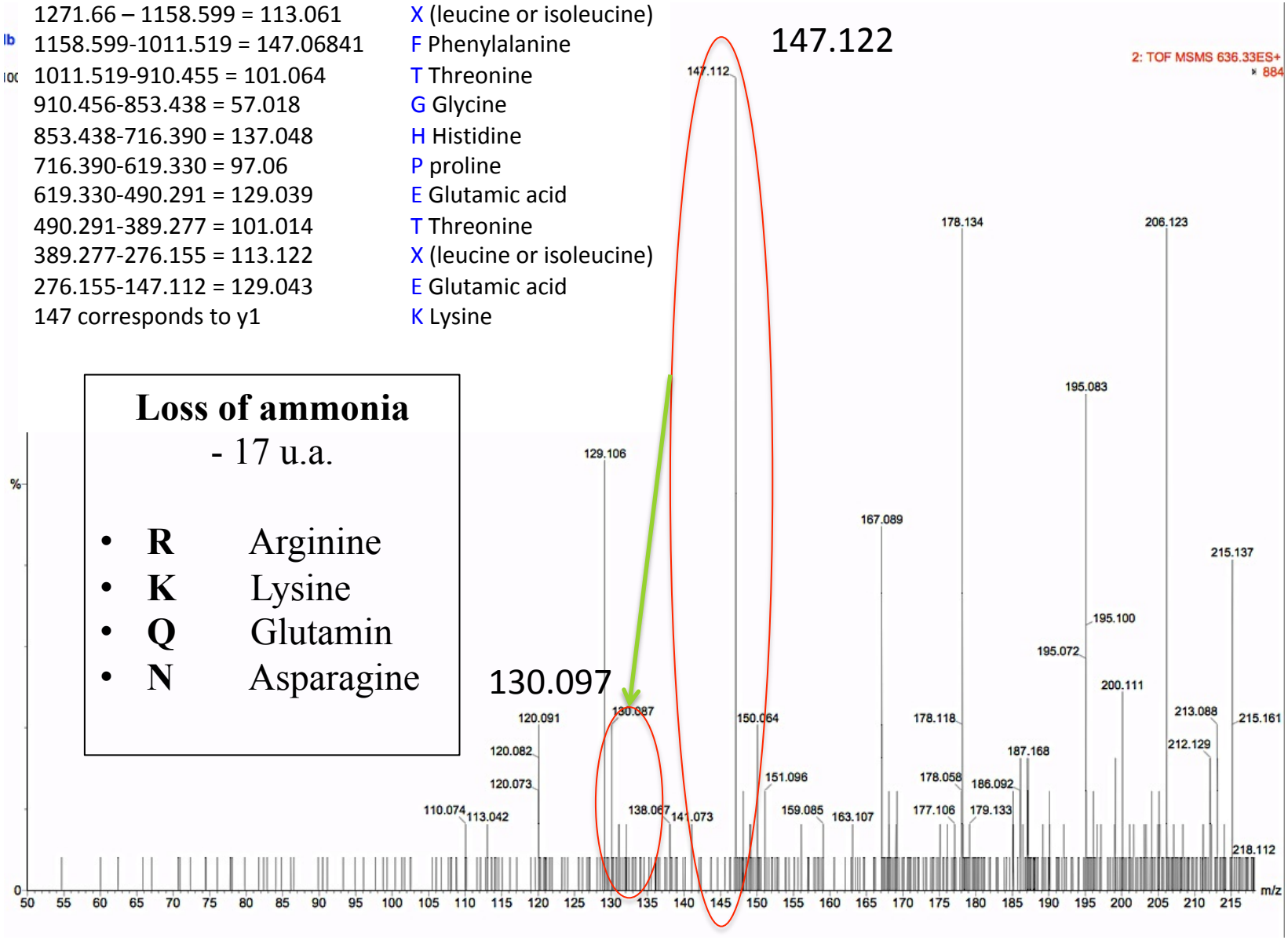
1271.66 – 1254.543 = 17	
1271.66 – 1158.599 = 113.061	X (leucine or isoleucine)
1158.599-1011.519 = 147.06841	F Phenylalanine
1011.519-910.455 = 101.064	T Threonine
910.456-853.438 = 57.018	G Glycine
853.438-716.390 = 137.048	H Histidine
716.390-619.330 = 97.06	P proline
619.330-490.291 = 129.039	E Glutamic acid
490.291-389.277 = 101.014	T Threonine
389.277-276.155 = 113.122	X (leucine or isoleucine
276.155-147.112 = 129.043	E Glutamic acid
147 corresponds to y1	K Lysine



1271.66 – 1254.543 = 17	
1271.66 – 1158.599 = 113.061	X (leucine or isoleucine)
1158.599-1011.519 = 147.06841	F Phenylalanine
1011.519-910.455 = 101.064	T Threonine
910.456-853.438 = 57.018	G Glycine
853.438-716.390 = 137.048	H Histidine
716.390-619.330 = 97.06	P proline
619.330-490.291 = 129.039	E Glutamic acid
490.291-389.277 = 101.014	T Threonine
389.277-276.155 = 113.122	X (leucine or isoleucine)
276.155-147.112 = 129.043	E Glutamic acid
147 corresponds to y1	K Lysine

Loss of ammonia
- 17 u.a.

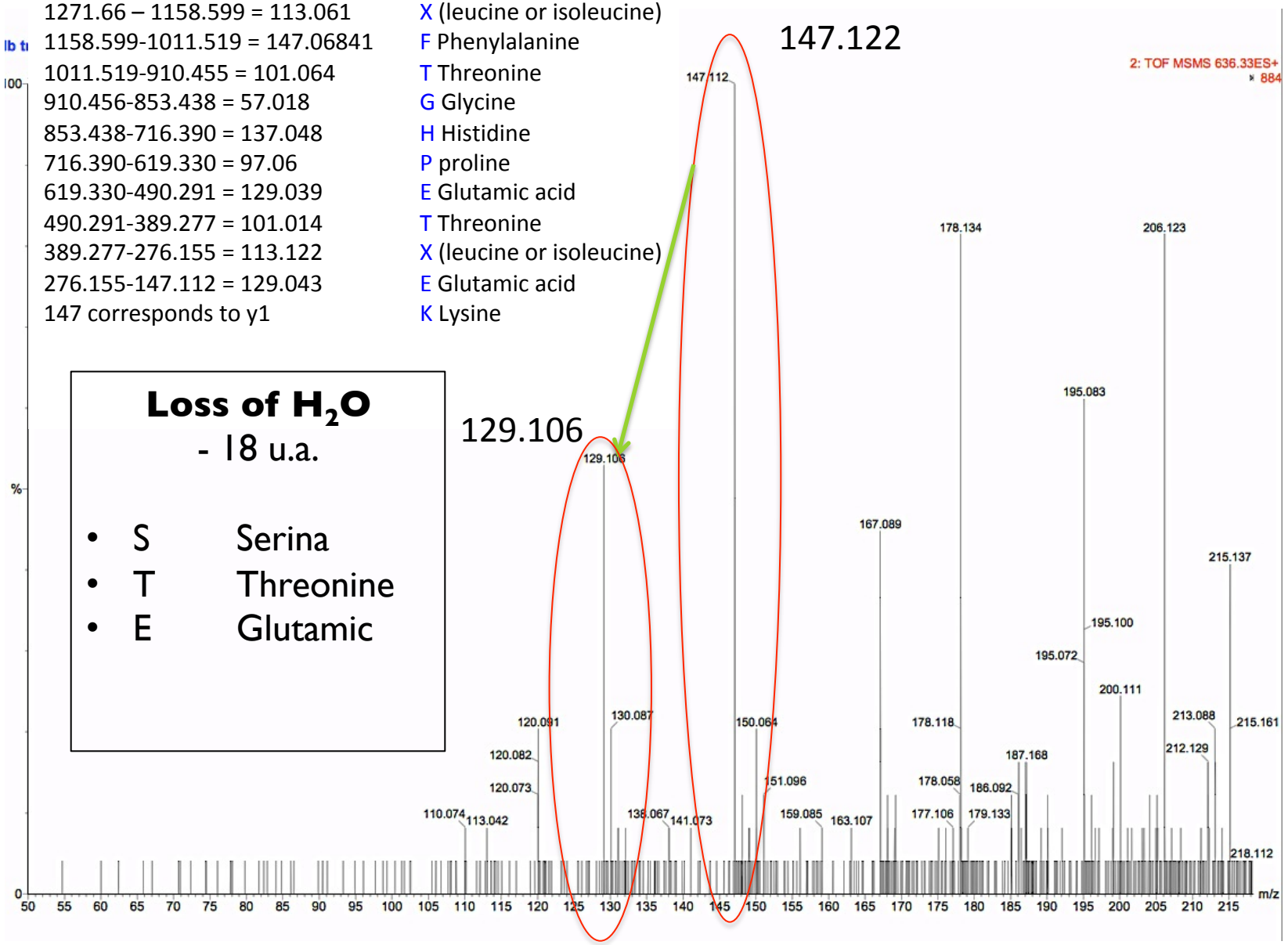
- **R** Arginine
- **K** Lysine
- **Q** Glutamin
- **N** Asparagine

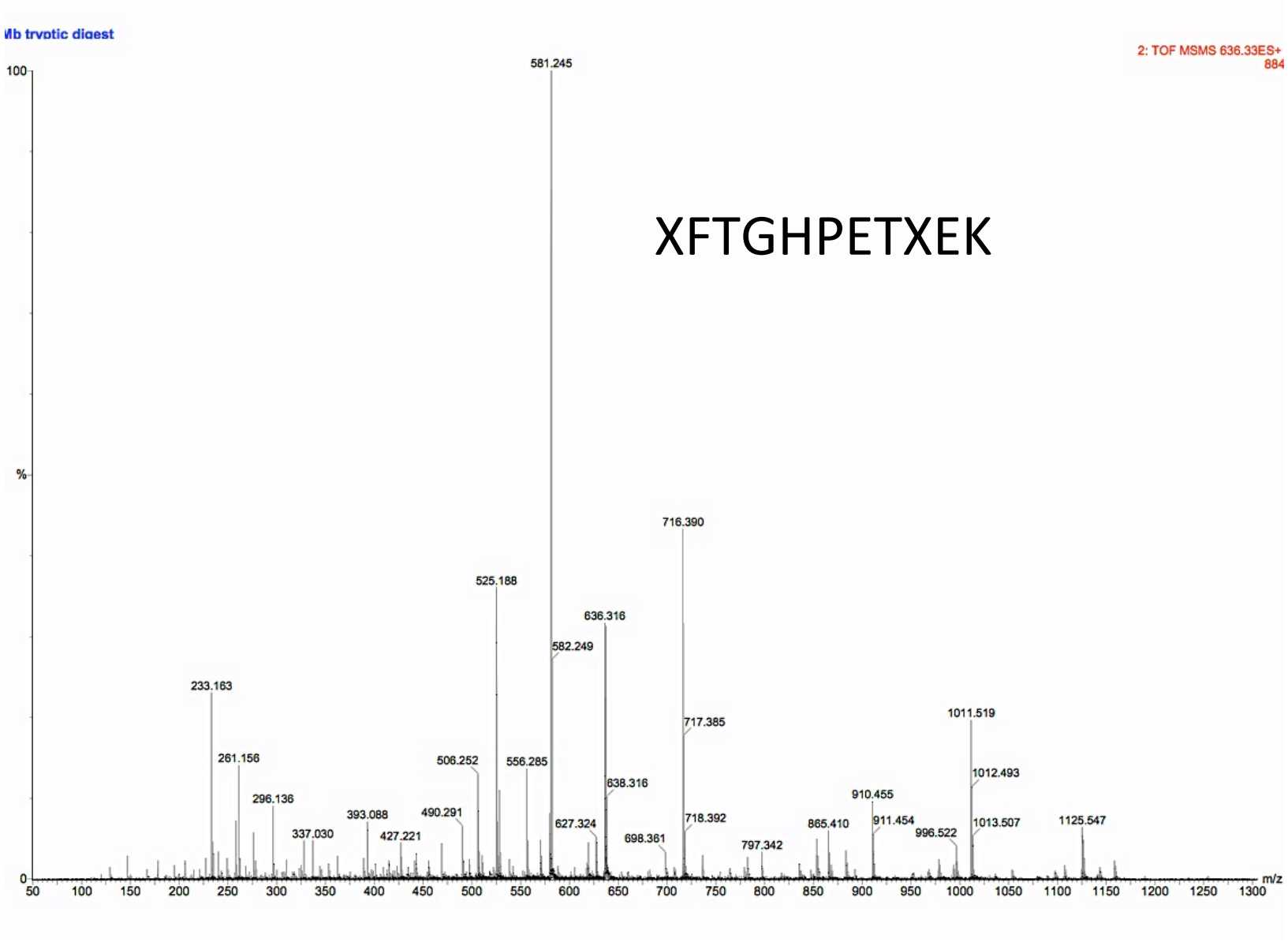


1271.66 – 1254.543 = 17	
1271.66 – 1158.599 = 113.061	X (leucine or isoleucine)
1158.599-1011.519 = 147.06841	F Phenylalanine
1011.519-910.455 = 101.064	T Threonine
910.456-853.438 = 57.018	G Glycine
853.438-716.390 = 137.048	H Histidine
716.390-619.330 = 97.06	P proline
619.330-490.291 = 129.039	E Glutamic acid
490.291-389.277 = 101.014	T Threonine
389.277-276.155 = 113.122	X (leucine or isoleucine)
276.155-147.112 = 129.043	E Glutamic acid
147 corresponds to y1	K Lysine

Loss of H_2O
- 18 u.a.

- S Serina
- T Threonine
- E Glutamic




Mb tryptic digest
2: TOF MSMS 636.33ES+
884
XFTGHPETXEK
100
%
0
581.245
716.390
525.188
636.316
582.249
233.163
717.385
1011.519
261.156
506.252
556.285
638.316
1012.493
296.136
910.455
393.088
490.291
718.392
627.324
865.410
911.454
1013.507
1125.547
337.030
427.221
698.361
996.522
797.342
m/z
50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800 850 900 950 1000 1050 1100 1150 1200 1250 1300

Key points

- It seems easy, but actually it is not.
- Leucine and isoleucine have isobaric masses and can not be distinguished.
- Lysine and glutamine have a difference of 0.03638. they can only be differentiated using high resolution mass spectrometry.

- Two
resic

Di-peptide Table

Note: Where di-peptides conflict with single amino acid the masses are noted in blue.
Masses in red are notations where di-peptides conflict with other di-peptides.

IS IonSource		G	A	S	P	V	T	C	L	I	N	D	Q	K	E	M	H	F	R	CMC	Y	W
		57	71	87	97	99	101	103	113	113	114	115	128	128	129	131	137	147	156	161	163	186
G	57	114	128	144	154	156	158	160	170	170	171	172	185	185	186	188	194	204	213	218	220	243
A	71	128	142	158	168	170	172	174	184	184	185	186	199	199	200	202	208	218	227	232	234	257
S	87	144	158	174	184	186	188	190	200	200	201	202	215	215	216	218	224	234	243	248	250	273
P	97	154	168	184	194	196	198	200	210	210	211	212	225	225	226	228	234	244	253	258	260	283
V	99	156	170	186	196	198	200	202	212	212	213	214	227	227	228	230	236	246	255	260	262	285
T	101	158	172	188	198	200	202	204	214	214	215	216	229	229	230	232	238	248	257	262	264	287
C	103	160	174	190	200	202	204	206	216	216	217	218	231	231	232	234	240	250	259	264	266	289
L	113	170	184	200	210	212	214	216	226	226	227	228	241	241	242	244	250	260	269	274	276	299
I	113	170	184	200	210	212	214	216	226	226	227	228	241	241	242	244	250	260	269	274	276	299
N	114	171	185	201	211	213	215	217	227	227	228	229	242	242	243	245	251	261	270	275	277	300
D	115	172	186	202	212	214	216	218	228	228	229	230	243	243	244	246	252	262	271	276	278	301
Q	128	185	199	215	225	227	229	231	241	241	242	243	256	256	257	259	265	275	284	289	291	314
K	128	185	199	215	225	227	229	231	241	241	242	243	256	256	257	259	265	275	284	289	291	314
E	129	186	200	216	226	228	230	232	242	242	243	244	257	257	258	260	266	276	285	290	292	315
M	131	188	202	218	228	230	232	234	244	244	245	246	259	259	260	262	268	278	287	292	294	317
H	137	194	208	224	234	236	238	240	250	250	251	252	265	265	266	268	274	284	293	298	300	323
F	147	204	218	234	244	246	248	250	260	260	261	262	275	275	276	278	284	294	303	308	310	333
R	156	213	227	243	253	255	257	259	269	269	270	271	284	284	285	287	293	303	312	317	319	342
CMC	161	218	232	248	258	260	262	264	274	274	275	276	289	289	290	292	298	308	317	322	324	347
Y	163	220	234	250	260	262	264	266	276	276	277	278	291	291	292	294	300	310	319	324	326	349
W	186	243	257	273	283	285	287	289	299	299	300	301	314	314	315	317	323	333	342	347	349	372

Conflicting Masses

Amino Acids	Mono		Mono	delta mass
Val	99.068414	acetyl-Gly	99.032034	0.03638
Leu	113.08406	Ile	113.08406	0
Lxx	113.08406	acetyl-Ala	113.047684	0.036376
Asn	114.04293	Gly-Gly	114.04298	0.000002
Gln	128.05858	Lys	128.09496	0.03638
Gln	128.05858	Gly-Ala	128.058578	0.000002
Lys	128.09496	Gly-Ala	128.058578	0.036382
Glu	129.04259	acetyl-Ser	129.042599	0.000009
Phe	147.06841	Met Sulfoxide	147.0354	0.033
Arg	156.10111	Val-Gly	156.089878	0.011232
Arg	156.10111	acetyl-Asn	156.0535	0.04761
Tyr	163.06333	Met Sulfone	163.0303	0.033
Trp	186.07931	Ala-Asp	186.064054	0.015256
Trp	186.07931	Ser-Val	186.100443	0.021133
Trp	186.07931	Gly-Glu	186.064054	0.015256

References / Further reading

[1] Edman, P.; Högfeldt, Erik; Sillén, Lars Gunnar; Kinell, Per-Olof (1950). "Method for determination of the amino acid sequence in peptides". Acta Chem. Scand. 4: 283–293.
doi:10.3891/acta.chem.scand.04-0283.

[2] Edman P, Begg G (March 1967). "A protein sequenator". Eur J Biochem. 1 (1): 80–91. PMID 6059350.

[3] Niall HD (1973). "Automated Edman degradation: the protein sequenator". Meth. Enzymol. Methods in Enzymology. 27: 942–1010.doi:10.1016/S0076-6879(73)27039-8. ISBN 978-0-12-181890-6. PMID 4773306.

[4] Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom. 1984 Nov; 11 (11):601.

References / Further reading

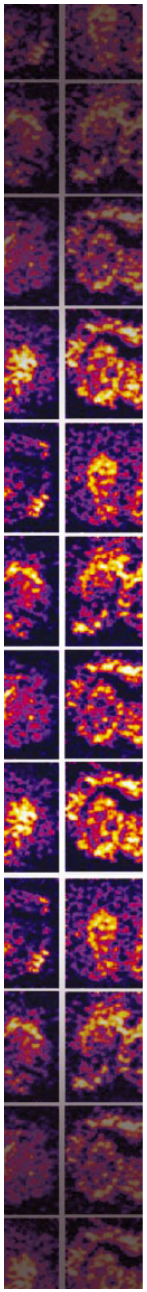
[5] K. Biemann In: J.A. McCloskey, Editor, Methods in Enzymology 193, Academic, San Diego (1990), pp. 886–887.

[6] K. Biemann. Contributions of mass spectrometry to peptide and protein structure. Biomed Environ Mass Spectrom. 1988 Oct; 16(1-12):99-111.

[7] J. Seidler, et al. De novo sequencing of peptides by MS/MS. Proteomics 2010, 10, 634–649

[8] http://www.ionsource.com/tutorial/DeNovo/DeNovoTOC.htm.

[9] http://www.matrixscience.com/help/interpretation_help.html.



Mass spectrometry-based proteomics

Ruedi Aebersold* & Matthias Mann†

**Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA (e-mail: raebersold@systemsbiology.org)*
†Center for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark (e-mail: mann@bmb.sdu.dk)

Recent successes illustrate the role of mass spectrometry-based proteomics as an indispensable tool for molecular and cellular biology and for the emerging field of systems biology. These include the study of protein–protein interactions via affinity-based isolations on a small and proteome-wide scale, the mapping of numerous organelles, the concurrent description of the malaria parasite genome and proteome, and the generation of quantitative protein profiles from diverse species. The ability of mass spectrometry to identify and, increasingly, to precisely quantify thousands of proteins from complex samples can be expected to impact broadly on biology and medicine.

Proteomics in general deals with the large-scale determination of gene and cellular function directly at the protein level. But as the accompanying articles in this issue describe, the field is a collection of various technical disciplines, all of which contribute to proteomics. These include cell imaging by light and electron microscopy, array and chip experiments, and genetic readout experiments, as exemplified by the yeast two-hybrid assay. Another powerful proteomic approach focuses on the *de novo* analysis of proteins or protein populations isolated from cells or tissues. Such studies typically pose challenges owing to the high degree of complexity of cellular proteomes and the low abundance of many of the proteins, which necessitates highly sensitive analytical techniques. Mass spectrometry (MS) has increasingly become the method of choice for analysis of complex protein samples. MS-based proteomics is a discipline made possible by the availability of gene and genome sequence databases and technical and conceptual advances in many areas, most notably the discovery and development of protein ionization methods, as recognized by the 2002 Nobel prize in chemistry.

Here we survey the state of the field, particularly as it has evolved over the three years since the last review in these pages[1]. Already, many of the dreams of the discipline have at least been partly realized. MS-based proteomics has established itself as an indispensable technology to interpret the information encoded in genomes. So far, protein analysis (primary sequence, post-translational modifications (PTMs) or protein–protein interactions) by MS has been most successful when applied to small sets of proteins isolated in specific functional contexts. The systematic analysis of the much larger number of proteins expressed in a cell, an explicit goal of proteomics, is now also rapidly advancing, due mainly to the development of new experimental approaches.

Today, proteomics still remains a multifaceted, rapidly developing and open-ended endeavour. Although it has enjoyed tremendous recent success, proteomics still faces significant technical challenges. Each breakthrough that either allows a new type of measurement or improves the quality of data made by traditional types of measurements expands the range of potential applications of MS to molecular and cellular biology. Indeed, this field is already too expansive for a comprehensive, single review; thus we apologize in advance for the many omissions. However, we do hope that this article captures the excitement of recent achievements in MS-based proteomics, and points the way towards the direction future developments will likely take.

Principles and instrumentation

Mass spectrometric measurements are carried out in the gas phase on ionized analytes. By definition, a mass spectrometer consists of an ion source, a mass analyser that measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector that registers the number of ions at each m/z value. Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two techniques most commonly used to volatize and ionize the proteins or peptides for mass spectrometric analysis[2,3]. ESI ionizes the analytes out of a solution and is therefore readily coupled to liquid-based (for example, chromatographic and electrophoretic) separation tools (Fig. 1). MALDI sublimates and ionizes the samples out of a dry, crystalline matrix via laser pulses. MALDI-MS is normally used to analyse relatively simple peptide mixtures, whereas integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples.

The mass analyser is, literally and figuratively, central to the technology. In the context of proteomics, its key parameters are sensitivity, resolution, mass accuracy and the ability to generate information-rich ion mass spectra from peptide fragments (tandem mass or MS/MS spectra) (see Fig. 1 and refs 1,4,5). There are four basic types of mass analyser currently used in proteomics research. These are the ion trap, time-of-flight (TOF), quadrupole and Fourier transform ion cyclotron (FT-MS) analysers. They are very different in design and performance, each with its own strength and weakness. These analysers can be stand alone or, in some cases, put together in tandem to take advantage of the strengths of each (Fig. 2).

In ion-trap analysers, the ions are first captured or 'trapped' for a certain time interval and are then subjected to MS or MS/MS analysis. Ion traps are robust, sensitive and relatively inexpensive, and so have produced much of the proteomics data reported in the literature. A disadvantage of ion traps is their relatively low mass accuracy, due in part to the limited number of ions that can be accumulated at their point-like centre before space-charging distorts their distribution and thus the accuracy of the mass measurement. The 'linear' or 'two-dimensional ion trap'[6,7] is an exciting recent development where ions are stored in a cylindrical volume that is considerably larger than that of

 DOI 10.1002/pmic.200600528

Orbitrap Mass Analyzer – Overview and Applications in Proteomics

Michaela Scigelova[1] and Alexander Makarov[2]

The orbitrap mass analyzer is proving itself as a useful addition to a proteomics tool box. The key attributes of this analyzer are accurate mass and high resolution similar to those achievable with FT ICR instrumentation. The basic principles underlying these capabilities, and how they translate into benefits in real-life proteomics experiments are discussed. The focus is on reviewing examples of protein identification with bottom-up and top-down approaches, and detection of post-translational modifications.

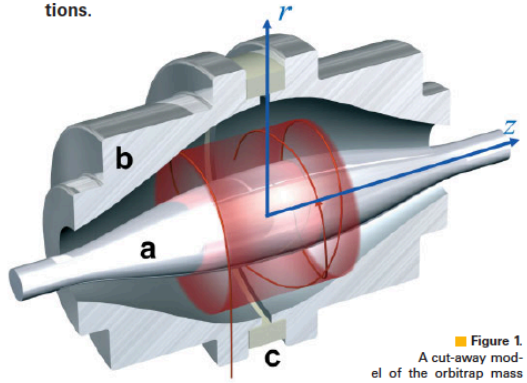


Figure 1. A cut-away model of the orbitrap mass analyzer. Ions are moving in spirals around a central electrode (a). An outer electrode (b) is split in half by an insulating ceramic ring (c). An image current induced by moving ions is detected via a differential amplifier between the two halves of the outer orbitrap electrode. The *m/z* if different ions in the orbitrap can be determined from respective frequencies of oscillation after a Fourier transform.

[1] Thermo Electron, Hemel Hempstead, UK

[2] Thermo Electron, Bremen, Germany

Michaela Scigelova
Thermo Electron Corp.
1 Boundary Park
Hemel Hempstead
HP2 7GE, UK
Tel.: +44-7977-984244
Fax: +44-1442-233667
michaela.scigelova@thermo.com

Introduction

Recent advances in many biological disciplines are closely related to application of proteomics techniques, in large part enabled by developments in mass spectrometry. An orbitrap mass analyzer is the most recent addition to the set of tools that can be applied to identification, characterisation and quantitation of components in biological systems. With its ability to deliver low-ppm mass accuracy and extremely high resolution, all within a time scale compatible with nano-LC separations, the orbitrap has become an instrument of choice for many proteomics applications since its commercial introduction in 2005. Let's revisit the basics of its operation, and review how this accurate mass detector has been used by some of the leading groups in the proteomics field.

As its name suggests, orbitrap is an ion trap. But it is not a conventional ion trap; there is neither RF nor a magnet to hold ions inside. Instead, moving ions are trapped in an electrostatic field [1, 2]. The electrostatic attraction towards the central electrode is compensated by a centrifugal force that arises from the initial tangential velocity of ions: very much like a satellite on orbit. The electrostatic field which ions experience inside the orbitrap forces them to move in complex spiral patterns (see supporting Microsoft PowerPoint presentation material for animation and more detail). The axial component of these oscillations is independent of initial energy, angles and positions, and can be detected as an image current on the two halves of an electrode encapsulating the orbitrap (Figure 1). A Fourier transform is employed to obtain oscillation frequencies for ions with different masses, resulting in an accurate reading of their *m/z*. Such measurements achieve very high resolution rivaling that of FT ICR instruments, and surpassing, by an order of magnitude, the resolution presently obtainable with orthogonal time-of-flight analyzers.

LESSONS IN *DE NOVO* PEPTIDE SEQUENCING BY TANDEM MASS SPECTROMETRY

Katalin F. Medzihradszky[1,2]* and Robert J. Chalkley[1]

[1]*Mass Spectrometry Facility, Department of Pharmaceutical Chemistry, School of Pharmacy, University of California San Francisco, 600 16th Street, Genentech Hall N472A, San Francisco, CA, 94158-2517*

[2]*Laboratory of Proteomics Research, Institute of Biochemistry, HAS Biological Research Centre, Szeged, Hungary*

Received 11 April 2013; revised 10 July 2013; accepted 10 July 2013

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/mas.21406

Mass spectrometry has become the method of choice for the qualitative and quantitative characterization of protein mixtures isolated from all kinds of living organisms. The raw data in these studies are MS/MS spectra, usually of peptides produced by proteolytic digestion of a protein. These spectra are "translated" into peptide sequences, normally with the help of various search engines. Data acquisition and interpretation have both been automated, and most researchers look only at the summary of the identifications without ever viewing the underlying raw data used for assignments. Automated analysis of data is essential due to the volume produced. However, being familiar with the finer intricacies of peptide fragmentation processes, and experiencing the difficulties of manual data interpretation allow a researcher to be able to more critically evaluate key results, particularly because there are many known rules of peptide fragmentation that are not incorporated into search engine scoring. Since the most commonly used MS/MS activation method is collision-induced dissociation (CID), in this article we present a brief review of the history of peptide CID analysis. Next, we provide a detailed tutorial on how to determine peptide sequences from CID data. Although the focus of the tutorial is de novo sequencing, the lessons learned and resources supplied are useful for data interpretation in general.

Keywords: *peptides; fragmentation; de novo sequencing; CID; HCD*

I. INTRODUCTION

A. Present/Future Role of *De Novo* Sequencing

With the ever-increasing number of complete genomes published, one might think that there is now less need for *de novo* protein sequence determination from mass spectrometry fragmentation data. However, each species features slightly different sequences due to single nucleotide/residue variants and splice-variants. The increased sensitivity of instrumentation is also revealing a multitude of unpredicted post-translational and sample-handling modifications, which if not specified as possibilities during database searching, will not be identified. In addition, protein prediction from genomes is partly based on homology. Thus, really unique, species-specific sequences might stay undiscovered. For example, a BLAST search performed with the first 29 amino acids of a snake venom toxin could not find any similar sequence in the NCBI database. Even the full-length sequence (61 residues) produced only one remotely similar structure: a venom peptide from another snake (Bohlen et al., 2011). Similarly, other relatively small, biologically active polypeptides, such as toxins and antibacterial agents, although coded in the genome, cannot be readily predicted. Thus, at minimum, determining sequence tags might be necessary.

Last, but not least, a generation of proteomics researchers has grown up relying heavily on automated data interpretation, and does not know enough about the fragmentation processes that underlie the results. This lack of hands-on experience prevents the critical evaluation of automated search results, and still frequently manifests itself in the acceptance and publishing of dubious or obviously incorrect assignments. The situation is more problematic for post-translational modification (PTM) analysis, especially when multiple different modifications are considered during a search, and permitted on a single peptide. Some common-sense rules have been suggested when someone should be suspicious about the automated sequence assignments and look for an alternative interpretation (Stevens, Prokai-Tatrai, & Prokai, 2008; Chalkley, 2013). The varied experience of researchers is one of the reasons why proteomics journals request assigned spectra and raw data to be deposited for single-peptide-based protein identifications and especially when reporting PTMs.

B. Historical Overview of *De Novo* Sequencing

Enkephalins are frequently used as mass spectrometry standards, or convenient small peptides to study (Sztáray et al., 2011). Most researchers are not aware that these structures were deciphered using mass spectrometry (Hughes et al., 1975). At that time, peptide structural elucidation using mass spectrometry was no small feat; extensive derivatization was required to make even small peptides volatile enough to be detected in a mass spectrometer. Mass spectrometry was an "exotic" analytical technique for protein chemists, because peptides, just like most other biologically interesting compounds, decomposed rather than ionized when the available ionization techniques (electron

Contract grant sponsor: NIH; Contract grant number: NIGMS 8P41GM103481.
**Correspondence to*: Katalin F. Medzihradszky, Mass Spectrometry Facility, UCSF, 600 16th Street, Genentech Hall, suite N472A, Box 2240, San Francisco, CA 94158-2517.
E-mail: folkl@cgl.ucsf.edu