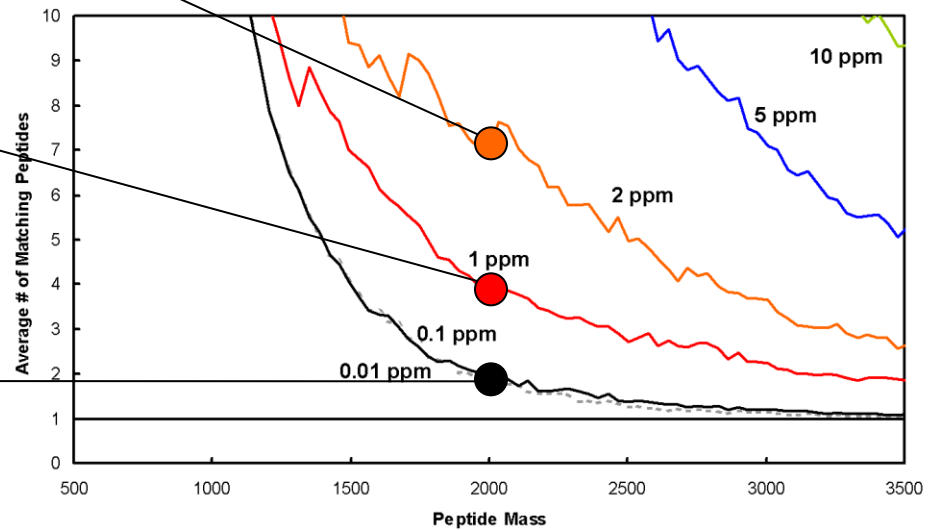
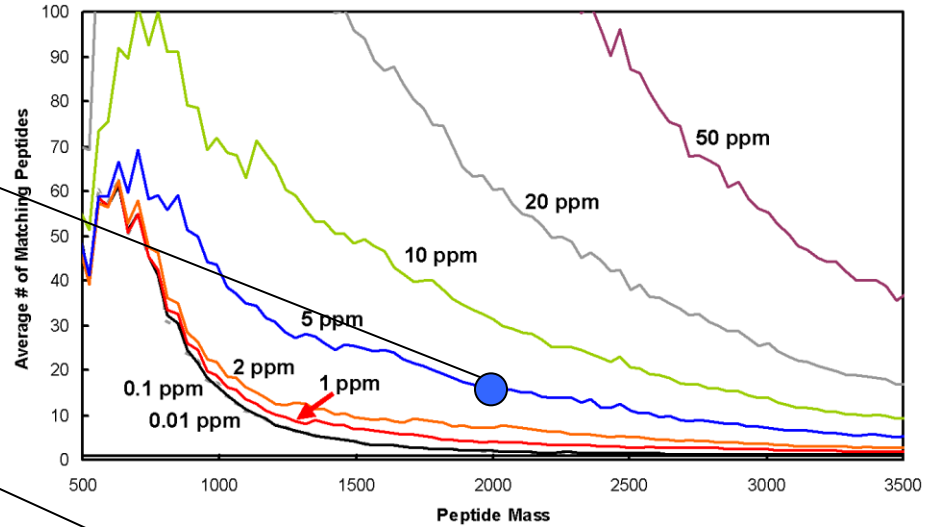
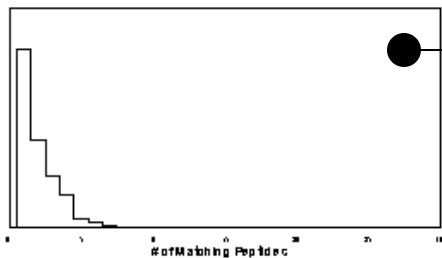
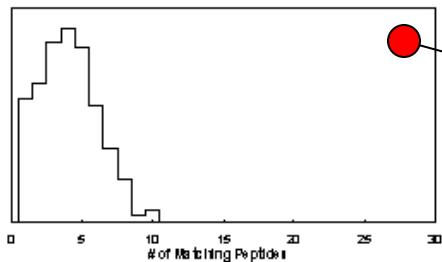
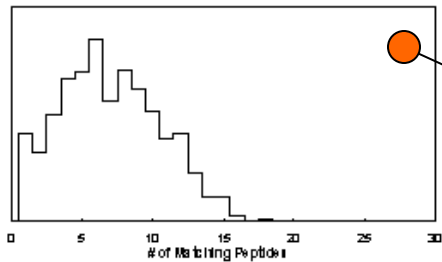
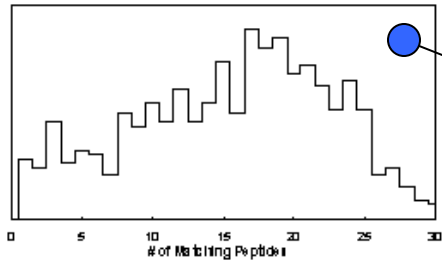


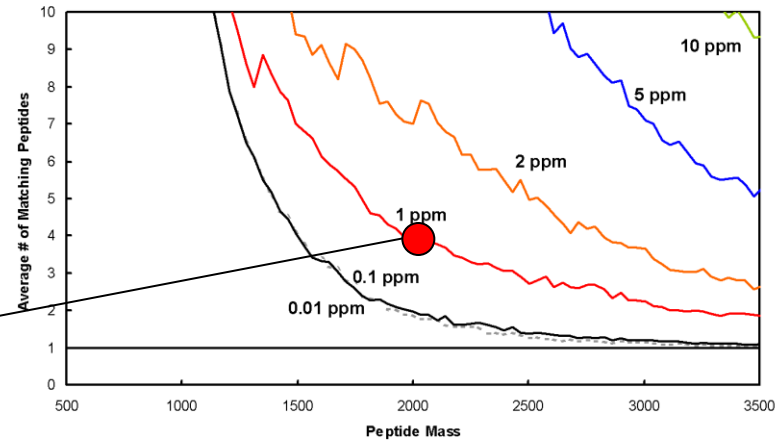
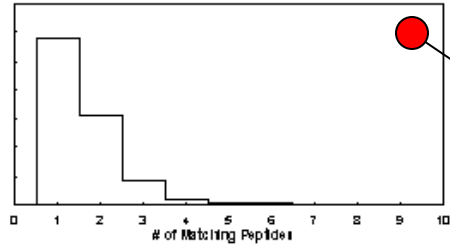
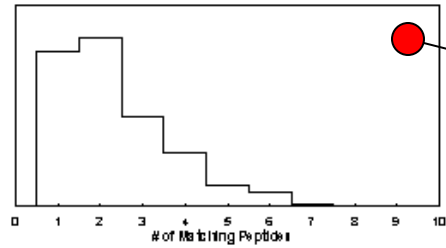
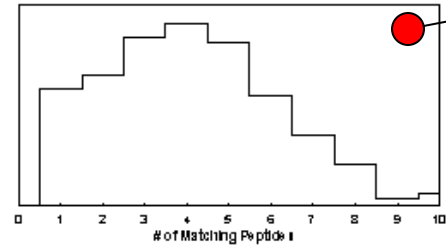
Proteomics Informatics -

**Protein identification I: searching protein sequence
collections and significance testing (Week 4)**

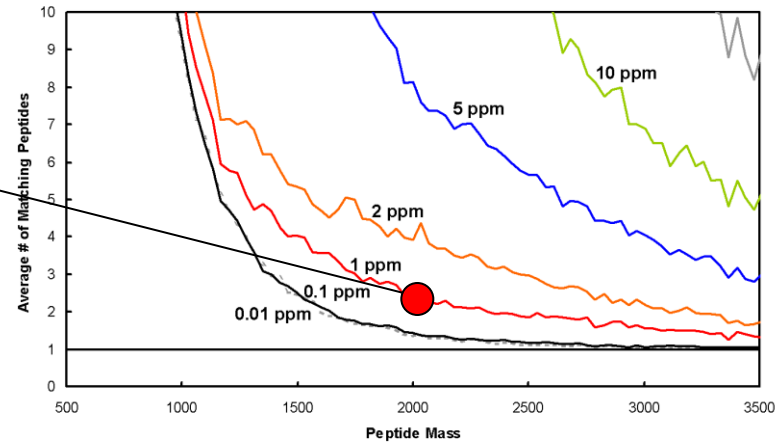
Peptide Mapping - Mass Accuracy



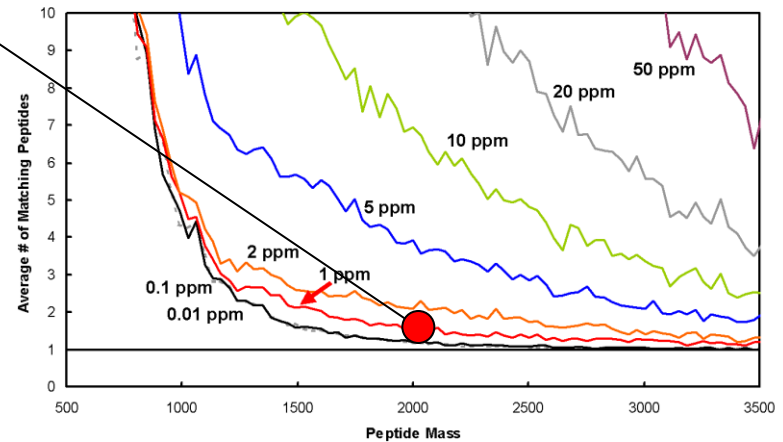
Peptide Mapping Database Size



Human

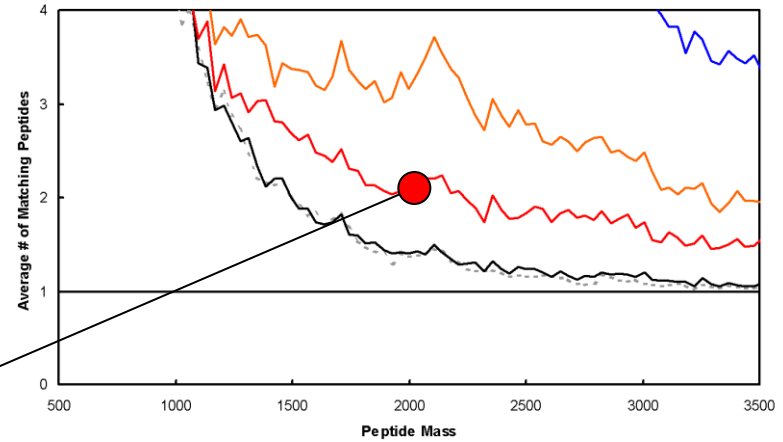
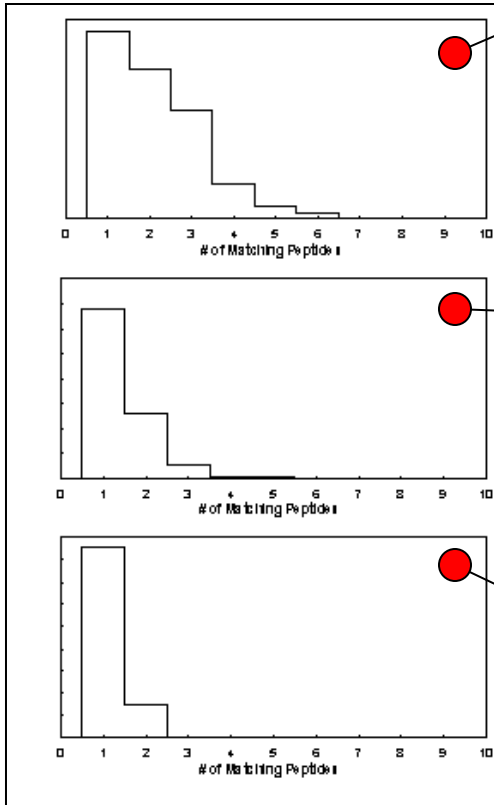


C. elegans

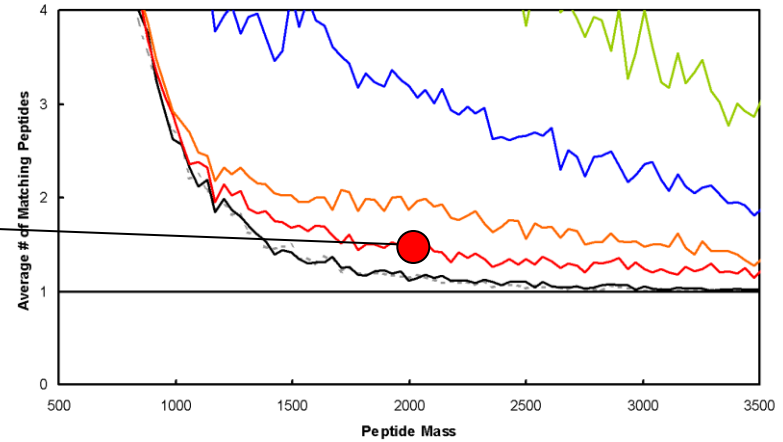


S. cerevisiae

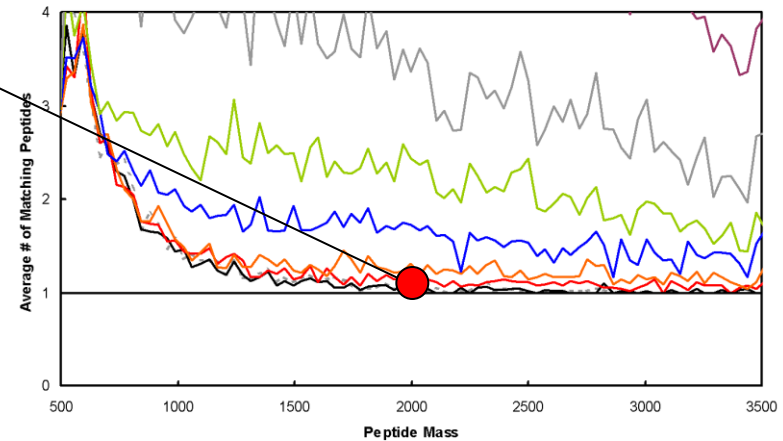
Peptide Mapping Cys-Containing Peptides



Human

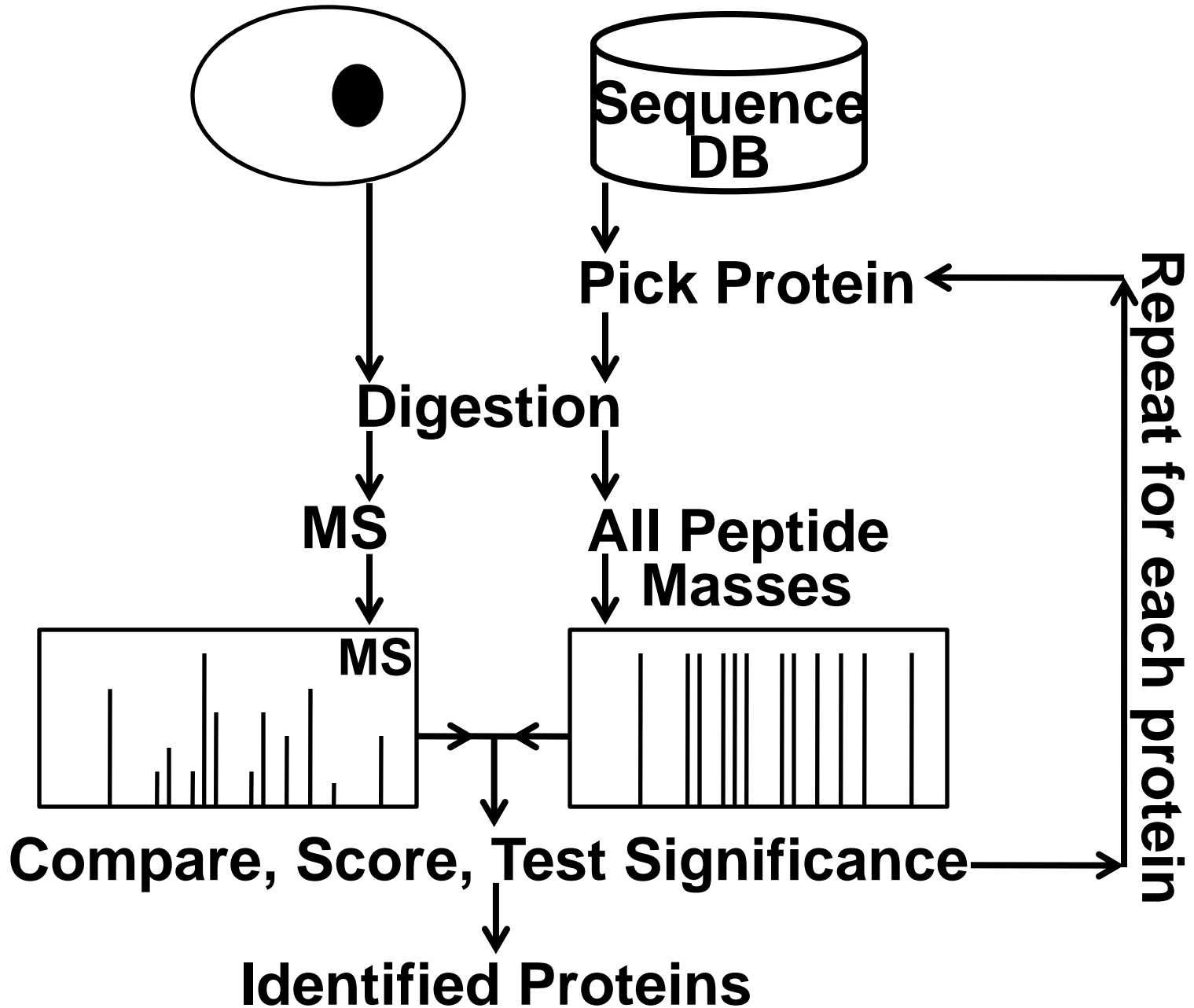


C. elegans



S. cerevisiae

Identification - Peptide Mass Fingerprinting

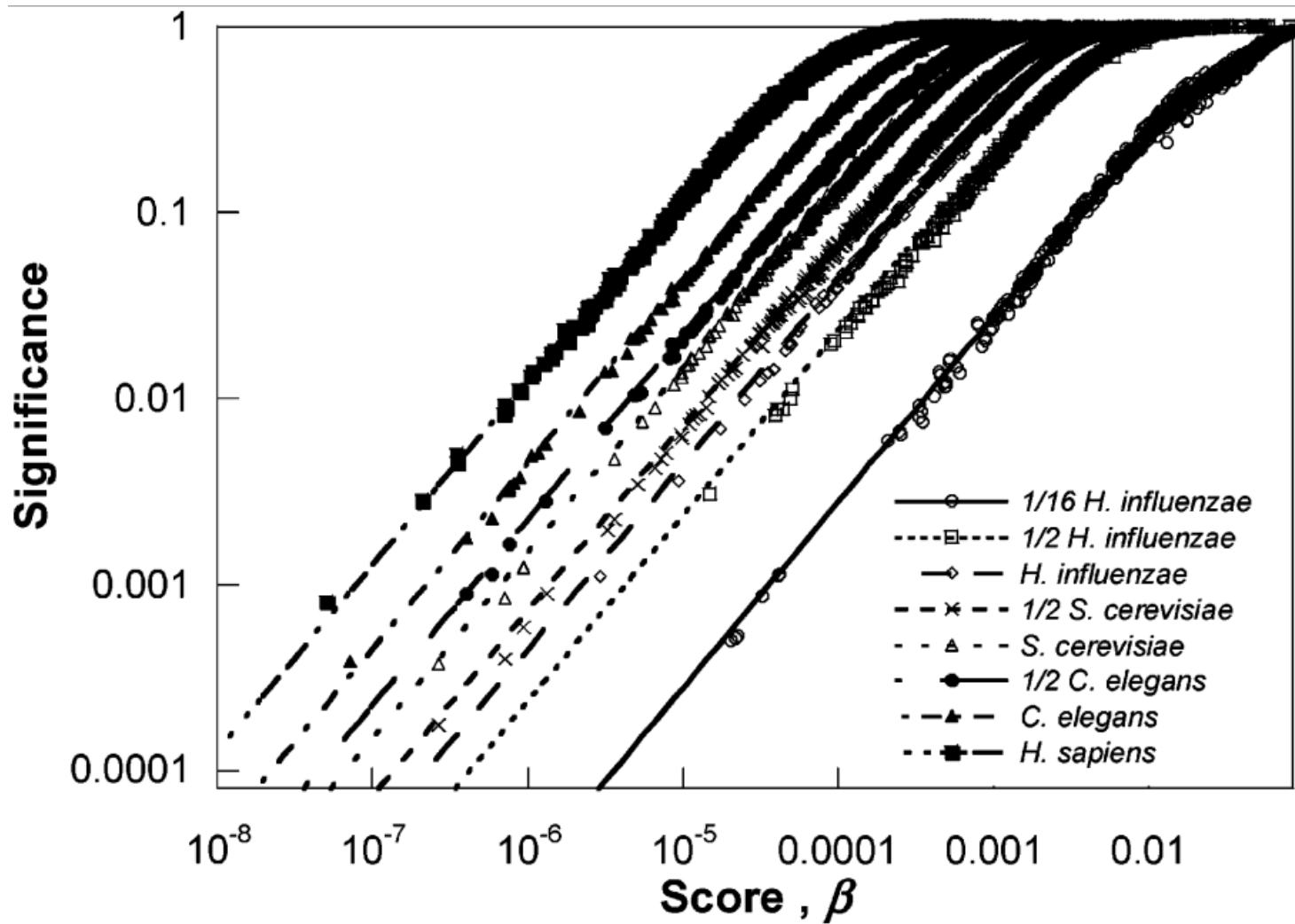


ProFound Results

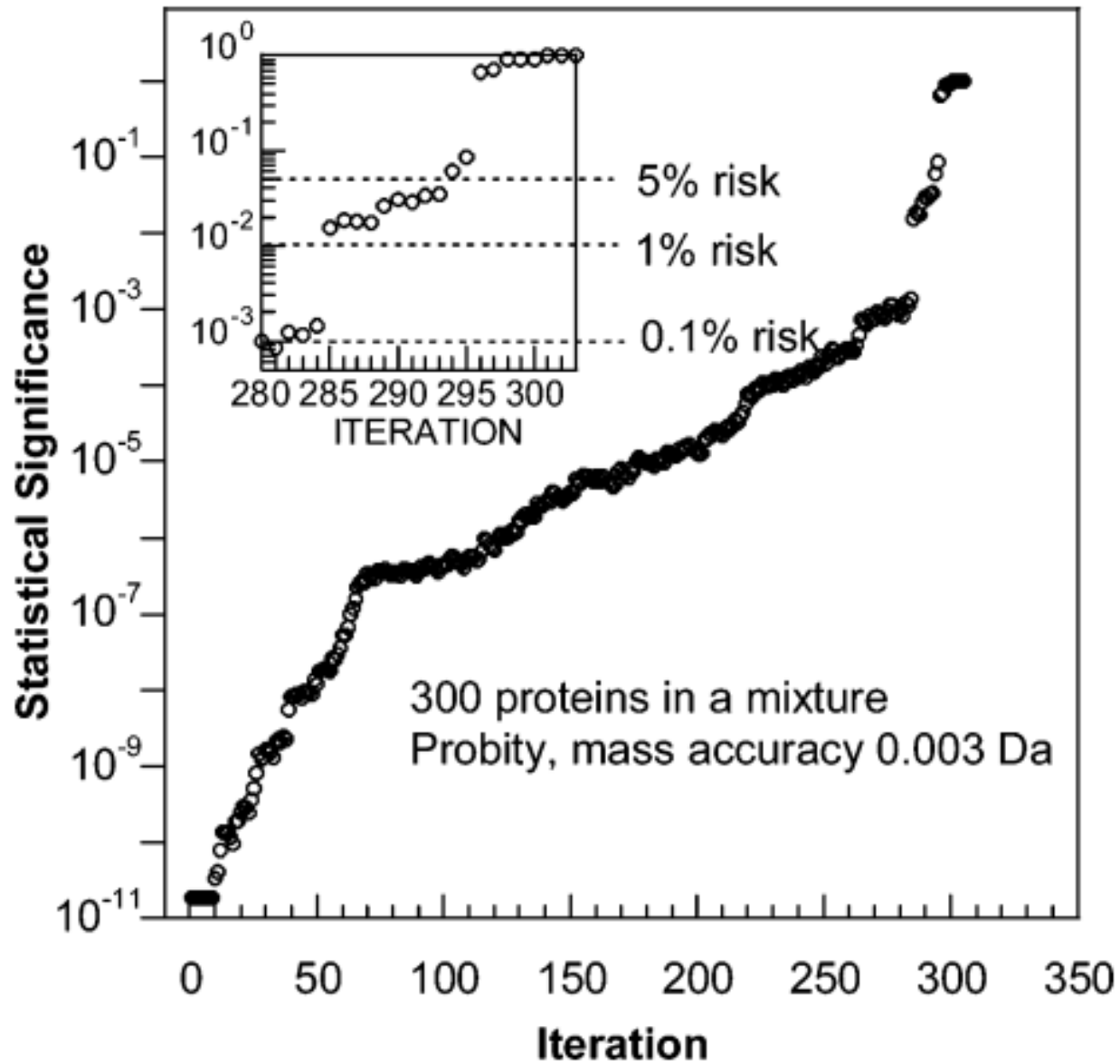
Protein Candidates

Rank	Expectation	Protein Information and Sequence Analyse Tools (T)	%	pI	kDa
+1	5.110 ⁻⁷	gi 148236543 ref NP_001081565.1 serine/threonine-protein kinase 6-A [Xenopus laevis]	36	9.6	46.35
+2	0.057	gi 213626249 gb AAI70128.1 Unknown (protein for MGC:196855) [Xenopus laevis]	8	5.3	147.73
3	0.094	gi 147905824 ref NP_001086865.1 WD repeat-containing protein 67 [Xenopus laevis]	9	7.5	126.81

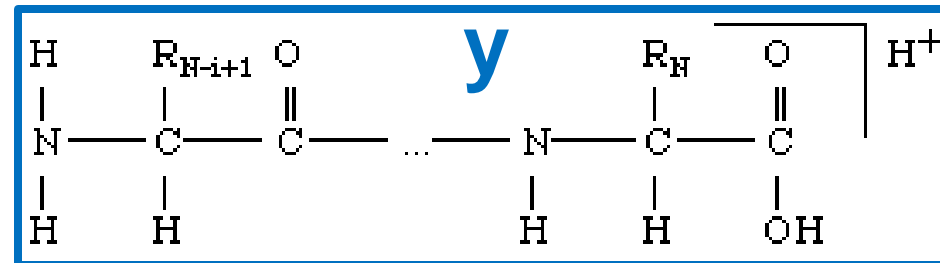
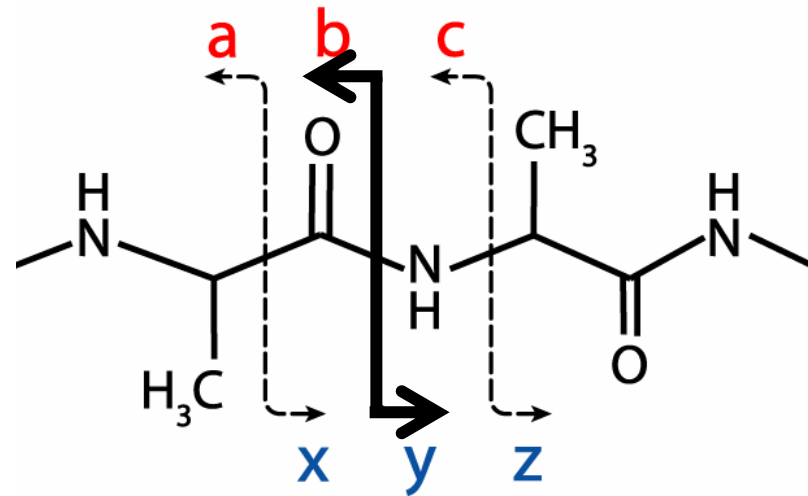
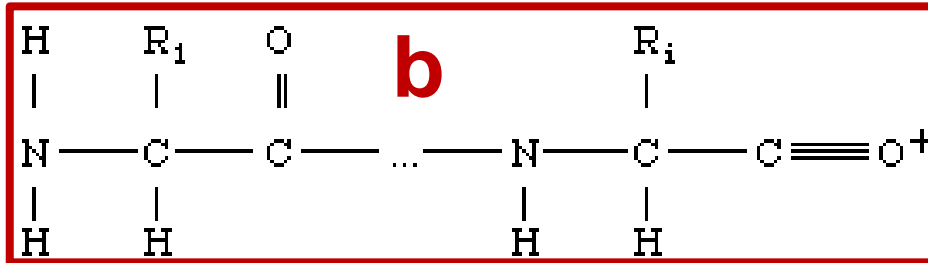
Database size



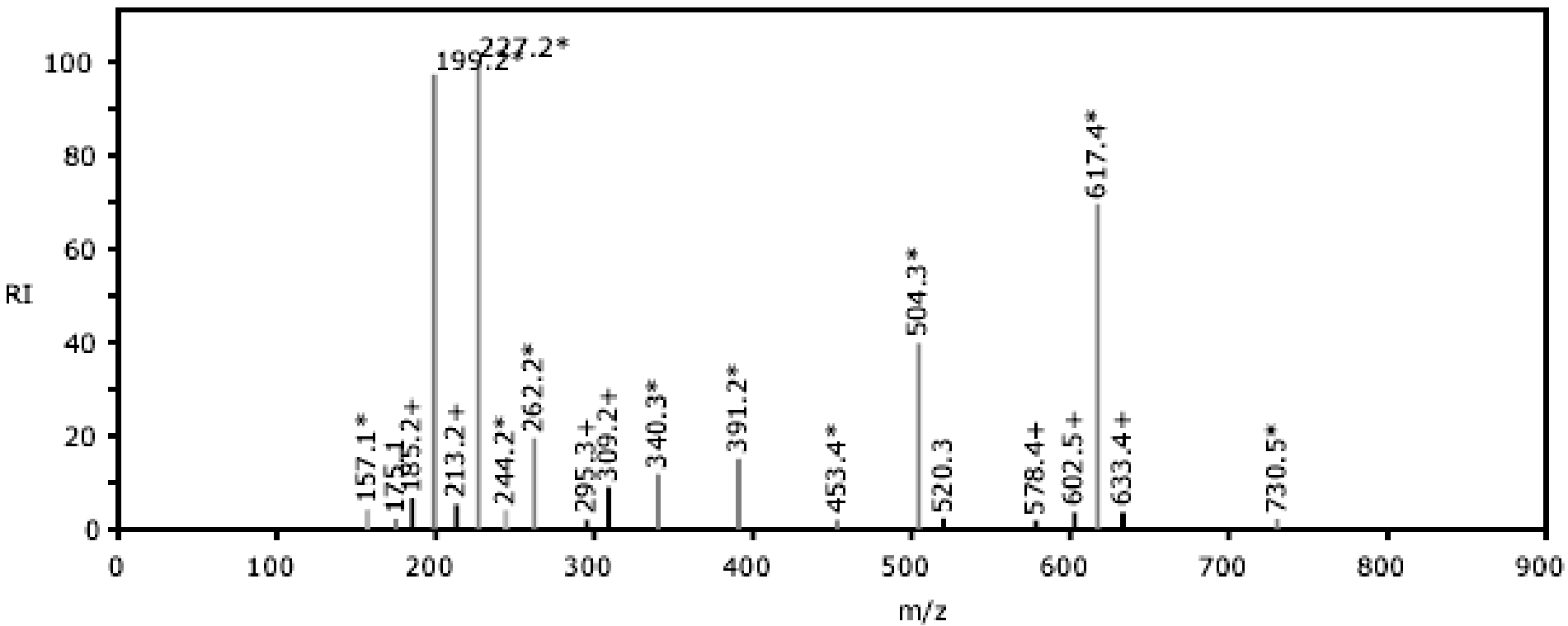
Mixtures



Peptide Fragmentation

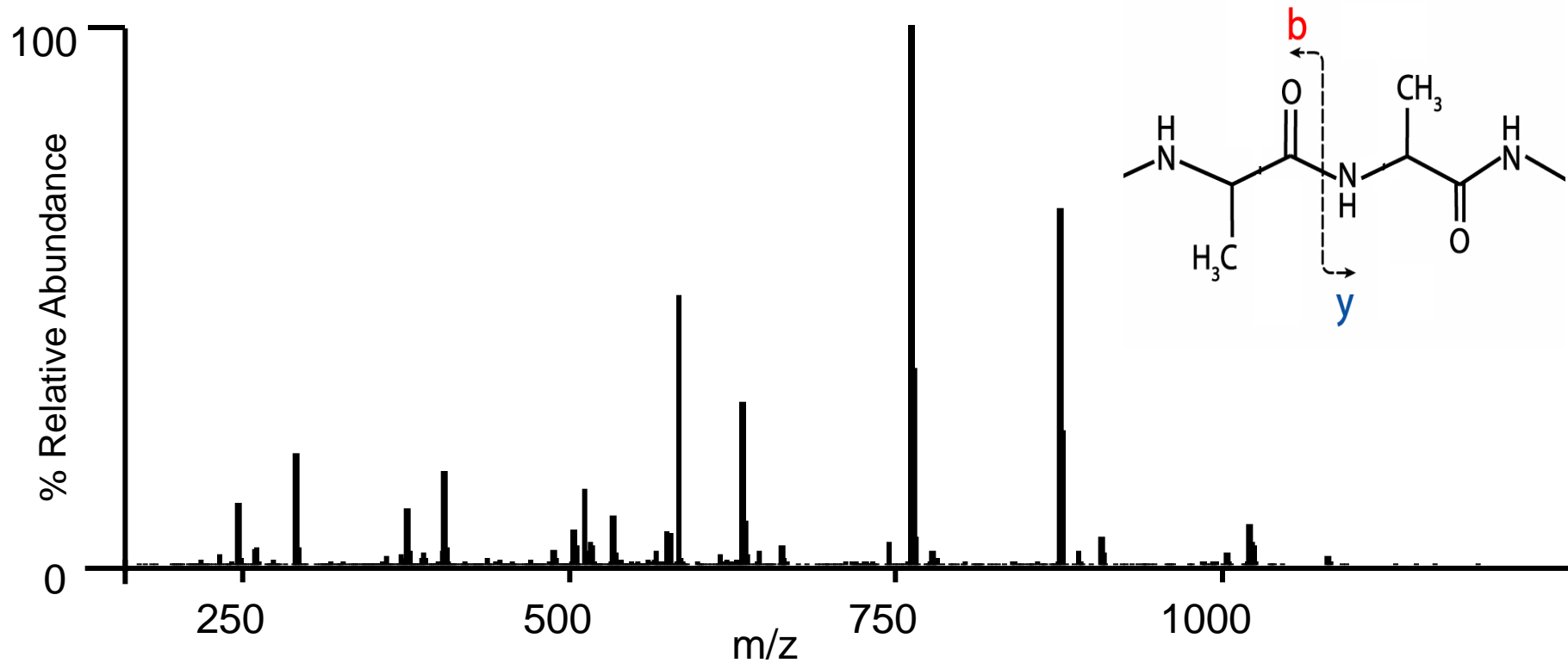


Identification - Tandem MS



Tandem MS - Sequence Confirmation

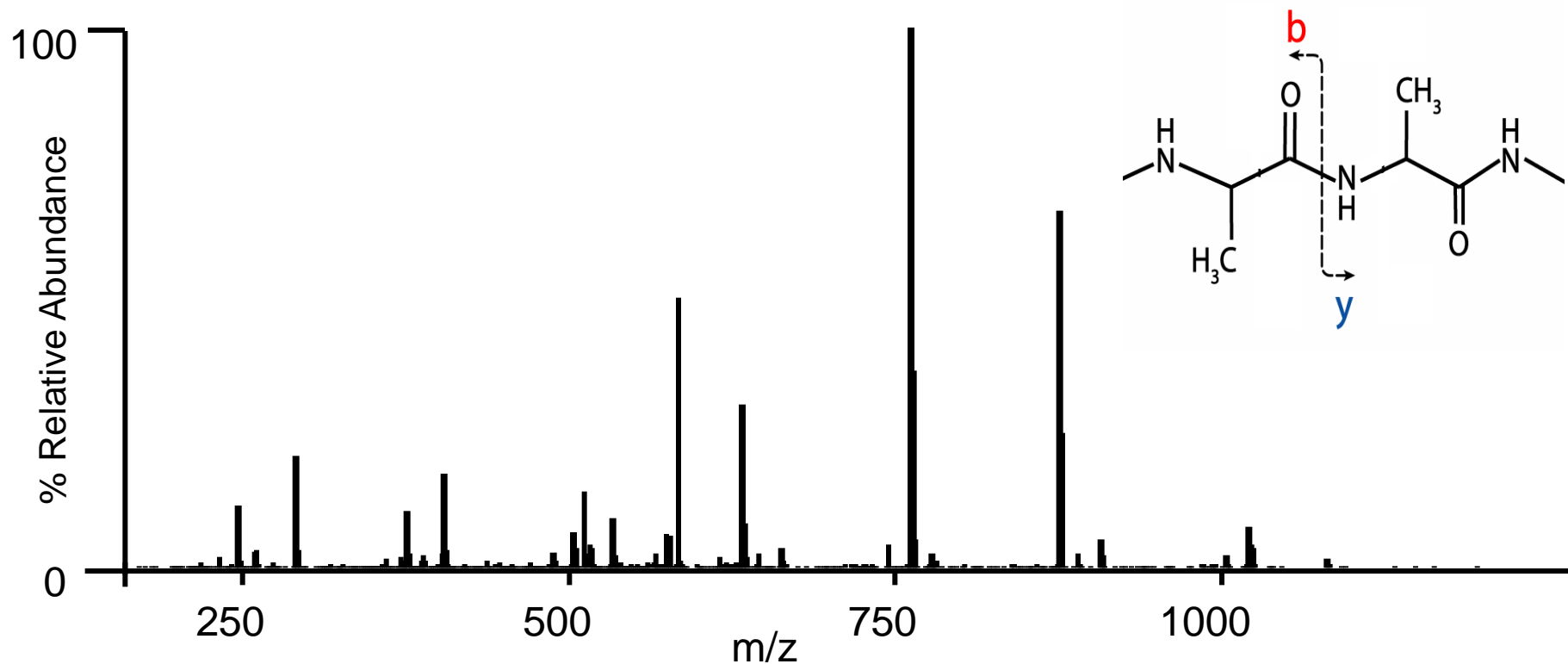
S G F L E E D E L K



Tandem MS - Sequence Confirmation

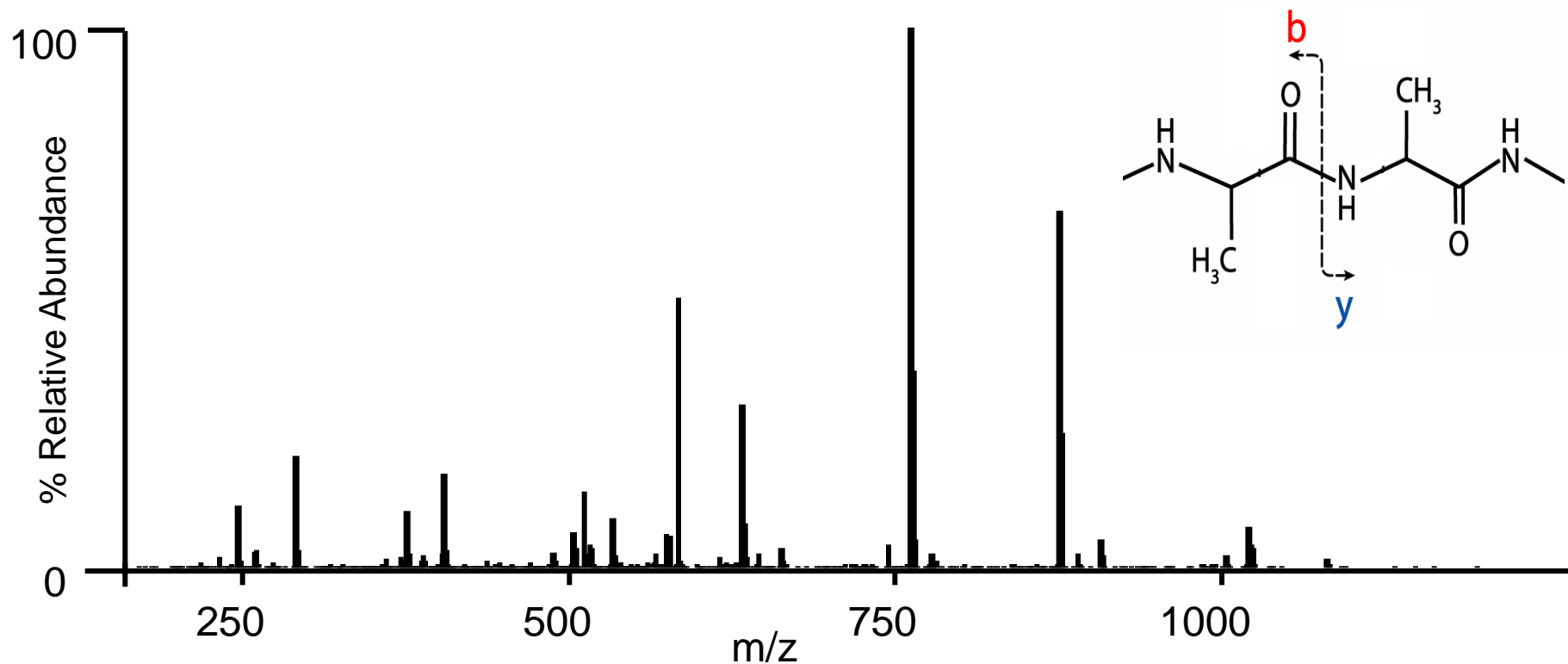
S G F L E E D E L K

88 145 292 405 534 663 778 907 1020 1166 b ions



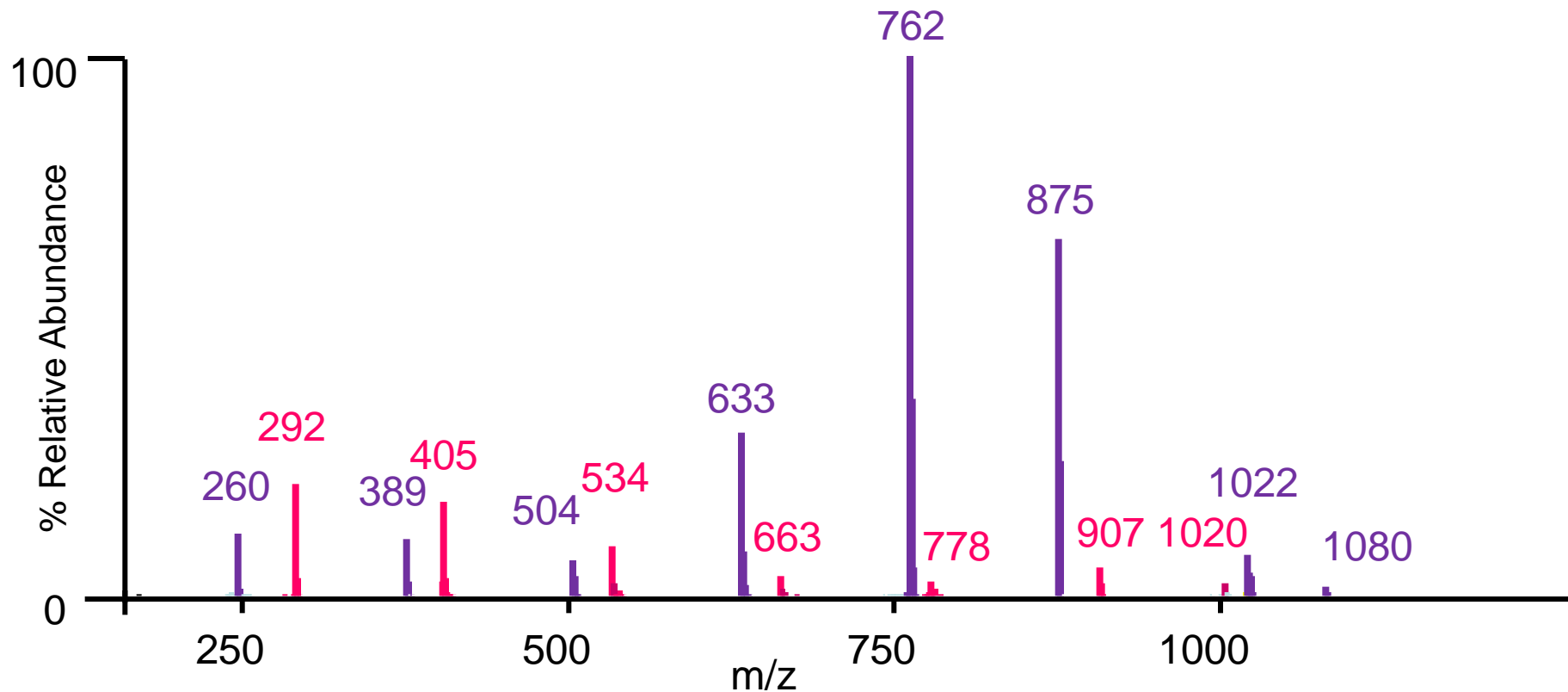
Tandem MS - Sequence Confirmation

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



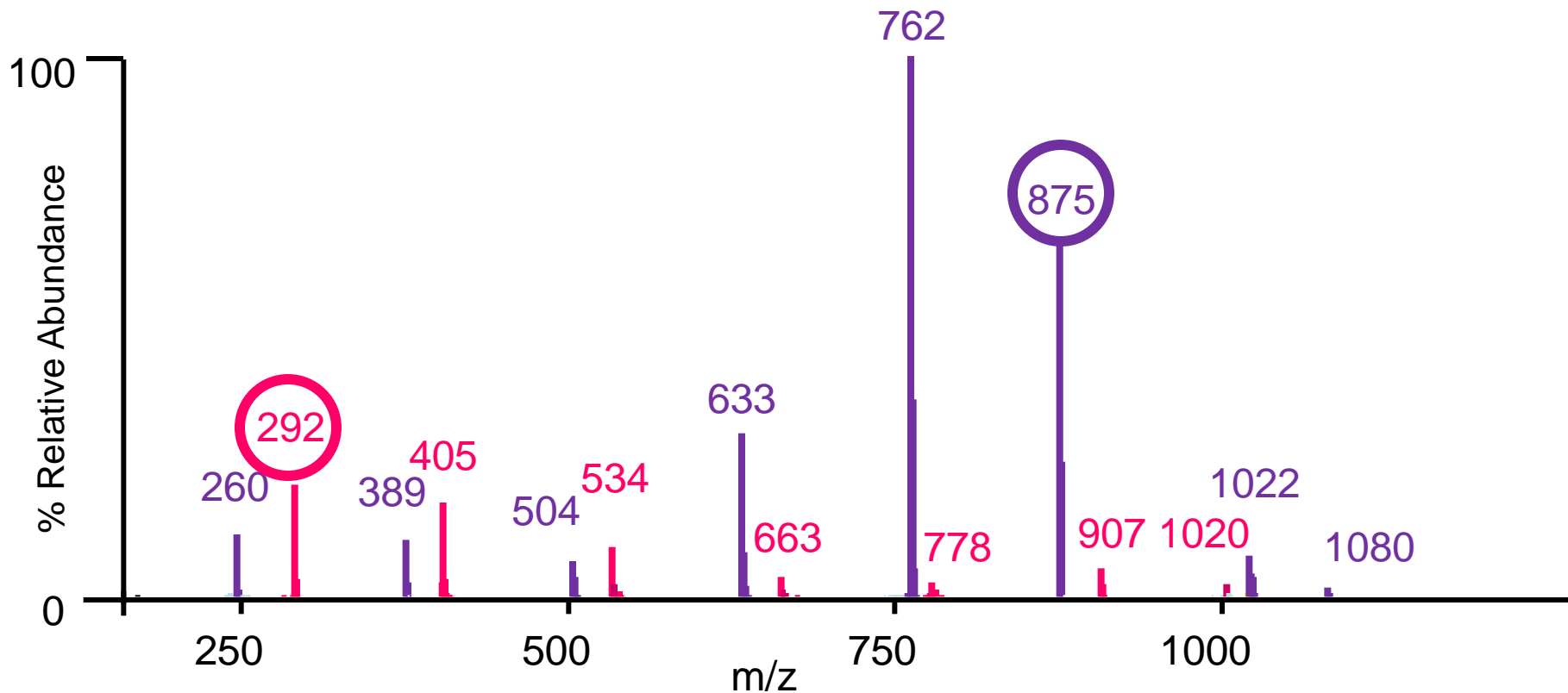
Tandem MS - Sequence Confirmation

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

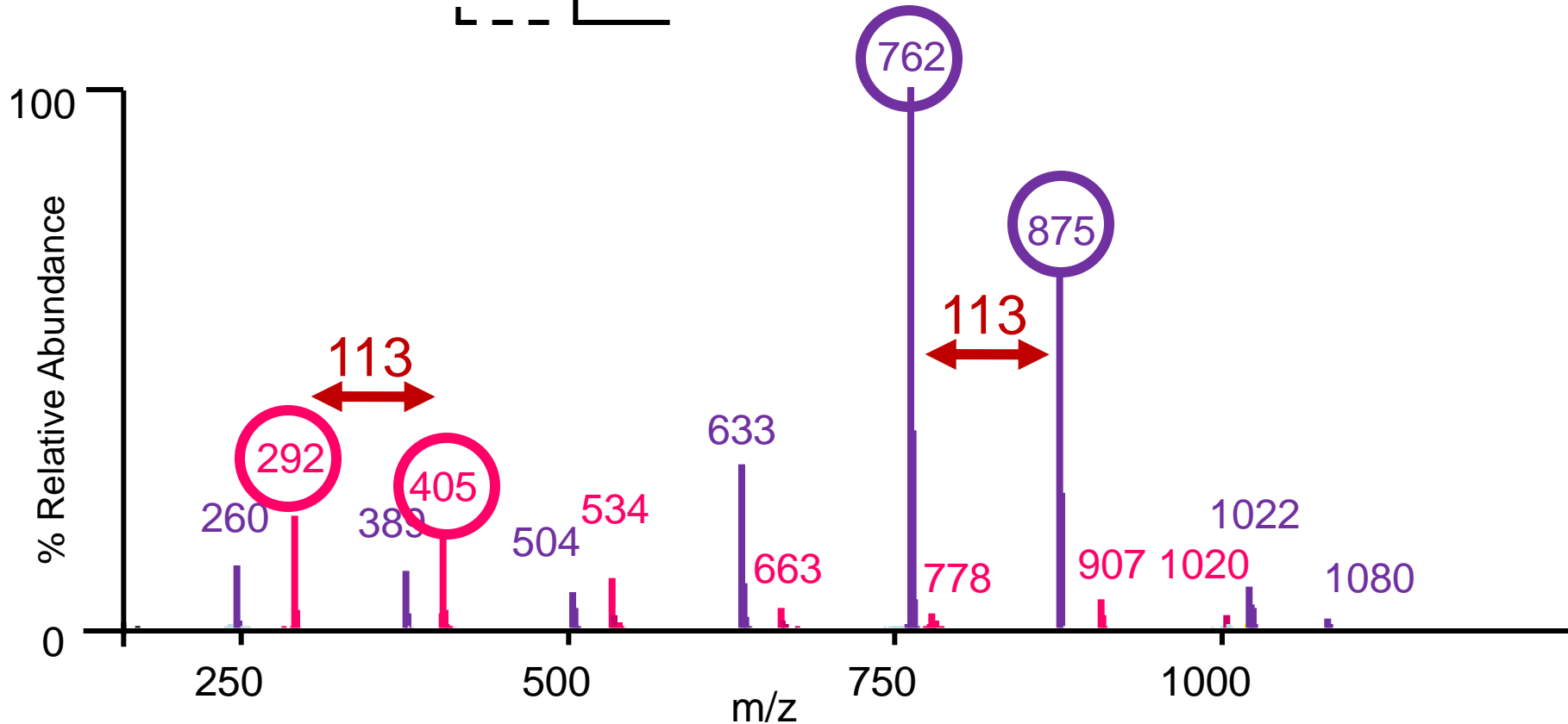
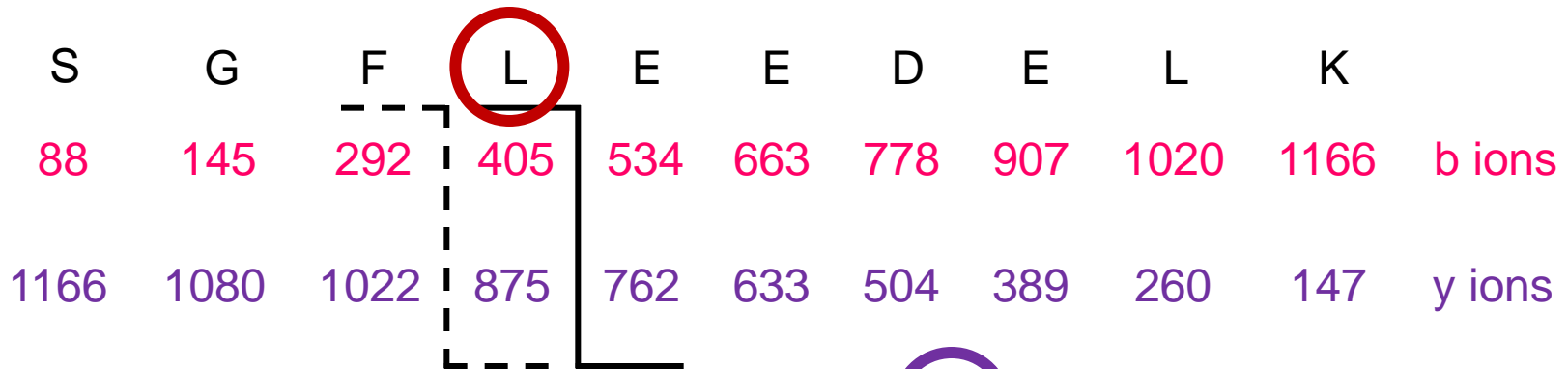


Tandem MS - Sequence Confirmation

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

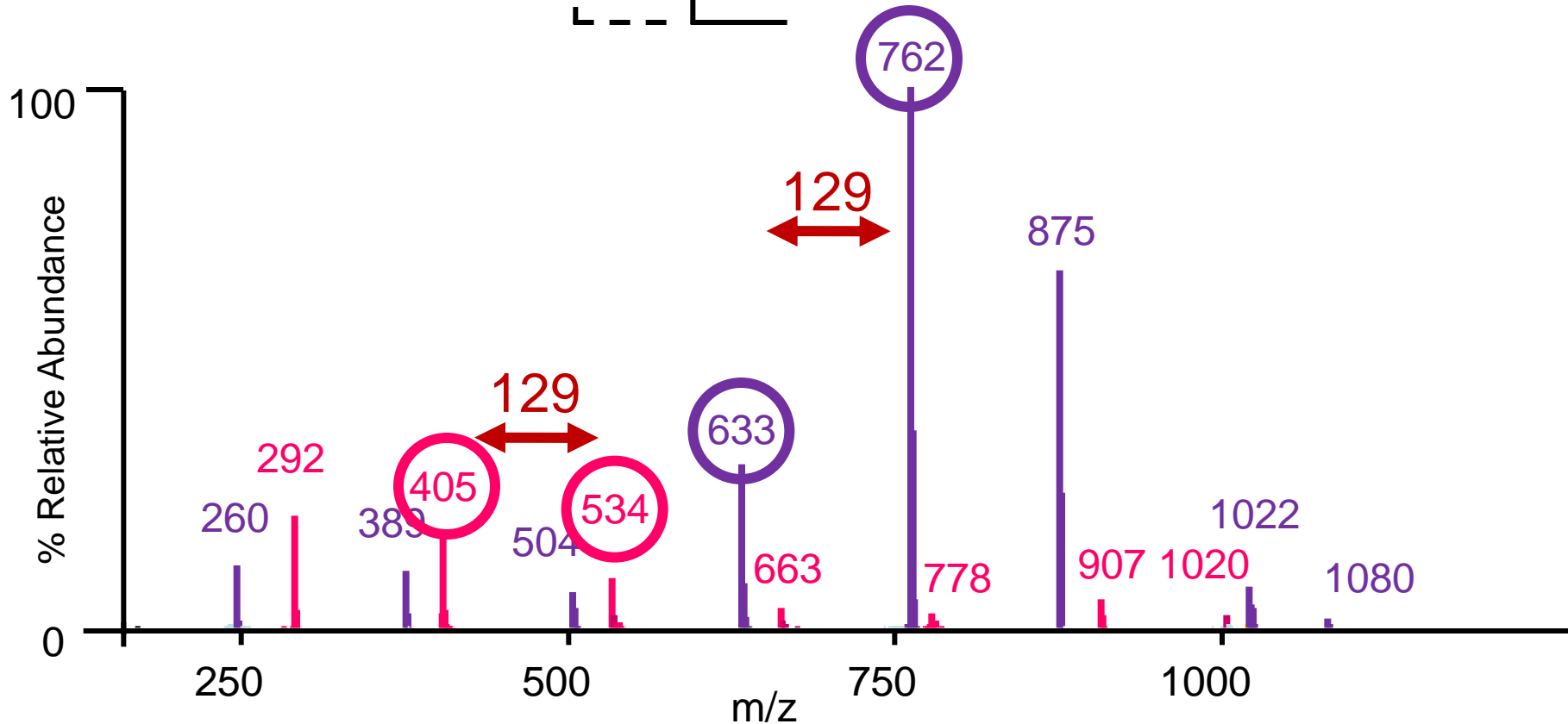


Tandem MS - Sequence Confirmation

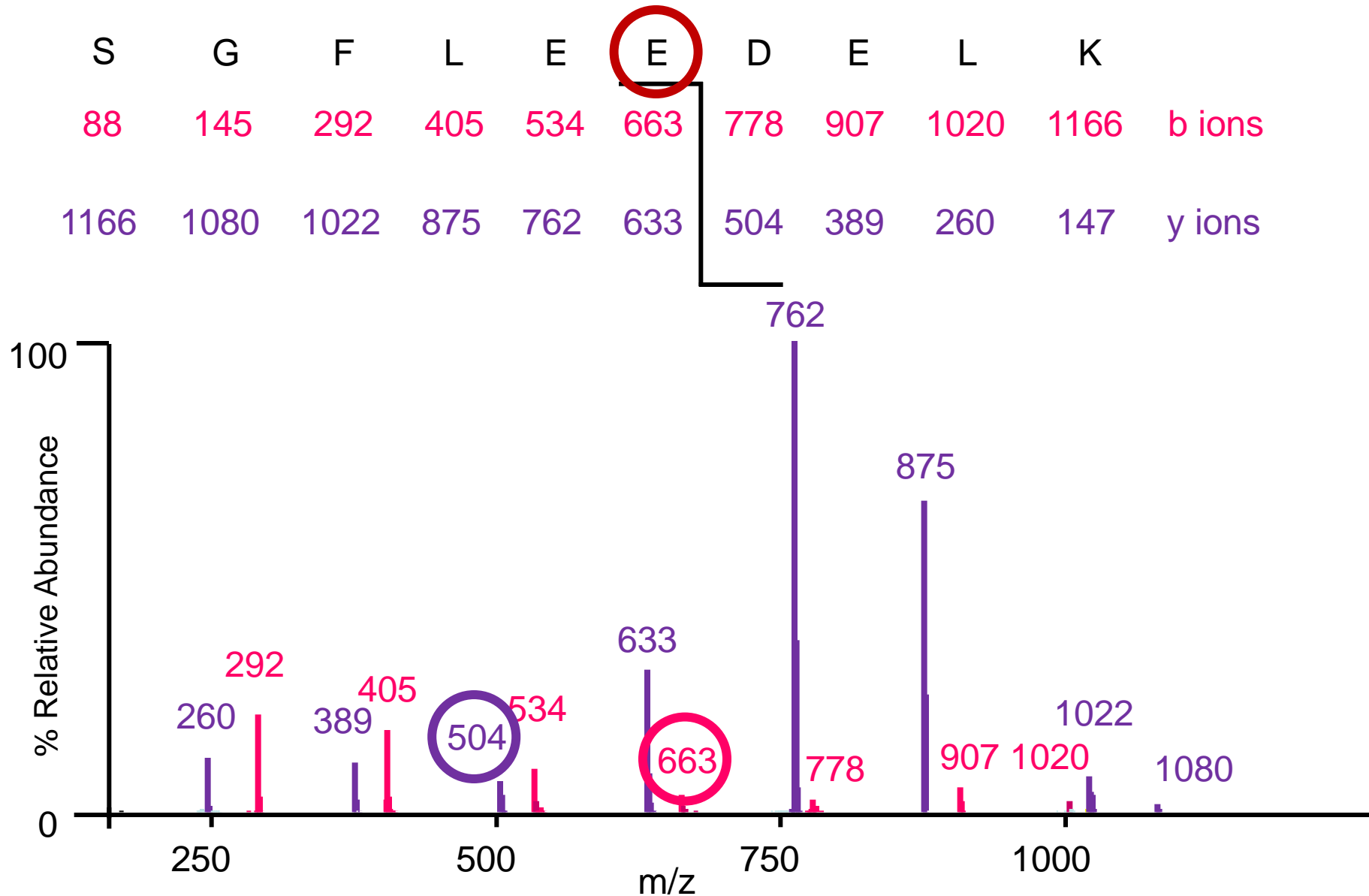


Tandem MS - Sequence Confirmation

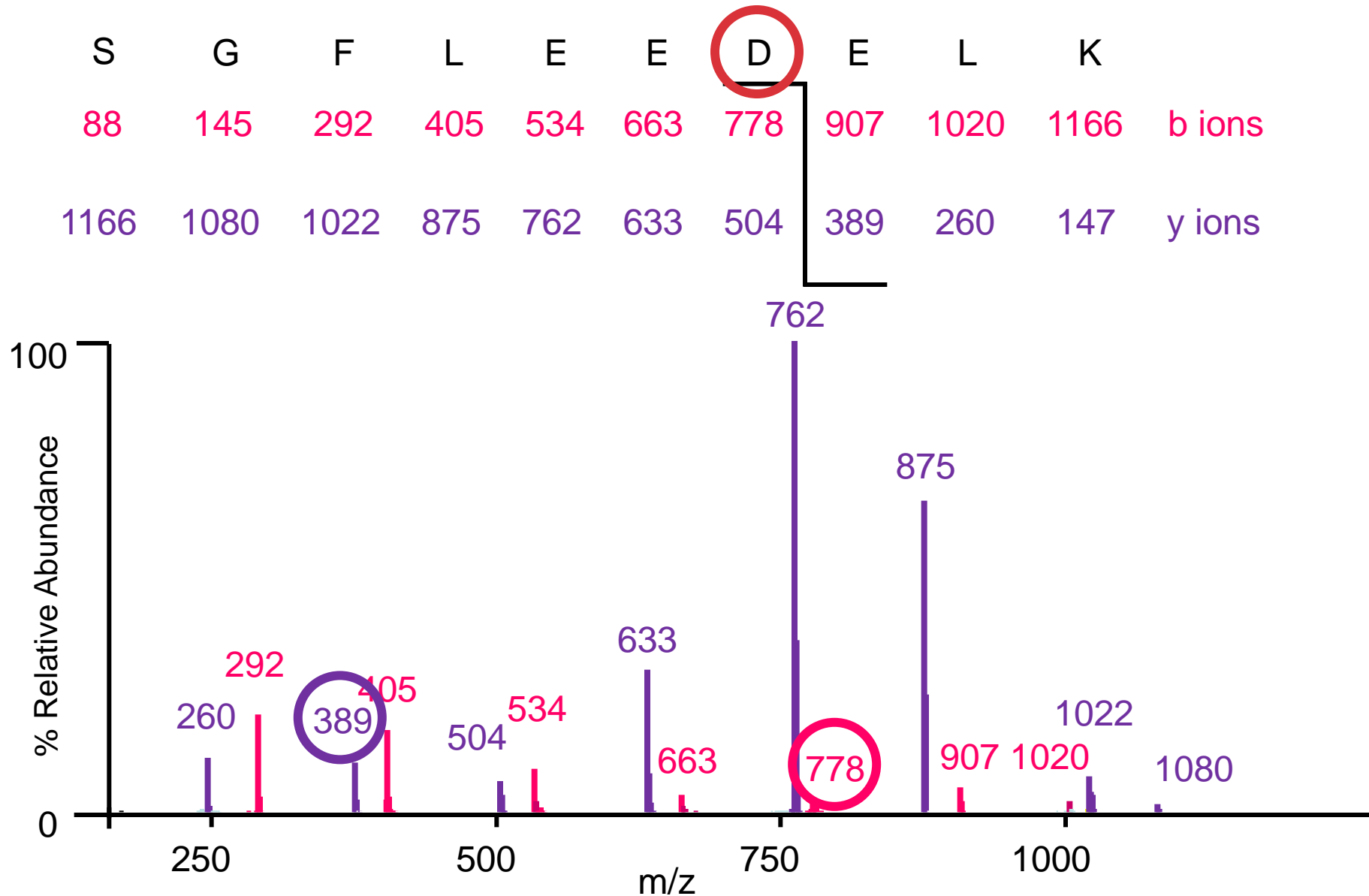
S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions
1166	1080	1022	875	762	633	504	389	260	147	y ions



Tandem MS - Sequence Confirmation

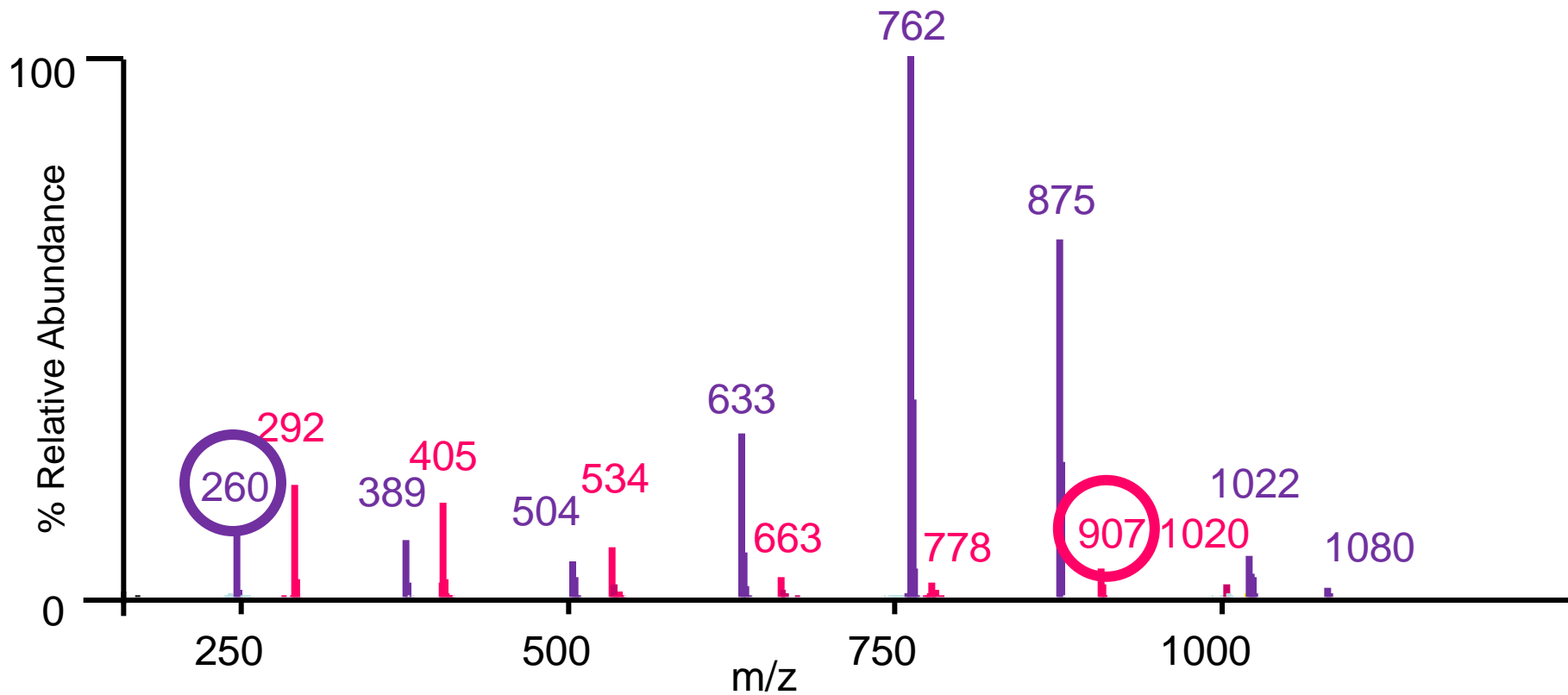


Tandem MS - Sequence Confirmation



Tandem MS - Sequence Confirmation

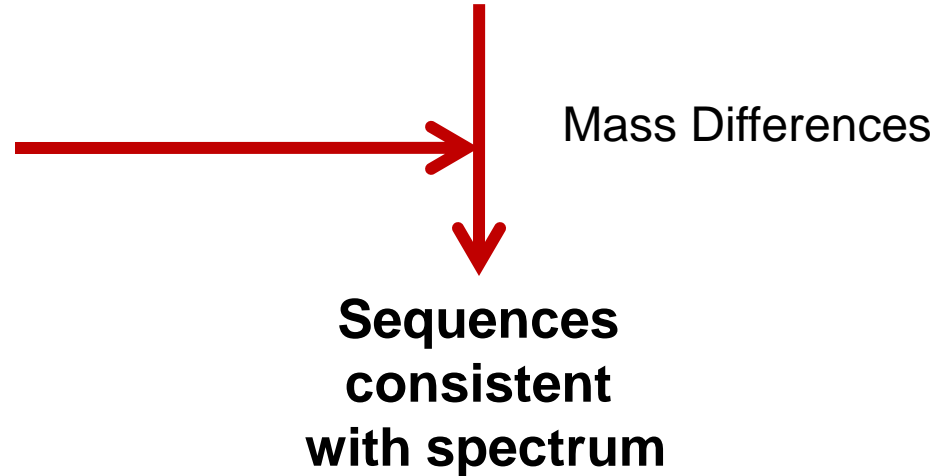
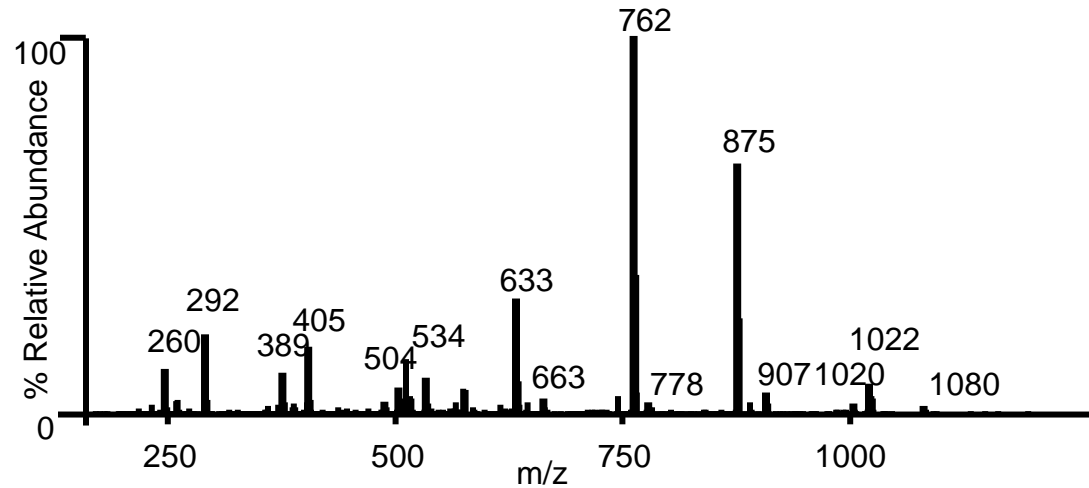
S	G	F	L	E	E	D	E	L	K	
88	145	292	405	534	663	778	907	1020	1166	b ions
1166	1080	1022	875	762	633	504	389	260	147	y ions



Tandem MS - de novo Sequencing

Amino acid masses

1-letter code	3-letter code	Chemical formula	Monoisotopic	Average
A	Ala	C ₃ H ₅ ON	71.0371	71.0788
R	Arg	C ₆ H ₁₂ ON ₄	156.101	156.188
N	Asn	C ₄ H ₆ O ₂ N ₂	114.043	114.104
D	Asp	C ₄ H ₅ O ₃ N	115.027	115.089
C	Cys	C ₃ H ₅ ONS	103.009	103.139
E	Glu	C ₅ H ₇ O ₃ N	129.043	129.116
Q	Gln	C ₅ H ₈ O ₂ N ₂	128.059	128.131
G	Gly	C ₂ H ₃ ON	57.0215	57.0519
H	His	C ₆ H ₇ ON ₃	137.059	137.141
I	Ile	C ₆ H ₁₁ ON	113.084	113.159
L	Leu	C ₆ H ₁₁ ON	113.084	113.159
K	Lys	C ₆ H ₁₂ ON ₂	128.095	128.174
M	Met	C ₅ H ₉ ONS	131.04	131.193
F	Phe	C ₉ H ₉ ON	147.068	147.177
P	Pro	C ₅ H ₇ ON	97.0528	97.1167
S	Ser	C ₃ H ₅ O ₂ N	87.032	87.0782
T	Thr	C ₄ H ₇ O ₂ N	101.048	101.105
W	Trp	C ₁₁ H ₁₀ ON ₂	186.079	186.213
Y	Tyr	C ₉ H ₉ O ₂ N	163.063	163.176
V	Val	C ₅ H ₉ ON	99.0684	99.1326



Tandem MS - de novo Sequencing

	260	292	389	405	504	534	633	663	762	778	875	907	1020	1022	1079
→	260	32	E	145	244	274	373	403	502	518	615	647	760	762	819
→	292		X	/L	212	242	341	371	470	486	583	615	728	730	787
→	389			16	D	145	244	274	373	389	486	518	631	633	690
→	405				X	E	228	258	357	373	470	502	615	617	674
→	504					30	E	159	258	274	371	403	516	518	575
→	534						X	E	228	244	341	373	486	488	545
→	633						30	E	145	242	274	387	389	446	
→	663							X	D	212	244	357	359	416	
→	762								16	/L	145	258	260	317	
→	778									X	E	242	244	301	
→	875										32	145	F	204	
→	907											/L	X	172	
→	1020													2	59
→	1022														G

SGF(I/L)EEDE(I/L)...

$$1166 - 1020 - 18 = 128$$

⇒ K or Q

SGF(I/L)EEDE(I/L)(**K/Q**)

Tandem MS - de novo Sequencing

Challenges in de novo sequencing

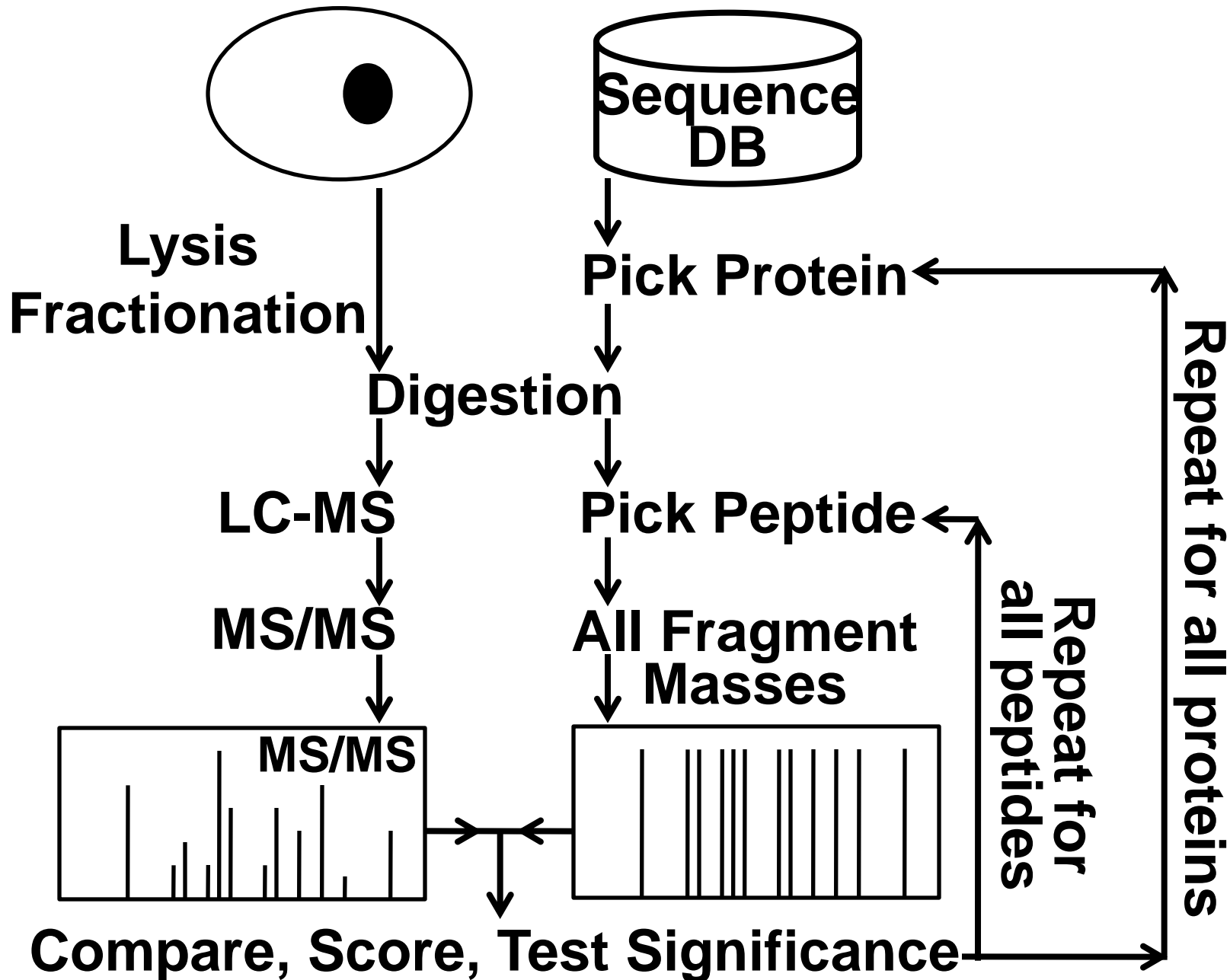
Neutral loss (-H₂O, -NH₃)

Modifications

Background peaks

Incomplete information

Tandem MS - Database Search



Search Results

1 match for *GPM33080001549*,

Display: [model](#) | [metadata](#) | [group](#) | [peptide](#) | [aaa](#) | [gel](#) | [GO](#) | [BTO](#) | [path](#) | [snaps](#) | [mh](#) | [ζ](#) | [wiki](#)

BRENDA cell culture: none

BRENDA tissue: none

CELL cell type: none

GO subcellular: none











institution: University of Toronto

name: Kislinger Lab

project: In-depth Proteomic Analyses of Direct Expressed Prostatic Secretions

project comment: Prostatic secretion 4, [Tranche](#) Fluids that are proximal to organs contain a repertoire of secreted proteins and shed cells reflective of the physiological state of that tissue, and thus represent potential sources for biomarker discovery and investigation of tissue-specific biology. Proximal fluids of the prostate are seminal plasma and expressed prostatic secretions (EPS). MudPIT-based proteomics was applied to EPS obtained from men with prostate cancer and resulted in the identification of 916 proteins. *J. Prot. Res.* DOI [10.1021/pr1001498](#) (PubMed).

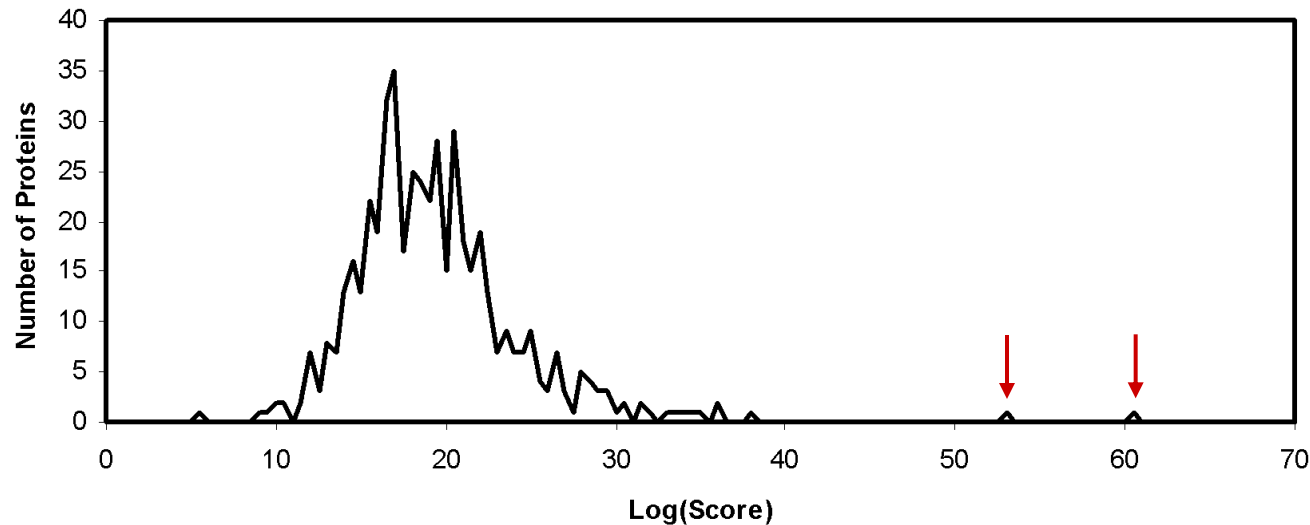
Best models for *GPM33080001549* [Show all](#), or display as

#	log(e)	accession	coverage	
1.	-2281.6	ALB		[31/13757]
2.	-2207.4	ALB		[12/10080]
3.	-1574	FCGBP		[1/1066]
4.	-1139.5	ACPP		[3/325]
5.	-1078.5	LTF		[5/2428]
6.	-1041.1	KLK3		[4/217]
7.	-760.5	TGM4		[0/68]
8.	-699.4	ANPEP		[9/958]
9.	-695.5	TF		[85/5619]
10.	-684.4	AZGP1		[3/2526]

Significance Testing

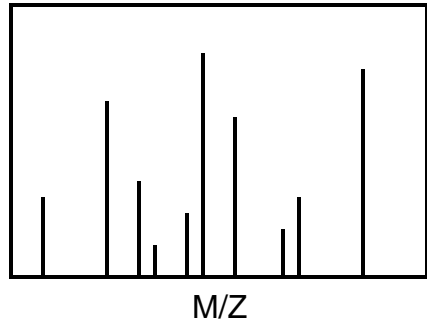
False protein identification is caused by random matching

Significance Testing - Expectation Values



The majority of sequences in a collection will give a score due to random matching.

Significance Testing - Expectation Values

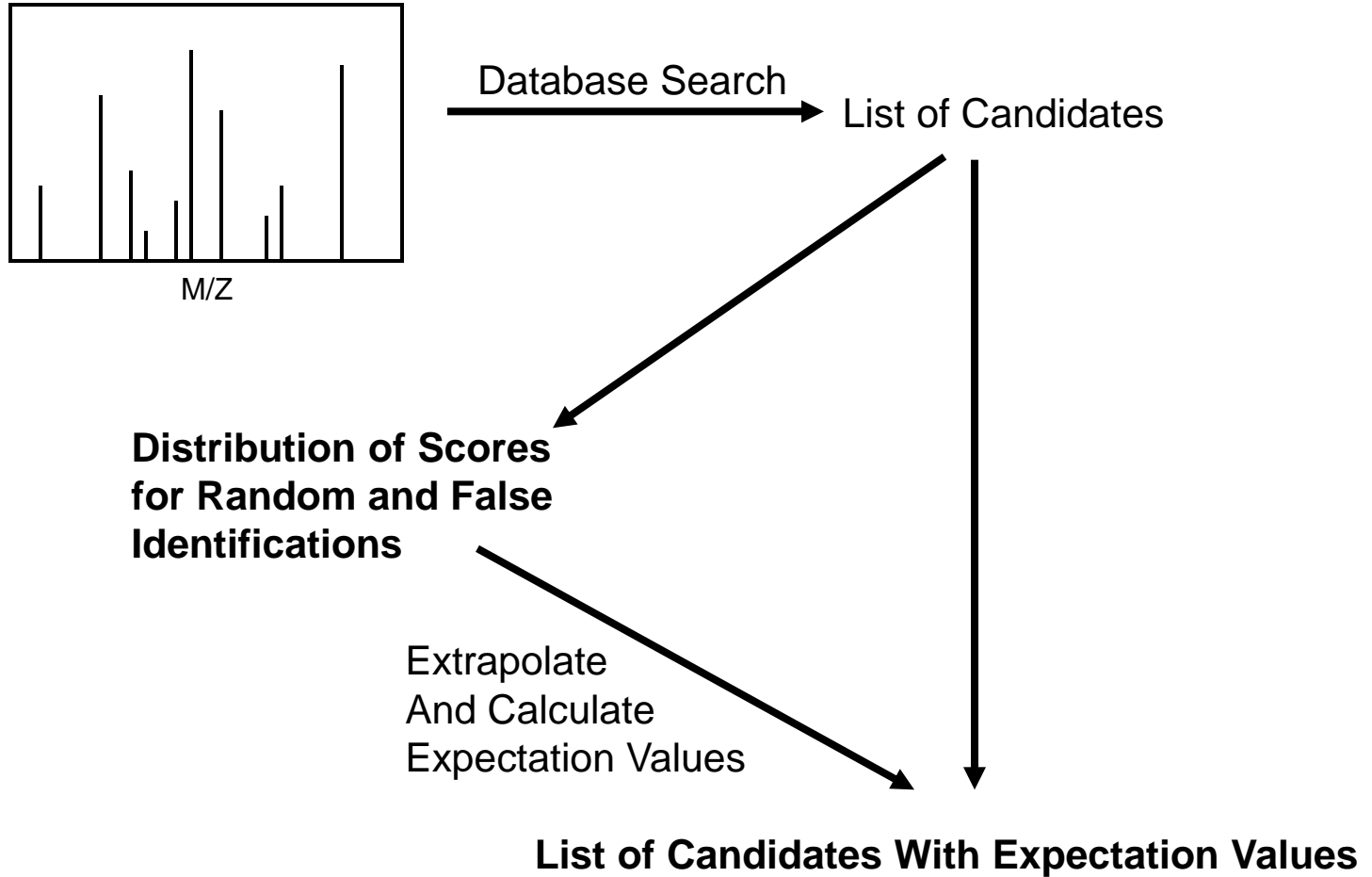


Database Search → List of Candidates

**Distribution of Scores
for Random and False
Identifications**

Extrapolate
And Calculate
Expectation Values

List of Candidates With Expectation Values



Rho-diagrams: Overall Quality of a Data Set

Expectation values as a function of score for random matching: $e(s) \propto \exp(-\beta s)$

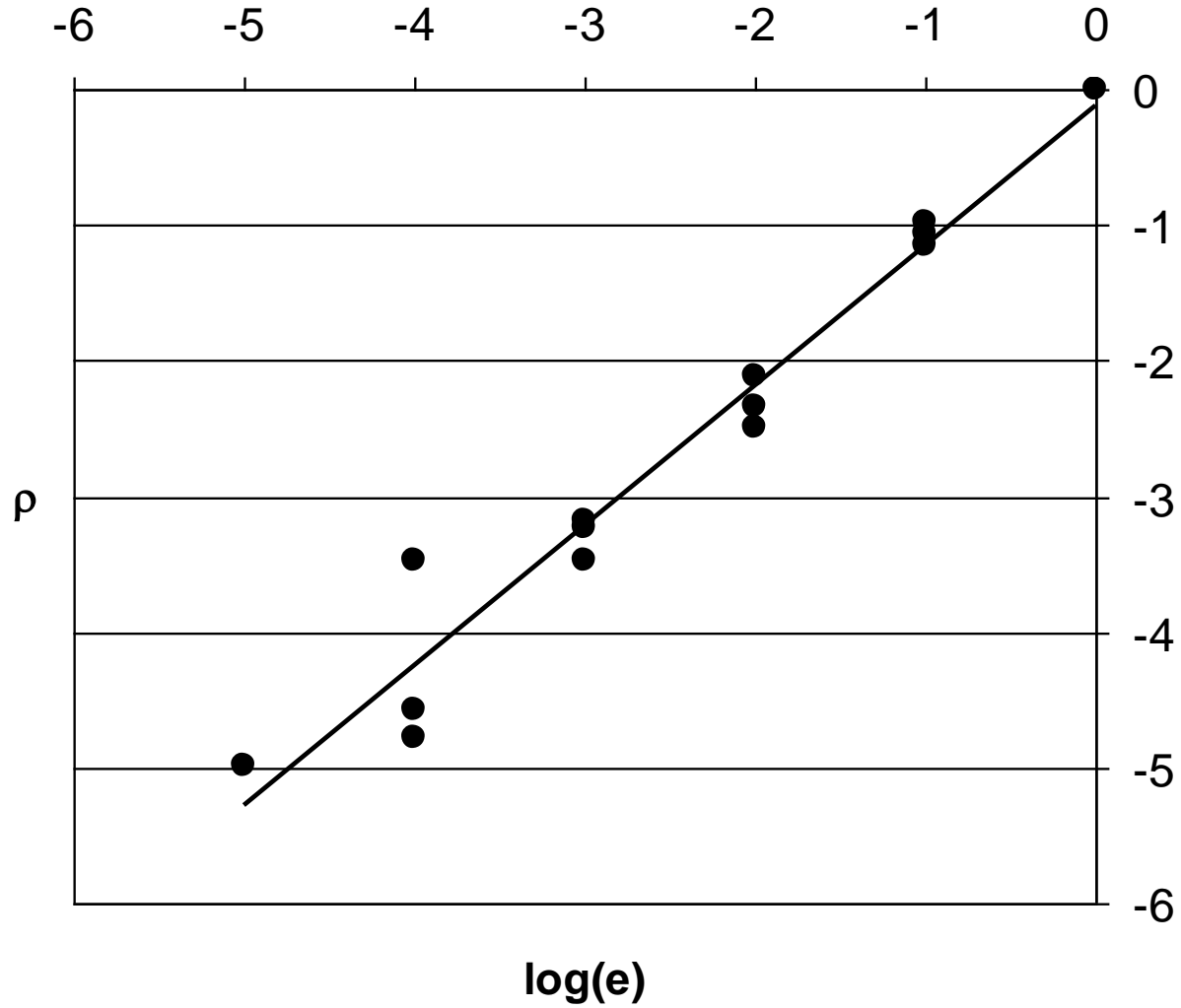
Definition: E_i ($i=0,-1,-2,\dots$) is the number of spectra that has been assigned an expectation value between $\exp(i)$ and $\exp(i-1)$. For random matching:

$$E_i = \int_{e=\exp(i-1)}^{e=\exp(i)} Nde = N\{\exp(i) - \exp(i-1)\}$$

$$\rho(i) = \log\left(\frac{E_i}{E_0}\right) = \log\left(\frac{N \exp(i) \{1 - \exp(-1)\}}{N \{1 - \exp(-1)\}}\right) = -i$$

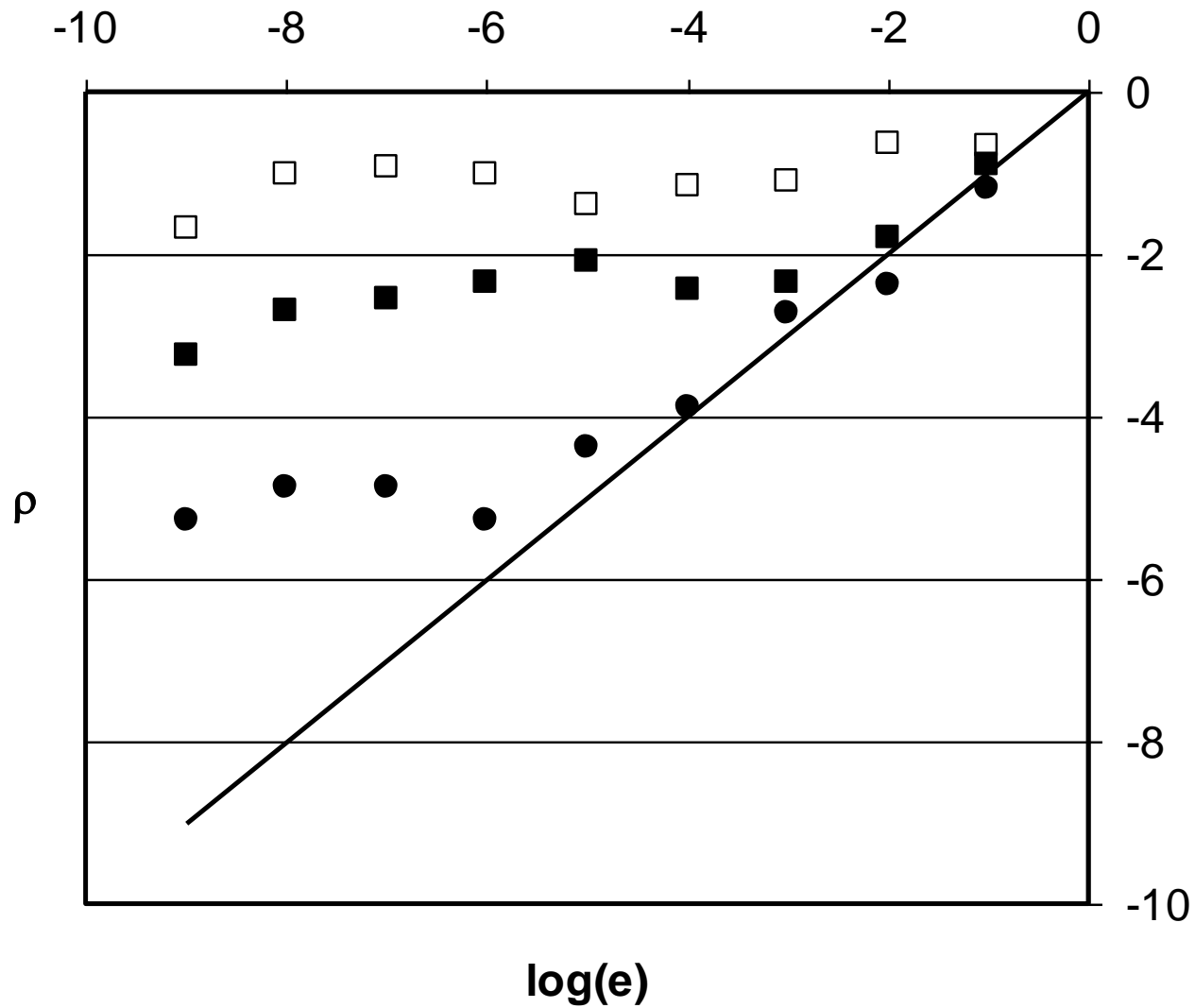
Rho-diagram

Random Matching

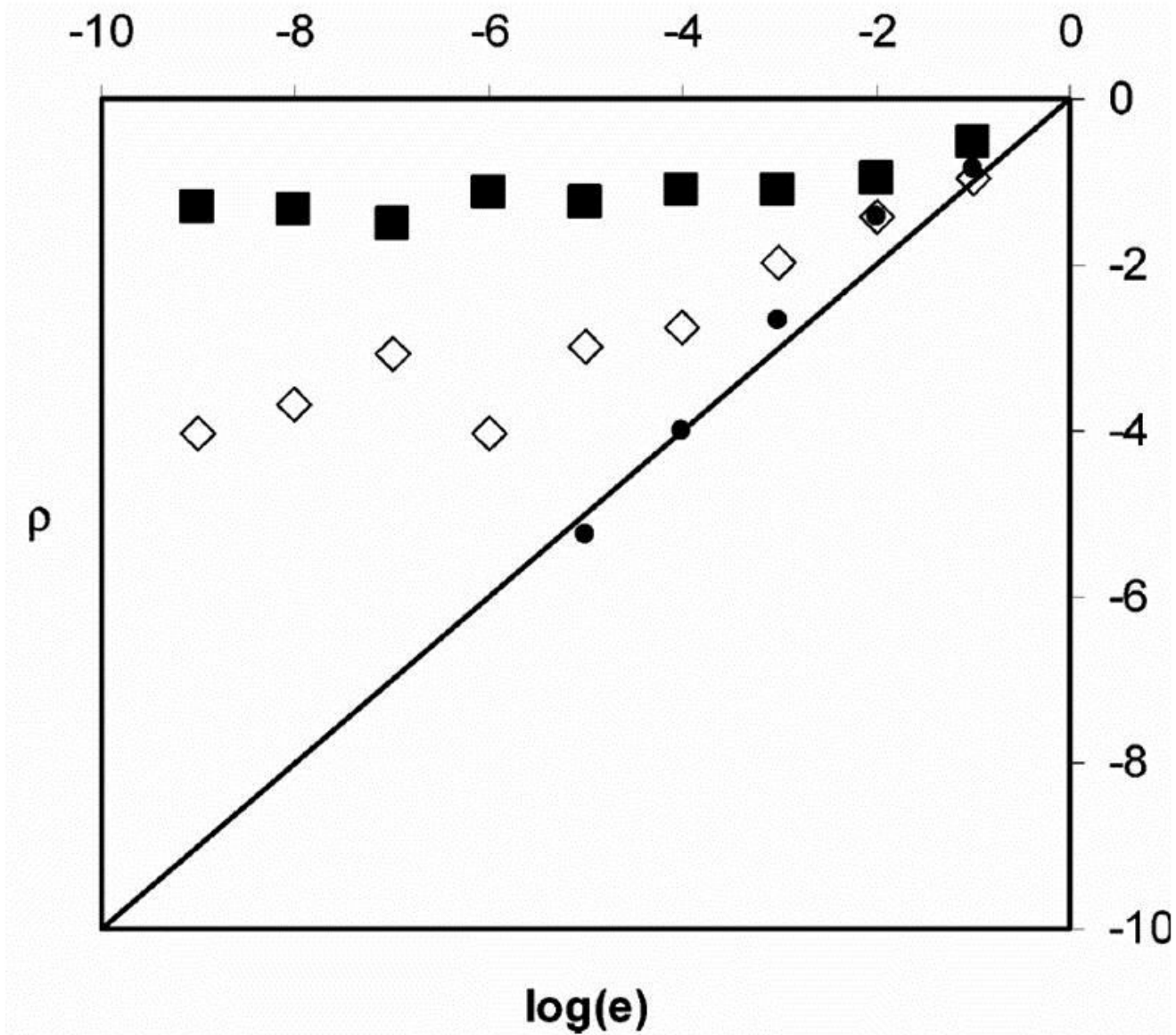


Rho-diagram

Data Quality



Rho-diagram Parameters



How many fragments are sufficient?

To identify an unmodified peptide?

To identify a modified peptide?

To localize a modification on a peptide?

How many fragments are sufficient?

How does it depend on different parameters?

- Precursor mass
- Precursor mass error
- Fragment mass error
- Background peaks

Simulations using synthetic spectra



Select a peptide sequence

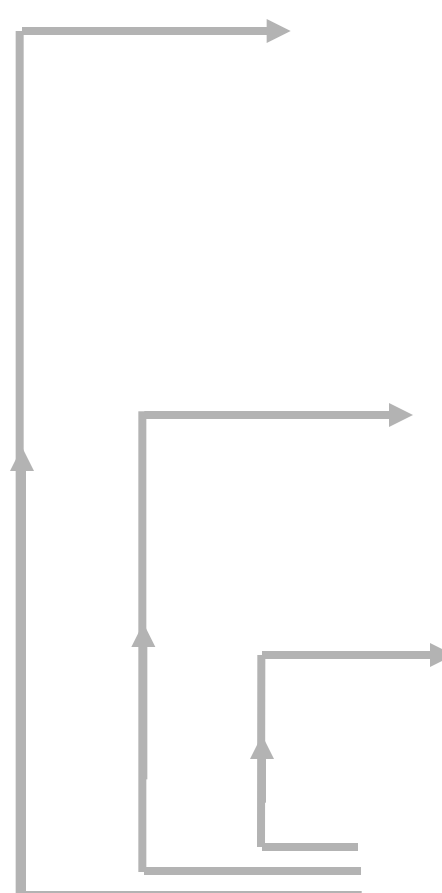
Calculate possible
fragment ion masses

Choose number of
fragment ions to select

Randomly select
fragment ions

Search and store result

Average over peptides



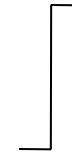
Simulations using synthetic spectra

Select a peptide sequence

LSDPGVSPAVLSLEMLTDR



↓
Calculate possible
fragment ion masses

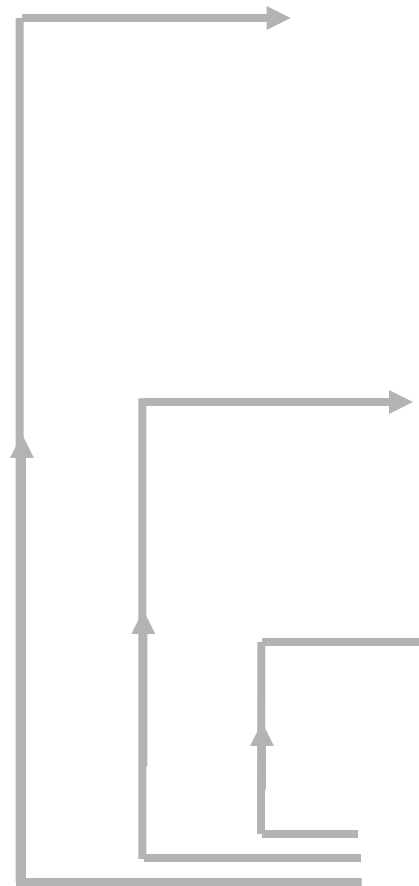


↓
Choose number of
fragment ions to select

↓
Randomly select
fragment ions

↓
Search and store result

↓
Average over peptides



Simulations using synthetic spectra

Select a peptide sequence

Calculate possible
fragment ion masses

LSDPGVSPAVLSLEMLTDR

Choose number of
fragment ions to select

Randomly select
fragment ions



Search and store result

Average over peptides

1825.92	175.12
1710.89	290.15
1609.84	391.19
1496.76	504.28
1365.72	635.32
1236.68	764.36
1123.59	877.44
1036.56	964.48
923.48	1077.56
824.41	1176.63
753.37	1247.67
656.32	1344.72
569.29	1431.75
470.22	1530.82
413.20	1587.84
316.15	1684.89
201.12	1799.92
114.09	1886.95

Simulations using synthetic spectra

Select a peptide sequence

Calculate possible
fragment ion masses

Choose number of
fragment ions to select

Randomly select
fragment ions

Search and store result

Average over peptides



1825.92	175.12
1710.89	290.15
1609.84	391.19
1496.76	201.12
1365.76	504.28
1236.67	964.48
1123.58	1123.59
1036.49	1247.67
923.40	1496.76
824.41	1530.82
753.32	1710.89
656.33	
569.24	1431.75
470.22	1530.82
413.20	1587.84
316.15	1684.89
201.12	1799.92
114.09	1886.95

Simulations using synthetic spectra

Select a peptide sequence

LSDPGVSPAVLSLEMLTDR



Calculate possible fragment ion masses

Choose number of fragment ions to select

Randomly select fragment ions

Search and store result

Average over peptides

Is the identified sequence identical to the one used to generate the synthetic data?

201.12
504.28
964.48
1123.59
1247.67
1496.76
1530.82
1710.89



Is it significant?



Simulations using synthetic spectra

Select a peptide sequence

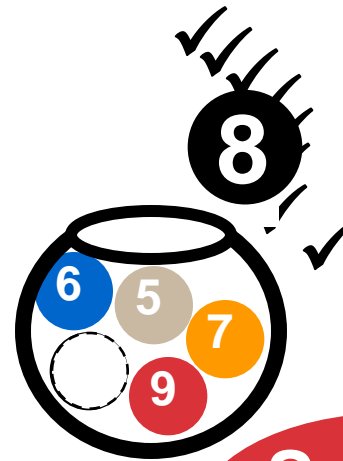
Calculate possible
fragment ion masses

Choose number of
fragment ions to select

Randomly select
fragment ions

Search and store result

Average over peptides



1825.92	175.12
1710.89	290.15
1609.84	391.19
1496.76	201.12
1365.72	504.28
1236.68	964.48
1123.59	1123.59
1036.56	1247.67
923.48	1496.76
824.41	1530.82
753.37	1710.89
656.32	1431.75
569.29	1530.82
470.22	1587.84
413.20	1684.89
316.15	1799.92
201.12	1886.95
114.09	



Simulations using synthetic spectra

Select a peptide sequence

Calculate possible
fragment ion masses

Choose number of
fragment ions to select

Randomly select
fragment ions

Search and store result

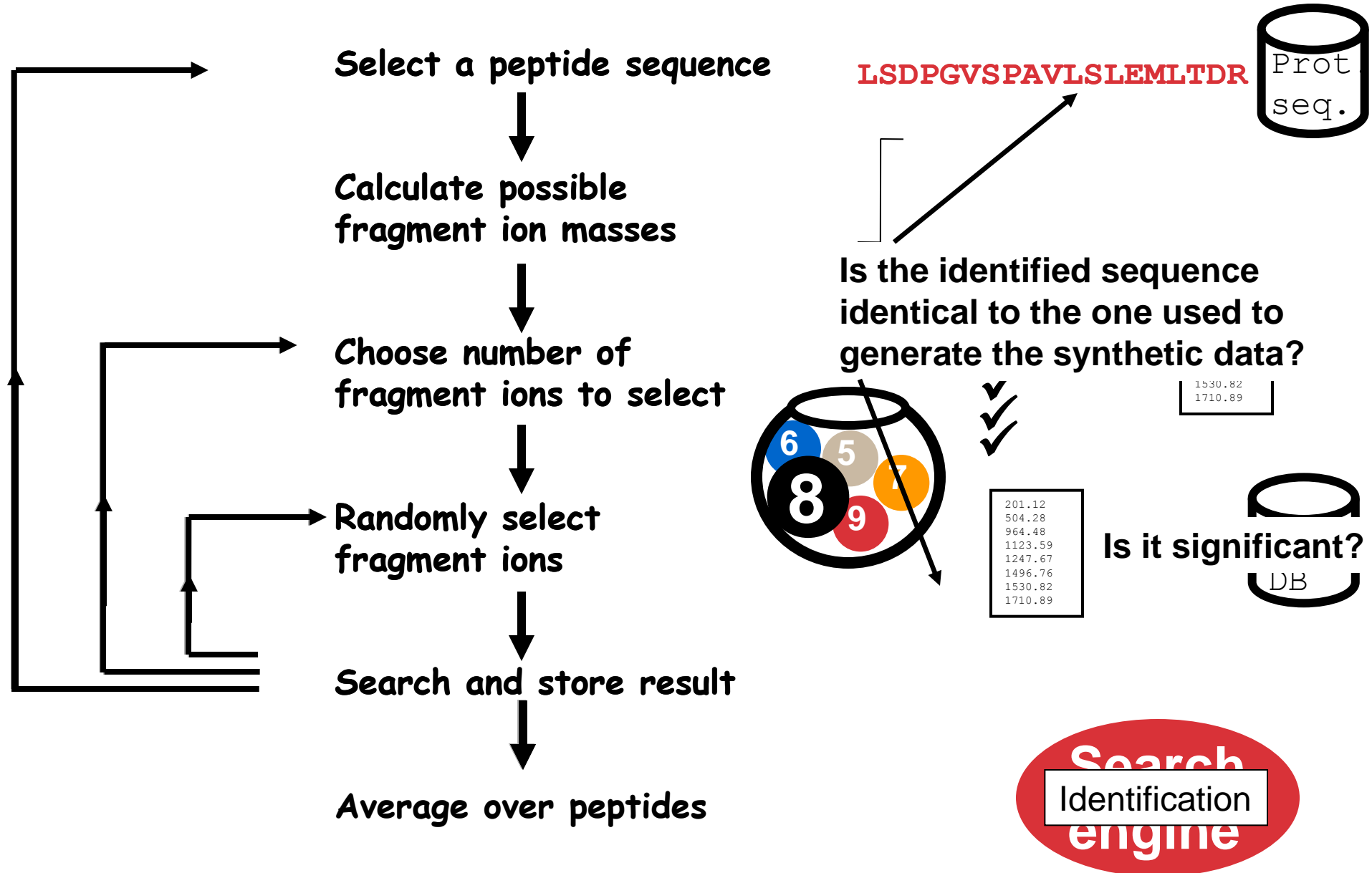
Average over peptides



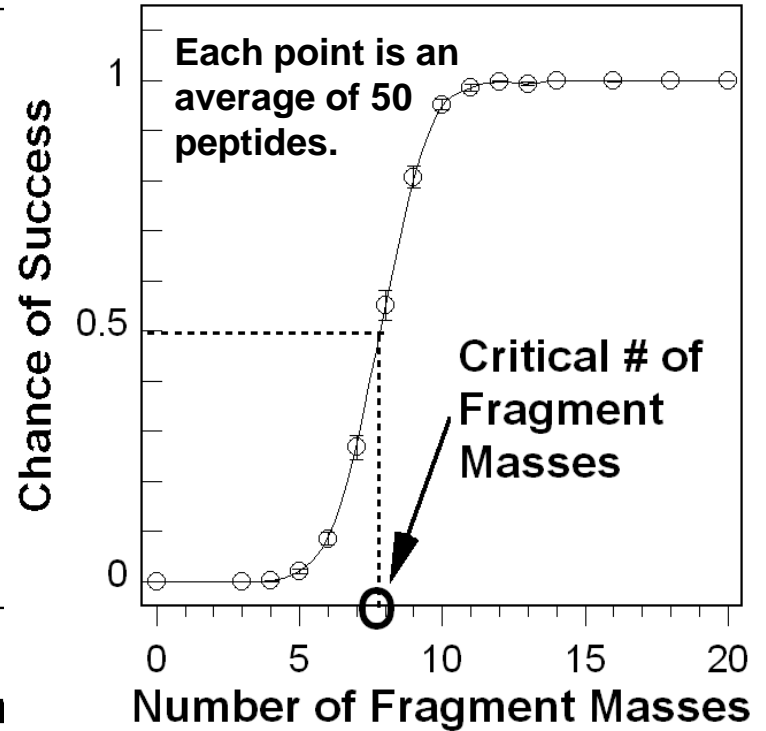
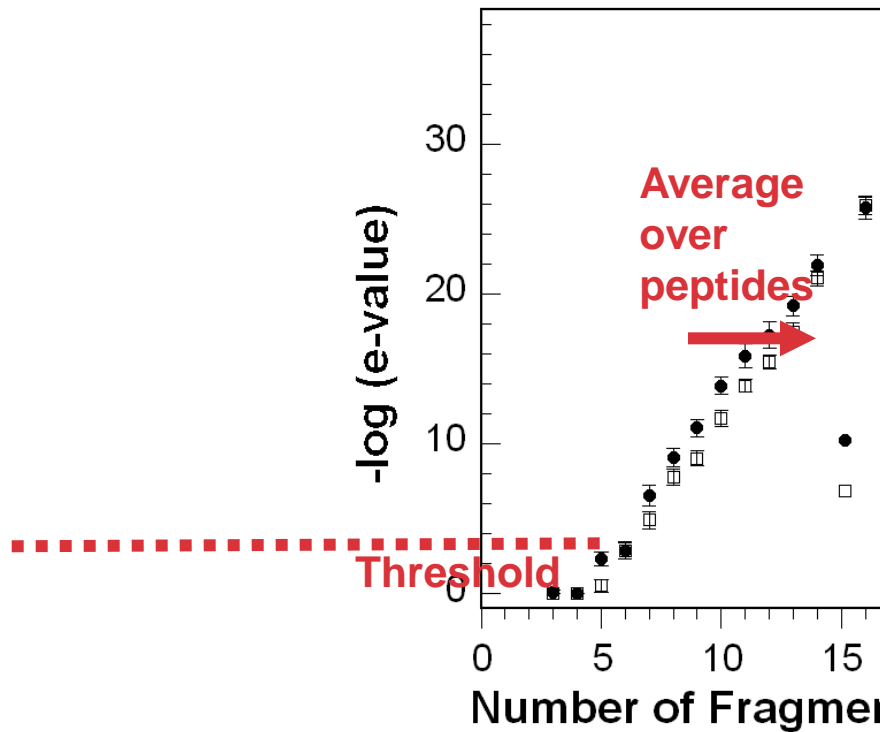
1825.92	175.12
1710.89	290.15
1609.84	391.19
1496.76	201.12
1365.72	504.28
1236.68	964.48
1123.59	1123.59
1036.56	1247.67
923.48	1496.76
824.41	1530.82
753.37	1710.89
656.32	1431.75
569.29	1530.82
470.22	413.20
413.20	1587.84
316.15	1684.89
201.12	1799.92
114.09	1886.95



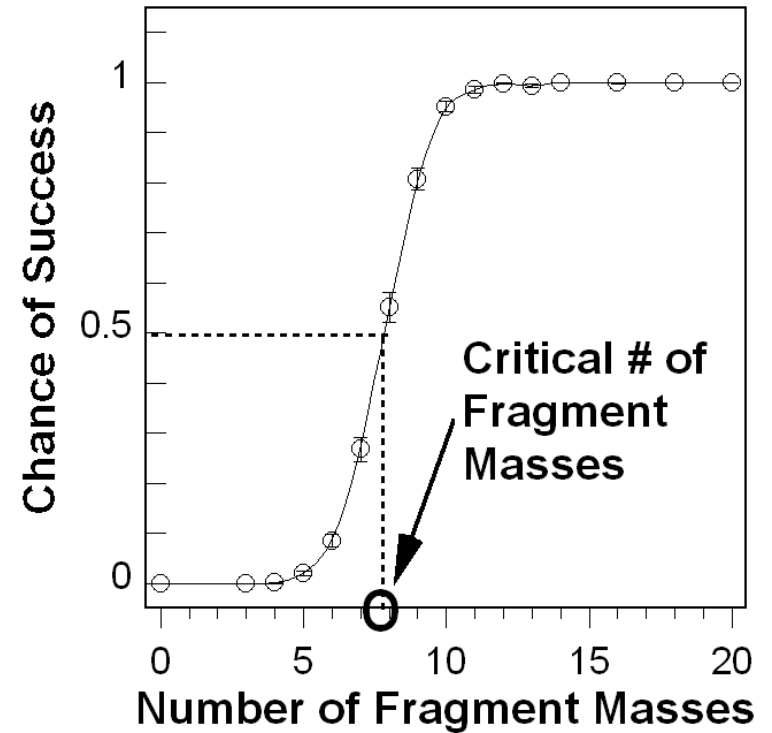
Simulations using synthetic spectra



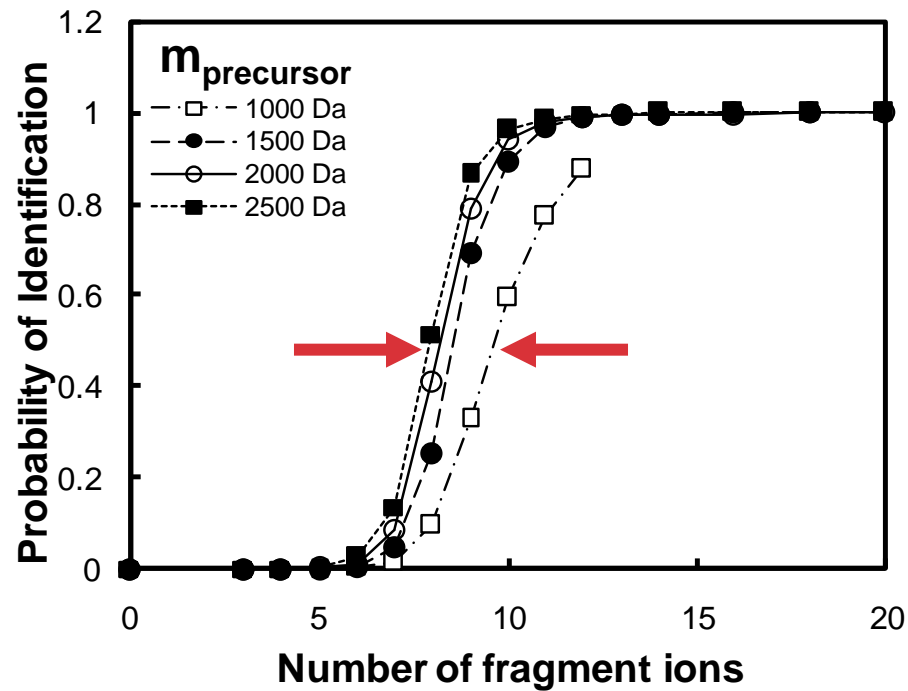
Simulations using synthetic spectra



Critical number of fragment masses

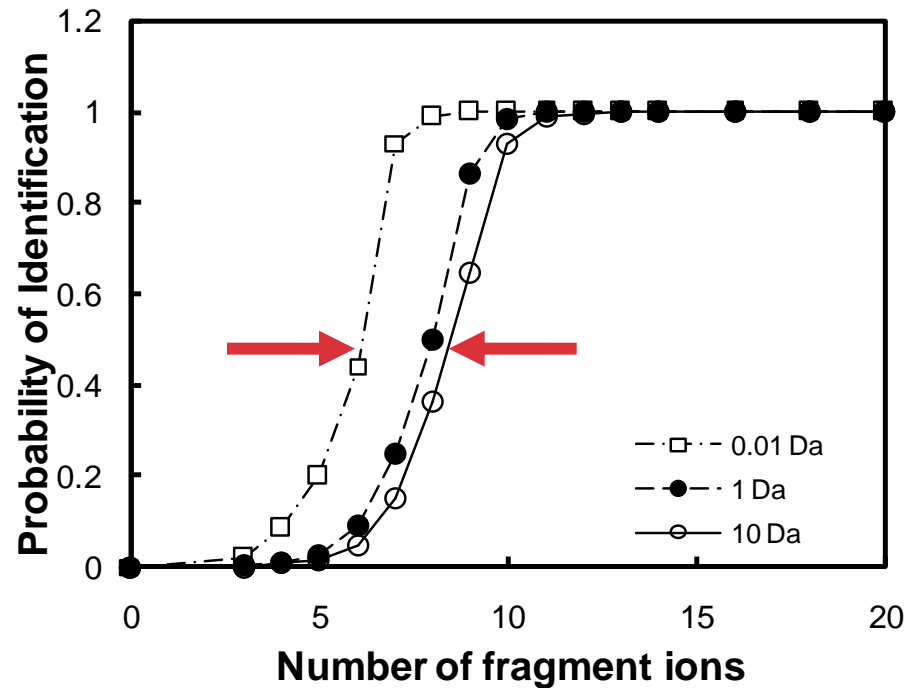


Small peptides are slightly more difficult to identify



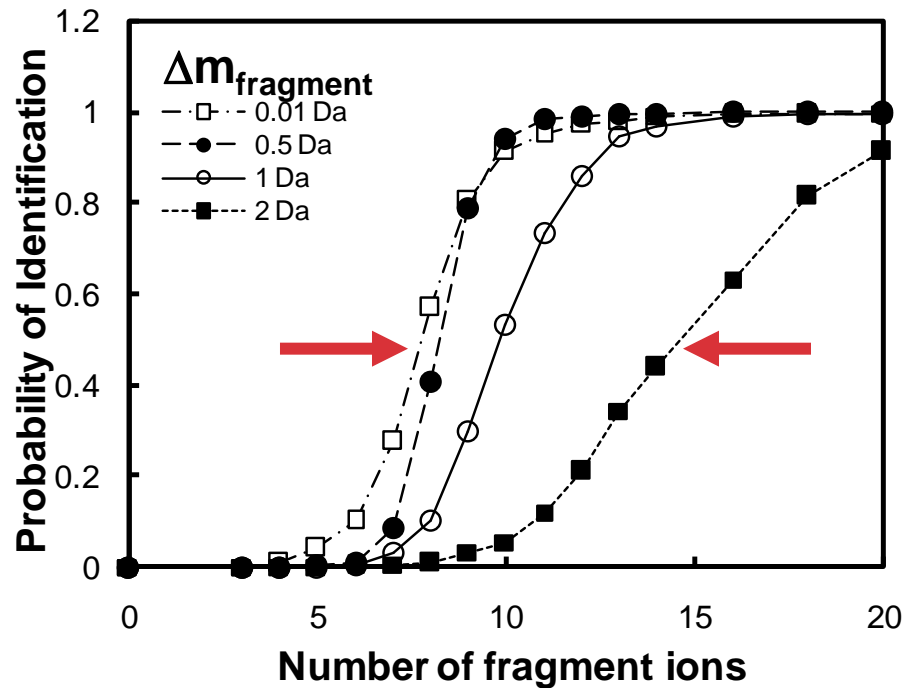
$\Delta m_{\text{precursor}} = 1 \text{ Da}$
 $\Delta m_{\text{fragment}} = 0.5 \text{ Da}$
No modification

A lower precursor mass error requires fewer fragment masses for identification of unmodified peptides



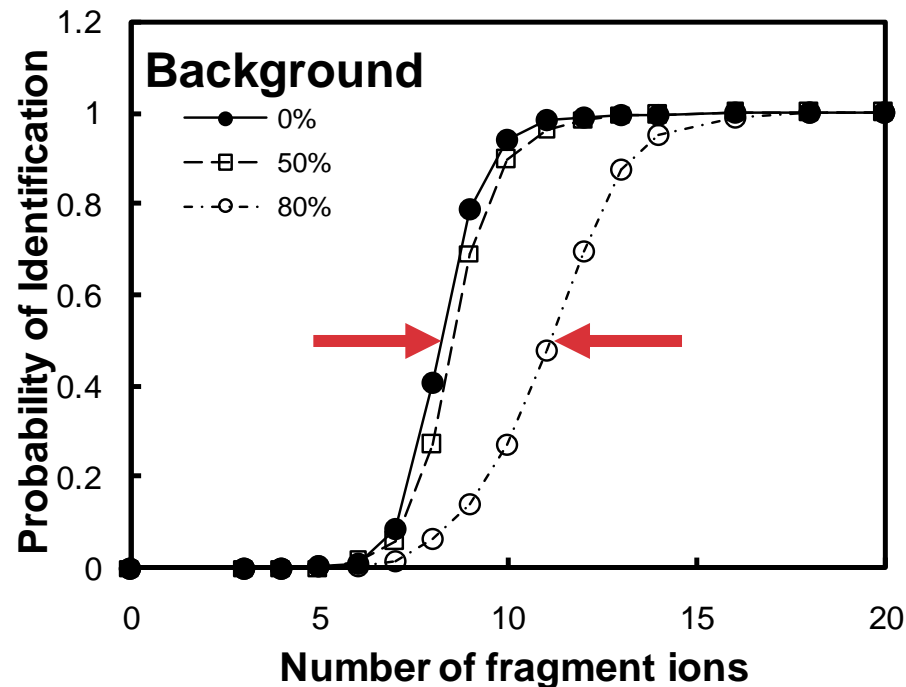
$m_{\text{precursor}} = 2000 \text{ Da}$
 $\Delta m_{\text{fragment}} = 0.5 \text{ Da}$
No modification

The dependence on the fragment mass error is weak below a threshold for identification of unmodified peptides



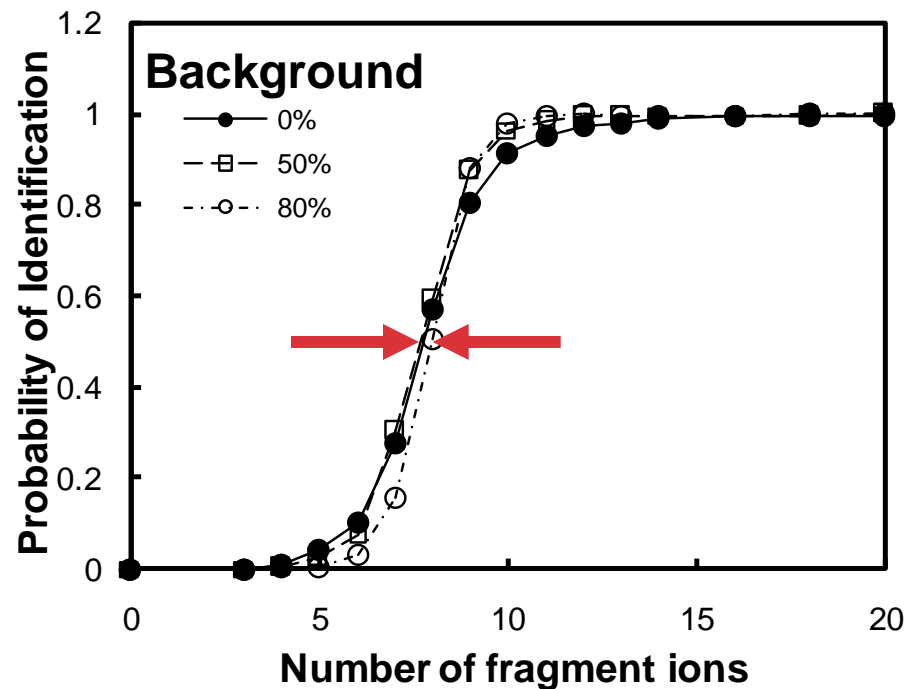
$m_{\text{precursor}} = 2000 \text{ Da}$
 $\Delta m_{\text{precursor}} = 1 \text{ Da}$
No modification

A moderate number of background peaks can be tolerated when identifying unmodified peptides



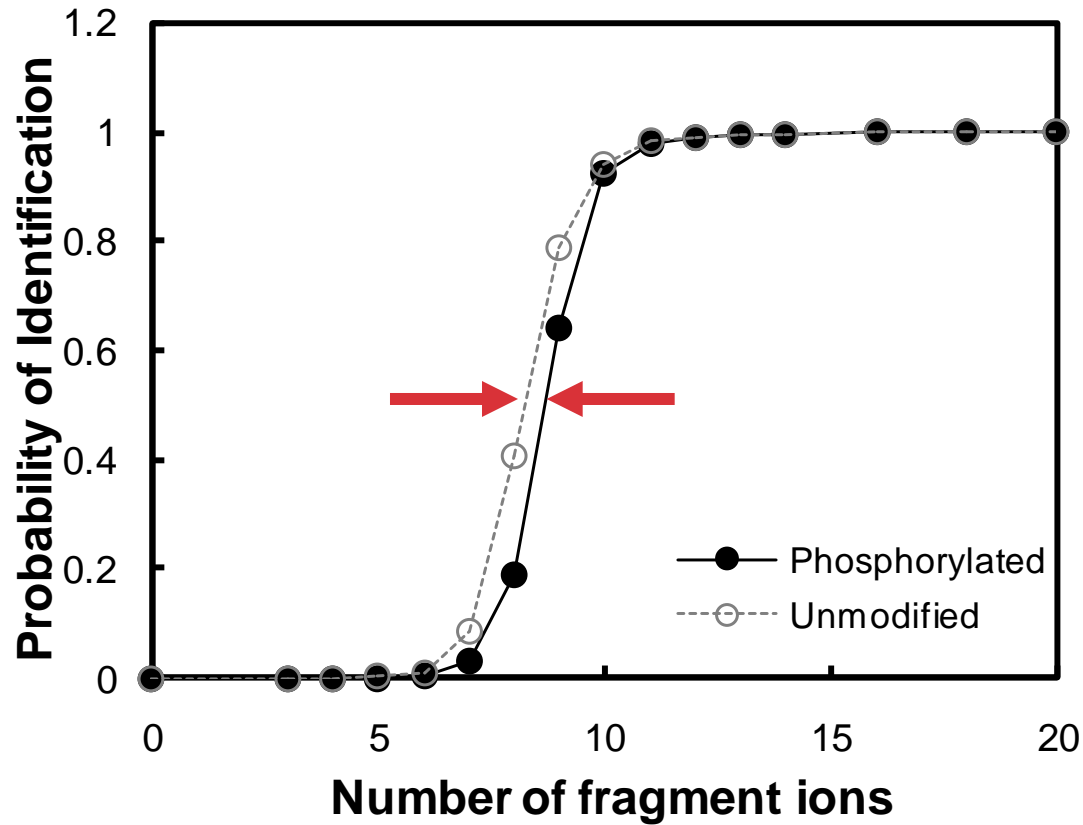
$m_{\text{precursor}} = 2000 \text{ Da}$
 $\Delta m_{\text{precursor}} = 1 \text{ Da}$
 $\Delta m_{\text{fragment}} = 0.5 \text{ Da}$
No modification

A large number of background peaks can be tolerated if the fragment mass is accurate



$m_{\text{precursor}} = 2000 \text{ Da}$
 $\Delta m_{\text{precursor}} = 1 \text{ Da}$
 $\Delta m_{\text{fragment}} = 0.01 \text{ Da}$
No modification

Identification of phosphopeptides is only slightly more difficult



$m_{\text{precursor}} = 2000 \text{ Da}$
 $\Delta m_{\text{precursor}} = 1 \text{ Da}$
 $\Delta m_{\text{fragment}} = 0.5 \text{ Da}$

Proteomics Informatics -

**Protein identification I: searching protein sequence
collections and significance testing (Week 4)**
