Proteomics Informatics – Protein identification II: search engines and protein sequence databases (Week 5)

General Criteria for a Good Protein Identification Algorithms

The response to random input data should be random.

Maximum number of correct identification and minimum number of incorrect identifications for any data set.

Maximal separation between scores for correct identifications and the distribution of scores for random matching proteins for any data set.

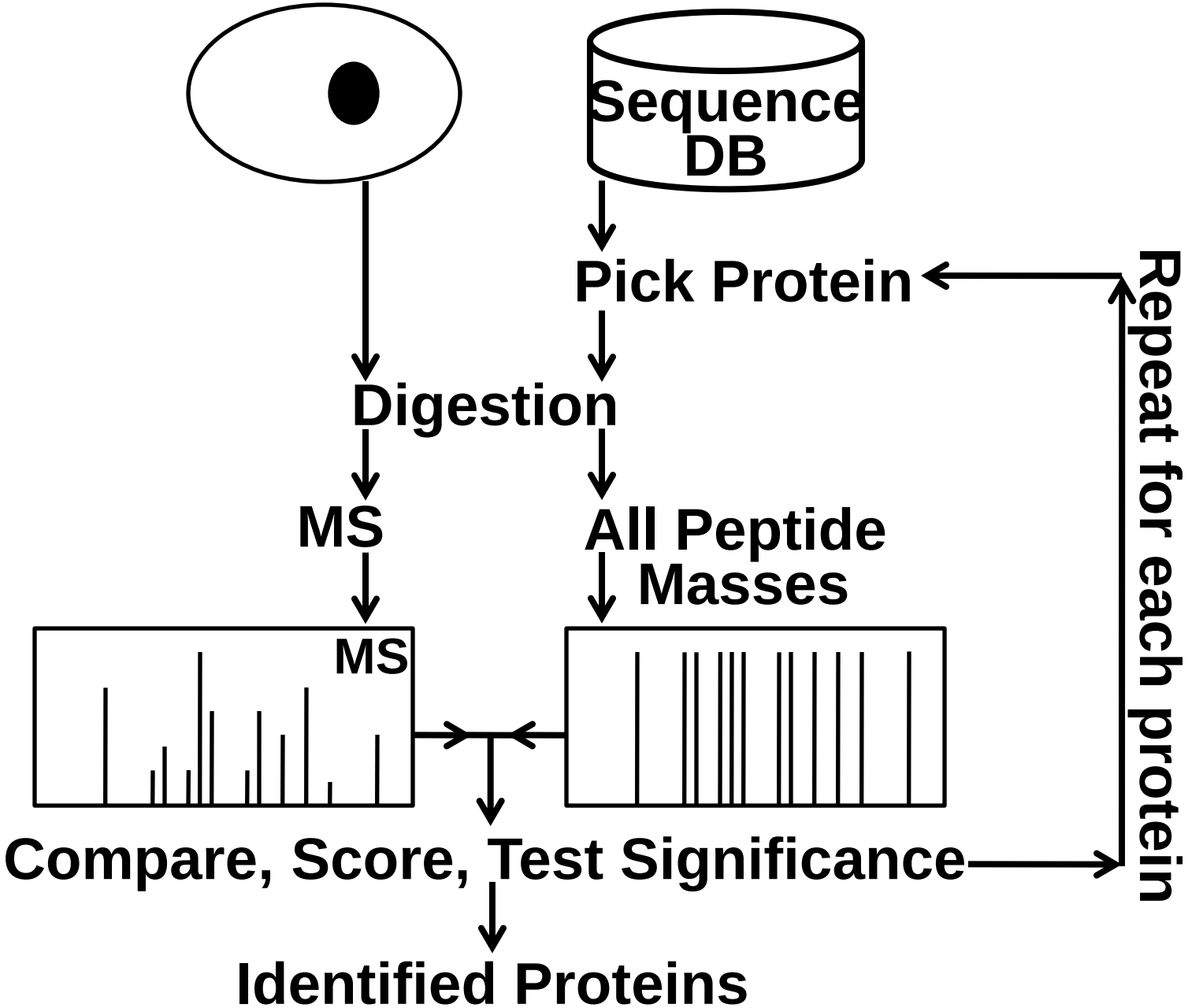
The statistical significance of the results should be calculated.

The searches should be fast.

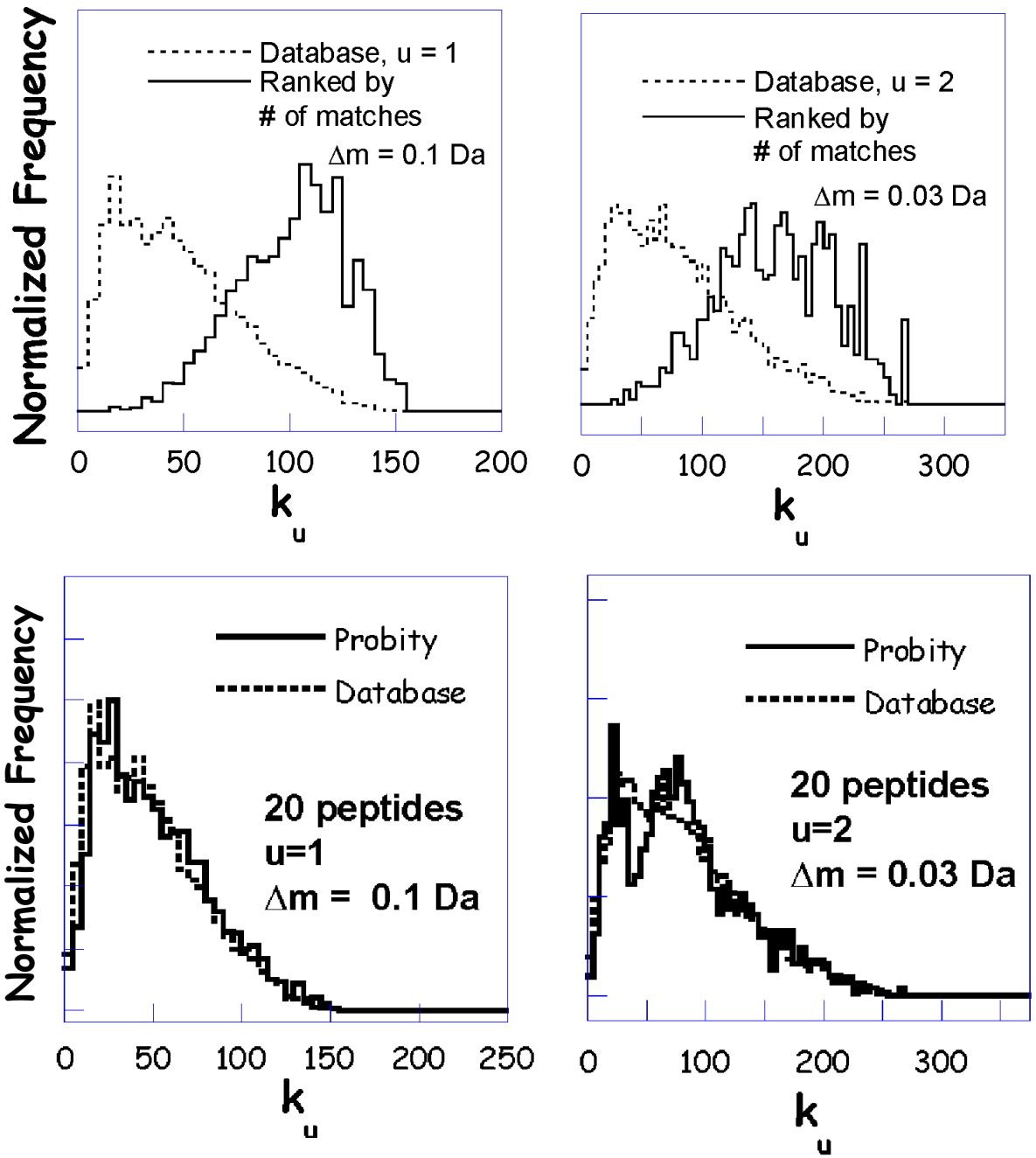
Search Parameters

Parent tolerance	+/- daltons/ppm
Frag. Tolerance	+/- daltons/ppm
Complete mods	Cys alkylation
Potential mods (artifacts)	Met/Trp oxidation, Gln/Asn deamidation
Potential mods (PTMs)	Phosphoryl, sulfonyl, acetyl, methyl, glycosyl, GPI
Cleavage	Trypsin ([KR]|{P})
Scoring method	Scores or statistics
Sequences	FASTA files

Identification – Peptide Mass Fingerprinting



Response to Random Data



ProFound - Search Parameters

General

Sample ID

Database: NCBI nr (2010/11/09)

Taxonomy: . . . . Xenopus laevis

Protein Mass: 0 - 3000 kDa

Protein pI: 0 - 14

Expect: 1

Z: show 10 candidates

Digestion

Allow maximum 1 missed cleavages

Enzyme: Trypsin

For user-defined cleavage, click here.

Modifications

Complete Modification(s):
- Unmodified
- 4-vinyl-pyridine (Cys)
- Acrylamide (Cys)
- Iodoacetamide (Cys)
- Iodoacetic acid (Cys)

Partial Modification: Methionine oxidation

For more partial modifications, click here.

Masses

Average Masses:

Monoisotopic Masses:

```
2672.328
2693.422
2723.384
3006.654
```

Mass tolerance (average): +/- 1

Mass tolerance (monoisotopic): +/- 10

Tolerance unit: Da % ppm

Charge state: M MH+

Identify Protein | Extra Settings | Example | Reset Form

http://prowl.rockefeller.edu/

ProFound – Protein Identification by Peptide Mapping

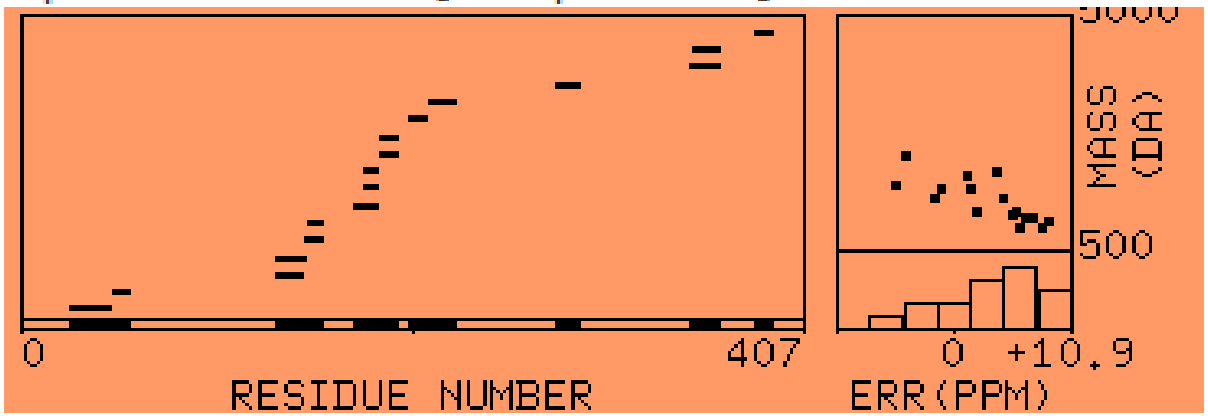
$$P(k \mid DI) \propto P(k \mid I) \frac{(N-r)!}{N!} \prod_{i=1}^{r} g_i \left(\frac{m_{\max} - m_{\min}}{2\sigma} \right)^r \exp\left[\frac{r}{2} - \frac{\sum_{i=1}^{r} (m_i - m_{i0})^2}{2\sigma^2} \right] F_{pattern}$$

W. Zhang & B.T. Chait,
Analytical Chemistry
72 (2000) 2482-2489

ProFound Results

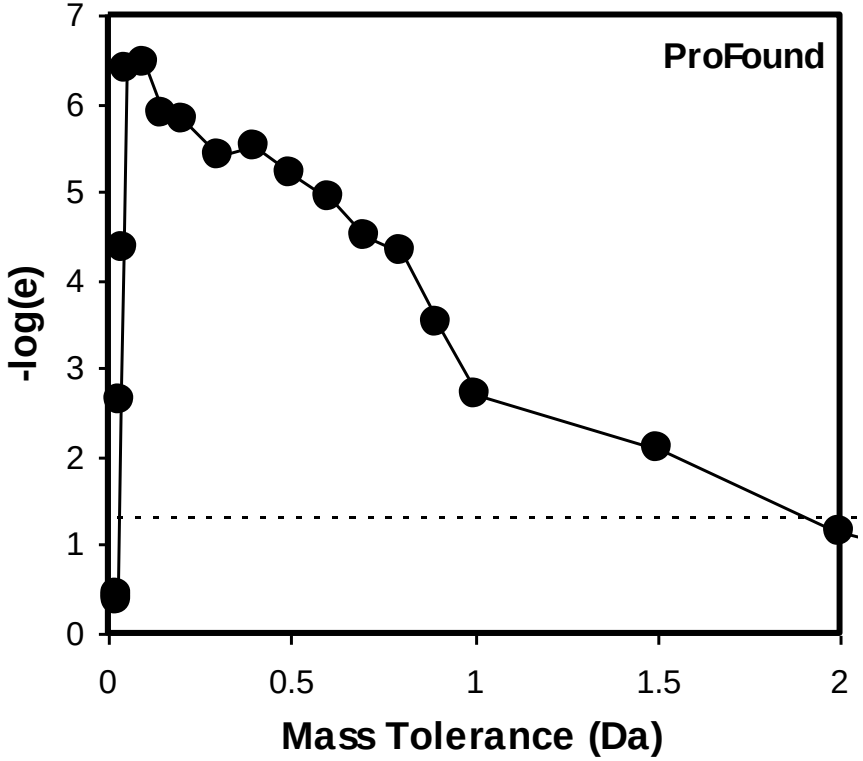
Protein Candidates

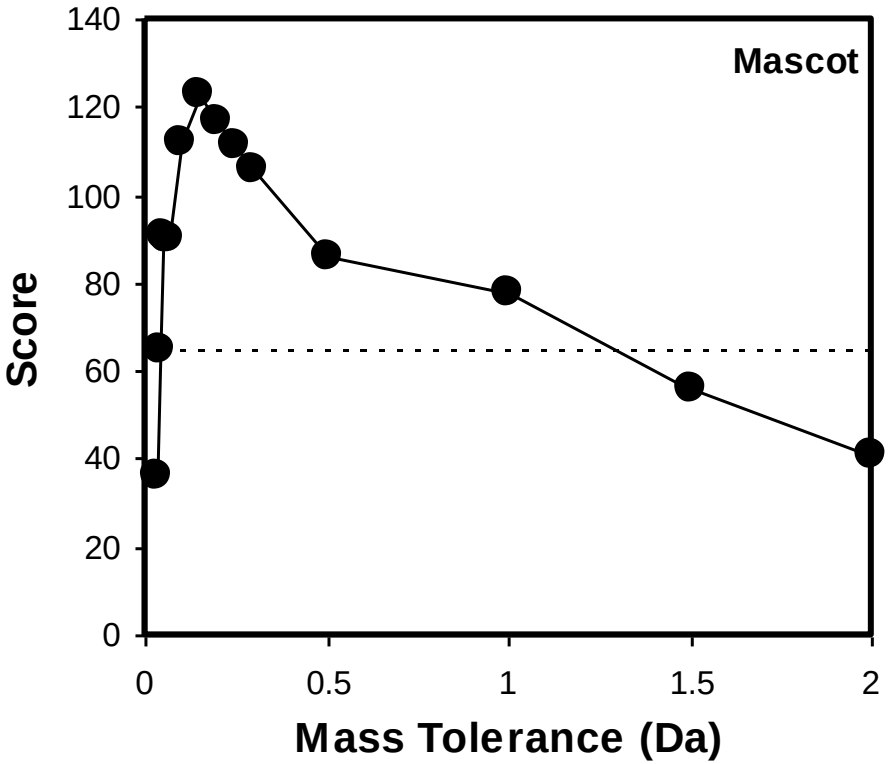
Rank	Expectation	Protein Information and Sequence Analyse Tools (T)	%	pI	kDa
+1	5.110^{-7}	gi|148236543|ref|NP_001081565.1| serine/threonine-protein kinase 6-A [Xenopus laevis]	36	9.6	46.35
+2	0.057		8	5.3	147.73
3	0.094		9	7.5	126.81



Measured Mass(M)	Avg/ Mono	Computed Mass	Error (ppm)	Residues Start	To	Missed Cut	Peptide sequen
908.490	M	908.482	8	179	186	0	AGVEHQLR
938.503	M	938.497	6	151	158	0	FGNVYLAR
1064.593	M	1064.583	9	179	187	1	AGVEHQLRR
1079.618	M	1079.608	9	49	58	0	ILGPSNVPQR
1109.590	M	1109.582	7	188	196	0	EVEIQSHLR
1123.622	M	1123.613	7	149	158	1	GKFGNVYLAR
1190.687	M	1190.681	5	383	392	0	GVLEHPWIIK
1227.570	M	1227.567	2	203	212	0	LYGYFHDASR
1265.691	M	1265.683	6	187	196	1	REVEIQSHLR
1493.792	M	1493.794	-2	174	186	1	SQLEKAGVEHQLR
1528.749	M	1528.742	5	279	292	0	IADFGWSVHAPSSR

Peptide Mapping – Mass Accuracy



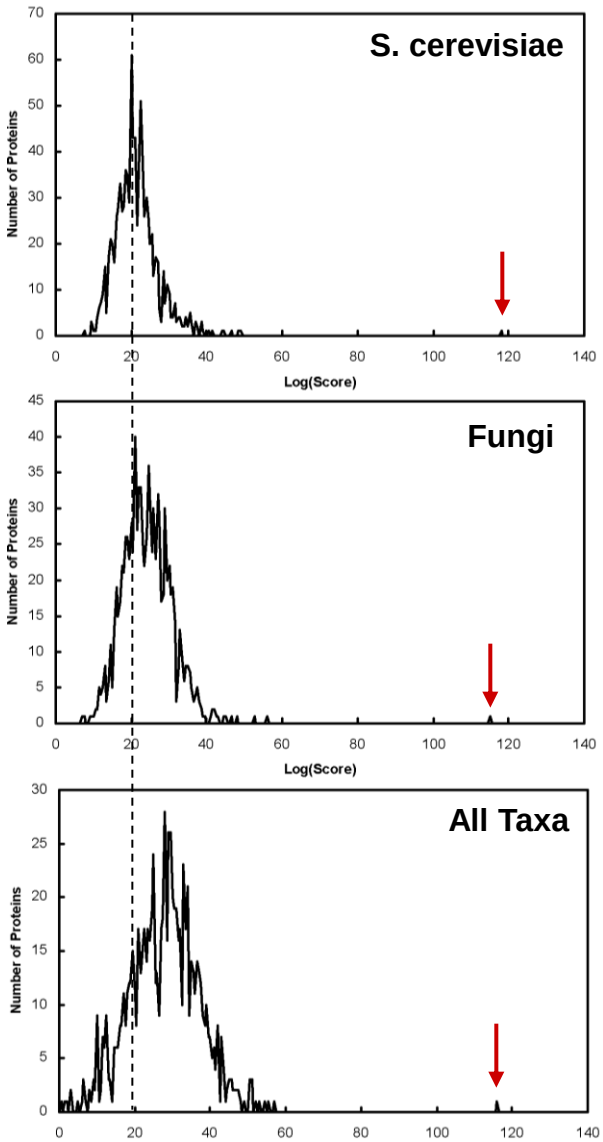


Peptide Mapping - Database Size

Expectation Values

Peptide mapping example:

S. Cerevisiae	4.8e-7
Fungi	8.4e-6
All Taxa	2.9e-4

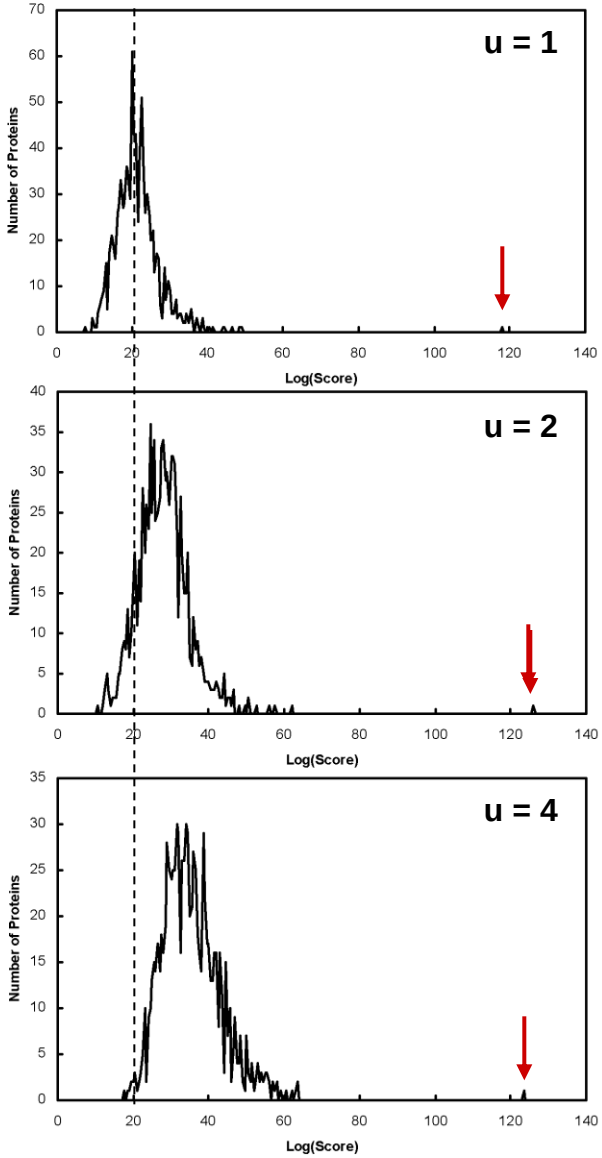


Missed Cleavage Sites

Expectation Values

Peptide mapping example:

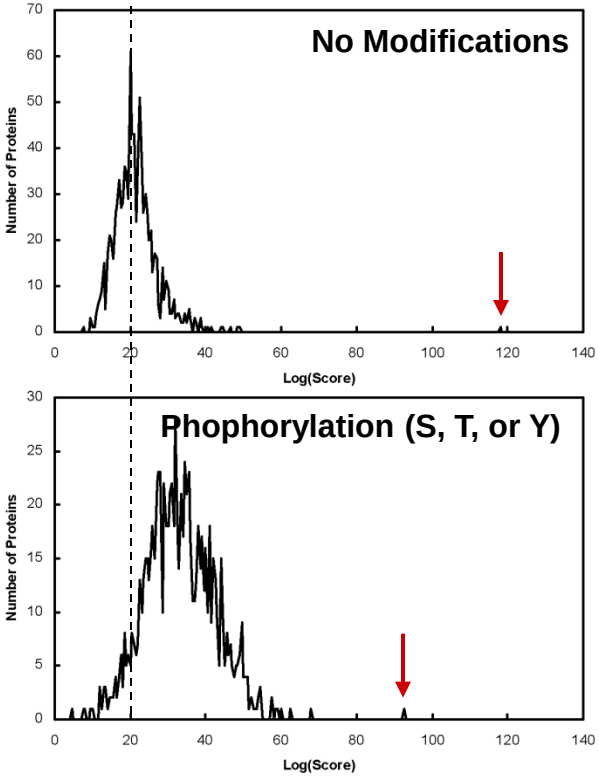
u=1	4.8e-7
u=2	1.1e-5
u=4	6.8e-4



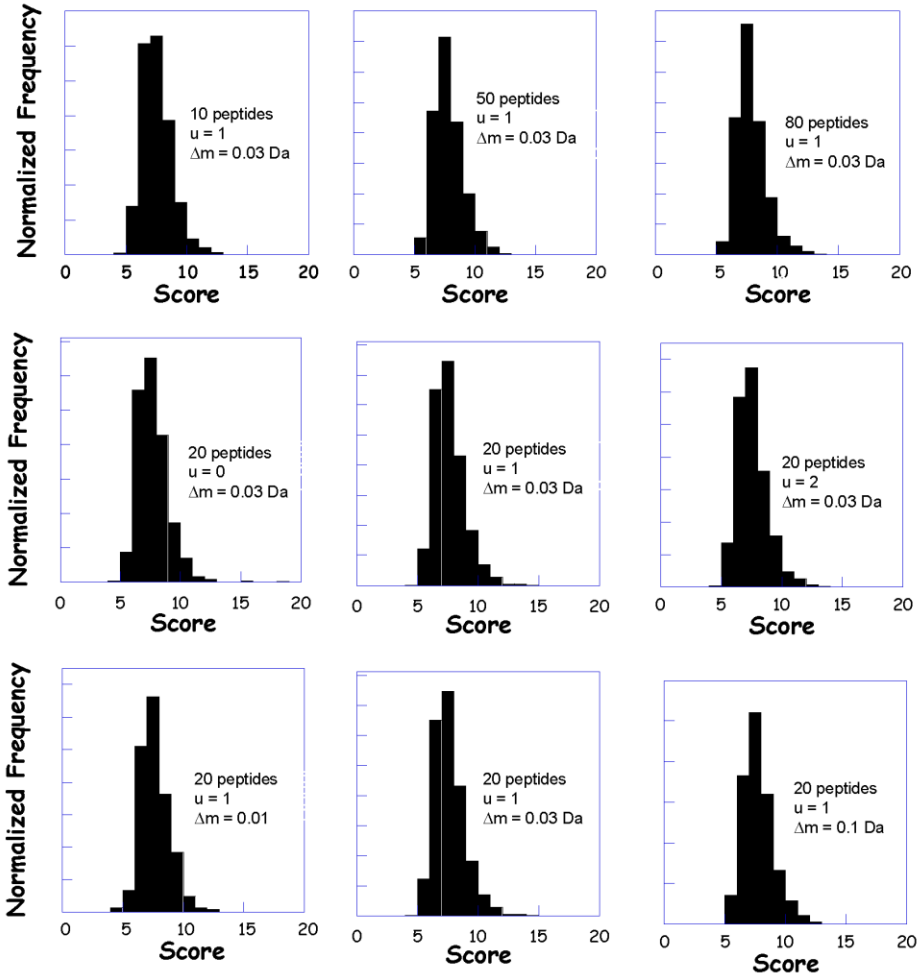
Peptide Mapping - Partial Modifications

	Searched Without Modifications	Searched With Possible Phosphorylation of S/T/Y
DARPP-32	0.00006	0.01
CFTR	0.00002	0.005

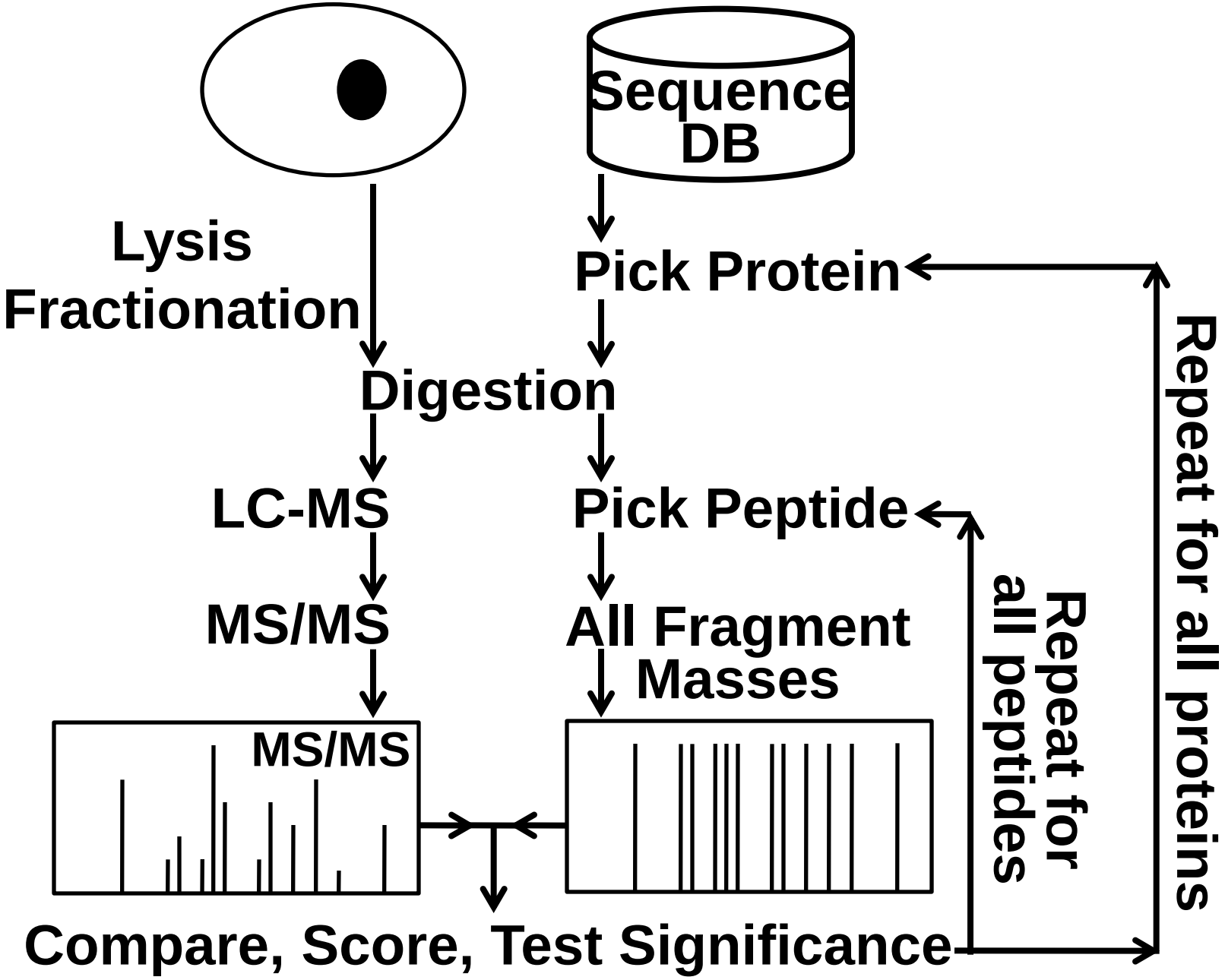
Even if the protein is modified it is usually better to search a protein sequence database without specifying possible modifications using peptide mapping data.



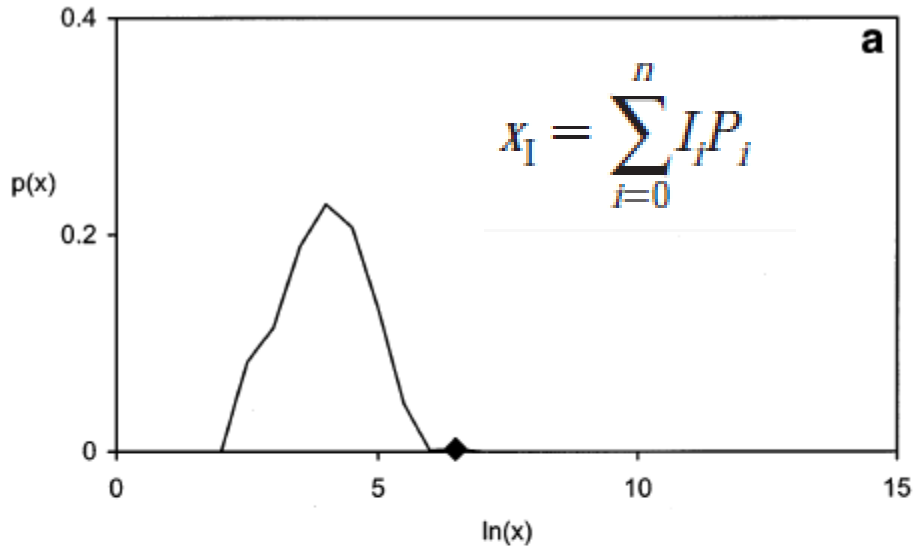
Peptide Mapping - Ranking by Direct Calculation of the Significance

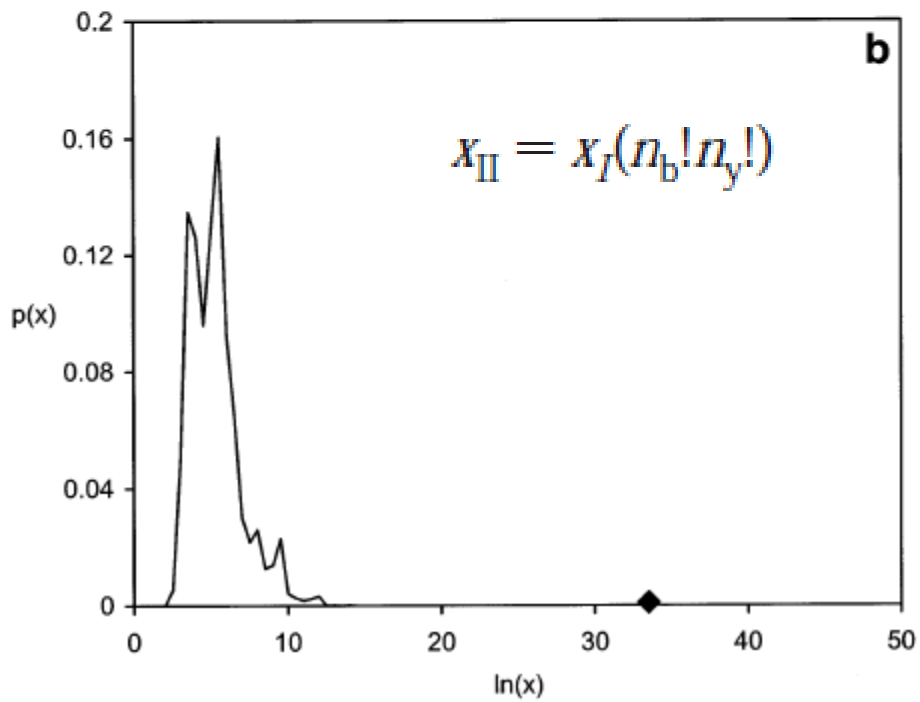


Tandem MS – Database Search

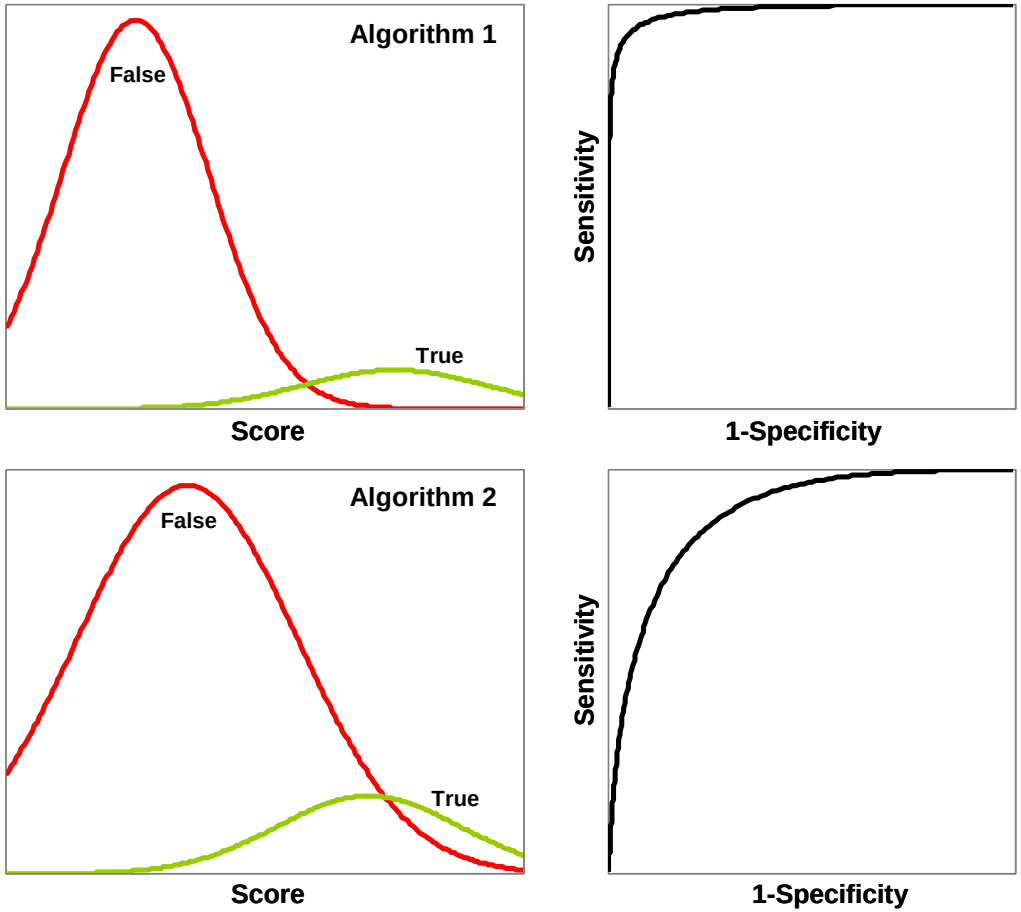


Algorithms


a
$x_{\mathrm{I}} = \sum_{i=0}^{n} I_i P_i$
p(x)
0.4
0.2
0
0
5
10
15
ln(x)


b
$x_{\mathrm{II}} = x_I(n_{\mathrm{b}}!n_{\mathrm{y}}!)$
p(x)
0.2
0.16
0.12
0.08
0.04
0
0
10
20
30
40
50
ln(x)

Comparing and Optimizing Algorithms

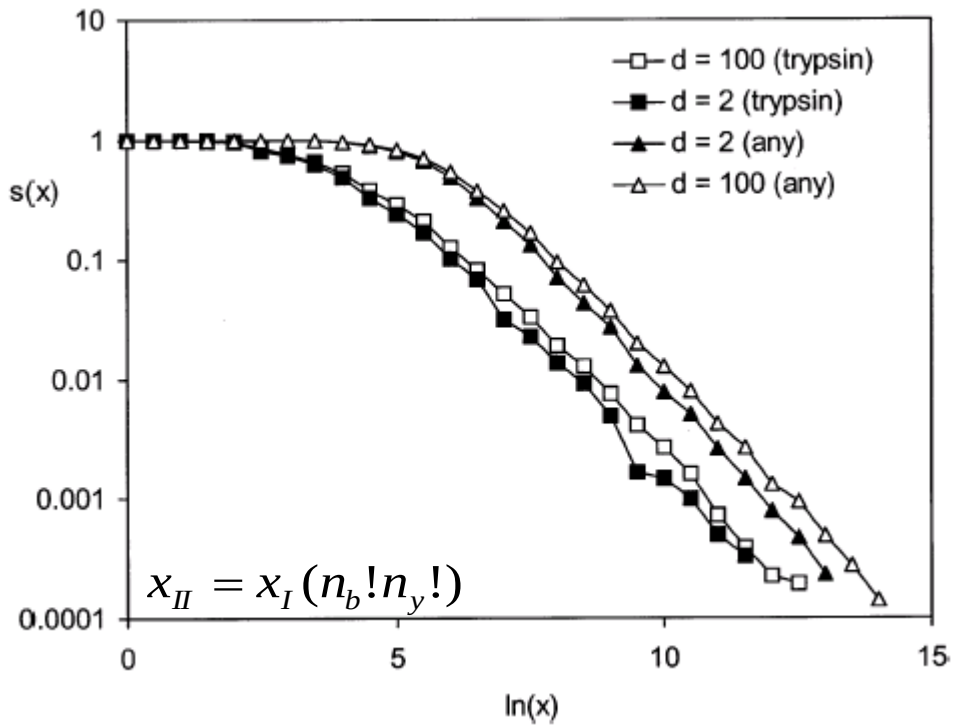


MS/MS - Parent Mass Error and Enzyme Specificity

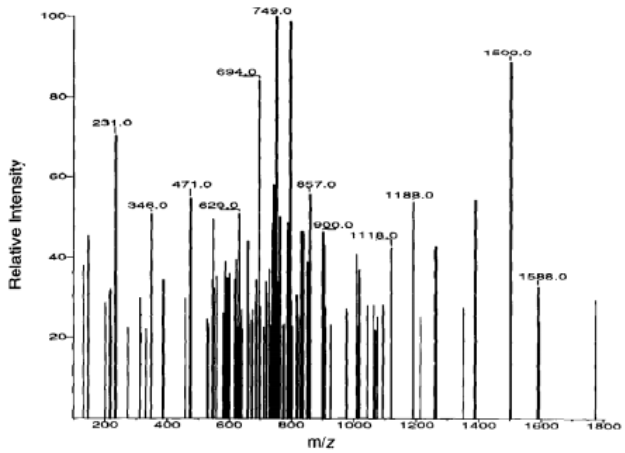
Expectation Values

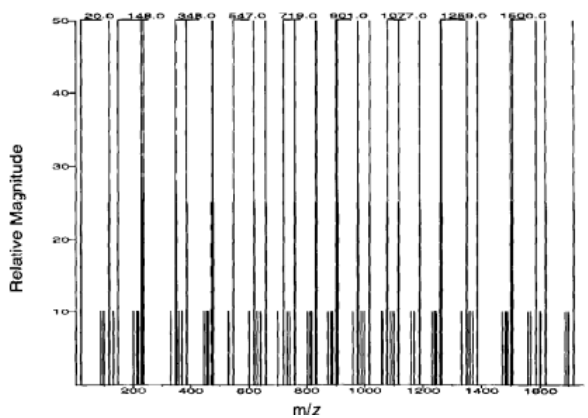
MS/MS example:

Δm=2, Trypsin	2.5e-5
Δm=100, Trypsin	2.5e-5
Δm=2, non-specific	7.9e-5
Δm=100, non-specific	1.6e-4



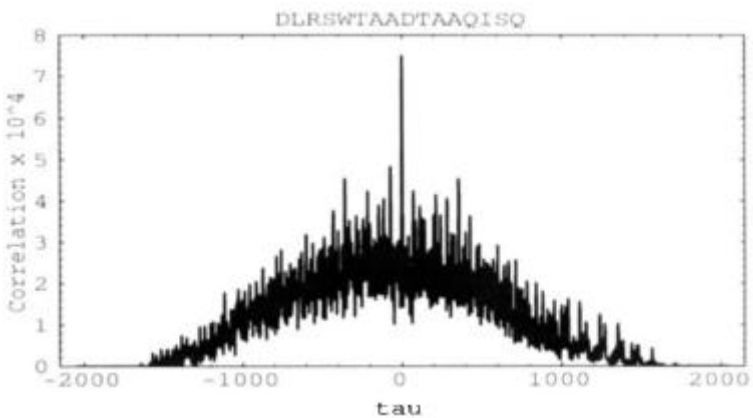
Sequest





Cross-correlation

$$R_\tau = \sum_{i=0}^{n-1} x[i]y[i+\tau]$$



X! Tandem - Search Parameters

http://www.thegpm.org/

simple page
view saved xml data

Lookup model:
GPM
go

what is the gpm
powered by tandem
send us email

Eukaryote proteomes
1 2 3 4 5 6 7

Boutique proteomes
human mouse frog
cow bacteria plant
fish rat

Algorithms
X! P3 X! Hunter

Information
gpmDB wiki
review lists

GPM Cyclone, advanced search form

1. spectra & taxon | 2. measurment errors | 3. signal processing
4. protein modifications | 5. refinement | 6. protein cleavage
Show all | Click to start search | FIND PROTEINS

1. spectra
common, mzXML, mzData, DTA, PKL or MGF only
Browse...

taxon
Select one or more.
☑ Eukaryotes ☑ Prokaryotes ☐ Viruses

none
H. sapiens, male
H. sapiens, female
M. musculus, male
M. musculus, female
R. norvegicus (rat)
S. cerevisiae (budding yeast)
--chordates--

none
Acaryochloris marina MBIC11017
Acetobacter pasteurianus IFO 3283 01
Acetohalobium arabaticum DSM 5501
Acholeplasma laidlawii PG 8A
Achromobacter xylosoxidans A8
Acidaminococcus fermentans DSM 20731
Acidilobus saccharovorans 345 15

1. Include reversed sequences: | ◉ none | ○ mixed | ○ only |
2. all ^{15}N amino acids ☐

Find proteins with peptide log(e) < -1 with protein log(e) < -1

gpmdb
1. Add to gpmDB: ◉ yes ○ restricted ○ no
2. Archive MS/MS information ◉ yes ○ no
3. Anonymous contribution: ◉ yes ○ no

⊞ more ...

X! Tandem - Search Parameters

2. measurement errors

1. Fragment mass error: 0.4 Da
2. Parent mass error: + 100 - 100 ppm
3. Isotope error: (•) yes ( ) no
4. Fragment type: (•) monoisotopic ( ) average

3. signal processing

1. Remove redundant: ( ) yes (•) no, angle: 40 (0-90)
2. Maximum parent charge: 4
3. Spectrum synthesis: (•) yes ( ) no
4. Noise suppression: ( ) yes (•) no
5. Minimum parent M+H: 500.0
6. Minimum fragment m/z: 150.0
7. Total peaks: 50
8. Minimum peaks: 15
9. Fragment types: [ ] a [x] b [ ] c [ ] x [x] y [ ] z

4. protein modifications

1. Complete modifications (unimod)

Set 1: Carbamidomethyl (C)
57.021464@C
specify your own

Set 2: No further mods
more sets ...

2. Potential modifications (unimod)

none
Oxidation (M)
Oxidation (W)
Deamidation (N)

specify your own 15.994915@M

3. Potential motif:
4. Protein N-terminus: 0.0 Da
5. Protein C-terminus: 0.0 Da
6. Use sequence annotations ( ) yes (•) no

X! Tandem - Search Parameters

5. refinement specification

1. Refine model: (•) yes ( ) no
2. Point mutations: ( ) yes (•) no
3. Use sequence annotations (•) yes ( ) no
4. Semi-style cleavage: ( ) yes (•) no
5. Potential modifications (unimod):

round 1

none
Oxidation (M)
Dioxidation (M)
Oxidation (W)

mods: 15.994915@M,15.994915@W,
motifs:

round 2

none
Oxidation (M)
Dioxidation (M)
Oxidation (W)

mods: 31.98983@M,31.98983@W
motifs:

round 3

none
Oxidation (M)
Dioxidation (M)
Oxidation (W)

mods:
motifs:

round 4

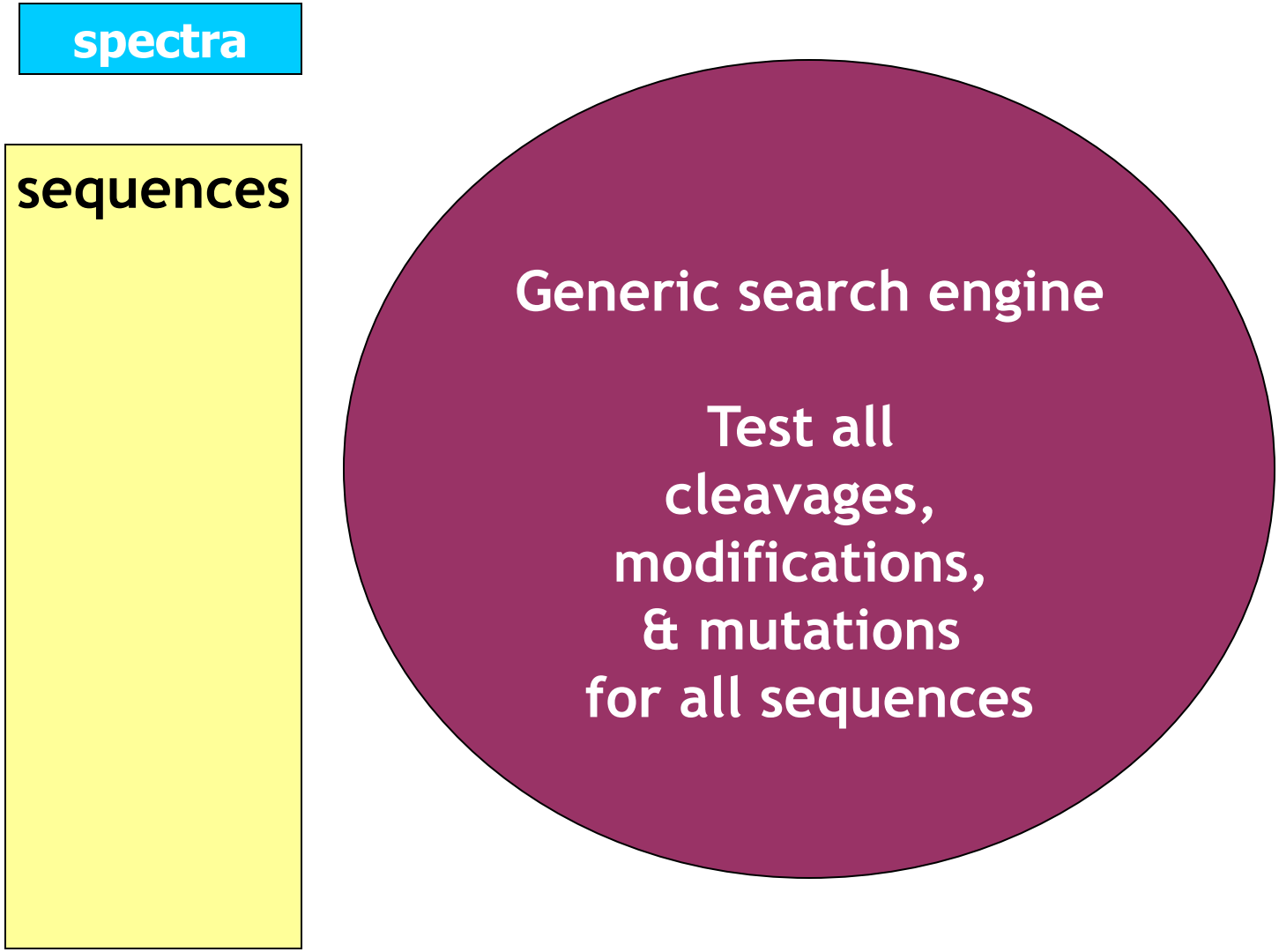
none
Oxidation (M)
Dioxidation (M)
Oxidation (W)

mods:
motifs:

6. Use these modifications throughout: ( ) yes (•) no
7. Unantipated cleaves ([X]|[X]): (•) yes ( ) no
8. Potential N-terminus modifications:
9. Potential C-terminus modifications:
10. Valid expectation: < -2

6. protein cleavage specification

1. Cleavage site:
   trypsin, [RK]|{P}
   specify your own
2. Semi-style cleavage: ( ) yes (•) no
3. Missed cleavage sites allowed: 1
4. Cleavage C-terminal change: +17.002735 Da
5. Cleavage N-terminal change: +1.007825 Da



Conventional,
single stage searching

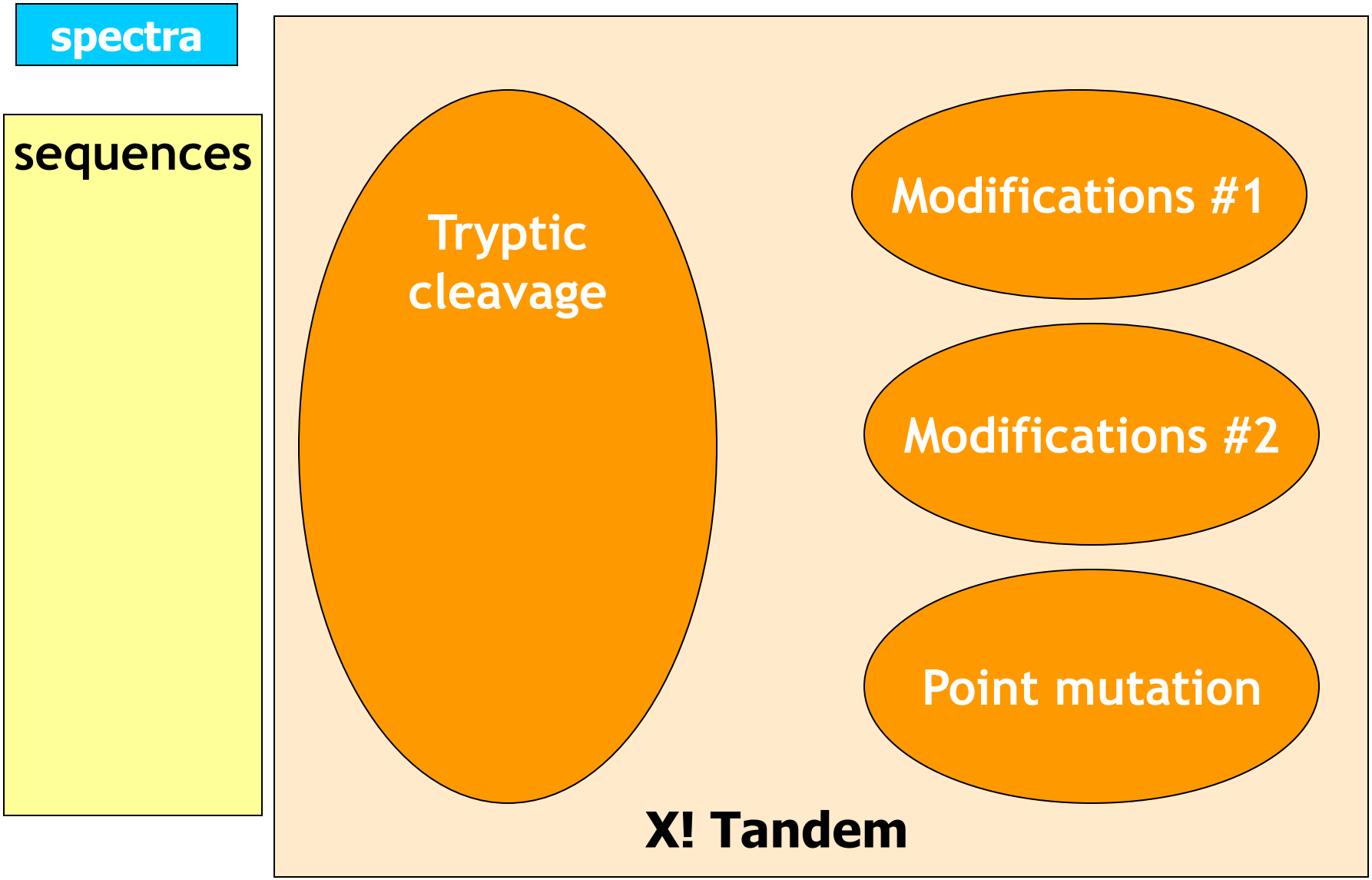
Some hard problems in MS/MS analysis in proteomics

Allowing for unanticipated peptide cleavages
- e.g., chymotryptic contamination in trypsin
- calculation order ~ 200 × tryptic cleavage
- "unfortunate" coefficient

Determining potential modifications
- e.g., oxidation, phosphorylation, deamidation
- calculation order 2^n
- NP complete

Detecting point mutations
- e.g., sequence homology
- calculation order 18^N
- NP complete

Multi-stage searching



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 hgnc go

#	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]

Search Results

ALB: albumin
log(e) = -2281.6 [Source: HGNC 399]
IPR001703 Alpha-fetoprotein
IPR000264 Serum albumin
IPR020858 Serum albumin-like
IPR020857 Serum albumin CS
IPR014760 Serum albumin N
IPR021177 Serum albumin subgroup

```
  1 mkwvtfisllflfssaysrgvfrrdahksevahrfkdlgeenfkalvliafaqylqqcpf 60
    MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPF
 61 edhvklvnevtefaktcvadesaencdkslhtlfgdklctvatlretygemadccakqep 120
    EDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEP
121 ernecflqhkddnpnlprlvrpevdvmctafhdneetflkkylyeiarrhpyfyapellf 180
    ERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLF
181 fakrykaafteccqaadkaacllpkldelrdegkassakqrlkcaslqkfgerafkawav 240
    FAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAV
241 arlsqrfpkaefaevsklvtdltkvhtecchgdllecaddradlakyicenqdsisskik 300
    ARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLK
301 eccekplleksh ciaevendempadlpslaadfveskdvcknyaeakdvflgmflyeyar 360
    ECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYAR
361 rhpdysvvlllrlaktyettlekccaaadphecyakvfdefkplveepqnlikqncelfe 420
    RHPDYSVVLLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFE
421 qlgeykfqnallvrytkkvpqvstptlvevsrnlgkvgskcckhpeakrmpcaedylsvv 480
    QLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVV
481 lnqlcvlhektpvsdrvtkcctesl vnrrpcfsalevdetyvpkefnaetftfhadictl 540
    LNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTL
541 sekerqikkqtalvelvkhkpkatkeqlkavmddfaafvekcckaddketcfaeegkklv 600
    SEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLV
601 aasqaalgl                                                    609
    AASQAALGL
```

Sequence Annotations

show legend

mvdqp	lower case sequence is the latest sequence from ENSEMBL for this accession number
reklqee	lower case transition from black to blue letters indicates an exon boundary; a red residue indicates a triplet shared between exons
MVDQP	upper case sequence is the protein sequence originally analyzed
dvdnas	synonymous SNP with no residue change and non-synonymous SNP which changes the residue
DIMR	residues part of at least one observed peptide domain
LREEQ	residues predicted to be difficult to observe by standard techniques
HFQL	residue found is a single amino-acid polymorphism
AYNG	residue found is chemically modified

Complete mods:
i. Carbamidomethyl@C, Carbamidomethyl@U

Potential mods:
i. Oxidation@M, Label:+6 Da@K, Label:+6 Da@R
ii. Oxidation@M, Oxidation@W, Deamidated@N, Deamidated@Q
iii. Dioxidation@M, Dioxidation@W

Protein-specific PTMs:
i. Phospho@S, Phospho@T, Phospho@Y

N-terminal:
i. Ammonia-loss@Q, Ammonia-loss@C, Dehydrated@E (peptide)
ii. ragged, Acetyl (protein)

Search Results

Identified Peptides

spectrum	log(e)	log(I)	m+h	delta	ζ	sequence	n
14014.1	-7.4	3.34	1149.5759	-0.0007	2/5	vfrr^{25} DAHKSEVAHR ^{34}fkdl	(5097)
16362.1	-2.1	3.82	1006.5177	0.0018	2/5	rrda^{27} HKSEVAHR ^{34}fkdl	(206)
6222.1	-5.4	4.10	1226.6052	0.0025	2/3	vahr^{35} FKDLGEENFK ^{44}alvl	(55404)
3243.1	-2.8	5.80	1226.6052	0.0024	3/3	vahr^{35} FKDLGEENFK ^{44}alvl	(55404)
18750.1	-8.6	3.73	2533.2908	-0.0002	2/3	enfk^{45} ALVLIAFAQY LQQCPFEDHV K ^{65}lvne	(84854)
						fk^{45} ALVLIAFAQY LQQCPFEDHV K ^{65}lvne	(84854)
						al^{47} VLIAFAQYLQ QCPFEDHVK ^{65}lvne	(1004)
						lv^{48} LIAFAQYLQQ CPFEDHVK ^{65}lvne	(1537)
						vl^{49} IAFAQYLQQC PFEDHVK ^{65}lvne	(2586)
						vli^{50} AFAQYLQQCP FEDHVK ^{65}lvne	(1886)
						ia^{51} FAQYLQQCPF EDHVK ^{65}lvne	(1377)
						ia^{51} FAQYLQQCPF EDHVK ^{65}lvne	(1377)
						af^{52} AQYLQQCPFE DHVK ^{65}lvne	(3958)
						af^{52} AQYLQQCPFE DH ^{63}vklv	(30)
						fa^{53} QYLQQCPFED HVK ^{65}lvne	(777)
						aq^{54} YLQQCPFEDH VK ^{65}lvne	(1701)
						aq^{54} YLQQCPFEDH VK ^{65}lvne	(1701)
						aq^{54} YLQQCPFEDH ^{63}vklv	(24)
						qy^{55} LQQCPFEDHV K ^{65}lvne	(1287)

Column notes.

1. **spectrum**: written in the form "X.Y", where X is a unique identifier for a particular tandem mass spectrum in this data set and Y is an identifier for this particular sequence solution.
2. **log(e)**: the base-10 log of the expectation that any particular peptide assignment was made at random (*E*-value).
3. **log(I)**: the base-10 log of the sum of the fragment ion intensities in the tandem mass spectrum used to make this assignment.
4. **m+h**: the calculated mass of the protonated parent ion for this sequence assignment.
5. **delta**: the difference between the measured and calculated protonated parent ion masses.
6. **ζ**: the ratio of the measured charge of the parent ion to the number of basic sites in the assigned peptide sequnce.
7. **sequence**: the sequence of the assigned peptide sequence. The sequences immediately N-terminal and C-terminal to the assigned peptide in the protein sequence are also shown.
8. **n**: the number of observations of this peptide sequence in GPMDB.
9. **ω**: the frequency of observation for this peptide in this protein (only available for some species).

Display modes:

1. **best**: the peptide assignment with the best expectation value for a particular sequence and parent ion charge is shown.
2. **all**: all peptide assignments are shown.
3. **modified**: all peptide assignments that have at least one modified residue are shown.
4. **homologues**: all peptides assignments unique to this protein sequence are shown.

Search Results

GPM33080001549: peptide model: 6227.1.1 of ENSP00000295897

| model | protein | homologues | XML | gpmDB | wiki | Peptide Atlas | SwedCAD |

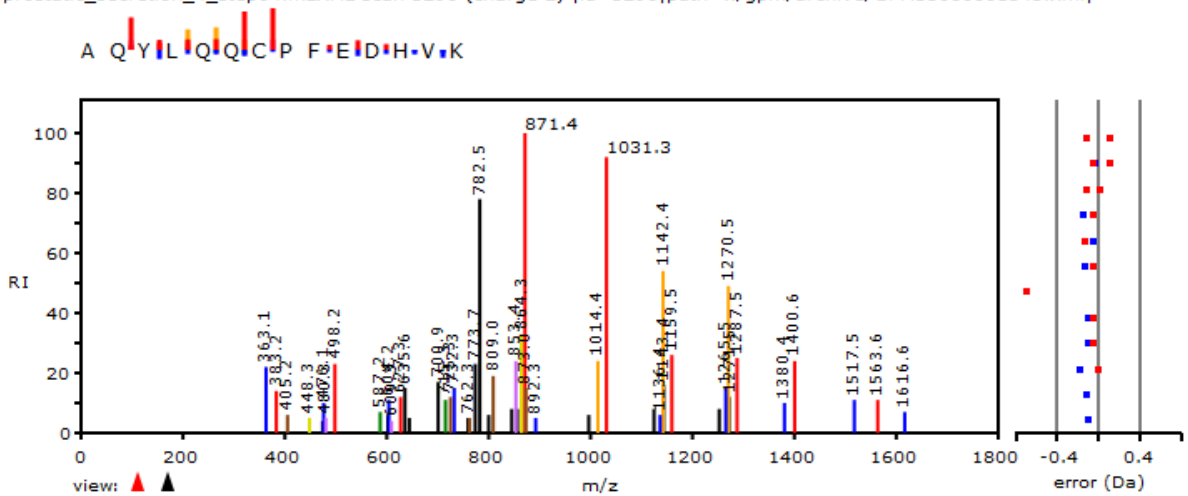
ENSP00000295897: albumin [Source: HGNC 399]

Sample information

#	log(e)	log(I)	m+h	delta	ζ	sequence | validate | studio | mgf | mrm | details |
6227	-9.0	5.06	1762.8216	0.0035	2/3	^{52}AQYLQQCPFEDHVK^{65} (3958) 0.0012

mods: ^{58}C+57.0215

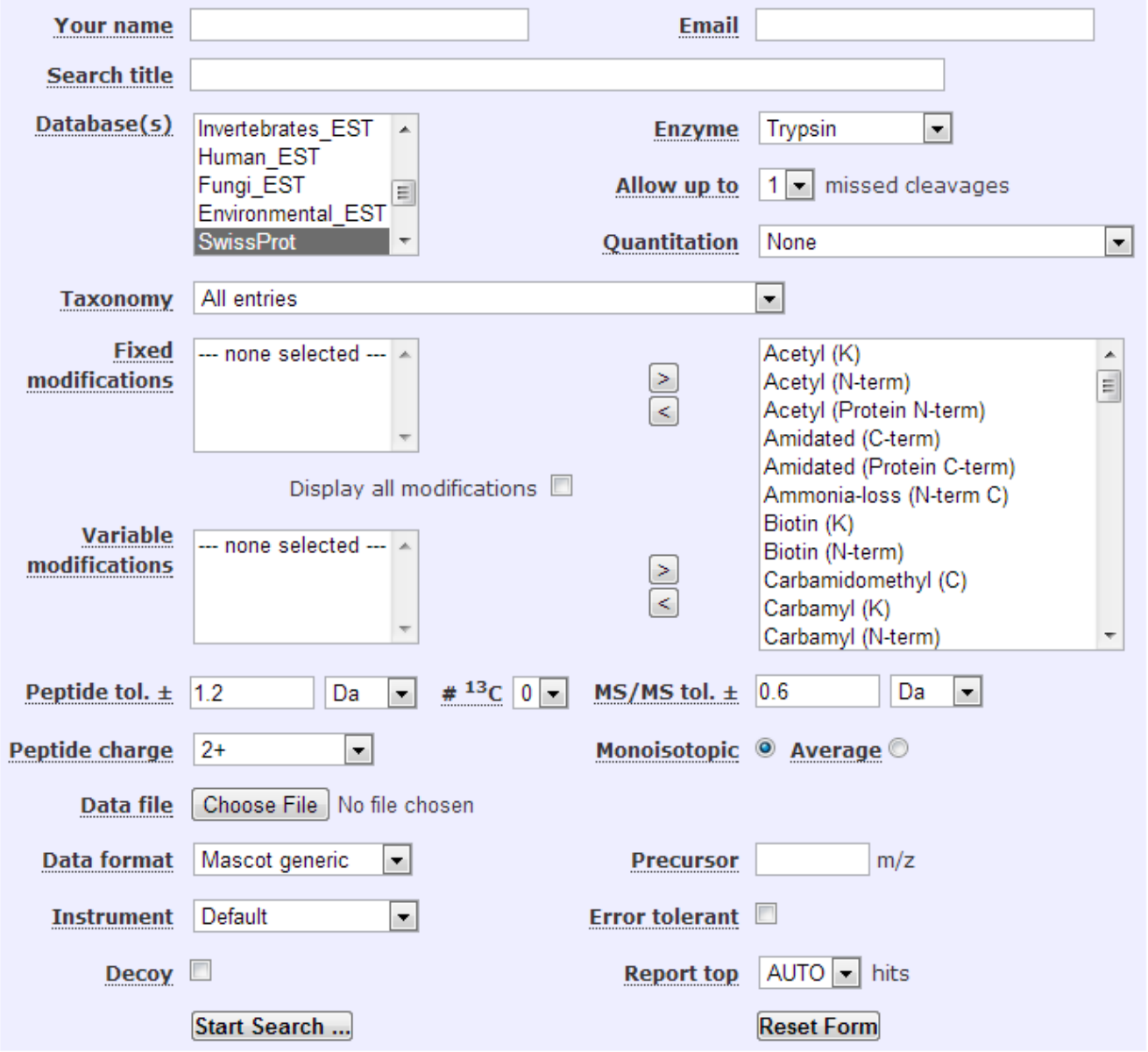
prostatic_secretion_4_step04.mzXML scan 5296 (charge 2) |id=5296|path=../gpm/archive/GPM33080001543.xml|



matched/total: # ions: 68% intensity: 75% σ: 0.17 Da

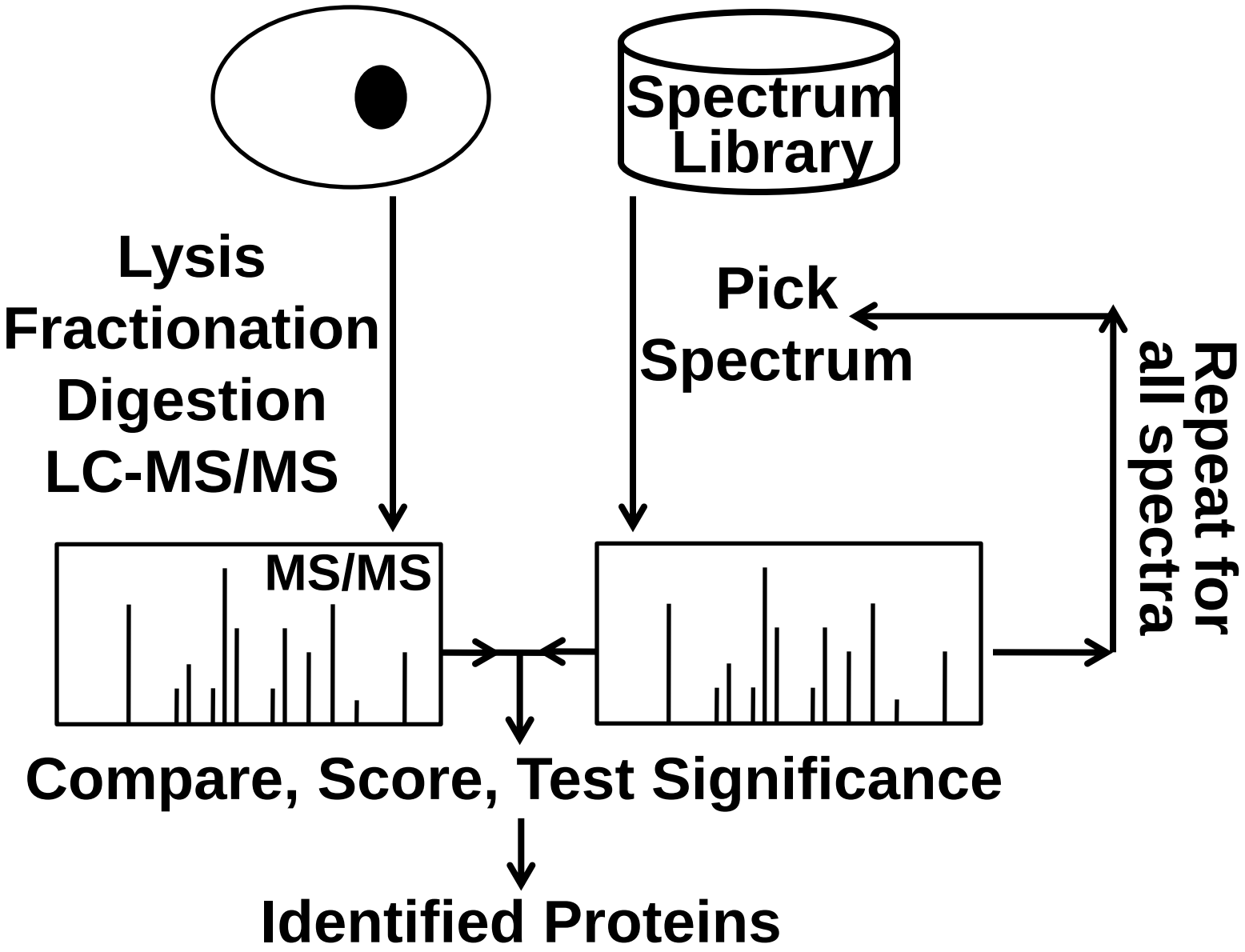
bond	^{+1}y	$^{+1}y^{-17}$	$^{+1}y^{-18}$	^{+1}b	$^{+1}b^{-17}$	$^{+1}b^{-18}$
A_1	1691.785	1674.758	1673.774	72.044	55.018	54.034
Q_2	$^{+1,2}$1563.726	1546.700	1545.715	200.103	183.076	182.092
Y_3	$^{+1,2}$1400.663	1383.636	1382.652	363.166	346.140	345.156
L_4	$^{+1,2}$1287.579	1270.552	1269.568	476.250	459.224	458.240
Q_5	1159.520	1142.494	1141.510	604.309	587.282	586.298
Q_6	1031.462	1014.435	1013.451	732.368	715.341	714.357
C_7	871.431	854.404	853.420	892.398	875.372	874.388
P_8	774.378	757.352	756.367	989.451	972.424	971.440
F_9	627.310	610.283	609.299	1136.519	1119.493	1118.509
E_{10}	498.267	481.241	480.256	1265.562	1248.535	1247.551
D_{11}	383.240	366.214	365.230	1380.589	1363.562	1362.578
H_{12}	246.181	229.155	228.171	$^{+1,2}$1517.648	1500.621	1499.637
V_{13}	147.113	130.086	129.102	$^{+1,2}$1616.716	1599.690	1598.706

Mascot



http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS

Identification – Spectrum Library Search



Steps in making an Annotated Spectrum Library (ASL):

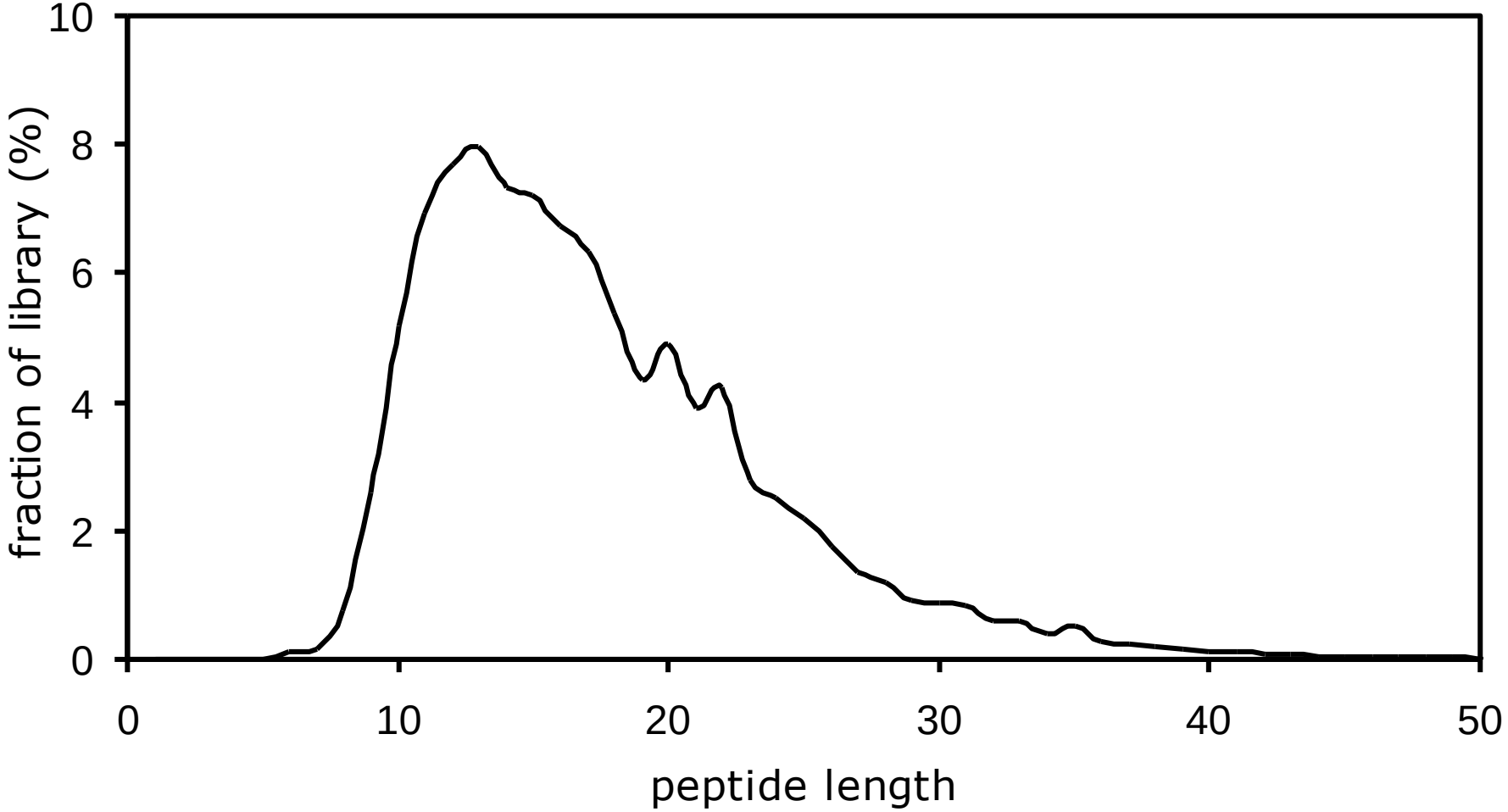
1. Find the best 10 spectra for a particular sequence, with the same PTMs and charge.

2. Add the spectra together and normalize the intensity values.

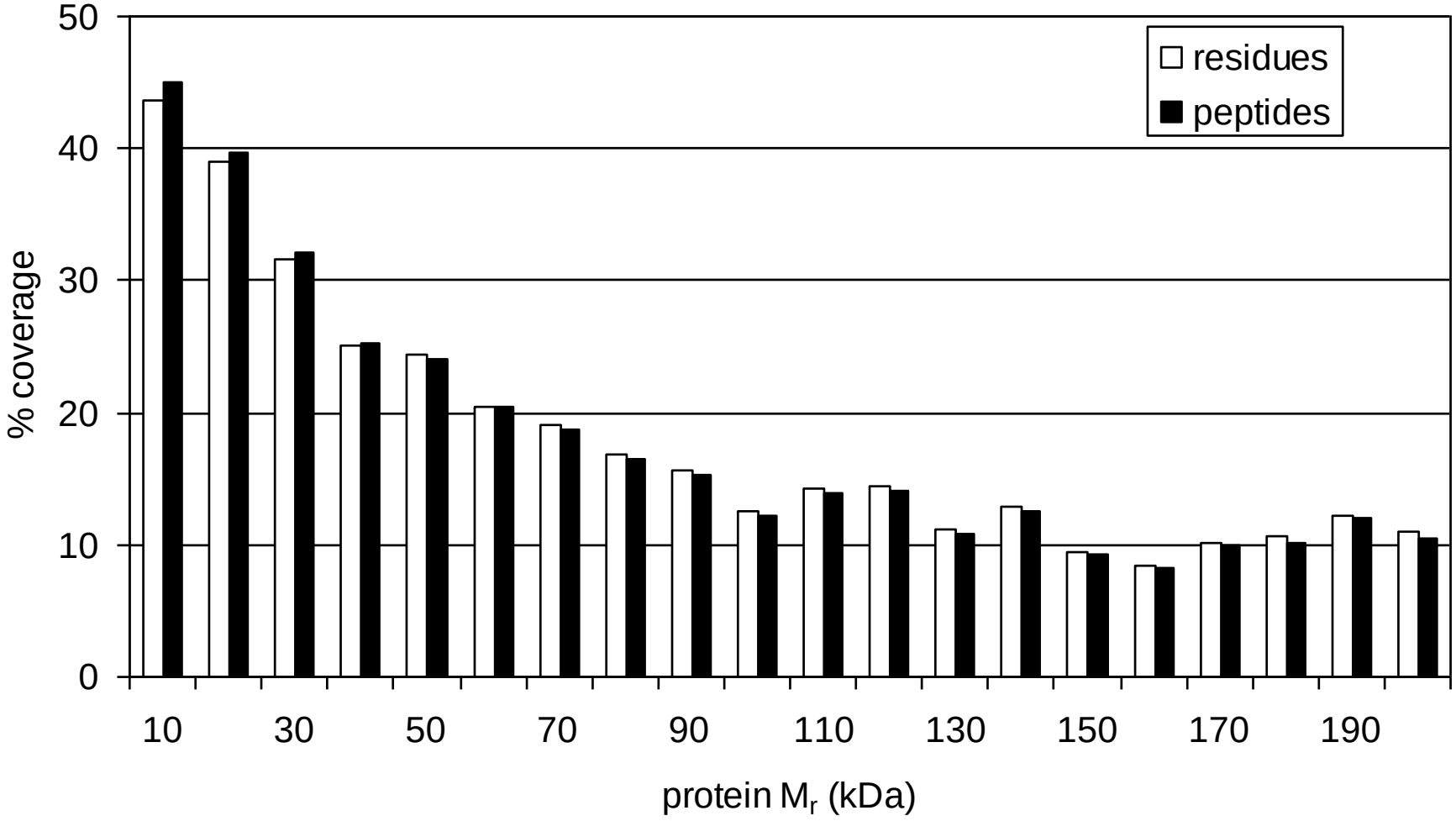
3. Assign a "quality" value: the median expectation value of the 10 spectra used.

4. Record the 20 most intense peaks in the averaged spectrum, it's parent ion *z*, *m*/*z*, sequence, protein accessions & quality.

Spectrum Library Characteristics – Peptide Length

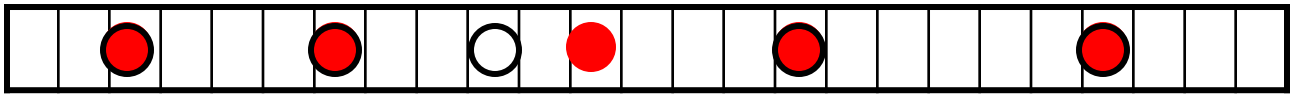


Spectrum Library Characteristics – Protein Coverage

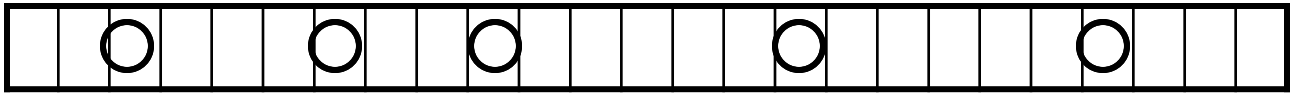


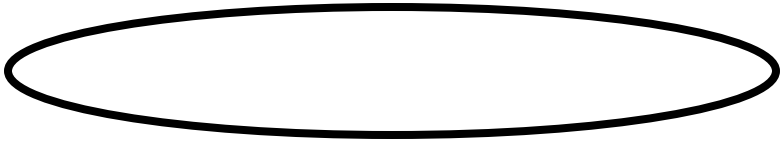
Identification – Spectrum Library Search

Library spectrum (5:25)



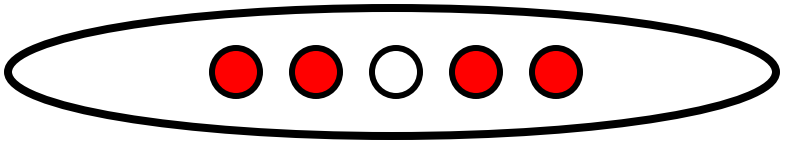
Test spectrum (5:25)





Results: 4 peaks selected, 1 peak missed

Identification – Spectrum Library Search



How likely is this?

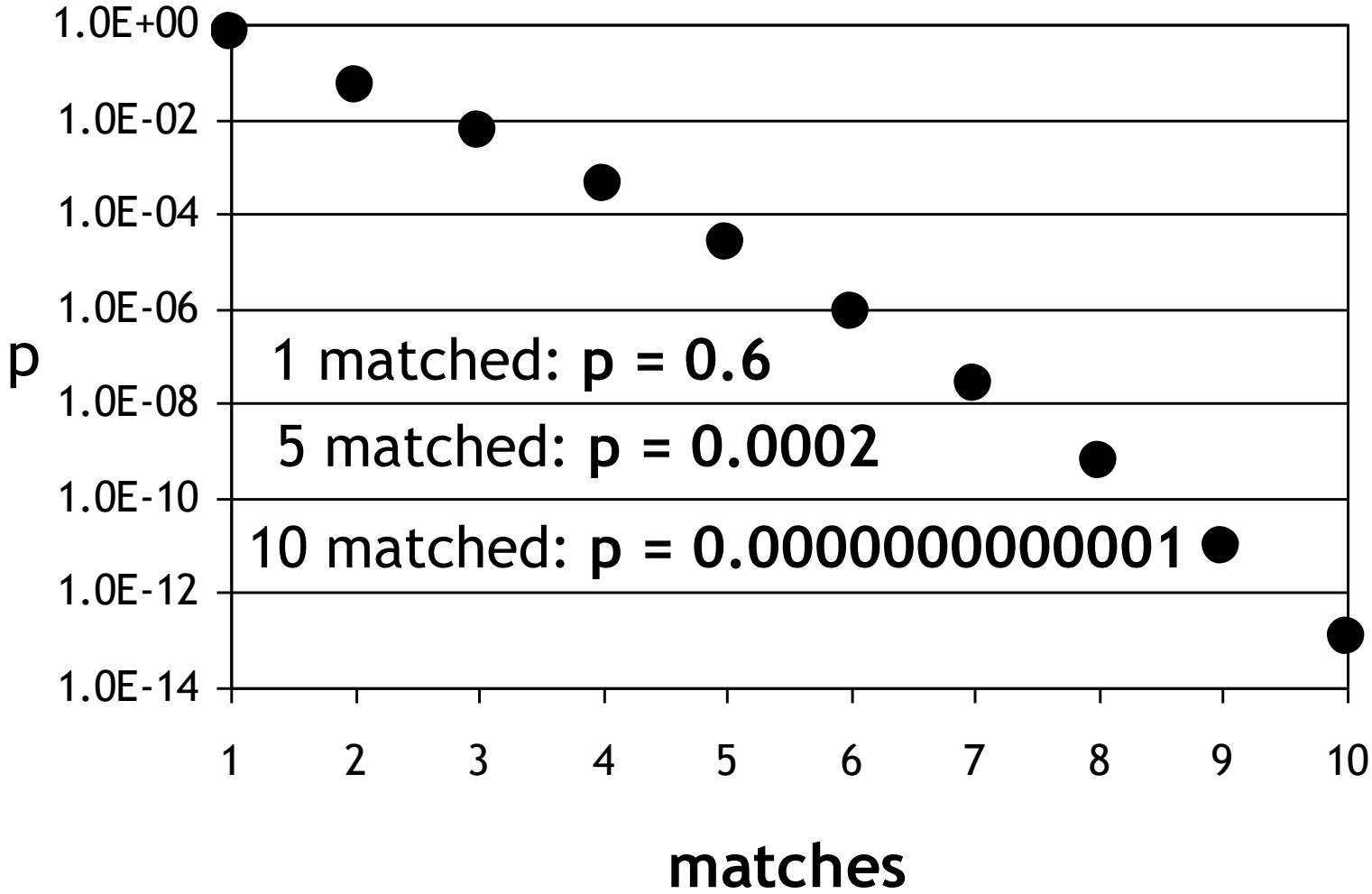
Apply a hypergeometric probability model:

- 25 possible m/z values;
- 5 peaks in the library spectrum; and
- 4 selected by the test spectrum.

Matches	Probability
1	0.45
2	0.15
3	0.016
4	0.00039
5	0.0000037

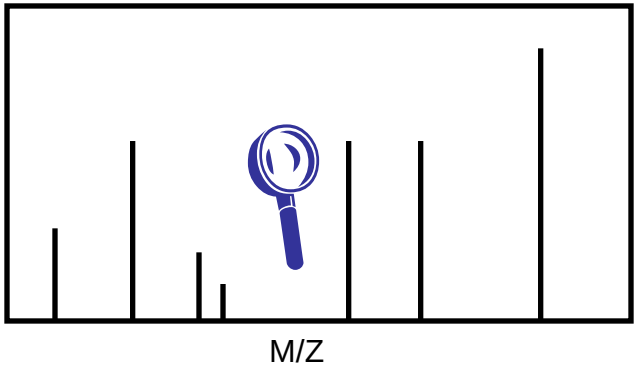
Identification – Spectrum Library Search

If you have 1000 possible m/z values and 20 peaks in test and library spectrum?

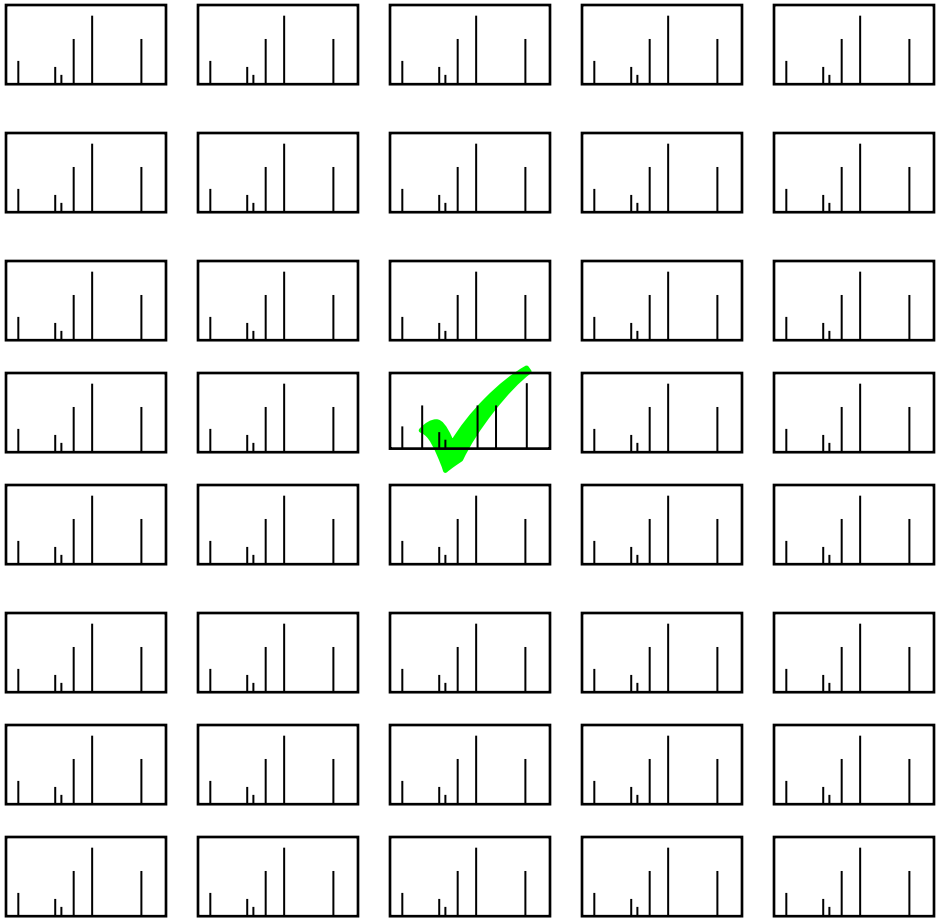


Identification – Spectrum Library Search

Experimental
Mass Spectrum



Library of Assigned
Mass Spectra



Best search result

X! Hunter

the gpm

1. X! Tandem 2013.02.01 successfully passes on-line tests
   The testing phase of the most recent X! Tandem release is complete.
2. Update of human sequences
   The human protein sequences used for the public GPM have been updated to ENSEMBL v.70 and dbSNP v.137

This site
saved xml data

Lookup GPM #
go

Information
about the GPM
about X! Hunter
send us email

More search sites

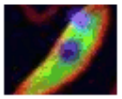
Eukaryote proteomes
1 2 3 4 5 6 7

Boutique proteomes
human mouse
cow bacteria
plant rat

Algorithms
X! P3 X! Hunter

Information
gpmDB wiki
review lists

Some species

GPM Cyclone, X! Hunter search form

X! Hunter is a search engine that compares experimentally observed spectra directly with consensus mass spectra obtained from the GPMDB. It can identify proteins for human, budding yeast, mouse and thale cress samples. Because the sequence modifications and cleavage sites for the peptides in the sequence library are already known, it is not necessary to specify as many parameters for this type of search as in more conventional search engines.

1. Spectra: Choose File No file chosen
2. Taxon:

Eukaryotes:

- H. sapiens, male
- H. sapiens, female
- H. sapiens: SILAC, male
- H. sapiens: SILAC, female
- M. musculus, male
- M. musculus, female
- M. musculus: SILAC, male
- M. musculus: SILAC, female

Prokaryotes:

- Bacillus anthracis A0248
- Bacillus anthracis Ames
- Bacillus anthracis Ames 0581
- Bacillus anthracis CDC 684
- Bacillus anthracis str Sterne
- Brucella abortus bv 1 9 941
- Brucella abortus S19
- Brucella melitensis

Viruses:

- Human immunodeficiency virus 1
- Influenza A virus_ A Puerto Rico 8 34 H1N1
- Monkeypox virus Zaire 96 I 16
- Respiratory syncytial virus

3. Parent mass error: + 100 - 100 ○ Da or ◉ ppm
4. Parent ion isotope error: ◉ yes ○ no
5. cos(θ) >: 0.3
6. Check all charges: ☐ yes
7. peptide log(e) < -1 and protein log(e) < -1
8. peptide sequences:
9. protein accessions:
10. Perform search: Find proteins

X! Hunter algorithm:

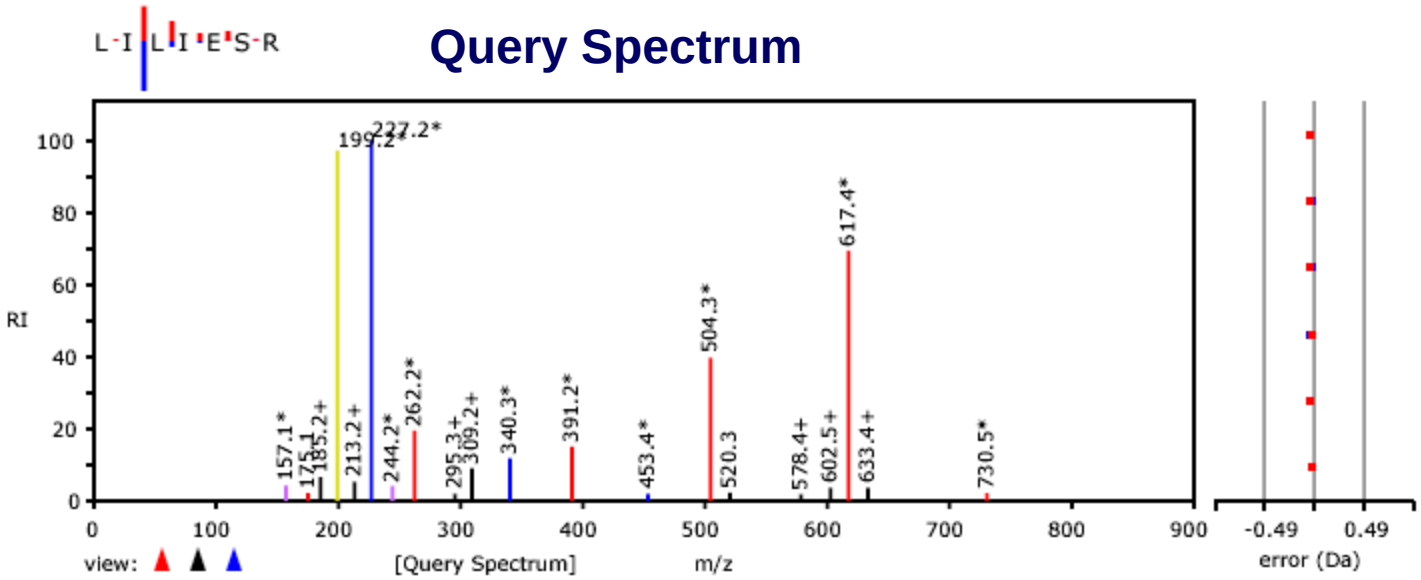
1. Use dot product to find a library spectrum that best matches a test spectrum.

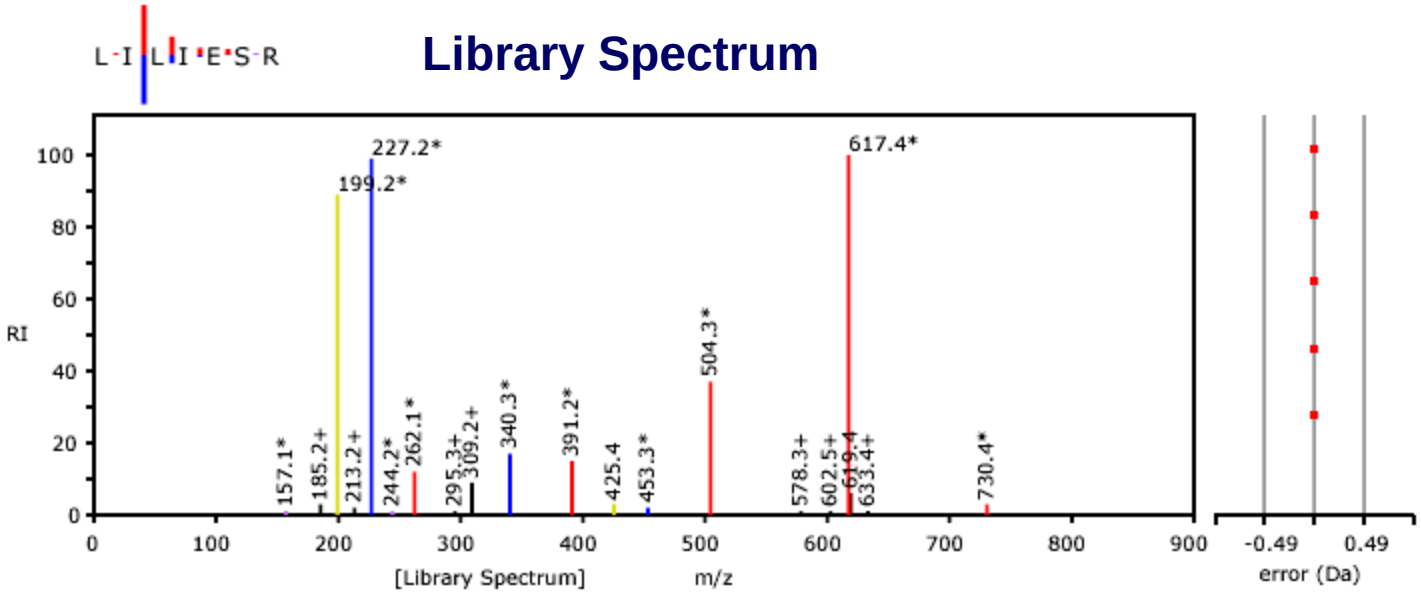
2. Calculate p-value with hypergeometric distribution.

3. Use p-value to calculate expectation value, given the identification parameters.

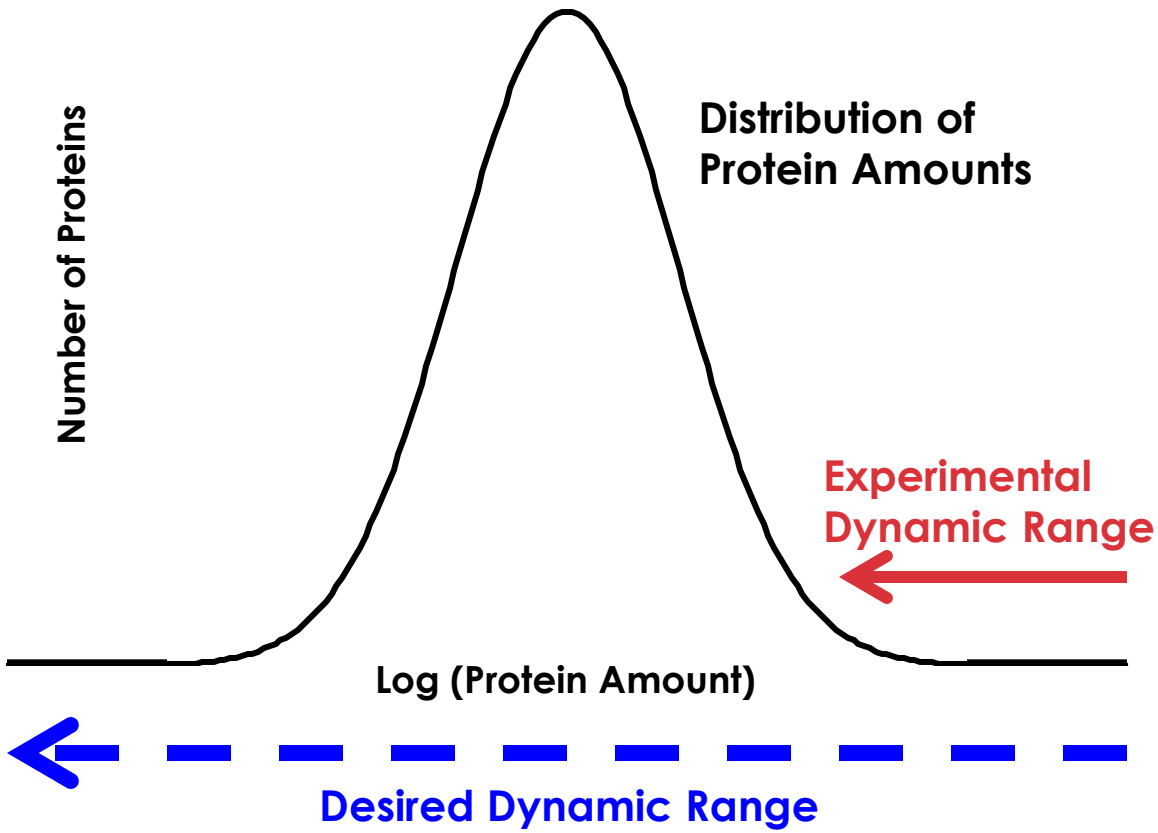
4. If expectation value is less than the median expectation value of the library spectrum, report the median value.

X! Hunter Result

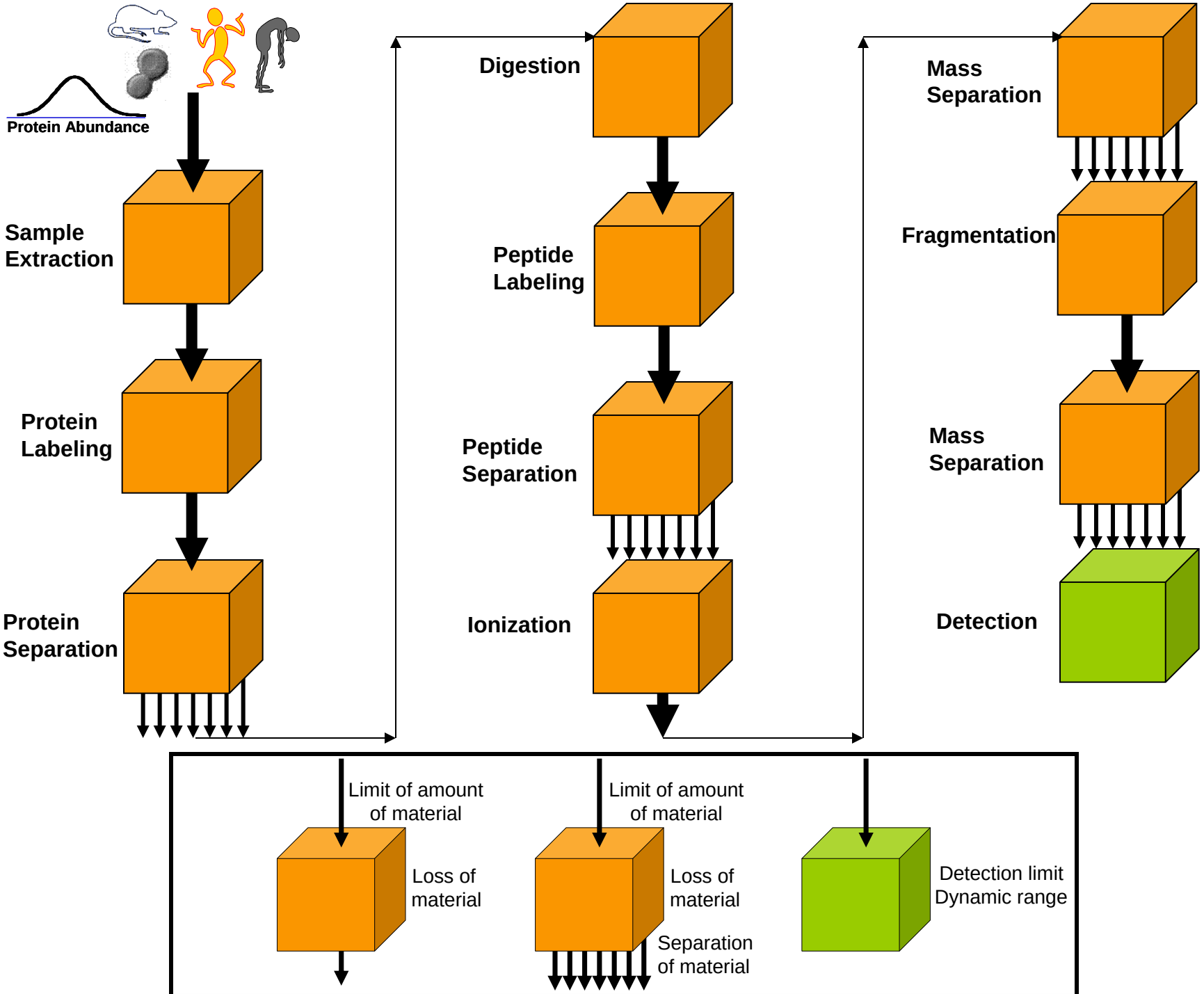




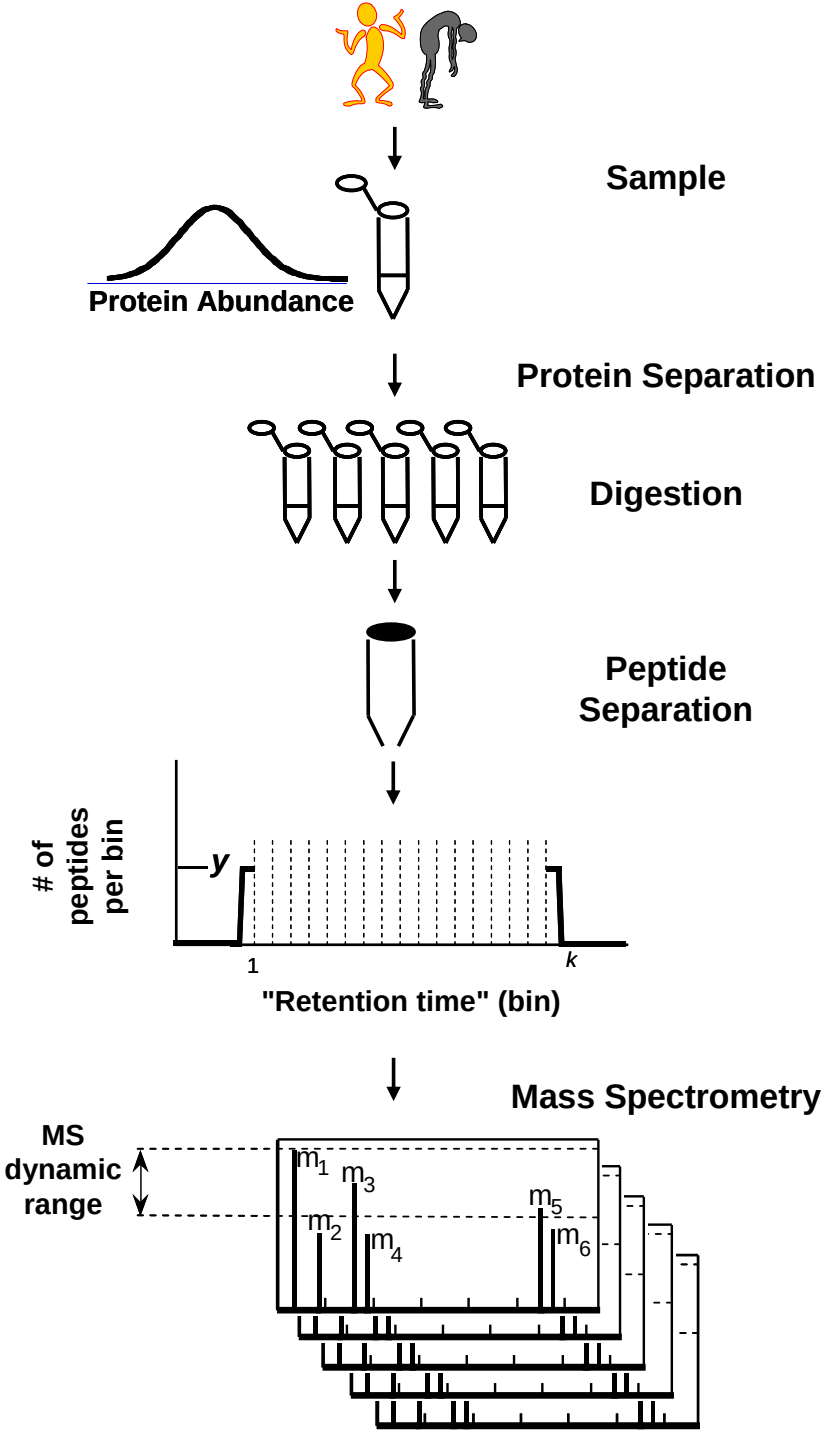
Dynamic Range In Proteomics



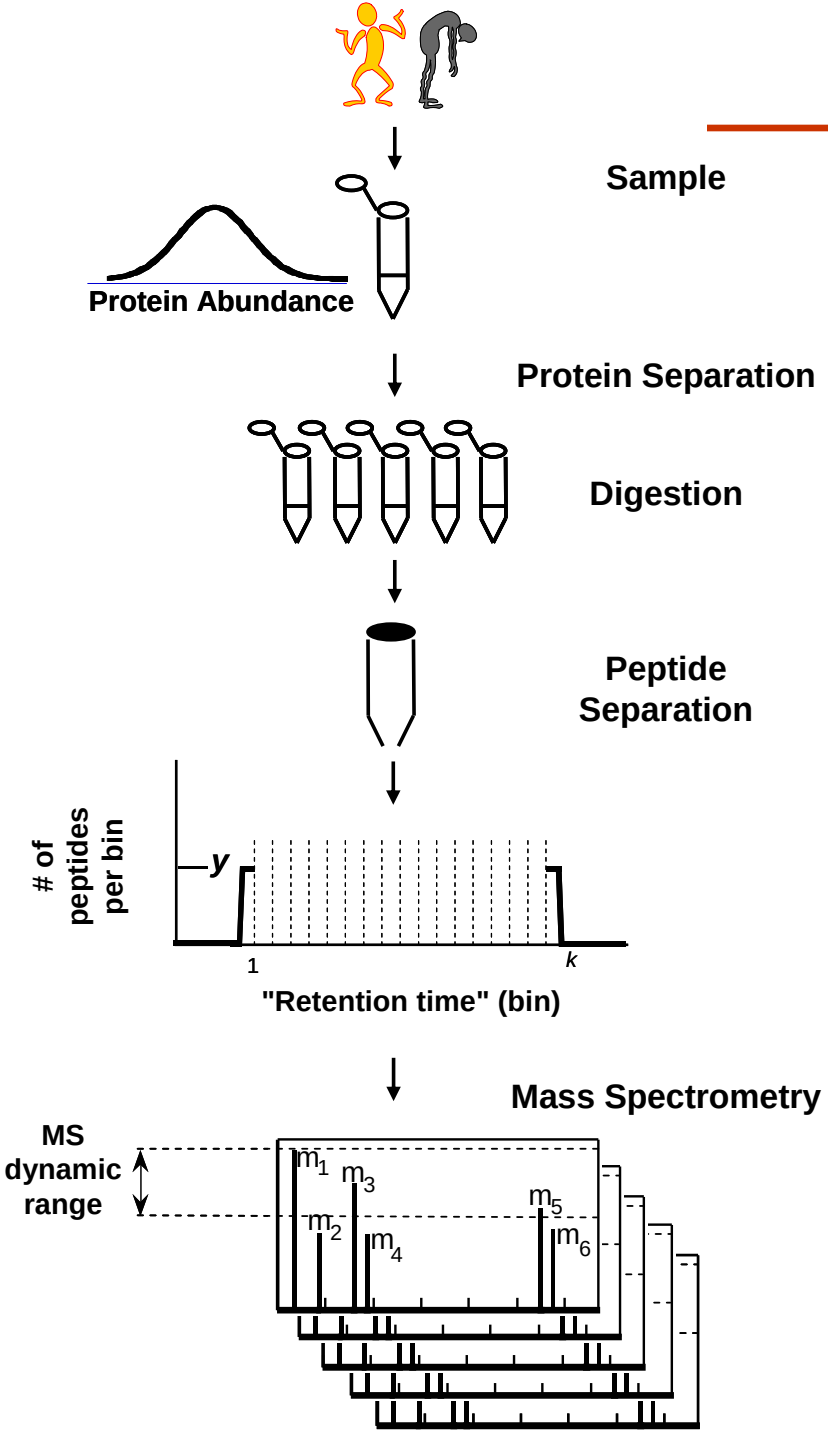
The goal is to identify and characterize all components of a proteome


Protein Abundance
Sample Extraction
Protein Labeling
Protein Separation
Digestion
Peptide Labeling
Peptide Separation
Ionization
Mass Separation
Fragmentation
Mass Separation
Detection
Limit of amount of material
Loss of material
Limit of amount of material
Loss of material
Separation of material
Detection limit
Dynamic range

Experimental Designs Simulated

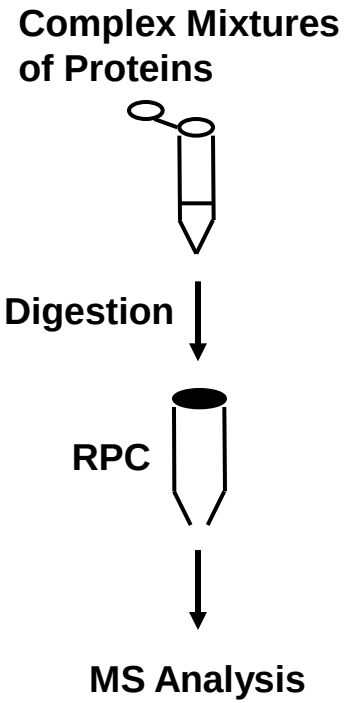


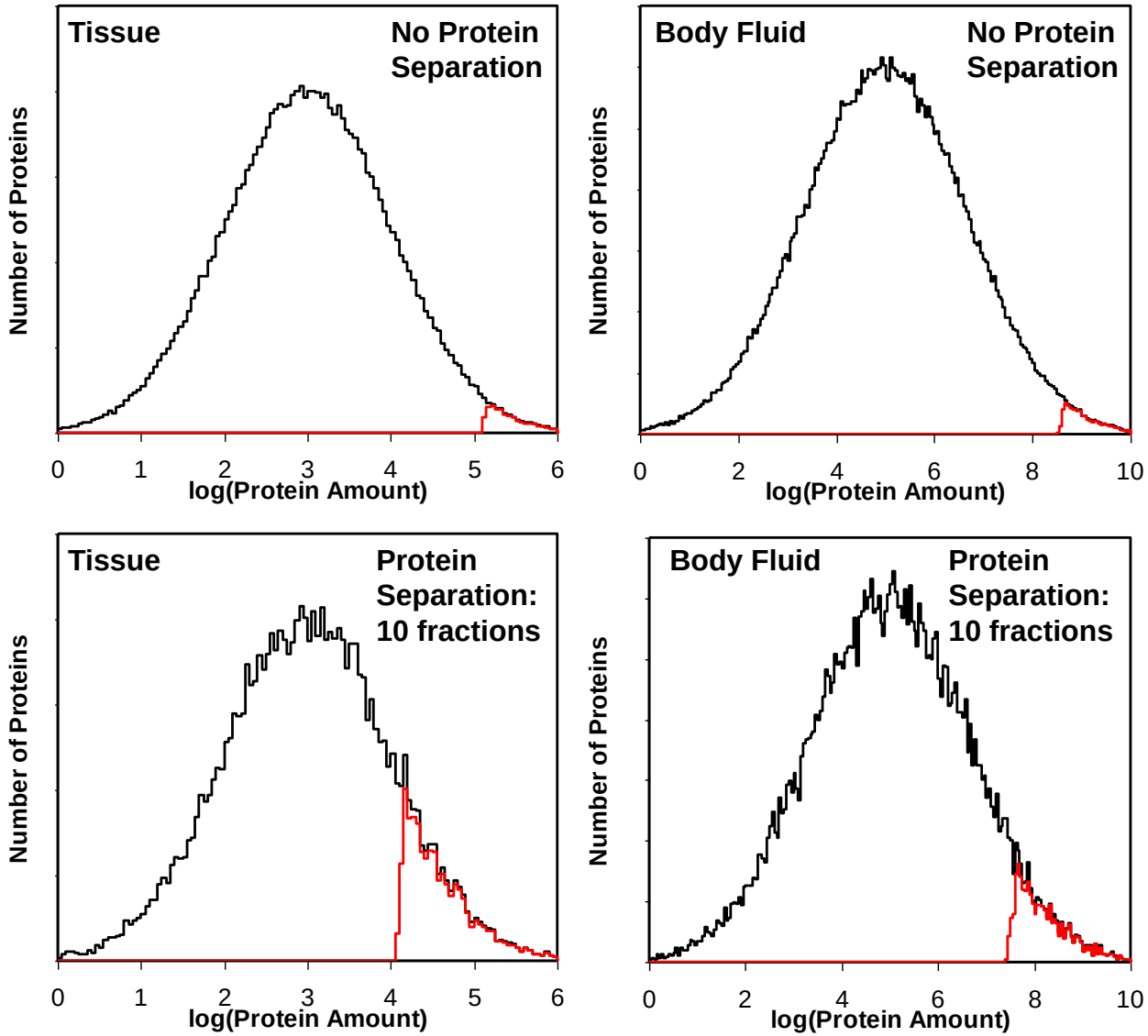
Parameters in Simulation



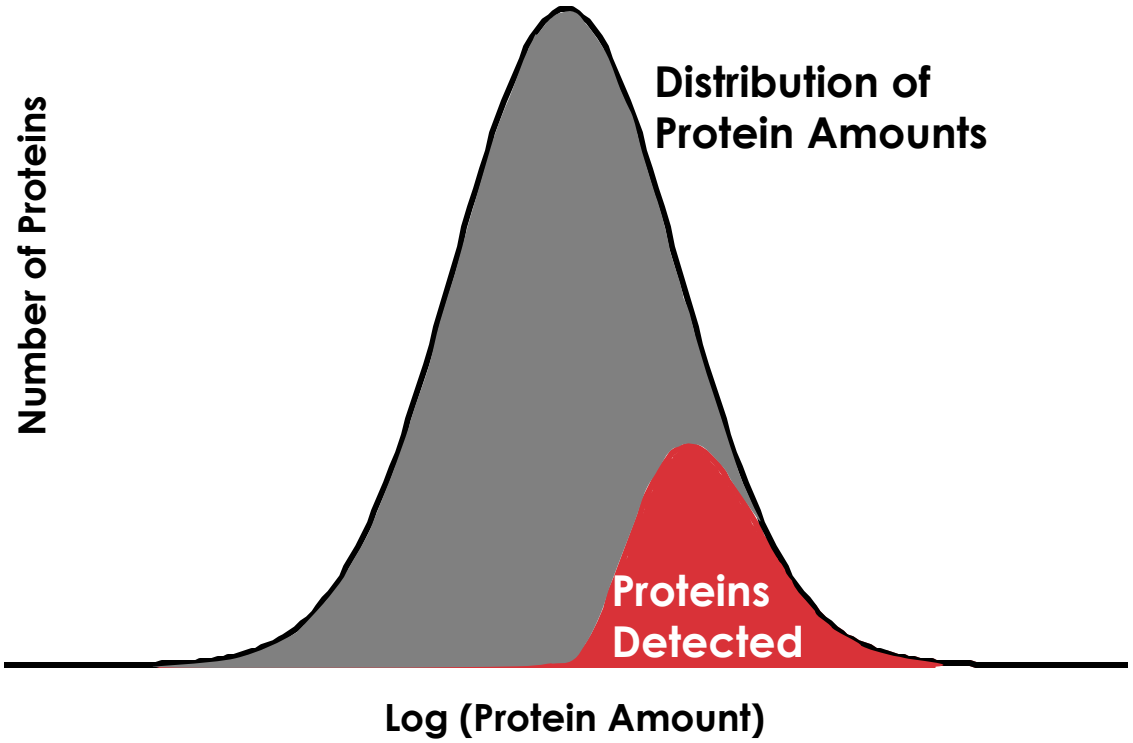
- Distribution of protein amounts in sample
- # of Proteins in each fraction
- Total amount of peptides that are loaded on column (limited by column loading capacity)
- Loss of peptides before binding to the column
- # of peptide fractions
- Loss of peptides after elution off the column
- Distribution of mass spectrometric response for different peptides present at the same amount
- Dynamic range of mass spectrometer
- Detection limit of mass spectrometer

Simulation Results for 1D-LC-MS



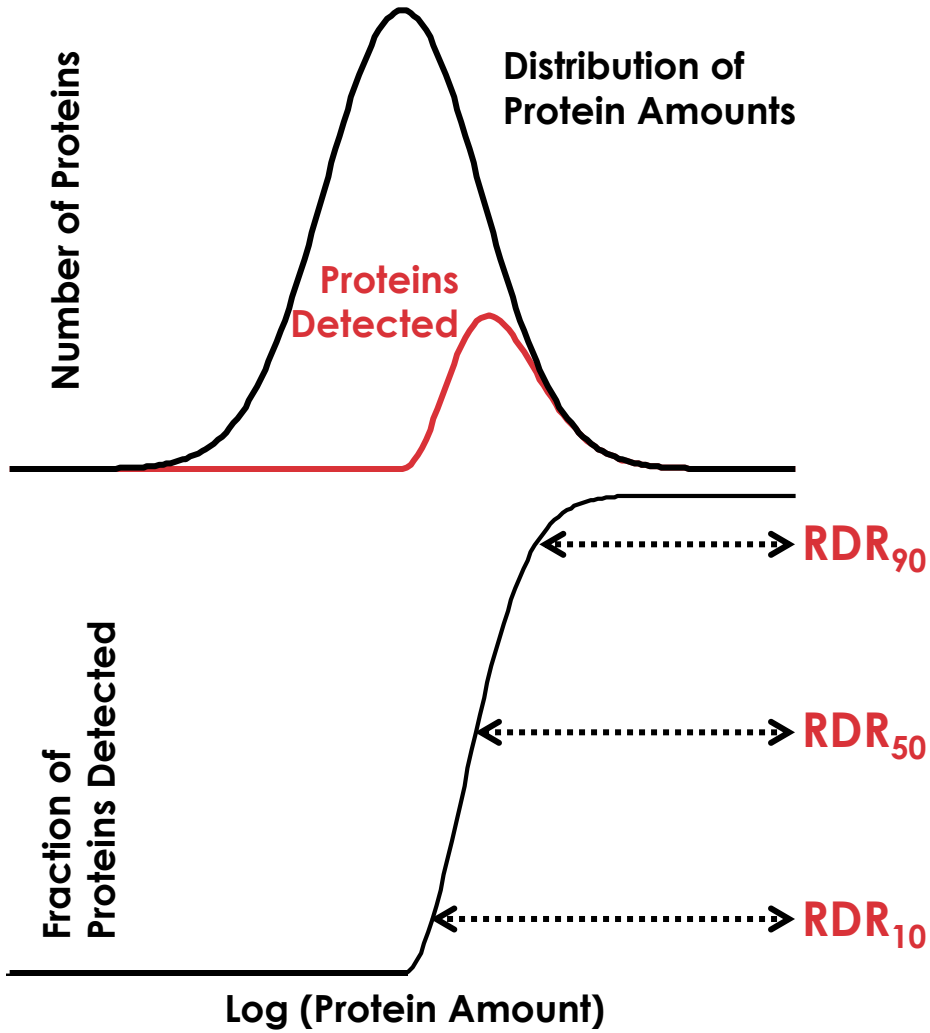


Success Rate of a Proteomics Experiment



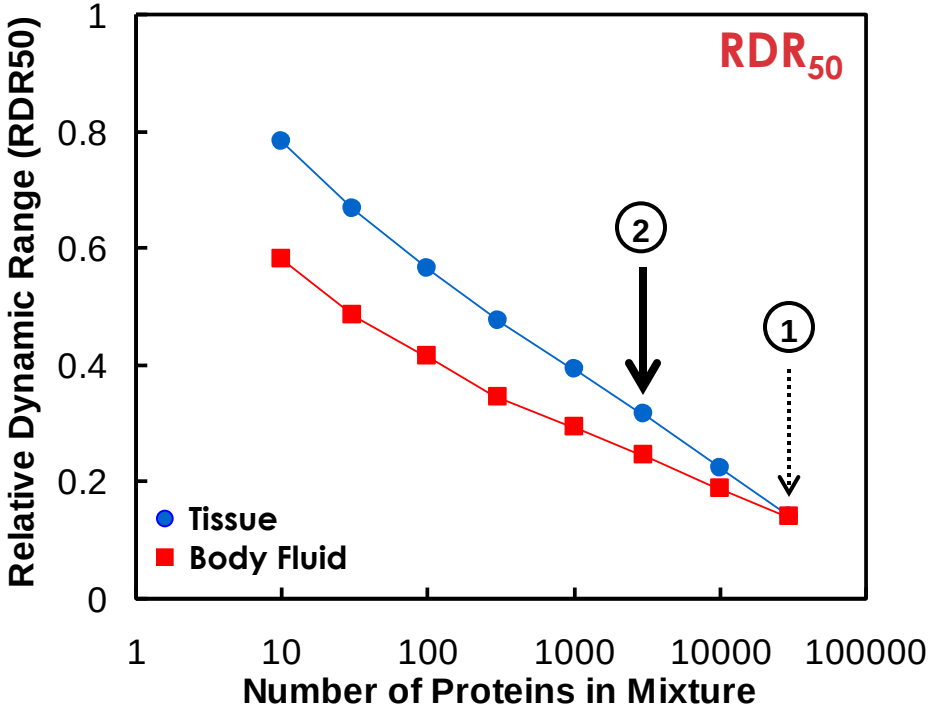
<u>*DEFINITION*</u>: The success rate of a proteomics experiment is defined as the number of proteins detected divided by the total number of proteins in the proteome.

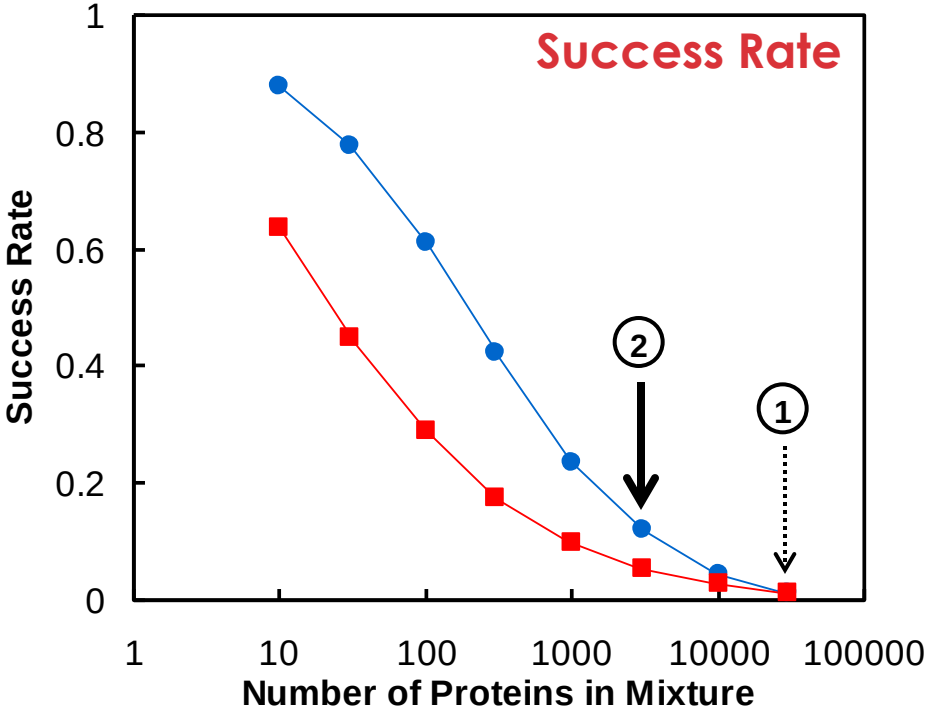
Relative Dynamic Range of a Proteomics Experiment

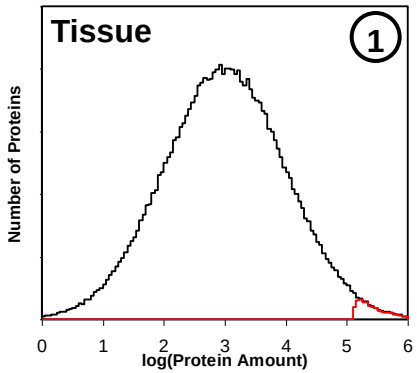


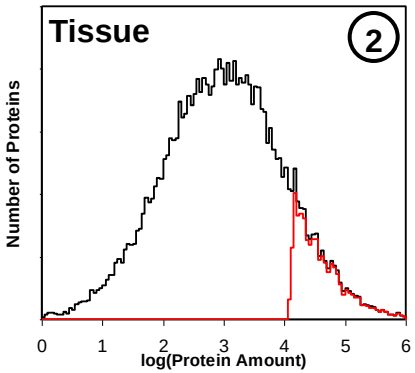
DEFINITION: **RELATIVE DYNAMIC RANGE**, **RDR_x**, where x is e.g. 10%, 50%, or 90%

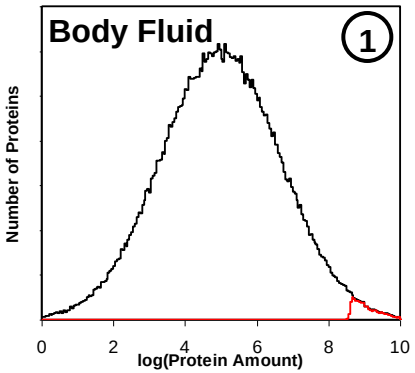
Number of Proteins in Mixture

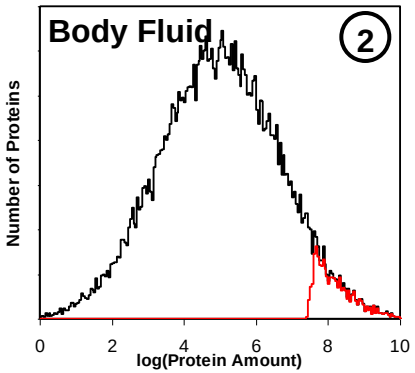




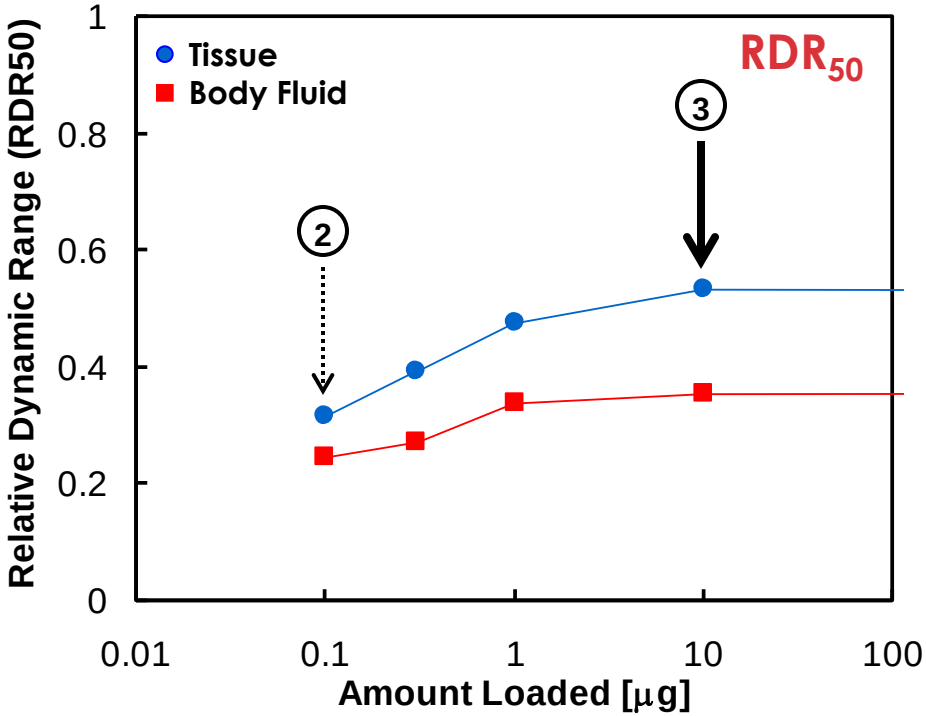


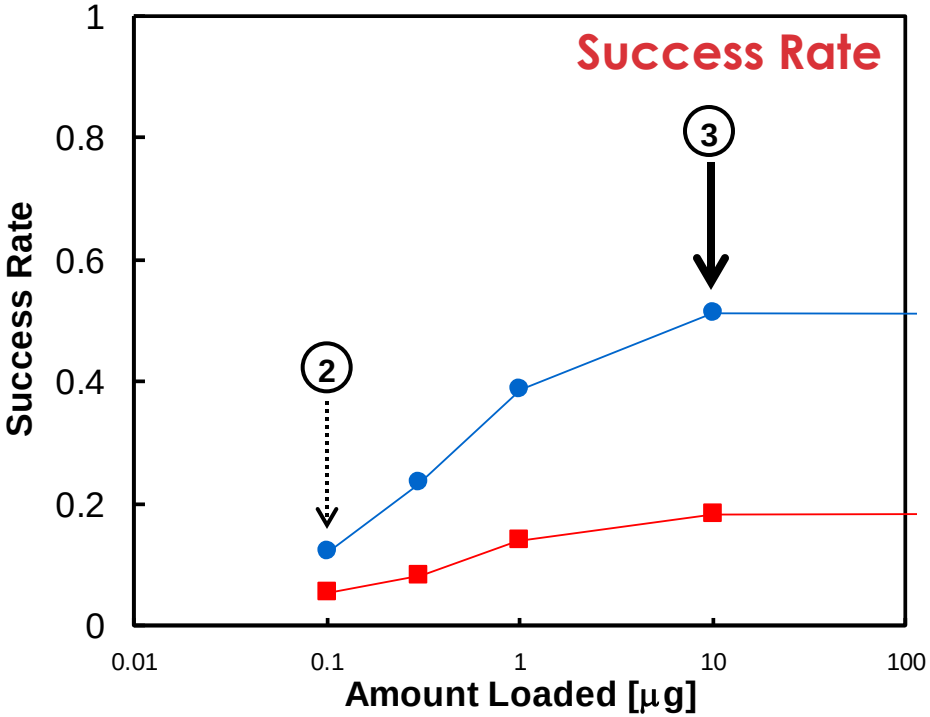


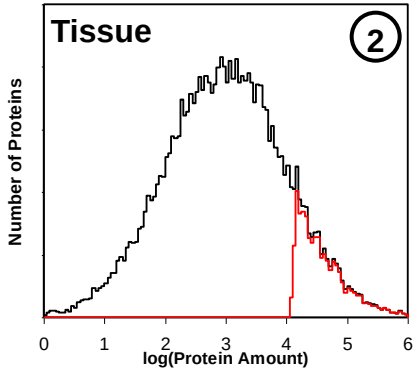


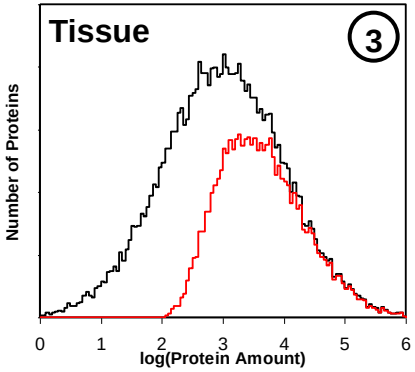


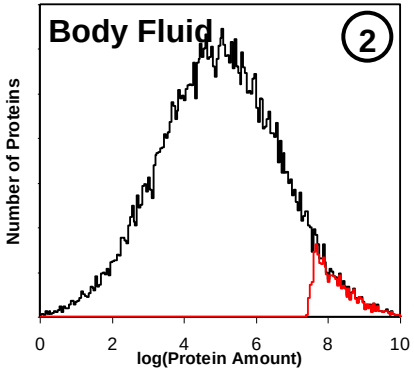
Amount of Peptides Loaded on the Column

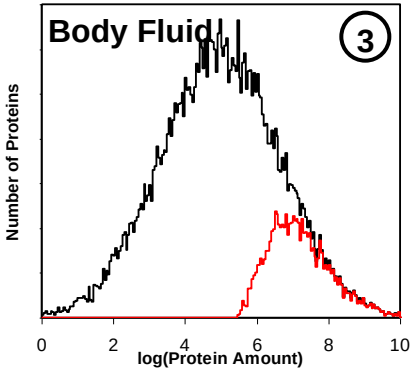




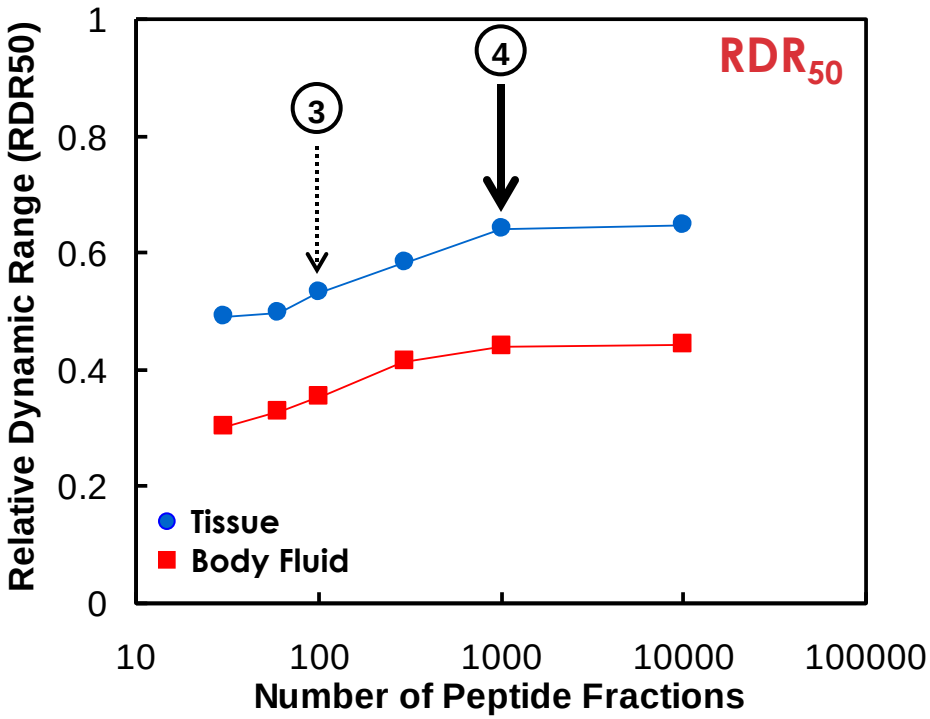


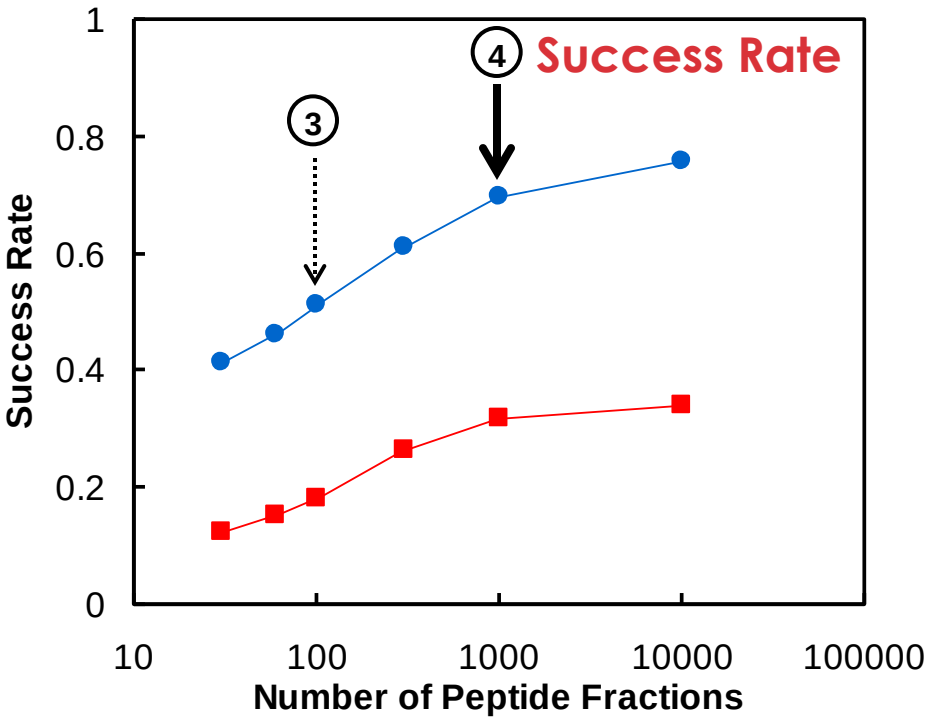


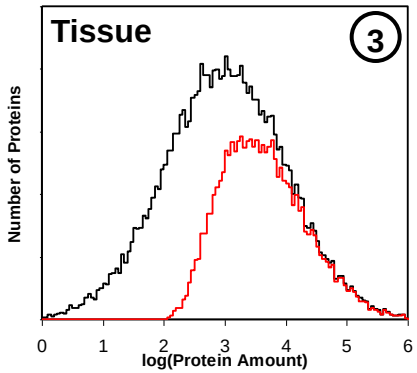


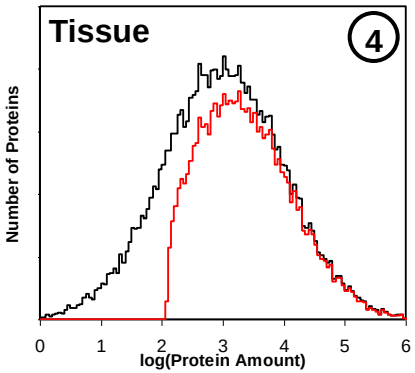


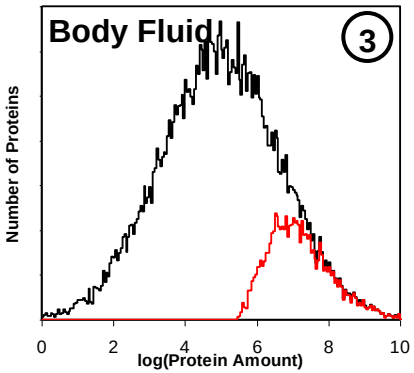
Peptide Separation

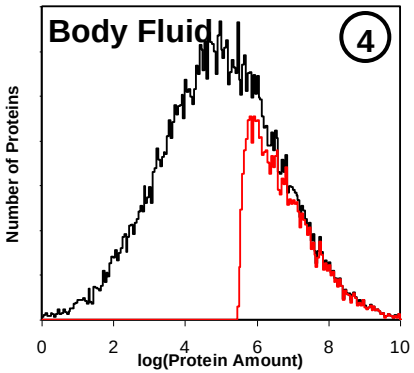








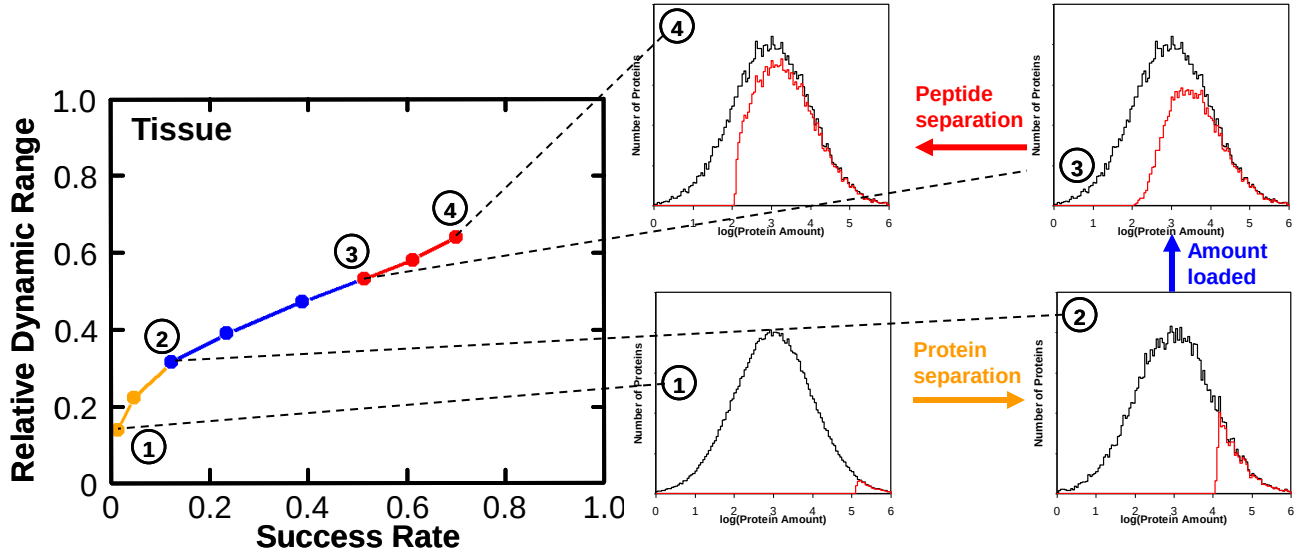




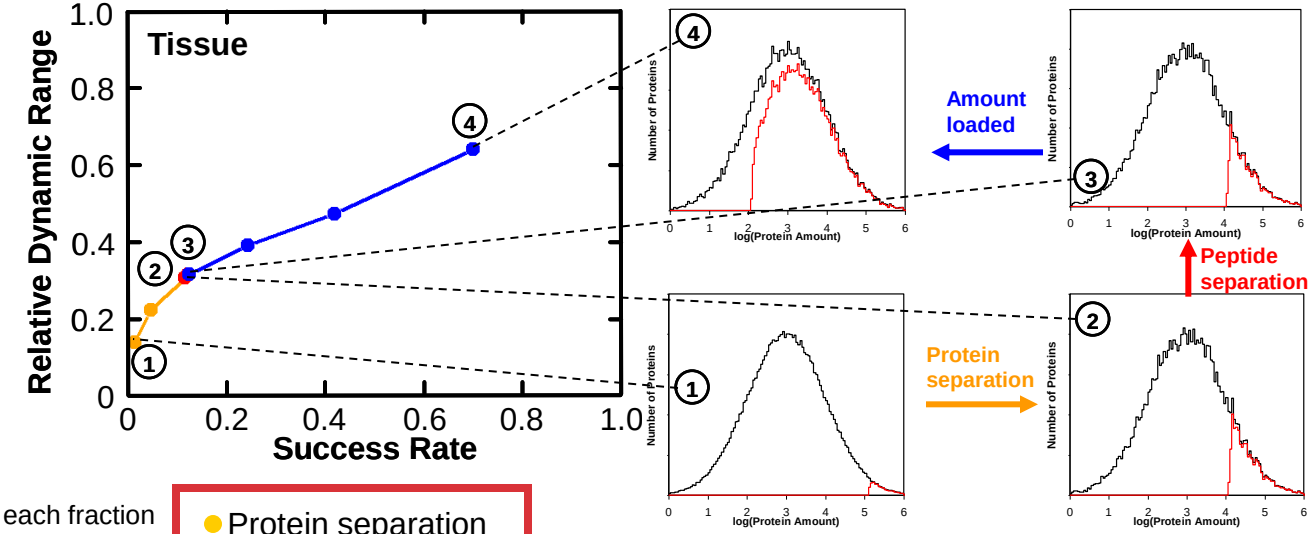
Amount loaded and peptide separation

Order:

1. Protein separation
2. **Amount loaded**
3. **Peptide separation**



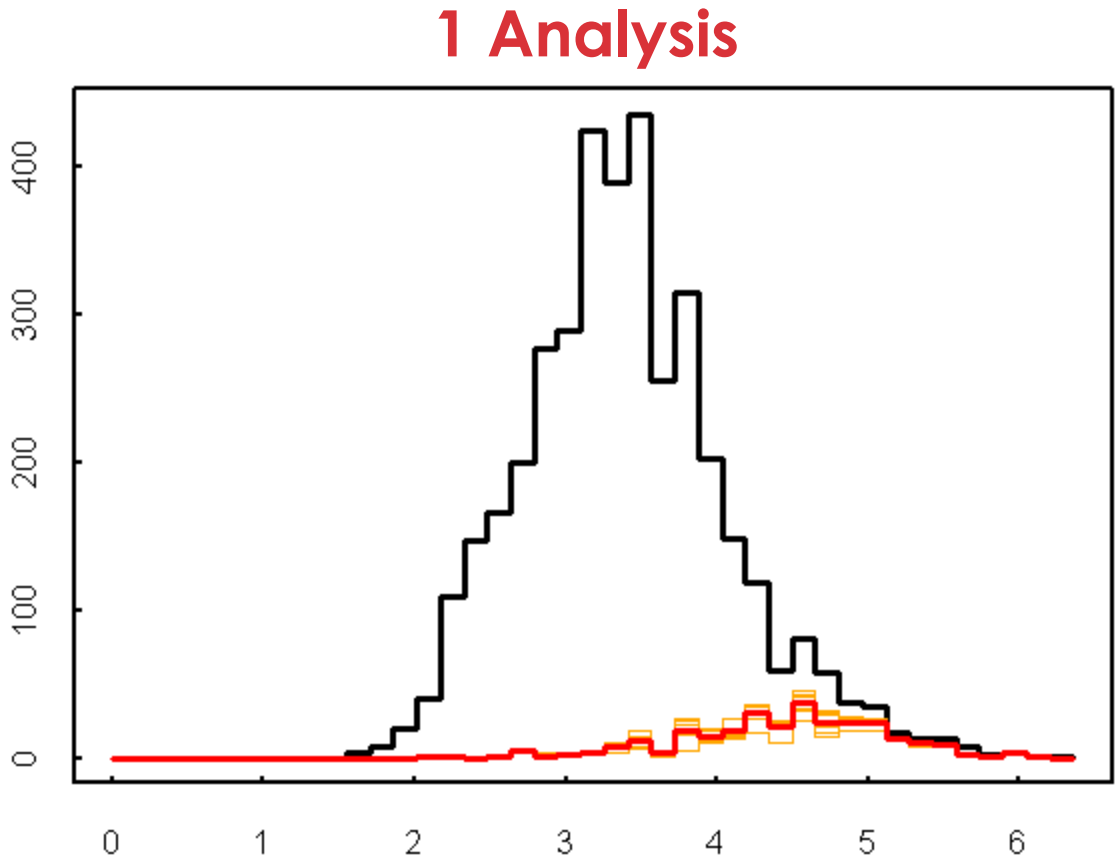
1. Protein separation
2. **Peptide separation**
3. **Amount loaded**



Ranges:
Protein separation: 30000 – 3000 proteins in each fraction
Amount loaded: 0.1 ug – 10 ug
Peptide separation: 100 – 1000 fractions

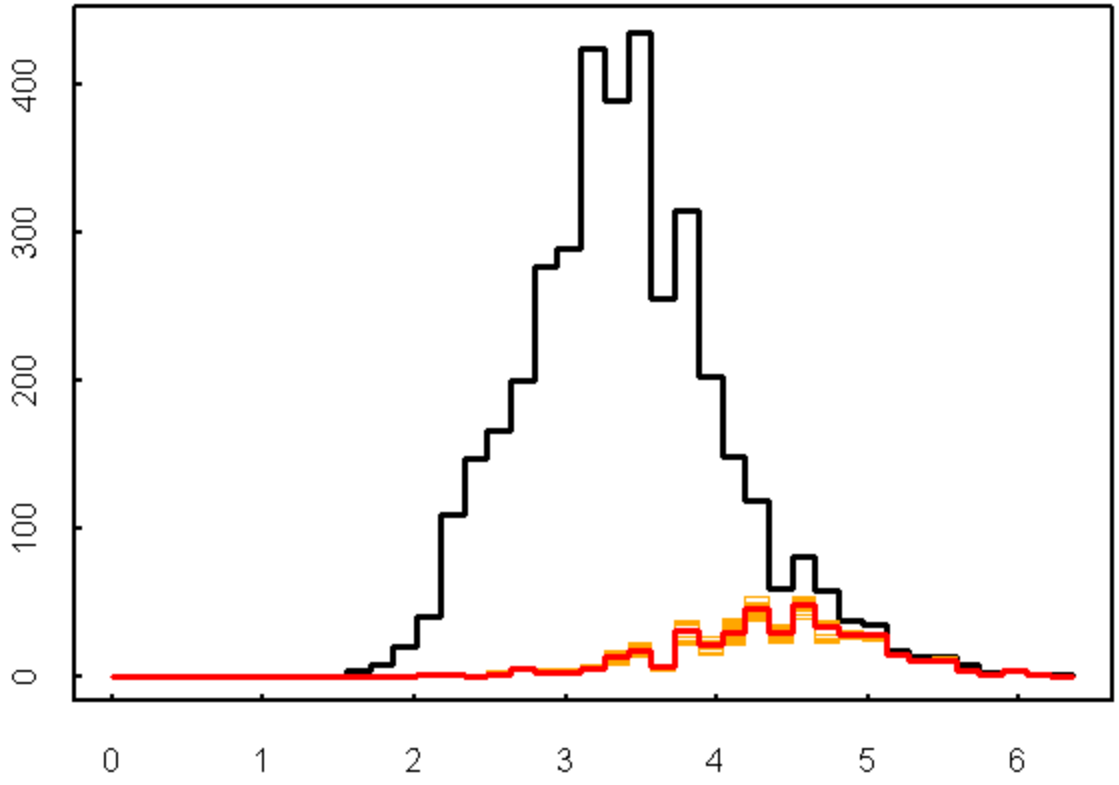
- Protein separation
- Amount loaded
- Peptide separation

Repeat Analysis



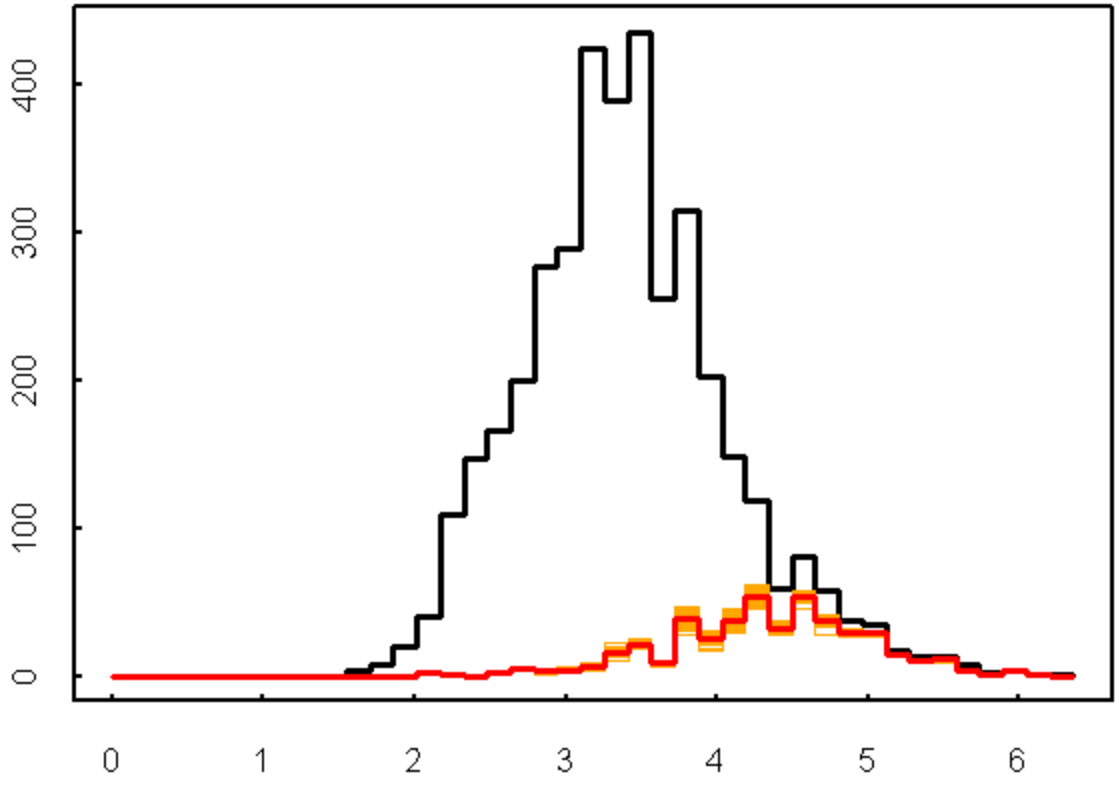
Repeat Analysis

2 Analyses



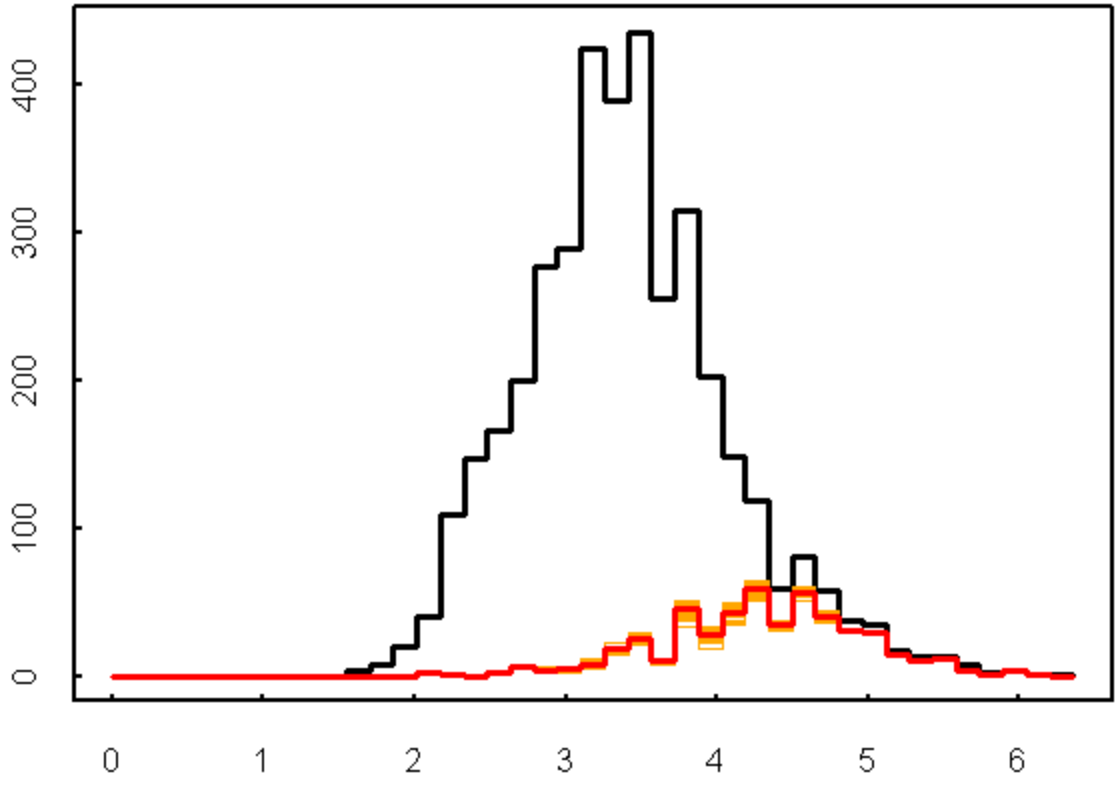
Repeat Analysis

3 Analyses



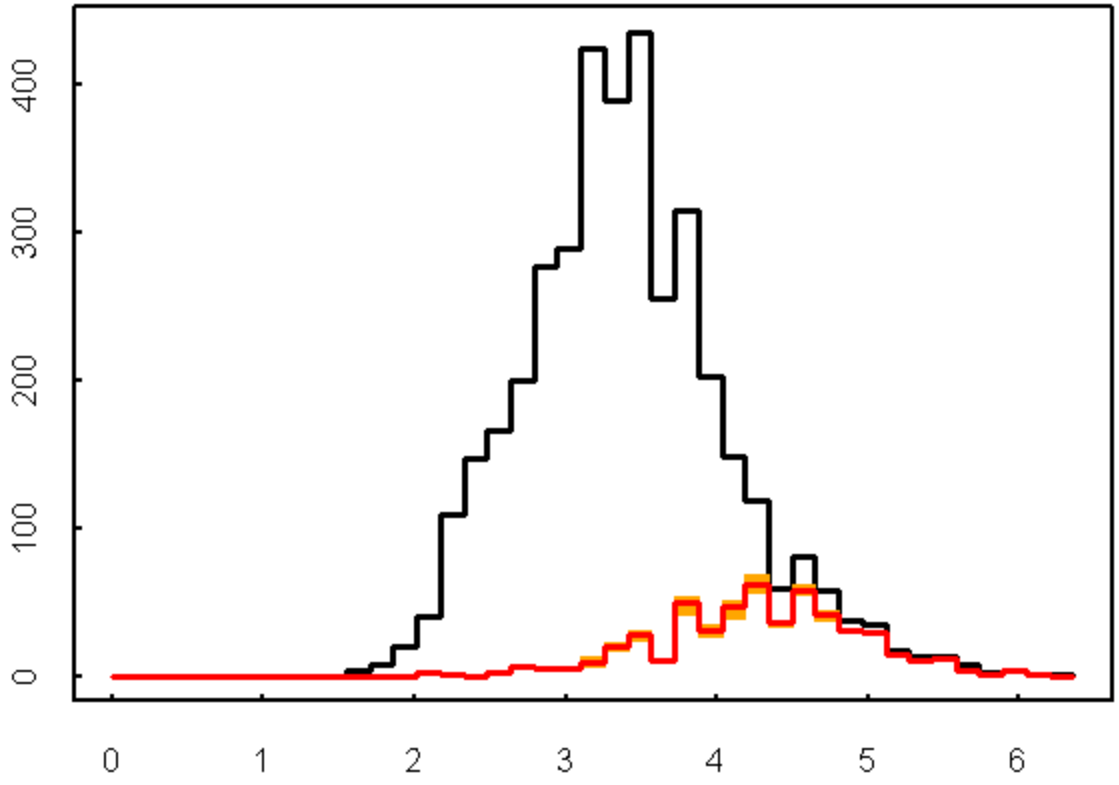
Repeat Analysis

4 Analyses



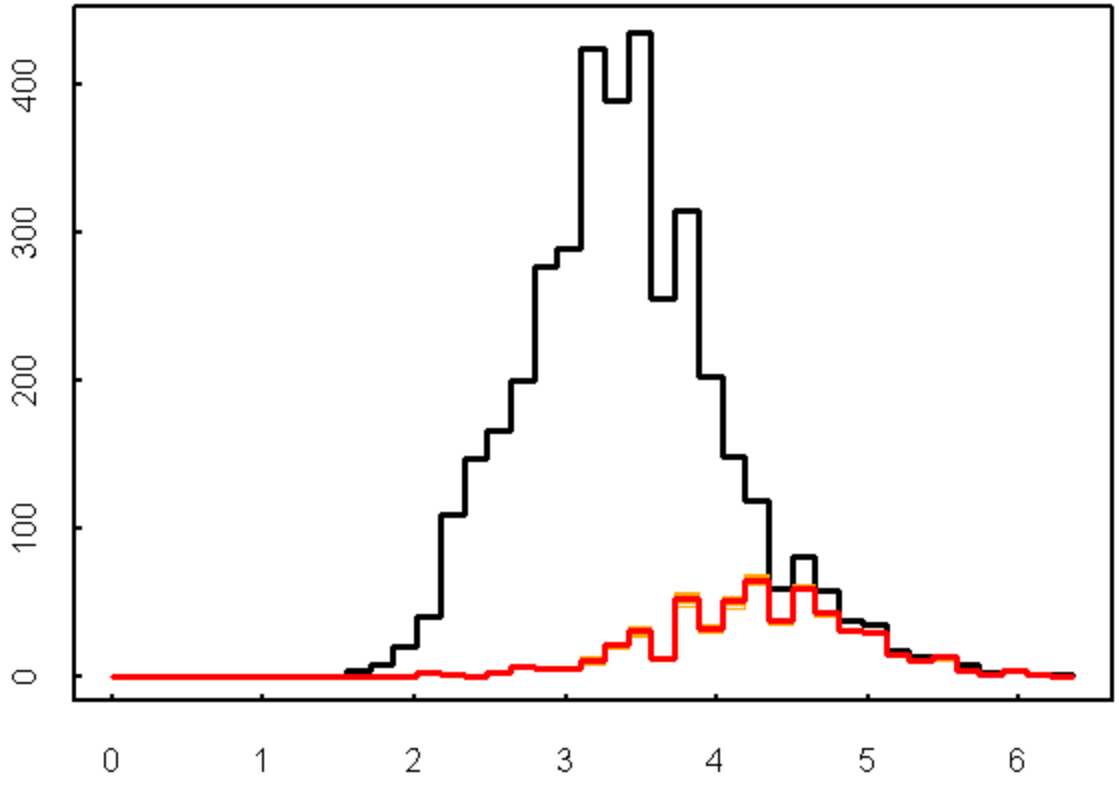
Repeat Analysis

5 Analyses

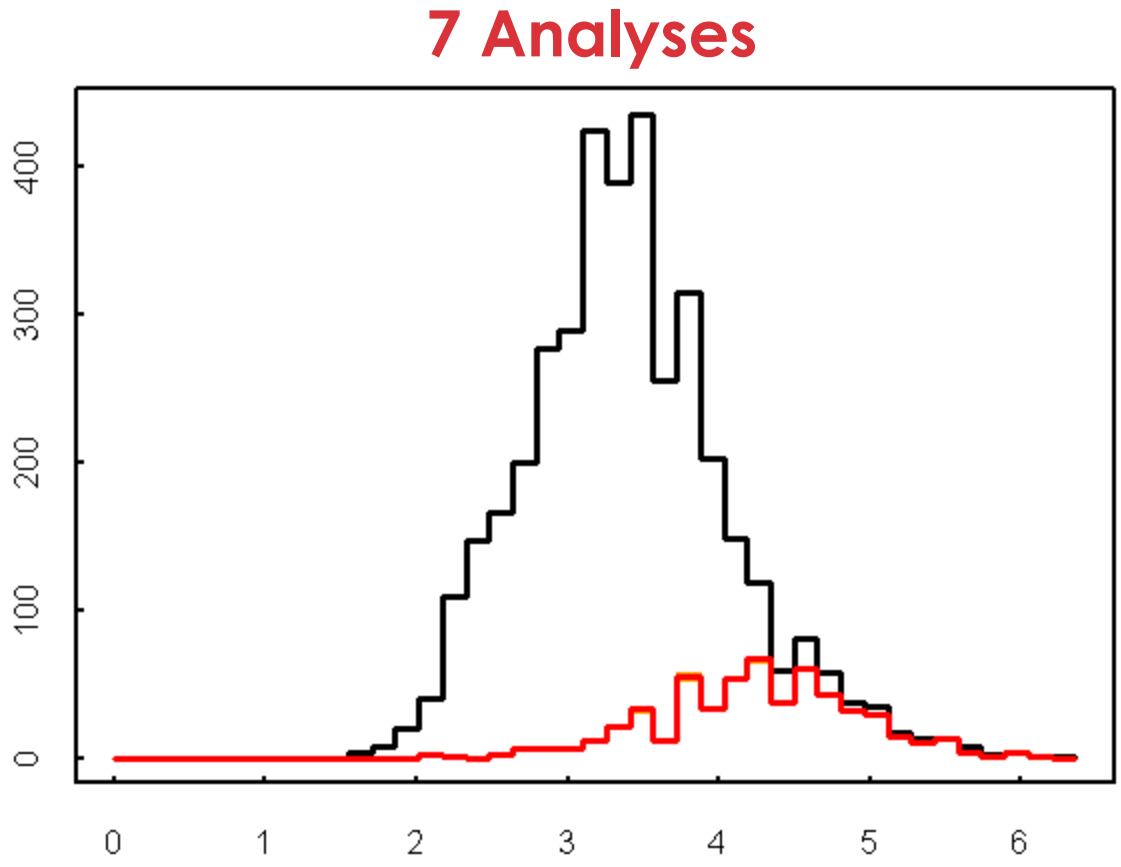


Repeat Analysis

6 Analyses

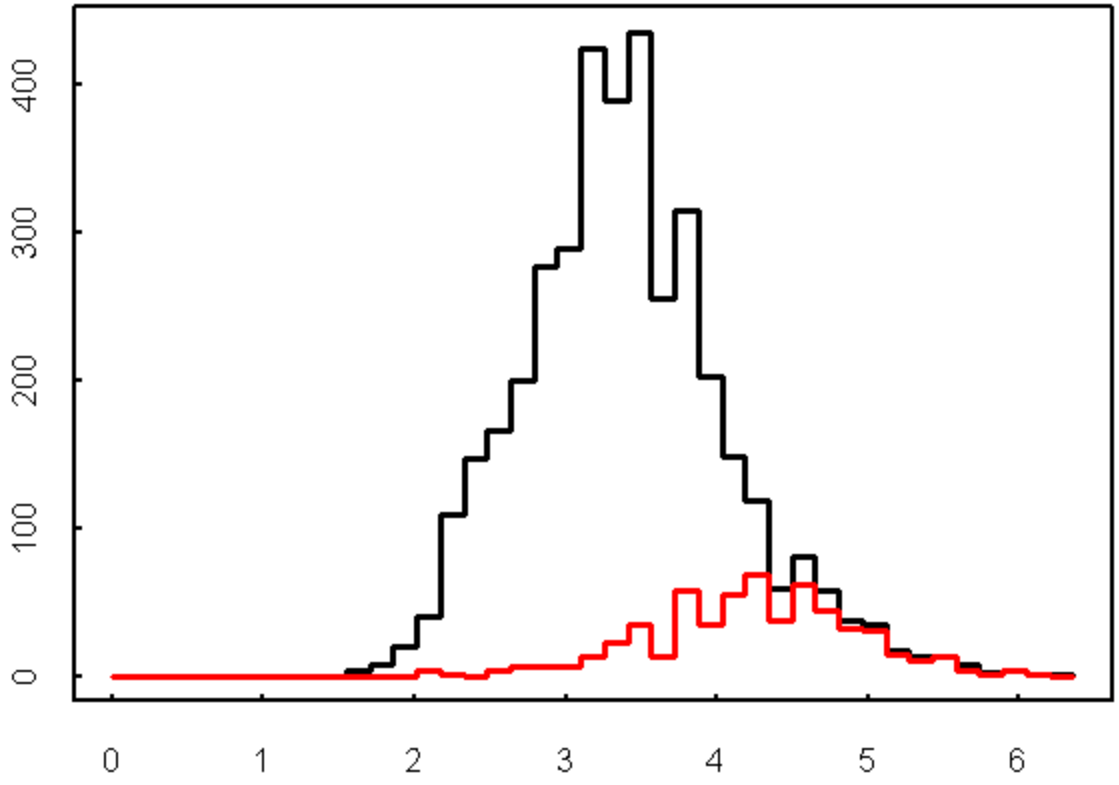


Repeat Analysis

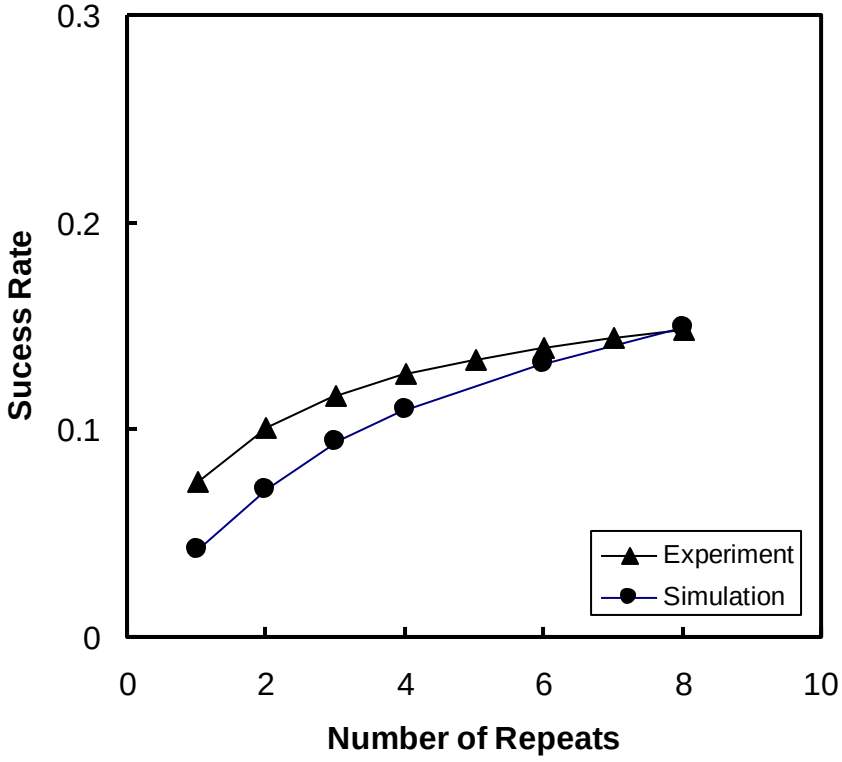


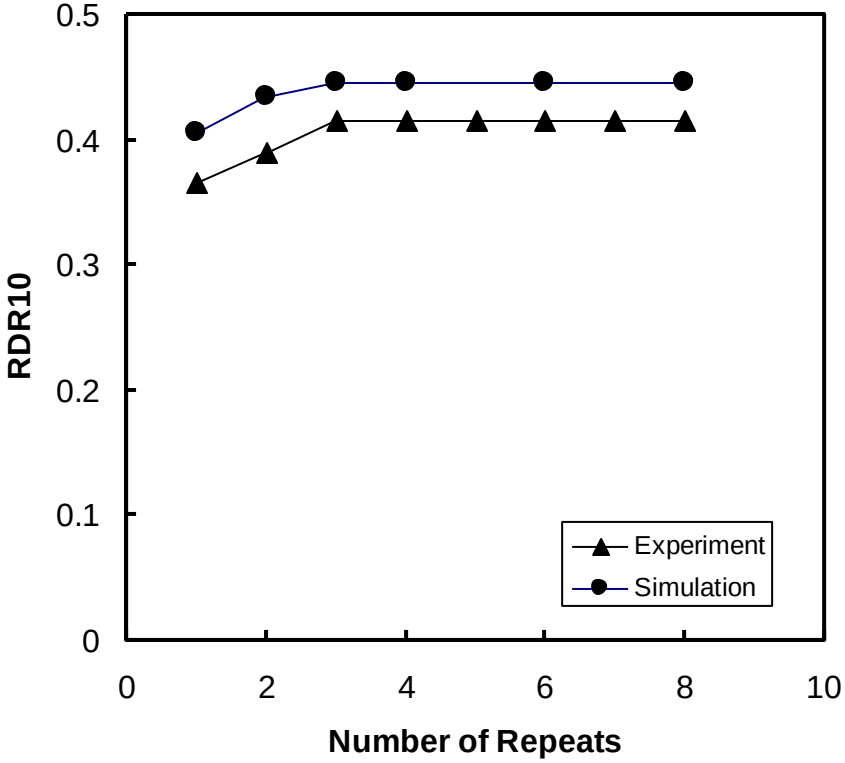
Repeat Analysis

8 Analyses



Repeat Analysis: Simulations





Summary

- The success rate of proteome analysis is influenced by the following factors (listed in order of importance):
- **The degree of protein separation**
- **Amount of peptides loaded on column or mass spectrometric detection limit**
- **The degree of peptide separation or mass spectrometric dynamic range**

Proteomics Informatics – Protein identification II: search engines and protein sequence databases (Week 5)