

## Bioinformatics/Search Engines/FDRs

Archer Smith IV, PhD  
Department of Biochemistry and Molecular Genetics  
[adsmith4@uab.edu](mailto:adsmith4@uab.edu)

## Biological Mass Spectrometry

### Bioinformatics:

Search Engines  
False Discovery Rates  
Post search Engine Processing

### Methods/Tools of Proteomics:

Fractionation  
Separation  
Enrichment

## Application of Mass Spectrometry in Biology

### 3 Major Levels For Biological Mass Spectrometry

#### **Protein Identification (Simplest)**

#### **Protein Modification**

Post Translational Modification (PTM)  
AA mutation(s)

#### **Protein Quantitation**

Global  
Affinity Directed

## Some Useful Techniques/Tools for Biological MS *(MS Instrumentation Not Included Here)*

### Common techniques:

#### **Separations:**

**2-D Gels** (Highly compatible with MALDI sources)  
**1-D Gels** (Separate first by size – then digest “bottom up” approach)  
**Mudpit** (Multidimensional Protein Identification Technology)  
**IEF** (Isoelectric focusing – useful for both “top down” and “bottom up”)

#### **Enrichment:**

**Antibodies** (Just about anything)  
**Metal Affinity Columns** (Typically phospho-specific)

#### **Quantitation:**

**MRM** (Multiple Reaction Monitoring - Triple Quads)  
**SILAC/iTRAC/iCAT** (Very broad instrument choices)

## Two Major Approaches To Biological Mass Spectrometry

**“Top Down”** – Examine whole protein(s)

**Cons:**

- i. The more complex the species – the more isoforms per protein
- ii. Requires a large magnet/field (the larger the protein the larger the magnet necessary)
- iii. Very high charge states are required for ionization and this makes interpretation much more complex
- iv. Other things not mentioned here

**“Bottom Up”** – Digest proteins and examine fragments

**Cons:**

- i. Requires significant processing of the protein(s) post purification
- ii. Proteins are not equally susceptible/vulnerable to protease cleavage
- iii. Proteases can be non-specific or erratic (we will discuss this topic later)
- iv. Adds complexity with overlapping peptides/missed cleavages (related to item # iii.)
- v. Other things not mentioned here

## The Bottom Up Approach

**“Bottom Up”** – Common Proteases

**Common Enzymes:**

Trypsin*	K-X, R-X
Chymotrypsin*	X-L, -F, -Y, -W (+)
Lys-C*	K-X
Arg-C*	R-X
Asp-N	X-D
Glu-C*	E-X

\* a tailing proline may inhibit digestion  
(+) may be much more erratic than AA listed

**The usage of multiple enzymes should increase the amount of sequence coverage! Very Important!**

# Digestion Example – BSA – Trypsin

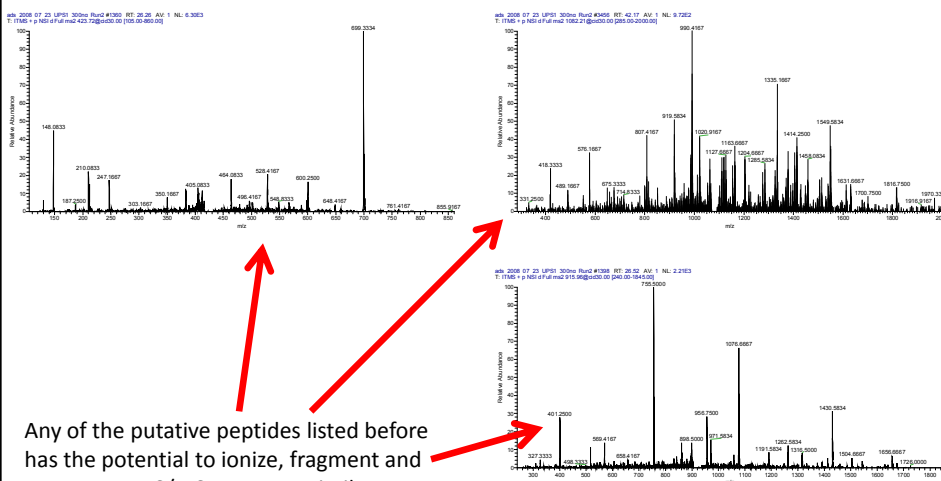
BSA Sequence – 607 AA. MW= 69293.4 DA

MKWVTFISLLLLFSSAYSRGVFRDRTHKSEIAHRFKDLGEEHFKGLVLIASFQYLQCC  
 PFDEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCKVASLRETYGDMADC  
 CEKQEPERNECFLSHKDDSPDLPLKLPDPNTLCEFAKDEKFKWGLYKLYEARRHP  
 YFYAPELKYANKYNGVFQEQCAEDKGACLLPKIETMREKVLASSARQLRRCASIQ  
 KFGERALKAWSVARLSQKFKAEFVETKLVDTLTKVHKECCHGDLLECAADDRADL  
 AKYICDNQDTISSKLECCDKPLLEKSHCIAEVEKDAIPENLPLTADFAEDKDVCKN  
 YQEAKDAFLGSLYYSRRHPEYAVSVLLRLAKEYEATLECCAKDDPHACYSTVFDK  
 LKHLVDEPQNLKQNCDFEKLGEYGFQNALIVRYTRKVPQVSTPLVEVSRSLGKV  
 GTRCCTKPESERPCTEDYLSLNLRLCVLHEKTPVSEKVKCCTESLVNRRPCFSAL  
 TPDETYVPKAFDEKLFTHADICTLPDTEKQIKKQTLVLLKHKPKATEEQLKVTMVE  
 NFVAFVDKCAADDKEACFAVEGPKLVVSTQTALA

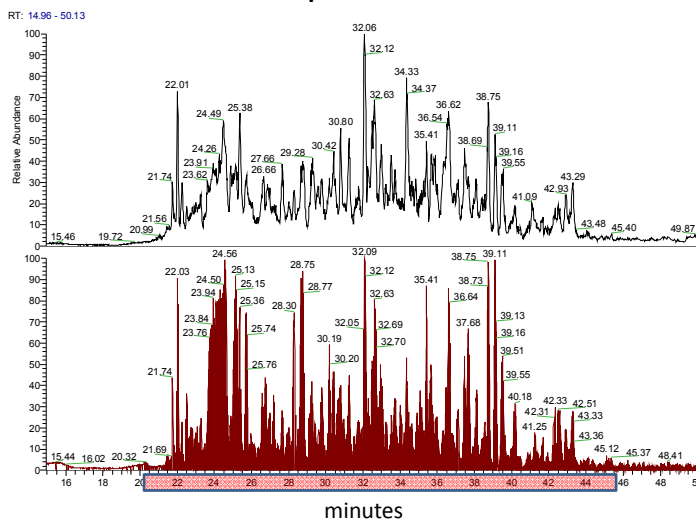
#	m/z (mi)	m/z (av)	Start	End	Missed Cleavages	Sequence
1	1817.489	817.9693	452	459	1	(R)ISGKIVETK(C)
2	2818.4254	818.9074	562	568	0	(K)ATEEQ(LK)T
3	3820.4676	820.9726	229	235	1	(K)FGERAL(K)A
4	4839.3022	839.9267	581	587	0	(K)CCAADK(E)
5	5847.5036	848.0391	242	248	1	(R)LSKQPK(A)
6	6886.4153	886.9394	131	138	0	(K)DSDPLK(P)L
7	7898.4815	899.1065	483	489	0	(R)ILVLEK(T)
8	8906.4713	907.088	205	211	1	(K)IETMREK(V)
9	9922.408	923.06	249	256	0	(K)KPFVETK(L)
10	10927.4934	928.0823	161	167	0	(K)YLVEAR(R)
11	11960.5473	961.1126	210	218	1	(R)IKVLASSAR(Q)
12	12974.4578	975.0521	37	44	0	(K)DGLGEEHFK(G)
13	13975.5404	976.1929	221	228	1	(R)LRCASSQ(K)F
14	14987.537	988.1411	29	36	1	(K)SEIAHRFK(D)
15	15987.5694	988.1414	212	220	1	(K)VLASSAR(Q)L
16	16988.5673	989.1637	490	498	1	(K)TPVSEK(V)C
17	17001.5891	1002.2118	233	241	1	(R)ALKAVSAR(L)
18	18002.583	1003.1908	598	607	0	(K)LVSTQTAL(-)
19	19014.6194	1015.2455	549	557	0	(K)QTLVELLK(H)
20	20034.4724	1035.1736	123	130	0	(R)INCFLSH(K)D
21	21068.4415	1069.1446	413	420	0	(K)QNDQFER(L)
22	221072.5092	1073.2201	310	318	0	(K)SHGIAEVEK(D)
23	231083.5946	1084.2709	161	168	1	(K)YLVEAR(H)
24	241107.5119	1108.2658	588	597	0	(K)EACFAVEK(P)L
25	251138.498	1139.3022	499	507	0	(K)KCTESLVN(R)
26	261142.7143	1143.4207	548	557	1	(K)KQTLVELLK(H)
27	271145.6436	1146.3427	236	245	1	(R)KAVSARLSQ(K)F
28	281153.6939	1154.4035	257	266	1	(K)LVDTLTKV(K)E
29	291163.6307	1164.3521	66	75	0	(R)NLNELTEFAK(T)
30	301166.4929	1167.3127	460	468	0	(R)KCTKPESE(R)M
31	311535.6022	1194.3001	25	34	1	(R)ITTHESIAE(R)E
32	321195.5888	1196.3788	223	232	1	(R)IASIQKFER(A)
33	331249.6212	1250.4051	35	44	1	(R)FKDLGEEHFK(G)
34	341254.5783	1255.4003	337	346	1	(R)DVKRVQEK(D)
35	351283.7106	1284.51	361	371	0	(R)HYAVSAR(L)R
36	361291.6021	1292.5249	300	309	0	(K)EEDKPLK(S)
37	371294.7042	1295.5305	246	256	1	(K)PKAEFVETK(L)
38	381305.7161	1306.5136	402	412	0	(K)HLVDLPQNLK(Q)
39	391308.727	1309.5172	558	568	1	(R)KPKATEEQ(L)K(T)
40	401388.7389	1389.7327	198	209	1	(K)GACLLPKIETM(R)E



# Examples of MS/MS spectra



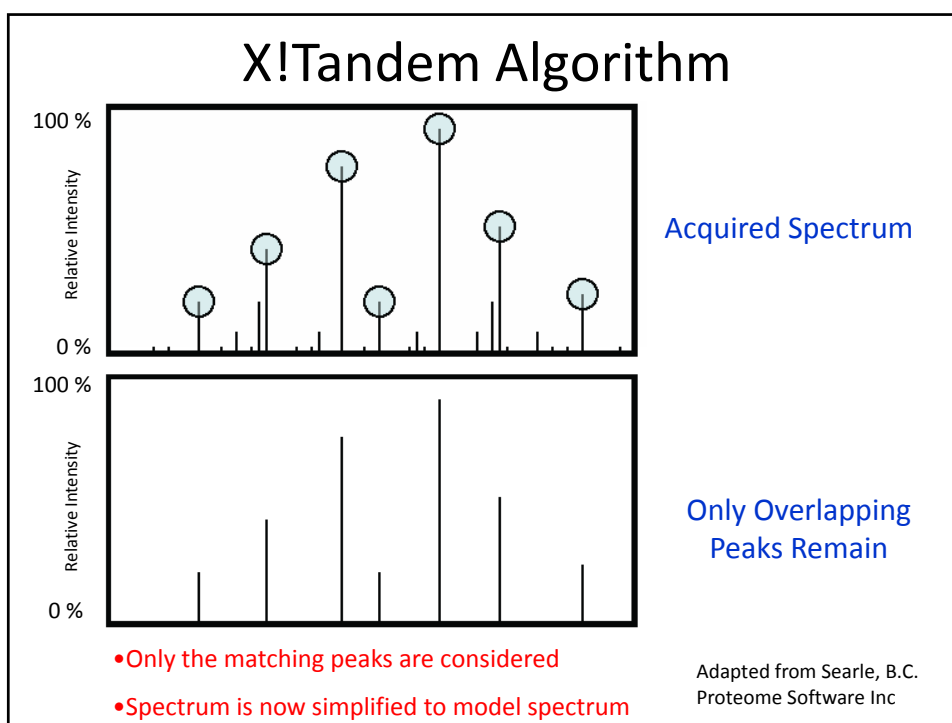
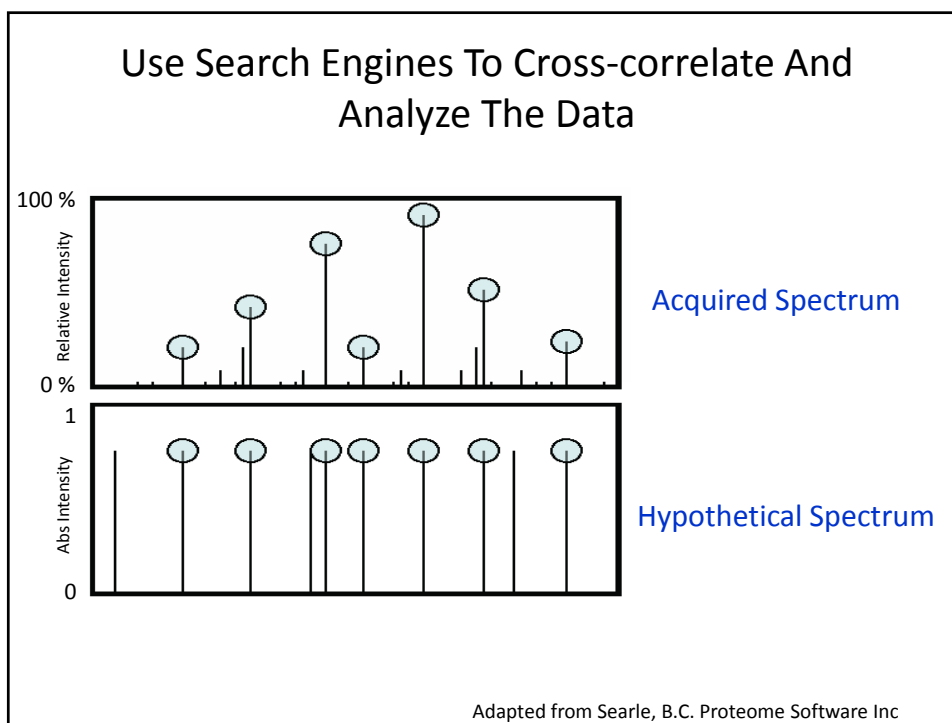
## Example Chromatogram - Tryptic Digest of a Complex Mixture



25 min ~ 3000 Spectrums Acquired

## Data Analysis - Pt. 1

- Now there is a significant amount of data to be analyzed >3000 spectra
- This is too much data to interpret manually
- It is necessary to use computers to process and analyze the data down to a manageable level

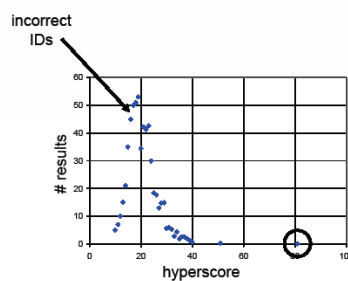


## X!Tandem Hyper-score

$$\text{Hyper-score} = \sum(I_{\text{obs}} * P_{\text{ints}}) * N_b! * N_y!$$



Simulated Spectrum



- The hyperscore indicates how closely peptide model is relative to the actual spectrum
- The match with the highest hyperscore is considered the “real” match
- Another set of statistics is performed on the hyperscore and generates an E-value

Adapted from Searle, B.C. Proteome Software Inc

## A Few Examples of Search Engines

### Two major approaches:

#### De Novo

PEAKS/  
PepNovo



#### Cross-correlation (or pattern recognition)

OMSSA

OMSSA

PHENYX

Phenyx



Protein Prospector

X!

X!Tandem



Sequest

MATRIX  
SCIENCE

Mascot

#### Statistics/Visualization:



Protein  
Prophet

BIOINQUIRE

PROTEOIQ



Scaffold

Further Refinement

## Search Engine Outputs Vary

<u>X!Tandem</u>	<u>Sequest</u>	<u>Mascot</u>	<u>Protein Prospector</u>
Hyperscore	Xcorr	Mascot Score	Peptide Score
E-value	Sp		Expectation Value
	Sf		
	$\Delta Cn$		

### Each Search Engine Has Unique Output Files/Formats

**Q1. *What is the proper way to search for my protein(s) of interest?***

**Q2. *How do I utilize the output information in the best manner?***

## Data Analysis - Pt. 2 Q1.

**Q1. *What is the proper way to search for my protein(s) of interest?***

**What are you looking for?**

protein identification? – all mods off  
 a modification? – which modifications?  
 quantitation? – which type of label?

**Which database should I use?**

entire proteome or subset database?

- **Pro** - Smaller databases reduce search time
- **Con** -Smaller databases “encourage” the algorithm for “force” a fit



## Data Analysis - Pt. 2 Q1 Cont.

**Q1. *What is the proper way to search for my protein(s) of interest?***

**What type of enzymatic cleavage?**

enzyme specific/semi/or non specific (No enzyme)

- **Pro** – Enzyme specific searches require less time to search
- **Pro** – Enzyme specific cleavages increase the confidence in the identification
- **Con** – Enzymes are not always 100% reliable/predictable
- **Compromise** – “SEMI”- specific at one termini

**Proteomics/Complex Mixtures?**

need to search with a concatenated decoy database

## Data Analysis - Pt. 2 Q1 Cont.

**Q1. *What is the proper way to search for my protein(s) of interest?***

**What mass window do I use?**

The larger the mass window – the more possibilities there are to match your protein(s) of interest - (can be good or bad thing)

Example Parameters:

Trap (LTQ) – typically 0.45Da to 2.0 Da  
 High Res TOFs – 50<ppm to 0.1 Da  
 FT's & Orbitraps – 7-10 ppm

**Larger mass windows:**

- **Pro** – Includes more peptides
- **Con** – Dramatically increases search time

# Sequest Example

**Run TurboSequest**

Sequest Protein Utilities Help Home

Searchflow: Default, Trypsin, rabbit (Indexed), 2009-01-24 [Save]

Directory: Single Multiple  
ads\_2009\_01\_22\_EKLF\_Band1

Dta Files: All Selected Select... Clear OUTs?

Database & Enzyme  
First: rabbit.fasta.hdr  
Second: [ ]  
Options: Auto Protein Nucleotide  
Enzyme: Trypsin  
Cleaves At: Both Ends

Server Options  
Sequest Queue: 0 Dir 0 Procs  
Priority: 5  
Continue [ ]  
Oper: [ ] Run SEQUEST Help

Differential Modifications  
Symbol AA DiffMass  
\* M 15.9949  
# 14.0157  
@ 79.9663  
^ 31.9898  
~ [ ]  
\$ [ ]  
] N-term  
[ C-term

Options Edit Add-Mass Edit Advanced

Parent Mass Type: Mono Avg  
Fragment Mass Type: Mono Avg  
Peptide Mass Units: amu mmu ppm  
Peptide Mass Tolerance: 2.5000 amu  
Fragment Ion Tolerance: 1.0000  
Output Lines: 12  
Description Lines: 12

Neutral Losses (H<sub>2</sub>O/NH<sub>2</sub>)  
a: [ ] b: [x] y: [x]

Ion Series Weightings  
a: [0.0] b: [1.0] c: [0.0]  
d: [0.0] v: [0.0] w: [0.0]  
x: [0.0] y: [1.0] z: [0.0]

Normalize XCorr Values: [ ]  
Sequest Header Filter: [ ]  
Partial Sequence: [ ]  
Use different parameters to continue: [ ]  
Load params from selected directory: Refresh  
Show Fragment Ions: [ ]

Eng, J.K. et al 1994 JASMS 976 – 989.

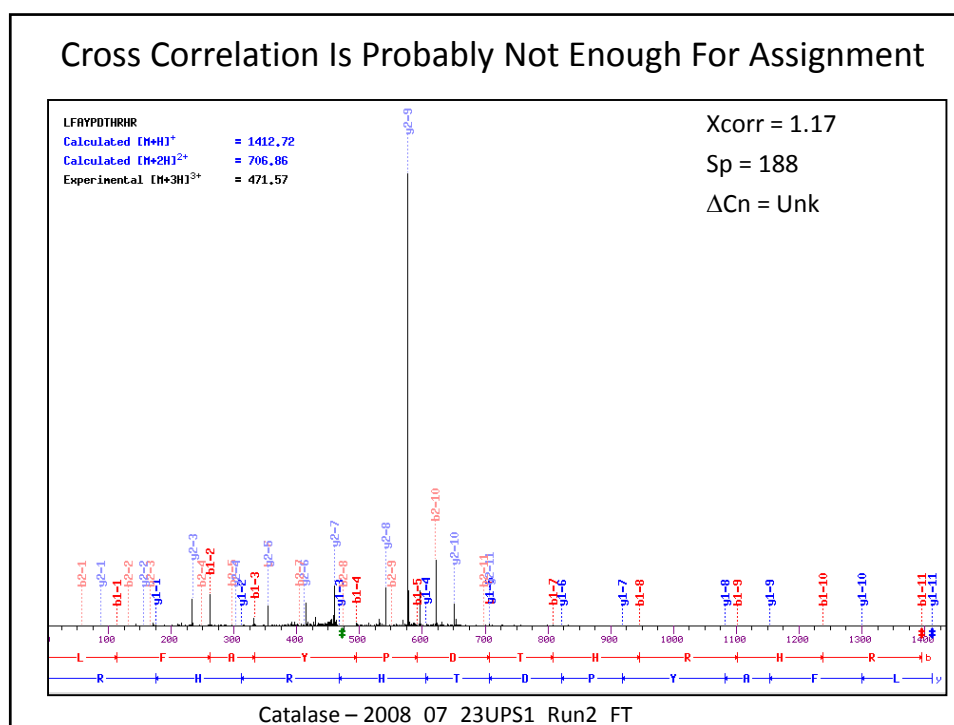
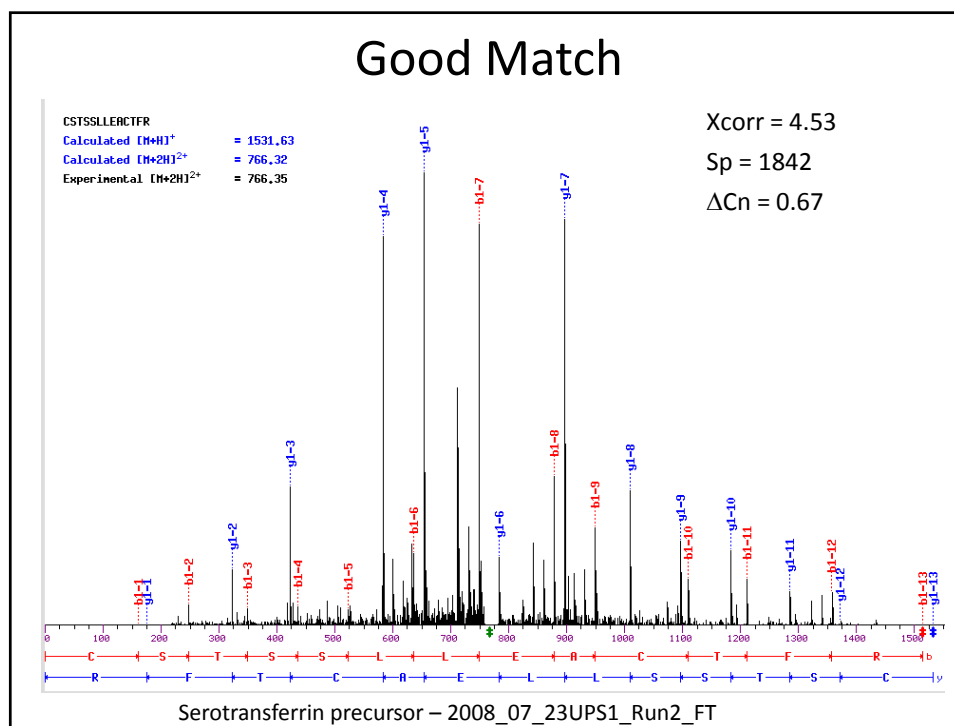
# Sequest Output File

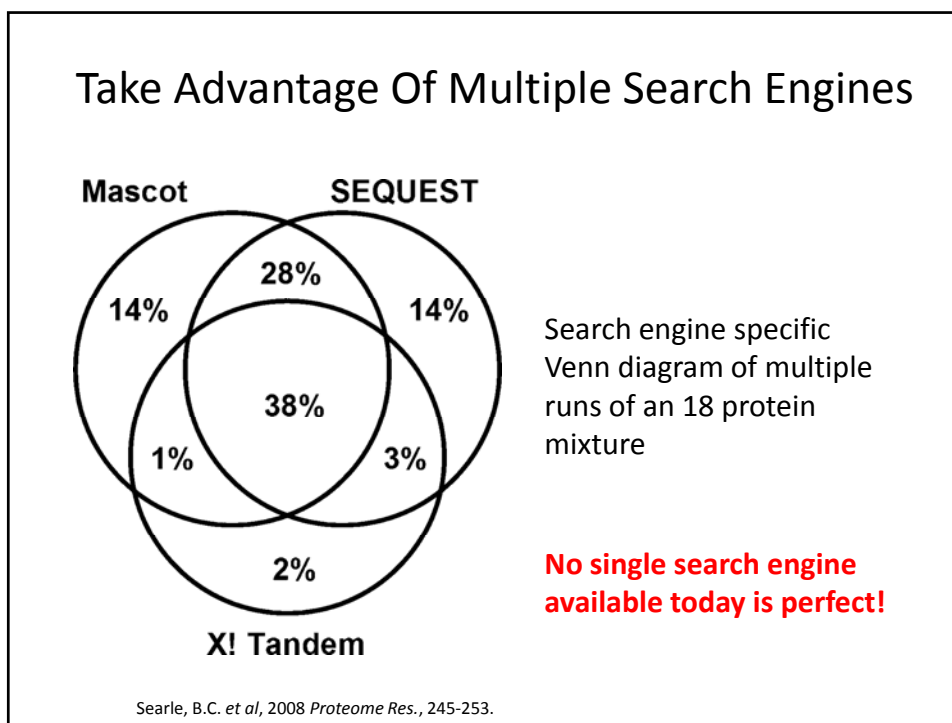
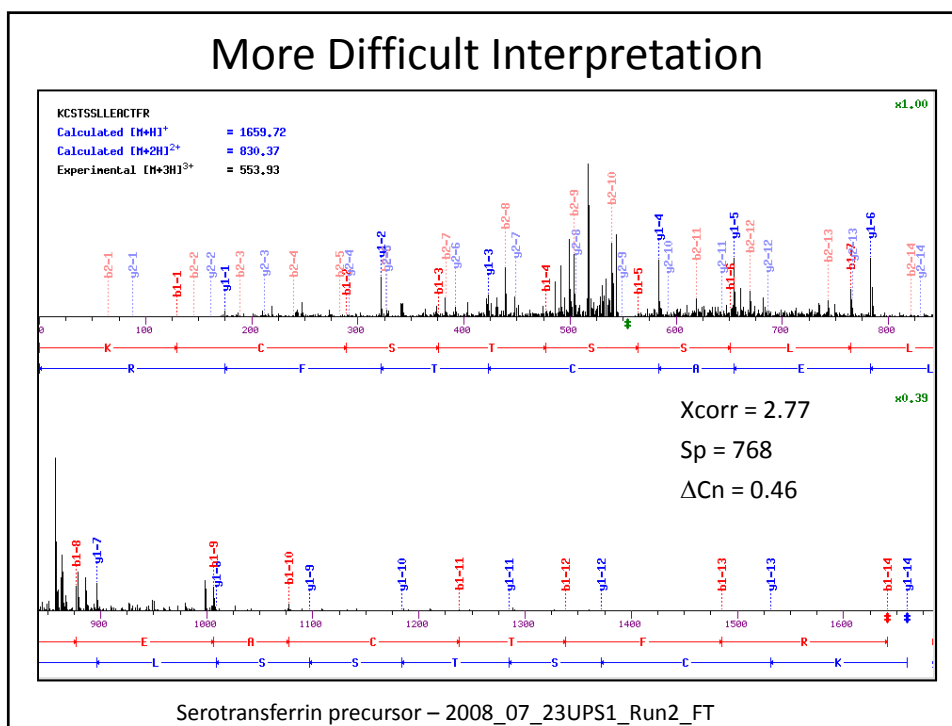
Xcorr  
ΔCn  
Sp

```
*Cluster: Myoglobin(n=1) Homo sapiens(Rep: Myoglobin - Homo s
[ ] Unref100_C00762 Covr: 0.0% Uniq: 0 Sequences: 10 Score: 0.00 Avg: 0.00 TIC: 0% I: 0.040 AvgI: 0.040 {10,0,0}
*Cluster: Ubiquitin-conjugating enzyme E2 C(n=4) Homo/Pan/Gor
2622 ----- 02329 ----- 3 -0.0160 3030.4781 6.02 0.39 911 1 37/103 -- 0 [ ] Unref100_c00762 +2 (R)IQGEMTIPKGGKGIAPFSDNLFK(W)
1600 ----- 02817 ----- 2 -0.0084 1882.9410 4.28 0.58 710 1 20/22 -- 0 [ ] Unref100_c00762 +1 (R)LSFFPSDFYVHPTFK(F)
2189 ----- 01131 ----- 2 -0.0009 1800.8659 3.75 0.64 471 1 19/28 -- 0 [ ] Unref100_c00762 +3 (R)YIGETKSGVDTGDF(-)
2446 ----- 03411 ----- 4 -0.0212 3106.5644 3.54 0.37 694 1 41/162 -- 0 [ ] Unref100_c00762 +2 (R)PIAQETMIDGSDGDIATFESDNTFK(W)
961 ----- 03554 ----- 4 -0.0278 3160.1224 3.46 0.40 500 1 39/204 -- 0 [ ] Unref100_c00762 +2 (R)IILLSIYVLAGRPLDGFNTDANLHNDPTAFK(V)
1430 ----- 02507 ----- 3 -0.0189 2174.1319 3.03 0.39 485 1 21/72 -- 0 [ ] Unref100_c00762 +1 (R)YKSLFPPSYVHPTFK(F)
2295 ----- 01783 ----- 2 -0.0009 1009.5105 2.76 0.48 1361 1 14/24 -- 0 [ ] Unref100_c00762 +2 (R)RSLVPTK(F)
1600 ----- 03542 ----- 5 -0.0061 3860.1018 2.36 0.36 508 1 47/232 -- 0 [ ] Unref100_c00762 +2 (R)IILLESQSLGEPHIDSEPLFTHAETLHNDPTAFK(V)
2331 ----- 01831 ----- 4 0.0138 2011.0177 2.07 0.09 395 10 29/146 -- 0 [ ] Unref100_c00762 +1 (R)DPAATVMAARHGRFSSGAR(L)
922 ----- 02482 ----- 4 0.0112 2875.4162 1.97 0.27 280 1 27/138 -- 0 [ ] Unref100_c00762 +1 (R)FLFCTKRPDVTQWICLDLTKK(W)
[ ] Unref100_P01008 Covr: 0.0% Uniq: 0 Sequences: 11 Score: 0.00 Avg: 0.00 TIC: 0% I: 0.040 AvgI: 0.040 {10,1,0}
*Cluster: Anfibombin-III precursor(n=2) Homo sapiens(Rep: A
[ ] Unref100_C00742_12 Covr: 0.0% Uniq: 0 Sequences: 11 Score: 0.00 Avg: 0.00 TIC: 0% I: 0.040 AvgI: 0.040 {7,5,7,4} Shared:
*Cluster: Tibin(n=1) Homo sapiens(Rep: Tibin - Homo sapiens (
[ ] Unref100_P10145 Covr: 0.0% Uniq: 0 Sequences: 7 Score: 0.00 Avg: 0.00 TIC: 0% I: 0.040 AvgI: 0.040 {7,0,0,0}
*Cluster: Interleukin-8 precursor (IL-8) (C-X-C motif chemokine
[ ] Unref100_P00781 Covr: 0.0% Uniq: 0 Sequences: 7 Score: 0.00 Avg: 0.00 TIC: 0% I: 0.040 AvgI: 0.040 {7,0,0,0}
*Cluster: Trypsin precursor(n=1) Sus scrofa(Rep: Trypsin prec.
[ ] Unref100_LP01001SAB0A5 Covr: 0.0% Uniq: 0 Sequences: 11 Score: 0.00 Avg: 0.00 TIC: 0% I: 0.040 AvgI: 0.040 {6,8,4,7} Shared:
```

**UbcE2 protein group is expanded to show how the data is organized**

- Cutout represents 6 protein groups out of > 200 putative protein groups
- Peptide hits are organized by highest to lowest Xcorr value





## Post Translational Modification(s)

**PTM's are very ubiquitous in nature however.... they can be problematic to detect**

### Things to consider about PTMs:

- PTM's are typically in a lower dynamic range than non-modified proteins
- Many PTM's are labile chemically – partly because they are so dynamic in nature
- On the peptide level, some PTM's inhibit “traditional” ionization techniques
- Some PTM's inhibit protease activity at adjacent AA residues (bottom up approach)
- Each PTM will add  $2^n$  more search time per modification
- Or some PTM's are falsely identified upon multiple factors – i.e. mass resolution or species/organism

## Searching For PTMs

Run TurboSequest

Sequest Protein Utilities Help Home

Searchflow: Default, Trypsin, rabbit (Indexed), 2009-01-24 Save

Directory:  Single  Multiple  
 ▼

Dta Files:  All  Selected Select...  Clear OUTs?

Database & Enzyme  
 First:  ▼  
 Second:  ▼  
 Options:  Auto  Protein  Nucleotide  
 Enzyme:  ▼  
 Cleaves At:  ▼

Server Options  
 Sequest Queue: 0 Dir 0 Procs  
 Priority:  ▼  
 Continue

Oper:  Run SEQUEST Help

Differential Modifications

Symbol	AA	DiffMass
*	N	15.9949
#	DE	14.0157
@	STY	79.9663
^	STY	97.9769
~	ST	365.132
\$	ST	203.079
	N-term	
[	C-term	14.0157

Options Edit Add-Mass Edit Advanced

Parent Mass Type:  Mono  Avg  
 Fragment Mass Type:  Mono  Avg  
 Peptide Mass Units:  amu  mmu  ppm  
 Peptide Mass Tolerance:  amu  
 Fragment Ion Tolerance:   
 Output Lines:   
 Description Lines:

Neutral Losses (H<sub>2</sub>O/NH<sub>3</sub>)  
 a:  b:  y:

Ion Series Weightings  
 a:  b:  c:   
 d:  v:  w:   
 x:  y:  z:

Normalize XCorr Values:   
 Sequence Header Filter:   
 Partial Sequence:   
 Use different parameters to continue   
 Load params from selected directory Refresh  
 Show Fragment Ions:

There are a large amount of known PTMs available at [www.unimod.org](http://www.unimod.org)

## Data Analysis - Pt. 2 Q2.

**Q2. How do I utilize the output information post search engine?**

### A Few Examples:

#### A. High Purity Gel Band:

You should be able to stop here and analyze the data directly from the search engine(s) output file.

#### B. Modified Protein (PTM):

- i. re-analyze for any mass modifications (Protein Prospector)
- ii. re-search with different modification(s) (UNIMOD)

#### C. Complex Mixture/Whole Proteome/Quantitation:

**You have only made the first step – next we perform statistics.**

## Data Analysis - Pt. 3

**Q3. How do I pick the correct hits from a large to massive data set?**

**Most common strategy** - Decoy database searches.

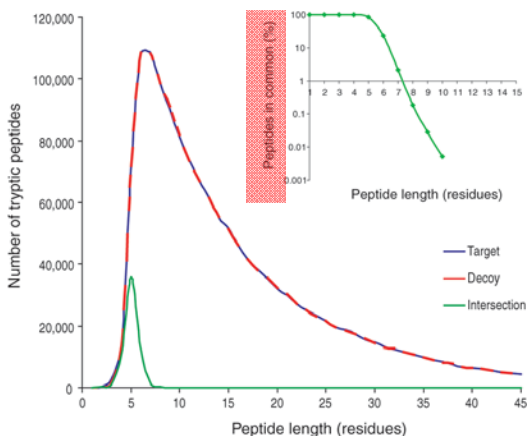
**Background:** The decoy database strategy allows one to “estimate” the number of false positives in the target database by using the number of correct hits in the decoy database as a metric – **calculate a False Discovery Rate (FDR)**

Measurement	Formulation	Description of estimate
False positive (FP)	2 times passing decoy assignments	Number of incorrect assignments above score threshold
True positive (TP)	Total passing assignments – number of FPs	Number of correct assignments above score threshold
Total correct (TC)	Maximum TPs for all evaluated score criteria combinations	Number of total correct assignments in the data set
Total incorrect (TI)	Total assignments – TC	Number of total incorrect assignments in the data set
False negative (FN)	TC – TP	Number of correct assignments falling below score threshold
True negative (TN)	TI – FN	Number of incorrect assignments falling below score threshold
Precision	$TP / (TP + FP)$	Fraction of correct assignments above score threshold
FP rate	$FP / (TP + FP)$ or $1 - \text{precision}$	Fraction of incorrect assignments above score threshold

Elias, J. E. *et al.*, 2007 Nat. Methods, 207-214.

# The Concatenated Database Approach

Generate a Decoy Database By Reversing the Sequence



Analysis comparing the statistical possibility of there being the same peptide in both databases as a function of peptide length

Elias & Gygi, 2007 Nat. Methods, 207-214.

## Hits Arranged by Xcorr

Xcorr

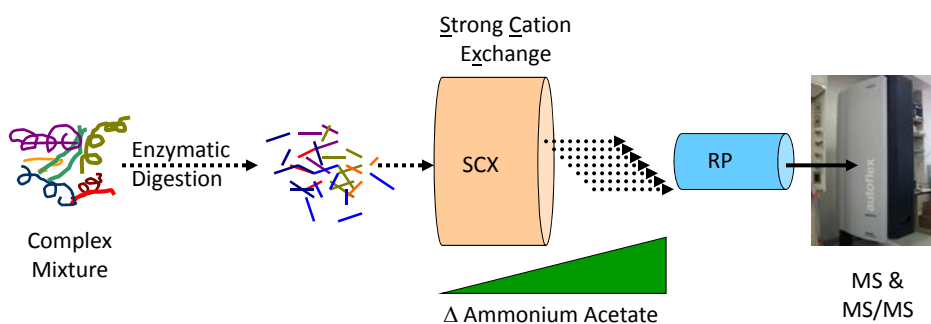
Rank	Xcorr	Sequence	Reference	Ext. FFP	Ext. TP	Ext. FN	Ext. TN	precision	sensitivity	accuracy
1	8.43	RLELLTADK	IP190006822 2	0	1	2152	156	1.000	0.000	0.170
2	8.15	QSGPFGQPR	IP190007352 1	0	3	2451	456	1.000	0.001	0.176
3	7.92	KIRFEDPQK	IP190007188 1	0	3	2190	466	1.000	0.001	0.176
4	7.90	KGRFTEGALVGVDPQWIK	IP190007321 1	0	4	2189	456	1.000	0.002	0.176
5	7.72	KDPVYVEEPPFGGQWVWQK	IP190046243 4	0	6	2198	456	1.000	0.002	0.177
6	7.60	KDLQFNLDDGGQVQVTFQAK	IP190046292 1	0	6	2447	456	1.000	0.003	0.177
7	7.08	RLKLPNDLDDGGQVQVTFQAK	IP190021790 1	0	7	2198	456	1.000	0.003	0.177
8	7.08	RLEKPDNLDDGGQVQVTFQAK	IP190008981 2	0	8	2145	456	1.000	0.004	0.178
9	7.07	KITQPMKQGVVDTFQVYK	IP190046299 1	0	9	2144	456	1.000	0.006	0.178
10	6.95	KRTEGQDQVDTFQVYK	IP190046299 1	0	10	2143	456	1.000	0.006	0.179
241	3.63	KVLAGDVPKATGADGGTNRK	IP190010330 3	0	241	1417	456	1.000	0.344	0.450
242	3.64	KEDDQVGGTQVPR	IP190030518 1	0	242	1411	456	1.000	0.345	0.450
243	3.64	KDVTMPPVPR	IP190030525 7	0	243	1410	456	1.000	0.346	0.460
244	3.64	KELHMPANALYVWVKA	IP190046228 1	0	244	1439	456	1.000	0.346	0.460
245	3.63	KLDQFNLDDGGQVQVTFQAK	IP190010328 1	0	245	1438	456	1.000	0.346	0.450
246	3.63	KVLAGDVPKATGADGGTNRK	IP190046292 2	2	144	1000	456	0.997	0.345	0.450
247	3.63	KVLAGDVPKATGADGGTNRK	IP190046221 1	2	247	1438	454	0.997	0.346	0.460
248	3.63	KVLAGDVPKATGADGGTNRK	IP190007003 3	2	248	1437	454	0.997	0.346	0.460
249	3.63	KVLAGDVPKATGADGGTNRK	IP190010328 1	3	249	1436	454	0.997	0.347	0.460
250	3.62	KDPVYVEEPPFGGQWVWQK	IP190046292 1	2	249	1436	454	0.997	0.347	0.461
1221	2.93	KELAGLQSTDEKE	IP190009279 5	15	1227	916	462	0.989	0.375	0.644
1222	2.92	KDVTMPPVPR	IP190030525 7	15	1222	915	462	0.987	0.375	0.644
1223	2.92	KELAGLQSTDEKE	IP190030525 7	15	1223	915	462	0.987	0.375	0.644
1224	2.92	KVLAGDVPKATGADGGTNRK	IP190010330 3	16	1230	915	463	0.986	0.375	0.643
1225	2.92	KVLAGDVPKATGADGGTNRK	IP190009279 5	16	1230	915	463	0.986	0.375	0.643
1226	2.92	KVLAGDVPKATGADGGTNRK	IP190030525 7	16	1230	915	463	0.986	0.375	0.643
1227	2.92	KVLAGDVPKATGADGGTNRK	IP190046292 2	16	1230	915	463	0.986	0.375	0.643
1228	2.92	KVLAGDVPKATGADGGTNRK	IP190010330 3	16	1230	915	463	0.986	0.375	0.643
1229	2.92	KVLAGDVPKATGADGGTNRK	IP190009279 5	16	1230	915	463	0.986	0.375	0.643
1230	2.91	KVLAGDVPKATGADGGTNRK	IP190030525 7	16	1230	915	463	0.986	0.375	0.644
1231	2.27	KVLAGDVPKATGADGGTNRK	IP190009279 5	175	1225	428	300	0.950	0.631	0.807
1232	2.27	KVLAGDVPKATGADGGTNRK	IP190030525 7	175	1225	428	300	0.950	0.631	0.807
1233	2.26	KVLAGDVPKATGADGGTNRK	IP190046292 2	175	1225	428	300	0.950	0.631	0.807
1234	2.26	KVLAGDVPKATGADGGTNRK	IP190010330 3	175	1225	428	300	0.950	0.631	0.807
1235	2.26	KVLAGDVPKATGADGGTNRK	IP190009279 5	175	1225	428	300	0.950	0.631	0.807
1236	2.26	KVLAGDVPKATGADGGTNRK	IP190030525 7	175	1225	428	300	0.950	0.631	0.807
1237	2.26	KVLAGDVPKATGADGGTNRK	IP190046292 2	175	1225	428	300	0.950	0.631	0.807
1238	2.26	KVLAGDVPKATGADGGTNRK	IP190010330 3	175	1225	428	300	0.950	0.631	0.807
1239	2.26	KVLAGDVPKATGADGGTNRK	IP190009279 5	175	1225	428	300	0.950	0.631	0.807
1240	2.26	KVLAGDVPKATGADGGTNRK	IP190030525 7	175	1225	428	300	0.950	0.631	0.807
1241	2.26	KVLAGDVPKATGADGGTNRK	IP190046292 2	175	1225	428	300	0.950	0.631	0.807
1242	2.26	KVLAGDVPKATGADGGTNRK	IP190010330 3	175	1225	428	300	0.950	0.631	0.807
1243	2.26	KVLAGDVPKATGADGGTNRK	IP190009279 5	175	1225	428	300	0.950	0.631	0.807
1244	2.26	KVLAGDVPKATGADGGTNRK	IP190030525 7	175	1225	428	300	0.950	0.631	0.807
1245	2.26	KVLAGDVPKATGADGGTNRK	IP190046292 2	175	1225	428	300	0.950	0.631	0.807
1246	2.26	KVLAGDVPKATGADGGTNRK	IP190010330 3	175	1225	428	300	0.950	0.631	0.807
1247	2.26	KVLAGDVPKATGADGGTNRK	IP190009279 5	175	1225	428	300	0.950	0.631	0.807
1248	2.26	KVLAGDVPKATGADGGTNRK	IP190030525 7	175	1225	428	300	0.950	0.631	0.807
1249	2.26	KVLAGDVPKATGADGGTNRK	IP190046292 2	175	1225	428	300	0.950	0.631	0.807
1250	2.26	KVLAGDVPKATGADGGTNRK	IP190010330 3	175	1225	428	300	0.950	0.631	0.807
1251	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1252	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1253	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1254	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1255	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1256	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1257	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1258	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1259	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1260	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1261	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1262	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1263	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1264	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1265	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1266	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1267	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1268	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1269	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1270	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1271	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1272	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1273	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1274	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1275	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1276	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1277	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1278	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1279	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1280	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1281	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1282	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1283	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1284	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1285	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1286	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1287	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1288	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1289	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1290	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1291	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1292	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1293	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1294	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1295	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	294	0.923	0.931	0.866
1296	1.94	KVLAGDVPKATGADGGTNRK	IP190030525 7	182	1229	214	294	0.923	0.931	0.866
1297	1.94	KVLAGDVPKATGADGGTNRK	IP190046292 2	182	1229	214	294	0.923	0.931	0.866
1298	1.94	KVLAGDVPKATGADGGTNRK	IP190010330 3	182	1229	214	294	0.923	0.931	0.866
1299	1.94	KVLAGDVPKATGADGGTNRK	IP190009279 5	182	1229	214	2			

## Recap – Thus Far

1. Use multiple search engines – increases coverage
2. The larger the database – the longer the search
3. The more PTMs – the longer the search
4. Large data sets require a FDR to reduce the amount of error to an acceptable level.
5. Usage of multiple enzymes should increase the amount of sequence coverage

## MudPIT

### Multidimensional Protein Identification Technology



- MudPIT is highly effective on complex mixtures
- Increased fractionations at the SCX level increases the sensitivity

Washburn, M.P. *et al*, 2001, Nat. Biotechnol. 24:2-7.



## MudPIT – Common Uses

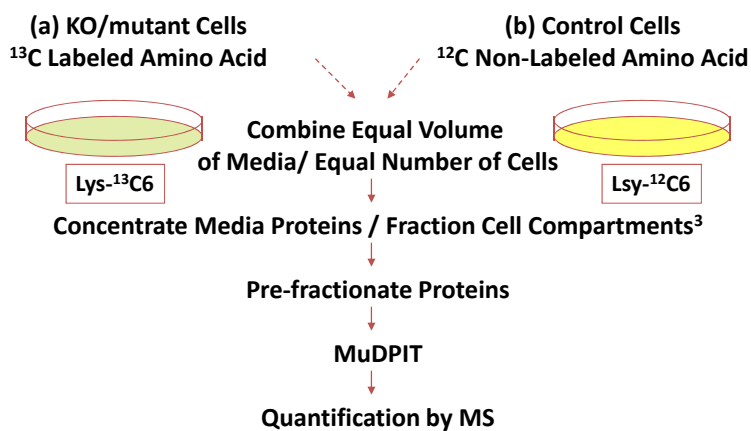
### Most commonly used for:

Highly complex mixtures/Proteomics

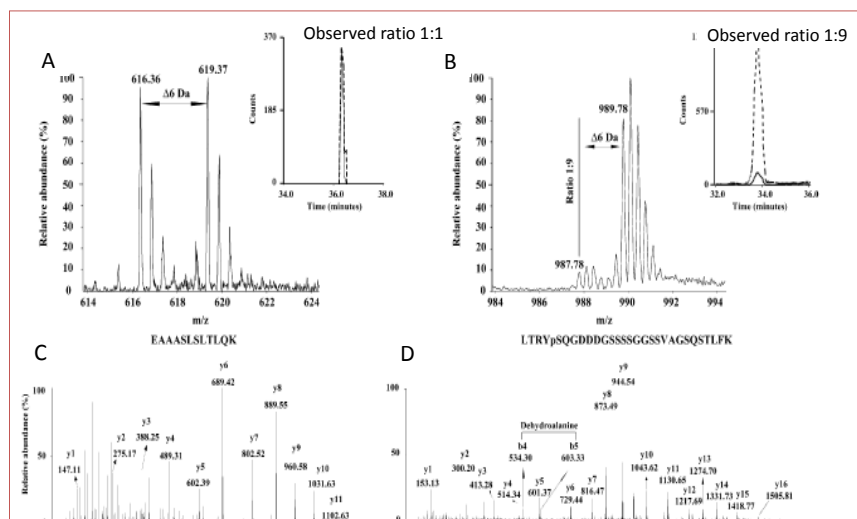
Quantitative Proteomics:

ICAT  
SILAC  
ITRAQ

## Stable Isotope Labeling with Amino Acids (SILAC)



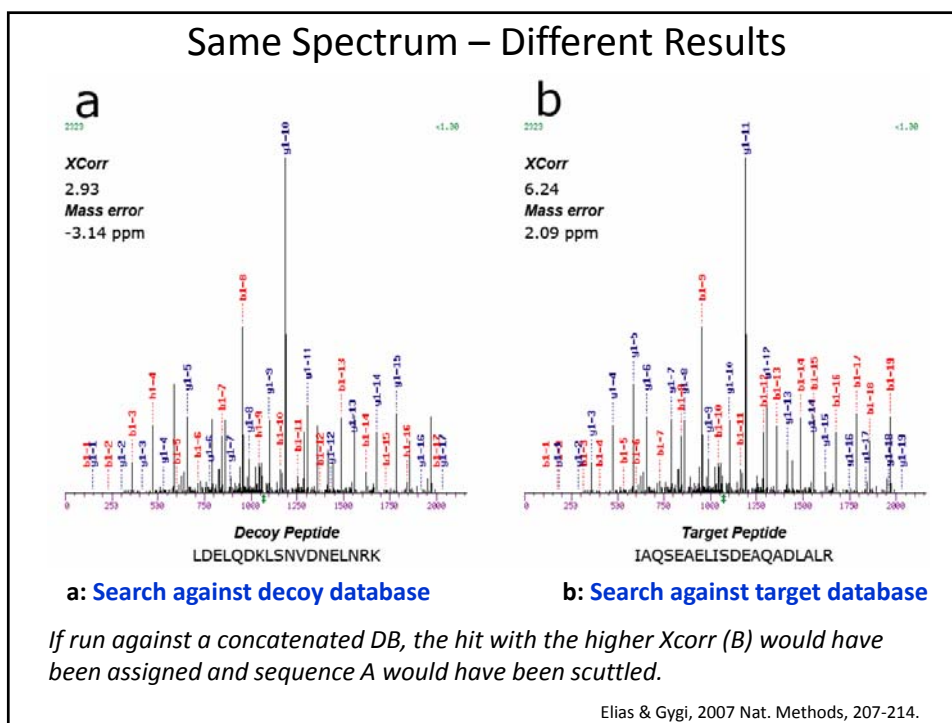
## Protein Quantitation - SILAC



SILAC pair example - Ibarrola, N. *et al*, Anal. Chem. 2003

## Useful Links!

- ▶ [i-mass.com](http://i-mass.com)
- ▶ [spectroscopynow.com](http://spectroscopynow.com)
- ▶ [expasy.ch/tools](http://expasy.ch/tools)
- ▶ [cprmap.com](http://cprmap.com)
- ▶ [psidev.sourceforge.net](http://psidev.sourceforge.net)
- ▶ [prospector.ucsf.edu](http://prospector.ucsf.edu)
- ▶ [jeolusa.com/ms/docs/ionize.html](http://jeolusa.com/ms/docs/ionize.html)
- ▶ [asms.org](http://asms.org) (become a member!)
- ▶ [hupo.org](http://hupo.org)
- ▶ [matrixscience.com](http://matrixscience.com)
- ▶ [proteomecenter.org/software.php](http://proteomecenter.org/software.php)
- ▶ [UNIMOD.org](http://UNIMOD.org)
- ▶ [ionsource.com](http://ionsource.com)
- ▶ [bruker.com](http://bruker.com)
- ▶ [thermo.com](http://thermo.com)
- ▶ [appliedbiosystems.com](http://appliedbiosystems.com)
- ▶ [shimadzu.com](http://shimadzu.com)
- ▶ [luminexcorp.com](http://luminexcorp.com)



## Sequest Discriminant Scoring – TPP/ISB

variable	[M + 2H] <sup>2+</sup>		[M + 3H] <sup>3+</sup>	
	coefficient	correlation	coefficient	correlation
Xcorr'	8.362	0.798	9.933	0.698
$\Delta C_n$	7.386	0.746	11.149	0.806
ln SpRank	-0.194	-0.510	-0.201	-0.491
$\Delta Mass$	-0.314	-0.306	-0.277	-0.251
constant	-0.959		-1.460	

$$\text{Discriminant score} = 8.362(\text{Xcorr}') + 7.386(\Delta C_n) - 0.194 (\ln(\text{Sp Rank}) - 0.314 (\Delta \text{mass}) - 0.959$$

**Xcorr'** normalizes Xcorr for peptide length

Keller, A. et al 2002 Anal. Chem. p 5383-5392