Bioinformatics/Search Engines/FDRs

Archer Smith IV, PhD
Department of Biochemistry and Molecular Genetics
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

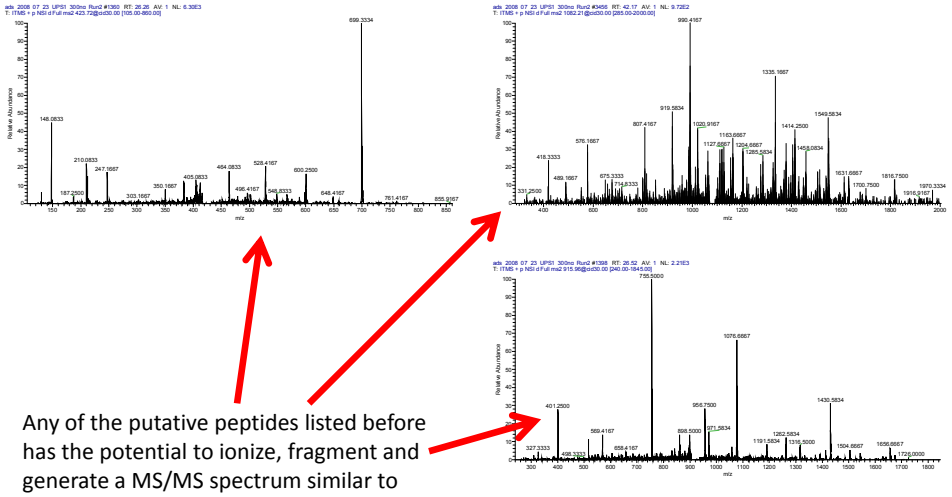
MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQC
PFDEHVKLVNELTEFAKTCVADESHAGCEKSLHTLFGDELCKVASLRETYGDMADC
CEKQEPERNECFLSHKDDSPDLPKLKPDPNTLCDEFKADEKKFWGKYLYEIARRHP
YFYAPELLYYANKYNGVFQECCQAEDKGACLLPKIETMREKVLASSARQRLRCASIQ
KFGERALKAWSVARLSQKFPKAEFVEVTKLVTDLTKVHKECCHGDLLECADDRADL
AKYICDNQDTISSKLKECCDKPLLEKSHCIAEVEKDAIPENLPPLTADFAEDKDVCKN
YQEAKDAFLGSFLYEYSRRHPEYAVSVLLRLAKEYEATLEECCAKDDPHACYSTVFDK
LKHLVDEPQNLIKQNCDQFEKLGEYGFQNALIVRYTRKVPQVSTPTLVEVSRSLGKV
GTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTKCCTESLVNRRPCFSAL
TPDETYVPKAFDEKLFTFHADICTLPDTEKQIKKQTALVELLKHKPKATEEQLKTVME
NFVAFVDKCCAADDKEACFAVEGPKLVVSTQTALA

Digest with trypsin

First 40 of 114 Possible Peptides

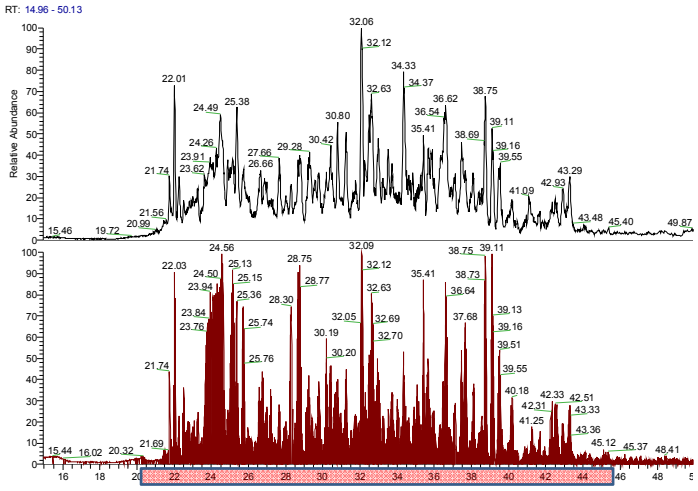
#	m/z (mi)	m/z (av)	Start	End	Missed Cleavages	Sequence
1	817.489	817.9693	452	459	1	(R)SLGKVGTR(C)
2	818.4254	818.9074	562	568	0	(K)ATEEQLK(T)
3	820.4676	820.9726	229	235	1	(K)FGERALK(A)
4	839.3022	839.9267	581	587	0	(K)ccAADDK(E)
5	847.5036	848.0391	242	248	1	(R)LSQKFPK(A)
6	886.4153	886.9394	131	138	0	(K)DDSPDLPK(L)
7	898.4815	899.1065	483	489	0	(R)LcVLHEK(T)
8	906.4713	907.083	205	211	1	(K)IETMREK(V)
9	922.488	923.06	249	256	0	(K)AEFVEVTK(L)
10	927.4934	928.0823	161	167	0	(K)YLYEIAR(R)
11	960.5473	961.1126	210	218	1	(R)EKVLASSAR(Q)
12	974.4578	975.0521	37	44	0	(K)DLGEEHFK(G)
13	975.5404	976.1929	221	228	1	(R)LRcASIQK(F)
14	987.537	988.1411	29	36	1	(K)SEIAHRFK(D)
15	987.5694	988.1414	212	220	1	(K)VLASSARQR(L)
16	988.5673	989.1637	490	498	1	(K)TPVSEKVTK(C)
17	1001.5891	1002.2118	233	241	1	(R)ALKAWSVAR(L)
18	1002.583	1003.1908	598	607	0	(K)LVVSTQTALA(-)
19	1014.6194	1015.2455	549	557	0	(K)QTALVELLK(H)
20	1034.4724	1035.1736	123	130	0	(R)NEcFLSHK(D)
21	1068.4415	1069.1446	413	420	0	(K)QNcDQFEK(L)
22	1072.5092	1073.2201	310	318	0	(K)SHcIAEVEK(D)
23	1083.5946	1084.2709	161	168	1	(K)YLYEIARR(H)
24	1107.5139	1108.2658	588	597	0	(K)EAcFAVEGPK(L)
25	1138.498	1139.3022	499	507	0	(K)ccTESLVNR(R)
26	1142.7143	1143.4207	548	557	1	(K)KQTALVELLK(H)
27	1145.6426	1146.3427	236	245	1	(K)AWSVARLSQK(F)
28	1153.6939	1154.4035	257	266	1	(K)LVTDLTKVHK(E)
29	1163.6307	1164.3521	66	75	0	(K)LVNELTEFAK(T)
30	1166.4929	1167.3127	460	468	0	(R)ccTKPESER(M)
31	1193.6022	1194.3001	25	34	1	(R)DTHKSEIAHR(F)
32	1195.5888	1196.3788	223	232	1	(R)cASIQKFGER(A)
33	1249.6212	1250.4051	35	44	1	(R)FKDLGEEHFK(G)
34	1254.5783	1255.4003	337	346	1	(K)DVcKNYQEAK(D)
35	1283.7106	1284.51	361	371	0	(R)HPEYAVSVLLR(L)
36	1291.6021	1292.5249	300	309	0	(K)EccDKPLLEK(S)
37	1294.7042	1295.5305	246	256	1	(K)FPKAEFVEVTK(L)
38	1305.7161	1306.5136	402	412	0	(K)HLVDEPQNLIK(Q)
39	1308.727	1309.5172	558	568	1	(K)HKPKATEEQLK(T)
40	1388.7389	1389.7327	198	209	1	(K)GAcLLPKIETMR(E)

Examples of MS/MS spectra



Any of the putative peptides listed before has the potential to ionize, fragment and generate a MS/MS spectrum similar to these examples

Example Chromatogram - Tryptic Digest of a Complex Mixture

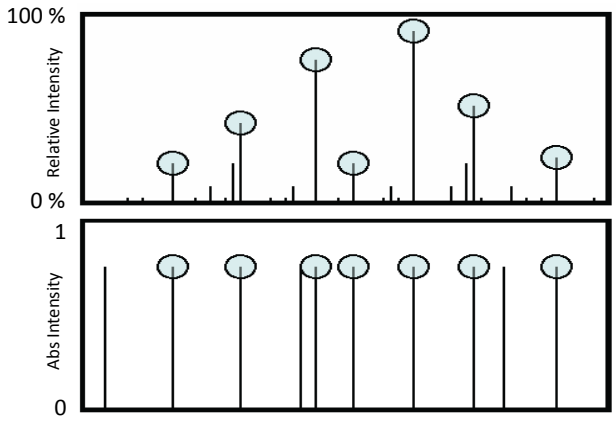


25 min ~ 3000 Spectrums Acquired

Data Analysis - Pt. 1

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

Use Search Engines To Cross-correlate And Analyze The Data

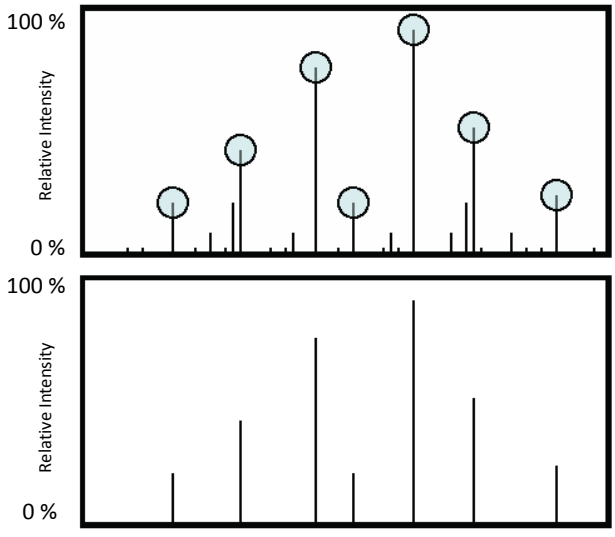


Acquired Spectrum

Hypothetical Spectrum

Adapted from Searle, B.C. Proteome Software Inc

X!Tandem Algorithm



Acquired Spectrum

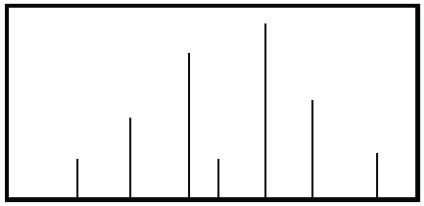
Only Overlapping Peaks Remain

- Only the matching peaks are considered
- Spectrum is now simplified to model spectrum

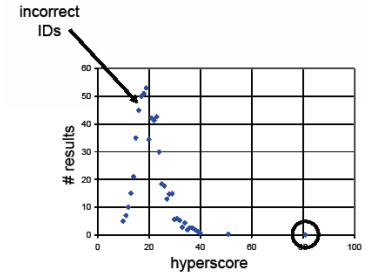
Adapted from Searle, B.C. Proteome Software Inc

X!Tandem Hyper-score

Hyper-score = sum(I_{obs} * P_{ints}) * Nb!*Ny!



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!Tandem

Sequest

MATRIX SCIENCE

Mascot

Statistics/Visualization:

Protein
Prophet

BIOINQUIRE

PROTEOIQ

Scaffold 2

Scaffold

Further Refinement

Search Engine Outputs Vary

X!Tandem	**Sequest**	**Mascot**	**Protein Prospector**
Hyperscore	Xcorr	Mascot Score	Peptide Score
E-value	Sp		Expectation Value
	Sf		
	ΔCn		

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)

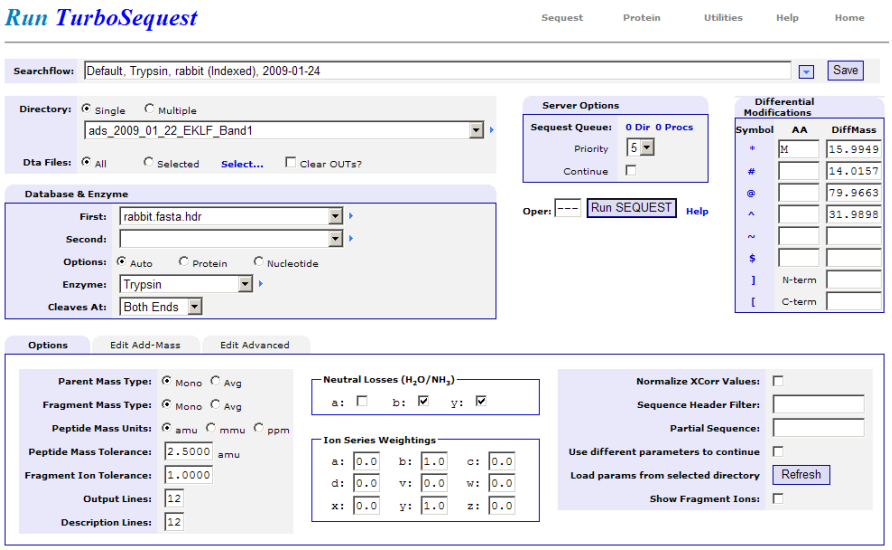
Example Parameters:

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

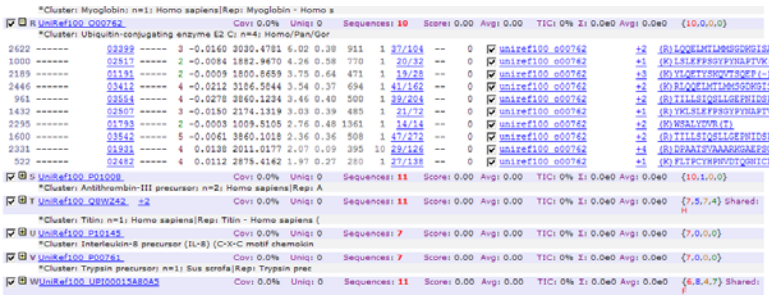
Sequest Example



Eng , J.K. *et al* 1994 JASMS ***976 – 989 .***

Sequest Output File

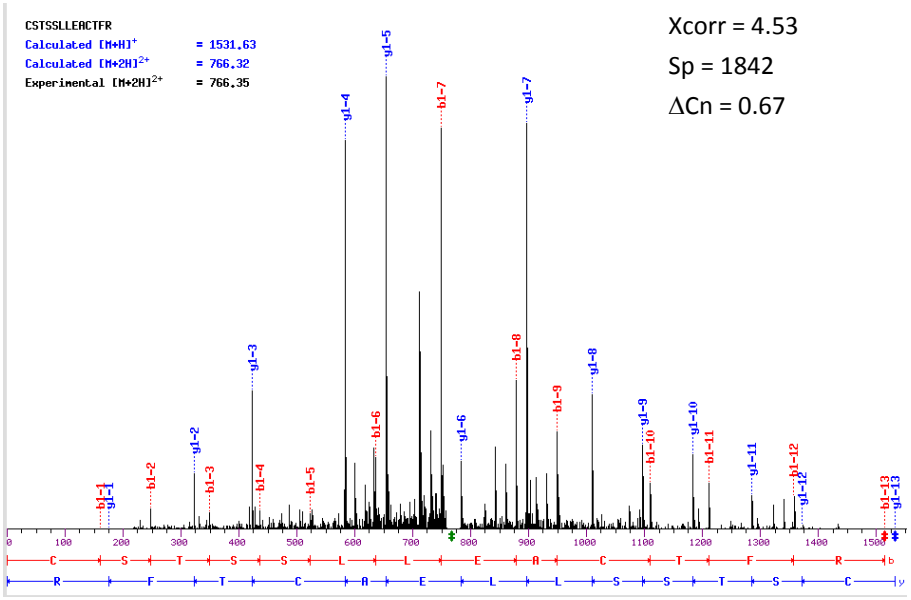
Xcorr
ΔCn
Sp



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

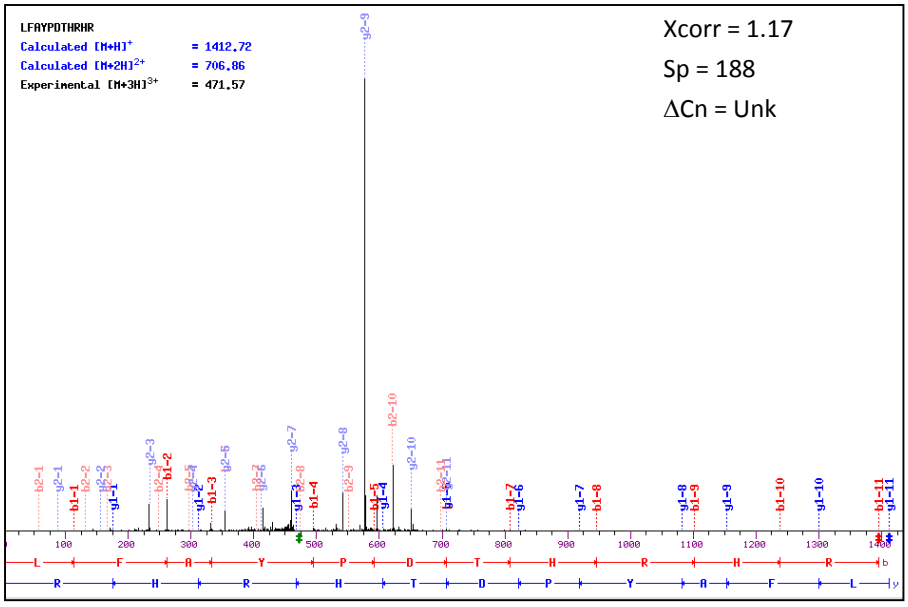
- Cutout represents 6 protein groups out of **> 200** putative protein groups
- Peptide hits are organized by highest to lowest Xcorr value

Good Match



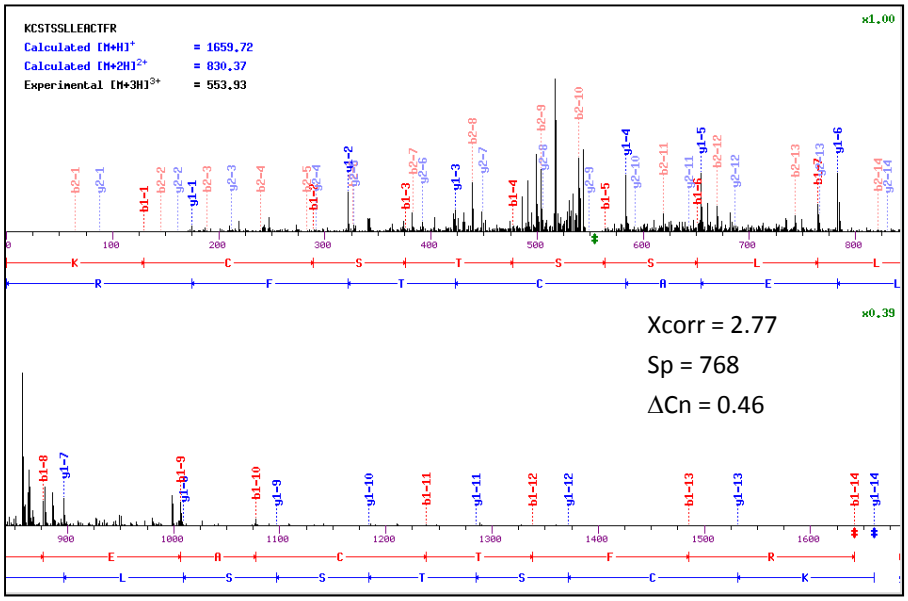
Serotransferrin precursor – 2008_07_23UPS1_Run2_FT

Cross Correlation Is Probably Not Enough For Assignment



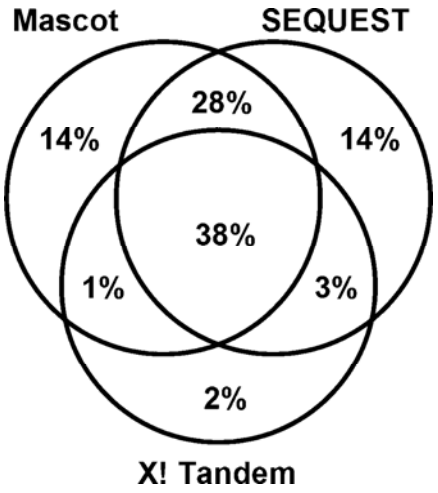
Catalase – 2008_07_23UPS1_Run2_FT

More Difficult Interpretation



Serotransferrin precursor – 2008_07_23UPS1_Run2_FT

Take Advantage Of Multiple Search Engines



Search engine specific Venn diagram of multiple runs of an 18 protein mixture

No single search engine available today is perfect!

Searle, B.C. *et al*, 2008 *Proteome Res.*, 245-253.

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

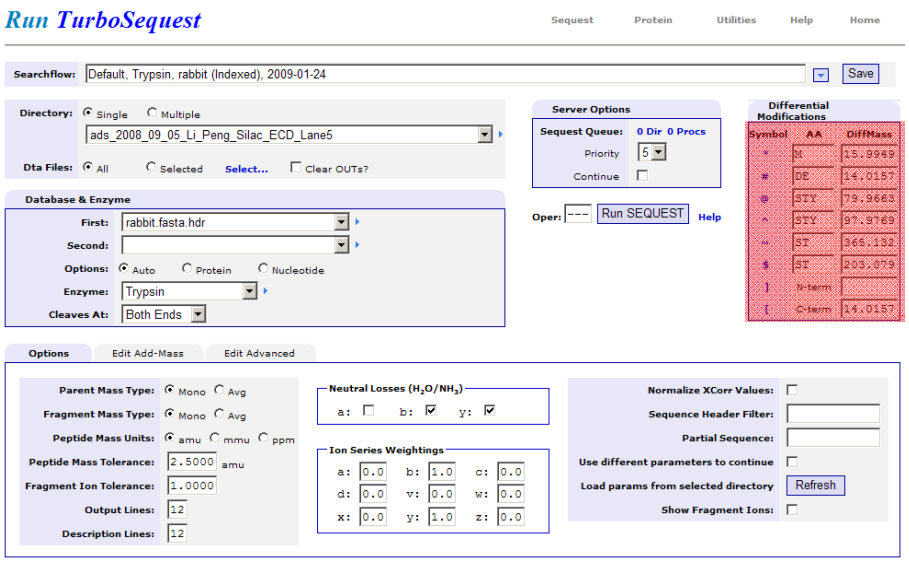
•On the peptide level, some PTM's inhibit "traditional" ionization techniques

•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



There are a large amount of known PTMs available at 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.

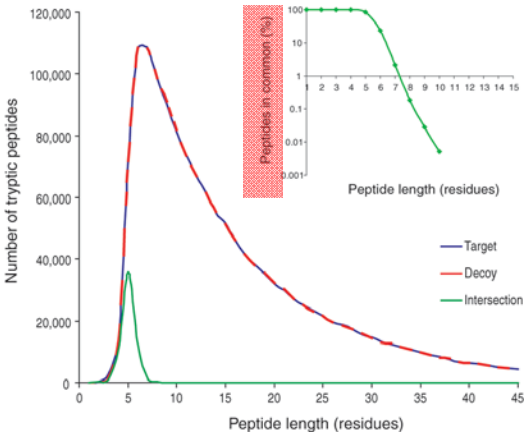
Background: The decoy database strategy allows one to "estimate" the number of false positives in the target database by using the number of correct hits in the decoy database as a metric – **calculate a False Discovery Rate (FDR)**

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 – 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

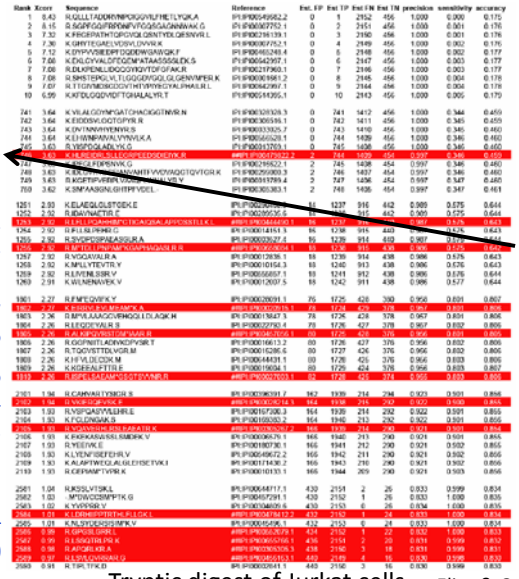


Analysis comparing the statistical possibility of there being the same peptide in both databases as a function of peptide length

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

Hits Arranged by Xcorr

Xcorr



Red lines indicate peptides from decoy database

3.63
2.92
2.92
2.27
2.26
2.26
1.94
1.93
1.01
<1.0

Relatively high Xcorr value for a false positive

Tryptic digest of Jurkat cells

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

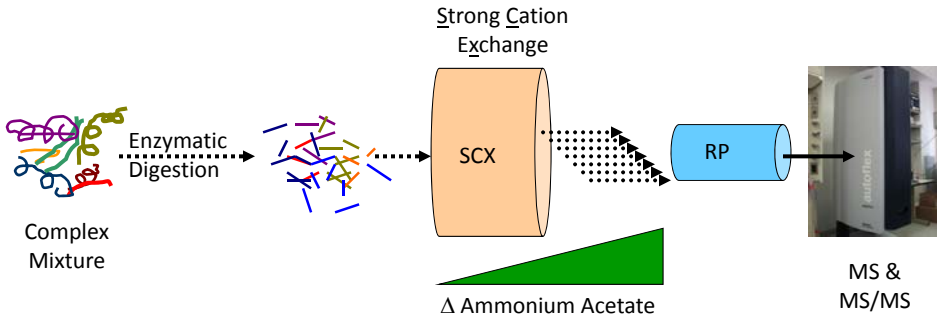
Recap – Thus Far

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. 242-7.

MudPIT – Common Uses

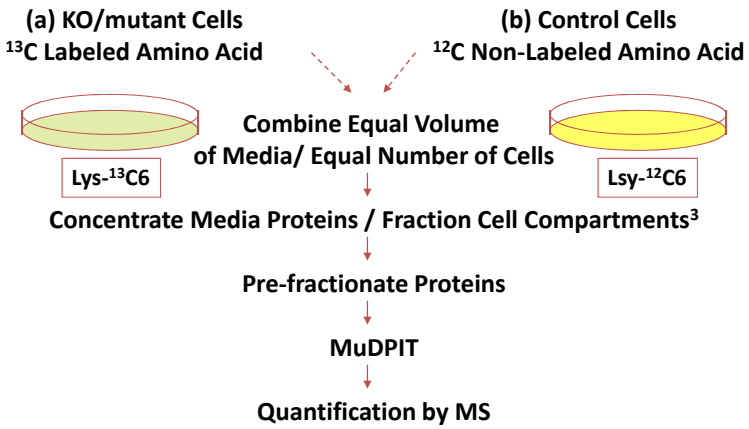
Most commonly used for:

Highly complex mixtures/Proteomics

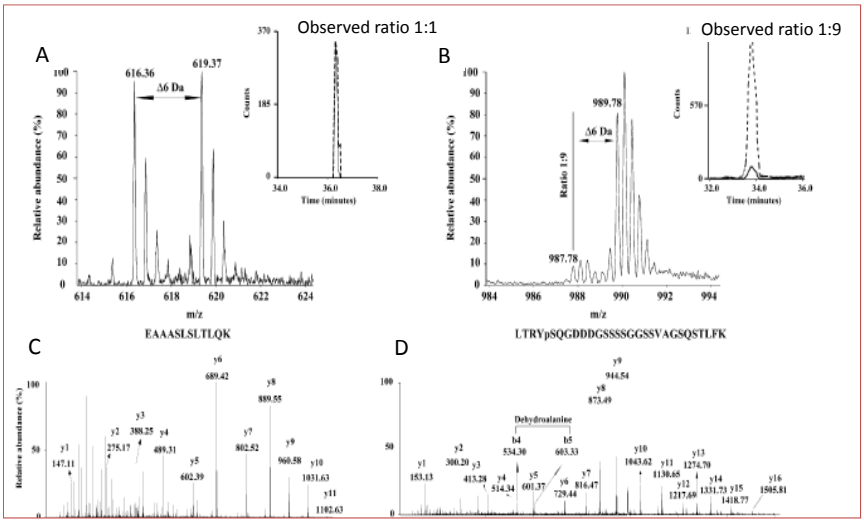
Quantitative Proteomics:

ICAT
SILAC
ITRAQ

Stable **I**sotope **L**abeling with **A**mino Acids (SILAC)



Protein Quantitation - SILAC



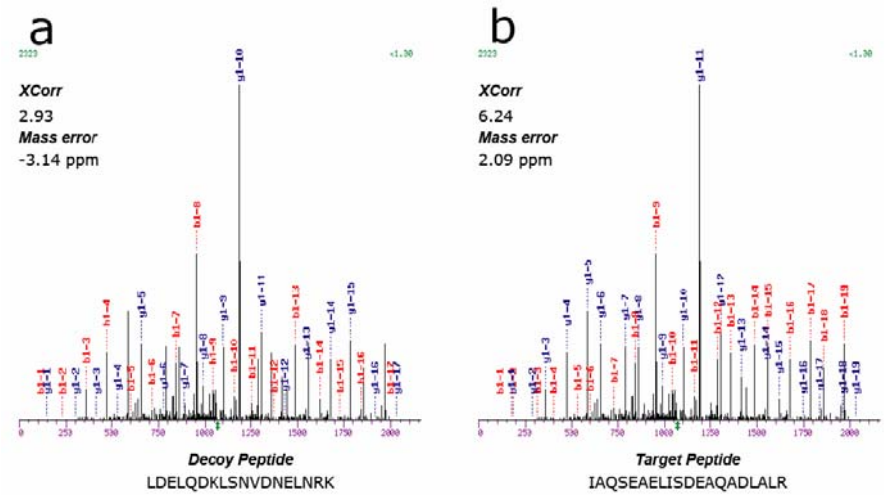
SILAC pair example - Ibarrola, N. *et al*, Anal. Chem. 2003

Useful Links!

- i-mass.com
- spectroscopynow.com
- expasy.ch/tools
- cprmap.com
- psidev.sourceforge.net
- prospector.ucsf.edu
- jeolusa.com/ms/docs/ionize.html
- asms.org (become a member!)
- hupo.org
- matrixscience.com
- proteomecenter.org/software.php
- UNIMOD.org

ionsource.com
bruker.com
thermo.com
appliedbiosystems.com
shimadzu.com
luminexcorp.com

Same Spectrum – Different Results



a: Search against decoy database **b: Search against target database**

If run against a concatenated DB, the hit with the higher Xcorr (B) would have been assigned and sequence A would have been scuttled.

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

Sequest Discriminant Scoring – TPP/ISB

	$[M + 2H]^{2+}$		$[M + 3H]^{3+}$	
variable	**coefficient**	**correlation**	**coefficient**	**correlation**
Xcorr'	8.362	0.798	9.933	0.698
ΔC_n	7.386	0.746	11.149	0.806
ln SpRank	−0.194	−0.510	−0.201	−0.491
ΔMass	−0.314	−0.306	−0.277	−0.251
constant	−0.959		−1.460	

Discriminant score = 8.362(Xcorr') +7.386(ΔCn) – 0.194 (ln(Sp Rank) – 0.314 (Δmass)-0.959

Xcorr' normalizes Xcorr for peptide length

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