INSTITUTE FOR Systems Biology™

Revolutionizing science. Enhancing life.

Proteomics Informatics Course Syllabus

PeptideAtlas

Trans Proteomic Pipeline

Human SRMAtlas

Seattle Proteome Center
Institute for Systems Biology

Seattle, Washington
October 25 – October 29, 2010

National Institutes of Health

National Institute of General Medical Sciences

National Human Genome Research Institute

Duchy of Luxembourg

Proteomics Informatics Course Syllabus

Seattle Proteome Center
Institute for Systems Biology
October 25th- October 29th, 2010

Instructors:
Rob Moritz
Luis Mendoza
Eric Deutsch
David Shteynberg
Mi-Youn Brusniak
Rich Johnson
Joe Slagel
Hector Ramos
Jeff Ranish

www.proteomecenter.org/

Seattle Proteome Center

Institute for Systems Biology
Proteomics Informatics Course
October 25 – October 29, 2010
Lakeview Facility

Time	Day One – October 25	
9:00 – 9:30	**Introduction** Robert Moritz, ISB Faculty	1
9:30 – 10:30	**Fundamentals of MS/MS Interpretation** Richard Johnson	7
10:30 – 1:00	**Sequence Database Searching** Luis Mendoza	13
1:00 - 2:00	**Catered lunch**	
2:00 – 2:30	**GUI Intro** Luis Mendoza	
2:30 - 3:00	**Data Formats and Conversion** Luis Mendoza	23
3:00 - 3:30	**Pep3D** Luis Mendoza	31
3:30 - 5:00	**Tutorials** Luis Mendoza	40
5:00 – 6:00	**Installation and Support**	47
6:00	**Clinic**	

Time	Day Two – October 26	
9:00 – 12:30	**PeptideProphet** David Shteynberg	**51**
12:30 - 2:00	**Lunch Break**	
2:00 - 3:30	**PeptideProphet** David Shteynberg	
3:30 – 5:30	**InterProphet** David Shteynberg	**73**
5:30 – 6:00	**QualScore** Luis Mendoza	**81**
6:00	**Clinic**	

Time	Day Three – October 27	
9:00 – 12:30	**ProteinProphet** Luis Mendoza	**87**
12:30 – 2:00	**Lunch Break**	
2:00 - 5:00	**SpectraST** Eric Deutsch	**105**
5:00 - 6:00	**TPP on the Cloud** Joe Slagel	**123**
6:00	**Clinic**	

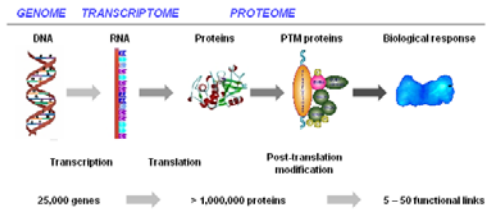
Time	Day Four – October 28	
9:00 – 10:30	**Quantitative Proteomics Applications** Jeff Ranish, ISB Faculty	
10:30 – 12:30	**Xpress & ASAPRatio** David Shteynberg	**129**
12:30 – 2:00	**Lunch break**	
2:00 – 3:00	**ASAPRatio** David Shteynberg	
3:00 – 4:00	**Libra** Luis Mendoza	**147**
4:00 – 6:00	**Corra** Mi-Youn Brusniak	**153**
6:00	**Clinic**	

Time	Day Five – October 29	
9:00 – 12:30	**SBEAMS, PeptideAtlas, & SRMAtlas** Eric Deutsch	**171**
12:30 – 2:00	**Lunch Break**	
2:00 – 3:00	**PIPE2** Hector Ramos	**187**
3:00 – 5:00	**TIQAM & ATAQS** Mi-Youn Brusniak	**207**
5:00	**BBQ**	

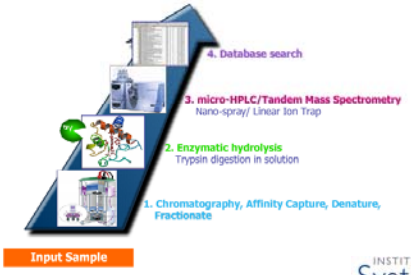
Introduction

Robert Moritz
Day 1
October 25, 2010

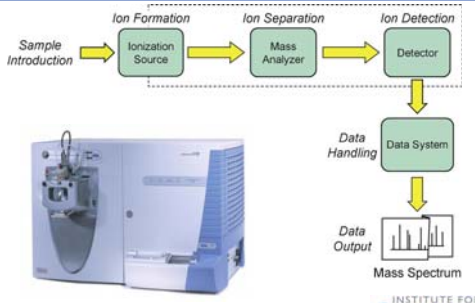
Proteomics – the study of protein identities



Proteome Pipeline - analysis by mass spectrometry



Schematic of a Mass Spectrometer



Data Interrogation:

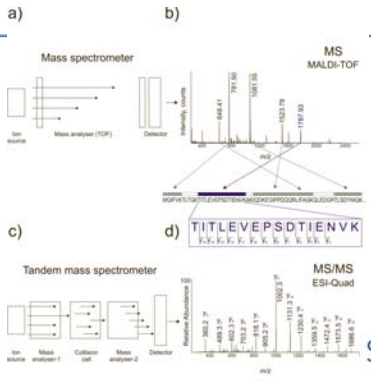
The Achilles heel of proteomics!

Fundamentals of Proteomics

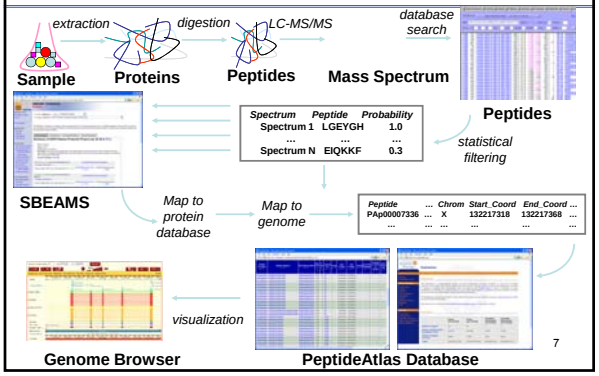
~~Modern~~ current day

Protein identification and quantitation

Digital Information of Proteomics

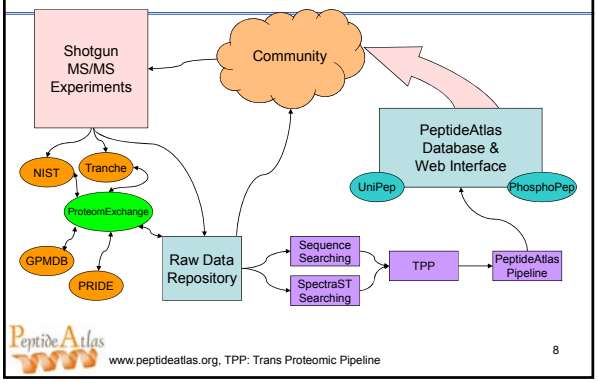


From Peptides to Genome Annotation



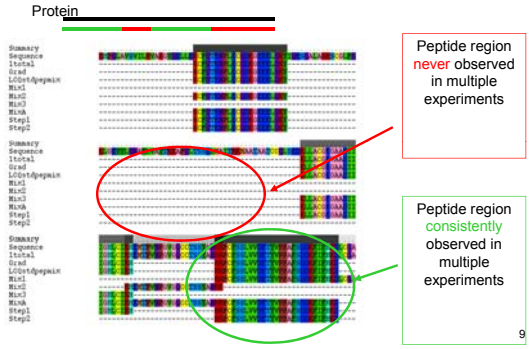
7

PeptideAtlas Workflow



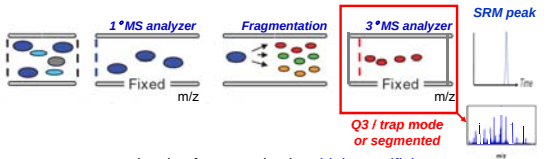
8

What is a proteotypic peptide?



9

Selected/Multiple Reaction Monitoring (S/MRM)

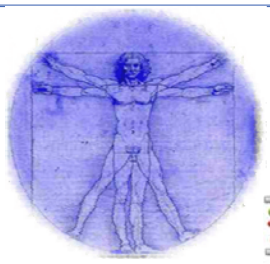


- *two levels of mass selection: high specificity*
- *not scanning (Q1/Q3 static), high duty cycle: high sensitivity*
- *the most sensitive mass spectrometry method known (low amole)*

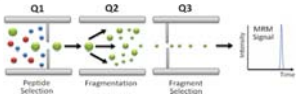
....you need to know what to look for!

(the mass spectrometrist's ELISA)

Human SRMAtlas

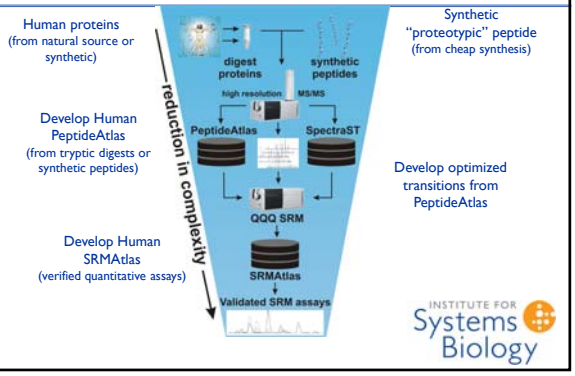


20,333 proteins (20,277 2010 version)
32,562 proteins incl. isoforms
658,684 tryptic peptides (any length)
480,284 distinct peptides (7-30aa)



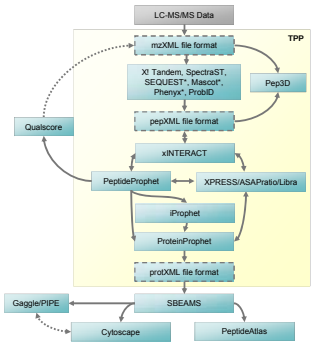
INSTITUTE FOR Systems Biology

Developments at ISB - SRMAtlas



INSTITUTE FOR Systems Biology

What is the Trans-Proteomic Pipeline (TPP)?



The TPP is a open-source and free collection of ***tools*** and supporting ***data formats*** which enable *shotgun proteomics* data analysis.

13

Proteomics Informatics Course Agenda

5 Days of fun

- Fundamentals of MS/MS Interpretation
- Sequence Database Searching, Data Formats and Conversion
- Installation and Support, Clinic
- PeptideProphet, InterProphet, ProteinProphet
- QualScore
- SpectraST
- TPP on the Cloud
- Quantitative Proteomics Applications
- Xpress & ASAPRatio, Libra
- Non-labeling quantitation - Corra
- SBEAMS, PeptideAtlas, & SRMAtlas
- PIPE2
- SRM workflow tools - TIQAM & ATAQS

INSTITUTE FOR Systems Biology

Manual MS/MS Data Interpretation

Rich Johnson

Day 1

October 25, 2010

Outline

1. **Background**
   - **Why bother?**
   - **Types of fragment ions**
   - **Mobile protons**
   - **Traps versus quadrupole collision cells**
   - **Annoying things to remember when sequencing peptides by MS**
2. **Examples:**
   a. **Sequencing from the middle and working towards the termini**
   b. **Sequencing from the C-terminus and working towards the N-terminus**

Additional information at www.hairyfatguy.com (follow the Lutefisk link).

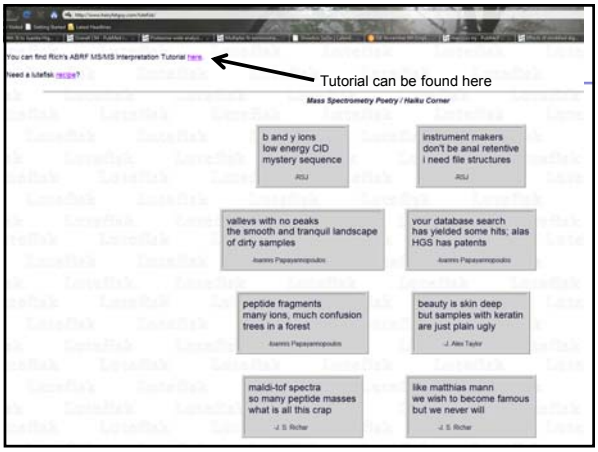
Welcome to

hairyfatguy.com

~~Where do you want to go today?~~
Shut up and get in the car.

Click on Lutefisk link

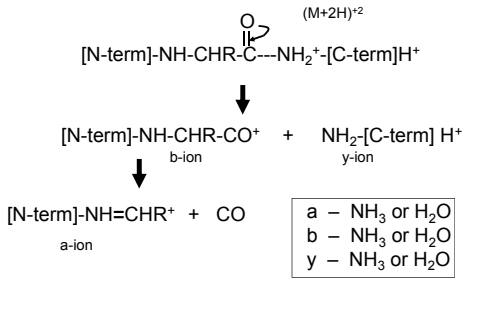
Sherpa | Lutefisk



Why bother interpreting MS/MS spectra?

- unsequenced genomes (e.g., I once identified a protein from an unsequenced species of mycoplasma)
- validate a database match (particularly important for protein identifications based on a single peptide)
- sometimes people are curious about unmatched spectra (e.g., I once found that I had a carbamylation problem; carryover problems)
- automated *de novo* software helps identify high quality spectra (find high quality peptide spectra that did not have a database match)
- it is good to look at raw data occasionally in order to develop a sense of mass spectrometric aesthetics (i.e., is your data good or bad?)

Sequence-specific fragment ions



Non-sequence-specific fragmentations

Aaa-Bbb-Ccc-Ddd-Eee-Fff-Ggg-Hhh$^+$ (a generic peptide ion)

H_2N-CHR$_{Ccc}$-CO-NH-CHR$_{Ddd}$-CO$^+$

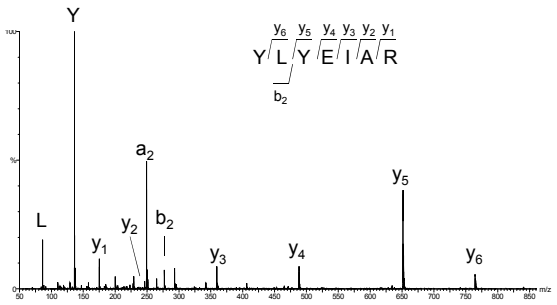
Immonium ions

$H_2N{=}CHR_{Ccc}{}^+$

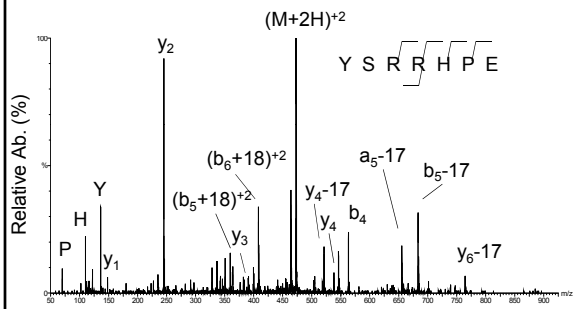
Why does everyone analyze positively-charged tryptic peptides?

- Usually better sensitivity from positively-charged peptide ions.
- "Mobile protons" protonate peptide bonds and promote b/y fragmentation
  - Arg sequesters protons in gas phase
  - Tryptic peptides typically have 0 -1 Arg
  - Tryptic peptide ions typically have two protons
  - Therefore, tryptic peptides usually have b/y ions
- Placing Arg's at the C-terminus makes it more likely that a complete series of y-ions will be observed.

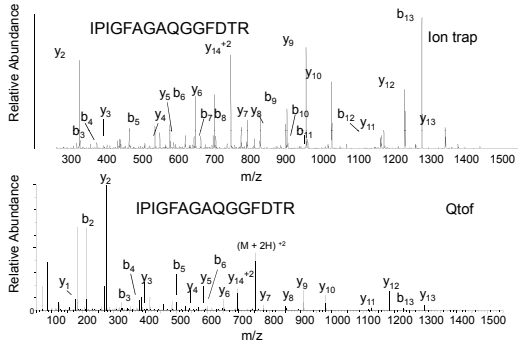
MS/MS spectrum of doubly-charged tryptic peptide (one Arg and two protons)



MS/MS spectrum of a doubly-charged non-tryptic peptide(two Arg's and two protons)



CID in traps vs quadrupoles



Annoying things to remember when sequencing peptides by MS/MS

- Leucine and isoleucine have the same mass
- Glutamine and lysine differ by 0.036 u
- Phenylalanine and oxidized methionine differ by 0.033 u
- Cleavages do not occur at every bond (more often than not, there is no cleavage between the first and second residues)
- Certain amino acids have the same mass as pairs of other amino acids: G + G = N, A + G = Q, G + V ~ R, A + D ~ W, S + V ~ W
- However: mass accuracy resolves many of these ambiguities

c and z ions result from ETD

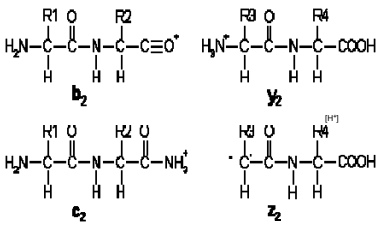


Table 1: Mass and abundance values for some biochemically relevant elements.

Element	Average Mass	Isotope	Monoisotopic Mass	Abundance (%)
Hydrogen	1.008	^{1}H	1.00783	99.985
		^{2}H	2.01410	0.015
Carbon	12.011	^{12}C	12	98.90
		^{13}C	13.00335	1.10
Nitrogen	14.007	^{14}N	14.00307	99.63
		^{15}N	15.00011	0.37
Oxygen	15.999	^{16}O	15.99491	99.76
		^{17}O	16.99913	0.04
		^{18}O	17.99916	0.200
Phosphorus	30.974	^{31}P	30.97376	100
Sodium	22.990	^{23}Na	22.98977	100
Sulfur	32.064	^{32}S	31.97207	95.02
		^{33}S	32.97146	0.75
		^{34}S	33.96787	4.21
		^{36}S	35.96708	0.02

Table 2: Amino acid residue masses (-NH-CHR-CO-)

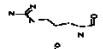
Residue	3-letter code	1-letter code	Mono-isotopic mass	Average mass	Structure
Alanine C_3H_5NO	Ala	A	71.03712	71.08	
Arginine $C_6H_{12}N_4O$	Arg	R	156.10112	156.19	
Asparagine $C_4H_6N_2O_2$	Asn	N	114.04293	114.10	
Aspartic acid $C_4H_5NO_3$	Asp	D	115.02695	115.09	
Asn or Asp	Asx	B			
Cysteine C_3H_5NOS	Cys	C	103.00919	103.14	
Glutamic acid $C_5H_7NO_3$	Glu	E	129.04260	129.12	
Glutamine $C_5H_8N_2O_2$	Gln	Q	128.05858	128.13	
Glu or Gln	Glx	Z			
Glycine C_2H_3NO	Gly	G	57.02147	57.05	
Histidine $C_6H_7N_3O$	His	H	137.05891	137.14	
Isoleucine $C_6H_{11}NO$	Ile	I	113.08407	113.16	

Table 2 continued

Residue	3-letter code	1-letter code	Mono-isotopic mass	Average mass	Structure
Leucine $C_6H_{11}NO$	Leu	L	113.08407	113.16	
Lysine $C_6H_{12}N_2O$	Lys	K	128.09497	128.17	
Methionine C_5H_9NOS	Met	M	131.04049	131.19	
Phenylalanine C_9H_9NO	Phe	F	147.06842	147.18	
Proline C_5H_7NO	Pro	P	97.05277	97.12	
Serine $C_3H_5NO_2$	Ser	S	87.03203	87.08	
Threonine $C_4H_7NO_2$	Thr	T	101.04768	101.10	
Selenocysteine C_3H_5NOSe	SeC	U	150.95364	150.03	
Tryptophan $C_{11}H_{10}N_2O$	Trp	W	186.07932	186.21	
Tyrosine $C_9H_9NO_2$	Tyr	Y	163.06333	163.18	
Unknown	Xaa	X			
Valine C_5H_9NO	Val	V	99.06842	99.13	

unmodified peptides and 43.0184 Da for acetylated N-terminus). [C] is the mass of the C-terminus (e.g., 17.0027 Da for unmodified peptides and 16.0187 Da for amidated C-terminus). [M] is the sum of the amino acid residue masses (see Table 2.1) that are contained within the fragment ion. CO is the combined mass of oxygen plus carbon atoms (27.9949 Da) and H is the mass of a proton (1.0078 Da). To calculate the m/z value of a fragment ion, add the mass of the protons to the neutral mass calculated from the table, and divide by the number of protons added.

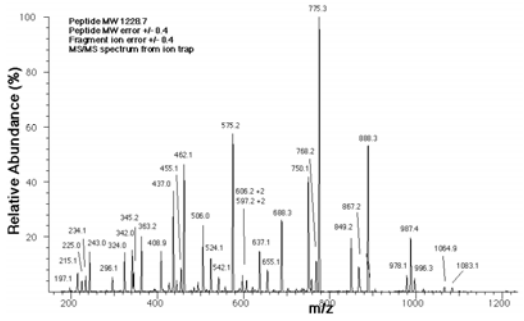
Ion Type	Neutral MW of the fragment
a	[N] + [M] - CO - H
a-H_2O	a – 18.0106
a-NH_3	a – 17.0266
b	[N] + [M] - H
b- H_2O	b – 18.0106
b- NH_3	b – 17.0266
c	[N] + [M] + NH2
d	a – partial side chain
x	[C] + [M] + CO - H
y	[C] + [M] + H
y- H_2O	y – 18.0106
y-NH_3	y – 17.0266
z	[C] + [M] - NH
v	y – complete side chain
w	z – partial side chain

Two approaches to manually sequencing peptides from MS/MS spectra

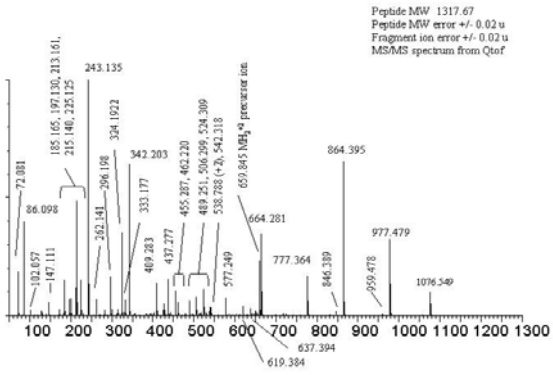
1. Finding a series of ions in the middle of the peptide, and working out towards one of the termini (illustrated using ion trap data)

2. Finding the C-terminus and working towards the N-terminus (illustrated using qtof data)

Derive a sequence from this spectrum:



Now derive a sequence from this spectrum:



MS/MS Database Searching

Luis Mendoza

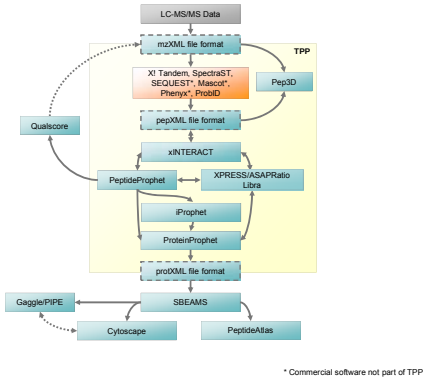
Day 1

October 25, 2010

Lecture topics

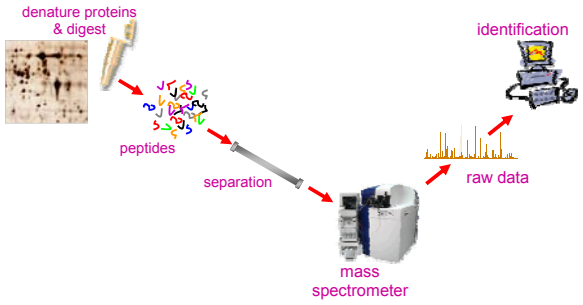
- Basic background & motivation
- Peptide fragmentation, nomenclature
- Peptide vs. tandem mass spectra
- MS/MS sequence database searching and search tools
- Interpretation of search results

1



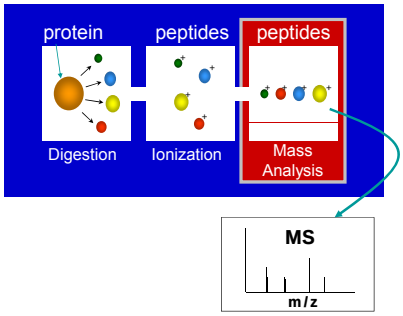
2

General MS-based proteomics workflow



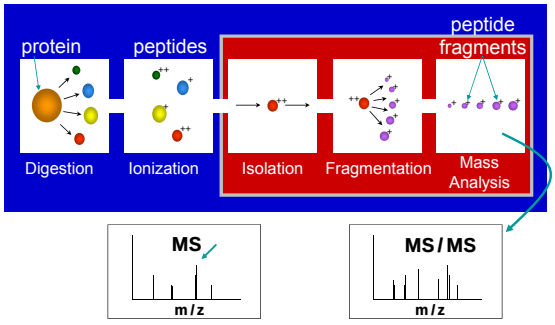
3

Single stage MS



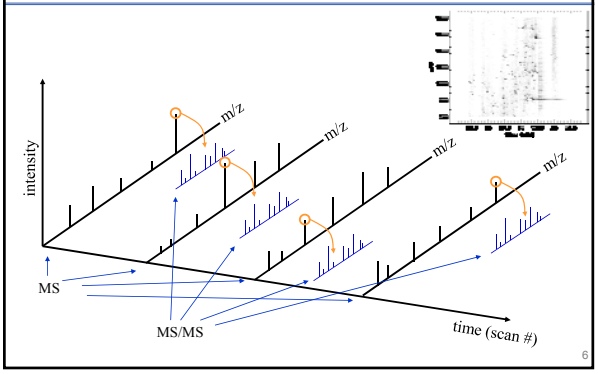
4

Tandem MS



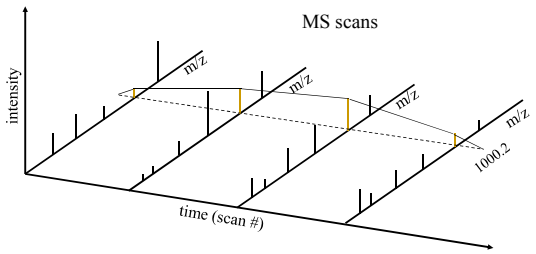
5

MS vs. MS/MS



6

Mass vs. Intensity vs. Time



7

Amino acids

amino acid	code	composition	mono mass	avg mass
glycine	G	C_2H_3NO	57.021463724	57.05132
alanine	A	C_3H_5NO	71.037113788	71.0779
serine	S	$C_3H_5NO_2$	87.032028410	87.0773
proline	P	C_5H_7NO	97.052763852	97.11518
valine	V	C_5H_9NO	99.068413916	99.13106
threonine	T	$C_4H_7NO_2$	101.047678474	101.10388
cysteine	C	C_3H_5NOS	103.009184478	103.1429
leucine	L	$C_6H_{11}NO$	113.084063980	113.15764
isoleucine	I	$C_6H_{11}NO$	113.084063980	113.15764
asparagine	N	$C_4H_6N_2O_2$	114.042927447	114.10264
aspartic acid	D	$C_4H_5NO_3$	115.026943032	115.0874
glutamine	Q	$C_5H_8N_2O_2$	128.058577511	128.12922
lysine	K	$C_6H_{12}N_2O$	128.094963018	128.17228
glutamic acid	E	$C_5H_7NO_3$	129.042593096	129.11398
ornithine	O	$C_5H_{12}N_2O_2$	132.089877640	132.16098
methionine	M	C_5H_9NOS	131.040484606	131.19606
histidine	H	$C_6H_7N_3O$	137.058911862	137.13928
phenylalanine	F	C_9H_9NO	147.068413916	147.17386
arginine	R	$C_6H_{12}N_4O$	156.101111028	156.18568
tyrosine	Y	$C_9H_9NO_2$	163.063328538	163.17326
tryptophan	W	$C_{11}H_{10}N_2O$	186.079312954	186.2099

element	code	mono mass	avg mass
hydrogen	h	1.0078250321	1.00794
carbon	c	12.0000000	12.0107
nitrogen	n	14.0030740052	14.0067
oxygen	o	15.9949146221	15.9994
phosphorus	p	30.973762	30.973761
sulphur	s	31.97207069	32.065
proton		1.00727646688	

8

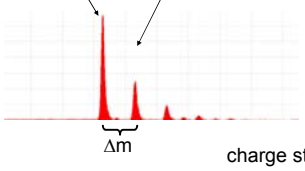
Average vs. monoisotopic mass

For example:
DIGSESTEDQAMEDIK
Mono MH+: 1767.7589 Da
Avg MH+: 1768.8272 Da

monoisotopic mass

average mass – centroid of isotopic envelope

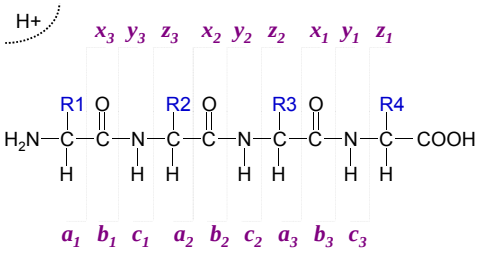
difference in mass grows as sequence grows and can be significant!



charge state = 1 / Δm

9

Fragment Ions

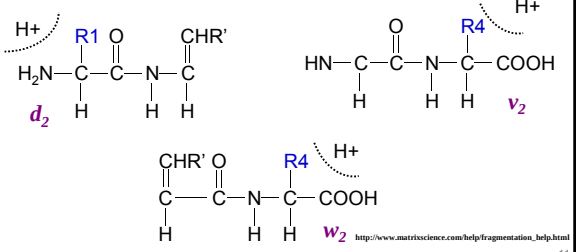


http://www.matrixscience.com/help/fragmentation_help.html

10

Fragment ion types

d-, *v*-, and *w*-ions are created by side chain cleavage. These ions are typically generated during high energy collision induced dissociation conditions. Of note, *d*- and *w*- ions allow the isobaric residues leucine and isoleucine to be differentiated.



http://www.matrixscience.com/help/fragmentation_help.html

11

Immonium ions

An internal fragment with just a single side chain formed by a combination of *a* type and *y* type cleavage is called an immonium ion. The presence of these ions can be a diagnostic to the presence of the corresponding amino acid in the peptide sequence.

Amino Acid	Residue Mass	Immonium ion mass	
Glycine	57.02147	30.03438	-
Alanine	71.03712	44.05003	-
Serine	87.03203	60.04494	+
Proline	97.05277	70.06568	++
Valine	99.06842	72.08133	++
Threonine	101.04768	74.06059	+
Cysteine	103.00919	76.0221	-
- carbamidomethylated	160.03065	133.0436	+
- carboxymethylated	161.01466	134.0276	+
- acrylamide adduct	174.0643	147.0772	+
Isoleucine	113.08407	86.09698	++
Leucine	113.08407	86.09698	++

Amino Acid	Residue Mass	Immonium ion mass	
Asparagine	114.04293	87.05584	+
Aspartic acid	115.02695	88.03986	+
Glutamine	128.05858	101.0715	+
Lysine	128.09497	101.1079 (84.08136)	
Glutamic acid	129.0426	102.0555	+
Methionine	131.04049	104.0534	+
- oxidized methionine	147.0354	120.0483	+
Histidine	137.05891	110.0718	++
Phenylalanine	147.06842	120.0813	++
Arginine	156.10112	129.114	-
Tyrosine	163.06333	136.0762	++
Tryptophan	186.07932	159.0922	+

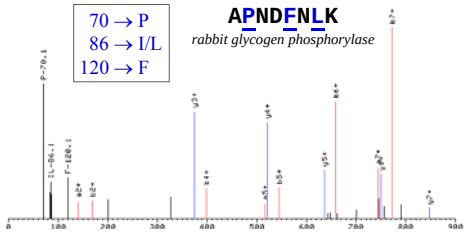
http://www.matrixscience.com/help/fragmentation_help.html
http://www.abrf.org/ResearchGroups/MassSpectrometry/EPosters/ms97quiz/residueMasses.html

12

Immonium Ions

70 → P
86 → I/L
120 → F

APNDFNLK
rabbit glycogen phosphorylase



MALDI-TOF-TOF tandem mass spectrum

13

Peptide fragmentation

A-P-N-D-F-N-L-K
(MH+ 918.5)

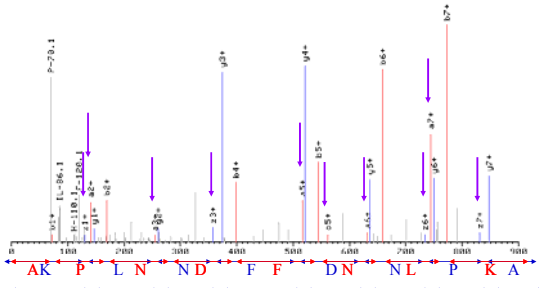
B-ions		Y-ions
72.0	A P-N-D-F-N-L-K	847.4
169.1	A-P N-D-F-N-L-K	750.4
283.1	A-P-N D-F-N-L-K	636.3
398.2	A-P-N-D F-N-L-K	521.3
545.2	A-P-N-D-F N-L-K	374.2
659.3	A-P-N-D-F-N L-K	260.2
772.4	A-P-N-D-F-N-L K	147.1

$b\text{-ions} = \Sigma AA + H^+$
$y\text{-ions} = \Sigma AA + H_2O + H^+$

monoisotopic masses

14

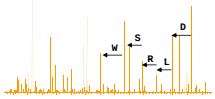
Sequence vs. Tandem Mass Spectrum



15

De novo sequencing

- Computationally (or manually) determine peptide sequence using just the MS/MS spectrum
- Requires good spectra to have half a chance of being successful
- Accurate mass very helpful to limit sequence analysis space
- Many tools tie in automated *de novo* sequencing with sequence database searching

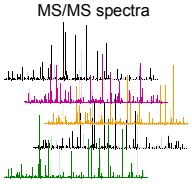


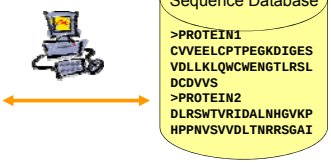
Examples: Lutefisk, PEAKS, AUDENS, PepNovo

16

Uninterpreted MS/MS database search

Raw, uninterpreted MS/MS spectra

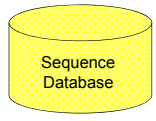




17

Uninterpreted MS/MS database search

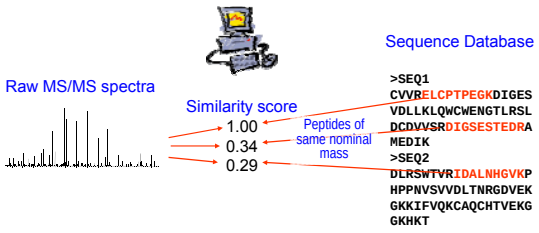
Input:
- Fragmentation spectrum
- Precursor mass, charge state



1. From database, select peptides that equal the input mass
2. Theoretically fragment peptides
3. Compare theoretical fragments to acquired spectrum
4. Generate score
5. Rank by score and display best matches

18

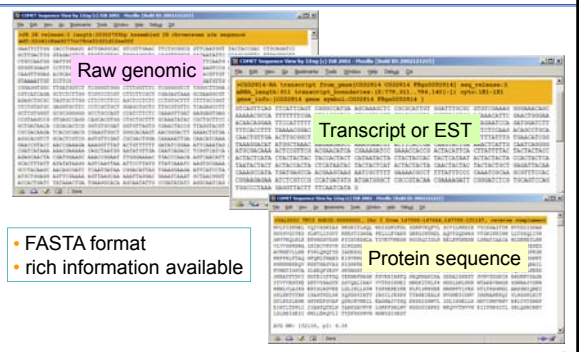
Uninterpreted MS/MS database search



MS/MS database search parameters

- Protein, nucleic acid, and EST sequence databases
- Optionally include enzyme specificity in the search
- Post-translation modifications can be identified
- Search software

Sequence databases

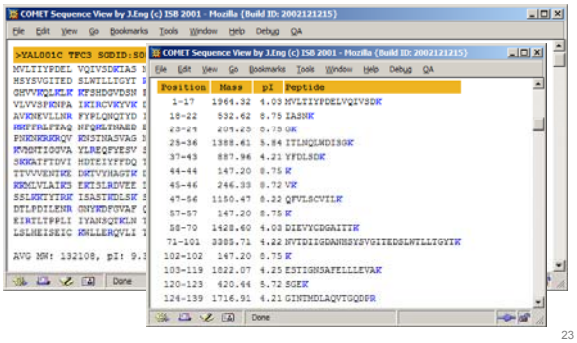


- FASTA format
- rich information available

MS/MS database search parameters

- Protein, nucleic acid, and short EST sequence databases can all be searched
- Optionally include enzyme specificity in the search
- Post-translation modifications can be identified
- Search software

DB: enzyme constraint



DB: enzyme constraint

tryptic peptides:

GDVEKGTKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGSK

TGQAPGFSYTDANKNKGITWGEETLMEYLENPKSYIPGT

enzyme-unconstrained peptides:

GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRK

TGQAPGFSYTDANKNKGITWGEETLMEYLENPKKYIPGT

DB: tryptic peptides vs. unconstrained search

mass	# tryptic peptides	# unconstr. peptides	factor
1000 Da	1,430	321,999	225x
2000 Da	466	325,096	697x
3000 Da	249	317,750	1276x

human IPI database, 47,754

25

MS/MS database search parameters

- Protein, nucleic acid, and short EST sequence databases can all be searched
- Optionally include enzyme specificity in the search
- Post-translation modifications can be identified
- Search software

26

Post-translation modifications

- Static Modification
  - ***All*** occurrences of an amino acid is modified
- Variable/Differential Modification
  - One or more occurrences of an amino acid ***may*** be modified
- Modifications can typically be specified on any residue(s) or termini.

27

Variable modifications

Serine phosphorylation:

1. DIGSESTEDQAMEDYK
2. DIGSESTEDQAMEDYK
3. DIGSESTEDQAMEDYK
4. DIGSESTEDQAMEDYK

How many peptide forms are possible if you consider serine and threonine phosphorylation for the above peptide? Serine + threonine + tyrosine?

28

Variable modification search

mass	# tryptic peptides	phos STY tryptic	factor	unconstr phos STY
1000 Da	1,430	5,093	3.5x	1,167,740
2000 Da	466	7,283	15.6x	4,538,383
3000 Da	249	16,761	67.3x	15,641,722

human IPI database, 47,754

29

Uninterpreted MS/MS database search

- Protein, nucleic acid, and short EST sequence databases can all be searched
- Optionally include enzyme specificity in the search
- Post-translation modifications can be identified
- Search software

30

SEQUEST

- First tool to perform peptide sequencing by searching uninterpreted tandem mass spectra against sequence databases
- Still currently widely used in nearly original implementation

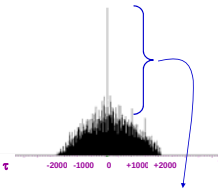
Two step approach in each search:

1. preliminary score (Sp) – a variation on the "shared peaks count" theme
2. final cross-correlation score (Xcorr) – dot product with correction

31

SEQUEST cross correlation

- Drop off rate in correlation function takes into account noise and unmatched peaks in spectra
  - faster the drop off, smaller the correction factor term, larger the Xcorr score



Xcorr = ***R*** [0] - AVG(***R*** [-75:+75])

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

τ = offset as one signal is translated across the other

32

Mascot

- Likely the most widely used search engine these days
- Probabilistic scoring
- Nice summary reports and web based tools
- Good identification performance
- Integrated or compatible with all major instrument manufacturers
- Fast searches

www.matrixscience.com

33

X! Tandem

- Open source search engine
- Very fast
- Integrated iterative searching
- Nice web-based tools and related resources
- Growing user base
- Score based on hypergeometric distribution
  - Hyperscore = dot product * ($N_{b\text{-ions}}$!) * ($N_{y\text{-ions}}$!)

www.thegpm.org
www.thegpm.org/TANDEM/api/

34

PepProbe Hypergeometric model

- If a peptide:
  - Has n_0 total fragment ions and m_0 matched fragment ions
- What the probability that this is a random match?
  - The probability that there are m_0 matches if we select n_0 times from a set that has n elements with m matches

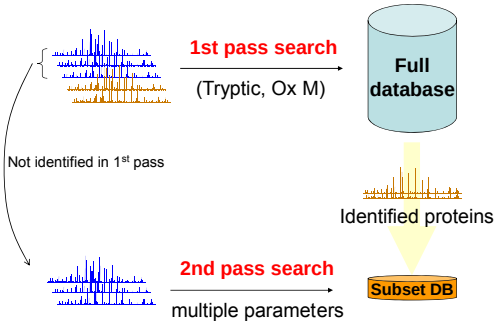
Hypergeometric probability: $P = \frac{\binom{m}{m_0} * \binom{n-m}{n_0-m_0}}{\binom{n}{n_0}}$ where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$

n: total # of calculated fragments ions in the database
m: total # of fragment ions (from n) that match a peak in the spectrum
n_0: # of fragment ions from a peptide (e.g. # b-, y-ions)
m_0: # of fragment ion matches of the peptide

Sadygov et al., Anal Chem. 2003 Aug 1;75(15):3792-8

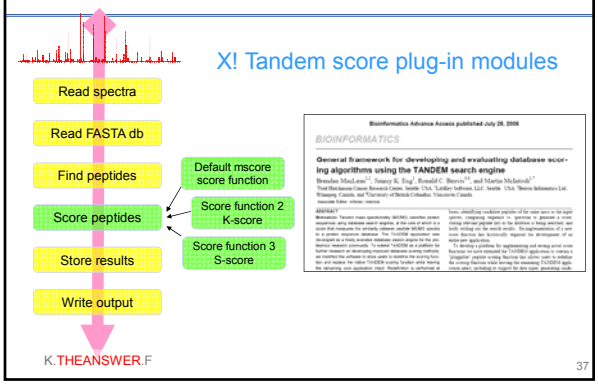
35

X! Tandem refinement mode:



36

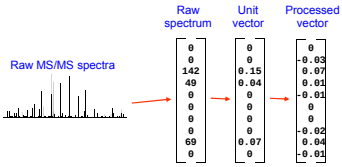
Open source = X!Tandem mod



37

k-score plug-in to Tandem

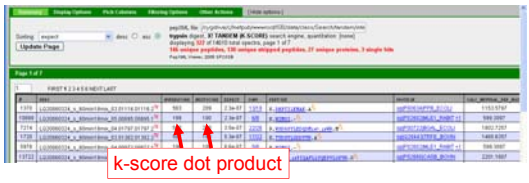
Spectral processing: $X[i] = X[i] - \frac{\sum_{j=-50}^{j=50} X[j]}{101}$



k-score = dot product of processed vector & theoretical spectrum

38

k-score plug-in to Tandem

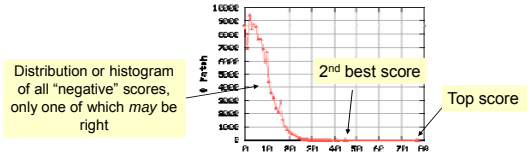


- The k-score dot product is stored and displayed in the "HYPERSCORE" column.
- The "NEXTSCORE" column shows the second highest k-score dot product score.

39

What's a good score metric?

- Every search will return the best matching peptide
- Notion of a 'good' score is typically relative
  - Look at distribution of all scores for a search
  - Look at separation of top score from the next best score
- Many search tools will estimate the likelihood that a match is incorrect (e.g. a member of the negative distribution) via p-value or e-value calculation.



40

Other search engines

- There are *many* other search tools to choose from including: OMSSA, InsPecT, MyriMatch, Phenyx, SpectrumMill, ProteinPilot, etc.
- Each tool has its own unique features and advantages/disadvantages
- Regarding the question of what tool is right for you, assuming decent or outstanding performance for each, the key questions are:
  - Which can you afford?
  - Can you easily get to the results you're looking for?
  - Is it compatible with other tools (e.g. quantification software)?
  - Is it compatible with your data and easily integrated into your data work flow?

41

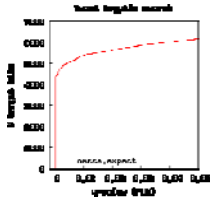
Use of 'decoy' databases

- Search target + decoy databases
  - How to generate decoys?
  - Search together (against an target + decoy appended database)
  - Search databases independently
- Decoy matches are used to estimate false discovery rate (FDR defined as expected % of false identifications):
  - Apply a score cutoff
  - Target peptide hits are what you will publish; toss out decoy matches as those are only used to estimate false IDs in your target set
  - # false IDs in the target set equals # decoys passing cutoff
  - FDR = N_{decoy} / N_{target}

42

Use of 'decoy' databases

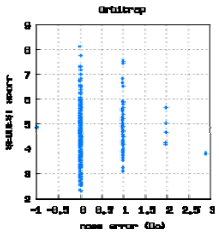
- What score cutoff should you apply?
- What FDR value is acceptable for reporting results?
- Is there a way to maximize IDs for a given FDR target?



43

What about mass accuracy?

- FTs and TOFs can measure masses very accurately (<10 ppm)
- Specifying a narrow peptide or precursor mass tolerance can greatly reduce the search times
- However, might not always be wise to use an extremely narrow mass tolerance in the primary search for a couple of reasons
  - Accurate monoisotopic mass often incorrectly determined
  - Opportunity to take advantage of mass accuracy in validation step
  - Extremely sparse search space can lead to increased false positives



44

Interpretation rules

K.LLGNQATFSPIVTVEPR.R

K.SPSDVKPLPSPDTDVPLSSVE.I

D.PEDVFTENPDEKSIITY.V

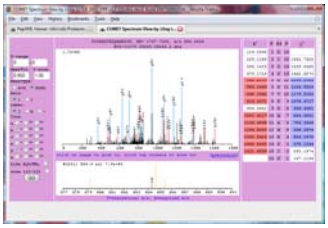
An enzyme un-restricted (or semi-enzyme) search can greatly assist in the interpretation process.

Look for peptides that exhibit the expected cleavage at both the N- and C-terminus.

Don't bother with peptides that exhibit no correct cleavage.

45

Interpretation rules



Match all fragment ions!

Correct identifications don't exhibit random fragment ion matches. Look for a series of y-ions or b-ions.

Trypsin leaves a basic residue (K or R) at the C-terminus which translate to strong y-ions so hopefully the big peaks match y-ions.

46

Interpretation rules

If a big peak matches a y-ion from an N-terminal cleavage of proline, that is a good indication of a correct identification.

The reverse is not true: a proline in a peptide that does not correspond to a big peak is not an indication of an incorrect identification.

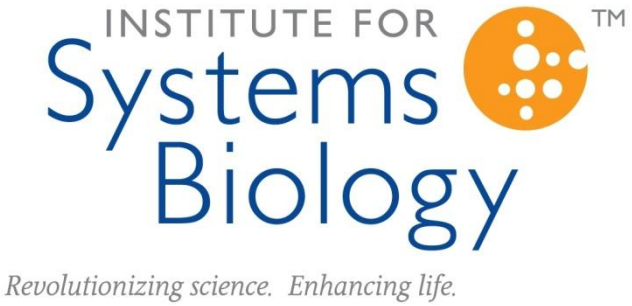
47

Proteomics Data Formats and Conversion: Generating and visualizing mzML, mzXML, pepXML, and protXML

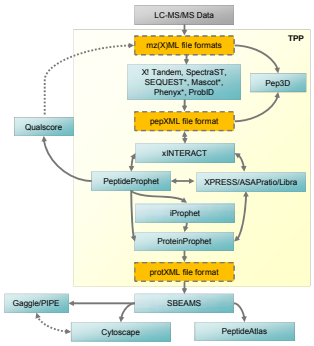
Luis Mendoza

Day 1

October 25, 2010



What is the TPP?



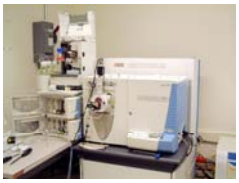
The TPP is a open-source and free collection of ***tools*** and supporting ***data formats*** which enable *shotgun proteomics* data analysis.

TPP: Tools and Data formats

Tools
- Validation
  - *PeptideProphet*
  - *ProteinProphet*
- Quantitation
  - *Xpress*, *ASAPRatio*
  - *Libra*
- Visualization
- Data Converters

Data Formats
- **mzML, mzXML**: Mass-Spec data
- **pepXML**: peptide IDs and statistics
- **protXML**: protein IDs and statistics

Why open formats?



MS Instrument Vendors: proprietary instrument formats require proprietary analysis/processing software ($)

Data (search results, etc) cannot be compared between machines / vendors / labs

vs.

TPP and related tools use open formats (mz-, pep-, prot- XML, etc) which enable:
- Sharing data between tools
- Sharing data between labs
- Applying common processing pipeline to data from different sources (machines, workflows, labs, etc)

Background: XML

- XML (eXtensible Markup Language) allows the creation of self-describing formats.
- XML consists of text organized within "tags".
  - "Human readable"
  - easily machine readable
  - example:

```
<book type="SciFi"
  location="A-13">
  <author>Douglas Adams</author>
  <title>The Hitchhiker's Guide to the Galaxy</title>
</book>
```

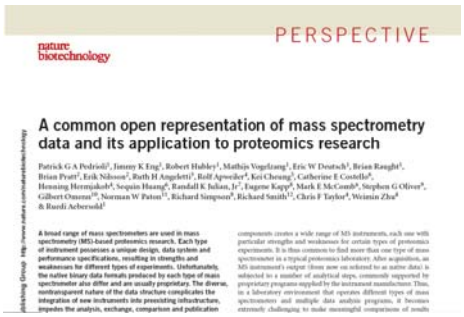

mz(X)ML: MS Data

- mzML/mzXML: MS/MS Data
  - converting from instrument formats to open formats
  - visualizing LC MS/MS with pep3D
- mzML: new MS/MS standard
- pepXML: peptide data
  - converting from search engine results to pepXML
  - visualizing with the pepXML viewer
- protXML: protein data
  - visualizing with protXML viewer

mzXML and other open formats

- Decouple your spectral data from reliance on vendor's proprietary format
- TPP-compatible MS/MS open data formats:
  - mzData 1.05 (PSI)
  - mzXML 2.1, 3.0 (SPC/ISB)
  - mzML 1.0, 1.1 (HUPO/PSI, SPC/ISB, vendors, others)

mzXML: publication


nature biotechnology

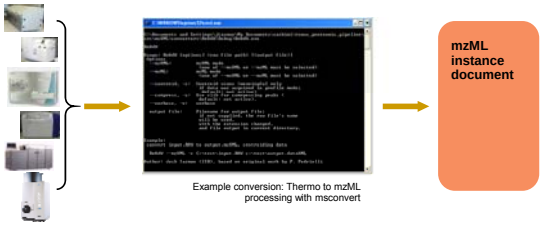
PERSPECTIVE

A common open representation of mass spectrometry data and its application to proteomics research

Pedrioli et al., Nature Biotechnology, 22, 1459-66, 2004

7

MS/MS Data Conversion Overview



specifc vendor format

appropriate mzML converter

mzML file

8

Available Converters

- *msconvert*: ProteoWizard project; included in TPP; handles Thermo to mzML, and much more
- *ReAdW*– XCalibur (**Thermo**) .raw files
- *massWolf*– MassLynx (**Waters**) .raw directories
- *compassXport** (**Bruker**) analysis.baf files (*produced by Bruker, not in TPP)
- *mzWiff*: Analyst (**ABI, Agilent**) .wiff files
- *trapper*: MassHunter (**Agilent**) .d directories

May compress peak lists and g-zip files (.mzML.gz)

9

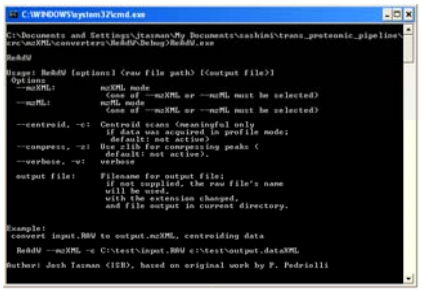
ReAdW: Thermo to mzXML

- Converts .RAW files generated by LTQ/LCQ
- Requires XCalibur
- Centroiding option

http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW

10

ReAdW: Command-line mode



11

massWolf

- Waters .raw directories
- Will require different versions for massLynx 4.0 vs 4.1
- Command-line interface and in Petunia

http://tools.proteomecenter.org/wiki/index.php?title=Software:massWolf

12

mzWiff

- ABI/Agilent .wiff files
- Command-line and in Petunia
- *Much* faster (than previous mzStar converter)
- Single program, auto-detects installed version of Analyst libraries
  - *NOTE:* ensure you're converting with the libraries you acquired the file with (e.g. converting 1.0QS with 1.4 not guaranteed.)
- Various options

13

trapper

- Agilent's recent MassHunter format (.d directories)
- Agilent has been extremely helpful with code and documentation support.
- Command-line and Petunia

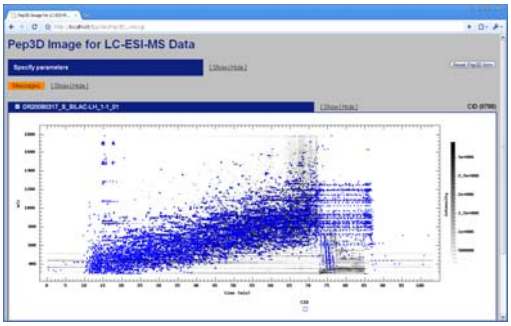
14

mzXML: MS Data

- mzXML/mzML: MS/MS Data
  - converting from instrument formats to open formats
  - visualizing LC MS/MS with Pep3D
- mzML: new MS/MS standard
- pepXML: peptide data
  - converting from search engine results to pepXML
  - visualizing with the pepXML viewer
- protXML: protein data
  - visualizing with protXML viewer

15

Pep3D



16

mzML

- mzXML/mzML: MS/MS Data
  - converting from instrument formats to open formats
  - visualizing LC MS/MS with pep3D
- mzML: new MS/MS standard
- pepXML: peptide data
  - converting from search engine results to pepXML
  - visualizing with the pepXML viewer
- protXML: protein data
  - visualizing with protXML viewer

17

mzML

- New data format for mass-spec based proteomics
- Joint effort between Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI), ISB/SPC, vendors, software houses
- More flexible, more detailed annotation than mzXML
- mzML 1.1 standardization and TPP migration: begun Spring/Summer 2009; all TPP tools can now process mzML

18

mzML Design

- XML and Controlled vocabulary
  - Like mzData
- Ontology: very thoroughly defines terminology of MS proteomics acquisition and spectrum description
- Hopefully allows inclusion of new knowledge (instrument types, ionization, scan methodology) without requiring formalization of new XML schema

19

Software: Raw-to-mzML converter

- Use **msconvert** from the ProteoWizard project
- Supported formats:

Vendor	Formats	Vendor Required Software	Required Runtime Libraries	License
Agilent	MassHunter .d	distributed with ProteoWizard	none	Apache
Applied Biosystems	WIFF	Protein Pilot 3 free trial	none	Protein Pilot
Applied Biosystems	T2D	Data Explorer	none	Data Explorer
Bruker	Compass .d, YEP, BAF, FID	distributed with ProteoWizard	MSVC 2005 SP1 x86 re-distributable	MDA
Thermo Fisher	RAW	distributed with ProteoWizard	none	MSFileReader
Waters	MassLynx raw	MassLynx	none	MassLynx

http://proteowizard.sourceforge.net/technical/formats/

20

pepXML: peptide data

- mzXML/mzML: MS/MS Data
  - converting from instrument formats to open formats
  - visualizing LC MS/MS with pep3D
- mzML: new MS/MS standard
- pepXML: peptide data
  - converting from search engine results to pepXML
  - visualizing with the pepXML viewer
- protXML: protein data
  - visualizing with protXML viewer

21

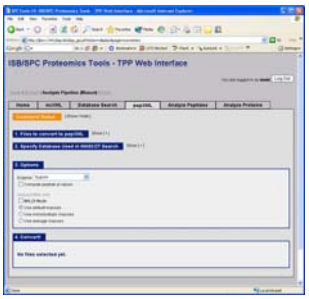
pepXML: features

- Standard format for representing **searched data**
- Stores information about peptide assignments to CID spectra
  - Mascot, Sequest, X!Tandem, ProbID, *SpectraST*, Pheynx, OMSSA
- Stores *PeptideProphet* validation results
- Stores results of Peptide Quantitation
  - *XPress, ASAPRatio, Libra*
- Stores DB Search results for one or more MS Runs
  - Search parameters stored separately for each run
- References one or more mz(X)ML files

22

pepXML: converters

- Mascot2XML
  - .dat file input
  - Petunia pepXML tab
- Out2XML
  - Converts a directory of Sequest .out files
  - Petunia pepXML tab
- Tandem2XML
  - X!Tandem result files
  - Petunia pepXML tab
- More!



23

pepXML readers/writers

- **Writers**
  - **Converters**: Mascot2XML, Sequest2XML, Out2XML, Tandem2XML
  - **Search Tools**: SpectraST
  - **Validation Tools**: PeptideProphet, iProphet
  - **Peptide Quantitation Tools**: Xpress, ASAPRatio, Libra
- **Readers**
  - **Visualization Tools**: Pep3D, PepXMLViewer
  - **Validation Tools**: PeptideProphet, iProphet, ProteinProphet
  - **Protein Quantitation Tools**: Xpress, ASAPRatio, Libra
  - **Spectrum Filtering Tools**: QualScore

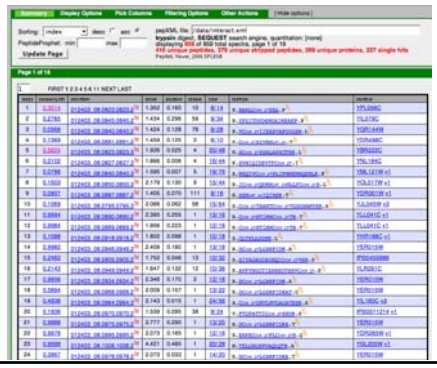
24

pepXML: peptide data

- mzXML/mzML: MS/MS Data
  - converting from instrument formats to open formats
  - visualizing LC MS/MS with pep3D
- pepXML: peptide data
  - converting from search engine results to pepXML
  - visualizing with the pepXML viewer
- protXML: protein data
  - visualizing with protXML viewer
- mzML: new MS/MS standard

25

pepXML Viewer



protXML: Protein Data

- mzXML/mzML: MS/MS Data
  - converting from instrument formats to open formats
  - visualizing LC MS/MS with pep3D
- mzML: new MS/MS standard
- pepXML: peptide data
  - converting from search engine results to pepXML
  - visualizing with the pepXML viewer
- protXML: protein data
  - visualizing with protXML viewer

27

protXML: Protein Data

- Open format for representing proteins identified by LC-MS/MS
- Stores **ProteinProphet** validation results
  - Inference of proteins based on peptide assignments to CID spectra
- Stores results of Protein Quantitation
  - XPress, ASAPRatio, Libra
- References one or more pepXML files
- Input for database storage and archiving systems
  - e.g. SBEAMS, CPAS

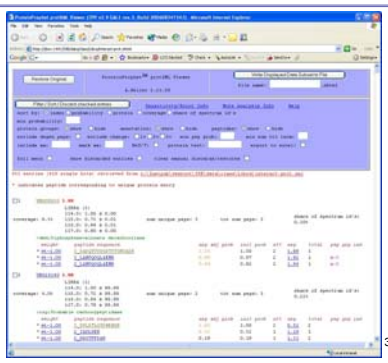
28

protXML: readers/writers

- **Readers**
  - **Visualization tools**: ProteinProphet Results Viewer (XSLT translation)
- **Writers**
  - **Protein Inference tools**: ProteinProphet, iProphet
  - **Protein Quantitation tools**: Xpress, ASAPRatio, Libra

29

protXML: visualization



30

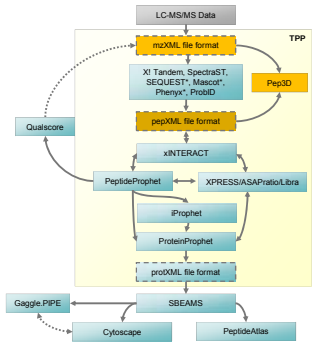
Pep3D

Luis Mendoza

Day 1

October 25, 2010

Visualizing LC-MS data



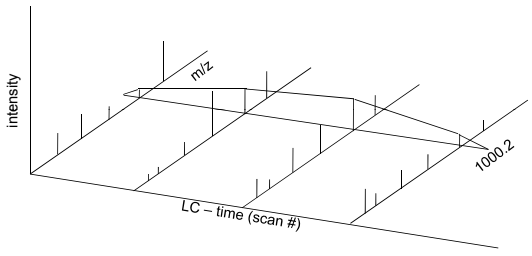
1

Why visualize LC-MS data?

- A single MS spectrum provides only a snapshot in time.
- MS/MS spectra lead to peptide identifications, but typically only the most abundant signals are examined.
- How do we know what else is occurring?
- For a more holistic view we can look at the entire LC MS space.
- The most intuitive method is visualization.

2

LC-MS Mass vs. Intensity vs. Time

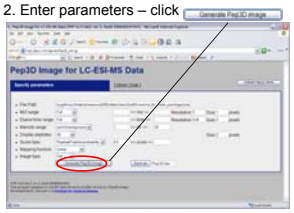


3

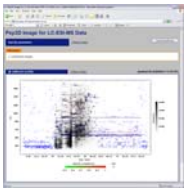
Pep3D from PepXMLViewer

1. Click Generate Pep3D button under Other Actions

2. Enter parameters – click

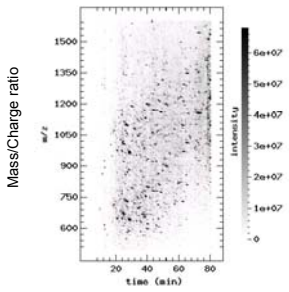


3. Visualize data!



4

That looks like a gel…

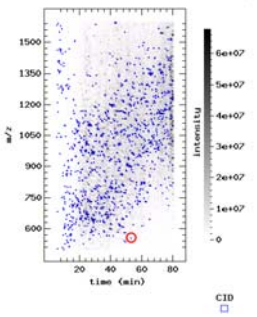


Chromatographic retention time

5

Image types: CID

CID (1198)



6

Image types: Peptide

Peptide (221)

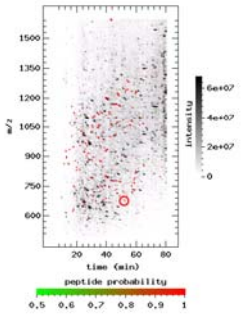
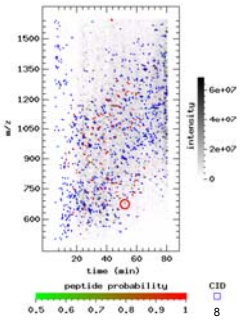


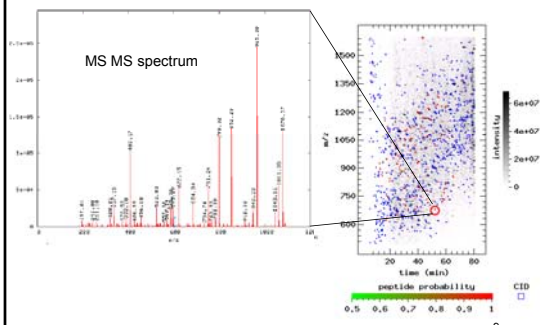
Image types: All

Peptide / CID
(221 / 1198 = 0.18)



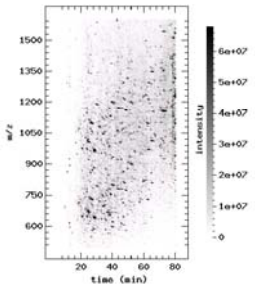
8

Display CID Spectrum



9

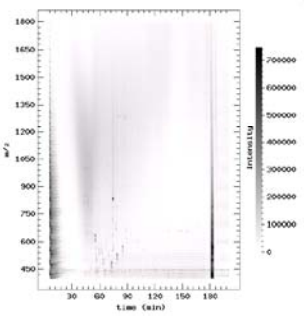
Good Sample / LC Separation



Plenty of well-localized spots without any particular large-scale pattern.

10

Where's the Sample?

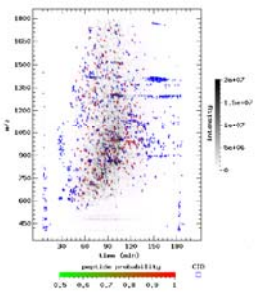


Peptide ID / CID (0/311 = 0)

- Very few localized spots
- Mainly background noise
- Distinguishable from no-spray

11

Insufficient Sample Separation

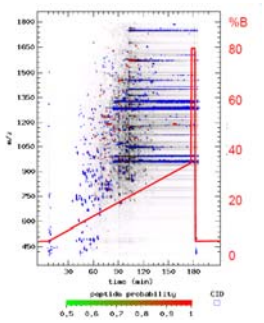


peptide ID / CID
(527/2150 = 0.25)

- Peptide ions elute in a narrow zone of the gradient
- Many intense ions not fragmented
- Insufficient fractionation / separation

12

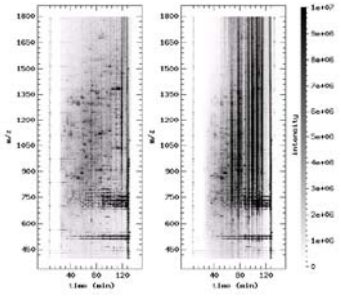
Non-optimal Gradient



- Empty elution space at the beginning
- Crowded in the middle
- Horizontal streaks at the end

13

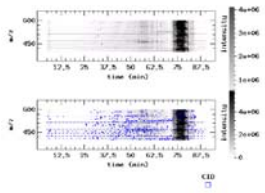
Bad RP Column



- Same sample run on 2 different columns
- peptide/CID
- (354/3004 = 0.12)
- (224/2760 = 0.09)
- 37% less IDs
- Quantitation also suffers

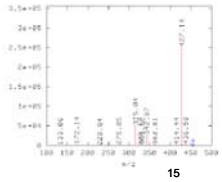
14

Chemical Contamination



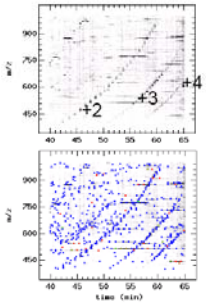
- Long horizontal streaks
- Low m/z values
- Singly charged ions
- Wasted CID attempts

CID (588); no positive ID



15

Polymer Contamination



- Localized spots running off diagonally
- Equidistant in m/z
- Almost equidistant in time
- May be ionized in multiple charge states
- Wasted CID attempts

Repeat unit of 44 amu

16

Pep3D Summary

- Pep3D can be used to:
  - Evaluate and optimize sample quality and LC-ESI-MS system performance.
  - Check reproducibility and consistency of different sample analysis

"A tool to visualize and evaluate data obtained by liquid chromatography/electrospray ionization/mass spectrometry", X-J Li et al., Analytical Chemistry, 76(13), 2004.

17

PETUNIA:
The Graphical User Interface for the TPP

Luis Mendoza

Day 1

October 25, 2010

PETUNIA Features

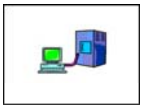
- Client/Server architecture (remote login)
- Familiar interface via web browser
- Use of advanced web technologies to render complex pages and notify user when jobs are done
- Remotely Browse/Copy/Delete files
- Controlled access via username/password
  - *NOT* secure. (Use https:// if needed)
- IIS, Apache Webservers. Firefox, IE browsers
- It's Cool!

Configuration Options

- Flexibility in Tools set-up:

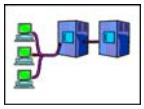
Single server

- Search software, data, TPP on a single user computer.



Search server

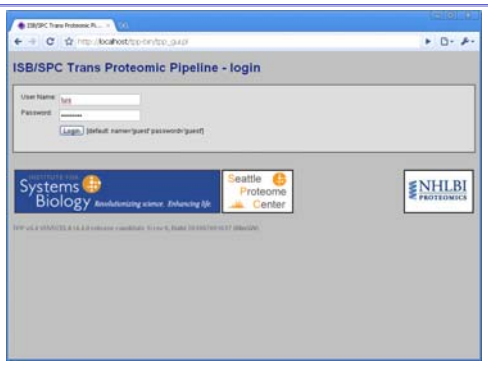
- Search software on server (cluster?)
- Data and TPP on user workstation



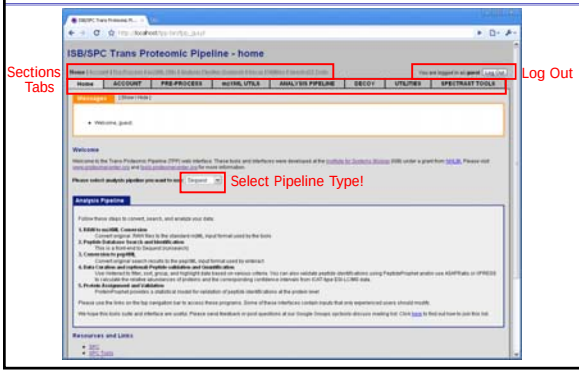
Remote access

- Search software on server (cluster?)
- Data and TPP on dedicated server
- Remote access via web browser (http)

Quick Tour: Login



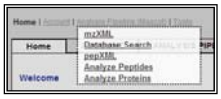
Quick Tour: Home



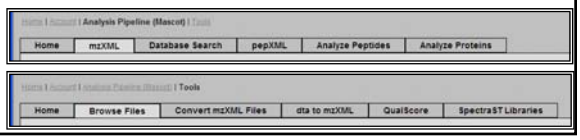
Quick Tour: Navigation

- Select Pipeline type: (default can be re-set)

- Mouse-over sections to reveal menus



- Tabs are context-sensitive:



PETUNIA: some tips & URL

- Choose Files first, then options
- OK to navigate while commands run; keep an eye on the Command Status bar for updates:

- Previous command output available until next command is run.
- Open all results files using web browser
- Use section menus for quick navigation
- Got a suggestion? Let us know!

http://<HOST>/tpp-bin/tpp_gui.pl

6

Database Searching Tutorial/Exercises

Use the PETUNIA interface to convert a RAW file to mzML, run a search, generate a pepXML file that is analyzed through the PepXML Viewer, and view the MS run and identifications with Pep3D

This tutorial will walk you through the steps needed to process data, starting from data in the mzML format through an X!Tandem search and the TPP visualization tools.

- convert raw mass-spec data to mzML
- search data with X!Tandem
- the search results will then be converted to pepXML files
- lastly the multiple pepXML files will be combined into a single view and opened in the PepXML Viewer for analysis
- visualize the MS run and ID's with Pep3D

0. A guided tour of Petunia

To start, open up a web browser and click on the home page icon in your browser and access the PETUNIA link get to the tools user interface at URL http://localhost/tpp-bin/tpp_gui.pl. Log in to the TPP graphical user interface using the username "guest" and password "guest".

On the "Home" tab, select "Tandem" as the analysis pipeline to use.

Quick tour of sections and tabs

1. Conversion to mzML

The TPP bundles the *msconvert* file conversion utility and is integrated in the web GUI (as well as converters for other vendors). Here, we will convert a Thermo file to a corresponding mzML file for subsequent analysis.

- Click on **Analysis Pipeline**. The **mzML** tab should be selected by default.
  (Note that you could also choose to convert to the mzXML format via a different tab, but you must have the appropriate converters installed!)

- Select the RAW input file
  Click **Add Files**, and navigate to the **class->Formats** folder. Select the checkbox for the two OR…RAW files and click **Select**. Leave the compression and centroiding options unselected.

- Convert to mzML
  Simply click **Convert to mzML**.

- Monitor conversion
  While the conversion runs, you can click **update this page** to refresh the page, or **show/hide** to conveniently monitor conversion progress. When the conversion is complete, the **Command Status** bar turns to orange.

- View Results
  From the same page, you can view the mz(X)ML file as text by clicking on the link from the orange box. More interestingly, you can visualize the file with Pep3D. We'll cover this in more detail in a moment. You can go to **Tools->Browse Files**, and click on the **Pep3D** link. Accept default options and click **Generate Pep3D image**, wait a moment and the gel-like image should appear.

2. Search with X!Tandem

The X!Tandem search involves three steps: (1) choosing input mz(X)ML files, (2) choosing the Tandem search parameters, and (3) selecting the sequence database to search against.

We will be using a different set of mzXML files because they contain a subset of an entire acquisition so that we can search them quickly for this exercise.

First, click on the **Add Files** button to choose the input mzXML files. Select the two mzXML files that are present in the ***data\class\Search*** directory.

Next, choose the Tandem parameters file. This file is named **tandem-params.xml** and exists in the same directory. Tandem has a rich set of search parameters and we'll discuss what some of these parameters mean.

Lastly, select the sequence database to search: ***dbase\ yeast_orfs_all_REV.20060126.fix.fasta***

Confirm everything is set correctly and click on the **Run Tandem Search** button to launch X!Tandem. On these laptops with these search parameters, the two mzXML files should complete searching in a few minutes.

3. Convert search results to pepXML

Once the searches are completed, proceed to the next tab, **pepXML**. This next step in the analysis pipeline is to convert the native search results, in this case Tandem XML output, to our pepXML format.

Simply choose the two OR2008*.tandem files in the search directory (***class\Search***) and click on the **Convert to PepXML** button.

This will launch the data conversion which takes the Tandem output and creates corresponding pepXML files that have .pep.xml extensions.

4. Combine multiple (related) search results into a single view

We're just about through with our pipeline processing steps. Next hover over the **Analysis Pipeline** section and click on the **Analyze Peptides** link (or tab). Under the **Select File(s) to Analyze** pane, choose the two OR2008*.tandem.pep.xml files that we just created. In the **PeptideProphet Options** pane, unselect the checkbox that says **RUN PeptideProphet** as that's a feature you'll learn about tomorrow. Keep all other defaults, scroll down to the bottom of the page, and hit the **Run XInteract** button. This effectively combines together the identifications from the two Tandem searches and will present them in the PepXML Viewer interface for analysis.

When the command has completed, select the **Click here to view log file and output files** link to get to the output results. Then, select the **View** link within the **Output Files** pane.

5. Explore results in pepXML Viewer

The data being analyzed are two Orbitrap runs of a SILAC labeled yeast dataset. However, we did not specify the SILAC-heavy modification in the search parameters so we are only going to identify the normal, unlabeled peptides. The data were searched against a database composed of the yeast + decoy sequences. The yeast protein sequences are denoted by their ORF identifiers (proteins beginning with "YAL", "YOR", etc.) while the decoy entries have protein identifiers that begin with "REV0" or "REV1". *A point to keep in mind is that we are searching X!Tandem with the k-score score plug-in which is different from native X!Tandem analysis.*

- The first step before we get started is to remove some unnecessary columns from the default view and add a peptide mass column. Click the **Pick Columns** pane. You can add, remove and reorder columns to view here. Remove the **bscore** and **yscore** columns (if present) and then click the **Update Page** button.

- Sort the results in ascending order based on the EXPECT score column. You can do this in either the **Summary** pane or the **Display Options** pane. Better identifications have lower EXPECT scores.

  Notice that the **IONS**, **PEPTIDE** and **PROTEIN** columns contain hypertext links. These open up an MS/MS spectrum viewer, NCBI blast link, and a sequence viewer respectively. Go ahead and click on the spectrum link for the first entry in the results view. You should see the screen on the right which will open in a new tab (in Firefox/Chrome browsers). What is notable about the spectrum shown? Does it look like a good identification? What features do you notice to support this?

  Go ahead and click on the **PEPTIDE** and **PROTEIN** links as well to see what they lead to.

- Now let's look for the best scoring decoy matches. Decoy hits are the known wrong ones!

  You could scroll through every page looking for a protein that begins with 'REV'. However, let's use the viewer's filtering options to select just these entries. To do this, go to the **Filtering Options** pane and enter '**^REV**' in the **required protein text** entry box.

  Look at the spectra of a few of these identifications. How plausible do they appear?

- What is the false discovery rate (FDR) in the dataset if you apply an EXPECT cutoff of 0.1? What about 0.01? What about 1.0?

  Remember, we calculated FDR as num_decoy_hits ÷ num_target_hits. To find these values, clear out any existing applied filters and enter a value of 0.1 or 0.01 in the **max expect** search results filter box. The **Summary** pane will tell you how many entries remain in the list after applying your filters.

  For a 0.1 cutoff, you should see "displaying 1417 of total 4158 total spectra" indicating 1417 peptide-to-spectrum matches pass this cutoff.

  So how many of these are target entries versus decoy entries? The easy way to determine this is to add an additional protein filter of '**^REV**' to determine the number of decoy entries. When you do so, you will see that there are 4 decoy hits which consequently mean that there are 1413 target hits. So at a 0.1 expectation score cutoff, the estimated FDR of the target hits would be 4 ÷ 1413 = 0.0028.

  What are the estimated FDRs at 0.01 and 0.1 cutoffs? More importantly, how would you choose an appropriate cutoff to apply?

- Remove all filtering options, add the **PPM** column, remove **HYPER** and **NEXT** columns, and sort by ascending order via the **expect** score.

  Not the PPM mass error of the good scoring IDs in the first few pages. And then browse to the last few pages and note their corresponding calculated PPM mass error. What is the highest scoring peptide with a 'bad' PPM mass error (where bad is say >50 PPM). Is there a plausible explanation for the large mass error or is this an incorrect match?

- Lastly, go ahead and explore the rest of the interface. Take a look at the available columns and how you can re-order them. Explore your ability to sort/filter based on the various properties. In order to start all over from scratch, including restoring data and views to default, choose **Restore Original** from the **Other Actions** pane.

  You will be using this viewer in the upcoming days so you should definitely become comfortable navigating the interface today.

6. Visualize searched data using Pep3D

Pep3D is a tool for visualizing LC MS data, along with CID attempts, certain analysis scores, and MS/MS spectra.

- Visualizing mz(X)ML files directly using Pep3D
  When Petunia detects an mzXML /mzML file in a file listing, it allows the user to open it directly using Pep3D. Open the File Browser in Petunia, navigate to the **/data/class/Search/** directory (if not already there) and find the file named **OR…_01-trimmed.mzXML**. Click on the **[Pep3D]** link to launch Pep3D. A new browser window will open showing the Pep3D interface. This interface allows the user to adjust various parameters that determine the way that the Pep3D image is rendered. For now, accept the default parameters and click the **Generate Pep3D Image** button. After a short time, an image representing the LC MS data will appear.

  Note the large amount of 'empty space' at the right of the image. This area is blank because the mass spectrometer did not acquire much data in that time range. Adjust the time range that Pep3D renders by entering an upper time range of 75 (minutes); notice that the value for **Elution time range** drop-down box automatically changed from **Full** to **Selected**. Click the **Generate Pep3D Image** button. After a short time, an adjusted image will appear.

  Note that the separation appears compressed between 15 to 70 minutes. Stretch this section of the image by adjusting the **Elution time range** with these values, and generate the image. You will notice that it is considerably narrower that the original. To increase the width or height of an image you simply adjust the values of the **Resolution** and **Size** parameters. It helps to think of these parameters as 'units per pixel'. Change the **minutes** increment setting parameter of the **Elution time range** from 0.5 to 0.2and again click the **Generate Pep3D Image**. Can you notice some light/heavy SILAC pairs as they co-elute?

  Pep3D can also overlay the co-ordinates of MS/MS events on the image. Select **CID** from the **Display peptides** drop-down box and generate image. The blue squares represent CID events (2078 of them). Click on a blue square CID. A new browser window will open showing an image of the selected CID spectrum. Also, notice the distribution of the blue squares within the m/z and elution time space; as mentioned above, we have "trimmed" the MS-2 data from the original file so that the search and analysis can be finished in a timely manner on these machines.

- Visualizing LC MS data with peptide identifications
  Pep3D can also be opened from within PepXMLViewer. Return to Petunia and use the Browse Files tab locate the **interact-OK.xml** file in the **Pep3D\raftflow31-39.sequest** directory. Click on the **[PepXML]** link to open the PepXML Viewer. Click on the **Generate Pep3D button** found under the **Other Actions** tab. A new browser window will open containing the Pep3D user interface. Accept the default parameters and click

Generate Pep3D Image. When Pep3D is given a pepXML file as input, it will generate an image for each mz(X)ML that is referenced in that pepXML.

Unlike the previous example, CID attempts are now color coded to reflect the PeptideProphet probability of their corresponding peptide assignments. Below each Pep3D image there is a colored scale ranging from probability 1 to 0.5, red to light green respectively.

(Note that because multiple images are being displayed, the CID links are disabled. If you are interested in viewing spectra, you can "hide" all images, deselect all but one, and click on **Generate Pep3D Image**. Then you can click on a colored square, and as before, a new browser window will open and display an image of the MS/MS spectrum. Note that now the assigned peptide sequence and its search scores are additionally displayed.)

Examine the peptide/CID ratio shown at the top of each Pep3D image. This is the ratio of PeptideProphet-confirmed peptide identifications (P>0.5) to CID events; the greater the number of successful identifications, the higher the ratio. A low ratio indicates either wasted CID attempts on non-peptide precursors or PeptideProphet did not find many high confidence identifications.

1. Which run has the highest ratio? _____ ratio: _____
2. Which run has the lowest ratio? _____ ratio: _____
3. Examine the Pep3D of these two runs. Are there any major differences?

- Troubleshooting low Peptide / CID ratios
  Return to Petunia and use the **Browse Files** tab locate the interact-PROBLEM.xml file in the **Pep3D\raftflow31-39.sequest.problem** directory. Click on the **[PepXML]** link to open the PepXMLViewer and generate Pep3D images as described previously.

  Examine the peptide/CID ratio shown at the top of each Pep3D image and compare to those from 1&2.

  4. What is the highest ratio? _____
  5. Given that the LC-MSMS data is the same as above, why are the ratios different? [HINT: examine the search parameters by clicking the Additional Analysis Info button under the **Other Actions** tab in PepXMLViewer].

TPP Installation & Support

Luis Mendoza

Day 1

October 25, 2010

TPP Installation

- Installs on Linux and Windows systems
- Open-Source: Source code available at Sourceforge: *http://sf.net/projects/sashimi/*

LINUX

- Distributed as a zip archive
- Consult the included README for build and usage instructions, and requirements
- Apache web server

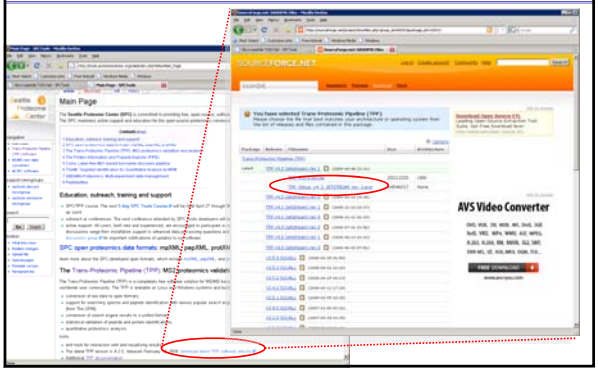
TPP Windows Installation

Windows

- Native Windows distribution
- Also uses Apache web server (bundled)
- Simple one-click download and installation (+ reboot)
- All instructions online at our website: *http://tools.proteomecenter.org/wiki/*

- *Extra downloads may be required for mz(X)ML file conversions*

TPP Windows Download



TPP Support

- Support is provided via "Google-groups" email distribution lists and a wiki:

spctools-announce

- Low-volume, moderated list
- New releases, software updates, etc.

spctools-discuss

- Open forum for discussing SPC proteomics tools, asking questions, and suggesting new features
- Over 2000 topics; 950+ members
- Archived at google: easy to search!
- Tip: Search first before sending a question!

TPP Support

Sign up at:

http://tools.proteomecenter.org/help.php

Browse archives:

http://groups.google.com/group/spctools-announce

http://groups.google.com/group/spctools-discuss

- **SPCTools Wiki:**
  - Installation and software help
  - FAQ
  - Contribute your own knowledge!
  - *http://tools.proteomecenter.org/wiki/*

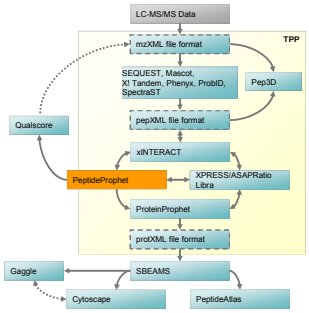
PeptideProphet: Statistical Validation of Peptide Identifications

David Shteynberg

Day 2

October 26, 2010

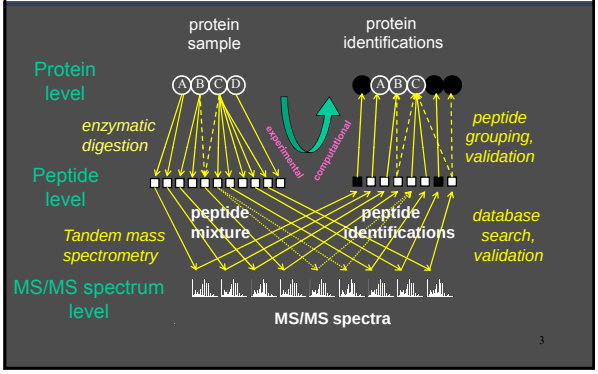
Next TPP Step: PeptideProphet



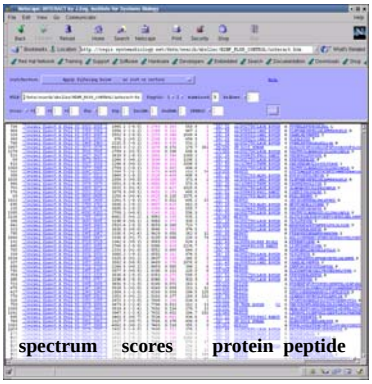
Outline

- Need to validate peptide assignments to MS/MS spectra
- Statistical approach to validation
- Running PeptideProphet software
- Interpreting results of PeptideProphet
- Demo

Shotgun Protein Identification



Peptide validation using PeptideProphet



MS/MS database search algorithms: SEQUEST, Mascot, Sonar, etc.

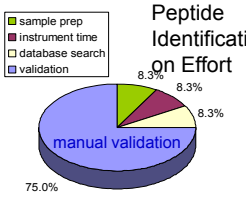
- thousands of spectra
- which peptide identification are correct?
- how to filter data?

Most search results are wrong

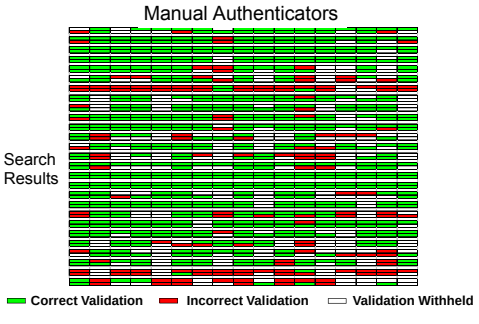
- $[M+2H]^{2+}/[M+3H]^{3+}$ uncertainty (LCQ)
- Non-peptide noise
- Incomplete database
  - post-translational modifications
  - polymorphisms
- Multiple precursors
- Limitation of database search algorithm

Validation of Peptide Assignments

- In the past, a majority of analysis time was devoted to identifying the minority of correct search results from the majority of incorrect results
- Required manual judgment



(Un)reliability of Manual Validation



7

Need for Objective Criteria

- Manual scrutiny of search results is not practical for large datasets common to high throughput proteomics
- As an alternative to relying on human judgment, many research groups employ search scores and properties of the assigned peptides to discriminate between correct and incorrect results

8

Traditional Filtering Criteria

- Each SEQUEST search result has:

  Xcorr, dCn, Sp, NTT (number of tryptic termini)

  – Accept all results that satisfy:

  [M+2H]2+: Xcorr ≥ 2, dCn ≥ 0.1, Sp ≤ 50 (NTT ≥ 1)

  [M+3H]3+: Xcorr ≥ 2.5, dCn ≥ 0.1, Sp ≤ 50 (NTT ≥ 1)

  [M+2H]2+: Xcorr ≥ 2, dCn ≥ 0.1, Sp ≤ 50, (NTT ≥ 1)

  [M+3H]3+: Xcorr ≥ 2, dCn ≥ 0.1, Sp ≤ 50, (NTT ≥ 1)

- Each Mascot search result has:

  Ionscore, Identityscore, Homologyscore, NTT

  – Accept all results that satisfy:

  Ionscore > Identityscore (NTT = 2)

  Ionscore > Homologyscore (NTT = 2)

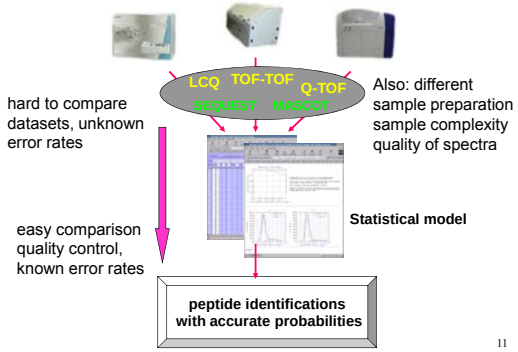
9

Problems with Traditional Filtering

- Different research groups use different thresholds
- Divides data into correct and incorrect-- no in between
- Unknown error rates (fraction of data passing filter that are incorrect)
- Unknown sensitivity (fraction of correct results passing filter)
- Appropriate threshold may depend on database, mass spectrometer type, sample, *etc*.

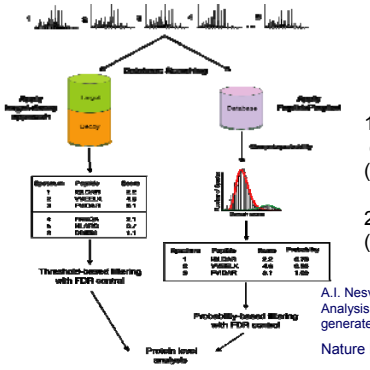
10

PeptideProphet Software



11

Target-decoy strategy



Commonly used methods:

1. Target-decoy database searching (global FDR)

2. PeptideProphet (probability – localFDR)

A.I. Nesvizhskii, O. Vitek, R. Aebersold Analysis and validation of proteomic data generated by tandem mass spectrometry

Nature Methods 4, 787-97 (2007)

12

Use of Forward/Reverse Database to Estimate False Positive Error Rates

- Do search against single Forward/Reverse database containing usual entries along with their sequence-reversed counterparts
- Forward and Reverse protein sequences each comprise 50% of the database peptides
- Incorrect results, taken at random from the database, are predicted to correspond with Reverse protein sequences on average 50% of the time
- Number of incorrect results passing any score filter calculated as twice the number of accepted results corresponding to Reverse proteins
- Search takes more time

13

Use of Separate Forward and Reverse Database Searches

- Do searches against Forward and Reverse databases separately
- Number of incorrect results in Forward search passing any score filter calculated as the number of results passing the same filter applied to the Reverse search
- Gives an overestimate of the number of incorrect results passing a filter since compares the Reverse search which has no correct results with the Forward search which may have up to 100% correct results
- Results of 2 searches must be analyzed in parallel

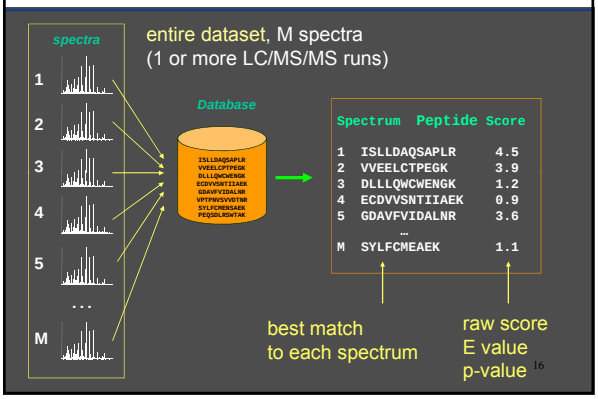
14

Statistical Approach

- Use search scores and properties of the assigned peptides to compute a probability that each search result is correct
- Desirable model properties:
  - Accurate
  - High power to discriminate correct and incorrect results
  - Robust

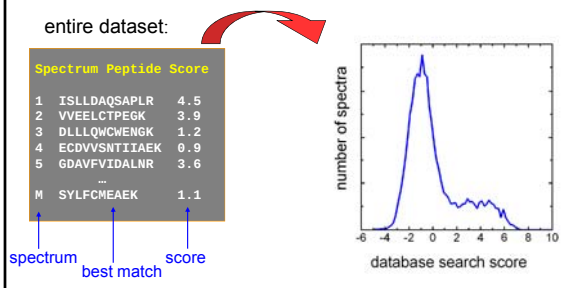
15

Statistical Model



16

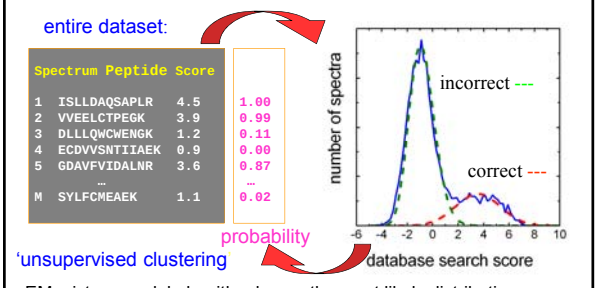
Bayesian Classification



A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold *Anal. Chem.* 74, 5383 (2002)

17

Statistical Model for Computing Peptide Probabilities (PeptideProphet)



EM mixture model algorithm learns the most likely distributions among correct and incorrect peptide assignments given the observed data

18

Training Dataset

- Want dataset of search results for which the true correct and incorrect peptide assignments are known
- Sample of 18 control proteins (bovine, yeast, bacterial)
- Collect ~40,000 MS/MS spectra, and search with engine of choice against a *Drosophila* database appended with sequences of 18 control proteins and common sample contaminants

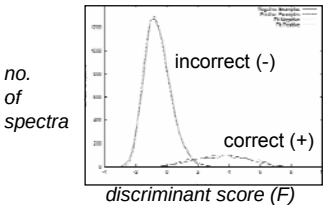
19

Derive Discriminant Function

- Derive single search score best at discriminating correct from incorrect search results
  - Generally, can combine together multiple search engine scores, when available, into single linear combination score using Linear Discriminant Function Analysis (*e.g.* SEQUEST's Xcorr, DeltaCn, and SpRank, X! Tandem K-score's Hyperscore, DeltaHyperscore)
  - Use search engine score directly if only one (e.g. Mascot's Ionscore – Identityscore)
- Derive separately for search results of each parent ion charge

20

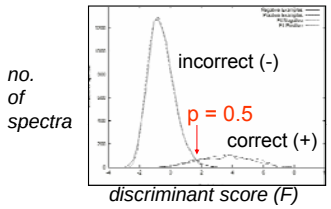
Discriminant Score Distributions



SEQUEST training dataset $[M+2H]^{2+}$ spectra

21

Computing probabilities from discriminant score distributions

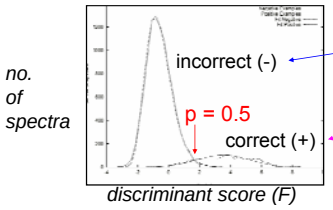


Probability of being correct, given discriminant score F_{obs}, is:

$$p = \frac{\text{Number of correct search results with } F_{obs}}{\text{Total number of search results with } F_{obs}}$$

22

Computing probabilities from discriminant score distributions



Model Incorrect results as ***Gamma Distribution***

Model Correct results as ***Normal Distribution***

Probability of being correct, given discriminant score F_{obs}, is:

$$p = \frac{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total correct}}{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total correct} + \text{Gamma}_{\alpha,\beta,\text{zero}}(F_{obs}) * \text{Total incorrect}}$$

23

Employing peptide properties

- Properties of the assigned peptides, in addition to search scores, are useful information for distinguishing correct and incorrect results.
- For example, in unconstrained searches with MS/MS spectra collected from trypsinized samples, a majority of correct assigned peptides have 2 tryptic termini (preceded by K,R), whereas a majority of incorrect assigned peptides have 0 tryptic termini.

24

Number of Tryptic Termini (NTT)

NTT can equal 0, 1, or 2:

G.HVEQLDSSS.D NTT = 0

K.HVEQLDSSS.D NTT = 1
G.HVEQLDSSR.D NTT = 1

K.HVEQLDSSR.D NTT = 2

- Computed in an analogous manner for any enzyme other than trypsin

25

Number of Tryptic Termini (NTT)

For the same value of F, assigned peptides with *higher* NTT values are *more* likely to be correct

Example: training dataset
Correct: 0.03 NTT=0, 0.28 NTT=1, 0.69 NTT=2
Incorrect: 0.80 NTT=0, 0.19 NTT=1, 0.01 NTT=2

Probability of being correct, given discriminant score F_{obs} with **NTT=2** is:

$$p = \frac{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total corr} * 0.69}{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total corr} * 0.69 + \text{Gamma}_{\alpha,\beta,zero}(F_{obs}) * \text{Total incorr} * 0.0}$$

F_{obs}: p = 0.5 without NTT becomes p=0.99 using NTT

26

Number of Tryptic Termini (NTT)

For the same value of F, assigned peptides with *lower* NTT values are *less* likely to be correct

Example: training dataset
Correct: 0.03 NTT=0, 0.28 NTT=1, 0.69 NTT=2
Incorrect: 0.80 NTT=0, 0.19 NTT=1, 0.01 NTT=2

Probability of being correct, given discriminant score F_{obs} with **NTT=0** is:

$$p = \frac{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total corr} * 0.03}{\text{Normal}_{\mu,\sigma}(F_{obs}) * \text{Total corr} * 0.03 + \text{Gamma}_{\alpha,\beta,zero}(F_{obs}) * \text{Total incorr} * 0.8}$$

F_{obs}: p = 0.5 without NTT becomes p=0.04 using NTT

27

Additional Peptide Properties

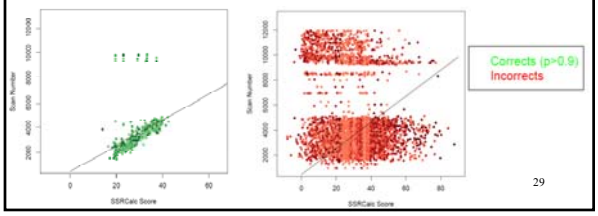
- Number of missed tryptic cleavages (NMC)
- Mass difference between precursor ion and peptide
- Difference between observed (scan no.) and calculated retention time
- Presence of light or heavy cysteine (ICAT)
- Presence of N-glyc motif (N-glycosylation capture)
- Calculated pI (FFE)

Incorporate similar to NTT above, assuming independence of peptide properties and search scores among correct and incorrect results

28

RT Model

- Use early estimates of correct results to derive linear correlation [diagonal lines below] between scan number and calculated retention time (SSRCalc score)
- Compute a new score of the difference [along y-axis] between the actual scan number and that correlating with the calculated retention time for the peptide



29

Computed Probabilities

Given training dataset distributions of F, NTT, NMC, Massdiff, RTdiff, ICAT, N-glyc, and pI among correct and incorrect search results,…

…then the probability of any search result with F_{obs}, NTTobs, NMCobs, Massdiffobs, $RTdiff_{obs}$, ICATobs, N-glycobs, and pI_{obs} can be computed as described above, with terms for each piece of information

- Accurate
- Discriminating

30

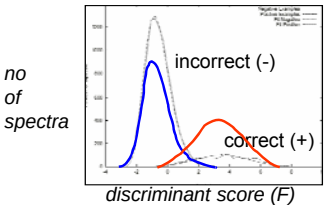
Robust Model

- One cannot rely on the ***training dataset*** distributions of F, NTT, NMC, Massdiff, RTdiff, ICAT, N-glyc, and pI among correct and incorrect search results
- These distributions are expected to vary depending upon:
  - Database used for search
  - Mass spectrometer
  - Spectrum quality
  - Sample preparation and purity
  - Chromatography

31

Variations in Discriminant Score Distributions

Different proportion of correct results in dataset

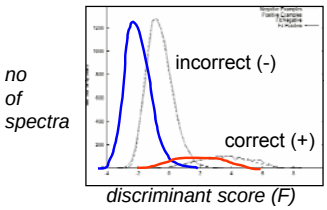


vs. training dataset $[M+2H]^{2+}$ spectra

32

Variations in Discriminant Score Distributions

Different distribution means



vs. training dataset $[M+2H]^{2+}$ spectra

33

EM Algorithm

- PeptideProphet learns the distributions of F and peptide properties among correct and incorrect search results in each dataset
- It then uses the learned distributions to compute probabilities that each search result is correct
- Expectation-Maximization (EM) algorithm: unsupervised learning method that ***iteratively*** estimates the distributions given probabilities that each search result is correct, and then computes those probabilities given the distributions
- Initial settings help guide algorithm to good solution

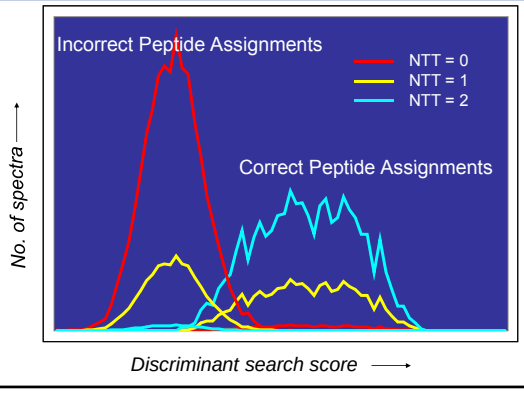
34

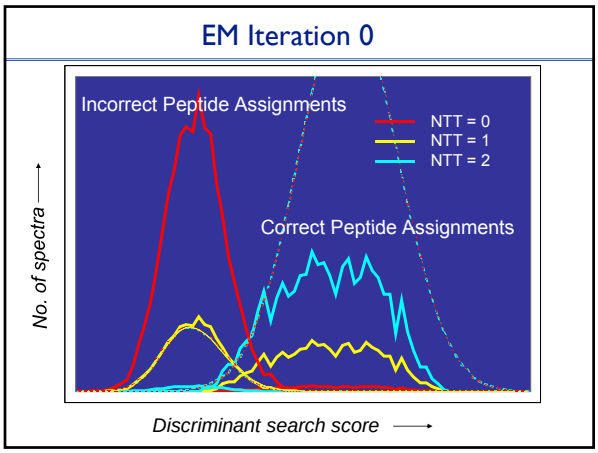
Guiding the EM with Decoy Results

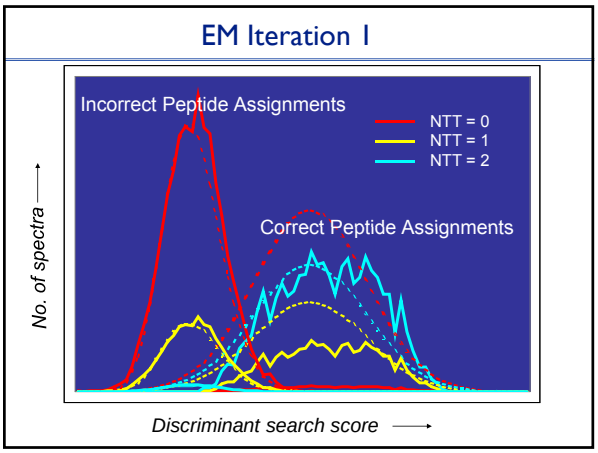
- Searches with Forward/Reverse databases provide extra information to the EM algorithm
- Results corresponding to decoy proteins (i.e. Reverse protein sequence entries) are known to be incorrect and assigned a probability of 0
- This can help the model correctly learn F-score distributions among correct and incorrect results even when they are highly overlapping
- Decoy results can also be used by themselves to model the distributions among incorrect results, eliminating the need for additional iterations

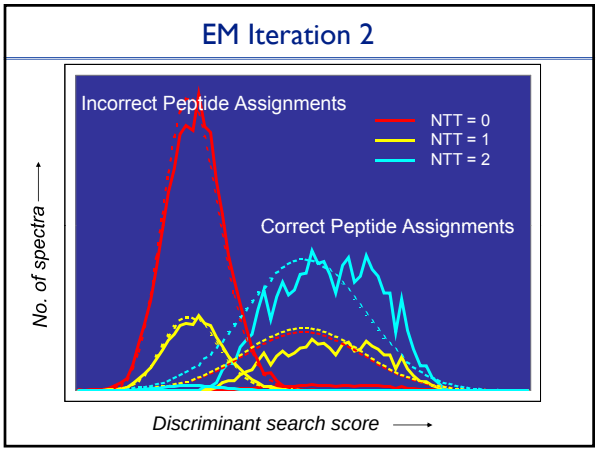
35

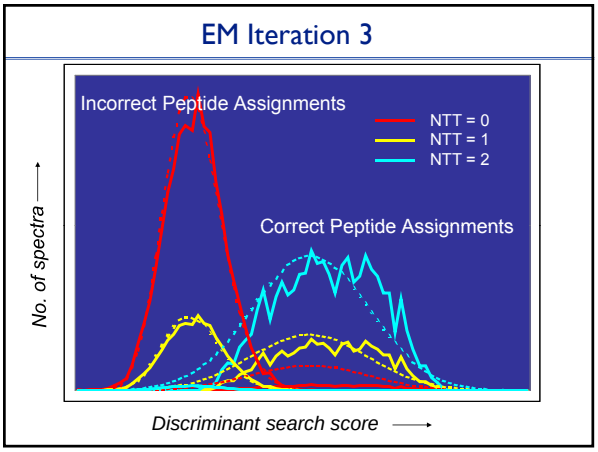
EM Algorithm learns test data score distributions

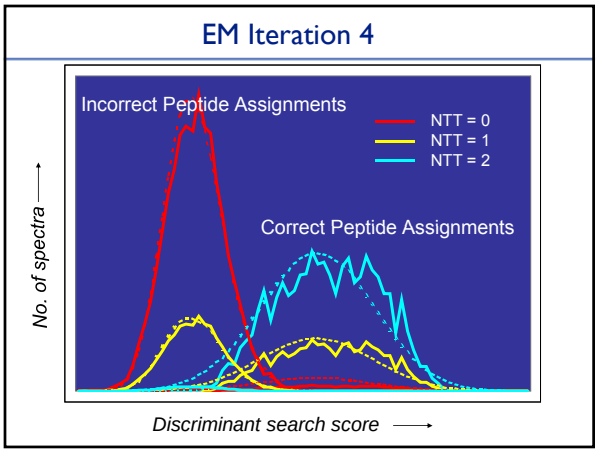


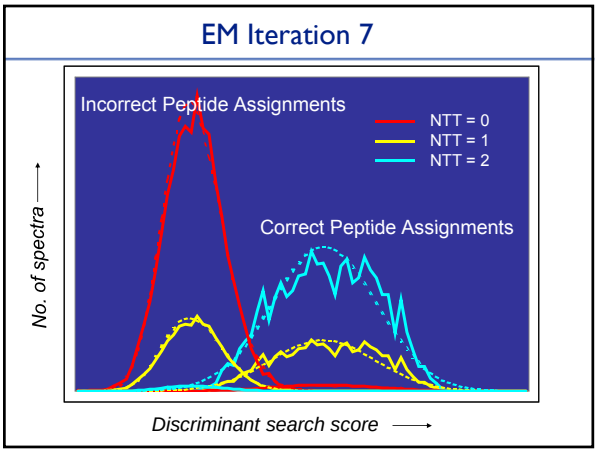

EM Iteration 0
Incorrect Peptide Assignments
NTT = 0
NTT = 1
NTT = 2
Correct Peptide Assignments
No. of spectra
Discriminant search score


EM Iteration 1
Incorrect Peptide Assignments
NTT = 0
NTT = 1
NTT = 2
Correct Peptide Assignments
No. of spectra
Discriminant search score


EM Iteration 2
Incorrect Peptide Assignments
NTT = 0
NTT = 1
NTT = 2
Correct Peptide Assignments
No. of spectra
Discriminant search score


EM Iteration 3
Incorrect Peptide Assignments
NTT = 0
NTT = 1
NTT = 2
Correct Peptide Assignments
No. of spectra
Discriminant search score


EM Iteration 4
Incorrect Peptide Assignments
NTT = 0
NTT = 1
NTT = 2
Correct Peptide Assignments
No. of spectra
Discriminant search score

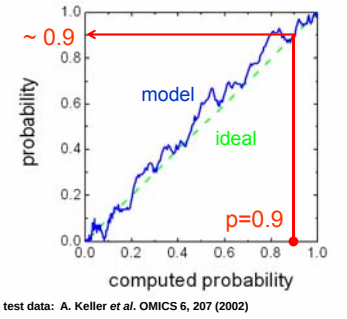

EM Iteration 7
Incorrect Peptide Assignments
NTT = 0
NTT = 1
NTT = 2
Correct Peptide Assignments
No. of spectra
Discriminant search score

Charge State Ambiguity

- LCQ/LTQ: Assumed a spectrum can be in either charge 2+ or 3+ (if not 1+)
- ETD: Spectrum can be in any number of predicted charge states
- ASSUMPTION: Only one charge state is correct

2009-06-11

Accuracy of the Model



100 spectra with computed p ~ 0.9

90% of them (90) should be correct

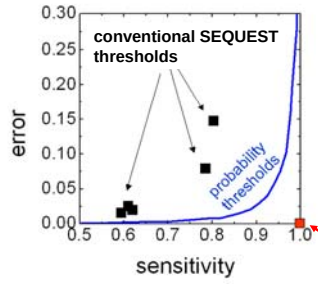
Observed probability is around 0.9

Model is accurate

test data: A. Keller *et al.* OMICS 6, 207 (2002)

44

Discriminating Power of Computed Probabilities: SEQUEST



Sensitivity: fraction of all correct results passing filter

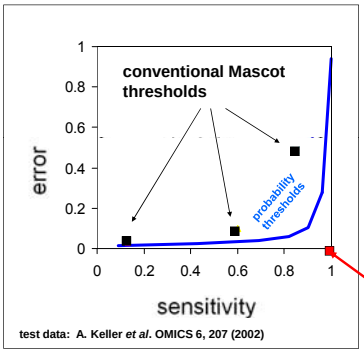
Error: fraction of all results passing filter that are incorrect

Ideal Spot

test data: A. Keller *et al.* OMICS 6, 207 (2002)

45

Discriminating Power of Computed Probabilities: Mascot



Sensitivity: fraction of all correct results passing filter

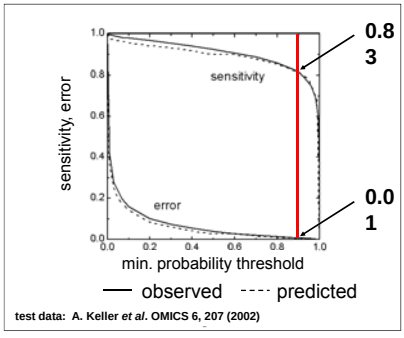
Error: fraction of all results passing filter that are incorrect

Ideal Spot

test data: A. Keller *et al.* OMICS 6, 207 (2002)

46

Discriminating Power Example: p ≥ 0.9



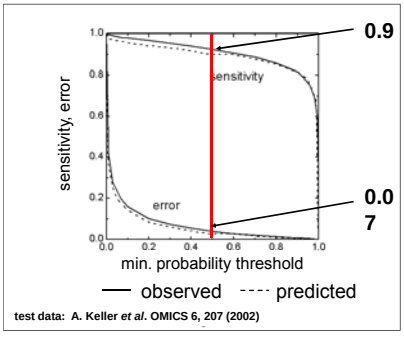
Sensitivity: fraction of all correct results passing filter

Error: fraction of all results passing filter that are incorrect

test data: A. Keller *et al.* OMICS 6, 207 (2002)

47

Discriminating Power Example: p ≥ 0.5



Sensitivity: fraction of all correct results passing filter

Error: fraction of all results passing filter that are incorrect

test data: A. Keller *et al.* OMICS 6, 207 (2002)

48

Use of PeptideProphet Probabilities to Compare Searches

- False positive error rate predicted by PeptideProphet is an objective criterion for comparing different searches
  - Sample preparation and LC/MS/MS
  - Search conditions
  - Search engine
- Compare the number of results of each search passing its minimum probability threshold to achieve a fixed predicted false positive error rate
  - Reflects both search engine and PeptideProphet performance

49

PeptideProphet Software Tutorial

- How to run PeptideProphet through the TPP Web Interface
- Interpretation of analysis results
- User options

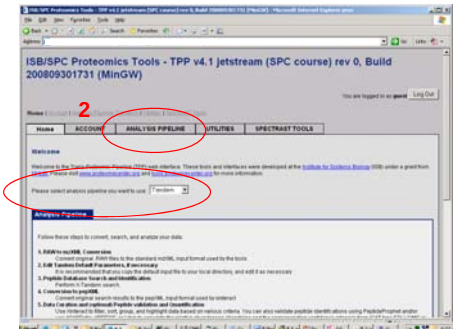
Getting started with PeptideProphet

- Input: pepXML files (**`file1.xml, file2.xml`**…)
- XInteract program first merges files together into single file **`interact.xml`**, then PeptideProphet runs model, computes probabilities, and writes probabilities as first column
- Combine together runs that are similar (sample, database, search constraints, mass spectrometer)

51

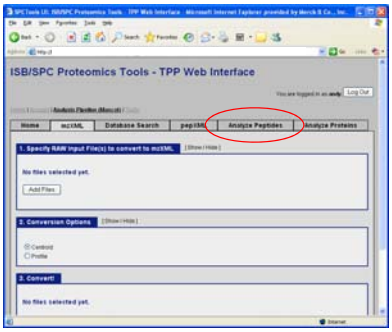
Getting started with PeptideProphet

Specify search engine and select Analysis Pipeline



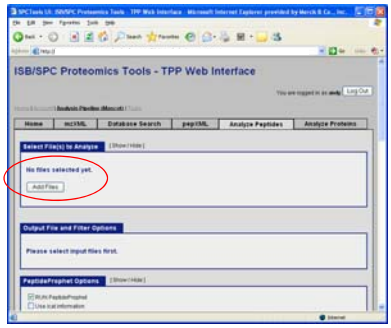
Getting started with PeptideProphet

Select peptide level analysis



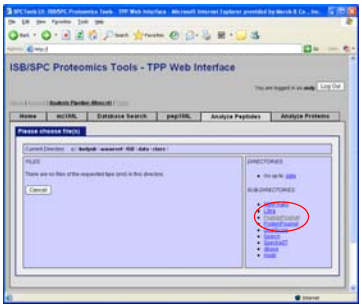
Getting started with PeptideProphet

Specify search results to analyze



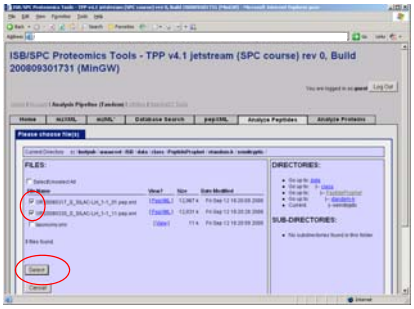
Getting started with PeptideProphet

Navigate data directories



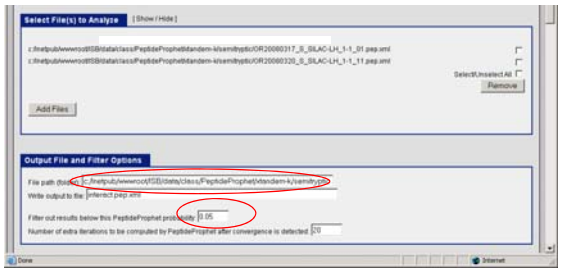
Getting started with PeptideProphet

Add each search run pepXML included in analysis



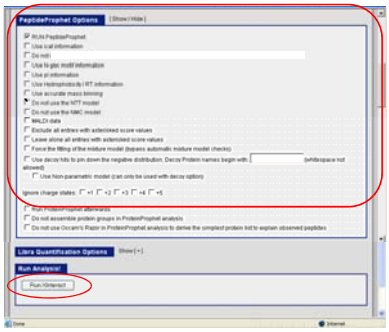
Getting started with PeptideProphet

Specify output file name and minimum probability filter



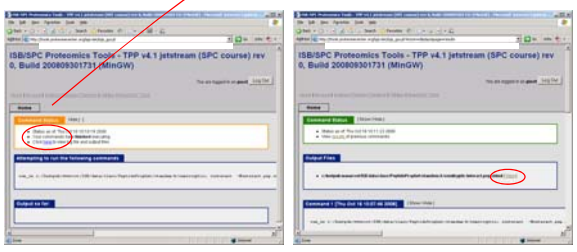
Getting started with PeptideProphet

Specify to run PeptideProphet, its optional parameters, and run analysis

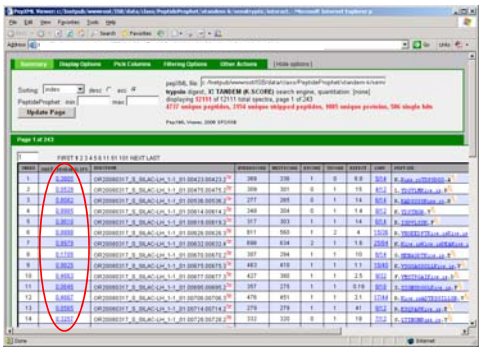


Getting started with PeptideProphet

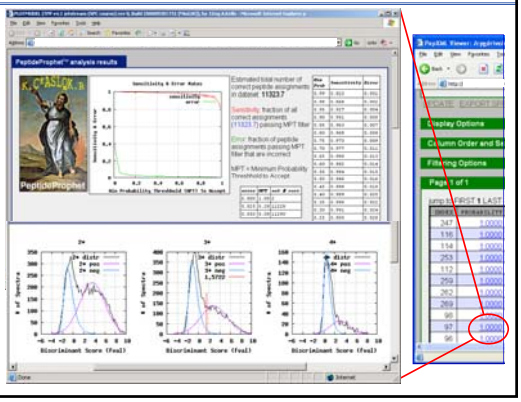
Click on links to view results of analysis



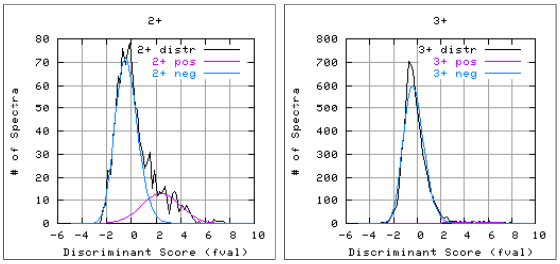
PeptideProphet Results



PeptideProphet Results: Model Summary

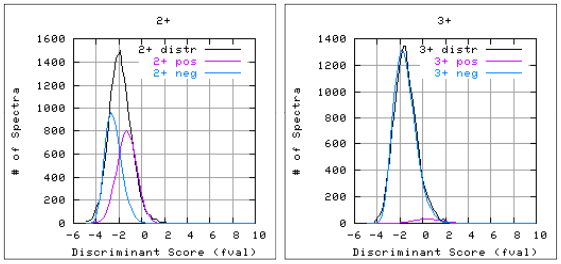


Reasonable Learned Discriminant Score Distributions



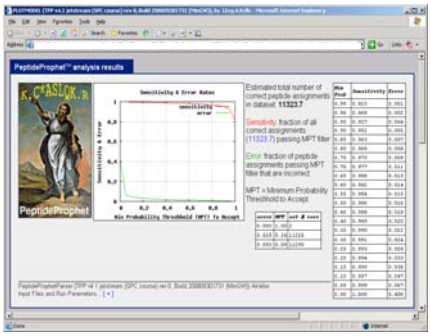
62

Suspicious Looking Learned Discriminant Score Distributions



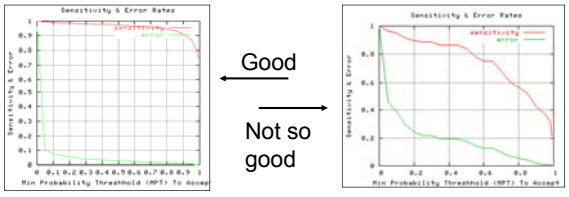
63

PeptideProphet Results: Model Summary



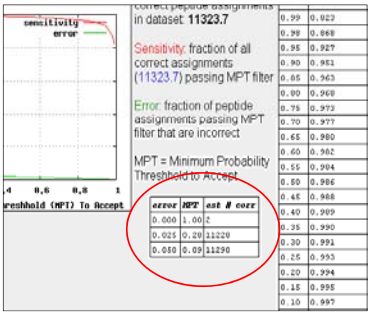
64

PeptideProphet Results: Model Summary



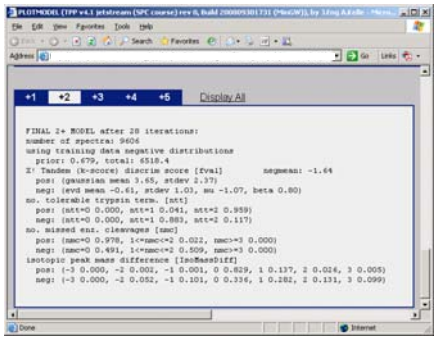
65

PeptideProphet Results: Predicted Numbers of Correct Peptides

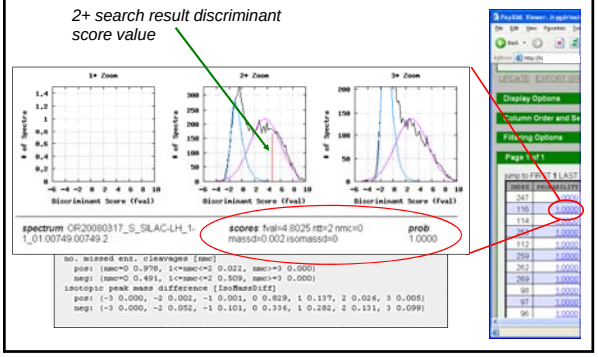


66

PeptideProphet Results: Model Summary



PeptideProphet Results: Contributing Score and Peptide Properties

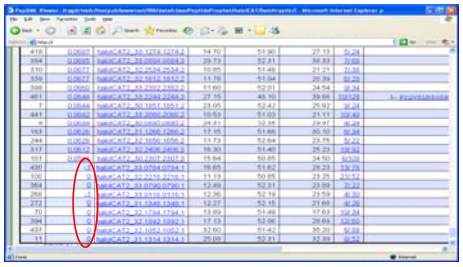


PeptideProphet $[M+2H]^{2+}$ vs $[M+3H]^{3+}$ Precursor Ions

334	1.0000	haloICAT2_33.1062.1062.3	51.06	51.94	36.29	10/44
338	0.9984	haloICAT2_33.1042.1042.2	33.62	51.98	33.58	6/22
357	0.9996	haloICAT2_33.1034.1034.2	27.70	51.85	26.77	7/18
331	0.9696	haloICAT2_33.1024.1024.3	24.96	52.07	32.23	12/44
386	0.4275	haloICAT2_33.1014.1014.3	17.19	49.67	27.34	6/96
372	0.5726	haloICAT2_33.1014.1014.2	41.36	51.37	27.11	15/36
373	0.9992	haloICAT2_33.1012.1012.2	34.49	51.39	35.29	11/36
312	0.6340	haloICAT2_33.1004.1004.1	27.62	52.49	33.39	7/22
330	0.9992	haloICAT2_33.1002.1002.2	26.84	52.06	29.37	8/22

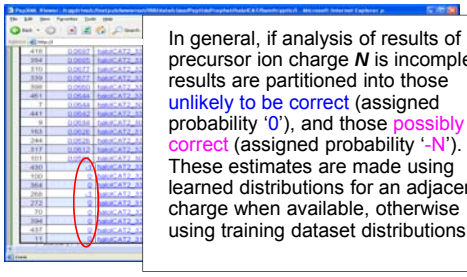
Spectrum searched as both 2+ and 3+ precursor received significant probability

PeptideProphet Results: Incomplete Analysis



Model incomplete for results of 1+ precursor ions

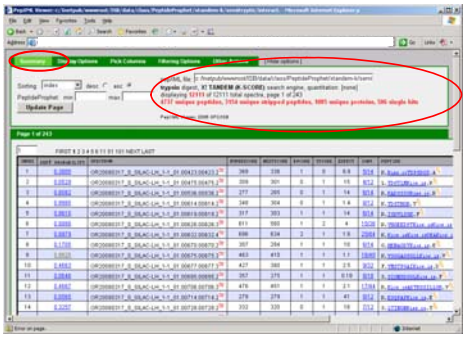
PeptideProphet Results: Incomplete Analysis



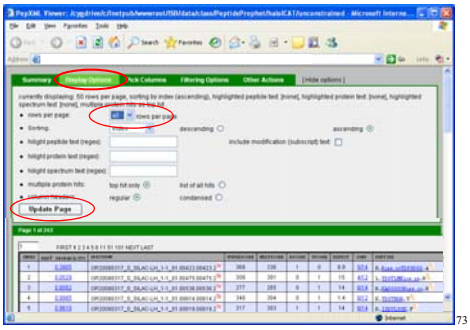
In general, if analysis of results of precursor ion charge ***N*** is incomplete, results are partitioned into those unlikely to be correct (assigned probability '0'), and those possibly correct (assigned probability '-N'). These estimates are made using learned distributions for an adjacent charge when available, otherwise using training dataset distributions

Model incomplete for results of 1+ precursor ions

Summary of Displayed Results

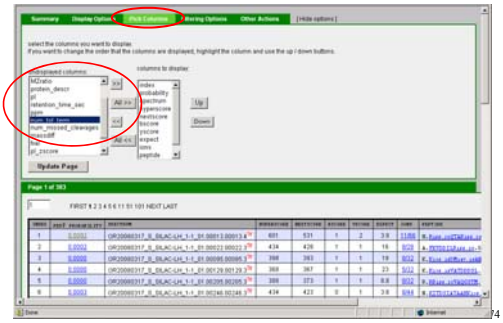


Number of Results to Display



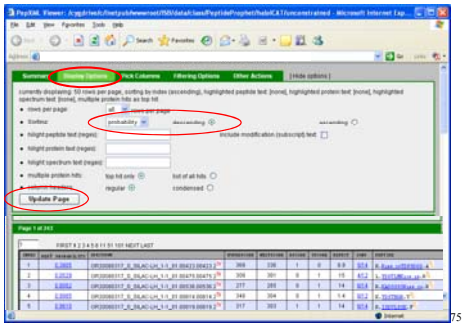
73

Pick Columns to Display



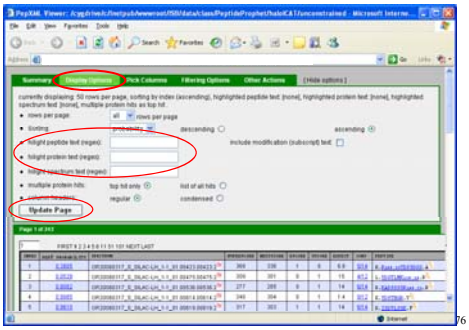
74

Sort Data by Computed Probability



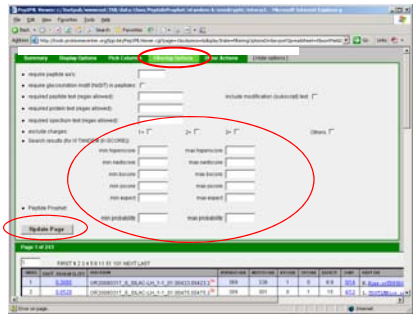
75

Color Specified AA's and Proteins



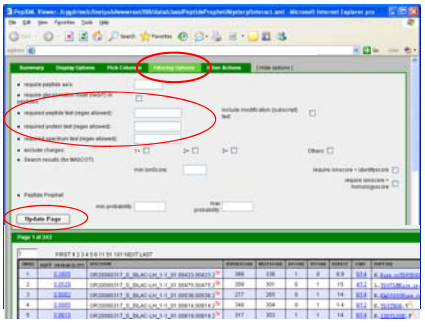
76

Filter Data by X! Tandem Scores



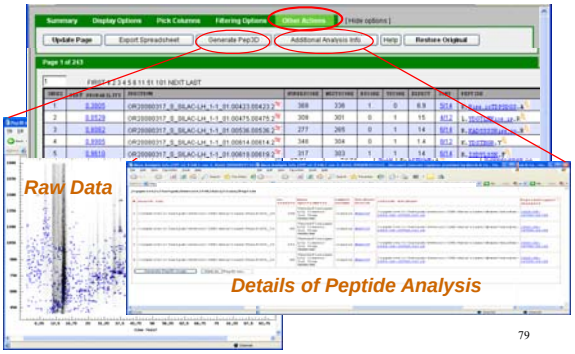
77

Select Specified AAs and Proteins



78

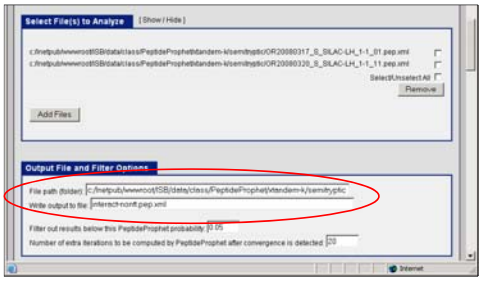
Pep3D and Analysis Summary Links



79

User Options for PeptideProphet

Rename Output File (*e.g.* to **`interact-nontt.xml`**):

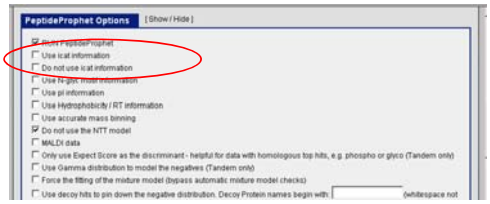


80

Use of Supplemental Discriminating Information

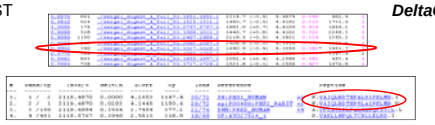
Use additional discriminating information, including ICAT or N-glyc, when relevant

- PeptideProphet automatically uses ICAT information when it thinks appropriate
- Nevertheless, you can explicitly set whether or not ICAT information is utilized

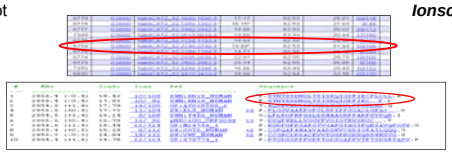


DeltaCn* and Ionscore* Examples

- Search results are marked with asterisked DeltaCn score (SEQUEST) or Ionscore (Mascot) when runner up peptide(s) share at least 75% sequence identity with top peptide
- SEQUEST ***DeltaCn****



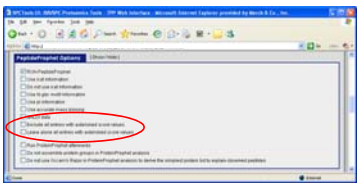
- Mascot ***Ionscore****



DeltaCn* and Ionscore* Options

There are three ways asterisked DeltaCn and Ionscores can be treated by PeptideProphet:

- Penalize (the default option, sets DeltaCn scores to 0, halves Ionscore values)
- Leave alone (suitable for the context of homologues)
- Exclude (the most conservative, assigns probability 0)



Run/Don't Run PeptideProphet

84

Exercises with PeptideProphet

- Accuracy of computed probabilities
- Utility of conventional search score thresholds and PeptideProphet analysis
- Model results for semi-tryptic data analyzed with and without NTT information
- Model results for Mystery dataset

85

Exercise Datasets

Many exercises utilize X! Tandem K-score search results of a ***SILAC yeast orbitrap*** dataset searched with a database containing equal numbers of non-decoy and decoy entries:

- Decoy results comprise 50% of incorrect results
- Error Rate = 2 * # decoy / # tot

Decoy results are easily distinguished from non-decoy results by their protein name beginning with 'REV'.

86

PeptideProphet Exercise

Introduction: Yeast SILAC Orbitrap semi-tryptic XTandem-K dataset

1. Analyze with PeptideProphet the semi-tryptic XTandem-K search results of the Yeast SILAC Orbitrap dataset (search conditions with the requirement that all peptides have at least one tryptic terminus), present in **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/semitryptic.** This dataset includes two XTandem-K pepXML files. Select both pepXML files (OR20080317_S_SILAC-LH_1-1_01.pep.xml and OR20080320_S_SILAC-LH_1-1_11.pep.xml) for analysis. Next, set a **"minimum probability of 0"** to retain even very low probability results. Make sure the location of the output file is set for **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/semitryptic**, select the option to **"Use accurate mass binning"** since this is a high-mass-accuracy analysis, and set the **"CLEVEL"** parameter to "**-1"** to allow modeling of lower scoring matches, then push the **'Run XInteract'** button at the bottom of the form. Select 'Show' on the Command Status to follow progress of the analysis. A link to the results **interact.pep.shtml** will appear when the analysis is complete.

Open the results link and click on any probability link. What is the total number of correct results predicted by the model? Do the learned discriminant score distributions among correct (pos) and incorrect (neg) results look reasonable, given the total distributions for the dataset?

Now scroll down the page. What distributions of discriminant score, numbers of tolerable tryptic termini (NTT), numbers of missed enzymatic cleavages (NMC), and isotopic mass offsets did the model learn for the correct, and for the incorrect, search results?

2. Using **Display Options** of the pepXML Viewer, sort the results by probability score (descending), and also color the decoy proteins red by typing 'REV' in the 'highlight protein text' box.

> *Note: The database in this search was appended with a set of decoy sequences (protein names beginning with REV) generated from the correct database by retaining the position of all potential cleavage targets (and specific non-cut targets) and reversing the sequence of the peptides. Because of the similar sizes of the decoy and non-decoy parts of the database we make the assumption that the matches to decoy sequences represent roughly half of the total number of incorrect hits.*

Open page 382 containing data with probability 0 identifications (if you don't see any data with probabilities below 0.05 it is because you forgot to specify a **minimum probability of 0**, so you need to redo step 1). Verify that approximately half of the 0 probability peptides correspond to REV sequences)

Do the same for page 227 of the data, containing identifications with probabilities close to 0.5.

Do the same for page 220 of the data, containing identifications with probabilities close to 0.8.

Do the same for pages 216 of the data, containing identifications with probabilities close to 0.9.

Finally do the same for page 1 of the data, containing identifications with probabilities close to 1

The TPP also provides a tool that will compare the PeptideProphet probability estimated FDR to the decoy estimated FDR. To access this tool use Petunia and go to the **Decoy → Decoy Peptide Validation** tab. As the input select the **interact.pep.xml** file that was generated in Step 1. Under Options change the decoy tag to **REV** then click run **Peptide Decoy Validation**.

> *Note: This tool generates a number of png image files all starting with **interact.pep**_* . Some of the images generated come from the FVAL models and kernel-density-based models in PeptideProphet (e.g. Accurate Mass, Retention Time, pI). For this exercise we want to open the images that compare FDR estimates based on PeptideProphet probabilities and Decoy information: interact.pep_FDR.png, and for an expanded view of the 0 to 0.05 FDR range, interact.pep_FDR_5pc.png*

Open the image file **interact.pep_FDR.png** in the output directory. How do PeptideProphet estimated FDRs compare against decoy estimated FDRs across the entire FDR range?

3. Filter the dataset using Tandem's conventional Expect Score thresholds: Max Expect 0.2

How many correct and incorrect peptide assignments result? Hint: The total number of results is displayed when you select **Summary** in the pepXML viewer. To determine the number of incorrect results, filter additionally for proteins that have REV in their name and multiply 2 times their number.

Assume the total number of correct search results in this dataset is 11279. Compute the sensitivity (fraction of correct results in dataset that pass filter) and false positive error rate (fraction of results passing filter that are incorrect) resulting from the use of the conventional threshold filter.

How does this sensitivity compare with that predicted by the PeptideProphet model for this dataset to achieve a similar error rate (click on any probability to view a detailed graph/table of predicted sensitivity and error values)?

Effect of NTT information on the model

4. Keep your previous pepXML viewer open so you can compare it with the next PeptideProphet analysis of the semi-tryptic search results of the Yeast SILAC Orbitrap dataset present in **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/semitryptic**, but this time ***not* using NTT information** and renamed 'interact-nontt.pep.xml'. As before, add all 4 XML files present in the directory. Make sure the location of the output file is set for **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/semitryptic**. Type 'interact-nontt.pep.xml' in the text box following 'Write output to file', exclude NTT information by selecting the checkbox before '**Do not use NTT model**' in the PeptideProphet Options section of the form, select the option to **"Use accurate mass binning"** as in Step 1, and then push the 'Run XInteract' button at the bottom of the form. A link to the results **interact-nontt.pep.shtml** will appear when the analysis is complete.
Compare the predicted number of correct peptide assignments as a function of predicted error rate for the models learned here and in Step 1 using NTT information (click on any probability to access this information). Which analysis yields more correct peptide assignments at an error rate of 2.5% or at an error rate of 5%?
If time permits, sort the results by probability (descending) and assess the model accuracy as done in Step 2, highlighting in red proteins with 'REV' and displaying 50 rows per page. Hint: Display pages 244, 231, 199, and 1 for average probabilities close to 0.2, 0.5, 0.9, and 1, respectively.

5. Pick out two search results of parent charge 2+ in this dataset that were assigned probabilities close to 0.5 without using NTT information: one result with an assigned peptide containing 2 tryptic termini, and one result with an assigned peptide containing 1 tryptic terminus. (Go to the **Columns** tab in the PepXMLViewer and add the **num_tol_term** parameter to the list of displayed columns, go to the **Filtering Options** tab and select results in the probability range **[0.47, 0.53]** and exclude charges other than 2, and finally select **Update Page**). Write down the spectrum name, computed probability, and number of tryptic termini of your two selected search results. Now look for each of those spectra in the analysis you performed previously that employed NTT information (**interact.pep.shtml**). To find a result, type the spectrum name into the 'required spectrum text' field in the **Filtering Options** tab and select **Update Page**. What probability was computed for the result with assigned peptide containing 2 tryptic termini in the analysis that used NTT information? For the result with assigned peptide containing only 1 tryptic terminus? What might explain these observations? [Note that if you can't find the search result for one of your spectra, it might have been filtered out if you neglected to set the minimum probability to 0 when launching the analysis in step 1

6. Look at the results in the middle probability range **[0.47, 0.53]** in **interact-nontt.pep.xml**. Be sure to adjust the filter to display results of all charges. In **Display Options**, select to color the decoy proteins red by typing 'REV' in the 'highlight protein text' box, and select to view **all rows per page**. Assess the accuracy of PeptideProphet probabilities among all 116 results Do you see the same accuracy when you first select for results with two tryptic termini by selecting in the **Filter Options** to **filter for results with NTT greater than or equal to 2**? What might account for the difference? Do you see a similar pattern when you assess in the same way results in the middle probability range **[0.47, 0.53]** of the analysis that used NTT information in step 1?

Mystery dataset: More comparisons of the model with standard score thresholds

7. Analyze *without* running PeptideProphet the XTandem-K search results of the Mystery dataset present in **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/mystery.** Go to 'Analysis Pipeline' and 'Analyze Peptides' again and go to the **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/mystery** directory and select for analysis the two XML files present there, **OR20080317_S_SILAC-LH_1-1_01.pep.xml** and **OR20080320_S_SILAC-LH_1-1_11.pep.xml**. Make sure the location of the output file is set for **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/mystery**, and rename the output by typing 'interact-noprob.pep.xml' in the text box following 'Write output to file'. Turn off PeptideProphet by unchecking the box following 'RUN PeptideProphet' in the form, then push the 'Run XInteract' button at the bottom of the page. A link to the results (**c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/interact-noprob.pep.shtml**) will appear when the analysis is complete.
Next, filter the dataset using conventional XTandem score thresholds:

Expect ≤ 0.05,
Expect ≤ 0.1,
Expect ≤ 0.2.
How many identifications pass the filter?

8. Now run XInteract *with* PeptideProphet on search results of the Mystery dataset present in **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/mystery**, setting a **"minimum probability of 0"** to retain even very low probability results. Look at the results in your browser at **c:/Inetpub/wwwroot/ISB/data/class/PeptideProphet/xtandem-k/mystery/interact.pep.shtml.**

How many search results of 1+, 2+, and 3+ precursor ions are predicted to be correct? What may explain these observations? Click on any probability link to view the models learned by PeptideProphet. Next, use Pep3D to view the Mystery RAW data by clicking on the 'Generate Pep 3D' bar in the **Other Actions** options of the pepXML Viewer. Set 'Display peptides' to **None**, then click the **Generate Pep3D image** button to display LC/MS data for the two runs. How does the quality of the two LC/MS runs look? How about the quality of sample MS/MS spectra viewed from **interact.pep.shtml**? What explanations for the results of PeptideProphet remain viable?

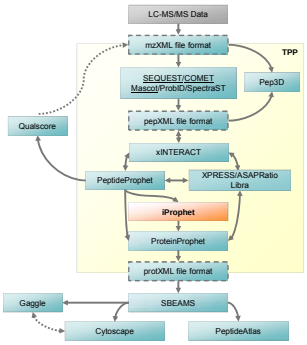
iProphet: Statistical Refinement of PeptideProphet Results

David Shteynberg

Day 2

October 26, 2010

Unique Peptide Sequence Validation

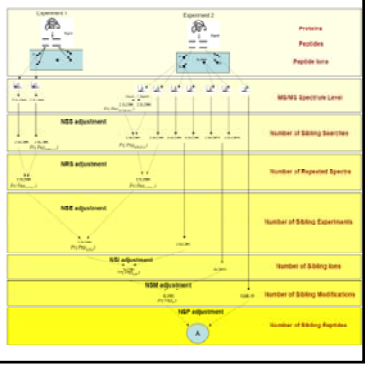


iProphet: What is it?

- An intermediate between PeptideProphet and ProteinProphet
- A statistical tool that considers additional information about Peptide Spectrum Matches not considered by PeptideProphet or ProteinProphet
- GOAL: Compute accurate unique peptide sequence probability

iProphet Statistical Models

- **NSS**
  - Number Sibling Searches
- **NRS**
  - Number Replicate Spectra
- **NSE**
  - Number Sibling Experiments
- **NSI**
  - Number Sibling Ions
- **NSM**
  - Number Sibling Modifications



NSS Model

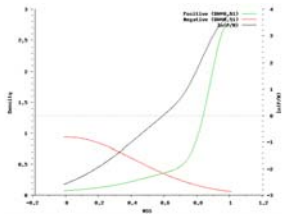
- Number of Sibling Searches
  - Statistic used to represent parallel searches of the same spectrum with different search engines
  - Ratio of probabilities of matching IDs to the total number of searches

$$NSS_d = \frac{\sum_{\{d' | d \neq d' \wedge Pep_d = Pep_{d'}\}} \Pr(Pep_{d'})}{\sum_{\{d' | d \neq d'\}} 1}$$

$$\Pr(+ \mid NSS_d) = \frac{\Pr(NSS_d \mid +)\Pr(Pep_d)}{\Pr(NSS_d \mid +)\Pr(Pep_d) + \Pr(NSS_d \mid -)(1 - \Pr(Pep_d))}$$

$$\Pr(NSS_d \mid +) = \frac{\sum_{\{\forall Pep \mid NSS(Pep) = NSS_d\}} \Pr(Pep)}{\sum_{\forall Pep} \Pr(Pep)} \qquad \Pr(NSS_d \mid -) = \frac{\sum_{\{\forall Pep \mid NSS(Pep) = NSS_d\}} (1 - \Pr(Pep))}{\sum_{\forall Pep} (1 - \Pr(Pep))}$$

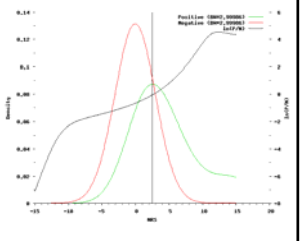
NSS Model Example



- Dataset searched with 5 different search engines: OMSSA, Sequest, XTandem, Myrimatch and Inspect

NRS Model

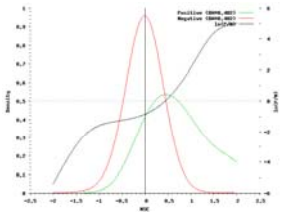
- Number of Replicate Spectra
  - Statistic used to represent repeated identifications of the same **peptide ion** in 1 experiment
  - Replicate spectra contribute a positive value to NRS if their PeptideProphet probability is above 0.5, and a negative value if their PeptideProphet probability is below 0.5
    - Prevent single spectrum wonders from getting punished by this model



$$NRS_i = \sum_{\{i' | i \neq i' \wedge Pep_i = Pep_{i'}\}} (\Pr(Pep_{i'}) - 0.5)$$

NSE Model

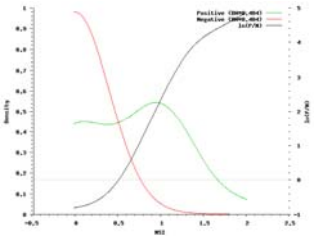
- Number of Sibling Experiments
  - Statistic used to represent repeated identifications of the same **peptide ion** across different experiment
  - Replicate spectra in a different experiment contribute a positive value to NSE if their PeptideProphet probability is above 0.5, and a negative value if their PeptideProphet probability is below 0.5
    - Prevents single experiment wonders from getting punished by this model



$$NSE_x = \sum_{\{x' | x \neq x' \wedge Pep_x = Pep_{x'}\}} (\Pr(Pep_{x'}) - 0.5)$$

NSI Model

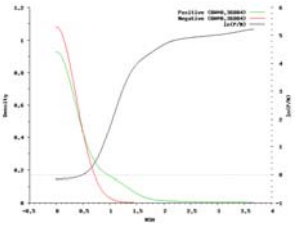
- Number of Sibling Ions
  - Statistic used to represent identifications of the same **peptide+mods** across different charge states
  - Same peptide sequence
  - Same peptide modification
  - Different charge state



$$NSI_z = \sum_{\{z' | z \neq z'\}} \Pr(Pep_{z'})$$

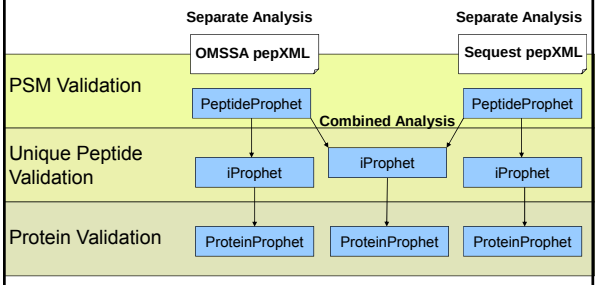
NSM Model

- Number of Sibling Modifications
  - Statistic used to represent identifications of the same **peptide** sequence across different modified states
  - Same peptide sequence
  - Different peptide modifications



$$NSM_m = \sum_{\{m' | m \neq m'\}} \Pr(Pep_{m'})$$

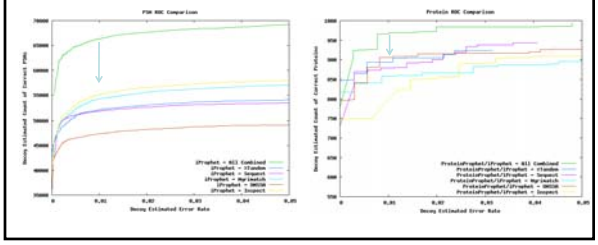
TPP Information Flow



PSM = Peptide Spectrum Match
Unique Peptide = Sequence Stripped of Charge and Modification

Results

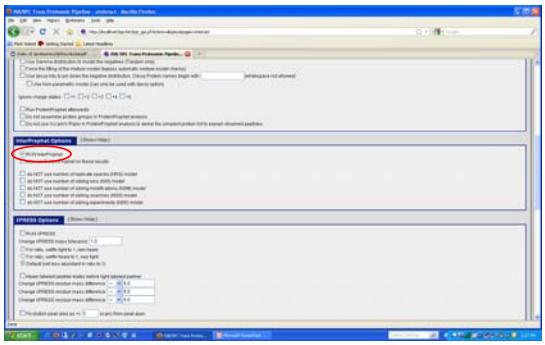
- *S.pyogenes* PeptideAtlas data
- *S.pyogenes* plus Human IPI database
  - (2:1:1) Forward:Decoy1:Decoy2
    - Two Decoy sets generated independently by randomizing tryptic sequences from the Forward database



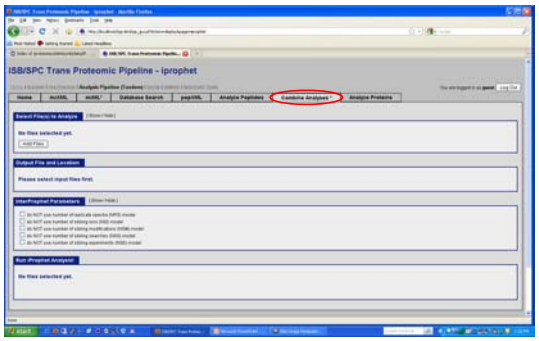
iProphet Considerations

- iProphet is a work in progress, publication is in prep due out this year
- The single search analysis runs in parallel to the standard TPP
- Need to look at the models carefully to make sure that they "make sense"
- Models that don't "make sense" need to be disabled and iProphet result regenerated manually
- Theoretically iProphet should always be as good or better than PeptideProphet
- iProphet performance improves with combined searches and large datasets

Running iProphet: Single Search Analysis



Running iProphet: Combined Search Analysis



iProphet – Tutorial

1. Using Petunia go to the "*Analyze Peptides*" tab. Add the following files to the analyses:

    a. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ iProphet \ xtandem \ OR20080317_S_SILAC-LH_1-1_01.pep.xml**
    b. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ iProphet \ xtandem \ OR20080317_S_SILAC-LH_1-1_11.pep.xml**

2. In the "*PeptideProphet Options*" pane select to "*Use accurate mass binning"*, and set the *"CLEVEL"* parameter to "-1". In the "*InterProphet Options*" pane select "*RUN InterProphet*".

3. To run this analysis Click on *"Run XInteract"*. When the program is finished the results can be accessed through files "**c:\ Inetpub \ wwwroot \ ISB \ data \ class \ iProphet \ xtandem \ interact.pep.shtml**" and "**c:\ Inetpub \ wwwroot \ ISB \ data \ class \ iProphet \ xtandem \ interact.iproph.pep.shtml**".

4. In Petunia use the "*Browse Files*" utility to examine the contents of the "**c:\ Inetpub \ wwwroot \ ISB \ data \ class \ iProphet \ xtandem**". Open the png files of the probability models:

File Name	Model
interact.iproph.pep_NRS.png	Number of Replicate Spectra
interact.iproph.pep_NSI.png	Number of Sibling Ions
interact.iproph.pep_NSM.png	Number of Sibling Modifications

5. In each plot, the green curve represents the positive distribution; the red curve represents the negative distribution. The black curve represents the log-ratio of the positive to negative probability densities. When the black curve is below the horizontal dotted line an ID with a corresponding value is less probable, when the black curve is above the dotted line an ID with a corresponding value is more probable. Examine each model. Do the models look reasonable? Why are the NSS and NSE models not displayed?

6. Open the link to file: **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Quantitation \ xtandem-k \ semitryptic \ interact.iproph.pep.shtml**. The **PeptideProphet** probabilities are displayed in the "*PEPP PROBABILITY*" column, the **iProphet** probabilities are displayed in the "*IP PROBABILITY*" column.

7. Count the number of IDs with a **PeptideProphet probability of atleast 90%**. How many of these IDs match the "REV" proteins?

8. Count the number of IDs with a **iProphet probability of atleast 90%**. How many of these IDs match the "REV" proteins?

9. Using Petunia open the "*Utilities → Decoy Peptide Validation*" tab, in the first section specify the iProphet pepXML file "**c: \ Inetpub \ wwwroot \ ISB \ data \ class \ iProphet \ xtandem \ interact.iproph.pep.xml**" as input. Under "*Options*" change the decoy tag to REV and leave all other options at the default settings.

   Note: This tool generates a number of png image files all starting with ***interact.iproph.pep_**** . *Some of the images generated come from the FVAL models and kernel-density-based models in PeptideProphet (e.g. Accurate Mass, Retention Time, pI). For this exercise we want to open the image that compares decoy estimated ROC curves: interact.iproph.pep_ROC.png*

10. According to the ROC curve in the file *interact.iproph.pep_ROC.png*, between 0 and 5% Decoy estimated FDR, which tool performs better in terms of the number of correct PSM hits?

11. Now open the images that compare Model-Estimated FDR to Decoy-Estimated FDR. Global FDRs are compared in files: interact.iproph.pep_FDR.png and interact.iproph.pep_FDR_5pc.png. Which tool provides a more conservative model as compared to decoy results?

12. Just like in step 9, using Petunia open the "*Utilities → Decoy Peptide Validation*" tab, in the first section specify the iProphet pepXML file "**c: \ Inetpub \ wwwroot \ ISB \ data \ class \ iProphet \ xtandem \ interact.iproph.pep.xml**" as input. Under "*Options*" change the decoy tag to REV and this time enable the checkboxes next to the options: "*Consider only best iProphet probability for each unique peptide sequence*" and "*Consider only best PeptideProphet probability for each unique peptide sequence*."

13. According to the ROC curve in the file *interact.iproph.pep_ROC.png*, between 0 and 5% Decoy-Estimated FDR, which tool performs better in terms of the number of correct unique peptide sequence hits?

14. Now open the images that compare Model-Estimated FDR to Decoy-Estimated FDR. Global FDRs are compared in files: interact.iproph.pep_FDR.png and interact.iproph.pep_FDR_5pc.png. Which tool provides a more conservative model as compared to decoy results?

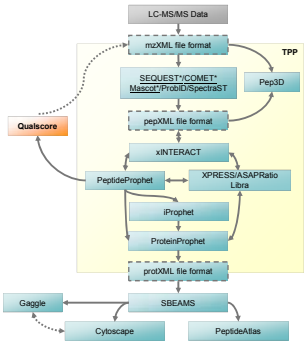
QualScore

Luis Mendoza

Day 2

October 26, 2010

QualScore



* Commercial software not part of TPP

Fraction of Spectra Left Unassigned in a Typical Search

Typical "bread and butter" MS/MS search
- SEQUEST, IPI database
- semi-constrained (tryptic on one end)
- Met + 16
- +/- 3 Da, average mass

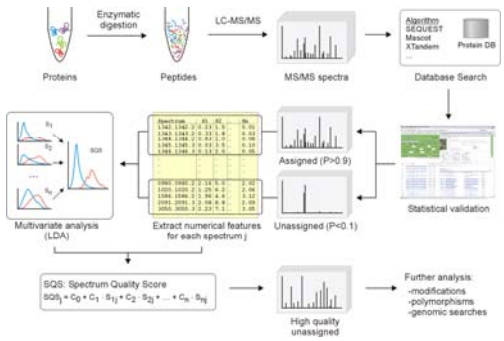
Average numbers (mix of ICAT/non-ICAT experiments):
- 10-15% of all spectra assigned peptide with high confidence
- 20-25 % of all high quality spectra are not assigned

What are these spectra?
- Biologically interesting peptides not in the database
  - Novel proteins
  - Novel splice variants
  - SNPs
- Modified peptide forms

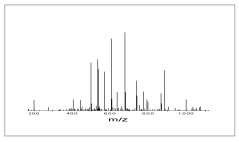
Why unassigned?

Possible causes of failure to assign peptide:
- Imperfect scoring scheme
- Constrained search (PTM, not tryptic etc.)
- Incorrect mass/ charge state
- Low spectrum quality / contaminant ion
- Correct sequence may not be in the database searched (e.g. SNP)
- Novel sequence (splice variants)

Finding and Mining High Quality Unassigned Spectra



QualScore



n spectral features

$x_1, x_2 \ldots x_n$

Composite score (QualScore):

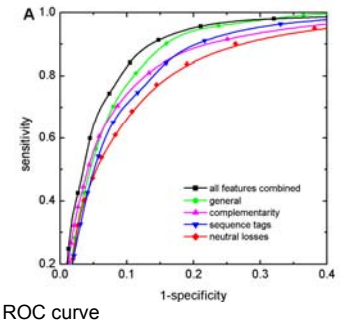
$$QS = c_1x_1 + c_2x_2 + c_3x_3 + \ldots + c_nx_n$$

Linear discriminant function approach
- Coefficients c_i indicate statistical significance of each spectral feature x_i for spectral classification
- Dynamically trained (robustness)

Spectrum Features

- 8 general descriptive features
  - Number of peaks in the spectrum
  - Distribution of peak intensities (mean and stddev)
  - Smallest m/z range containing 95% of peak intensity
  - Smallest m/z range containing 50% of peak intensity
  - Total ion current per m/z feature
  - Standard deviation of consecutive m/z gaps
  - Average number of neighborhood peaks within a 2-Da windows around each peak
- Sequence tags
  - length of the longest sequence tag extracted using *de novo* type algorithm
    - Fast dynamic programming algorithm that assumes that all peaks correspond to singly charged b^+ and y^+ fragment ions
- Complementary ion pairs
  - number of fragment ion pairs that sum to the precursor ion mass
- Neutral losses
  - Ammonia (-17 Da)
  - Water (-18 Da)
  - Carbon Monoxide (-28 Da)

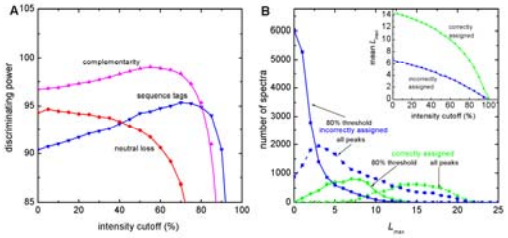
Statistical Significance of Spectrum Features



- Combining different classes of features improves performance of the classifier
- Individually: general spectrum features are best for filtering out bad quality spectra
- Assuming a peptide ion precusor:
  - Complementary ion pairs are best for finding high quality spectra of peptides

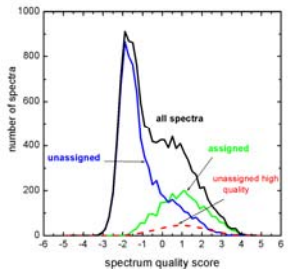
ROC curve

Optimization of Spectrum Features



- Complementary ion pairs, sequence tags and neutral losses are computed using high intensity peaks only
- Optimal signal / noise threshold is different for each class

Human Raft Dataset



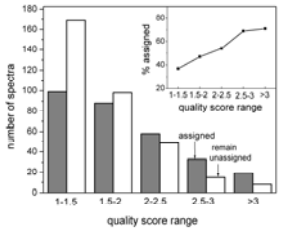
SEQUEST ('ISB typical')
PeptideProphet
↓
Create training dataset:
unassigned: P < 0.01
assigned: P > 0.95
↓
Dynamically train Classifier
↓
Apply to all spectra in the dataset; compute quality score for each spectrum

Potentially **~25%** gain in the number of high confidence IDs

Reanalysis of Unassigned High Quality Spectra

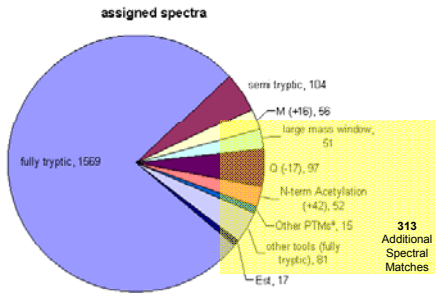
- Large mass tolerance search
  - SEQUEST, semi-tryptic, 4Da mass tolerance (previously 3Da).
- Q -17 search
  - SEQUEST, semi-tryptic, 3Da mass tolerance, allowing for conversion of glutamic residues to pyroglutamic acid (loss of 17 Da) as a variable PTM.
- Mascot search
  - Mascot, tryptic peptides only, 2 missed cleavages or less, 3Da mass tolerance, allowing for N-terminal acetylation as a variable PTM.
- Miscellaneous searches
  - XTandem with more than one type of PTM per peptide
  - SEQUEST and Mascot allowing for PTMs not specified in the previous searches (e.g., conversion of N-terminal glutamic acid residues to pyroglutamic acid, phosporylation, acetylation, guanidation, and etc.).
- EST database search

Percent of Previously Unassigned Spectra Assigned After All Additional Searchers

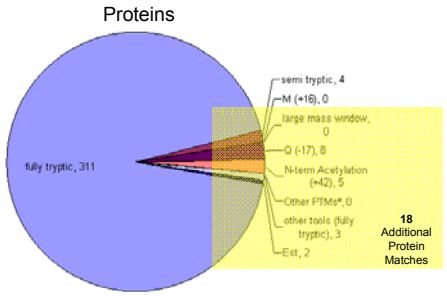


- The higher the spectrum quality, the more likely the spectrum is assigned with high confidence.
- Spectra of very high quality (QS>3) were unassigned in the initial search.
- More than 70% of these were eventually assigned

What Are Those Additional Identifications?

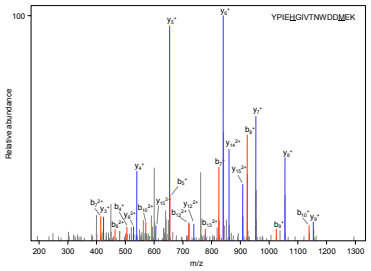


Do They Add Any New Proteins?



13

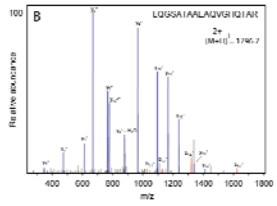
Any Biologically Interesting Peptides/Proteins?



YPIEHGIVTNWDDMEK from Actin, cytoplasmic 1 protein (SW:P02570) containing methylated histidine at position 5

14

Searching Genomic Databases



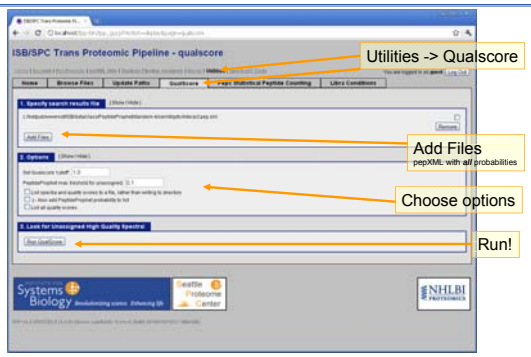
- Human lipid rafts
- Search against EST database

LQGSATAAEAQVGHQTAR (>10 EST sequences)

This intron-exon spanning peptide identifies a novel splice variant of the Lck-interacting transmembrane adaptor 1 protein (LIME1, NP_060276). LIME1 was shown to be a raft-associated protein in several recent studies.

15

Running QualScore



16

Getting Qualscore

- Requires Java
- Windows
  - Part of the TPP installation
- Linux
  - Must download separately
- Link to more info:

http://tools.proteomecenter.org/wiki/index.php?title=Software:QualScore

Nesvizhskii *et. al*. Mol Cell Proteomics. 2006 Apr;5(4):652-70. Epub 2005 Dec 12

17

Acknowledgements

Alexey Nesvizhskii

Mathijs Vogelzang
Franz Roos
Jonas Grossmann
Sasha Baginsky

James Eddes

Ruedi Aebersold

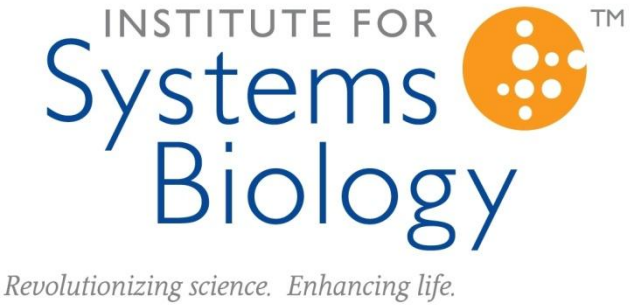
18

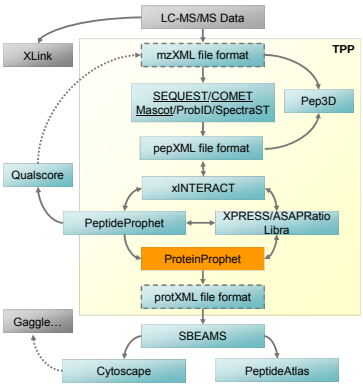
ProteinProphet: Statistical Validation of Protein Identifications

Luis Mendoza

Day 3

October 27, 2010





1

Outline

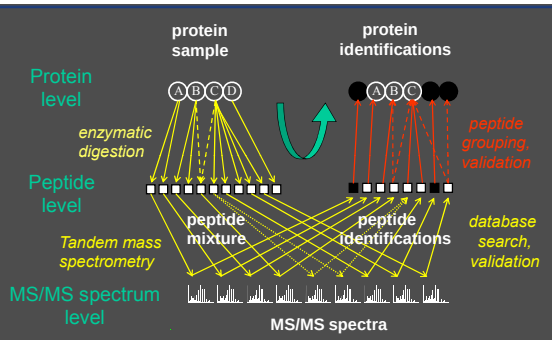
- ProteinProphet:
  1. peptide grouping
  2. adjusting peptide probabilities for protein grouping information
  3. protein inference problem
- Interpretation of shotgun proteomics data
- Data publication guidelines

References: A statistical model for identifying proteins by tandem mass spectrometry, A. I. Nesvizhskii *et al.*, *Anal. Chem.* 75, 4646-4658 (2003)

Interpretation of shotgun proteomic data: The Protein Inference Problem, A. I. Nesvizhskii, R. Aebersold, *Mol. Cell. Proteomics* 4, 1419-1440 (2005)

2

Shotgun Protein Identification



3

Statistical Model for Proteins (simple model)

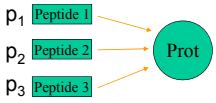
- Group peptides by protein
- Compute a probability that a protein is present in the sample based upon the evidence of corresponding peptides in the dataset

4

Computing Protein Probabilities

Compute probability that at least one corresponding peptide is a correct identification

$$P(\text{protein}) = 1 - \Pi_i (1 - p_i)$$



$$P_{\text{incorrect}} = (1 - p_1)(1 - p_2)(1 - p_3)$$

$$P = 1 - (1 - p_1)(1 - p_2)(1 - p_3)$$

5

Computing Protein Probabilities

>sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG) (ALLERGEN BOS D 5) - Bos taurus (Bovine).

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKLV
LDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHI
RLSFNPTQLEEQCHI

PTPEGDLEILLQK : p = 0.81

TPEVDDEALEK : p = 0.96

LSFNPTQLEEQCHI : p = 0.65

P = 1 – (1-0.81)(1-0.91)(1-0.65) = 0.99

Repeated Sequencing Events

>sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG) (ALLERGEN BOS D 5) - Bos taurus (Bovine).

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQK WENGECAQKKIIAEKTKIPAVFKIDALNENKLV
LDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHI
RLSFNPTQLEEQCHI

PTPEGDLEILLQK : p = 0.81
PTPEGDLEILLQK : p = 0.62
PTPEGDLEILLQK : p = 0.95

same peptide sequenced multiple times

P = ??

Repeated Sequencing: Example

4 MS/MS spectra

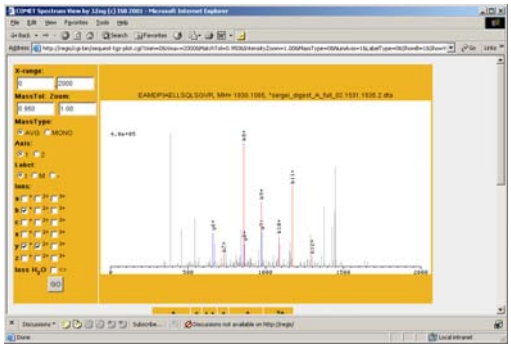
high probability peptide assignments

consistent retention time as indicated by the scan numbers

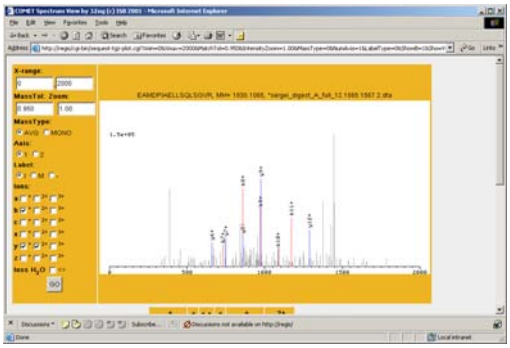
match to the same database peptide

note: spectra searched using SEQUEST with trypsin specified as the digestion enzyme

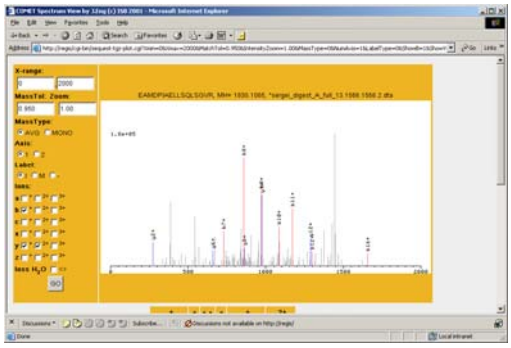
Spectrum 1



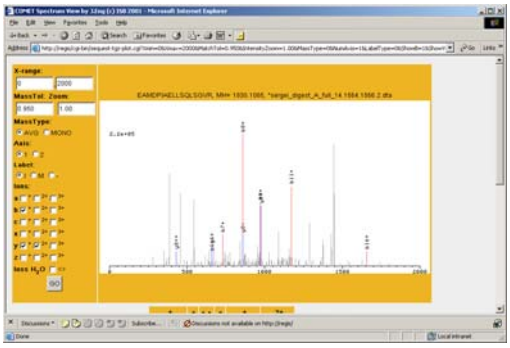
Spectrum 2



Spectrum 3



Spectrum 4



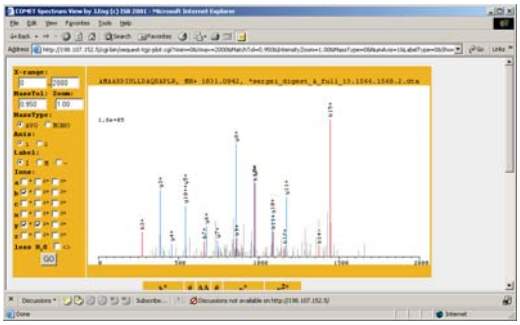
All 4 peptide assignments shown above are incorrect!

Search without enzyme specificity reveals the true peptide (partially tryptic):

Rank/Sp	(M+H)+	deltCn	XCorr	Sp	Ions	Reference	Peptide
1 / 1	1831.0942	0.0000	3.6893	1636.0	20/34	sp|P02754|LACB_BOVIN	L.AMAASDISLLDAQSAPLR.V
2 / 25	1830.1065	0.3403	2.4338	715.4	15/32	PIR2:T34540	+3 R.EAMDPIAELLDQLHQVR.R
3 / 10	1830.1035	0.3483	2.4042	827.1	16/34	SW:GBAZ_HUMAN	+1 L.TQPAEVRGRITPELLGVR.R
4 /126	1830.0453	0.3855	2.2671	554.0	15/32	GP:AR032956_1	T.EDNQDEVVSTLELKPAD.R
5 /344	1830.1956	0.4158	2.1553	451.5	13/34	GP:U49881_1	R.TPSVASKQVVLLMHQLYG.S
6 / 13	1828.0752	0.4214	2.1345	786.5	15/34	GP:BC007453_1	S.TPLASFKILALESADSHQ.G
7 / 74	1830.0079	0.4277	2.1114	610.4	13/30	SW:CO1_HUMAN	+1 I.RSLTPTSNLLDQQQQQ.Q
8 /188	1830.0483	0.4357	2.0820	508.2	13/30	GP:AF163751_1	+1 Q.YENNQALSLLSKFQEQQ.R
9 / 57	1830.0068	0.4363	2.0798	640.8	13/28	SWN:EFP1_HUMAN	G.QPVSAWELINSKYTS.K
10 / 49	1829.8743	0.4379	2.0736	653.2	15/32	GP:AF055994_1	+1 S.SSSGHSQSTLFDSDVFQ.T

sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG) (ALLERGEN BOS D 5) - Bos taurus (Bovine).
PIR2:T34540 hypothetical protein DKFZp434L1021.1 - human (fragment) [MASS=30460]

Correct Assignment

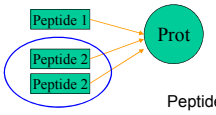


Assignments of Multiple Spectra to the Same Peptide

Assignments to the ***same*** peptide are not independent

conservative approach:

$$P(\text{protein}) = 1 - \Pi_i \, (1 - max_j \, p_i^j)$$



$$P = 1 - (1 - p_1)(1 - max\,(p_{2A}, p_{2B}))$$

Peptide assignment with highest probability is kept, other instances of the same peptide do not contribute to the protein probability.

Treatment of Repeated Sequencing Events

>sp|P02754|LACB_BOVIN BETA-LACTOGLOBULIN PRECURSOR (BETA-LG) (ALLERGEN BOS D 5) - Bos taurus (Bovine).

MKCLLLALALTCGAQALIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSA
PLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKLV
LDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHI
RLSFNPTQLEEQCHI

same peptide sequenced multiple times

PTPEGDLEILLQK : p = 0.81
PTPEGDLEILLQK : p = 0.62
PTPEGDLEILLQK : p = 0.95 → protein P = 0.95

Note: identifications of the same peptide sequence from MS/MS spectra of different charge state are still considered 'independent' events. Same in the case of modifications (light/heavy isotope tags, oxidation etc.)

Issues for Protein Identification

- Peptides corresponding to 'single-hit' proteins are less likely to be correct than those corresponding to 'multi-hit'proteins
- Many peptides are present in more than a single protein (isoforms, homologous proteins, etc.)

A. I. Nesvizhskii *et al.*, *Anal. Chem.* **75**, 4646-4658 (2003)
A. I. Nesvizhskii, R. Aebersold, *MCP* 4, 1419-1440 (2005)

Problem # 1

Non-random Grouping of Peptides according to Their Corresponding Protein

(single-hit protein identification problem)

Non-random Grouping of Peptides

Correct peptide assignments tend to correspond to "multi-hit" proteins, those to which other correctly assigned peptides correspond

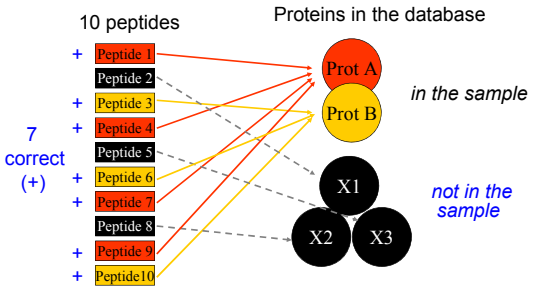
Incorrect peptide assignments tend to correspond to "single-hit" proteins to which no other correctly assigned peptide corresponds

False positive identification error rate on the protein level is higher than on the peptide level

Hard to distinguish single-hit correct proteins form the incorrect ones

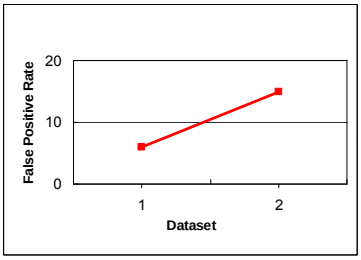
19

Non-random Grouping of Peptides according to Corresponding Protein



20

Protein FDR



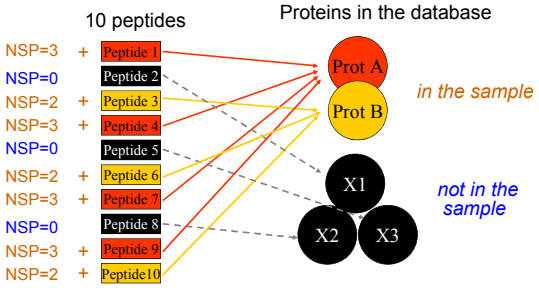
Apply same peptide-level threshold

Keep acquiring data and you will "identify" everything in the database!

Datasets:
1 Halobacterium, 4 runs
2 Halobacterium, 45 runs

21

Number of Sibling Peptides (NSP) as Measure of Protein Grouping



Number of sibling peptides (NSP): measure of how many other peptides in dataset correspond to same protein

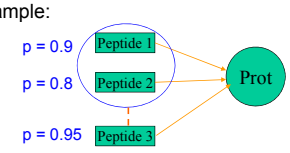
Number of Sibling Peptides (NSP)

peptide i ∈ protein P; sum over all other j ∈ protein P:

$$NSP_i = \sum_{j \neq i} p(+ \mid D_j)$$

peptide probability computed by PeptideProphet

example:



peptide 3 has 2 siblings, peptide 1 and peptide 2
It has NSP = 1.7 (calculated as 0.9 + 0.8)

23

Adjusting Peptide Probabilities for NSP

$$p(+ \mid D, NSP) = \frac{p(+ \mid D)\,p(NSP \mid +)}{p(+ \mid D)\,p(NSP \mid +) + p(- \mid D)\,p(NSP \mid -)}$$

adjusted probability

Amount of adjustment depends on $\frac{p(NSP \mid +)}{p(NSP \mid -)}$

- Learn NSP distributions from each dataset (EM algorithm)
- Adjust peptide probabilities to include NSP information

peptides with high *NSP* – more confidence → p ↑
peptides with low *NSP* – less confidence → p ↓

The appropriate amount of adjustment for NSP is determined by the model from the data

24

NSP Distributions

Why do we need to determine the appropriate amount of adjustment for NSP (e.g., by how much we should penalize single-hits) from the data?

NSP distributions among correct and incorrect peptide assignments, p(NSP|+) and p(NSP|-), vary from dataset to dataset. They strongly depend on sample "coverage".

coverage = dataset size / sample complexity

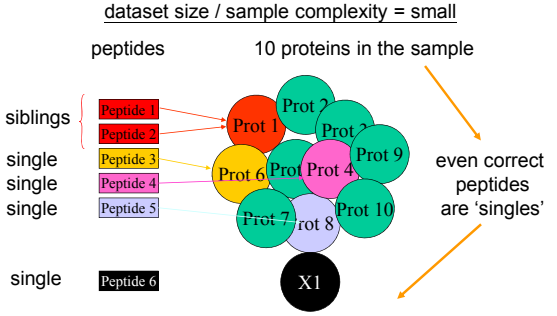
sample complexity: number of proteins

dataset size: number of peptide assignments (number of acquired MS/MS spectra)

25

Low Coverage Dataset

dataset size / sample complexity = small

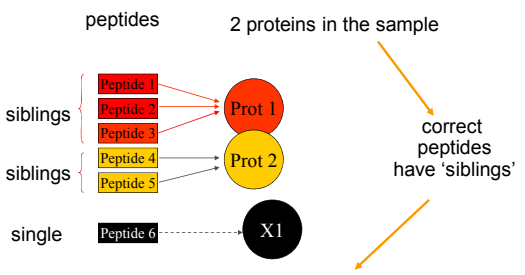


small difference between p(NSP|+) and p(NSP|-)

26

High Coverage Dataset

dataset size / sample complexity = large

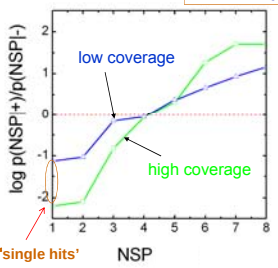


large difference between p(NSP|+) and p(NSP|-)

27

NSP distributions: dependence on sample coverage

before adjusting for *NSP*



Similar sample size (# of sequencing attempts)

Different sample complexity:

- complex *H. Influenzae* sample (low coverage)
- 18 protein sample (high coverage)

The higher the coverage, the bigger the adjustment for NSP

28

Adjusting Peptide Probabilities: Example 1

Given the initial peptide probabilities, the model "learns" that:

If a peptide has "1 sibling" (*NSP=0.5+0.5=1*), then *p(NSP|+) = 0.4* and *P(NSP|-) = 0.1* (that is, a peptide with 2 siblings is 4 times more likely to be correct assignment than incorrect one

10 peptides with p=0.5

Peptide 1
Peptide 2
Peptide 3
Peptide 4
Peptide 5
Peptide 6
Peptide 7
Peptide 8
Peptide 9
Peptide10

$$p' = \frac{p \cdot p(NSP \mid +)}{p \cdot p(NSP \mid +) + (1-p) \cdot p(NSP \mid -)}$$

$p(NSP|+) = 0.4 \quad p(NSP|-) = 0.1$

$$p' = \frac{0.5 \cdot 0.4}{0.5 \cdot 0.4 + (1-0.5) \cdot 0.1} = 0.8$$

Peptides 1, 4, and 7: adjusted p = 0.8

29

Adjusting Peptide Probabilities: Example 2

Given the initial peptide probabilities, the model "learns" that:

If a peptide has 0 sibling (*NSP=0*) , then *p(NSP|+) = 0.1* and *P(NSP|-) = 0.4* (that is, a peptide with no siblings is 4 times less likely to be correct assignment than incorrect one

10 peptides with p=0.5

Peptide 1
Peptide 2
Peptide 3
Peptide 4
Peptide 5
Peptide 6
Peptide 7
Peptide 8
Peptide 9
Peptide10

$$p' = \frac{p \cdot p(NSP \mid +)}{p \cdot p(NSP \mid +) + (1-p) \cdot p(NSP \mid -)}$$

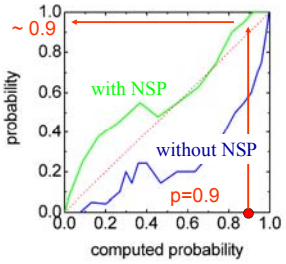
$p(NSP|+) = 0.1 \quad p(NSP|-) = 0.4$

$$p' = \frac{0.5 \cdot 0.1}{0.5 \cdot 0.1 + (1-0.5) \cdot 0.4} = 0.2$$

Peptides 2, 5, 6, 8, and 9: adjusted p = 0.2

30

Accuracy of Computed Protein Probabilities



100 proteins with computed p ~ 0.9

90% of them (90) should be correct

observed probability is around 0.9

Statistical model is more accurate with adjustment for protein grouping information (NSP)

31

ProteinProphet (with adjustment for NSP)

- Group peptides by protein
- Adjust peptide probabilities for grouping information (penalize 'single hits')
- Compute a probability that a protein is present in the sample based upon the evidence of corresponding peptides in the dataset (using adjusted peptide probabilities)

32

Problem # 2

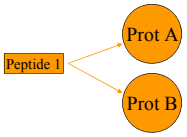
Protein Inference Problem

("degenerate" or "shared" peptides)

33

Protein Inference Problem

Shared peptides: correspond to more than a single entry in protein database



protein A or protein B ??

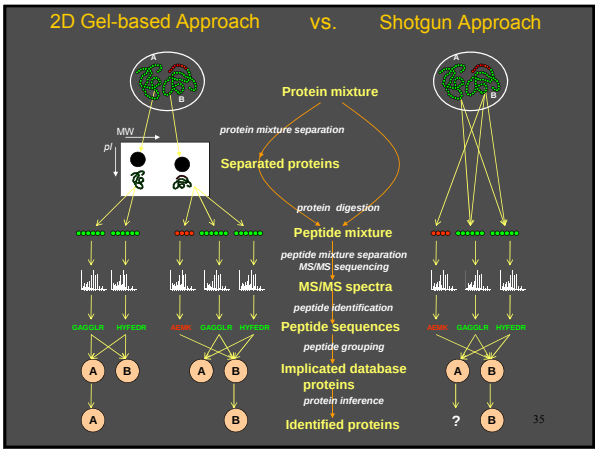
Or both?

In shotgun proteomics, the connectivity between peptides and proteins is lost

Shared peptides are more prevalent with databases of higher eukaryotes due to the presence of:

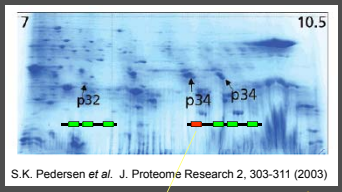
- related protein family members
- alternative splice forms
- partial sequences

34



35

Protein Isoforms and 2D-gels



S.K. Pedersen *et al.* J. Proteome Research 2, 303-311 (2003)

NADH-cytochrome b5 reductase

p32: soluble form, Inter-membrane space

p34: membrane bound, outer membrane of the mitochondrion

p32/p34 cleavage site

MFSRLSRSHSKALPIALGTVAIAAATAFYFANRNQHSFVFNESNKVFKGDDKWIDLPISK
IEEESHDTRRFTFKLPTEDSEMGLVLASALFAKFVTPKGSNVVRPYTPVSDLSQKGHFQL
VVWHYEGGKMTSHLFGLKPNDTVSFWGPIMKWKWQPNQFKSITLLGAGTGINPLYQLAHH
IVENPNDKTKVNLLYGNKTPQDILLRKELDALKEKYPDKFNVTYFVDDKQDDQDFDGEIS
FISKDFIQEHVPGPKESTHLFVCGPPPFMNAYSGEKSPKDQGELIGILNNLGYSKDQVF
KF

When purified proteins are digested, the connectivity between peptides and proteins is usually apparent

36

Challenge for Shotgun Proteomics: Alternative Splice Forms

Gene: EPLIN location: 12q13 Q9UHB6-1 (isoform Beta)

>IPI:IPI00220465 Splice isoform Alpha of Q9UHB6 Epithelial protein lost in neoplasm

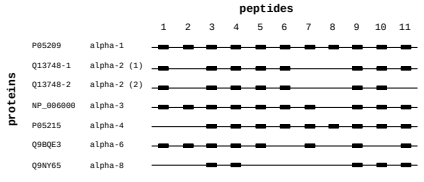
MESSPFNRRQWTSLSLRVTAKELSLVNKNKSSAIVEIFSKYQKAAEETNMEKKRSNTENLSQHFRKGTLTVLKKKWENP
GLGAESHTDSLRNSSTEIRHRADHPPAEVTSHAASGAKADQEEQIHPRSRLRSPPEALVQGRYPHIKDGEDLKDNSTES
KKMENCLGESRHEVEKSEISENTDASGKIEKYNVPLNRLKMMFEKGEPTQTKILRAQSRSASGRKISEAKSVSLDDLETG
PGQLSSSTFDSEKKESRRRLEIPRESETSIKDRWAKYQAAVSKQSSSTNYTNELKASSGETKTHKMEQKENVPPGPEVC
ITHQEGEKISANENSLAVRSTPAEDDSRDSQVKSEVQQPVHPKPLSPDSRASSLSESSPPKAMKKFQAPARETCVECQK
TVYPMERLLANQQVFHISCFRCSYCNNKLSLGTYASLHGRIYCKPHFNQLFKSKGNYDEGFGHRPHKDLWASKNENEEI
LERPAQLANARETPHSPGVEDAPIAKVGVLAASMEAKASSQQEKEDKPAETKKLRIAWPPPTELGSSGSALEEGIKMSK
PKWPPEDEISKPEVPEDVDLDLKKLRRSSSLKERSRPFTVAASFQSTSVKSPKTVSPPIRKGWSMSEQSEESVGGRVAE
RKQVENAKASKKNGNVGKTTWQNKESKGETGKRSKEGHSLEMENENLVENGADSDEDDNSFLKQQSPQEPKSLNWSSFVD
NTFAEEFTTQNQKSQDVELWEGEVVKELSVEEQIKRNRYYDEDEDEE

37

Challenge for Shotgun Proteomics: Gene Families (Paralogues)

C Tubulin-alpha protein family

1	TIGGGDDSFNTFFSETGAGK	5	IHFPLATYAPVISAEK	9	VGINYQPPTVVPGGDLAK
2	AVFVDLEPTVIDEVR	6	AYHEQLSVAEITNACFEPANQMVK	10	AVCMLSNTTAIAEAWAR
3	QLFHPEQLITGKEDAANNYAR	7	YMACCLLYR	11	LDHKFDLMYAK
4	NLDIERPTYTNLNR	8	SIQFVDWCPTGFK		



>IPI00387144 IPI:IPI00387144.2|SWISS-PROT:P05209 Tubulin alpha-1 chain

MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQ
LITGKEDAANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVV
EPYNSILTTHTTLEHSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLISQIVSSITASLRFDGALNVDLTEFQTNLVPYPRIHFPLATYA
PVISAEKAYHEQLSVAEITNACFEPANQMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRSIQFVDWCPTGFKVGINYQPPTVVP
GGDLAKVQRAVCMLSNTTAIAEAWARLDHKFDLMYAKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGVDSVEGEGEEEGEEY

38

Complication: Database Redundancies

Entrez Protein **Stress-70 protein, mitochondrial precursor**

Sequence entry 1:

gi|21264428|sp|P38646|GR75_HUMAN Stress-70 protein, mitochondrial precursor (75 kDa glucose regulated protein) (GRP 75) (Peptide-binding protein 74) (PBP74) (Mortalin) (MOT)

gi|24234688|ref|NP_004125.3| heat shock 70kDa protein 9B precursor; heat shock 70kD protein 9; **gi|477763|pir||B48127** dnaK-type molecular chaperone precursor, mitochondrial - human

Sequence entry 2: *G → R at position 540*

gi|292059|gb|AAA67526.1| MTHSP75
Human mRNA L15189

Sequence entry 3: *Q → R at position 74; H → R at position 184*

gi|12653415|gb|AAH00478.1| Heat shock 70kDa protein 9B, precursor [Homo sapiens].
Human mRNA BC000478
Human mRNA BC024034

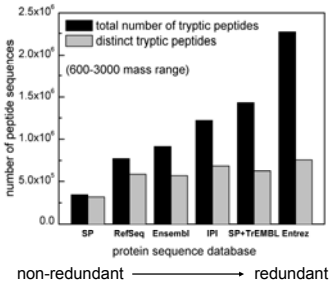
gi|18645123|gb|AAH24034.1| Heat shock 70kDa protein 9B, precursor [Homo sapiens].

Sequence entry 4: *M → V at N terminus*

gi|21040386|gb|AAH30634.1| heat shock 70kD protein 9B (mortalin-2) [Homo sapiens].
Human mRNA BC030634

redundant database entries

Protein Sequence Databases

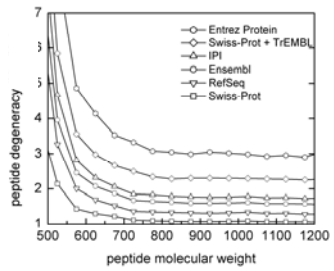


non-redundant ⟶ redundant

40

Protein Sequence Databases

Average number of database entries containing each peptide



choice of the database depends on the goals of the experiment

e.g. is ability to identify polymorphisms Important?

41

Mature Protein Forms

A

>IPI00019906.1|UniProt/Swiss-Prot:P35613 Basigin precursor

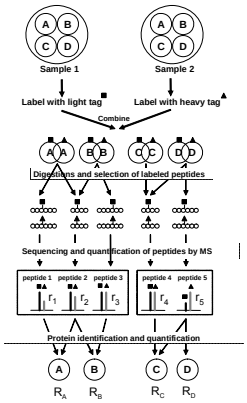
signal peptide **N-terminal peptide**

MAAALFVLLGFALLGTHGASGAAGTVFTTVEDLGSKILLTCSLNDSATEVTGHRWLKGGVVLKEDALPGQKTEFKVDSDDQWGEYSCVFL
PEPMGTANIQLHGPPRVKAVKSSEHINEGETAMLVCKSESVPPVTDWAWYKITDSEDKALMNGSESRFFVSSSQGRSELHIENLNMEADP
GQYRCNGTSSKGSDQAIITLRVRSHLAALWPFLGIVAEVLVLVTIIFIYEKRRKPEDVLDDDDAGSAPLKSSGQHQNDKGKNVRQRNSS

signal peptide removed

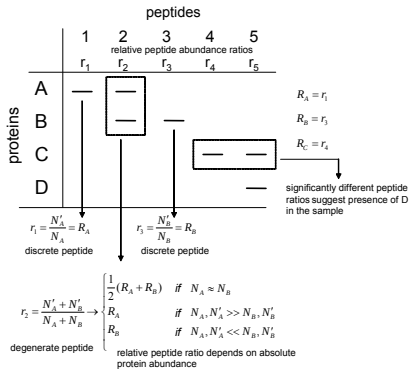
42

Quantitative Proteomics



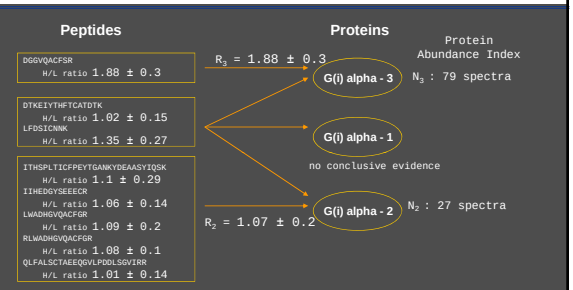
43

Quantitative Proteomics



44

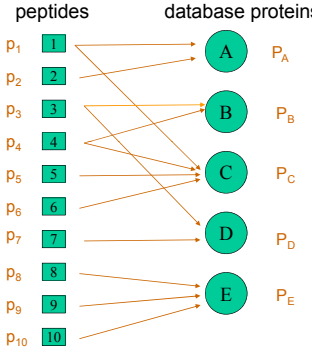
Quantitative Proteomics



Quantification and identification are interdependent problems

45

Protein Inference with ProteinProphet



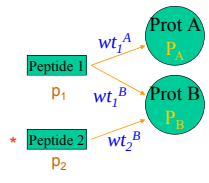
PeptideProphet computational tool:

Probability-based model aims at deriving minimum list of proteins that can explain all observed peptides

Resolves ambiguous cases (when possible) and presents results in a convenient "biologist-friendly"format

46

Apportionment of Shared Peptides



EM

$$wt_i^n = \frac{P_n}{\sum_{j=1\ldots N} P_j}$$

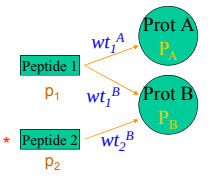
$$P_n = 1 - \prod_i (1 - wt_i^n \, p_i)$$

For each peptide, weights sum to 1

Model aims at deriving the simplest list of proteins sufficient to explain the observed peptides (Occam's razor)

47

Apportionment of Shared Peptides



Initialize:
$P_A = P_B$
$wt_1^A = wt_1^B = 0.5$
$wt_2^B = 1$

Run EM:
$wt_1^A = 0$
$P_A = 0$

$wt_1^A = 1$
$P_B = 1 - (1-0.8)(1-0.7) = 0.94$

Protein B is present in the sample

48

Protein Group: Example 1

peptides
1 2 3 4

proteins
A - -
B - -

distinct proteins

ProteinProphet output

#1 Prot A P=1
*wt=1 peptide 1
*wt=1 peptide 2

#2 Prot B P=1
*wt=1 peptide 3
*wt=1 peptide 4

*: no other sequence database entry has this peptide (wt=1)

49

Protein Group: Example 2

peptides
1 2 3

proteins
A - - -
B - - -

indistinguishable proteins

ProteinProphet output

#1 Prot A Prot B P=1
*wt=1 peptide 1
*wt=1 peptide 2
*wt=1 peptide 3

*: no other sequence database entry has this peptide (wt=1)

50

Protein Group: Example 3

peptides
1 2 3

protein
A - - -
B - - -

B is a subset protein

Occam's razor: Prot A is present in the sample. No conclusive evidence for the presence of Prot B.

ProteinProphet output

#1 a) Prot A P=1
*wt=1 peptide1
wt=1 peptide 2
wt=1 peptide 3

b) Prot B P=0
wt=0 peptide 2
wt=0 peptide 3

New in version 4.2:
subset protein is shown as part of the group

51

Protein Group: Example 4

peptides
1 2 3 4

proteins
A - - -
B - - -

differentiable proteins

Default output (MININDEP=0)

ProteinProphet output

#1 Prot A P=1
*wt=1 peptide 1
wt=0.5 peptide 2
wt=0.5 peptide 3

#2 Prot B P=1
wt=0.5 peptide 2
wt=0.5 peptide 3
*wt=1 peptide 4

52

Protein Group: Example 4 (optional)

peptides
1 2 3 4

proteins
A - - -
B - - -

differentiable proteins

ProteinProphet output

#1 a) Prot A P=1
*wt=1 peptide 1
wt=0.5 peptide 2
wt=0.5 peptide 3

b) Prot B P=1
wt=0.5 peptide 2
wt=0.5 peptide 3
*wt=1 peptide 4

command line option parameter:
If MININDEP is set to >0, then proteins identified by only a few distinct peptides (fraction of total peptides less than MININDEP value) are grouped

53

Protein Group: Example 5

peptides
1 2 3 4

proteins
A - -
B - -
C - -

B: subsumable protein

Even if MININDEP is set to 0 (default), proteins identified by distinct peptides but linked through one or more subsumable proteins are grouped

ProteinProphet output

#1 a) Prot A P =1
*wt=1 peptide 1
wt=1 peptide 2

b) Prot C P=1
wt=1 peptide 3
*wt=1 peptide 4

c) Prot B P=0
wt=0 peptide 2
wt=0 peptide 3

54

Protein Group: Example 6

peptides
1 2 3

proteins
A - - -
B - -
C - -

special case:

A has all three peptides (most likely protein). However, B and C combined can account for all the observed peptides as well

ProteinProphet output

#1 Protein group P=1
a Prot A *P=1*
wt=1 peptide 1
wt=1 peptide 2
wt=1 peptide 3

b Prot B *P=0*
wt=0 peptide 1
wt=0 peptide 3

c Prot C *P=0*
wt=0 peptide 2
wt=0 peptide 3

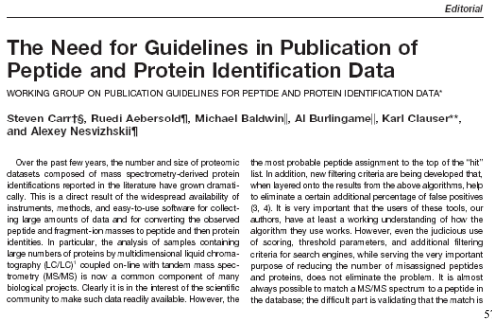
55

ProteinProphet (full analysis)

- Retrieves ***all*** proteins corresponding to each assigned peptide from search database
- Groups peptides by protein and computes a probability that a protein is present based upon the evidence of corresponding peptides in the dataset
- Adjusts peptide probabilities for grouping information (penalizes 'single hits')
- Apportions shared peptides, those corresponding to more than a single protein in the database, among ***all*** corresponding proteins
- Collapses redundant and indistinguishable protein database entries into one identification

56

Publishing Large-Scale Datasets



Editorial

The Need for Guidelines in Publication of Peptide and Protein Identification Data

WORKING GROUP ON PUBLICATION GUIDELINES FOR PEPTIDE AND PROTEIN IDENTIFICATION DATA*

Steven Carr¶§, Ruedi Aebersold¶‖, Michael Baldwin‖, Al Burlingame‖, Karl Clauser**, and Alexey Nesvizhskii¶

Over the past few years, the number and size of proteomic datasets composed of mass spectrometry-derived protein identifications reported in the literature have grown dramatically. This is a direct result of the widespread availability of instruments, methods, and easy-to-use software for collecting large amounts of data and for converting the observed peptide and fragment-ion masses to peptide and then protein identities. In particular, the analysis of samples containing large numbers of proteins by multidimensional liquid chromatography (LC/LC)[1] coupled on-line with tandem mass spectrometry (MS/MS) is now a common component of many biological projects. Clearly it is in the interest of the scientific community to make such data readily available. However, the the most probable peptide assignment to the top of the "hit" list. In addition, new filtering criteria are being developed that, when layered onto the results from the above algorithms, help to eliminate a certain additional percentage of false positives (3, 4). It is very important that the users of these tools, our authors, have at least a working understanding of how the algorithm they use works. However, even the judicious use of scoring, threshold parameters, and additional filtering criteria for search engines, while serving the very important purpose of reducing the number of misassigned peptides and proteins, does not eliminate the problem. It is almost always possible to match a MS/MS spectrum to a peptide in the database; the difficult part is validating that the match is

57

Publishing Large-Scale Datasets

Technology

The Application of New Software Tools to Quantitative Protein Profiling Via Isotope-coded Affinity Tag (ICAT) and Tandem Mass Spectrometry

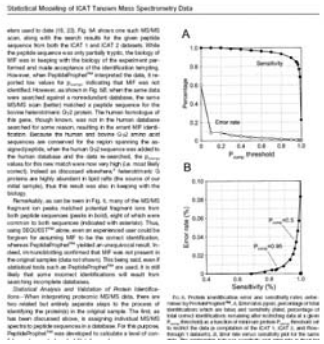
II. EVALUATION OF TANDEM MASS SPECTROMETRY METHODOLOGIES FOR LARGE-SCALE PROTEIN ANALYSIS, AND THE APPLICATION OF STATISTICAL TOOLS FOR DATA ANALYSIS AND INTERPRETATION*

Priska D. von Haller‡, Eugene Yi, Samuel Donohoe, Kelly Vaughn, Andrew Keller, Alexey I. Nesvizhskii, Jimmy Eng, Xiao-jun Li, David R. Goodlett, Ruedi Aebersold, and Julian D. Watts§

Molecular & Cellular Proteomics **3**, 532-534 (2004)

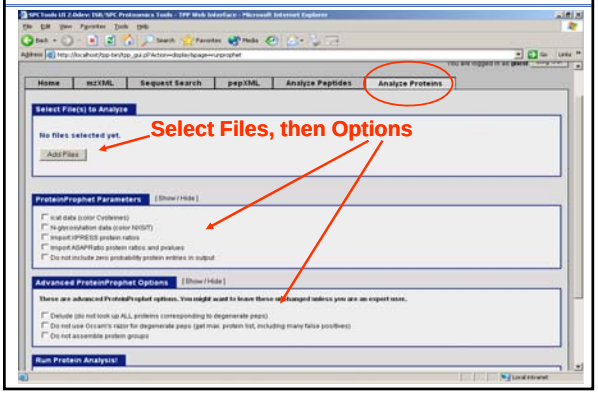
58

Publishing Large-Scale Datasets

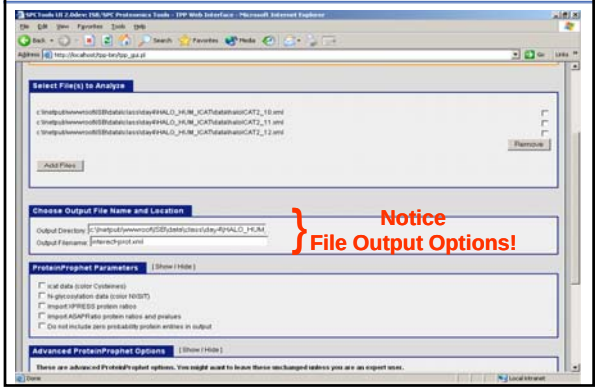


59

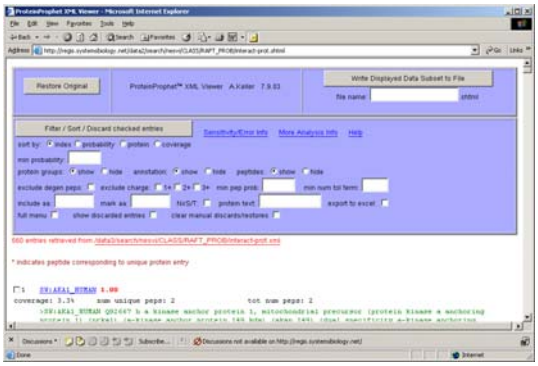
Running ProteinProphet



Running ProteinProphet (after file select)



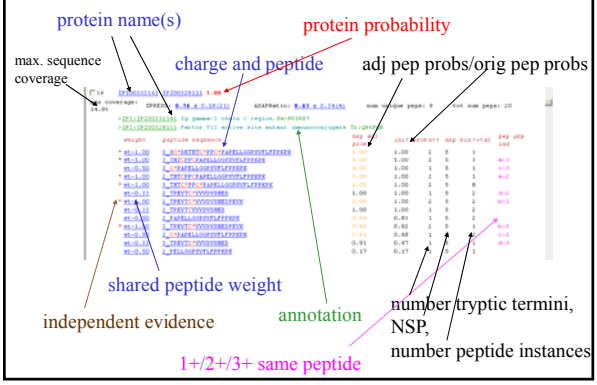
ProteinProphet Output



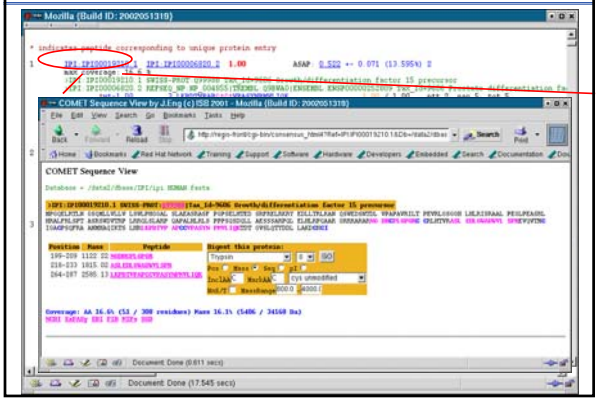
ProteinProphet Output

63

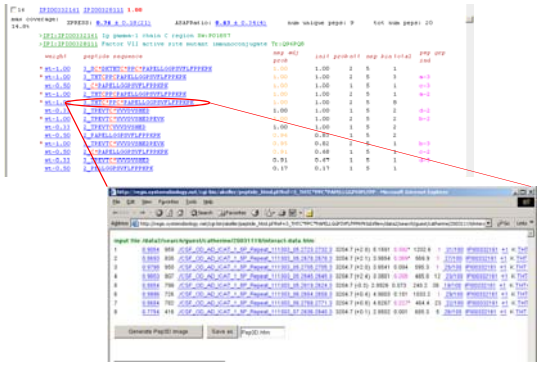
ProteinProphet Output



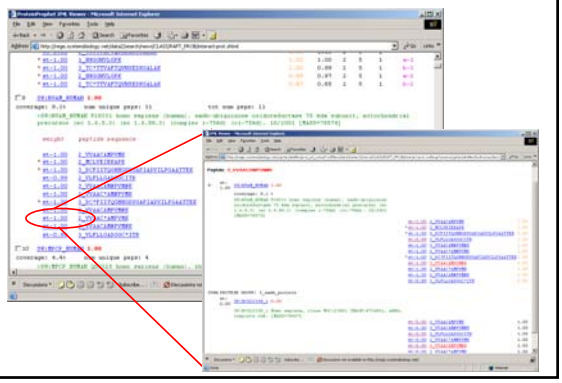
Protein Entry Link



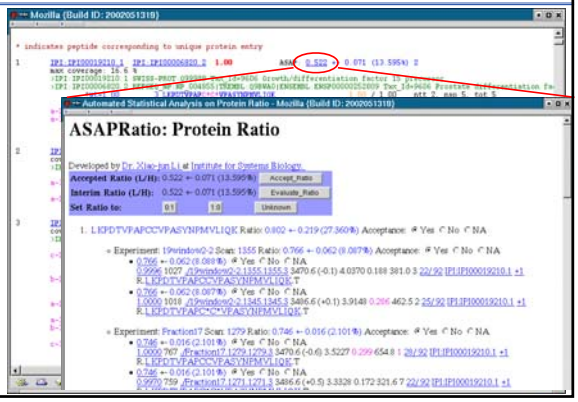
Peptide Link



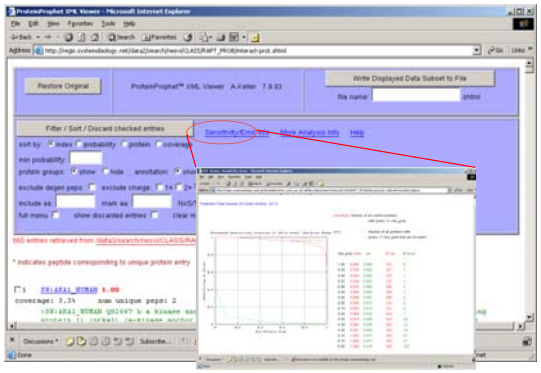
Peptide Weight Link



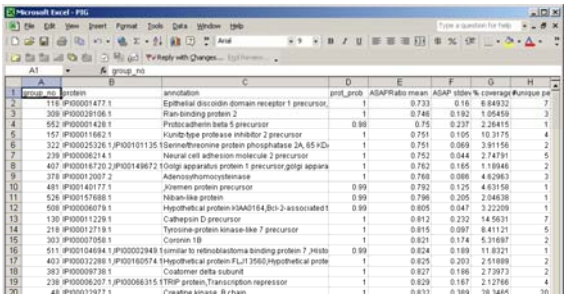
ASAPRatio Link (ICAT experiment)



Sensitivity/Error Rate Plot



Tab Delimited File Output



ProteinProphet – Tutorial and exercises

1. Yeast Orbitrap data

We will use the same Yeast LTQ-Orbitrap dataset to evaluate the performance of ProteinProphet.

All MS/MS spectra were searched using X-Tandem (with k-score plug-in) against a Yeast database appended with an equal number of decoy sequences and common contaminants. In this dataset all decoy sequences have names that start with REV0_ or REV1_. All decoy proteins are incorrect identifications.

The search results were analyzed using PeptideProphet. You do not need to run ProteinProphet, we have done it for you already.

Using the File Browser in Petunia, navigate down to the class/ProteinProphet/xtandem-k/semitryptic/ folder, and open the interact.prot.shtml file by clicking on the View link. This file is the main ProteinProphet output file.

Do the following:

i) It is always good to check first that PeptideProphet worked fine and that computed peptide probabilities are likely to be accurate. If there was a problem running PeptideProphet and peptide probabilities are not accurate, then ProteinProphet results are not going to be accurate as well. To do this, first follow **any** peptide link. A new window should open showing more information about that particular peptide identification. Click on the peptide probability link, and it should bring you to the PeptideProphet output page. Look at the discriminant score distributions learned by the model. Do they look Ok? Look at the model output below and check that other parameters learned by the model are reasonable (e.g., the distribution of NTT parameter among correct and incorrect identifications).

ii) Familiarize yourself with ProteinProphet output.
Find entry #**344a**, YEL034W.
- What is the probability assigned to this protein?
- How many different peptide sequences are identified that correspond to this protein?
- What is the number of "unique peptides", and how is it defined in ProteinProphet?
- Are there any peptides identified multiple times?
- Are there any *'shared peptides'*, i.e. peptides present not only in YEL034W but also in some other protein(s)?

Look at the peptide 2_NGFVVIK.
- How many siblings does it have?
- Compute the NSP value for that peptide by summing the probabilities of its sibling peptides. Does this computed NSP value agree with the NSP number shown for that peptide?
- Does adjustment for peptide grouping information (NSP) increase or decrease the probability that this particular peptide identification is correct?

iii) *Extra questions for those who like math and statistics*: Find entry #**474**. This one is a single-hit protein identification (and an incorrect one). The initial peptide probability computed by PeptideProphet was somewhat high, **0.9761**. However, ProteinProphet penalized this peptide identification (and, therefore, reduced the probability of the protein identification) for being a single hit. As a result, the adjusted peptide probability is only **0.5006**. Please repeat the calculations yourself. Follow the NSP link and find the values of p(NSP|+) and p(NSP|-) for the corresponding bin (NSP value 0, NSP bin 0). Plug the numbers in the expression (5) of the Anal. Chem. (2003) paper and see how the initial probability **0.98** gets reduced to **~0.50**.

iv) Check the accuracy of protein probabilities computed by ProteinProphet. This is a small dataset, and we do not really know for sure what identifications are correct. So this is just a simple estimate.

1. Consider all protein identifications in the probability range between 0.65 and 0.75. Count the total number of proteins in that range (N), and the number of decoys among them (N_d, names start with REV0_ or REV1_).
2. Estimate the number of correct proteins in the 0.65-0.75 range by assuming that the number of incorrect Yeast protein identifications in that range is equal to the number of decoy protein identifications, $N_c = N - N_d - N_d = N - 2*N_d$
3. Calculate the ratio of the estimated number of correct identifications to the total number of identifications N_c/N in this probability range. Is this ratio close to the expected value of ~ 0.7?

v) As discussed in the lecture, adjustment of peptide probabilities to account for peptide grouping information (NSP) makes peptide probabilities (and, therefore, protein probabilities) more accurate. Check if this is the case. Open the file *interact-nonsp.prot.shtml*, it is in the same directory as interact.prot.shtml. This file has protein probabilities computed without adjustment for NSP. Again, look at the same probability range (0.65-0.75). Are computed probabilities accurate? (no need to count proteins, the answer should be obvious).

vi) What are the ProteinProphet predicted sensitivity and false discovery rate (FDR) for this dataset when filtered using minimum protein probability threshold of 0.7? To find that, follow the Sensitivity/Error Info link at the top of the file (note FDR=err). Compare the predicted FDR with that estimated based on decoy counts, $FDR = 2N_d / N$, where N is the total number of proteins with probability above 0.7, and N_d is the number of decoys among them.

vii) Think about different sources of false positives. What are we NOT taking into account when performing target-decoy based FDR estimates, or when using ProteinProphet computed probabilities? How does it affect the error rate estimates?

Consider entry #**387**, protein YNL014W.
>HEF3 Translational elongation factor EF-3; paralog of YEF3 and member of the ABC superfamily.
This protein is identified with a very high probability, **0.9991**. However, this identification is likely to be a false positive. Investigate this case. On what peptide is this identification based? Find an alternative explanation that makes this identification questionable. Hint: this is a high mass accuracy data (LTQ-Orbitrap), which can help.

viii) What is the advantage of generating data on high mass accuracy MS instruments with respect to the source of false positive protein identifications discussed in the previous question? How could you modify the database search parameters to lessen this problem (although not eliminate completely).

2. Human Raft Dataset searches against the Human IPI database

This is a subset of a much larger dataset from a human raft protein profiling experiment. This dataset demonstrates the complexity often encountered in proteomics experiments on higher eukaryote organisms. The purpose of this exercise is to get familiar with the difficulty of inferring what proteins are present in the sample given the list of identified peptides. To view the results, use Petunia to navigate down to the **C:\inetpub\wwwroot\ISB\data\class\ProteinProphet\RAFT\IPI** directory, and open the file **interact2.prot.shtml** using View link (or run yourself, data are in **C:\inetpub\wwwroot\ISB\data\class\ProteinProphet\RAFT\IPI\data**).

Go through the following examples:

i) Indistinguishable proteins.
Find entry #**77**, IPI00026185 IPI00218782
This is a typical example of multiple proteins that cannot be distinguished on the basis of identified peptides. In this case, the two proteins are different isoforms of the F-actin capping protein beta subunit, SW:P47756-1 and SW:P47756-2. You can follow the Ensembl links (and from the Ensembl page Description sections, Swissprot links) to learn more about these proteins. What can you conclude about the presence of the isoforms in the sample?

For the most curious: spend some time playing with this example. Check the sequences of these two proteins. Are they significantly different? To do that, go to Swiss-Prot (the easiest way since this alternative splicing even is annotated in Swiss-Prot), or cut and paste protein sequences and align them using a sequence alignment program (e.g., utility bl2seq that can be found at http://www.ncbi.nlm.nih.gov/blast). In what situation would it be possible to determine which of the two proteins (or both) is actually present in the sample? (hint: are there tryptic peptides in these proteins that, if identified, would discriminate between the two isoforms?)

ii) Subset proteins
Find entry #**178a**, IPI00027500
This protein (Rho A, SW:P06749) is a member of a family of Rho proteins. Two of its peptides, IGAFGYMECSAK and modified form, are unique to this protein (marked with an asterisk). The other peptides are shared, i.e., they are also present in another protein from the same protein family. What is the name of the other protein that also contains these peptides? What probability did ProteinProphet assign to it? What can be concluded about the presence of that other protein in the sample?

iii) Differentiable proteins
Find entry #**167**
This is another interesting example. There are several members from the same protein family that are grouped together. For example, consider entry #**167e,** IPI00023138. This protein (Ras-related C3 botulinum toxin substrate 3) is identified by one unique peptide, HHCPHTPILLVGTK, and several

shared peptides. Some of the other peptides are shared between this protein and a different isoform (e.g. entry #**167b**: IPI00010270, Ras-related ... substrate 2). However, the other isoform is identified by several peptides that are unique to it, including HHCPSTPIILVGTK (note a two amino acid difference compared to the peptide that is unique to the first isoform). Thus, even though these proteins share a set of peptides, each of them has at least one unique peptide. What does ProteinProphet conclude about the presence of these proteins in the sample?

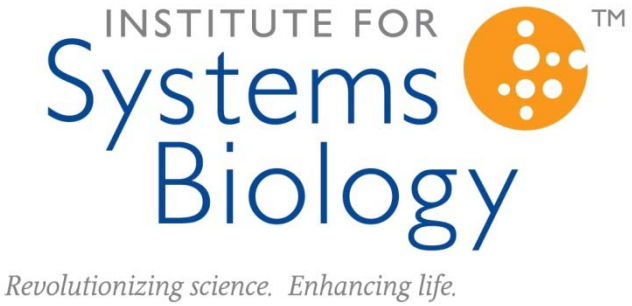
iv) Special case: a protein group containing proteins with no distinct peptides
Find entry # **165**, Protein Group **15**
This is an example of a special case where, strictly speaking, the parsimony rule (Occam's razor) cannot be applied. Four protein entries comprise this group with an assigned probability of 1. Is there definitive evidence that any particular group member is present in the sample? Which protein in this group can explain the presence of all peptides observed in the dataset that correspond to proteins from this group ('subsumes' all the others), and is therefore the most likely candidate? Can we be certain what protein(s) are present in the sample?

SpectraST:
A Spectral Library Building and Searching Tool for Proteomics

Eric Deutsch

Day 3

September 27, 2010

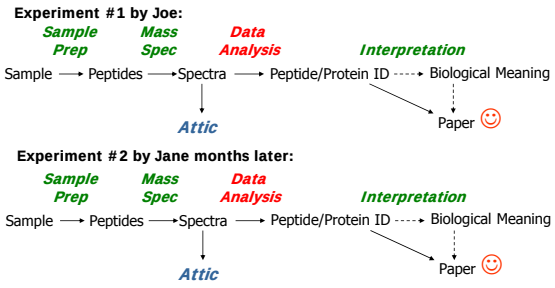


Outline

- Motivation
- Spectral library searching
- Spectral library building
- Challenges

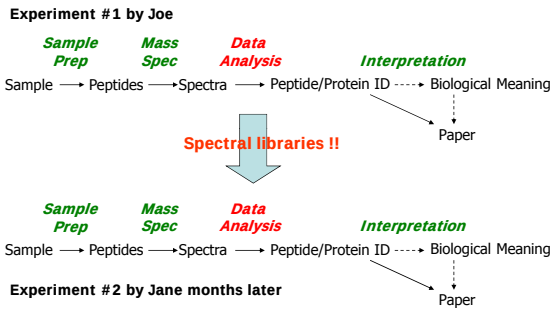
1

Memoryless Workflow



2

Learning from the Past



3

Spectral Libraries

- Simply, collections of identified spectra
  - Rely on sequence searching to make initial identifications
- Enables future identifications by *spectral matching* (similar spectra => same ID)
  - Pros: Extremely fast, very high sensitivity and specificity
  - Cons: Can only find things previously seen
- Other uses
  - As a living, retrievable record of the observed proteome
  - Data mining
  - Planning SRM experiments

4

Spectral Libraries

- Incomplete library coverage, but…
  - Limiting factor is MS technology
  - More interesting biological questions are answered with targeted approaches
  - Can be integrated with sequence searching to improve pipeline

5

Spectral searching

- NIST MS (NIST, Stein et al, JASMS 1994)
- LIBQUEST (Yates, Yates et al, Anal Chem. 1998)
- X!Hunter (GPM, Craig et al, JPR 2006)
- Bibliospec (MacCoss, Frewen et al, Anal Chem 2006)
- SpectraST (Aebersold, Lam et al, Proteomics 2007)

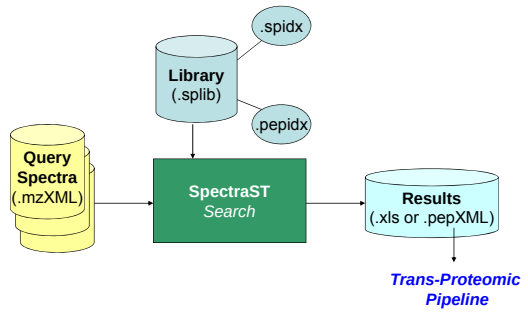
6

SpectraST

- Open Source (http://sourceforge.net/projects/sashimi under trans_proteomic_pipeline)
- LINUX / Windows (MSVC, MinGW) versions (http://tools.proteomecenter.org/TPP.php)
- Extensible, modular design
- Fully integrated with Trans-Proteomic Pipeline
- Modest processor and memory requirements

7

SpectraST *Search* Mode



8

SpectraST Search Algorithm

- Query spectrum processing
  - Basic spectrum filtering
  - Remove region around parent peak
  - Scale intensities (to deemphasize dominant peaks)
    - Scaled intensity = (intensity)$^{0.5}$
  - Assign peaks into unit-m/z bins
  - No deisotoping, no neutral loss removal

9

SpectraST Search Algorithm

- Similarity scoring
  - Dot product

    (j = bin number, I = normalized scaled intensity)

    $$Dot = \sum_{j=1}^{n} I_{query}(j) I_{library}(j)$$
  - Delta Dot

    $$Delta\ Dot = \frac{Dot(1) - Dot(2)}{Dot(1)}$$
  - Dot Bias

    1.0 = one bin accounts for entire dot product

    ~0.0 = all bins contribute equally

    $$Dot\ Bias = \frac{1}{Dot}\sqrt{\sum_{j=1}^{n} I_{query}^2(j) I_{library}^2(j)}$$
- Discriminant function
  - Fval = 0.6 (Dot) + 0.4 (Delta Dot) – Dot Bias penalty

10

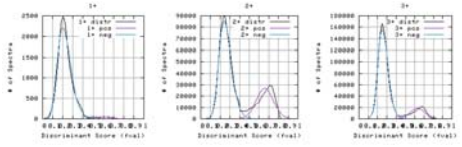
Spectral searching

- Human plasma dataset (Novartis-GeneProt)
  - 7,000 runs, 2.4 million MS/MS spectra
  - SEQUEST search
    - Semitryptic, 3 Da precursor mass window
    - Against human IPI
    - 350,000 p>0.99 IDs (PeptideProphet)
    - Weeks on a 80-CPU cluster
  - SpectraST search
    - 3 Th precursor m/z window
    - Against NIST human spectrum library (2006)
    - 430,000 p>0.99 IDs (PeptideProphet)
    - Overnight on 1 CPU

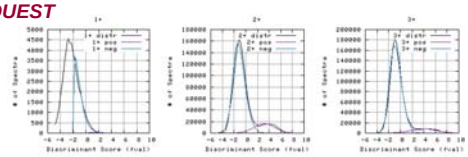
11

Score discrimination

SpectraST



SEQUEST



12

Lessons

- SpectraST identified 22% more spectra at same confidence cutoff
- Extra SpectraST identifications are mostly
  - Same IDs made by SEQUEST but at lower confidence
  - Repeated IDs of same peptide ions (from lower-quality spectra that SEQUEST failed to identify)
- SpectraST is rarely wrong with its confident IDs (~0.01% FDR)
- 99.5% of the missed IDs by SpectraST is not in the spectrum library

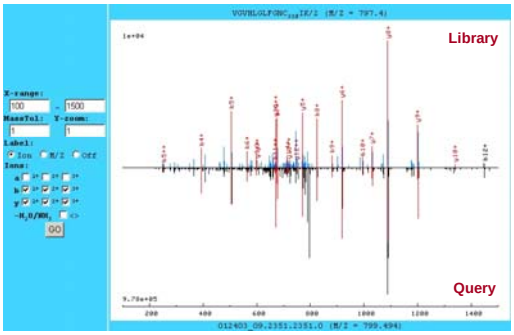
13

Why the improvement?

- Smaller search space
  - Only previously observed peptides are searched
  - Approximately 0.1x for very constrained searches, 0.001x for typical searches
- NIST library contains IDs of several sequence search engines
  - Implicitly combines multiple search algorithms
- More precise similarity scoring
  - Global similarity, not just presence of b- and y-ions

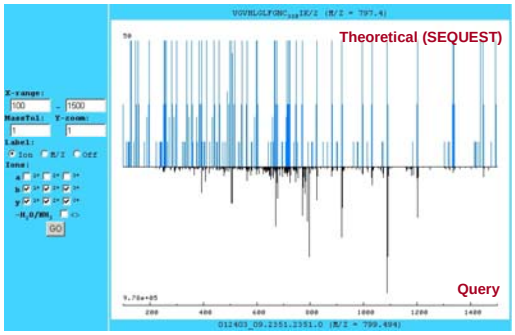
14

Global Similarity



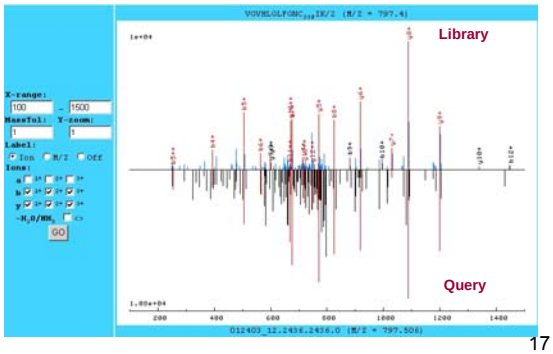
15

Theoretical vs Real Spectrum



16

Matching Noisy Spectrum



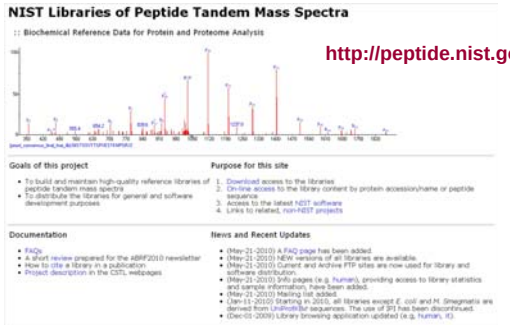
17

Spectrum Library Building

- Centralized (NIST, PeptideAtlas, GPM)
  - Greater coverage
  - Higher quality
- Do-it-yourself (X!Hunter, Bibliospec, SpectraST)
  - Specialized libraries
  - Proprietory data
  - Data organization and reusability

18

NIST Libraries



http://peptide.nist.gov/

NIST Libraries

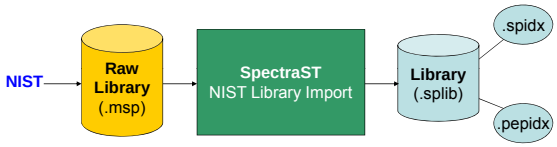
- Searchable with NIST MS / SpectraST (after import)
- Post-translational modifications considered:
  - CAM-cysteine, cleavable and uncleavable ICAT
  - Methionine oxidation
  - N-terminal acetylation
  - N-terminal pyro-glutamate
- Sequence search engines used: SEQUEST, Mascot, X!Tandem, OMSSA
- Stringent quality filters (only multiply observed peptide ions, numerous filters)

NIST Libraries

Library Name	Species	Num. Spectra	Current Release	Previous Releases
Libraries by Biological Species				
Human	*H. sapiens*	345,489	Info | browse | Download	archive
Human (qtof)	*H. sapiens*	14,827	Info | browse | Download	archive
Mouse	*M. musculus*	156,495	Info | browse | Download	archive
Drosophila	*D. melanogaster*	113,877	Info | browse | Download	archive
C. elegans	*C. elegans*	98,151	Info | browse | Download	archive
Yeast	*S. cerevesiae*	90,507	Info | browse | Download	archive
Yeast (qtof)	*S. cerevesiae*	2,418	Info | browse | Download	archive
E. coli	*E. coli*	58,967	Info | browse | Download	archive
D. radiodurans	*D. radiodurans*	11,913	Info | browse | Download	archive
M. smegmatis	*M. smegmatis*	5,217	Info | browse | Download	archive
Rat	*R. novicus*	20,992	Info | browse | Download	archive
Libraries for Protein Standards and Mixtures				
NCI-20	protein standard	3,898	Info | browse | Download	archive
chicken egg	*G. gallus*	4,472	Info | browse | Download	archive
Sigma UPS1	protein standard	3,542	Info | browse | Download	archive
Human serum albumin	*H. sapiens*	2,309	Info | browse | Download	archive
BSA	*B. tarus*	969	Info | browse | Download	archive
Beta-2-micoglobulin	*H. sapiens*	179	Info | browse | Download	archive
C-reactive protein	*H. sapiens*	94	Info | browse | Download	archive

Also available at http://www.peptideatlas.org/speclib/

SpectraST *Create* Mode I



Do-it-yourself Library Building

- Informatics
  - Bring the ID and the spectrum together
  - Merge many datasets or existing libraries of various formats and protocols
  - Record and propagate relevant information
  - Enable easy and fast retrieval
- Algorithm
  - Consensus building
  - Spectrum cleaning
  - Quality control
  - Updates, error correction…

DIY Library Building with SpectraST

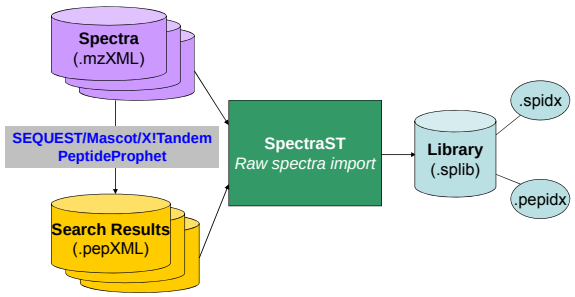
- Uses common open XML formats
  - Supports all major vendors
  - Supports all popular sequence search engines
- Allows import of other library formats
  - NIST, X!Hunter, Bibliospec
- Robust consensus algorithm
- Carries sample information
  - Sequence search scores, statistical and quality measures, sample source, instrument, RT, etc.
- Automatic quality control

DIY Library Building with SpectraST

- Step 1: Gather all identified spectra
  - Read pepXML files for confident IDs
  - Extract spectrum from mzXML files
- Step 2: Combine multiple observations (replicates)
  - Pick the "best" replicate
  - Consensus
- Step 3: Quality control
  - False positives
  - Low-quality spectra
- Step 4: Library manipulation
  - Merging, subtracting, filtering…

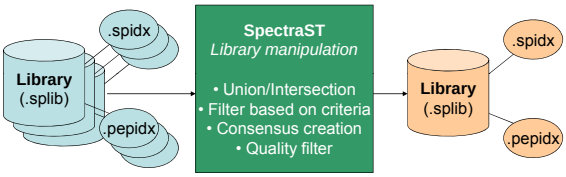
25

SpectraST *Create* Mode 2



26

SpectraST *Create* Mode 3



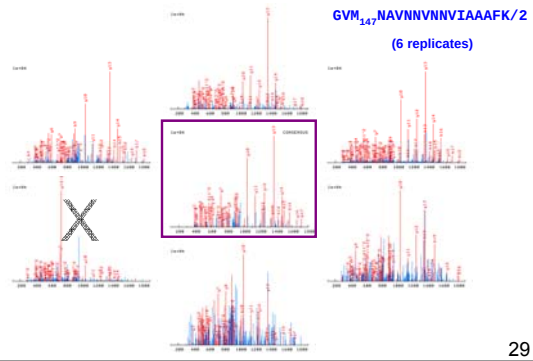
27

Consensus Spectrum Building

- Pool replicates (spectra identified to the same peptide ion)
- Remove dissimilar replicates
- Align slightly m/z-shifted peaks
- Use "peak voting" to decide if peak belongs in consensus
- Weighted-average peak intensities by a measure of replicate signal-to-noise

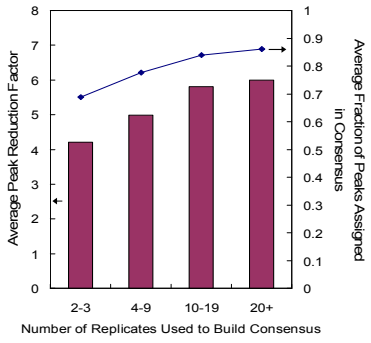
28

Consensus Spectrum



29

Noise Reduction

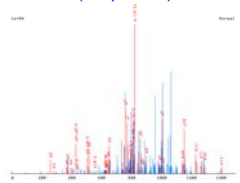


30

Quality control

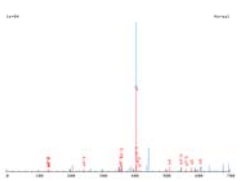
- Problem: There are still occasional false positives and terribly noisy spectra

AVDLLFFTDESGDSR/2
(2 replicates)



Possibly correctly identified, but impure spectrum

DFFTPNLFLK/3
(1 replicate)



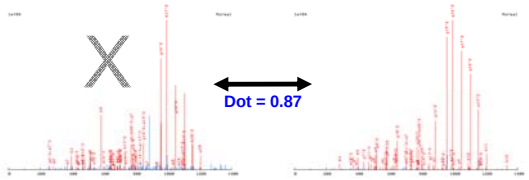
False-positive impure spectrum

31

Quality control

FFTAICDMVAWLGYTPYKVTY/3
(1 replicate)

ALVLIAFAQYLQQC$_{160}$PFEDHVK/3
(100 replicates)

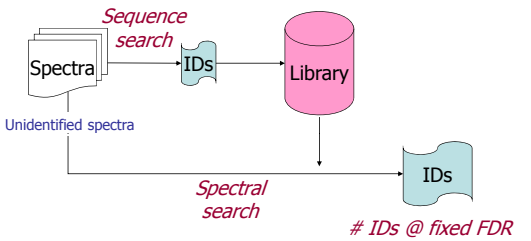


Similar spectra having conflicting IDs

- Potential source of false negatives!

32

Evaluating Library Building Methods



 # IDs @ fixed FDR as measure of libraries' performance

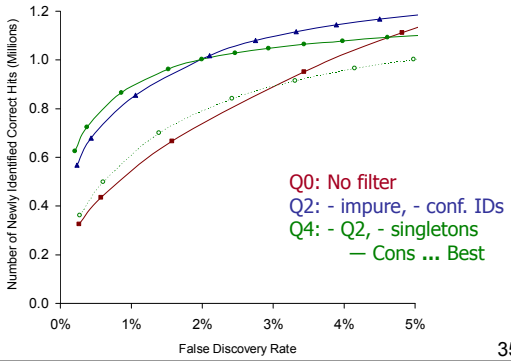
33

Evaluating Library Building Methods

- Use entire Human Plasma PeptideAtlas
  - 40 datasets, 1.3M identified spectra at p>0.9
  - Consensus vs best-replicate
  - Different quality levels
  - The remaining unidentified spectra (15M) used to evaluate libraries by spectral searching
- Speed
  - Library building took about 2 days on 1 CPU

34

Evaluating Library Building Methods



35

Lessons

- Up to 0.9 million (70%) more confident IDs identified at same FDR of 1% compared to initial SEQUEST search
  - 1000x faster!
- Quality of library matters
  - Higher quality means better discrimination, more IDs at same confidence cutoff
  - Eventually trade off with lower coverage
- Consensus is much better than best-replicate
  - Raw replicates are more similar to the consensus than to the best replicate

36

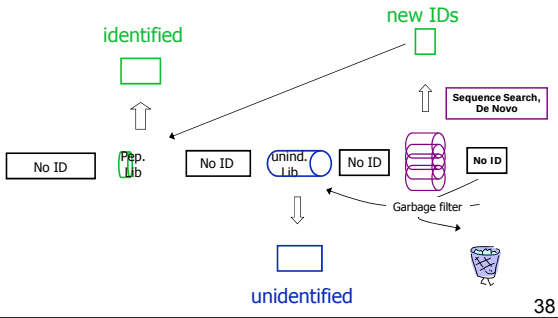
Ongoing developments

- Algorithm optimization
  - Speed
  - Scoring
  - Consensus building
  - Proper quality filters (coverage vs quality)
- Statistical validation
  - Decoy libraries
- Pipeline development
  - Integration with sequence searching

37

Spectral Libraries

- "Peptide ID Pipeline of the future" – Steve Stein, NIST



38

Ongoing development

- Centralized library building
  - How to gather data from community efficiently?
  - How to ensure quality of data?
- Growing libraries
  - More organisms, more instrument types, ETD
  - PTMs (empirical vs semi-empirical)
  - Non-peptides, impurities
- Library maintenance
  - Manual/automatic updates? How often?
  - How to correct mistakes?

39

Updates

40

Exercises – Spectral Library Building and Searching Using SpectraST

These exercises take you through the process of building a spectral library from sequence search results of the yeast SILAC dataset we have been using, searching the same dataset against it, validating the results using PeptideProphet, and comparing them to the original sequence search results.

Step 1. Extracting confident IDs from sequence search results and library building

In this tutorial, we will build a spectral library from the Tandem-K search results of the yeast SILAC dataset we have been using. In the first step, we extract the confident identifications from the PeptideProphet output (i.e., the interact.xml file you created in the PeptideProphet tutorial).

Log onto the Petunia interface, and select the **SpectraST Tools** link. Then click on the tab **SpectraST Library Import**. In the pane **1. Specify File Format**, select **.pepXML (Sequence Search Result)** from the drop-down menu.

Now, click **Add Files** in the pane **2. Specify files to import**, browse to the directory class/SpectraST. A pepXML file containing sequence search results of the SILAC dataset, as well as the mzXML files containing the query spectra, are already copied here. Select the interact.pep.xml file. (Note that when building spectral libraries, the mzXML files must be in the same directory as the pepXML file.)

In the pane **3. General Options**, type in "raw" in the **Enter name of output file** box; this will be the name given to output file. In the **Specify a dataset identifier** box, type in "course"; this will allow you to keep track of the sample source of your consensus spectra later on. In the **Specify a minimum probability to import** box, check that the value of 0.9 is set as the default. This is the minimum probability for an identified spectrum to be extracted by SpectraST.

Uncheck the **Annotate Peaks** option. Leave all other options as their defaults. Go to the bottom of the page and click **Import Library Files**. It will take about 3-6 minutes to run. During this time, SpectraST scans through all the identifications contained in the interact.pep.xml file, selecting only those above the probability threshold, goes back to the respective mzXML files to extract the query spectra, and then creates a raw spectral library of them. While the program is running, monitor the progress in the **Output so far** pane. You may have to click the **UPDATE THE PAGE** link at the bottom to force the page to refresh every now and then.

When SpectraST is done, the **Output so far** pane should contain information about the newly created raw library. How many spectra are extracted into the library? *(10,821)* How many distinct peptide ions do they represent? *(4,559)* How many spectra are identified to fully tryptic peptides? *(10,611)*

The **Command Status** pane should have turned orange by now, indicating the end of the execution of the command. Open the **Utilities** menu, click on **Browse Files**, and browse the directory **c:\Inetpub\wwwroot\ISB\data\class\SpectraST**. You will see that 5 files are created: raw.splib (a binary-format library used by SpectraST to search), raw.sptxt (a text-format equivalent of raw.splib for human viewing), raw.spidx (a library index on the precursor m/z value), raw.pepidx (a library index on the peptide ion), and spectrast.log (a log file). Click the **View** links to see how they look like. In particular, notice some useful information presented at the beginnings of the files raw.sptxt, raw.pepidx and spectrast.log. You probably want to use the "Stop" button of your browser to stop loading the enormous raw.sptxt in its entirety.

Step 2. Building a Consensus Library

The second step is to build a consensus library from the raw library you just created. Note that in the raw library, some peptide ions are represented by more than one raw spectrum; in consensus building, these "replicate" spectra will be combined into one.

Again click on the tab **SpectraST Library Import** under **SpectraST Tools**. From the drop-down menu under **1. Specify File Format**, select **.splib (perform join/build actions on SpectraST libraries)**. In the pane **2. Specify files to build/join**, remove the interact.pep.xml file, then browse to the directory class/SpectraST to add the raw.splib file that you just built in Step 1. In the **3. Select Actions** pane, select **Consensus** in the **Select Build Action** drop-down menu (leave "join action" as default). Type in "consensus" in the **Enter name of output file** box. Uncheck the **Annotate Peaks** box. Leave all other options as defaults. Scroll down to the bottom and hit the **Import Library Files** button.

It should take less than a minute to run. SpectraST will create a "consensus" spectrum for each peptide ion that has multiple replicates in the raw library raw.splib. For peptide ions with only a single replicate, SpectraST will also include them in the final library after some spectrum processing.

When it is done running, check the **Output so far** pane to see the statistics. How many spectra are there in this library? *(4,559)* Note that this is the same as the number of unique peptide ions in the raw library, as the replicates are now combined. How many spectra are from single observations (see the NREPS line)? *(1,848)*

Step 3. Evaluate the quality of the spectral library and applying quality filters

Usually, with a high probability cutoff of 0.9, most of the spectra contained in the consensus library are of decent quality, but occasionally there will be a few that are mis-identified by the sequence search engine in the first place, and/or highly impure. In this step, SpectraST will attempt to identify these spectra and (if you so choose) remove them from the library.

Go to the **SpectraST Library Import** page once again, and again select **.splib (perform join/build actions on SpectraST libraries)** as the file format. In the pane **2. Specify files to build/join**, remove the raw.splib file, and add the consensus.splib file. Select **Quality_Filter** in the **Select Build Action** drop-down menu. Uncheck the **Annotate Peaks** option. Scroll down and you should notice a few more quality filter options. In this case we will use the defaults, so just go ahead and click **Import Library Files**.

SpectraST will subject the spectra to each of its quality filters, and determine if they fail any of them. In this case, we are asking SpectraST to keep all spectra regardless of quality, but indicate in the output library which of these spectra have failed which filters, if any.

When it is done, browse (Utilities → Browse Files) to the same directory **c:\Inetpub\wwwroot\ISB\data\class\SpectraST**, and click to **View** the file spectrast.log. If you scroll down, you can see a log of what SpectraST has done in this session so far, including Step 1 through 3 in this tutorial. Towards the end there is information on the quality filters (prefixed with "QUALITY_FILTER") and at the very end there are some statistics. It tells you how many spectra would be left if you had selected various levels of quality.

How many spectra would be left if we set the quality level at 2? *(4,303)*. How many would be left if we set the quality level at 3? *(3,832)* The level designations are as follows:

Level 1: Remove impure spectra
Level 2: Level 1 + spectra that have a spectrally similar counterpart in the library with a conflicting identification
Level 3: Level 2 + spectra whose peptide sequence has no shared sub-sequence with any other peptides in the library
Level 4: Level 3 + singleton spectra
Level 5: Level 4 + inquorate spectra (with a user-defined quorum)

A utility called Lib2HTML can convert a SpectraST library to a web page for visualization. In the Petunia interface, you can run this program by going to the **SpectraST Tools** menu and clicking the tab **Lib2HTML**. Remove any other .splib file, add the newly created file consensus_quality.splib, and hit **Convert Library Files**. When it is done, un-hide the **Command Status** pane, and click to view the html file. Here you can click on the links in the **LibID** column (far right) to see the spectra themselves. The **Status** column specifies the least stringent among the failed quality filters for that entry. ("Normal" implies the spectrum passes through all quality filters.)

Click on some "Singleton" spectra and some "Normal" spectra. Which type of spectra appears to be of high-quality (in terms of criteria such as long consecutive ion series, fewer noise peaks, fewer unassigned peaks)? *(Normal)*

Recall that in the original Tandem-K search, there are some identifications mapped to decoy sequences, which we can be sure are false positives. We have not explicitly removed them during our consensus building process. They are indicated by a "REV" prefix in the protein name. Use the Find function of your browser to count the number of *spectra* identified

to "REV" proteins. (Note that some spectra map to multiple REV proteins.) How many spectra do you count? *(23)* How many library entries do you estimate to be false positives? *(23 × 2 = 52)* What is the estimated identification error rate of your library? *(52 / 4,559 = 0.01)*

A 1% error rate is certainly quite good, but we can do better. Of course, if we desire, we can first remove spectra identified to "REV" proteins, which we are certain to be false, and cut the error rate roughly in half. However, for this tutorial, let's keep these spectra around for educational purpose.

Another way to reduce the false positives is by using SpectraST's quality filters. Now, revisit the entries with "REV" proteins using the Find function of your browser. This time, determine how many of the 12 entries have a **Status** of "Impure" or "Conflicting_ID." *(5 and 6, respectively)* This implies that if we remove these questionable spectra, we can cut the number of falsely identified spectra in our library by about half. Click on a few of the "Impure" entries (the link in the LibID column) to get a feeling of what SpectraST considers an impure spectrum. Also click on a few of the "Conflicting_ID" entries and see if their identifications look questionable to you.

As you are counting, you may notice that the rest of the entries with "REV" proteins are marked "Inquorate_Unconfirmed" and none of them "Normal." In other words, if we are more conservative and choose to filter at quality level 3, we would have removed all of the known false positives, and by our assumption, all false positives as well. However, this comes with a price: we will also have to sacrifice some correctly identified spectra.

For this tutorial, let's decide that quality level 2 is where we will strike our balance between coverage and quality. Perform the quality filter again, this time telling SpectraST to actually remove "Impure" and "Conflicting_ID" spectra. To do so, go back to the **SpectraST Library Import** page again, and select **.splib (perform join/build actions on SpectraST libraries)** as the file format. In the pane **2. Specify files to build/join**, browse to select the file consensus.splib (remove any other files already there). Select **Quality_Filter** in the **Select Build Action** drop-down menu. This time, name your output library "consensus_Q2" in the **Enter name of output file** box in the **4. General Options** pane. In the pane **6. Quality Filter Options**, select **2: …+spectra with look-alikes having conflicting IDs** in the drop-down menu **Quality Level to Remove**. This tells SpectraST to remove all spectra that are found to be impure. Uncheck the **Annotate Peaks** option. Click **Import Library Files**.

How many spectra are there in the resulting library consensus_Q2.splib? *(4,303)* Note that this is the number promised at the end of the spectrast.log file. For the purpose of this course, it is didactic to be able to see the library spectra about to be removed, but in real-life library building, one does not always need to apply the quality filter in this two-step manner (flag everything first, determine a suitable quality level, then remove). For most applications, a quality level of 2 is generally suitable, although more advanced users may want to go through this process to find the right balance between coverage and quality.

Step 4: Concatenating the yeast library to a decoy library

The spectral library you just created is extremely small and unsuitable for spectral searching. First, the number of candidates considered in each search is not large enough to form a reliable background statistically. Second, since we are about to search the same dataset we used to create the library against itself, we need to have some negative control. Put differently, we want to give the search engine some room to make mistakes and see if it makes them, thereby giving us a sense of how reliable our spectral search is.

To do so, we can concatenate our yeast library to a much larger human library. A "decoy" library of 16,649 human spectra that do not have any isobaric identical or homologous counterparts in the yeast library has been created for you.

To join your yeast library to the decoy library, go to the **SpectraST Library Import** page, select **.splib (perform join/build actions on SpectraST libraries)** again. In the **2. Specify files to build/join** pane, remove all previously added files, and add the files consensus_Q2.splib and human_decoy.splib from the directory class\SpectraST. In the **3. Select Actions** pane, select **None** in the **Select Build Action**, and **Union** in the **Select Join Action** menu. Type in "consensus_Q2_plus_decoy" in the **Enter name of output file** box. Uncheck the **Annotate Peaks** option. Scroll down and hit **Import Library Files**.

While you are waiting for it to run (it probably will take a couple minutes), it perhaps is good to point out another method of generating decoys for spectral searching. SpectraST can also generate decoy spectra by taking real spectra and randomly moving peaks around. These artificial spectra will mimic real spectra, but matches to them necessarily signal a false positive. This is a more robust and more convenient way of applying the decoy approach to spectral searching. For details, please refer to Lam et al., *Journal of Proteome Research* **9**, 605-610 (2010).

Step 5: Searching the original dataset against this spectral library

Now we are ready to re-search our original dataset against our newly built spectral library. To speed things up and to make it more interesting, we are going to search *only those spectra that were not identified in the original Tandem-K search (with probability no less than 0.9)*. These were also the spectra that we did not use to build our libraries. Two mzXML files: OR20080317_S_SILAC-LH_1-1_01_FILTERED.mzXML and OR20080320_S_SILAC-LH_1-1_11_FILTERED.mzXML are created for you that only contain spectra unidentified by Tandem-K.

Click on the **Home** link at the top. Then in the **Analysis Pipeline** pane, select **SpectraST** in the drop-down menu. Click on the **Analysis Pipeline** link at the top, and click on the **SpectraST Search** tab. To set up the search, first click **Add Files** in the first pane **1. Specify mzXML Files**, browse to the directory class\SpectraST, and select the 2 filtered files: OR20080317_S_SILAC-LH_1-1_01_FILTERED.mzXML and OR20080320_S_SILAC-LH_1-1_11_FILTERED.mzXML. In the second pane **2. Specify Library File**, select consensus_Q2_plus_decoy.splib in the same directory. In the pane **3. Specify a sequence**

database to be printed to the output file for downstream processing, browse to the directory class\dbase, and select yeast_orfs_all_REV.20060126.fix.fasta.

We are using all default options, so we are all set to go. Scroll down to the bottom of the page and hit **Run SpectraST**. Monitor the progress by viewing the **Output so far** pane.

The search should take about a minute or two. In the meantime, check out the page http://www.peptideatlas.org/speclib/. This website links to all the spectral searching for proteomics projects that we are aware of. You will find download links to spectral libraries and spectral library searching tools here, so be sure you check back for updates. We will be posting various spectral libraries derived from the PeptideAtlas project here.

When the search is done, run PeptideProphet as you normally would. Click on the **Analyze Peptide** tab, and add the two pepXML files containing the search results: OR20080317_S_SILAC-LH_1-1_01_FILTERED.xml and OR20080320_S_SILAC-LH_1-1_11_FILTERED.xml. Name the output file interact-spec.pep.xml (so as not to overwrite interact.pep.xml). Check the option **Use accurate mass binning**. Then go ahead and hit **Run XInteract** at the bottom of the page. Un-hide the **Command Status** pane, and **View** interact-spec.shtml when it is done.

Filter the identifications for probabilities above 0.9. How many hits remained? *(1,300)* Click on a few of them and see if they look good to you. Recall that all of these identifications were *missed* by Tandem-K in the previous search. How does the quality of these spectra compare to what you are used to seeing in previous sessions? What does this say about the sensitivity of spectral searching compared to sequence searching?

Feel free to play with the displaying options to the left of the spectra. The coloring scheme for the spectrum viewer is as follows. In the library (top) spectrum: *Red lines* = peaks assigned to known ions; *Blue lines* = unassigned peaks; *Red peak labels* = the ion assignment (you can customize what ion types to display on the left panel); red color indicates that this peak is present in the query spectrum as well; *Black peak labels* = indicates that this peak is missing in the query spectrum. In the query (bottom) spectrum: *Red lines* = peaks that match assigned peaks in the library spectrum; *Black lines* = peaks that do not match any assigned peaks in the library spectrum. In the ion table underneath the spectra: *Red colored boxes* = ions that are present in both spectra; *Pink-colored boxes* = ions that are present in the library spectrum only; *White colored boxes* = ions that are not present in either spectra.

Visualize the PeptideProphet models by clicking on any of the probabilities. Do they look reasonable? What is the sensitivity and error rates are a probability cutoff of 0.9? *(Sensitivity = 0.831, Error = 0.021)* What is the expected number of incorrect identifications among your positives (hits with probability >= 0.9)? *(1,300 × 0.021 = 27)*

As a sanity check, let's see if SpectraST finds any positive identification from the decoy (human) library. Because we only specified a yeast database for subsequent TPP processing, TPP will fail to map human peptides in the decoy library to any proteins. With the probability filter still on, filter the SpectraST hits for "UNMAPPED" proteins. How many did you find? *(0)* What

is the highest probability of any UNMAPPED protein? *(0.82)* Given the decoy (human) library is about 4 times larger than the target (yeast) library, what does it say about the accuracy of SpectraST? Does it give you additional confidence in these identifications that SpectraST found but Tandem-K missed?

Lastly, recall that we deliberate left some known false positives in the spectral library. These are spectra identified to "REV" proteins in the original Tandem-K search. With the probability filter still on, filter the SpectraST hits for "REV" proteins. How many did you find? *(4)* These identifications were of course wrong, but we cannot blame the spectral search engine for them. Here, we are simply propagating the error we made at the sequence searching step. That is why it is so important to use quality filters when we build spectral libraries to minimize such errors.

TPP On The Cloud

Joe Slagel

Day 3

October 27, 2010

Lecture topics

- Introduction to Cloud Computing and Amazon Web Services
- Setup and Trial of the new TPP Web Launcher for Amazon (TWA)
- Future TPP Direction with the Cloud

1

So What is Cloud Computing?

Cloud computing is Internet-based computing, whereby shared resources, software, and information are provided to computer and other devices on demand, like the electricity grid.
--Wikipedia

2

Three Aspects of Cloud Computing

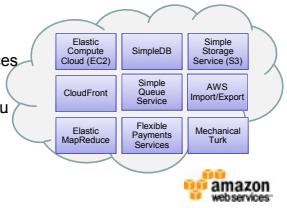
SaaS (Software as a Service)
Software applications available via the browser
E.g. Gmail, flickr, NCBI

IaaS (Infrastructure as a Service)
Storage, servers and networking components provided on demand through the internet
E.G Amazon EC2 & S3, Rackspace, IBM, HP,…

PaaS (Platform as a Service)
Hosted development environment for building and deploying cloud applications
E.g Google Apps, Microsoft Azure, Salesforce.com

Amazon Web Services

- Collection of web computing services offered by Amazon
- "Elastic" IT infrastructure – allocate computers, storage, and other services as needed
- Cost effective -- pay only for what you use
- Easy to use – simple API accessed over HTTP which supports almost every language
- Large number of tools available built for it



Amazon S3: Simple Storage Service

- S3 lets you store files/data on the "web" in "buckets"
  - Virtually unlimited storage, bandwidth, and # users
  - No loss of data – 99.999999999 % durability/yr
  - Always on - 99.99% availability/yr
- Files can range from 1 byte to 5 gigabytes in size with *no limit* to # files
- Authentication mechanisms ensure that data is kept secure and access rights can be granted to specific users.
- Uses standard http REST and SOAP interfaces to access the data that work with any language

First Tier S3 Pricing	
Storage	$0.150 per GB – first 50 TB / month of storage used
Data Transfer	In: $0.100 Out: $0.170 per GB first 10 TB/month
Requests	• $0.01 per 1,000 PUT, COPY, POST, or LIST requests • $0.01 per 10,000 GET and all other requests • Delete requests are free

Amazon S3: Management Console

Web based, secure management tool for managing S3 storage

Features include:

- Create/delete buckets
- Create/delete folders
- Upload or download files
- Modify properties (permissions)

6

Amazon EC2: Elastic Compute Cloud

- EC2 allows you launch new server instances in minutes
- Can choose from "Small" to "High-CPU Extra Large" instances
- Choice of large assortment of different OS images (Linux, Windows) or create your own image
- Billed only for actual usage + data transfers on a monthly basis
- Full control of the instance

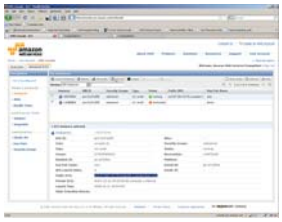
EC2 Pricing		
Small	1.7 GB 32-bit 1 Core 1 ECU[1] , 160 GB storage, Moderate I/O	$0.10/hr + I/O[2]
Large	7.5 GB, 32-bit 2 Core 2 ECU[1]each, 850 GB storage, High I/O	$0.40/hr + I/O[2]
High - CPU Extra Large	7GB, 64-bit 8 Core, 2.5 ECU[1] each 1690 GB storage, High I/O	$0.80/hr + I/O[2]

[1] EC2 Compute Units – One unit is equivalent CPU capacity of a 1.0-1.2 GHZ Operon or Xeon processor
[2] Data transfer in is $0.10/GB, transfer out is $0.17/GB for first 10 TB/month

Amazon EC2: Management Console

Web based, secure management tool for managing EC2

- Start and stop EC2 instances
- Find, manage, and create Amazon Machine Images (AMIs)
- Monitor instances with real time-operational metrics



Advantages of Cloud vs. Cluster

amazon web services

- Scalable
  - Unlimited amount of disk space
  - As many "servers" as needed
- Dependable
  - Large distributed system
  - Secure
  - Resilient
- Platform agnostic
  - CentOS, Debian, Ubuntu, Windows, etc.
  - 32-bit/64-bit
- No support costs

Traditional Cluster

- High initial startup cost
- Limited scalability
- Single point of failure
- Requires local IT personal for maintenance
- OS/Hardware lock-in
- Requires 3rd party grid software (PBS/GridEngine, etc)
- High initial costs and variable support costs
- Scheduling issues/complexity
- Users compete for resources

Advantages of Cluster vs. Cloud

amazon web services

- Performance issues
  - Bandwidth between instances
  - Bandwidth between S3 upload/downloads
  - Instance performance
  - Resource allocation (instances)
  - File I/O
- Cost model
  - Could get expensive
  - Pay as you go means you also pay for mistakes
- Lack of control
- Regulatory compliance

Traditional Cluster

- High Performance
  - Can finely tune network performance
  - Can finely tune hardware performance
  - Exceptional hardware capabilities
  - Exceptional File I/O
  - Dedicated resources
- Instances immediately available
- Host MPI Applications
- Security

Amazon Web Services Cost

Hypothetical Analysis

- 100 mzXML files
- Avg 100MB/file
- Avg 10 min/file

# EC2	Time (Hrs)	Cost
1	16.98	$8.10
5	3.7	$8.10
100	0.54	$11.34

Actual Results

Data Set	# EC2 / Threads	# Files	Scan Count	Time	Cost	Cost/ file
0021	20 / 8x	132	1,372,984	4:03	$ 50.05	$ 0.38
0049	20 / 8x	132	2,279,874	6:06	$ 46.24	$ 0.35

Using TPP on the Cloud

TPP Web Launcher for Amazon (TWA)

- Simple web based launcher to start petunia on a Amazon server
- Manages data flow and controls EC2 instances
- Doesn't require any software installation and is inexpensive to run
- Can use EC2 High CPU instance types
- Great tool for trying out TPP or increasing computing capabilities

Alpha software

TPP Amazon command line tools

- Advanced command line toolset: amztandem, amzomssa, amzinspect
- Launches parallel searches of files across multiple nodes
- Manage all aspects of cloud computing including data transfer, scheduling, and instances
- Great for quickly and inexpensively processing large amounts of data
- Production software

Getting Started: Account Creation

- Go to http://aws.amazon.com/ and click on the Sign Up Now" button 

1. Choose new user

2. Choose password

3. Enter Account Details

Getting Started: Product Sign Up

1. Click on the products menu

2. Choose Amazon Elastic Compute Cloud (EC2)
3. Click on the Sign Up Button 
   You will have to provide some sort of payment method (e.g. credit card, consolidated billings)
4. Repeat for Amazon Simple Service (S3)

Getting Started: Key Pairs

Amazon EC2 key pairs are needed to launch and securely access your Amazon EC2 Instances

Go to the Amazon console

1. Select Key Pairs
2. Click Create Key Pair
3. Enter a name for the key pair e.g. "TPP"
4. Save the key pair file.

15

Getting Started: Your Amazon Key ID and Secret Key

Your Amazon API key and secret key are used by 3rd party add-ons to access Amazon services

Go to the Amazon Web Services

1. Under the accout menu select "Security Credentials"
2. Choose the Access Keys tab (default)
3. Your Access Key ID should be displayed. Click "show" to temporarily show your secret key.

16

Using the TPP Web Launcher for Amazon (TWA)

1. Navigate to http://tools.proteomecenter.org/twa
2. Enter your Amazon Key ID and Secret
3. Click "Start Instance"
4. Welcome to Petunia
5. When you are done just click "Stop Instance"

17

TPP Cloud Future Directions

- Move TWA "alpha" to version 1.0
- Add capabilities to distribute TPP jobs in petunia to Amazon instances
- Include features for persistently storing MS data and TPP results in Amazon S3
- Include multiple file upload and download capability
- Ability to share data sets on the cloud

18

More Information

- Cloud computing with Amazon Web Services
  http://www.ibm.com/developerworks/library/ar-cloudaws1/
- Amazon Elastic Compute Cloud
  http://aws.amazon.com/ec2/
- Amazon Simple Storage Solution
  http://aws.amazon.com/s3
- TPP Cloud Services
  http://tools.proteomecenter.org/wiki/index.php?title=TPP:Cloud

19

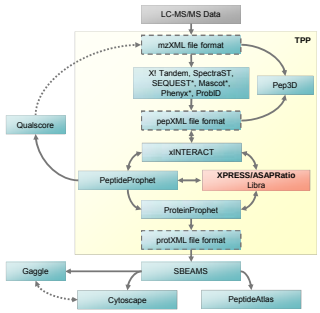
Quantitation with XPRESS and ASAPRatio

David Shteynberg

Day 4

October 28, 2010

Peptide and Protein Quantitation



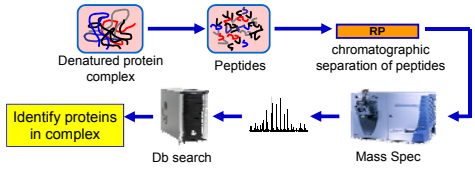
* Commercial software not part of TPP

Lecture Outline

- **Principles of quantitative proteomics using LC-ESI-MS/MS**
- Peptide and Protein Quantitation with XPRESS
  - Running XPRESS
  - Looking at results
- Peptide and Protein Quantitation with ASAPRatio
  - Running ASAPRatio
  - Looking at results
- Exercises

Summary of LC-ESI-MS/MS

- Protein mixtures are digested into peptides
- Peptides are concentrated and fractionated by separation technologies such as SCX, IEF, RP, etc.
- While eluting from RP column, peptides are ionized by ESI and analyzed by MS/MS
- Peptides are identified from CID spectra
- Peptides are usually quantified from MS signatures
  - Except in the case of iTRAQ



Complications

Shotgun MS detects peptides not proteins

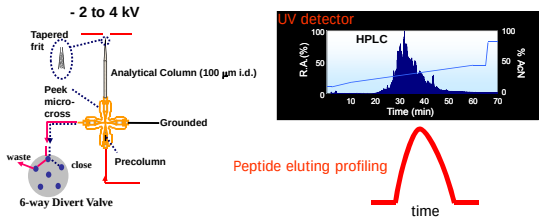
- Multiple peptides per protein
- Multiple proteins per peptide

Strong Cation-Exchange Chromatograph

- Fair but not great separation power
- Same peptide separated into several fractions

Reversed-Phase Chromatography

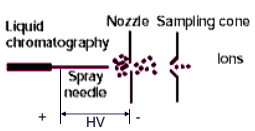
Reproducible: but a few erratic data points may exist

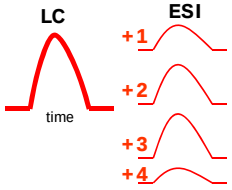


Electrospray Ionization

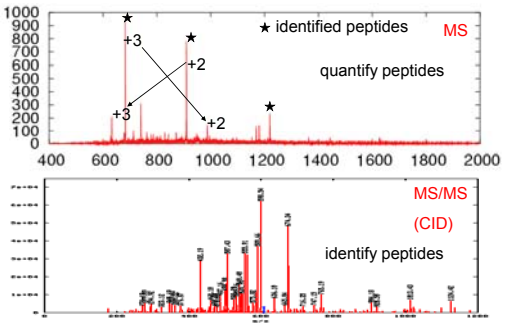
Multiple charge states: from +1 to +4

$M + z\,H^+ = M(H^+)_z$ $\quad m/z = (M+z*H)/z$





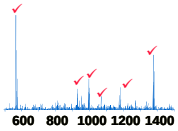
ESI-Tandem Mass Spectrometry

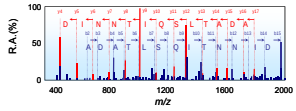


7

Peptide Identification

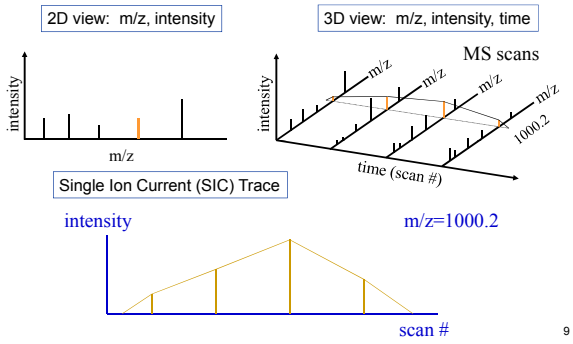
- Match CID (MS/MS) spectra with database
  - SEQUEST, MASCOT, X!Tandem, …
- Multiple IDs for the same peptide
  - different isotopes: light and heavy
  - different charge states: +1, +2, +3
  - repeating IDs: same isotope and same charge state





8

Single Ion Chromatogram



9

Peptide Quantitation

- Area under SIC is proportional to peptide abundance
- **PROBLEM**

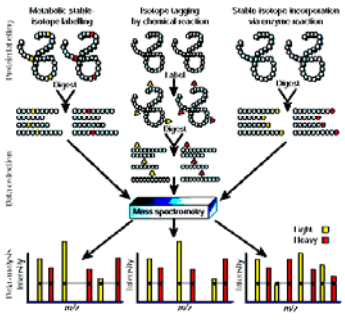
  Ionization efficiency of each peptide is different
  - Depends on the peptide molecular properties (e.g. number of basic residues)
- **ONE SOLUTION**

  Samples labeled with different stable isotopes
  - Chemically identical
  - Peptides are identified before quantification
  - Distinguishable by MS in mass shift
  - Peptide abundance ratio measured by ratio of SIC areas

10

Different Labeling Methods

- Metabolic labeling ^{13}C, ^{15}N, SILAC
- Chemical reaction ICAT, cleavable ICAT iTRAQ
- Enzyme reaction ^{18}O



1

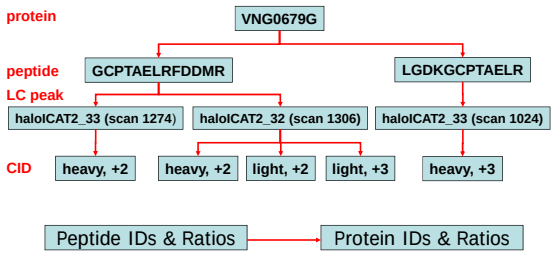
Summary of Quantitative LC-MS/MS Approach

- Samples are isotopically labeled
- Simultaneously identify & quantify thousands of proteins in complex samples
  - Peptide ion must be identified in MS^2 spectrum to be quantified
- Accuracy: ±10-30%
- Dynamic range: ~100 fold
- TPP provides 2 options: Xpress and ASAPRatio

12

Protein Identification and Quantification

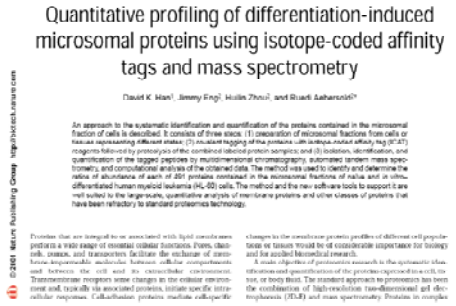
Hierarchy Structure



Lecture Outline

- Principles of quantitative proteomics using LC-ESI-MS/MS
- **Peptide and Protein Quantitation with XPRESS**
  - Running XPRESS
  - Looking at results
- Peptide and Protein Quantitation with ASAPRatio
  - Running ASAPRatio
  - Looking at results
- Exercises

XPRESS Publication

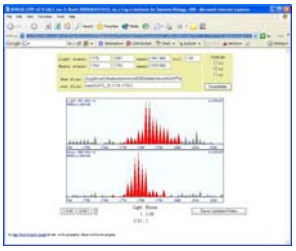


Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry

Han DK, Eng J, Zhou H, and Aebersold R. (2001) *Nature Biotechnology* 19:946-51.

XPRESS Peptide Ratio

- Calculated from SIC of charge state in which peptide was identified
- Smoothing done with a Butterworth low-pass filter
- No background estimation
- Works with different labeling methods
  - ICAT, SILAC



XPRESS Protein Ratio

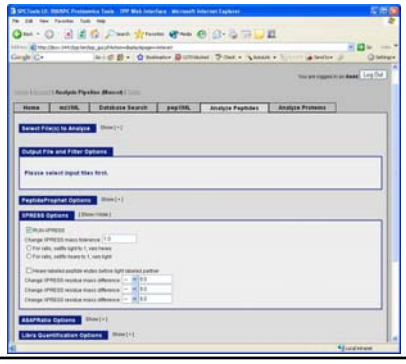
- Calculated as the Geometric Mean of the constituent peptide ratios

$$R = 10^{\frac{\sum \log r_i}{n}}$$

- Uncertainty is also calculated

$$\varepsilon = R \cdot \sqrt{\frac{\sum (\log r_i)^2}{n} - \left(\frac{\sum \log r_i}{n}\right)^2}$$

Running XPRESS: Petunia Interface



Running XPRESS: Command-line

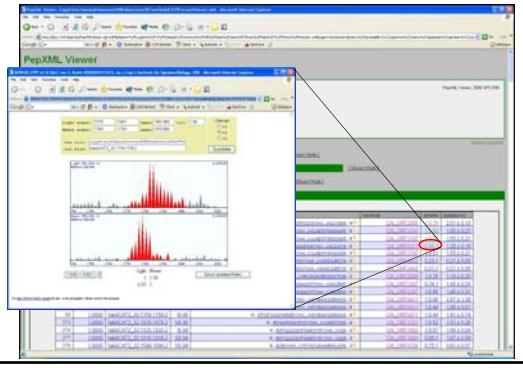
- Use the –X flag for xinteract

xpressoptions [will run XPRESS analysis with any specified options that follow the 'X']:

-m<num>	change XPRESS mass tolerance (default=1.0)
-l<str>	change labeled residues (default='C')
-n<str>,<num>	change XPRESS residue mass difference for <str> to <num> (default=9.0)
-b	heavy labeled peptide elutes before light labeled partner
-F<num>	fix elution peak area as +-<num> scans (<num> optional, default=5) from peak apex
-L	for ratio, set/fix light to 1, vary heavy
-H	for ratio, set/fix heavy to 1, vary light
-M	for metabolic labeling; ignore all other parameters, assume IDs are normal and quantify w/corresponding 15N heavy pair
-N	for metabolic labeling; ignore all other parameters, assume IDs are 15N heavy and quantify corresponding 14N light pair

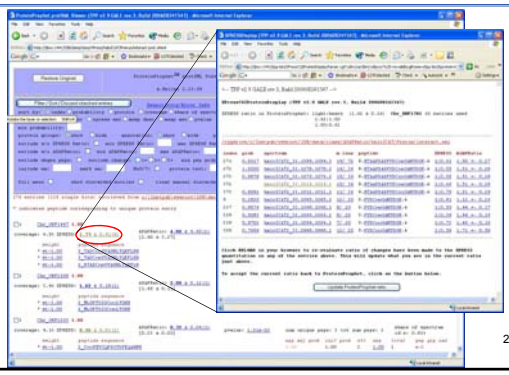
19

XPRESS PeptideProphet Results



20

XPRESS ProteinProphet Results



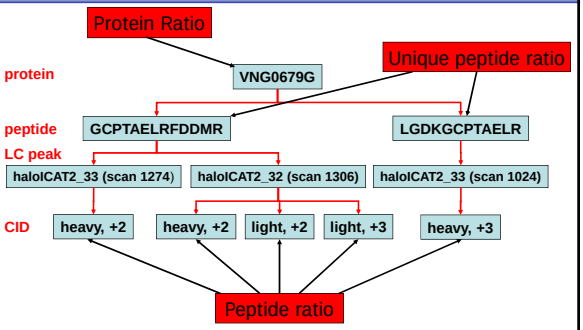
21

Lecture Outline

- Principles of quantitative proteomics using LC-ESI-MS/MS
- Peptide and Protein Quantitation with XPRESS
  - Running XPRESS
  - Looking at results
- **Peptide and Protein Quantitation with ASAPRatio**
  - Running ASAPRatio
  - Looking at results
- Exercises

22

Definitions



23

ASAPRatio Methodology

- Reconstruction of single-ion chromatograms
- Evaluation of peptide abundance ratios
- Evaluation of unique peptide abundance ratios
- Evaluation of protein abundance ratios
- Sample-dependent ratio normalization
- Large-scale protein profiling

Anal. Chem.; 2003; 75(23) pp 6648-6657.

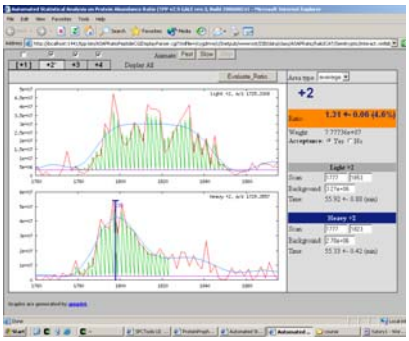
24

Reconstruction of Single-Ion Chromatogram

- Assume peptide identification correct
- Raw chromatogram
  - Summarize MS intensities within a m/z window and trace the sum in time
- Smooth chromatogram
  - Savitzsky-Golay smooth filter
- Subtract background and calculate area
- Estimate elution time of isotopic partner

25

Example on Single-Ion Chromatogram



Red: raw

Blue: fitting

Green: area

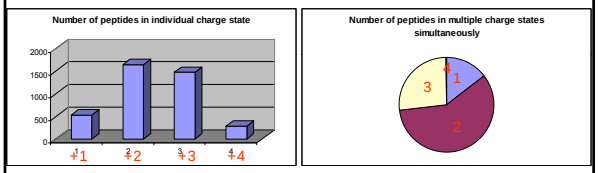
Pink: background

T-bar: CID

26

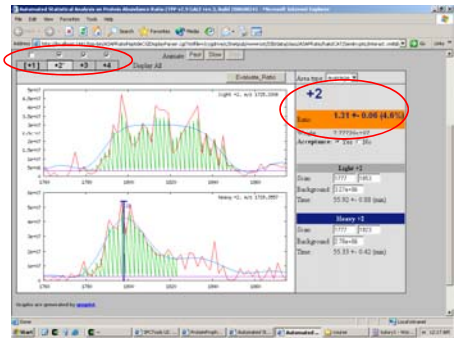
Peptide Charge Distribution

Out of 1857 peptides



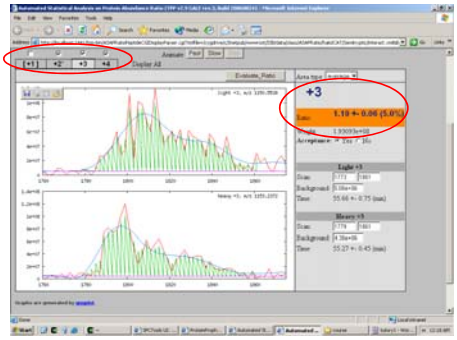
27

Single-Ion Chromatogram of +2 Ion



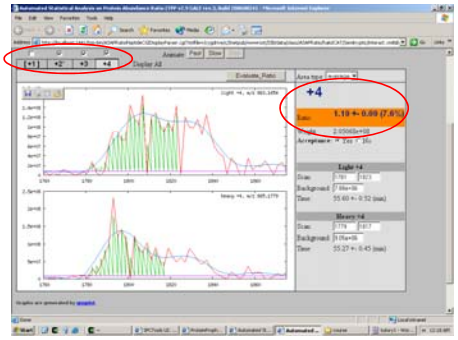
28

Single-Ion Chromatogram of +3 Ion



29

Single-Ion Chromatogram of +4 Ion



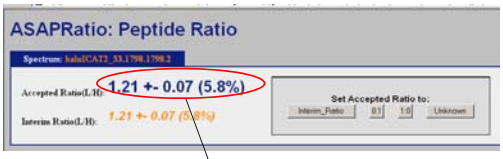
30

Evaluation of Peptide Abundance Ratio

- Evaluate a peptide ratio with error from each available charge states
- Use Dixon's test to identify any outliers
- Weight charge states by chromatogram areas
- Use statistical methods to calculate peptide ratio and error

31

Example on Peptide Ratio

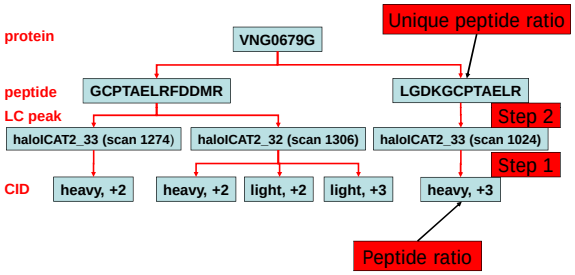


mean +- SD (CV%)
CV = SD/mean

SD: Std.Dev, CV: Coeff. Of Variation

32

Evaluation of Unique Peptide Abundance Ratio



33

Evaluation of Unique Peptide Ratio

Step 1

- Group abundance ratios of same peptide and same RP elution peak together
  - isotopic forms, charge states, repeats
- Most of them same
- If not
  - Weight data points by their largest chromatogram areas
  - Calculate mean and standard deviation
  - Use Dixon's test for outliers

34

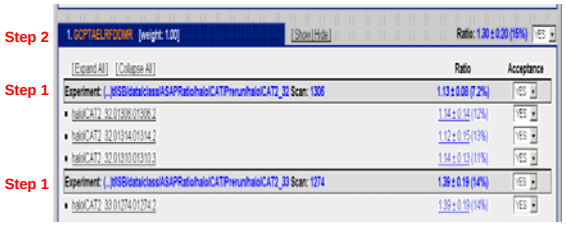
Evaluation of Unique Peptide Ratio

Step 2

- Group abundance ratios of same peptide but different RP elution peaks together
  - SCX fractions, RP elution times
- Weight data points by their largest chromatogram areas
- Calculate mean and standard deviation
- Use Dixon's test for outliers

35

Example of Unique Peptide Ratio



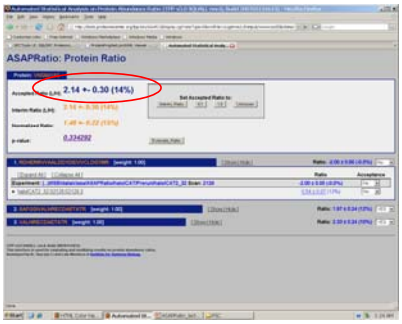
36

Evaluation of Protein Abundance Ratio

- Collect all unique peptide ratios of same protein together
- Use Dixon's test on outliers
  - misidentification, modification, etc.
- Weight data points by error
- Use statistical methods to calculate mean and standard deviation

37

Example on Protein Ratio

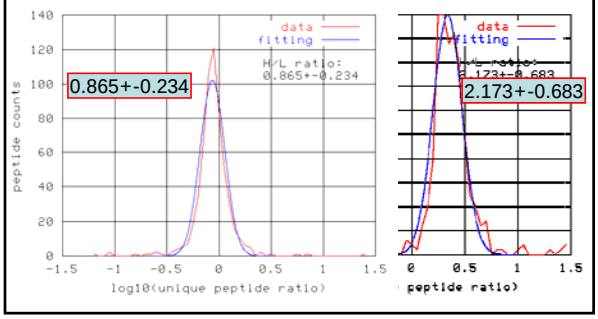


0.54 (outlier)
1.97
2.33

38

Sample-Dependent Ratio Normalization

To Correct Systematic Error Due to Sample Handling



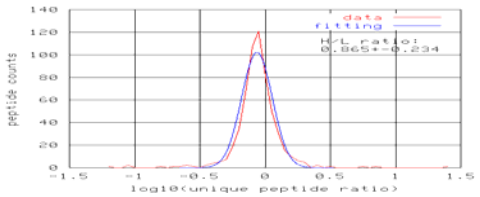
Sample-Dependent Ratio Normalization

Condition: Background Proteins Dominant

- Fit log10(unique peptide ratio) with normal distribution (Fig. 5, ASAPRatio paper)

$$n(r;A,r_0,\sigma) = A*\exp[-(\lg r/r_0)^2/2\sigma^2]$$

- Normalize protein ratios by peak ratio



$$\tilde{r} = r/r_0$$

40

Large-Scale Protein Profiling

- Evaluate p value for each protein

  p value: probability of a protein belonging to background group

$$p = erfc[|\lg r_P/r_0|/\sqrt{2((\Delta\lg r_P)^2+(\Delta\lg r_0)^2+\sigma^2)}]$$

- P value depends on: r_P Δr_P σ
- Specify significance level (by user)

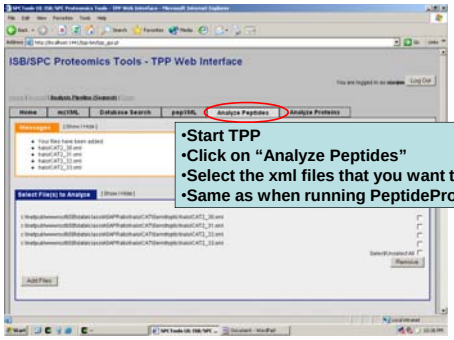
41

ASAPRatio Main Features

- Able to handle various labeling methods (except iTRAQ)
- Estimate error on peptide and protein ratios
- Calculate peptide ratios from multiple charge states
  - Not just from charge state in which the CID was matched
- Chromatogram signal background subtraction to increase the dynamic range
- Calculate protein ratios based on peptides that were assigned to proteins by ProteinProphet
- Evaluate p-value for protein profiling
- Detect outliers: Dixon's test
- Easy to use user interface for manual validation of ratios

42

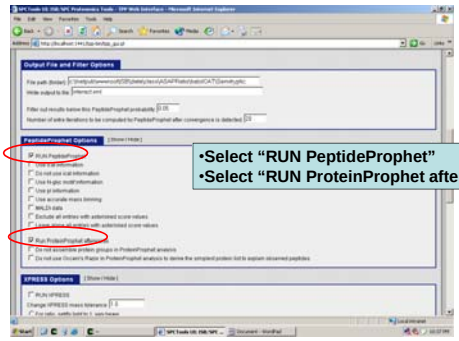
How to Use TPP for Data Analysis in Quantitative Proteomics



- Start TPP
- Click on "Analyze Peptides"
- Select the xml files that you want to analyze
- Same as when running PeptideProphet

43

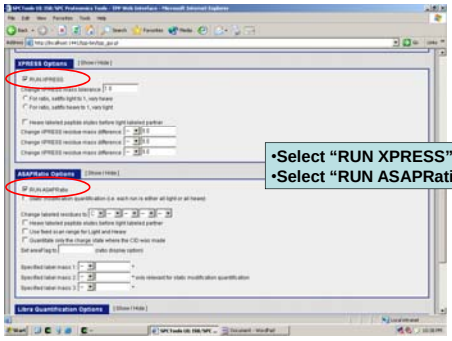
How to Use TPP for Data Analysis in Quantitative Proteomics



- Select "RUN PeptideProphet"
- Select "RUN ProteinProphet afterwards"

44

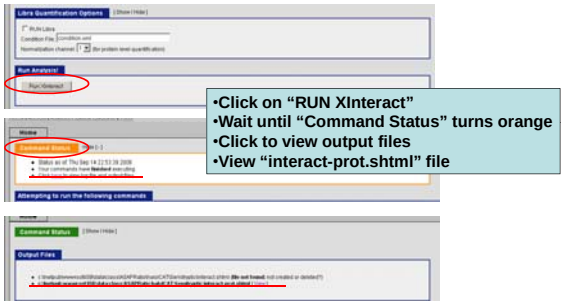
How to Use TPP for Data Analysis in Quantitative Proteomics



- Select "RUN XPRESS"
- Select "RUN ASAPRatio"

45

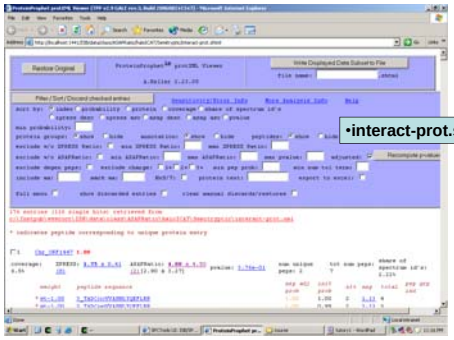
How to Use TPP for Data Analysis in Quantitative Proteomics



- Click on "RUN XInteract"
- Wait until "Command Status" turns orange
- Click to view output files
- View "interact-prot.shtml" file

46

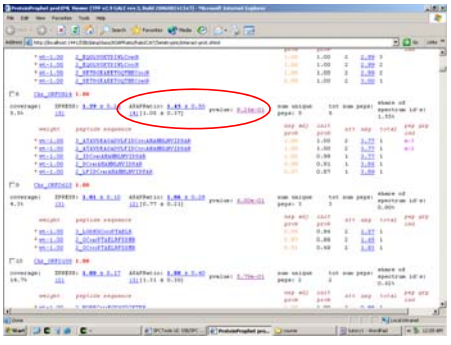
How to Use TPP for Data Analysis in Quantitative Proteomics



- interact-prot.shtml

47

How to Interpret ASAPRatio Results



48

How to Interpret ASAPRatio Results

Protein ratio and its standard deviation

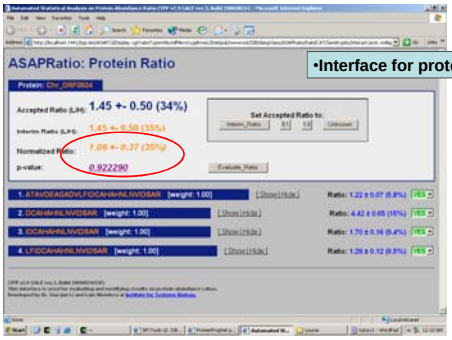
Protein p-value for differential expression

ASAPRatio: 1.45 ± 0.50 pvalue: 9.26e-01
(4) [1.05 ± 0.37]

Number of unique peptides

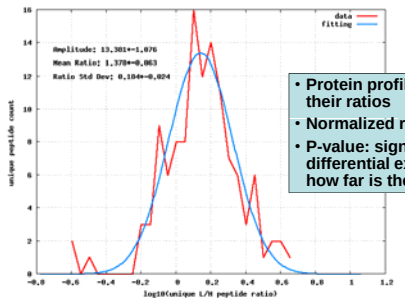
Normalized protein ratio and its standard deviation

How to Interpret ASAPRatio Results



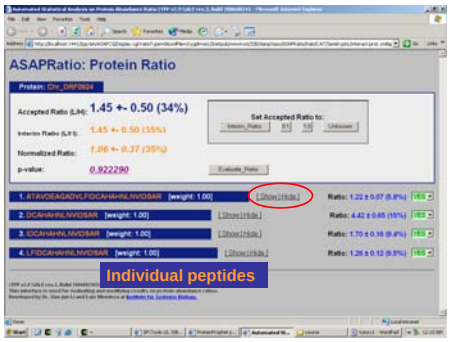
•Interface for protein ratio

How to Interpret ASAPRatio Results

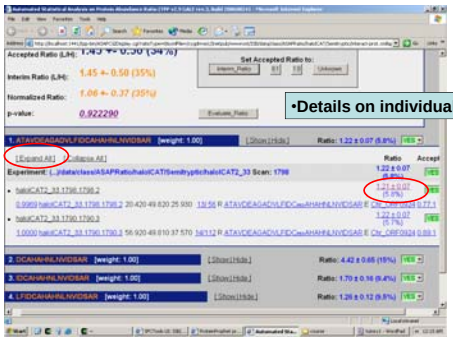


- Protein profiling based on their ratios
- Normalized ratio: $r^* = r/r_0$
- P-value: significance in differential expression; how far is the data from r_0

How to Interpret ASAPRatio Results

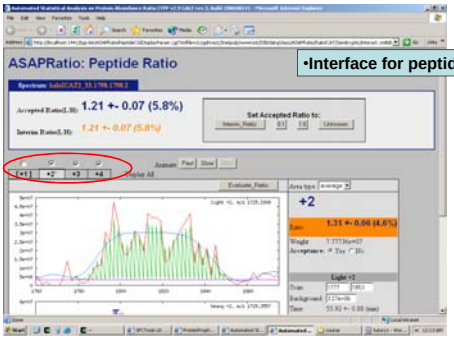


How to Interpret ASAPRatio Results



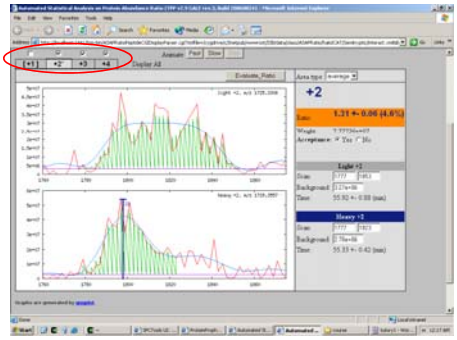
•Details on individual peptides

How to Interpret ASAPRatio Results



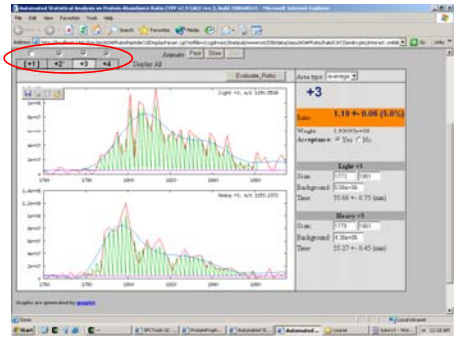
•Interface for peptide ratio

How to Interpret ASAPRatio Results



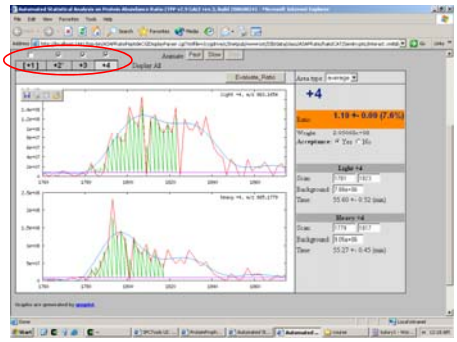
55

How to Interpret ASAPRatio Results



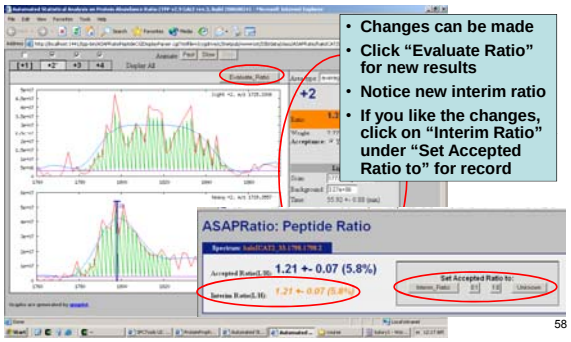
56

How to Interpret ASAPRatio Results



57

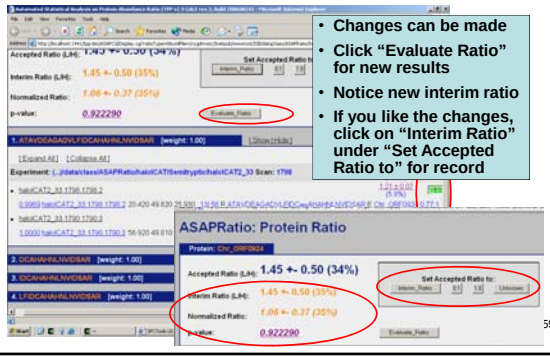
How to Interpret ASAPRatio Results



- Changes can be made
- Click “Evaluate Ratio” for new results
- Notice new interim ratio
- If you like the changes, click on “Interim Ratio” under “Set Accepted Ratio to” for record

58

How to Interpret ASAPRatio Results

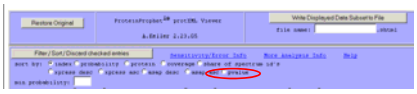


- Changes can be made
- Click “Evaluate Ratio” for new results
- Notice new interim ratio
- If you like the changes, click on “Interim Ratio” under “Set Accepted Ratio to” for record

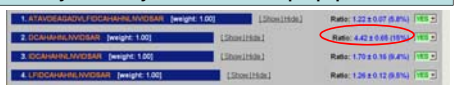
59

How to Interpret ASAPRatio Results

Sort by p values first and verify potentially interesting data



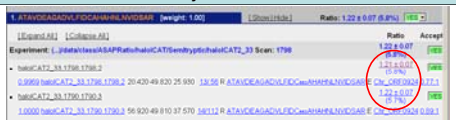
Identify and verify troublesome unique peptide ratios



60

How to Interpret ASAPRatio Results

For peptides of same experiment, verify one peptide ratio and reject others



Pay attention to unusual data: large error, 1:0, 0:1, or "unknown"

61

Lecture Outline

- Principles of quantitative proteomics using LC-ESI-MS/MS
- Peptide and Protein Quantitation with XPRESS
  - Running XPRESS
  - Looking at results
- Peptide and Protein Quantitation with ASAPRatio
  - Running ASAPRatio
  - Looking at results
- **Exercises**

62

Xpress and ASAPRatio -- Tutorial

Running Quantitation Tools - Please do the first three steps of the tutorial **before** the Quantitation lecture

1. Using Petunia go to the "*Analyze Peptides*" tab. Add the following files to the analyses:

    a. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Quantitation \ xtandem-k \ OR20080317_S_SILAC-LH_1-1_01.pep.xml**
    b. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Quantitation \ xtandem-k \ OR20080317_S_SILAC-LH_1-1_11.pep.xml**

2. In the "*PeptideProphet Options*" pane select to "*Use accurate mass binning"* and "*Run ProteinProphet Afterwards*". In the "*XPRESS Options*" pane select "*RUN XPRESS*". Change "*XPRESS Mass Tolerance"* to **0.1**, set "*Change XPRESS residue mass difference:"* with **K 8.0142** and **R 10.0083**. In the "*ASAPRatio Options*" pane select "*RUN ASAPRatio*". Change "*Labeled Residues"* to **K** and **R**, set "*m/z range to include in summation of peak:"* to **0.05**. Set "*Specified masses:"* to **M 147.035, K 136.10916** and **R 166.10941**. Check the "*Use fixed scan range*" option.

3. To run this analysis Click on *"Run XInteract"*. When the program is finished the results can be accessed through files "**c:\ Inetpub \ wwwroot \ ISB \ data \ class \ Quantitation \ xtandem-k \ semitryptic \ interact.pep.shtml**" and "**c:\ Inetpub \ wwwroot \ ISB \ data \ class \ Quantitation \ xtandem-k \ semitryptic \ interact.prot.shtml**".

For the next part of this tutorial open the file: **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Quantitation \ xtandem-k \ semitryptic \ interact.pep.shtml**

4. In the viewer, in the "*Pick Columns*" tab add the "*massdiff*" column to "*columns to display*". Next, on the "*Summary*" tab set a **minimum probability of 1,** and sort the spectra by "*peptide*" in "*descending*" order.

5. Click on the XPRESS ratio for spectrum - **OR20080320_S_SILAC-LH_1-1_01.03002.03002.2** (index 855 near the bottom of page 1). The page displayed shows the reconstructed chromatogram for the identified peptide. The raw signal is shown as vertical triangles, the smoothed chromatogram is displayed as a dotted line, and the region of the chromatogram used for quantitation is highlighted in red on the raw signal and blue on the smoothing curve. By default the program quantitates the chromatogram of the charge state in which the peptide was identified. In the top panel select a different charge state (change the Z parameter) and click the "Quantitate". What happens to the peak when you select charge state +3? What happens to the peak when you select charge state +1?

6. Go back to the PepXML Viewer, and click on the ASAPRatio link for the same spectrum. What is the ratio for this peptide, according to ASAPRatio? How does it compare to the one reported by XPRESS Unlike XPRESS, ASAPRatio tries to quantitate using chromatograms from all charge states where it can find a decent signal; in this case, it is using +2 and +3, even though the

identification was made in the +2 charge state. Look at the data for the rest of the charge states; did ASAPRatio do a good job of rejecting bad signals? You may have also noticed that ASAPRatio used a much larger scan range to quantitate than XPRESS; this is due to differences in the way these programs smooth the data to determine elution peaks. We'll learn how to adjust these boundaries below.

7. Close the ASAPRatio screen, as well as the XPRESS one if still open, without saving.

8. Using Petunia file browser open the file: **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Quantitation \ xtandem-k \ semitryptic \ interact.prot.shtml**.

9. Click on the Xpress ratio for protein **YAL044C** (the 4th entry). How many peptide ratios contribute to the ratio of this protein? Adjust the mass tolerance and the peak boundaries for the spectrum IDs that don't have a reasonable quantitation peaks displayed. For spectra that don't show reasonable elution profiles, set the ratios to unknown by clicking on the question mark button below the lower left corner of the chromatogram image. Do not close the Xpress pages, after changing each Peptide Ratio, refresh the Protein Ratio page and make sure the change gets correctly recorded, then go on to the next peptide ratio. What is the Xpress Protein Ratio and Error after you've corrected the peaks?

10. Click "Update ProteinProphet ratio" button on the Protein Ratio page. Close the Protein Ratio page and refresh the ProteinProphet page.

11. Now let's look at the ASAPRatio analysis; click on the ASAPRatio link for this protein. What is the protein ratio as evaluated by ASAPRatio? What is the normalized protein ratio? What is the protein p-value? How many peptide sequences contribute to the protein ratio?
12. Click on the p-value link. What is the computed mean peptide ratio in this dataset? What is the standard deviation?
13. Click on "[Expand All]" under the peptide sequence "LGEGVNVEQVEGLMSLEQYEK." How many independent LC peaks were detected for the peptide? How many times was the peptide identified? In what isotopic forms and charge states was the peptide identified? What are the peptide ratios at those identifications? Why are the last two peptide ratios so similar? What is the unique peptide ratio? What is the CV? How was the CV calculated?

    a. Click on the Peptide Ratio link of the identification **OR20080317_S_SILAC-LH_1-1_01.09309.09309.2**. In what isotopic form and charge state was the peptide identified? In what charge states were signals of the peptide detectable? What were the charge states that contributed to the calculation of peptide ratio? Adjust the peaks and charge states used to compute the peptide ratio to reduce the error and save the Interim Ratio you are happy with.
    b. Click on the Peptide Ratio link of the identification **OR20080320_S_SILAC-LH_1-1_11.09396.09396.2**. In what isotopic form and charge state was the peptide identified? In what charge states were signals of the peptide detectable? What were the charge states that contributed to the calculation of peptide ratio?
    c. Refresh the "ASAPRatio: Protein Ratio" page, what happens to the error in the Interim Protein Ratio?

14. Now let's look at the ASAPRatio analysis for protein **YAL012W**. What is the protein ratio as evaluated by ASAPRatio? How many independent peptides contribute to the evaluation? What is the normalized protein ratio? What is the protein p-value?

15. Which peptides contributed to the evaluation of protein ratio? What are their ratios?

16. Click on "[Show | Hide]" next to the 2nd peptide, "ISVGIEDTDDLLEDIKQALK", of the protein, and then click on "[Expand All]". How many independent LC peaks were detected for the peptide? How many times was the peptide identified? In what isotopic forms and charge states was the peptide identified?

    a. Click on the Peptide Ratio link of the identification OR20080320_S_SILAC-LH_1-1_11.10488.10488.3. In what isotopic form and charge state was the peptide identified? In what charge states were signals of the peptide detectable? What were the charge states that contributed to the calculation of peptide ratio? What went wrong with this quantitation? Adjust this peak and save the ratio.

    b. Refresh (reload) the Protein Ratio interface in order to view the recent changes. What is the new ("Interim") protein ratio?

17. The 4th peptide (QFLQNAIGAIPSPFDAWLTHR) has a ratio that is quite higher than the others; "show" the peptide section, and click on the ratio link therein. What is the peptide ratio? Now look at the areas that ASAPRatio picked for quantifying; what could be throwing off the ratio?

    a. Modify the peaks contributing to this peptide ratio to get a reasonable ratio or invalidate the ratio.

    b. Go back to the Protein Ratio interface. Click on "Evaluate_Ratio". What is the new protein ratio?

18. The 6th peptide (YINGHSDVVLGVLATNNKPLYER) has an error that is high "show" the peptide section, and click on the ratio link therein. What is the peptide ratio? Now look at the areas that ASAPRatio picked for quantifying; what could be throwing off the ratio?

    c. Modify the peaks contributing to this peptide ratio to get a reasonable ratio or invalidate the ratio.

    d. Go back to the Protein Ratio interface. Click on "Evaluate_Ratio". What is the new protein ratio?

19. Click on "Interim_Ratio" under "Set Accepted Ratio to". What is the Accepted Ratio now?

20. Go back to the "interact-prot.shtml" file and refresh the browser. What the ratio of the protein now?

iTRAQ Quantitation with Libra

Luis Mendoza

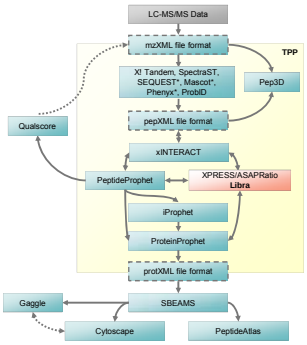
Day 4

October 28, 2010

INSTITUTE FOR Systems Biology™

Revolutionizing science. Enhancing life.

Peptide and Protein Quantitation



* Commercial software not part of TPP

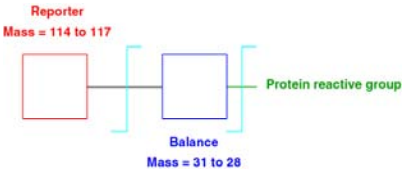
1

Libra Overview

- Libra MS^2-level Quantitation
  - Not alternative to ASAPRatio or XPRESS
  - An alternative labeling technique
- Background on iTRAQ
- Peptide Quantitation
- Protein Quantitation
- Running Libra
- Looking at results

2

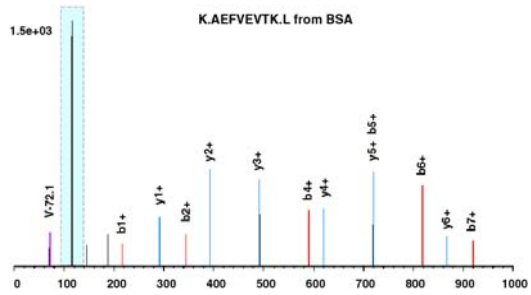
iTRAQ



Four reagents with the same mass and same retention time on reverse phase chromatography, but different **reporter ions** upon fragmentation.

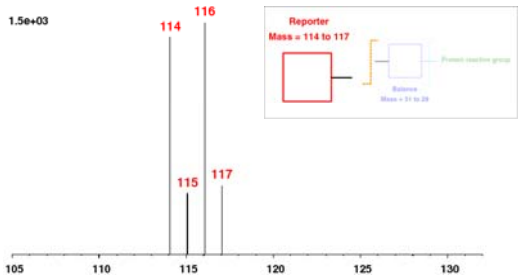
3

iTRAQ



4

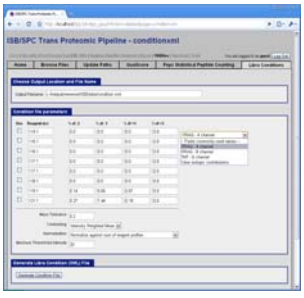
iTRAQ



5

Libra

- Parses mzXML files to **extract intensities** for a defined list of m/z values
- Controlled by an **XML parameter file** (condition.xml)
- Can be generated in Petunia (up to 8 channels)



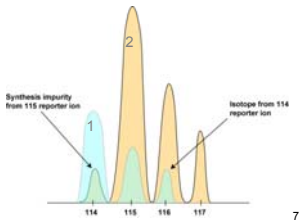
Old: http://db.systemsbiology.net/webapps/conditionFileApp/

6

Intensity Correction

- Raw **intensities** must first be **adjusted**
  - corrects for isotopic and synthesis by-products

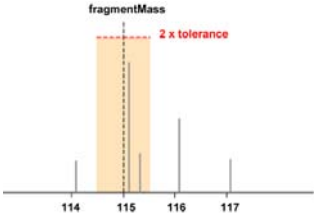
```
<isotopicContributions>
  <contributingMz value="1">
    <affected mz="2" correction="0.063" />
    <affected mz="3" correction="0.0020" />
  </contributingMz>
  <contributingMz value="2">
    <affected mz="1" correction="0.02" />
    <affected mz="3" correction="0.06" />
    <affected mz="4" correction="0.0010" />
  </contributingMz>
  .....
</isotopicContributions>
```



7

Mass Tolerance

- The m/z **tolerance** for matching a **target ion**
- If multiple ions are in the tolerance interval only the most intense one will be used



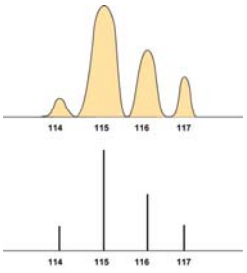
8

Centroiding

Controls conversion of **profile** data into **centroid** mode.

Types:

0. None
1. Mathematical
2. Intensity weighted



9

Other Libra Parameters

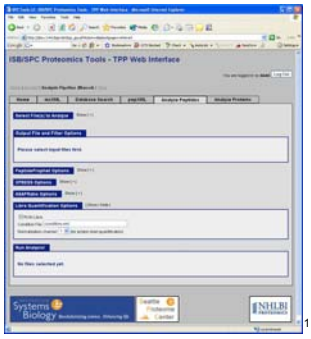
- Normalization
- Target MS level
  - default 2
- Output Options

http://tools.proteomecenter.org/wiki/index.php?title=Software:Libra

10

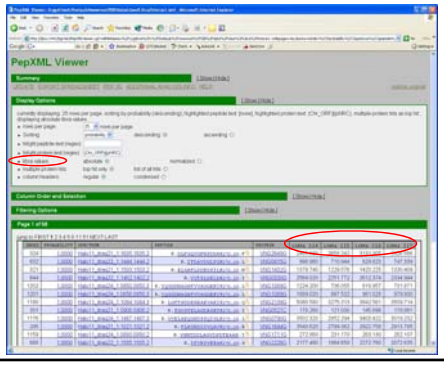
Running Libra

1. Generate a condition.xml file
2. Copy the condition file to data directory
3. Add input files and set Prophet options
4. Check “RUN Libra” and specify condition file by name



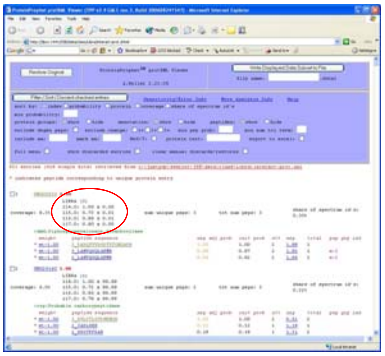
11

Libra PeptideProphet Results



12

Libra ProteinProphet Results



13

quantitation.tsv Auxiliary File

#protein	peptide	pepratio0	pepratio1	pepratio2	pepratio3	inten0	inten1	inten2	inten3	kept?	protratio0	protratio1	protratio2	protratio3	proterr0	proterr1	proterr2	proterr3
VNG0510G											0.25	0.26	0.26	0.22	0.04	0.03	0.01	0.01
VNG0510G	PHPNVGGR	0.18	0.31	0.27	0.23	590.52	1008.15	861.41	754.78	Yes								
VNG0510G	AVASQTDASFIK	0.32	0.23	0.24	0.21	1263.5	913.39	973.71	850.38	Yes								
VNG0510G	QHGNNQEALTEVTDELR	0.27	0.25	0.26	0.22	410.66	378.35	388.12	339.6	Yes								
VNG0147C											0.28	0.23	0.26	0.23	0.02	0	0.01	0.02
VNG0147C	NLLGEFEDK	0.26	0.23	0.26	0.25	569.2	485.99	551.74	546.48	Yes								
VNG0147C	LDDFADDHGYTGR	0.3	0.23	0.27	0.21	807.2	614.94	744.49	562.9	Yes								
VNG0166G											0.27	0.23	0.26	0.25	0.01	0.01	0.01	0.01
VNG0166G	GTASIGVR	0.26	0.24	0.23	0.28	156.84	143.47	139.12	167.04	Yes								
VNG0166G	TPEGVVLVVDK	0.25	0.23	0.27	0.24	1882.72	1728.55	2043.2	1831.15	Yes								
VNG0166G	LFETDPSGTSNEWK	0.28	0.22	0.26	0.24	1791.04	1382.09	1617.41	1546.37	Yes								
VNG0166G	SVTDHIQQYTQVGGAR	0.27	0.22	0.28	0.24	597.28	478.48	610.62	531.6	Yes								
VNG0177G											0.31	0.21	0.25	0.22	99.99	99.99	99.99	99.99
VNG0177G	TLNSYWVGEDGSQK	0.31	0.21	0.25	0.22	420.5	287.53	335.23	300.29	Yes								
VNG0258H											0.28	0.22	0.25	0.25	0	0	0	0.01
VNG0258H	ELEEYYGEEVNHGR	0.28	0.21	0.25	0.25	266.46	202.47	235.32	238.27	Yes								
VNG0258H	ELEEYYGEEVNHGR	0.28	0.22	0.26	0.24	1269.56	1003.8	1165.55	1083.88	Yes								
VNG0259G											0.25	0.19	0.29	0.27	0.01	0	0.01	0.01
VNG0259G	AEIAEFFETYK	0.59	0.41	0	0	6905.06	4811.98	0	0	No								
VNG0259G	AEIAEFFETYK	0.27	0.19	0.26	0.28	2278	1584.13	2215.04	2323.88	Yes								
VNG0259G	AEIAEFFETYK	0.21	0.18	0.35	0.26	216.8	185	354.91	271.43	Yes								
VNG0259G	ETAVLGWGDAQAAK	0.27	0.2	0.27	0.26	2068.88	1557.08	2101.8	2061.45	Yes								
VNG0259G	AEIAEFFETYK	0.24	0.2	0.28	0.28	1659.7	1398.34	1901.66	1933.42	Yes								
VNG0259G	PVALMEMDDDGEQDDK	0.26	0.2	0.28	0.27	210.4	159.4	225.78	219.78	Yes								
VNG0259G	AEIAEFFETYK	0.27	0.19	0.29	0.24	564	394.85	607.7	506.83	Yes								

14

Libra – Exercises

1. Using Petunia go to the "*Analyze Peptides*" tab. Add the following files to the analyses:

    a. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Libra \ Halo11_itraq21_1.xml**
    b. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Libra \ Halo11_itraq22_1.xml**
    c. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Libra \ Halo11_itraq24_1.xml**
    d. **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Libra \ Halo11_itraq25_1.xml**

2. In the "*PeptideProphet Options*" pane select "*Run ProteinProphet Afterwards*". In the "*Libra Options*" pane select "*RUN Libra"*. Click "*Run XInteract*" to launch the program.

3. The analysis of this data should take no more than a couple of minutes. When the program is finished open the peptide results file: **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Libra \ interact.shtml**, the results of running Libra quantitation software are captured in columns beginning with "Libra". In the viewer, select to sort the spectra by probability in descending order.

4. Under "*Display Options*" in the PepXMLViewer change libra values from "*absolute*" to "*normalized*" and click to "*Update Page*". What is the effect on the reported values? [ANS: normalized against the 114.1 channel]

5. Select a spectrum with high probability and click on its matched ions link. In the spectrum viewer controls check the "*zoom 112-122*" checkbox and click "*GO*". Compare the relative reporter ions intensities to the Libra values reported in the viewer

6. Examine several other spectra of various probabilities by visually comparing the reporter ion intensities to the values reported by Libra.

7. Using Petunia file browser open the file **c: \ Inetpub \ wwwroot \ ISB \ data \ class \ Libra \ interact-prot.shtml**.

8. For several proteins in the file, compare the displayed Libra values to the Libra values of their constituent peptides.
    a. HINT: Click on the peptide sequences to see the Libra peptide values

9. Using the Petunia File Browser tool find and View the file: **c:\Inetpub\wwwroot\ISB\data\class\Libra\quantitation.tsv**. This file contains the information about the protein abundance ratios calculated by Libra for each protein identified by ProteinProphet using the constituent peptide iTRAQ abundance ratios. The constituent peptide ratios for each protein are listed below the protein entry, along with information about whether or not a given peptide's ratio contributes to the protein ratio (the "kept?" column).

Discovery and Validation Tools for Biomarker Research: Corra

Mi-Youn Brusniak

Day 4

October 28, 2010

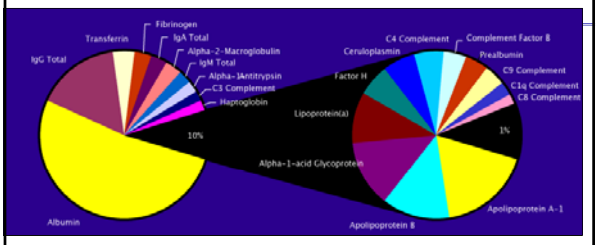
Outline of Discussion

- Day 1: Discussion of Discovery Tools for Biomarkers
  - Label Free Quantification
  - Introduction to Corra and Hands on Demo Using Tutorial
  - Introduction to PIPE

Outline of Discussion

- Day 1: Discussion of Discovery Tools for Biomarkers
  - Label Free Quantification
  - Introduction to Corra and Hands on Demo Using Tutorial

Challenge to Biomarker Discovery



Serum albumin represents >50% total serum protein itself
10 most abundant serum proteins represent 90% total protein
22 most abundant serum proteins represent 99% total protein

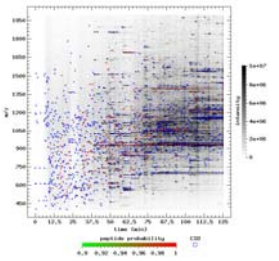
Challenges in Plasma/Tissue Based Biomarker Discovery in Proteomics

Challenge

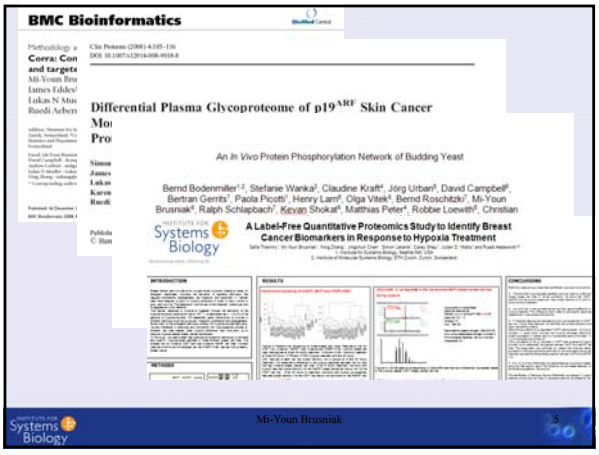
- High abundance proteins mask the 'interesting' lower abundance proteins
- Especially problematic for the study of plasma (albumin etc.)
- Tissue/cell line heterogeneity

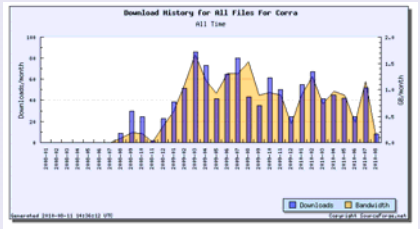
Our Approaches

- Selective enrichment of *N*-glycosylated proteins
- Development of an MS1 analytical workflow (*Corra*)
- Hypothesis driven validation workflow (TIQAM) using MRM



Features: 2720 CIDs: 1633 IDs: 363
ID/feature: 13%





- **ETH**
- **UW**
- **Harvard Medical School**
- **Integrated Analysis**
- **Yale Center for Statistical Genomics and Proteomics**
- **Internal ISB Collaborators**

Principles of Quantitative Proteomics

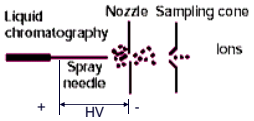
- Protein mixtures are digested into peptides
- Peptides are concentrated and fractionated by separation technologies such as SCX, IEF, RP, etc.
- While eluting from RP column, peptides are ionized by ESI and analyzed by MS/MS
- Peptides are identified from CID/ETD spectra
- Peptides are quantified from MS1, MS2 and SRM trace.

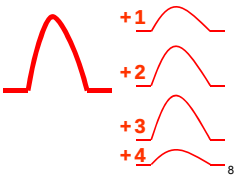
7

Principles of Quantitative Proteomics
MS1 Based Quantification

- Multiple charge states: from +1 to +6

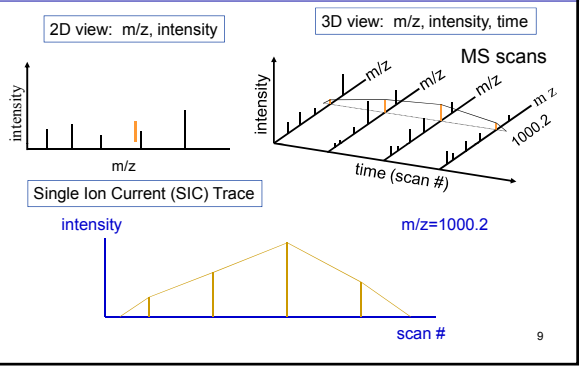
$M + z\,H^+ = M(H^+)_z$ $\quad m/z = (M+z*H)/z$





8

Principles of Quantitative Proteomics
Single Ion Chromatogram



9

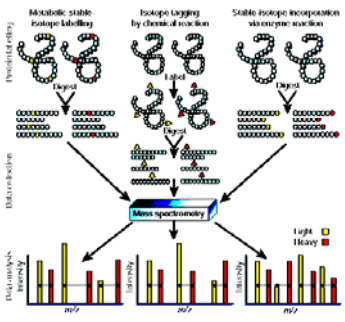
Principles of Quantitative Proteomics
Peptide Quantification

- Area of SIC is proportional to peptide abundance
- Ionization efficiency of each peptide is different
  - Depends on the peptide molecular properties (e.g. number of basic residues)
- MS Technology is NOT absolute quantitative measurement technology.

10

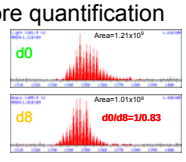
Labeling Approaches for Quantitative Proteomics

- Metabolic labeling ^{13}C, ^{15}N, SILAC
- Chemical reaction ICAT, cleavable ICAT iTRAQ
- Enzyme reaction ^{18}O



Labeling Approaches for Quantitative Proteomics

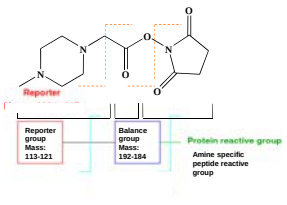
- Samples labeled with different stable isotopes
- Chemically identical
- Distinguishable by MS in mass shift
- Peptide abundance ratio measured by ratio of SIC areas
- Peptides are identified before quantification

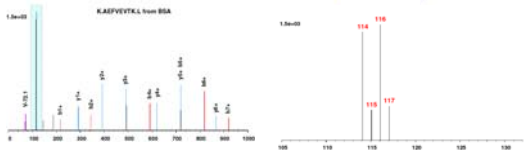


12

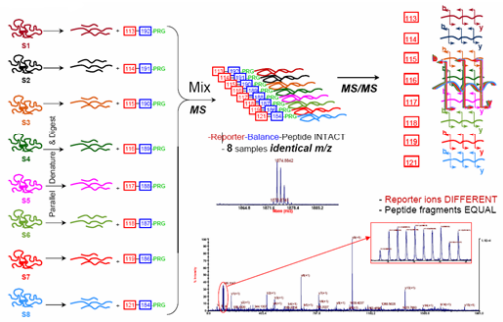
Labeling Approaches for Quantitative Proteomics

- **Eight** reagents with the same mass and same retention time on reverse phase chromatography, but different **reporter ions** upon fragmentation.
- **Libra** is a for analyzing iTRAQ data





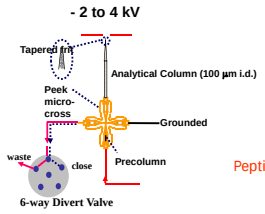
Labeling Approaches for Quantitative Proteomics

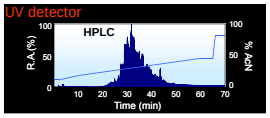


14

Principles of Quantitative Proteomics Reversed-Phase Chromatography

- Separate peptides by hydrophobicity
- Reproducible
- Automated, coupled online with MS





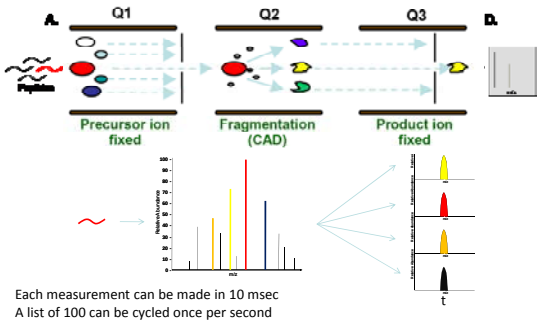
15

Principles of Quantitative Proteomics Peptide Quantification

- Area of SIC is proportional to peptide abundance
- Ionization efficiency of each peptide is different
  - Depends on the peptide molecular properties (e.g. number of basic residues)
- MS Technology is NOT absolute quantitative measurement technology.

16

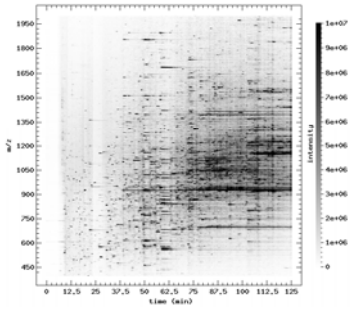
Principles of Quantitative Proteomics Transition Trace Based Quantification (SRM)



Each measurement can be made in 10 msec
A list of 100 can be cycled once per second
Elution over 20 seconds will have 20 points to generate a quantifiable peak

17

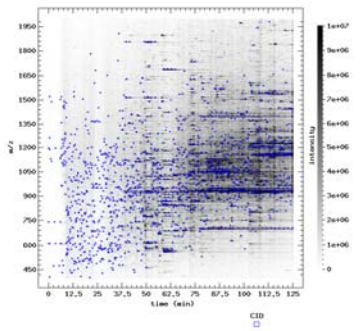
Principles of Quantitative Proteomics MS1 Based Quantification With Label Free



Features: 2720

18

Principles of Quantitative Proteomics
MS1 Based Quantification With Label Free

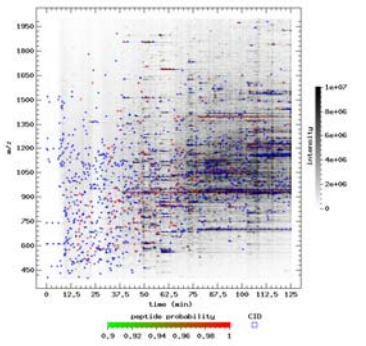


Features: 2720

CIDs: 1633

19

Principles of Quantitative Proteomics
MS1 Based Quantification With Label Free



Features: 2720

CIDs: 1633

IDs: 363

ID/CID: 22%

ID/feature: 13%

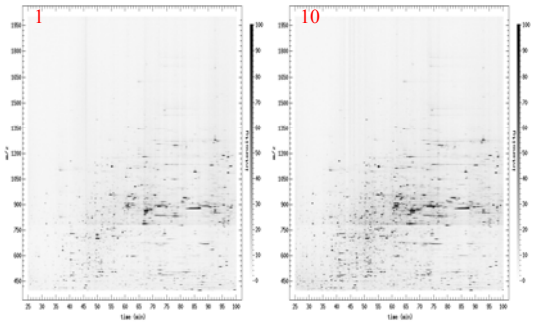
20

Principles of Quantitative Proteomics
MS1 Based Quantification With Label Free

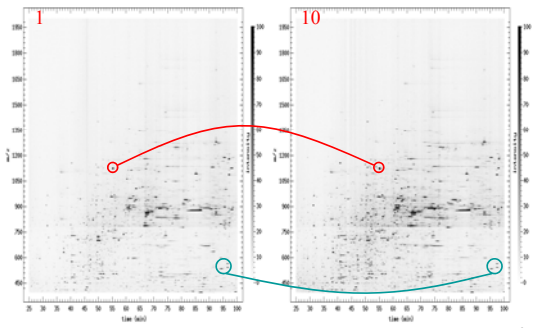
- Sample size limited:
  - ICAT (2), iTRAQ (8), …
- Difficult to trace protein abundance across a large number of samples
- Most peptides cannot be identified
- Difficult to identify & quantify low-abundance proteins

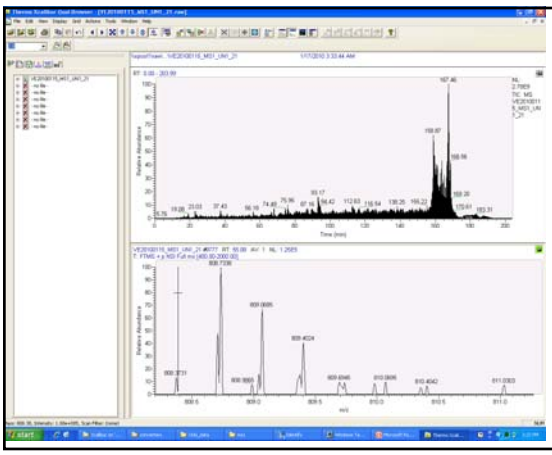
21

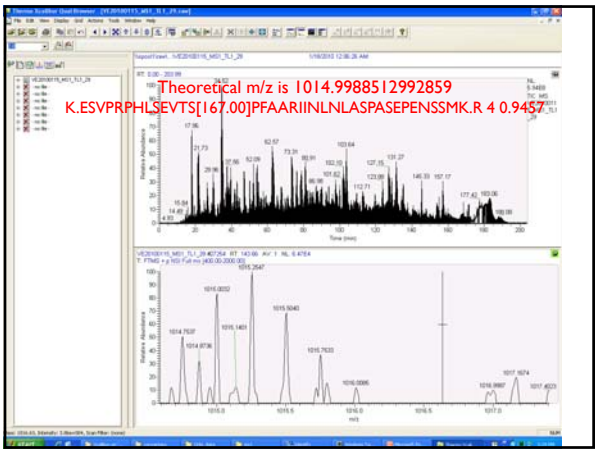
Non-Labeling Approaches for Quantitative Proteomics



Non-Labeling Approaches for Quantitative Proteomics





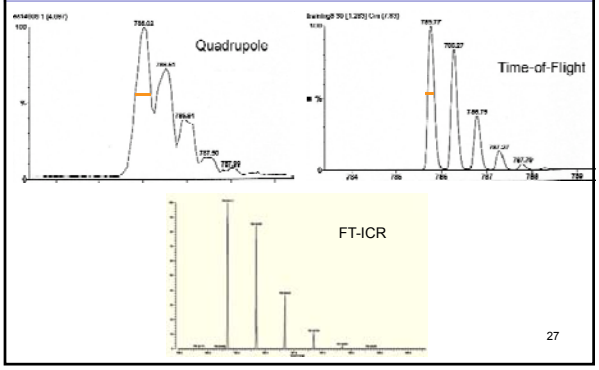


Principles of Quantitative Proteomics MS1 Based Quantification With Label Free

- Samples are NOT labeled
- Samples are analyzed under identical settings
- Peptide abundance is evaluated by MS signal intensity in different runs
- Reproducibility in LC-MS analysis critical
- Peptide alignment crucial
- Followed by target LC-MS/MS
- Suitable for large scale protein profiling

26

Mass Spectrum High Resolution and Mass Accuracy



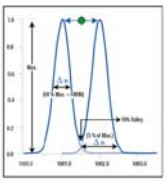
27

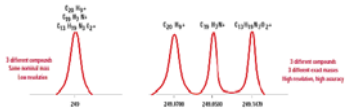
Mass Spectrum High Resolution and Mass Accuracy

Resolution: Represents the ability to separate two adjacent masses. It measures the "sharpness" of the MS peak.

Mass Accuracy: Indicates the accuracy of the mass information provided by the mass Spectrometer

$ppm = 10^6 * (m_{real}-m_{measured})/m_{measured}$





• A low resolution instrument cannot provide a high mass accuracy

28

Principles of Quantitative Proteomics MS1 Based Quantification With Label Free

Published Tools

SpecArray - Software Suite for the Generation and Comparison of Peptide Arrays from Sets of Data Collected by Liquid Chromatography-Mass Spectrometry
Xiao-Jun Li *et. al. Molecular & Cellular Proteomics 4.9, 2005*

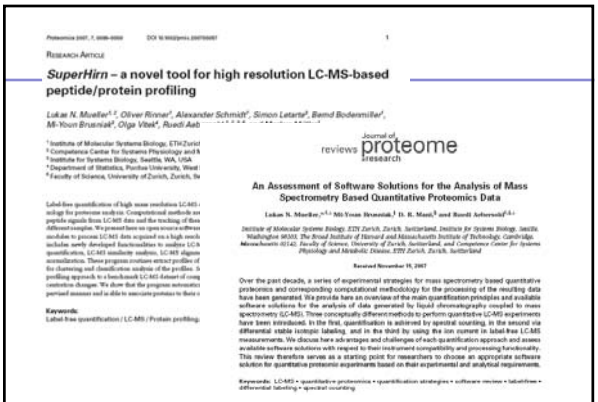
SuperHirn – a novel tool for high resolution LC-MS based peptide/protein profiling
Lukas N. Muller et. al. Proteomics, 2007

MZmine – toolbox for processing and visualization of mass spectrometry based molecular profile data
Mikko Katajamaa *et. al. Bioinformatics, 2006*

MSInspect – A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS
Matthew Bellew *et. al. Bioinformatics, 2006*

PEPPeR – A Platform for Experimental Proteomic Pattern Recognition
Jacob Jaffe *et. al Molecular & Cellular Proteomics 2006*

29



RESEARCH ARTICLE

SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling

reviews proteome research

An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data

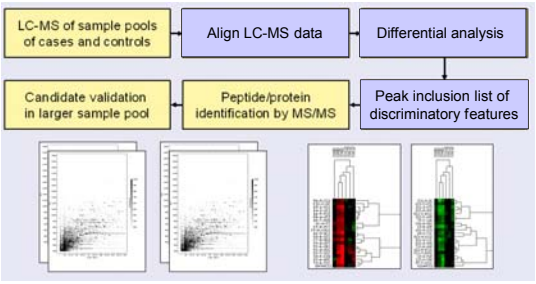
30

Outline of Discussion

- Day 1: Discussion of Discovery Tools for Biomarkers
  - Label Free Quantification
  - Introduction to Corra and Hands on Demo Using Tutorial

31

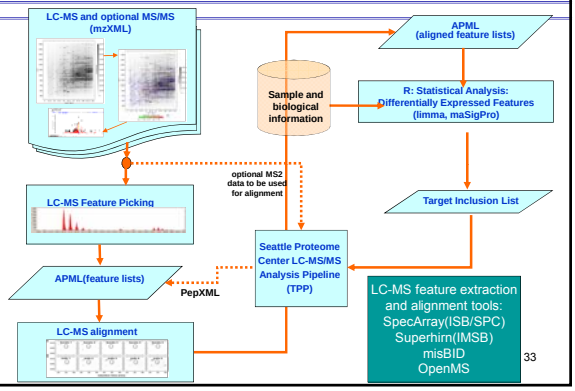
Corra: A Discovery-Based Approach for Generating Biomarker Candidate



Discriminatory peptide features initially determined just from MS1 data analysis
Follow-up targeted MS/MS subsequently determines peptide identity

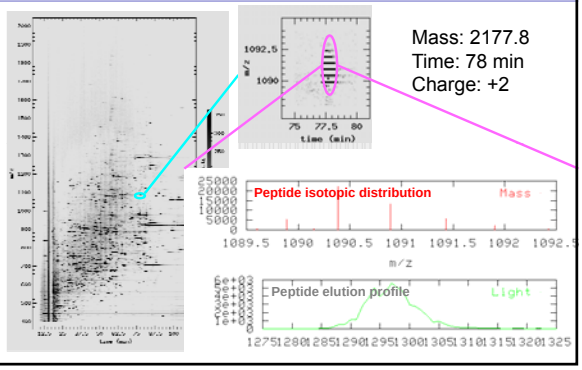
32

Corra Workflow

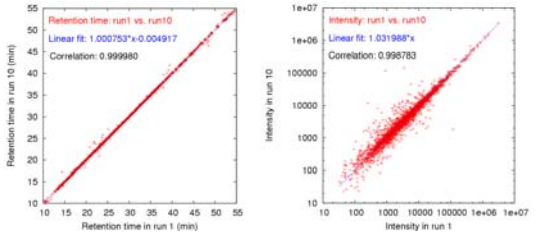


33

Peptide Feature Detection



Peptide Alignment



retention time intensity

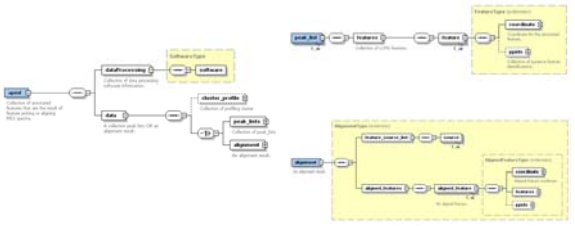
35

APML
(Annotated Putative Peptide Markup Language)

- APML is a MS1 data presentation of comparative protein expression profiling data.
- APML can facilitate interoperability of existing and new LC-MS and statistical tools.
- APML can help processed data management
- APML is used in Corra framework
- Current Corra computational tools (Superhirn, SpecArray, msInspect, msBID, OpenMS, CorraStatistics.R) have adapted APML
- APML parser library (org.systemsbiology.libs.apmlparser) has been implemented in Java 6 using both SAX and StAX parser for quantitative proteomics community to build easy plug-in modules.

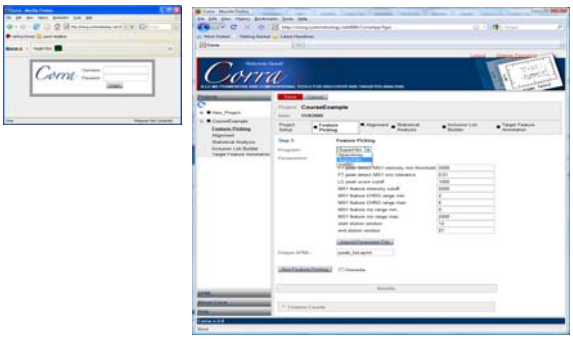
36

APML Schema
(Annotated Putative peptide Markup Language)



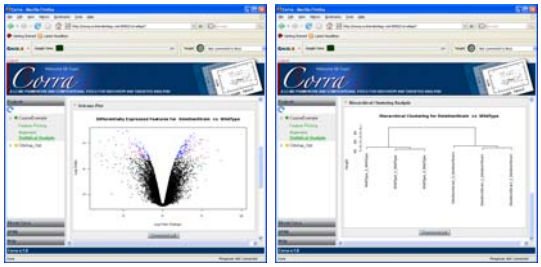
37

Corra: Graphical User Interface



38

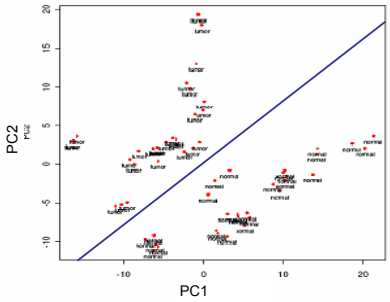
Corra: Graphical User Interface



Statistical Analysis GUI Panel

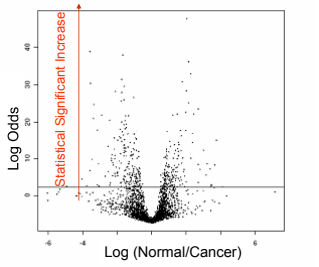
39

Corra: Application to Mouse Skin Cancer



40

Corra: Application to Mouse Skin Cancer



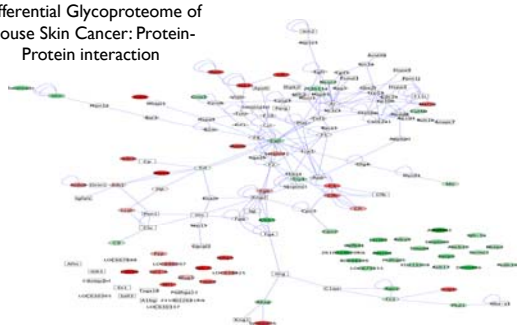
Limma analysis: Smyth, G. K. (2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." Statistical Applications in Genetics and Molecular Biology 3.

41

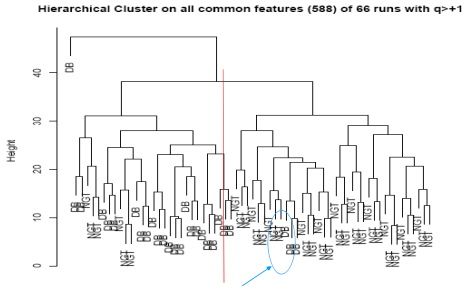
Targeted LC-MS/MS Analysis
(Inclusion List Setup)

Corra: Application to Mouse Skin Cancer

Differential Glycoproteome of Mouse Skin Cancer: Protein-Protein interaction



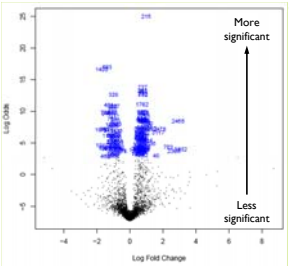
Corra: Application to Human Type 2 Diabetes



DB patient annotated as potentially not fasted prior to test

44

Corra: Application to Human Type 2 Diabetes



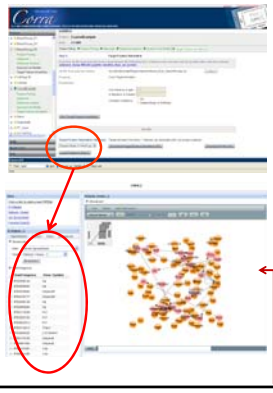
*Overall, we observed an average of 2130 features per run for all 105 runs (low of 1427 to a high of 2644)

Volcano plot shows the 4240 features that aligned across at least three of the 66 LC-MS runs (NGT and DB only)

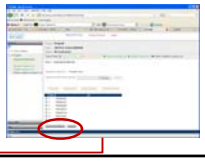
Outline of Discussion

- Day 1: Discussion of Discovery Tools for Biomarkers
  - Label Free Quantification
  - Introduction to Corra and Hands on Demo Using Tutorial

PIPE
(Protein Information and Property Exploration)



- **PIPE2** is a free web service for protein information and property exploration (PIPE) in context of biological network.
- **PIPE2** is designed to be friendly to both new informatics tool developers and data mining bioinformaticians
- **PIPE2** is by providing various standard operation such as GO enrichment and HPRD network visualization built in to be used in SRM based biomarker candidate protein generation steps.
- **PIPE2** is also integrated to firegoose to accept and broadcasting to firegoose-ready applications such as Corra and ATAQS.



Corra v2.0
User's Guide

INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

Original tutorial with easy to follow screenshot can be downloaded from http://tools.proteomecenter.org/Corra/corra.html

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics

Mi-Youn Brusniak[1], Bernd Bodenmiller[2,3], David Campbell[1], Kelly Cooke[1], James Eddes[1], Andrew Garbutt[1], Hollis Lau[1], Simon Letarte[1], Lukas N Mueller[2,3], Vagisha Sharma[1], Olga Vitek[4], Ning Zhang[1], Ruedi Aebersold[1,2,3,5] and Julian D Watts*[1]

Address: [1]Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103, USA, [2]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland, [3]Competence Center for Systems Physiology and Metabolic Disease, ETH Zurich, Zurich, Switzerland, [4]Department of Statistics and Department of Computer Science, Purdue University, West Lafayette, IN, USA and [5]Faculty of Science, University of Zurich, Zurich, Switzerland

Email: Mi-Youn Brusniak - mbrusnia@systemsbiology.org; Bernd Bodenmiller - bodenmiller@imsb.biol.ethz.ch; David Campbell - dcampbel@systemsbiology.org; Kelly Cooke - kcooke@systemsbiology.org; James Eddes - eddes@imsb.biol.ethz.ch; Andrew Garbutt - andga@microsoft.com; Hollis Lau - hlau@amgen.com; Simon Letarte - sletarte@systemsbiology.org; Lukas N Mueller - Lukas.Mueller@imsb.biol.ethz.ch; Vagisha Sharma - vsharma@u.washington.edu; Olga Vitek - ovitek@stat.purdue.edu; Ning Zhang - nzhang@u.washington.edu; Ruedi Aebersold - rudolf.aebersold@imsb.biol.ethz.ch; Julian D Watts* - jwatts@systemsbiology.org

* Corresponding author

Published: 16 December 2008

BMC Bioinformatics 2008, 9:542 doi:10.1186/1471-2105-9-542

Received: 20 May 2008
Accepted: 16 December 2008

Corra is an open source software Licensed under the Apache License, Version 2.0 and it's source code , demo data and this guide can be downloaded at the http://tools.proteomecenter.org/Corra/corra.html.

This user guide is written by Micheleen Harris (mharris@systemsbiology.org) and Mi-Youn Brusniak (mbrusniak@systemsbiology.org)

1. Introduction

Corra is a single, user-friendly, informatic framework, that is simple to use and fully customizable, for the enabling of LC-MS-based quantitative proteomic workflows of any size, able to guide the user seamlessly from MS data generation, through data processing, visualization, and statistical analysis steps, to the identification of differentially abundant or expressed candidate features for prioritized targeted identification by subsequent MS/MS. In the first published version of Corra software with the paper was v 1.5 in 2008 and since then, there were more update in the pipeline. In detail, Corra v1.5 pipeline ended by generating target list from statistical analysis. Corra v2.0 added additional feature extracting alignment tools as well as customized target list generation and annotation step using target LS-MS run. This guide uses the yeast gene knock out example used in Corra paper to illustrate the step of using v2.0 extended pipeline steps.

1.1 Login

Website: Ask administer in your institution which server the Corra is deployed to and ask Corra admin to add your account. For this guide, we will use guest account. The URL should be something like the following. http://corrademo.systemsbiology.net

1.2 Once logged in, click “New” to create a new project and give it a name (here it is “CourseExample”).

1.3 Choose the instrument type under drop down menu “MS Instrument”

1.4 Adding Data. Your data must be in mzXML format (if not, there are several converters from RAW data to mzXML, such as ReAdW and mzWiff). Click “Add” next to mzXML files to add mzXML formatted data to the project (required before you save the project). Select the mzXML files from the drop-down menu with which you want to run Corra (you can hold down the Shift key to select a group of files). Then press “Save” and reopen Project Setup by pressing “Edit.”

1.5 Defining Conditions, Sample IDs, Replicates and Time points. Make sure you have clicked “Edit” to continue setting up the project.

Click on “Condition_1” or “Condition_2” to rename these labels. If you wish to add any more conditions, click “Add.”

Check the files to label and use the drop-down menu to select the condition label appropriate for this group of files.

Each group of files (e.g. replicates belonging to a particular biological group) should share the same “Sample ID.” Assign a numerical ID by clicking on a number in the “Sample ID” column as shown left side.

Define replicates using the drop-down menu in the “MS Replicates” column. If you have more than 3 replicates increase the replicate count by clicking on the number next to “Max. Replicate Count.”

If you have more than one defined time point, add it by clicking the “Add” button next to “Time1” and rename by replacing “t_1” or “t_2” etc. Then specify them in the “Time Point” column using the drop-down menu.

Don’t forget to **Save your work**!

Alternatively, you could setup the project by importing a “Sample Information File.” This is useful if, say, you have a similar project with many mzXML files, as entering all of the setup information by hand could be a rather long process or you can use “Copy” project option which will create a new project with current project setup page. This “Copy” option can be used to analyze data using alternative Corra pipeline options.

2. Feature Picking

2.1 Click on “Feature Picking”, then “Edit”

2.2 Program for feature picking

Select the desired program (e.g. SpecArray is used for TOF-MS data and SuperHirn /msBID for FT-MS/Orbitrap)

For the Feature Picking step it might be useful to view the mzxml file(s) data. Using Pep3D, a .png (image) file can be created and viewed in a generic graphic viewer. For example, using the SuperHirn program, the elution window is set by default to begin at 12 and end at 87 minutes. Viewing the mzxml file in a program like Pep3D can help you decide if you wish to exclude (or include) parts of the experiment based on how the elution profile looks (Pep3D is a viewer of LC-MS or LC-MS/MS data in a general 2D “gel-like” format).

2.3 Set “Parameters” or import a parameter file

2.4 Click “Run Feature Picking”

> **Note**: Text in **yellow** indicates a process that is currently running and text in **green** is a process which has completed successfully. Text in **red** indicates an error has occurred and Corra log files may be referenced for further information.

When Feature Picking is done, you can scroll down to view the resulting feature counts for each input file.

> Note: The “FT peak detect MS1 intensity min threshold” could be increased in the case where there are too many features and/or you desire the subsequent runs to be faster (adjust the parameter and rerun Feature Picking).

Here is a picture of the result of the feature counting:

Note: These pictures can be downloaded as a .pdf file through the link below this graph.

3. Alignment

3.1 Click on the next step, "Alignment." Click "Edit" to setup the Alignment parameters.

3.2 Select a program (this should correspond to the program selected during Feature Picking)

3.3 Parameters

Adjust parameters to meet the specifications of your analysis and then click "Run Alignment."

Note: It might be a good idea to start with a value of 5 for the MS1 retention time tolerance.

3.4 Alignment Results

An APML (Annotated Putative Peptide Markup Language) file is created and maybe downloaded (by clicking the APML link in red) and viewed in an APML viewer comes with Corra (http://sourceforge.net/projects/corra/files/Corra-APML/APMLv2.0.1/APMLv2.0.1.tgz/download). This will help the user to view in a graphical way, the amount of aligned features. See next section for details about the APML Viewer.

3.4.1 APML Viewer

Open an .apml file in the viewer to see the aligned features in a *m/z* vs. Tr plot.
Try this: In the "Plotting Tool Bar", go to "Selected Plot View" -> "Times Aligned View" and click on a point in the graph to get a dialog box which shows the aligned features for that point (in this case there are three features aligning):

4. Statistical Analysis

4.1 Setup Statistical Analysis

Click on "Statistical Analysis" and "Edit." The program in use is a collection of R modules called CorraStatistics.R. Set the "B-Statistics Cutoff" (B = -[log odds ratio]) or use the default of 2.2. Here we change it to 0. Usually, having a "N/A Replace Method" of "none" is satisfactory. The "N/A Replace Method" is to be used if you wish to fill in missing features with a value, either a minimum value or user-defined value. Use the drop-down menu to select the type of N/A Replace Method to use.

Select the comparisons to be calculated (red circle on figure below). Save the setup and the Statistics step will begin.

4.2 Results

The result of the statistical analysis is displayed in two plots, a volcano plot and hierarchical cluster (unsupervised) as shown below. The red dots are features which are found in some of the samples, but not all and have a Log Odds ratio greater than the value set in the "B-Statistics Cutoff" field (I set it to 0, here, but the default is 2.2). The blue dots represent features that were found in all of the samples with

a Log Odds greater than the default or user defined limit. A Log Fold Change which is negative, indicates that a feature is more abundant in Condition 2 (here, WildType) whereas a positive Log Fold Change indicates that the feature is more abundant in Condition 1 (here, DeletionStrain).

A tab delimited file (.tsv) is created for the aligned features and can be downloaded by clicking "Differentially Expressed Feature List (.tsv)" link below the Volcano Plot. The data contained in the .tsv file comes from an analysis using CorraStatistics.R (Bioconductor) as a backend. It is used as input in the following step, "Inclusion List Builder."

The "Differentially Expressed Feature List (.tsv)" file is shown below, opened in MS Excel. Note in addition to the data there is statistical information such as logFC (log fold change), p-value and B value.

5. Inclusion List Builder

Go to the "Inclusion List Builder" section. (Note: Inclusion List Builder depends upon the Statistical Analysis step which must be completed successfully).

Click "Create Inclusion List" and give it a name.

Click "F" to add a filter and "Add Filter." Select a type of filter by clicking on the "+" sign (circled in blue in next figure) and using the drop down menu. To delete this filter click on the "-" sign (circled in orange).

A "#LC aligned" filter might be useful if you wish to focus on the number of features aligned across all samples a certain amount of times.

A "Mean Intensity" filter might be useful when features with low intensities wish to be excluded (Note: mean intensity is the $\log_2$ Intensity of a peak).

Hit "Save Inclusion List" and it will show how many total features and filtered features there are, plus some information such as min and max m/z (filters may be applied to limit these as well). Press "F" again to close the filter menu.

Click the red "S" to modify the segment settings. Segments can be useful when using the Thermofinnagin machine as these can be programmed into a target run. They can allow the machine to focus on certain parts of the run and not focus on others.

Segment Length is the "window" so to speak (minutes). The segment overlap is how many minutes one wishes to expand the window before and after the segment.

"First Segment Start" is usually just zero, but "First Segment End" is important as this first segment might capture parts of the run (usually at the beginning) where nothing very informative is happening. The "Min. Features per Segment" and "Max. Features per Segment" might be useful to play around with if there are too many features or too few.

Click "Create Segments."

To view the result of segmentation click on “View Segment Summary” and something like this should be displayed:

In this case I only had 100 features to begin with so there are very few features in my segments so I might try to increase my segment length.

You may save list inclusion list by clicking “Export to Excalibur.”

6. Target Feature Annotation

This module is to be used after MS/MS identification of peptide fragments to add sequence (and other protein descriptions) annotations to the original sample spectra, beginning the process of identifying proteins of interest. These could be the focus of future DDA or SRM analyses.

Target Feature Annotation annotates the statistical analysis output data (volcano plot data) based on the *m/z* values in a *.pep.xls* input file (provided by user).

6.1 Add an input file *somefile*.pep.xls

> Note: This input file can be created using a pepxml viewer to convert a *.pep.xml* file to *.pep.xls* (e.g. PepXML Viewer – part of the TPP, see tutorial at http://tools.proteomecenter.org/wiki/index.php?title=TPP_Tutorial#PepXML_Viewer).

This is a screenshot from PepXMLViewer (uses PeptideProphet analysis):

The input interact.pep.xls file should have **at least** these headers (but you may add more, like the index and spectrum for instance):

assumed_charge MZratio peptide retention_time_sec protein

In order to run “Target Feature Annotation” you must Add an .xls (e.g., interact.pep.xls) file which has all of the possible annotations that may be queried and added to your data (volcano.tsv file actually). See next figure for adding a xls file.

You may adjust the “m/z tolerance in ppm” which is set at a default to 25 ppm. Also, you may wish to adjust the” rt tolerance” in minutes. These are worth playing with if you do not get very many features annotated.

Then, hit “Run Target Features Annotation.”

6.2 Results of Target Feature Annotation

Once Target Feature Annotation has run, you will have an annotated volcano.tsv file from section 4.

At this point, you may click “Download Target Feature Annotation TSV,” (circled in red) a file which looks similar to the “Differentially Expressed Feature List (.tsv)” from section 4, but with additional information including peptide descriptions for some of the features. Below, the annotated *.tsv* file is shown opened in MS Excel.

You may also download the “IPI file TSV” which just contains just the features which are associated with IPI(s).

6.3 PIPE2

You may load the results of Targeted Feature Annotation into PIPE2 by clicking “Load Proteins to Pipe2.” Alternatively, you can copy and paste your IPIs, ORFs or other feature identifiers into PIPE2 to map them to several other databases, providing additional information about these important features. The PIPE2 link is here:

http://db.systemsbiology.net:8070/PIPE2/

Note: you must have firegoose extension installed in your computer when using Mozilla Firefox browser (http://gaggle.systemsbiology.org/docs/geese/firegoose/install/).

Here is a screenshot of our yeast DeletionStrain and Wildtype data into PIPE2 after Target Feature Analysis (remember we found 52 aligned and annotated features).

This links to a tutorial of PIPE2:

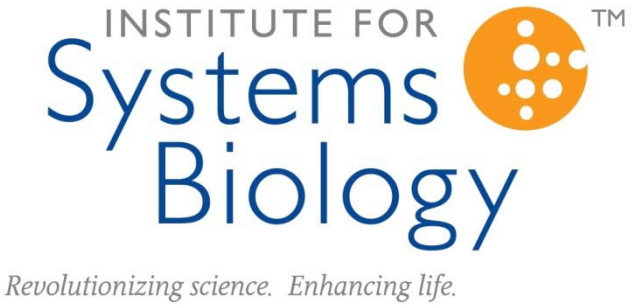
http://db.systemsbiology.net:8070/PIPE2/PIPE2/docs/PIPE2_tutorial.doc

SBEAMS, PeptideAtlas & SRMAtlas:
Database resources for Proteomics

Eric Deutsch

Day 5

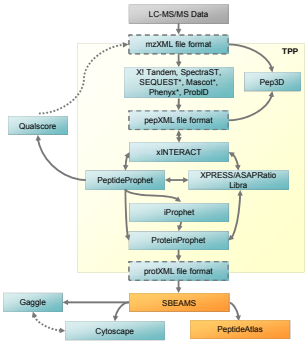
September 29, 2010



Outline

Topics

- Introduction to SBEAMS and SBEAMS-Proteomics
- PeptideAtlas: Compendium of peptides and proteins observed by MS/MS
- SRMAtlas: Enabling targeted proteomics experiments
- Tutorial and Exercises



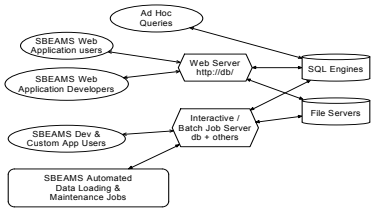
* Commercial software not part of TPP

A Need for Custom Database Front End Software

- Many databases for many Data Types
  - Many data types are generated and used at ISB, even as part of one project (Proteomics, Microarray, Genotyping, Immunostain, Interactions, ...)
  - Relational Database needed to keep track of it all
  - One grand unified database tough
  - Allow different databases to evolve under a common system using common software, database engines, interface
  - Integration of different data types relatively easy to integrate in this model
- Data Accessibility
  - Making data available to all levels of users
  - Reasonably simple web interface for data entry and queries
  - Client platform independence
  - UNIX command line interface for maintenance jobs and complex data mining
  - Remote data access via HTTP for scripting and automation
  - Relational back-end, flavor independent

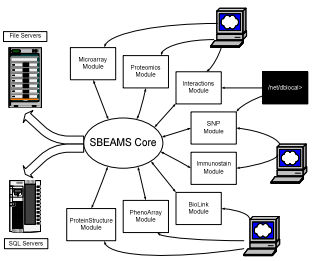
SBEAMS
Systems Biology Experiment Analysis Management System

- A framework for writing software for collecting, storing, accessing, and integrating data produced by various experiments using a relational database
- Tools for creating web front end entering data, queries, triggering batch jobs
- Programming interface for maintenance jobs, data loading and retrieval scripts, and interactive applications

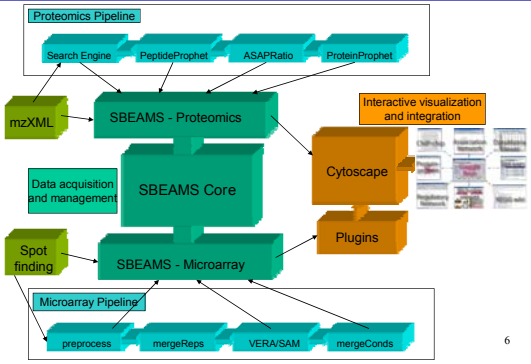


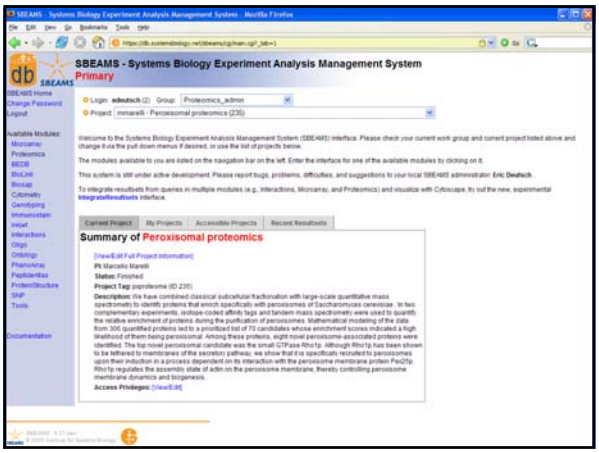
SBEAMS
Systems Biology Experiment Analysis Management System

- SBEAMS is designed as a core set of functionality around which individual modules can be built
- Each experiment or data type can have its own module
- Simultaneous "live" and development environments allow continual development
- Web interface accessible from any platform with a browser / no client installation
- Also UNIX command-line client and scriptable HTTP client API



SBEAMS
Integration of data acquisition, management, and analysis tools



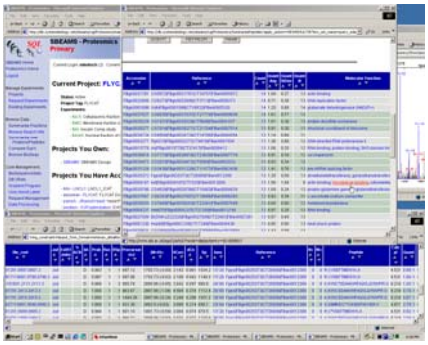


PeptideAtlas: Background

- There are many shotgun proteomic datasets of which only a small part of the information potential has been used
  - Only a limited set of proteins were of interest
  - Analysis software is still far from optimal
  - Experiment did not properly address hypothesis and is unpublished
- What further benefit can be extracted from a large group of heterogeneous experiments?

8

SBEAMS – Proteomics
Proteomics Analysis Database

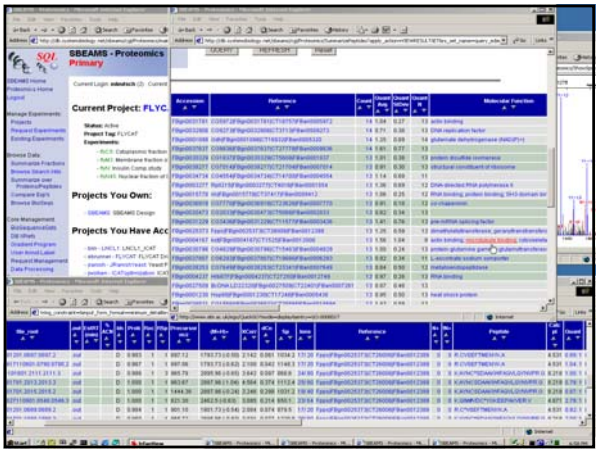


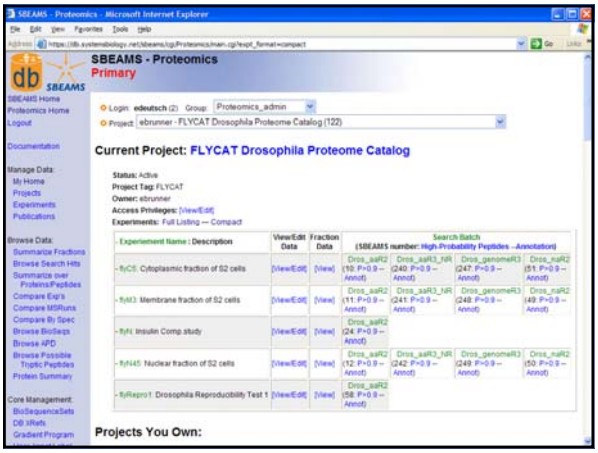
9

SBEAMS – Proteomics: Goals

- Organize many projects/experiments/searches into relational database schema
- Tools to explore the search results similar to existing ways plus lots of new ways including comparison of multiple experiments
- Allow users to store annotations of search hits after personal validation
- Allow queries across multiple experiments to capitalize on previous annotations
- Designed for ISB high-throughput 2DLC mass-spec Proteomics experiments
- Manage data collection and analysis pipeline
- Annotated Peptide Database (PeptideAtlas): library of observed peptides including properties and conditions under which they were seen
- Provide a platform for further software development and analysis
- Integration with other modules/databases (e.g., Microarray)
- Data visualization with Cytoscape

10





SBEAMS – Proteomics

PRO

- Central repository of organized data
- Annotate results and capitalize on annotations of others
- Queries to compare/combine experiments
- Queries to search for the needle in the haystack
- Write your own queries if you know/learn SQL
- Cytoscape integration

CON

- Not a robust, streamlined system
- Needs lots of work
- No full-time support
- In some ways, harder to "do your own thing"
- Beware the resultset that is different from what you thought you asked for

13

SBEAMS – Proteomics
Accessing the System (ISB internal site)

- http://db.systemsbiology.net/ and click on SBEAMS (SSL)

  or just
- https://db.systemsbiology.net/sbeams/
- Access is via SSL from outside the firewall
- Log on with your ISB username and either Windows or UNIX password or else a special account needs to be set up for you
- Test Drive at http://www.sbeams.org/sbeams/

14

How Can I Use It If Not at ISB?
Installing SBEAMS at another site

Requirements:

- SBEAMS Application Server
  - Perl 5.6+ required
  - Web server required
  - Developed under Linux + Apache
  - Anecdotes of running it on Windows exist but not yet at ISB
- RDBMS (separate machine recommended but not required)
  - Developed on SQL Server
  - SBEAMS Core tested on MySQL and PostgreSQL
  - Proteomics module has not. Could be ported with some effort
- Database programmer
  - Effort in getting it installed at your site should not be underestimated
  - And it will required on going management and development

Download It and Install It:

- http://www.sbeams.org/

15

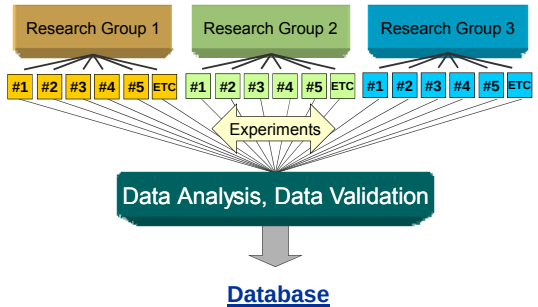
Outline

Topics

- Introduction to SBEAMS and SBEAMS-Proteomics
- PeptideAtlas: Compendium of peptides and proteins observed by MS/MS
- SRMAtlas: Enabling targeted proteomics experiments
- Tutorial and Exercises

16

PeptideAtlas
Combining Many Heterogeneous Experiments



17

PeptideAtlas
What Is It?

- PeptideAtlas is the integration of a large number of uniformly processed tandem mass spec experimental results into a master list of observed peptides mapped to the genome
- Currently includes ~1000 experiments from:
  - Aebersold lab
  - ISB Proteomics Facility (including data from external clients)
  - NHLBI Consortium members (Yale: Williams, JHU: Pandey)
  - Data from the Open Proteomics Database (OPD@UT: Marcotte)
  - Other contributors (Reising, Gygi, Haynes, Hogue, Conrads..)
  - Collaborations
- ISB was well suited to start this because of the large amount of in-house data and the general lack of publicly available data, although now the data is flowing more freely from the community

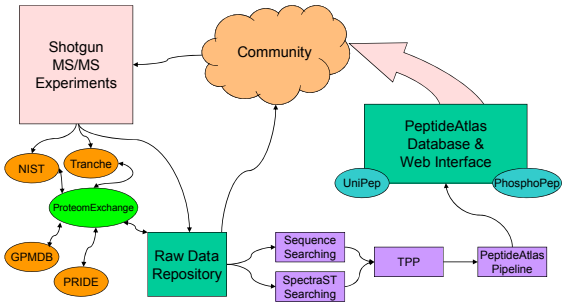
18

PeptideAtlas
Why?

- Genome Annotation:
  - Validating "predicted" proteins
  - Validating intron/exon boundaries and alternative splice forms
  - Validating the reference protein databases (e.g., we find many peptides that don't map to Ensembl)
- Experiment Planning:
  - Which proteins & peptides are observable with MS/MS
  - Targeted proteomics via inclusion lists
- Data Analysis Aid:
  - Use the web UI to examine whether a protein/peptide in your experiment is already in the PeptideAtlas, which samples, how often, etc.
  - Faster MS/MS analysis using spectrum libraries
- Data Mining:
  - Exploring which peptides are seen and which are not
  - Exploring MS/MS spectral patterns
- Defining the (MS/MS observable) Proteome

19

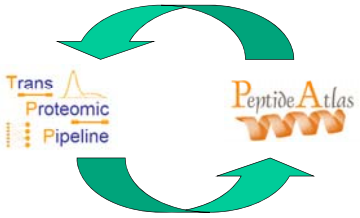
PeptideAtlas Workflow



20

TPP: Foundation for PeptideAtlas

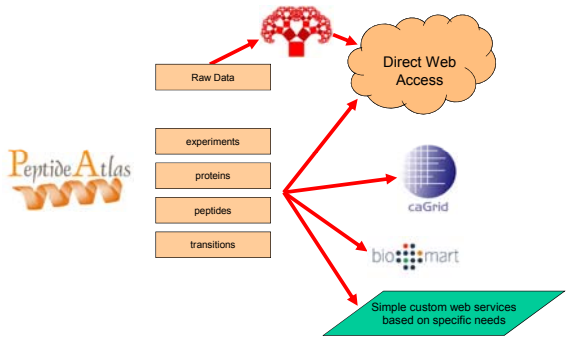
Drives tool development and optimization



Advanced, uniform processing of all data

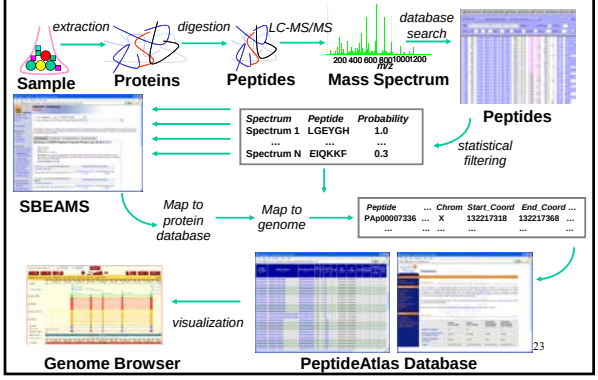
21

PeptideAtlas Web Services



22

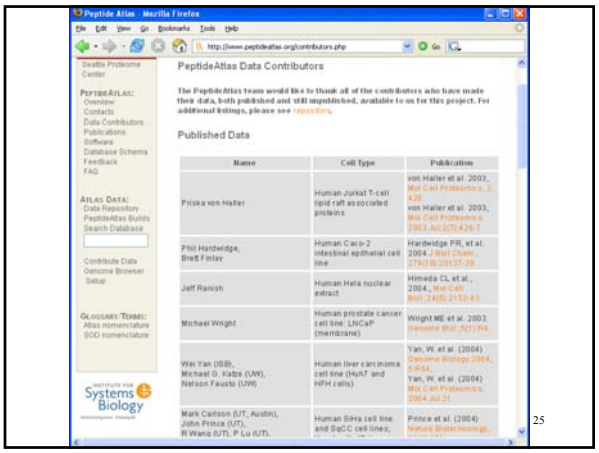
From Peptides to Genome Annotation



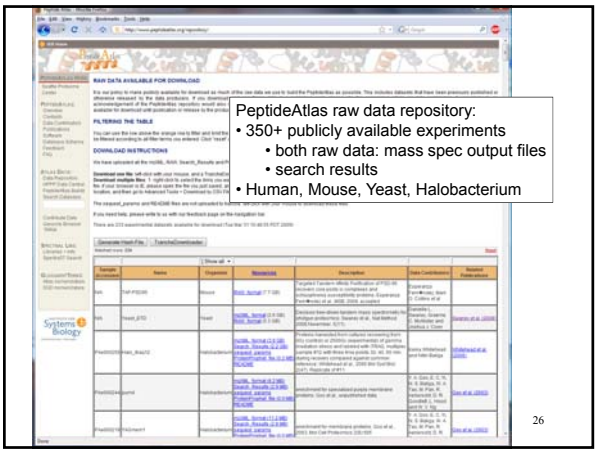
23

http://www.peptideatlas.org/

24


PeptideAtlas Data Contributors
Published Data

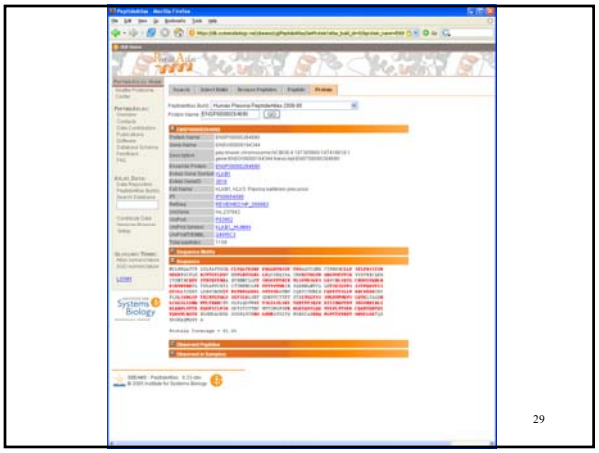
25


PeptideAtlas raw data repository:
• 350+ publicly available experiments
• both raw data: mass spec output files
• search results
• Human, Mouse, Yeast, Halobacterium

26


Build results available for download
• As flat files
• XML
• MySQL dump

28



29

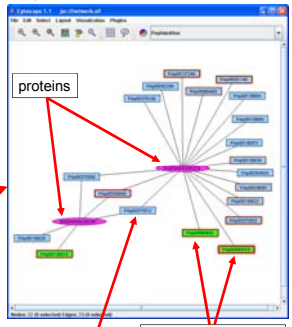
30

31

32

PeptideAtlas protein view page

Cytoscape view of proteins & peptides



ambiguously mapped peptide

proteotypic peptides
$N_{prot} = 1$ $N_{obs} > 3$
EPS > 0.3

What is a proteotypic peptide?

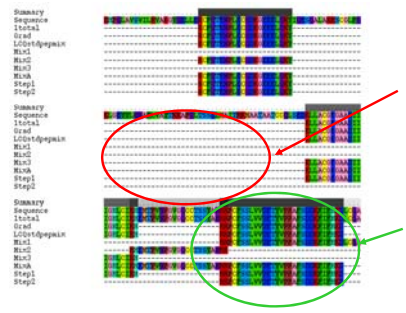
A proteotypic peptide:

1. Is likely to be observed with today's mass spectrometry technology in complex samples
2. Maps uniquely to a single protein (or isoform)

34

What is a proteotypic peptide?

Protein



Peptide region never observed in multiple experiments

Peptide region consistently observed in multiple experiments

35

Observability score

A measure of how likely it is that a peptide will be observed in an experiment (of a certain type)

$$\text{Observability Score} = \frac{N_{samples}(\text{peptide})}{N_{samples}(\text{parent protein})}$$

- *e.g.* if peptide X is observed in 5 samples, and its parent protein A is observed in 10 samples, score = 0.5

36

Why are proteotypic peptides useful?

- Shotgun MS/MS workflows have been very successful for discovery experiments
- But for quantitative proteomics, shotgun workflows leave holes
- For proteomics to be a rich tool for system biology requires a new workflow: Targeted Proteomics
- Proteotypic peptides are best used for creating target lists such as inclusion lists and SRM transitions
- Ideal peptides to be used as heavy isotope standards for quantification

37

Empirical Proteotypic Peptides

Best Peptides

Peptide Accession	Pre AA	Peptide Sequence	Fol AA	Suitability Score	Best Prob	Best Adjusted Prob	N Obs	Empirical Proteotypic Score	SSRCalc Relative Hydrophob
PAp00038844	K	VPLQQNFQDNQFQGK	W	1.10	1.000		92	0.40	26.72
PAp00038127	R	TFVPGCQPGEFTLGNIK	S	0.83	1.000		63	0.20	34.90
PAp00027993	K	SYNVTSVLFR	K	0.80	1.000		9	0.40	29.98
PAp00037688	K	SLGLPENHIVFPVPIDQCIDG	-	0.74	1.000		40	0.20	43.97
PAp00139942	K	MYATIYELK	E	0.74	1.000		16	0.30	29.31
PAp00139411	K	SYPGLTSYLVR	V	0.66	1.000		16	0.20	29.92
PAp00139413	K	WYVVGLAGNAILR	E	0.66	1.000		16	0.20	40.76
PAp00139412	R	VVSTNYNQHAMVFFK	K	0.64	1.000		10	0.20	32.32
PAp00024943	K	EDKSYNVTSVLFR	K	0.41	0.997		1	0.10	28.23
PAp00033583	K	ELTSELKENFIR	F	0.41	0.996		1	0.10	29.19
PAp00378369	K	CDYWIR	T	0.38	0.985		3	0.10	23.04
PAp00038222	R	TKELTSELKENFIR	F	0.34	0.976		1	0.10	29.62
PAp00139941	K	ITLYGR	T	0.15	0.919		1	0.10	16.15
PAp00036688	I	PAPPLSKVPLQQNFQDNQFQGK	W	0.04	0.933		1	0.10	33.34

38

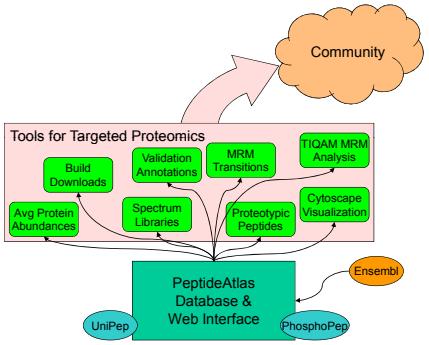
Predicted Proteotypic Peptides

Predicted Highly Observable Peptides

Peptide Accession	Pre AA	Peptide Sequence	Fol AA	Suitability Score	Detectability Predictor Score	PeptideSieve Score
	K	SSGWGNSGGSNNSSWW	-	0.87	0.77	0.97
PAp00015022	R	HVEDCDMTPVGER	Q	0.84	0.72	0.97
PAp00051450	R	AGNTGLATAFFNSENSNIVK	G	0.81	0.74	0.88
	R	GGYGNGGFFGGNNGGSR	S	0.80	0.75	0.84
PAp00111840	R	SGAATLLVATAVAAR	G	0.77	0.84	0.71
	R	NNSSNYNNNNGGYNGGR	G	0.77	0.69	0.84
PAp00014537	R	GCDLLVATPGR	L	0.74	0.53	0.95
PAp00052741	K	GLHEILTEANQEVPSFLK	D	0.71	0.66	0.77
PAp00001717	R	DLMACAQTGSGK	T	0.71	0.71	0.71
PAp00055176	K	TGGFLFPVLSESFK	T	0.71	0.76	0.65
	K	AGGASAGGWGSSR	S	0.60	0.63	0.57
PAp00052060	K	DVPEPITEFTSPPLDGLLLENIK	L	0.59	0.66	0.53
PAp00105916	K	AYPTAVIMAPTR	E	0.58	0.65	0.51
PAp00080212	R	VGSTSENITQK	V	0.54	0.75	0.32
PAp00046504	R	QTLMFSATFPADIQHLAR	D	0.52	0.57	0.48
	R	DFLSDYIFLSVGR	V	0.52	0.75	0.28
PAp00084399	K	TGPPPDPFSGQSFYQR	K	0.47	0.76	0.19

39

PeptideAtlas Access Tools

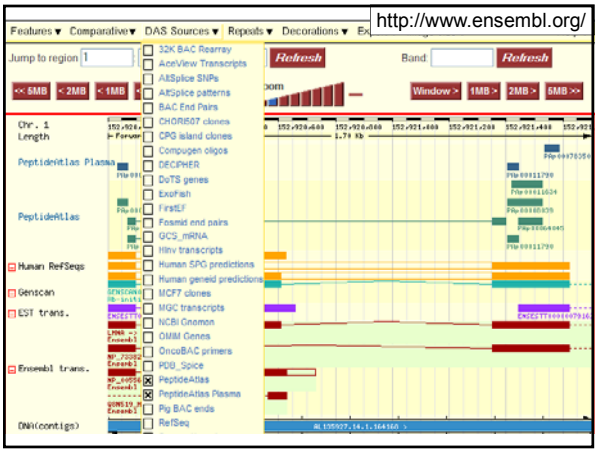


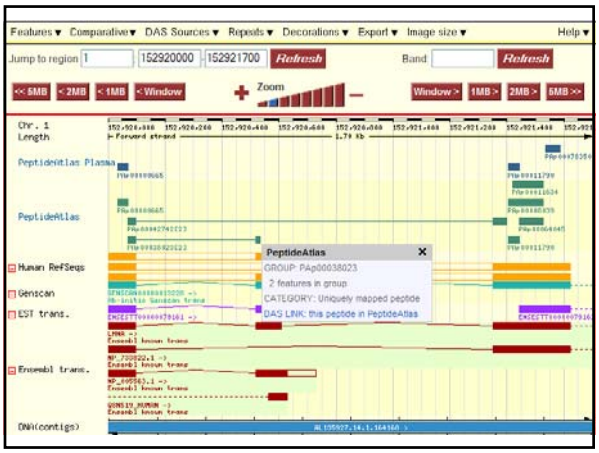
40

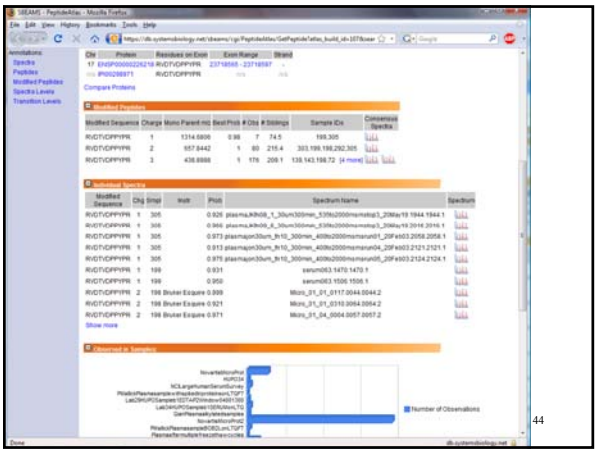
Annotations

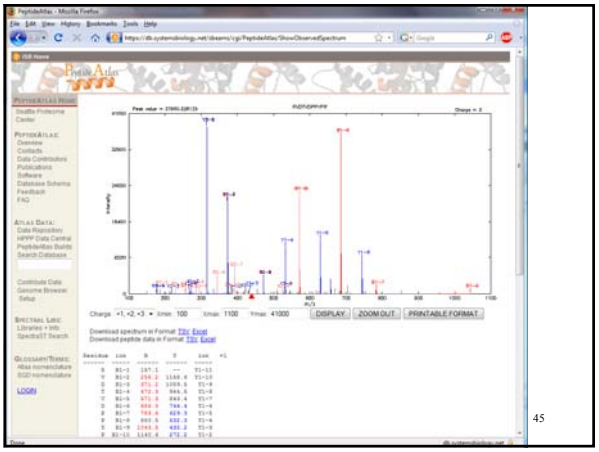
- Collect and display
  - Heavy peptides that have been synthesized and used
  - Transitions that have been validated and/or published

41









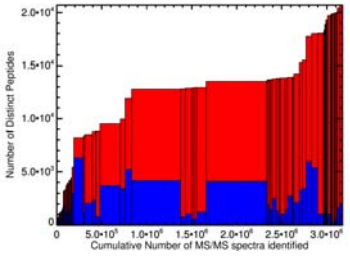
PeptideAtlas: Builds Summary

Build	# Exps	# MS Runs	Searched Spectra	IDs P>0.9	Distinct Peptides	Distinct Proteins
Human All	424	54 k	49 M	5.9 M	90 k	10355
Human Plasma	91	48 k	16 M	3.2 M	21 k	2563
Yeast	53	2957	6.5 M	1.1 M	36 k	4336
Mouse	59	3097	10 M	1.4 M	51 k	7686
Drosophila	43	1769	7.5 M	498 k	72 k	9124
Drosophila PhosphoPep	4	448	0.9 M	170 k	10 k	4583
Halobacterium	88	497	0.5 M	76 k	12 k	1518
S. pyogenes	5	64	215 k	52 k	7 k	1068
Mouse Plasma	4	568	7 M	0.8 M	9.4 k	2075
More…						

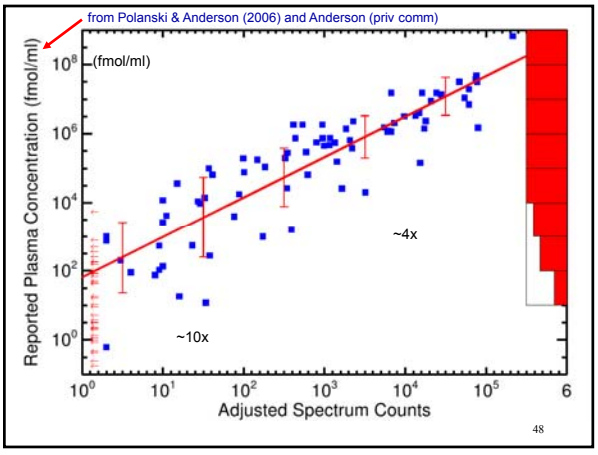
46

Human Plasma PeptideAtlas build 2009-12

- Public and released
- 91 experiments/samples
- PSM (peptide-spectrum match) FDR ~ 0.0001
- 3.2 M PSM IDs
- Distinct peptide FDR 0.0014
- 21k distinct peptides
- Final protein FDR 0.0094
- 1789 / 20333 canonical UniProt proteins
- 1975 / 20333 canonical + possibles UniProt
- 2563 canonical + possibles all

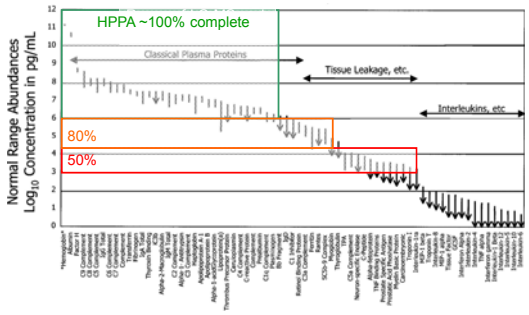


47



48

Plasma PeptideAtlas Range



(From Anderson & Anderson (2002) *MCP*, **1**, 845-67)

PeptideAtlas: Publications

- Initial PeptideAtlas Publication:
  Frank Desiere, Eric W. Deutsch, Alexey I. Nesvizhskii, Parag Mallick, Jimmy K. Eng, ...
  "Integration of Peptide Sequences Obtained by High-Throughput Mass Spectrometry with the Human Genome", Genome Biology 2004, 6:R9
- Human Plasma PeptideAtlas Publication:
  Eric W. Deutsch, Jimmy K. Eng, Hui Zhang, Nichole L. King, Alexey I. Nesvizhskii, ...
  "Human Plasma PeptideAtlas", Proteomics. 2005 Aug;5(13):3497-500
- Yeast PeptideAtlas:
  Nichole L. King, Eric W Deutsch, Jeff Ranish, Alexey I. Nesvizhskii, James S. Eddes, ...
  "Analysis of the S. cerevisiae proteome with PeptideAtlas", Genome Biology, 7, R106
- NAR Database Issue Update:
  Frank Desiere, Eric W. Deutsch, Nichole L. King, Alexey I. Nesvizhskii, Parag Mallick, Jimmy Eng, ...
  "The PeptideAtlas Project", Nucleic Acids Research, 2006, 34, D655-D658
- PeptideAtlas as a Resource for Targeted Proteomics:
  Eric W Deutsch, Henry Lam, & Ruedi Aebersold
  "A Resource for Target Selection for Emerging Targeted Proteomics Workflows", 2008, EMBO Reports, 9, 429
- Mouse Plasma PeptideAtlas:
  Zhang, Q., Menon, R, Deutsch, E. W., Pitteri, S. J., Faca, V. M., Wang, H., Newcomb, L., ...
  "A Mouse Plasma Peptide Atlas as a Resource for Disease Proteomics", 2008, Genome Biology, 9, 6, R93
- Halobacterium PeptideAtlas:
  Phu T. Van, Amy K. Schmid, Nichole L. King, Amardeep Kaur, Min Pan, Kenia Whitehead, ...
  "Halobacterium salinarum NRC-1 PeptideAtlas: Toward Strategies for Targeted Proteomics...", 2008, JPR, 7, 3755

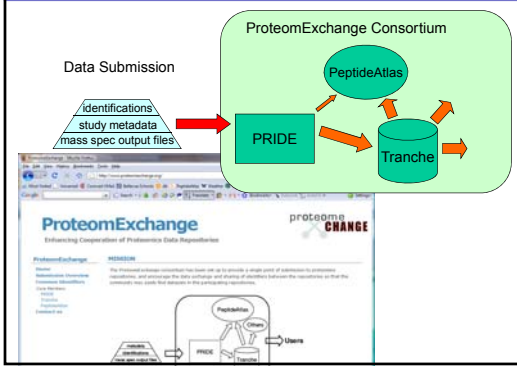
PeptideAtlas: How to use it

- Link to / paste in your (human, mouse, drosophila, yeast…) peptides or proteins of interest and see if they have been seen already and in what samples
- Download the PeptideAtlas build results and mine the data
- Contribute your data:
  - Published data. We'll put it up in the repository for others to download
  - Unpublished data. We'll include it in the PeptideAtlas with minimal annotation
  - Human or Mouse data of most interest right now
  - Data from other organisms. We'll take it, esp. Ensembl organisms
  - Preferably the raw files, we'll run it through the pipeline here
- Start your own PeptideAtlas for your favorite organism
  - We release all the tools to build your own PeptideAtlas for whatever you want to do

Proteomics Data Repositories

- PeptideAtlas (ISB)
  - Raw data submissions only
  - All data are reprocessed through search and TPP unless done elsewhere
  - Raw and processed data posted for easy download
  - Combined builds available for browsing and download
- PRIDE (EBI)
  - Peptide identifications only
  - Supporting MS/MS spectra okay but not required
  - No raw data. Pointers to Tranche
- Tranche (Proteomecommons.org – U Michigan)
  - Worldwide distributed file system
  - Can hold any type of file (proteomics data related only accepted)
  - Journal article data annotation and scrounging effort
- Peptidome (NCBI)
  - Just launched; Populated with some PeptideAtlas data
- OPD (U Texas)
  - Original repository; holds only proof-of-concept data from local lab

ProteomExchange Consortium

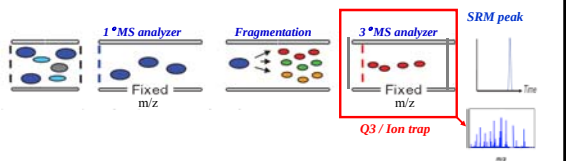


Outline

Topics

- Introduction to SBEAMS and SBEAMS-Proteomics
- PeptideAtlas: Compendium of peptides and proteins observed by MS/MS
- SRMAtlas: Enabling targeted proteomics experiments
- Tutorial and Exercises

Selected Reaction Monitoring (SRM)



- *two levels of mass selection: high specificity*
- *not scanning (Q1/Q3 static), high duty cycle: high sensitivity*
- *the most sensitive mass spectrometry method known (low amole)*

....you need to know what to look for!

(the mass spectrometrist's ELISA)

55

Targeted Proteomics

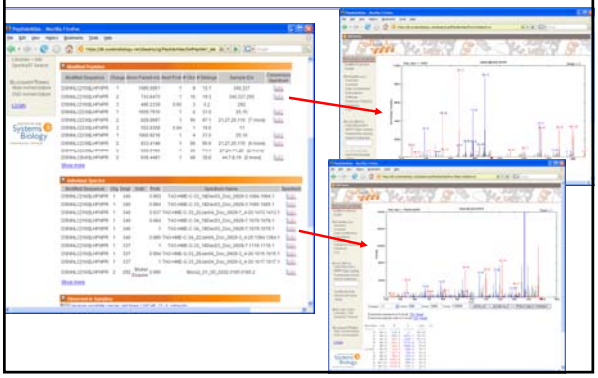
- Target selection
  - Target protein selection
  - Target peptide selection
  - Reference peptide selection
  - SRM transition selection
- PeptideAtlas: A Resource for Target Selection for Emerging Targeted Proteomics Workflows
- How?
  - Observed and predicted proteotypic peptides
  - Consensus and individual spectra
  - SRM transitions based on these spectra
  - Approximate abundance scales for proteins
  - Annotations from community contributed or published:
    - Synthetic peptides
    - Validated transitions

56

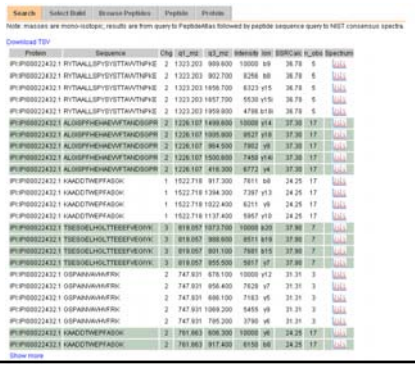
Transitions in PeptideAtlas

57

Consensus and Individual Spectra

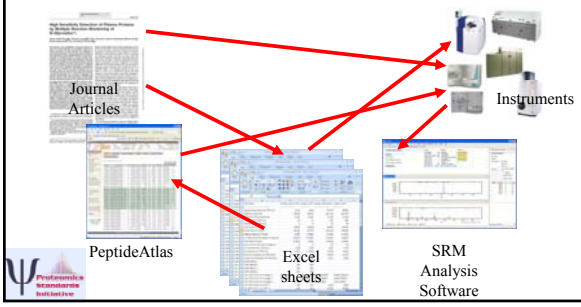


Export Transition Lists



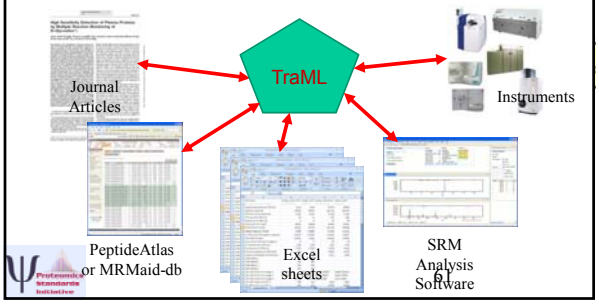
59

Many disparate formats for transitions



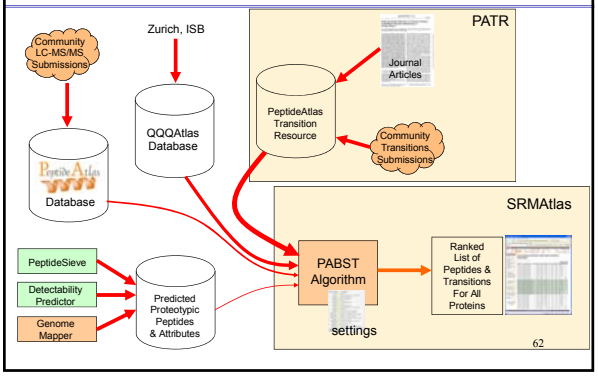
Unify exchange of transitions with TraML

- PSI's TraML (Transitions Markup Language) (similar to mzML et al.)
- Proposed format for encoding SRM/MRM transitions
- Version 0.9.4 currently under review

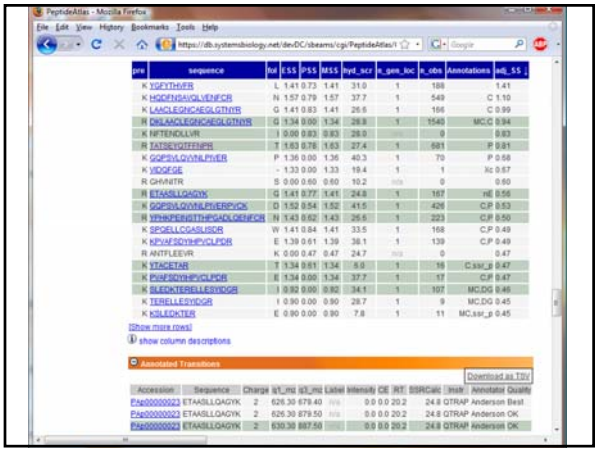


61

PeptideAtlas, PATR, SRMAtlas Overview



62

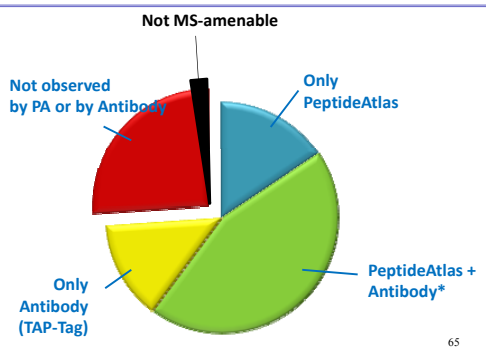


A complete yeast proteome SRM assay library

Picotti et al. Nature Methods 2007
Picotti et al. in preparation

64

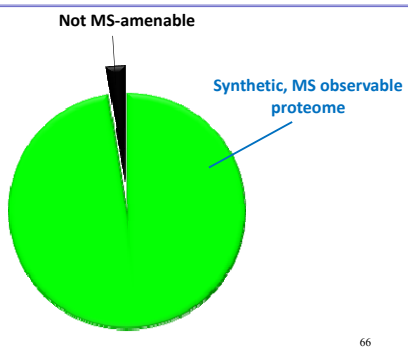
Yeast Proteome Coverage… before



65

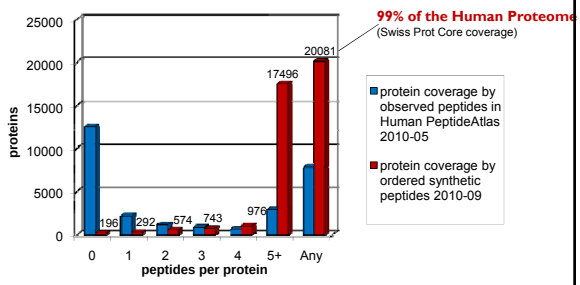
*Ghaemmaghami et al, *Nature 2003*

Yeast Proteome Coverage… after



66

Human Proteome SRM Assay Coverage



Equally, through orthologous peptide matching, 76% of the mouse proteome has SRM assays developed in SRMAtlas

Human Proteome SRM Coverage (subsets)

Single amino acid mutations

2726 Major SNP's (>30% frequency =1363) resulting in non-synonymous mutations (NCBI dbSNP Build 131)

Glycosylated proteins

5199 Membrane proteins with 11,269 N-glycosites

1748 secreted proteins

Bioinformatics construction of the human cell surfaceome. da Cunha JP, Galante PA, de Souza JE, de Souza RF, Carvalho PM, Ohara DT, Moura RP, Oba-Shinja SM, Marie SK, Silva WA Jr, Perez RO, Stransky B, Pieprzyk M, Moore J, Caballero O, Gama-Rodrigues J, Habr-Gama A, Kuo WP, Simpson AJ, Camargo AA, Old LJ, de Souza SJ. Proc Natl Acad Sci U S A. 2009 ;106(39):16752-7.

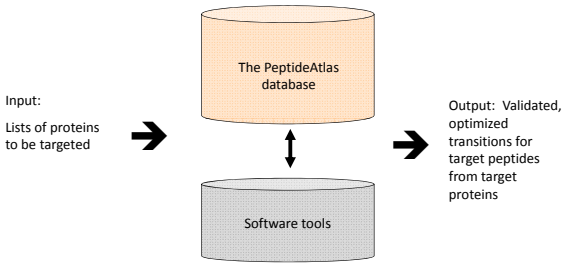
Prediction of the human membrane proteome. Fagerberg L, Jonasson K, von Heijne G, Uhlén M, Berglund L. Proteomics. 2010;10(6):1141-9.

Varsplice isoforms not present in SwissProt primary

11,309 additional proteotypic peptides for unique identification

68

The PeptideAtlas project: A resource supporting targeted proteomics



www.peptideatlas.org

Desiere et al. Genome Biology 2004. Deutsch et al et al. EMBO Reports, 2008

69

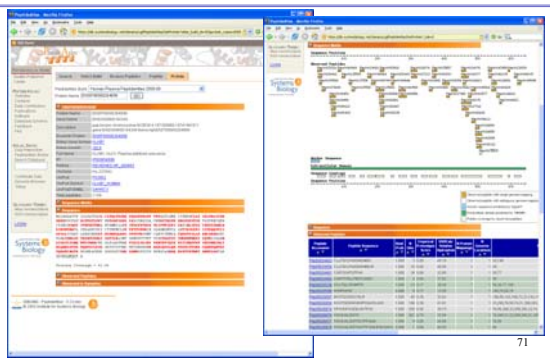
Outline

Topics

- Introduction to SBEAMS and SBEAMS-Proteomics
- PeptideAtlas: Compendium of peptides and proteins observed by MS/MS
- SRMAtlas: Enabling targeted proteomics experiments
- Tutorial and Exercises

70

PeptideAtlas: Tutorial



71

PeptideAtlas Tutorial

Please follow along in this tutorial as we go through it in the class. Feel free to add your own notes. If you make notes or have suggestions or bug reports that might be useful for others, please email them to edeutsch@systemsbiology.org.

1. Open a web browser and go to PeptideAtlas (recommend Firefox to use FireGoose plugin):
    - http://www.peptideatlas.org/
2. Explore the PeptideAtlas Raw Data Repository (link on left nav bar):
    - Select [Yeast] in the Organism drop-down list and see the 48 experiments
    - Click [Reset] link,
    - Type "liver" in the description field and press [Enter] and see the 18 experiments
3. Explore the PeptideAtlas Builds Download Area (link on left nav bar):
    - Examine the build summary table
    - Click [Download] under the C. elegans subsection
4. Explore the PeptideAtlas Build Summary:
    - Click on "Search Database" on left nav bar
    - Mouse-over the [All Builds] tab and choose [Select Build]
    - Click on "Yeast 2009-0522" PeptideAtlas build
    - Examine resulting page
5. Explore the PeptideAtlas Search interface ("Search Database" on left nav bar):
    - Type YDR502C into search box, set Build Type to [Any] and [GO]
    - Results are shown for main Yeast atlas and SRM atlas.
    - Click on the list for the main atlas ("Yeast") and examine the result
    - What is the protein coverage? (Answer A1 at bottom)
    - Find 2 peptides deemed highly proteotypic both empirically and predicted (A2)
    - Resort the observed peptides by "N Obs" descending. Note the new rank of IIVDAYGGASSVGGGAFSGK
    - How many other proteins and genome locations do the constituent peptides map to? (A3)
    - Click on Cytoscape link to visualize
    - Click on PAp00018124 (SLVAAGLCK) for the Peptide View
    - Click on [Compare Proteins] in Genome Mappings section
    - How do the proteins YLR180W and gi|172534 differ? (A4)
6. Explore the PeptideAtlas Search interface ("Search Database" on left nav bar) part 2:
    - Click on [Search] tab
    - Select Build Type: Yeast
    - Type in search box: %peroxisom% (leave off the e) and click GO
    - How many proteins match? And how many with at least 2 hits? (A5)
7. Explore the PeptideAtlas SRM Transitions interface:
    - Select [Query Transitions] under the [SRMAtlas] tab
    - Select Build Type: Yeast Public 2010-02
    - Protein name constraint: YAL00%
    - N fragment ions to keep: 3 and N peptides per protein: 3 and [QUERY]
    - How many proteins have transitions selected from real spectra (not just predictions?) (A6)

Answers:

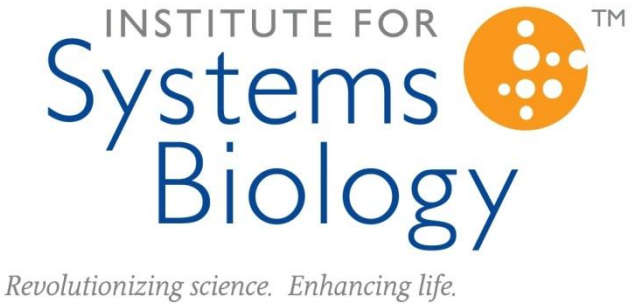
- A1: 92.9% (although 100% if one excludes regions unlikely to be observed)
- A2: TCNVLVAIEQQSPDIAQGLHYEK & ICDQVSDAILDACLEQDPFSK are ranked 1 & 3 observed and 7 & 6 predicted
- A3: Up to 5 proteins and up to 2 different genome locations
- A4: They differ by just two residues and both the YLR180W variants are seen (Hint: realign just these two proteins)
- A5: 57 matches, 43 of which have > 1 hits (sort increasing N Peptides and count)
- A6: 4 proteins: YAL003W, YAL005C, YAL007C, YAL008W

Using PIPE2 & Gaggle to Explore the Biological Significance of Protein Lists

Hector Ramos

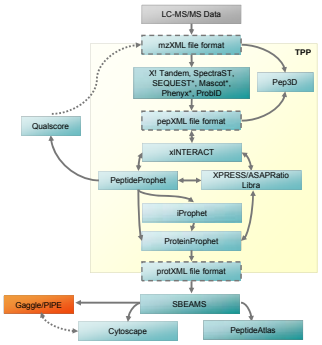
Day 5

October 29, 2010



Presentation Overview

- Motivation
- PIPE 2
- Firegoose
- Gaggle
- Tutorial



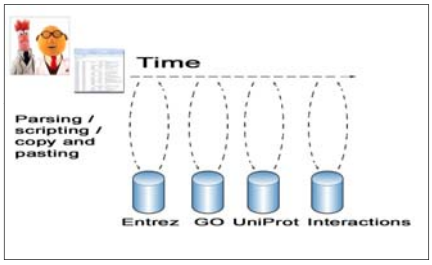
1

Motivation

- Now we have a list of identified proteins
- Perform some common/useful tasks to begin exploring biological significance
- Pull data from several different publicly available data sources to help make sense of your data
- Visualizing data
- **Advanced**: Use Gaggle tools to further explore

2

Why This Can Get Messy Quickly



3

A Few Specific Tasks

- Looking up bits of annotation data from various sources
  - Identifier mapping (from one type of ID to another)
  - Functional annotations
  - Interactions between proteins in your list
- Functional enrichment calculations
  - Gene Ontology Enrichment
- Visualizing
  - protein-protein interactions
  - functional associations
  - protein expression levels
  - set analysis (intersections, unions, etc.)

4

PIPE2 Software Goals

- To make these (and other) tasks as quick and easy for the user (assumed to be non-programmer) as possible
- To make it easy to move your data into the Gaggle software

5

PIPE2 Details

- Load your list of proteins into PIPE2
- From here you can:
  - Map protein identifiers to Entrez Gene IDs(IPI, Yeast ORFS)
  - Functionally annotate your IDs (GO, Uniprot)
  - Perform Gene Ontology enrichment on molecular functions, biological processes, and cellular component
  - Look up other proteins of similar function
  - Visualize HPRD (or Y2H) curated interactions in a network
  - Visualize functional associations in a network
  - Utilize powerful web resources and databases such as Entrez, KEGG, STRING, and DAVID

6

Workflow Philosophy of Gaggle and PIPE2

- There are already great software tools out there that do their jobs really well; we're not going to write new ones, we're just going to write the glue to stick them all together.
- Each of these tools then becomes a module that can be pieced together as the user sees fit.

7

e.g., in PIPE 2

- 5 Modules:
  - ID Mapper
  - GO Enrichment
  - Keyword Lookup
  - Network Visualizer
  - Venn Diagram

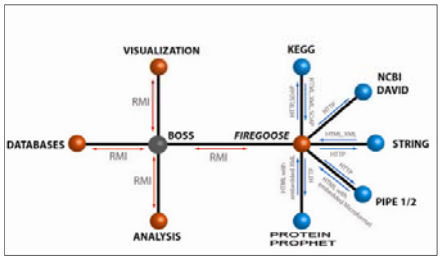
8

e.g., in Gaggle

- Several Modules:
  - DMV (Data Matrix Viewer)
  - Cytoscape
  - Matlab
  - MeV (Multiple Experiment Viewer)
  - R / Bioconductor
  - Firegoose
    - Connects several different web resources

9

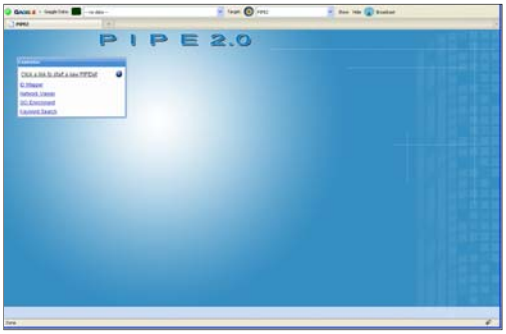
PIPE/Gaggle/Firegoose Big Picture



10

The Protein Information and Property Explorer 2 (PIPE2)

http://db.systemsbiology.net:8070/PIPE2/



11

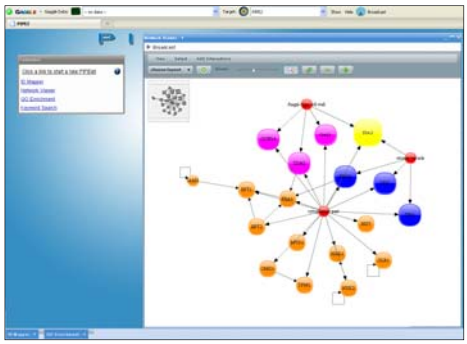
PIPE2 – ID Mapper PIPElet

12

PIPE2 – GO Enrichment PIPElet

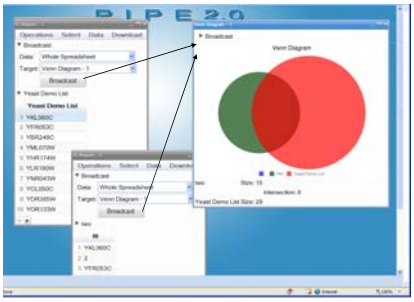
13

PIPE2 – Network Visualizer PIPElet



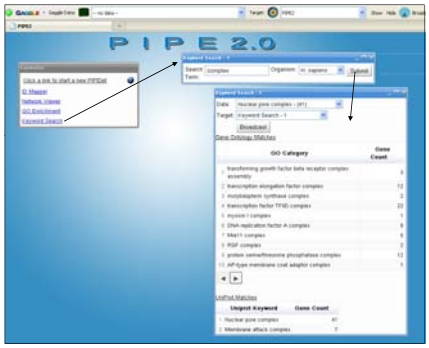
14

PIPE2 – Venn Diagram PIPElet



15

PIPE2 – Keyword Search PIPElet



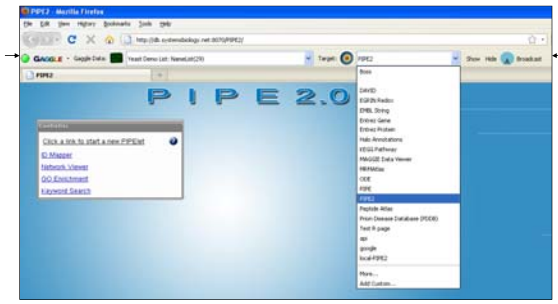
16

The Firegoose

- A toolbar for the Mozilla Firefox browser
- Exchanges data between Gaggle and the web
- It manages to do this through various different internet protocols
  - See Chris Bare if you're interested in details

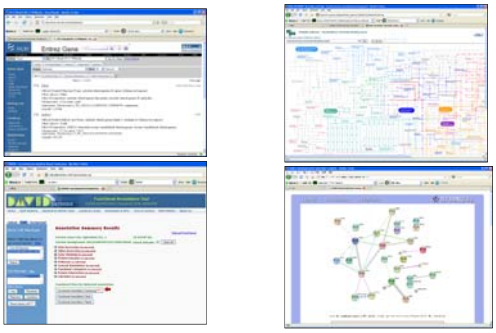
17

The Firegoose



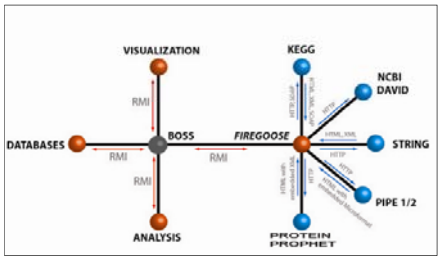
18

Some Firegoose Targets



19

PIPE/Gaggle/Firegoose Big Picture



20

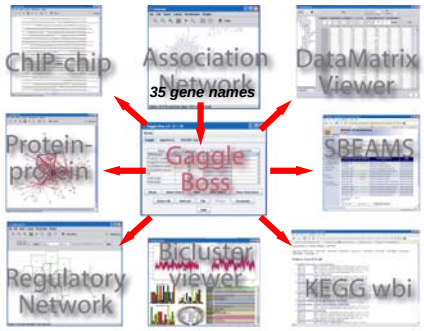
What is the Gaggle?

- A framework for exchange of data
- Via messaging
- Between independently developed tools
- Through a few simple data types.

- It's the GLUE sticking different tools together
- More info: http://gaggle.systemsbiology.net

21

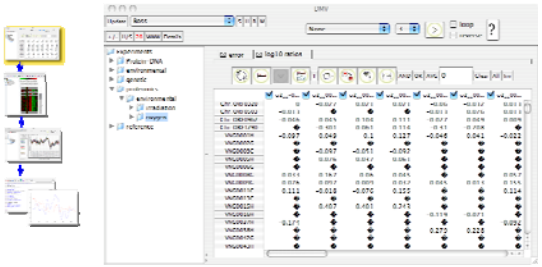
Gaggle Broadcasting



22

Example: Response of Halobacterium to changes in Oxygen

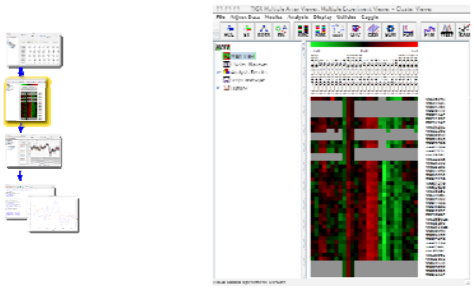
- Protein abundance measurements in DMV



23

Response of Halobacterium to changes in oxygen

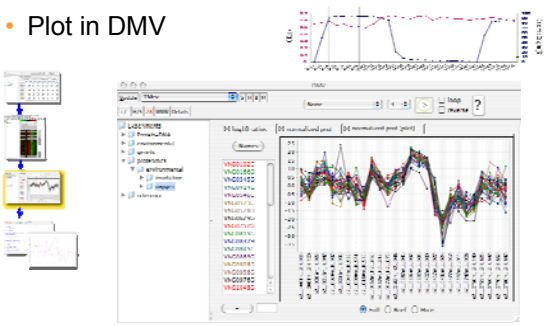
- Broadcast matrix to MeV and perform clustering



24

Response of Halobacterium to changes in oxygen

- Plot in DMV



25

Response of Halobacterium to changes in oxygen

- Plot mean expression levels (red) vs mean protein abundance (blue) in R statistical computing package

26

Credits

- PIPE
  - Ruedi Aebersold
    - Paul Shannon
    - Mi-Youn Brusniak
- Gaggle
  - Nitin Baliga
    - Chris Bare
    - Dan Tenenbaum

27

PIPE 2 and Gaggle

http://pipe2.systemsbiology.net/
(PIPE1: http://pipe.systemsbiology.net/)

Please follow along in this tutorial as we go through it in the class, or go your own pace if you choose. Feel free to add your own notes. If have suggestions or bug reports that might be useful for others, please email them to hramos@systemsbiology.org or mbrusniak@systemsbiology.org .

We will be using a set of Yeast proteins derived from a real ISB experiment. The researcher was interested in identifying any potential protein complex in the sample. These proteins had a protein prophet probability of > 0.9. We will use PIPE2, the Firegoose, Entrez gene, Kegg and STRING to explore the functions of and interactions between these proteins and come to a conclusion about any potential protein complex.

Before we get started, be sure to make sure that the most recent Gaggle Firefox extension is installed on your Firefox browser. At time of writing, this was version 0.8.270. (This step is already taken care of for the course laptops. For other computers, see http://gaggle.systemsbiology.net/docs/geese/firegoose).

I. Loading Data into PIPE2

For the sake of this tutorial, we will simply press a button to load our example set of proteins. However, options for importing your own lists of proteins (at a later time) include: broadcasting directly from ProteinProphet (through Firegoose), uploading a tab delimited text file, or copy and pasting tab delimited text directly into a new instance of the IDMapper PIPElet.

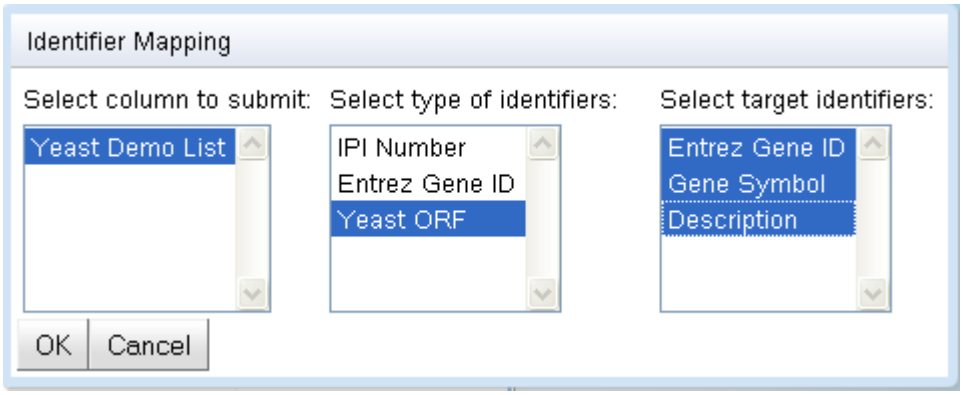
1. Within Firefox, go to: http://pipe2.systemsbiology.net/
2. Open a new instance of an IDMapper PIPElet.
3. Click the "Demo Yeast Proteins" button to load the set of proteins we will be working on in this tutorial.

II. Looking up Gene IDs

Entrez Gene IDs are used for a variety of functional annotations. In this step, we'll look up the Gene IDs for our yeast proteins, along with other bits of information.

1. Inside this new PIPElet, click the menu item "Operations" then "ID Mapping" to bring up the Identifier Mapping dialog box.

2. In the Identifier Mapping dialog box, make the following selections, then press OK:
   -Column to submit: Yeast Demo List
   -Type of identifiers: Yeast ORF
   -Target Identifiers: Entrez Gene ID, Gene Symbol, and Description (use the "Control" button to multi-select). The Identifier Mapping Dialog box should look like this:



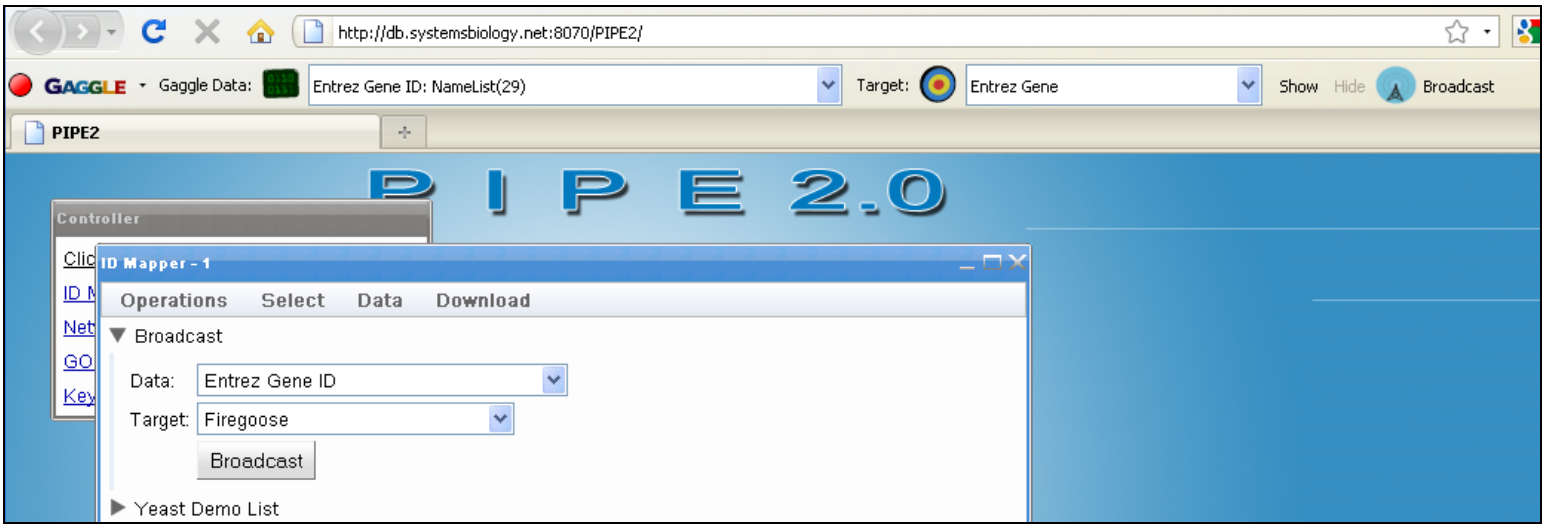
III. Entrez Gene Database (www)

Here we will query the online Entrez Gene database to gather a bit more information about some of our proteins.

To do this, we will use the "broadcasting" mechanisms of both PIPE2 and the Firegoose/Gaggle.

1. Send the Entrez Gene IDs to the Firegoose. Click the "Broadcast" arrow in the ID Mapper PIPElet to reveal the source and target fields. For data source, select "Entrez Gene ID". For target, select "Firegoose". Click "Broadcast" to send the list to the Firegoose.

2. Find the Gaggle toolbar (Firegoose) near the top of your browser window and select Entrez Gene as the target (and "Entrez Gene ID: NameList(29)" as the Data source), like this:



3. Press "Broadcast" button on the Firegoose. You will see the NCBI Entrez Gene index page for the genes. Click into these descriptions to find the answers to the following questions:

How many interactions are noted for each of the following genes (estimations are perfectly OK)?

1. FBA1 - _____
2. HXK2 - _____
3. MVD1- _____

Bonus question: How many of those interactions are with other genes in our list?

(You really don't have to answer that, but we'll visually answer this question in a minute)

IV. KEGG Database (www)

Let's see what KEGG has to say about our list of proteins.

1. Return to the PIPE2 tab in the Firefox browser. In the IDMapper PIPElet's broadcast panel, select "Yeast Demo List" as the data and "Firegoose" as the target, and press "Broadcast". In the Firegoose, change the Firegoose broadcast target to KEGG Pathway and press Broadcast.

2. In the resulting Pathway Search Results page, notice that 15 of our proteins are mapped onto the "Metabolic pathways" item. Click on it to open the pathways image.

In this image, the lines represent transitions/reactions (catalyzed by enzymes) of one compound transitioning into another (the circles). The red lines are those enzymes (proteins) contained in our list. If you hold the mouse over any of these objects on the screen, you will get more information about it. There are also labels scattered throughout describing the pathway in its proximity.

Which pathway has the highest concentration of our genes around it (most red lines in its proximity)? _____

What are the names of 2 of those proteins?
1. _____
2. _____

3. Press "Back" on the browser to go back to the Pathway Search Results.

What are items #3 and #5 on that list, and what proteins do they have in common? (hint: press "show all objects")

#3)_____

#5)_____

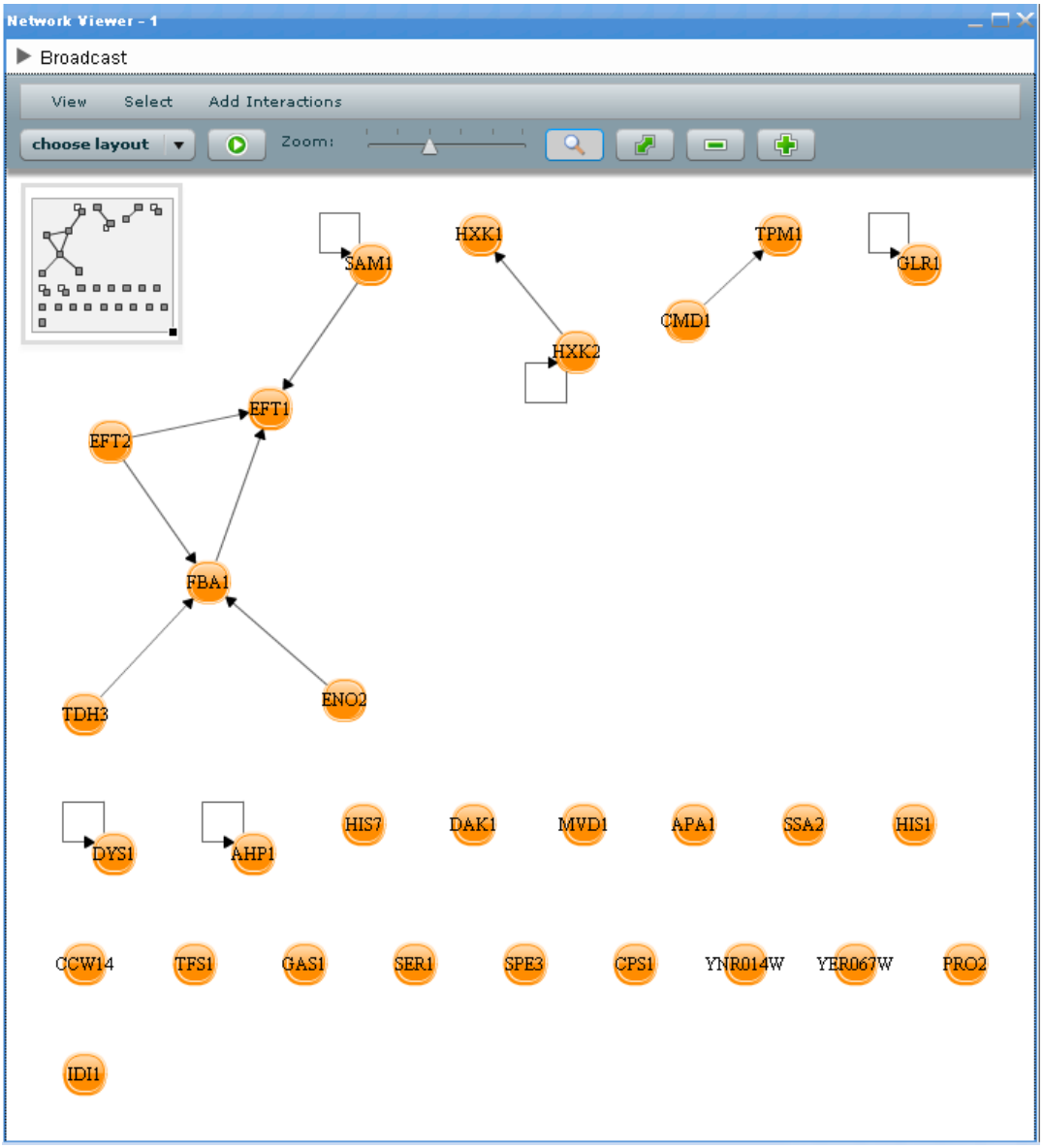
_____

V. Exploring interactions (Yeast-2-Hybrid + STRING)

Here we explore interactions through network views.

1. In PIPE2, open a new instance of a Network Viewer PIPElet.
2. Back in IDMapper -1, in the broadcast panel, select "Whole Spreadsheet" as the data source and "Network Viewer – 1" as the target, then hit "Broadcast".
3. In "Network Viewer – 1", in the menu bar, select "Add Interactions" -> "Yeast" -> "Add Yeast Two-Hybrid Interactions".
4. Press the layout button: 
5. From the menu bar, click "View" -> "Set Node Labels ->" -> "Gene Symbol".

Your Network Viewer PIPElet should look something like this:

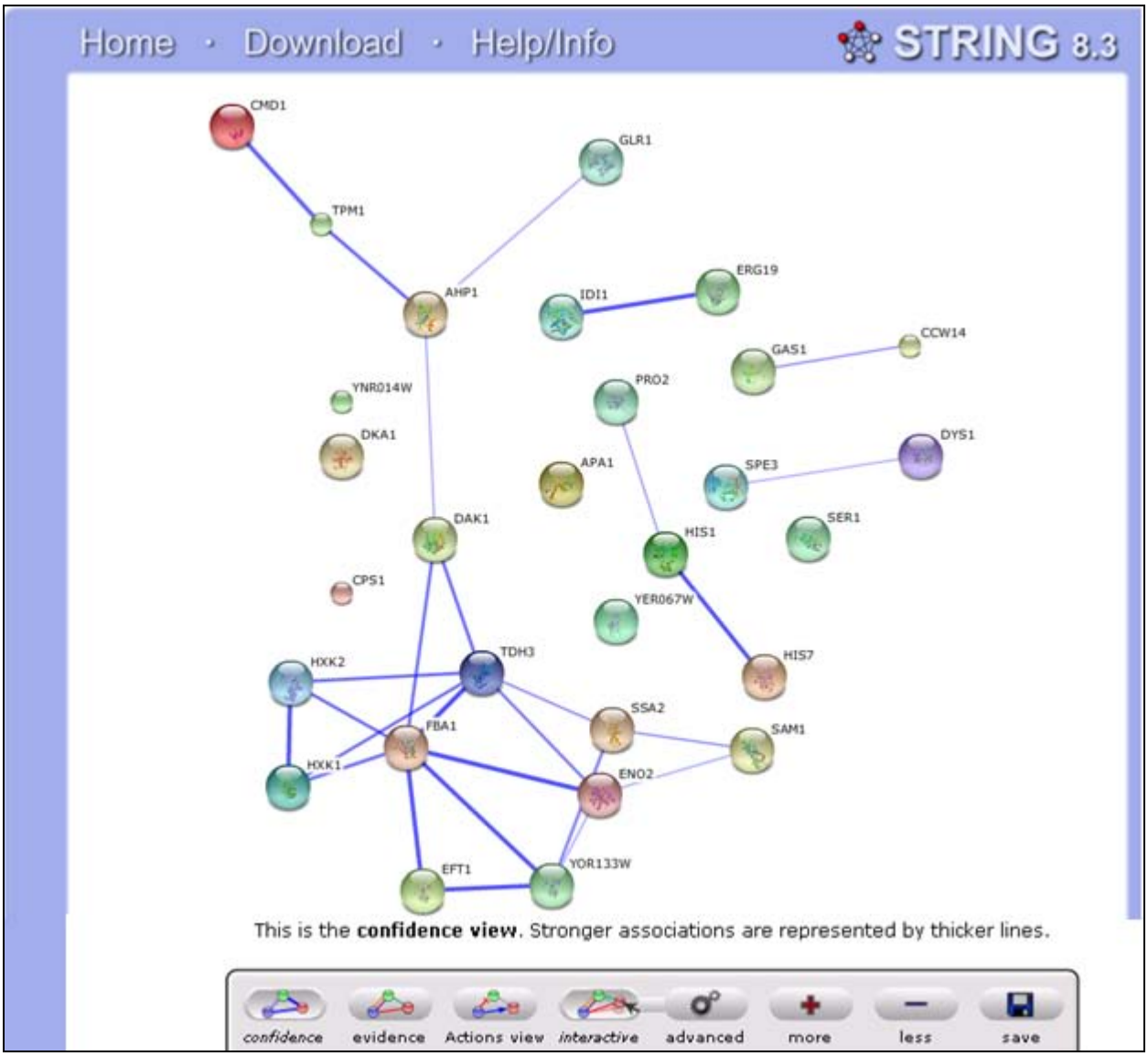


These interactions come from UW's Yeast-2-Hybrid interaction dataset.

Lets see what STRING says about these proteins.

6. Back in IDMapper – 1, broadcast the first column (By selecting "Yeast Demo List" from the broadcast panel data source field) of the data to the Firegoose, and from the Firegoose, broadcast to EMBL String. Press "Continue" until you get to this screen:

(Note: you may have to click the "confidence" icon to get this view.)



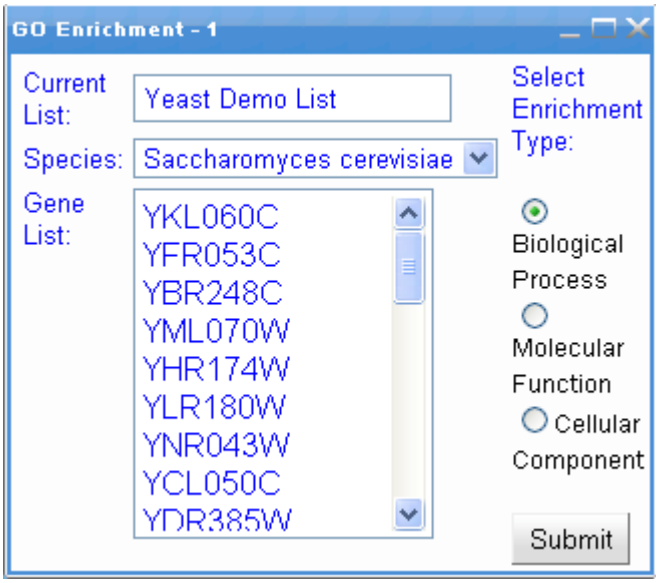
Locate the 3 proteins from the end of the last section (FBA1, HXK1, HXK2). Investigate the difference in connectivity between these 3 proteins. Where does STRING get the connections not found in PIPE2? (hint: click on the connecting edges) _____

We'll come back to these networks in a minute.

VI. Functional Enrichment with Gene Ontology Categories

Gene Ontology enrichment tells you which GO categories are significantly enriched for a list of proteins/genes.

1. In PIPE2, open a GO Enrichment PIPElet. From IDMapper – 1, broadcast the "Yeast Demo List" column to this new GO Enrichment PIPElet. The GO Enrichment PIPElet should look like this:



2. Hit "Submit". (When you get good at PIPE2, you can multitask and do other things while this process is completing, but for now, just relax.)

We are enriching for biological process GO categories. The p-value for each GO category corresponds to the hypergeometric distribution value based on the 4 parameters: # of items in your list that mapped to that category, your list's size, number of genes total (in the yeast genome) that map to the same category, and number of total genes possible in the organisms genome (for Yeast, ~6,000).
e.g., for "alcohol catabolic process":

hyperg(6, 29, 67, 6000) = 4.33804668787027e-07

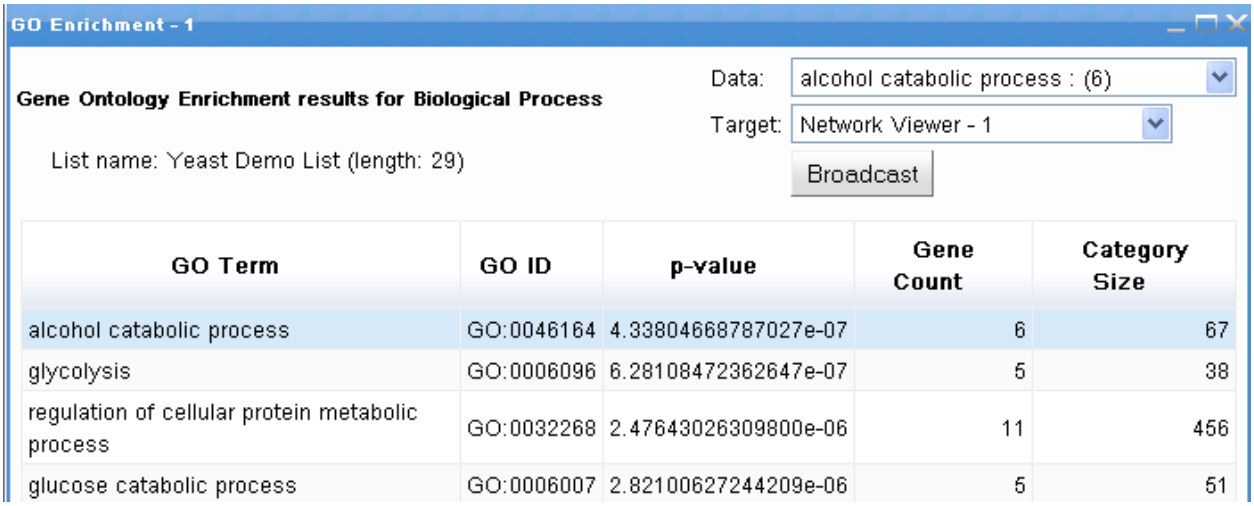
Notice that the results on the first page seem to also suggest a lot of sugar metabolic processes (like KEGG did).

VII. Integrating Annotation and Interaction Data

The Network Viewer is programmed to treat incoming GO terms uniquely. This might be useful in the following manner.

1. In the GO Enrichment – 1 PIPElet, select the "alcohol catabolic process" row of the table. This will add that category to the list of possible broadcast sources.

2. Open the Broadcast panel of the GO Enrichment – 1 PIPElet and select "alcohol catabolic process" as the data source and "Network Viewer – 1" PIPElet as the target, like so:



GO Term	GO ID	p-value	Gene Count	Category Size
alcohol catabolic process	GO:0046164	4.33804668787027e-07	6	67
glycolysis	GO:0006096	6.28108472362647e-07	5	38
regulation of cellular protein metabolic process	GO:0032268	2.47643026309800e-06	11	456
glucose catabolic process	GO:0006007	2.82100627244209e-06	5	51

3. Hit "Broadcast".

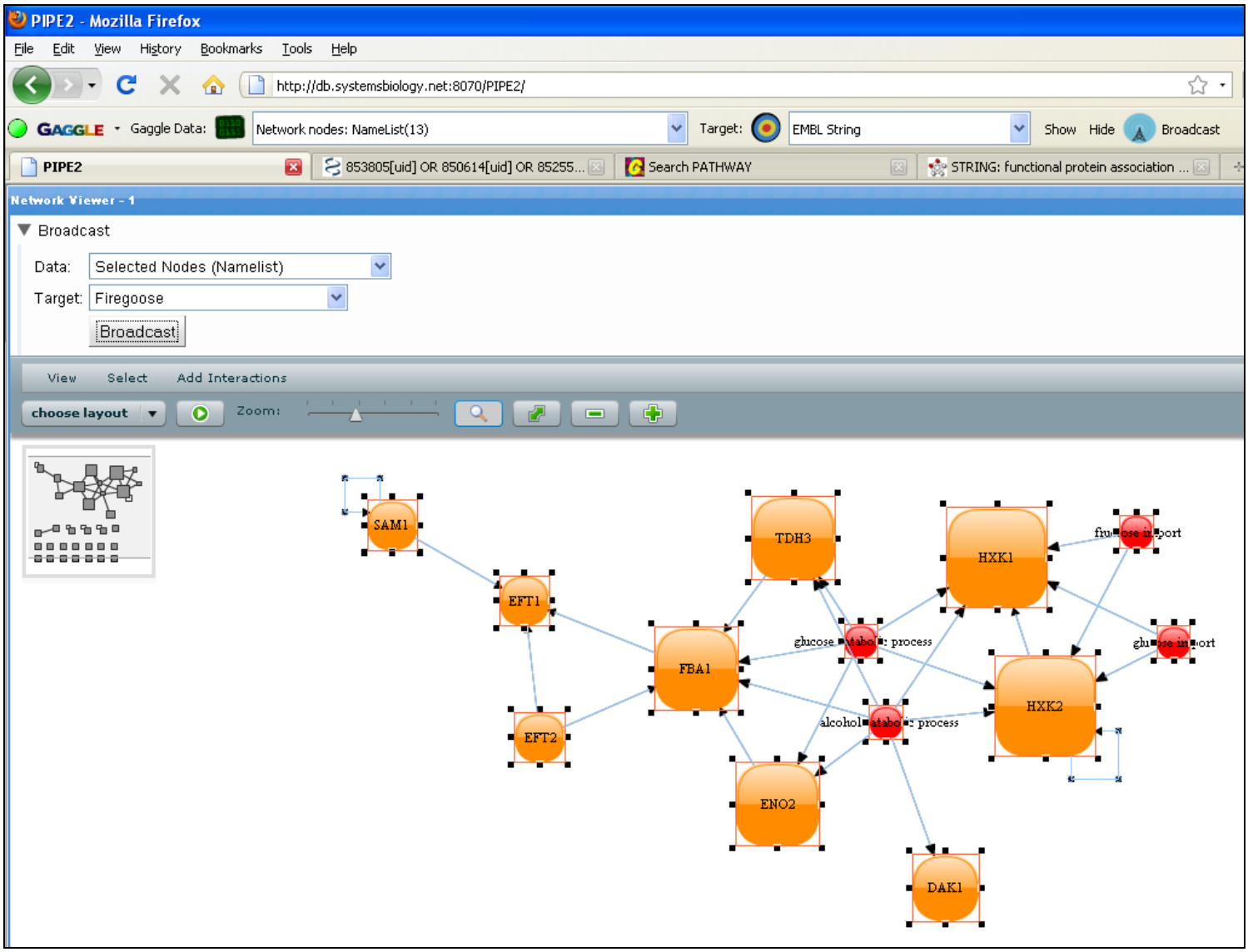
4. Do the same thing (broadcast to Network Viewer) with the following GO categories:
   - glucose catabolic process
   - fructose import
   - glucose import

5. Go back to the "Network Viewer – 1" PIPElet, maximize it (similar to windows on your desktop) and click the layout button: 

Now we have a cluster of proteins connected by direct interaction experiments (yeast-2-hybrid) and functional associations (GO terms). Let's see how STRING compares.

6. Select the cluster in the Network Viewer by clicking and dragging across it.

7. Expand the "Broadcast" panel of the Network Viewer PIPElet. Select "Selected Nodes (Namelist)" as the datasource and "Firegoose" as the target and hit "Broadcast". It should look something like this:

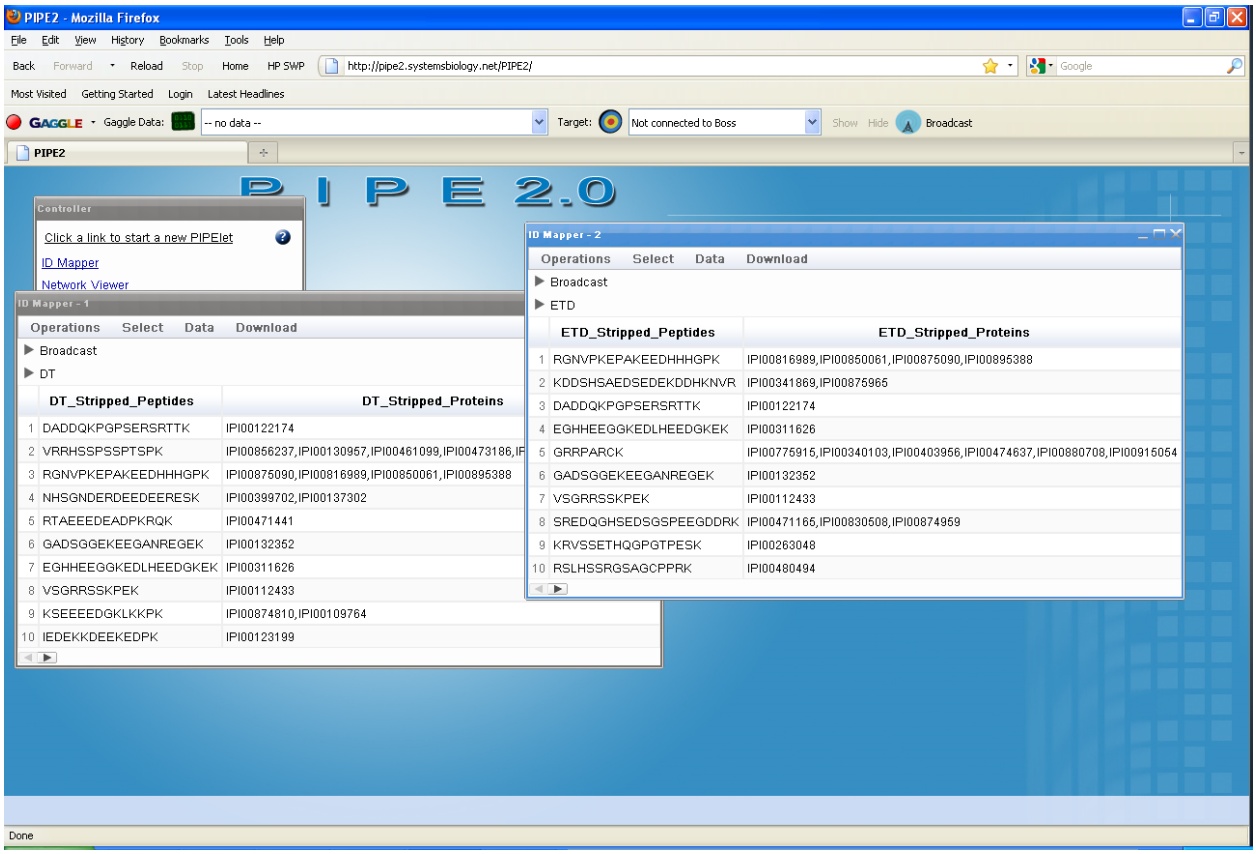


8. In the Firegoose, ensure "Network nodes: NameList(13)" is the data Source and "EMBL String" is the target, and hit "Broadcast".

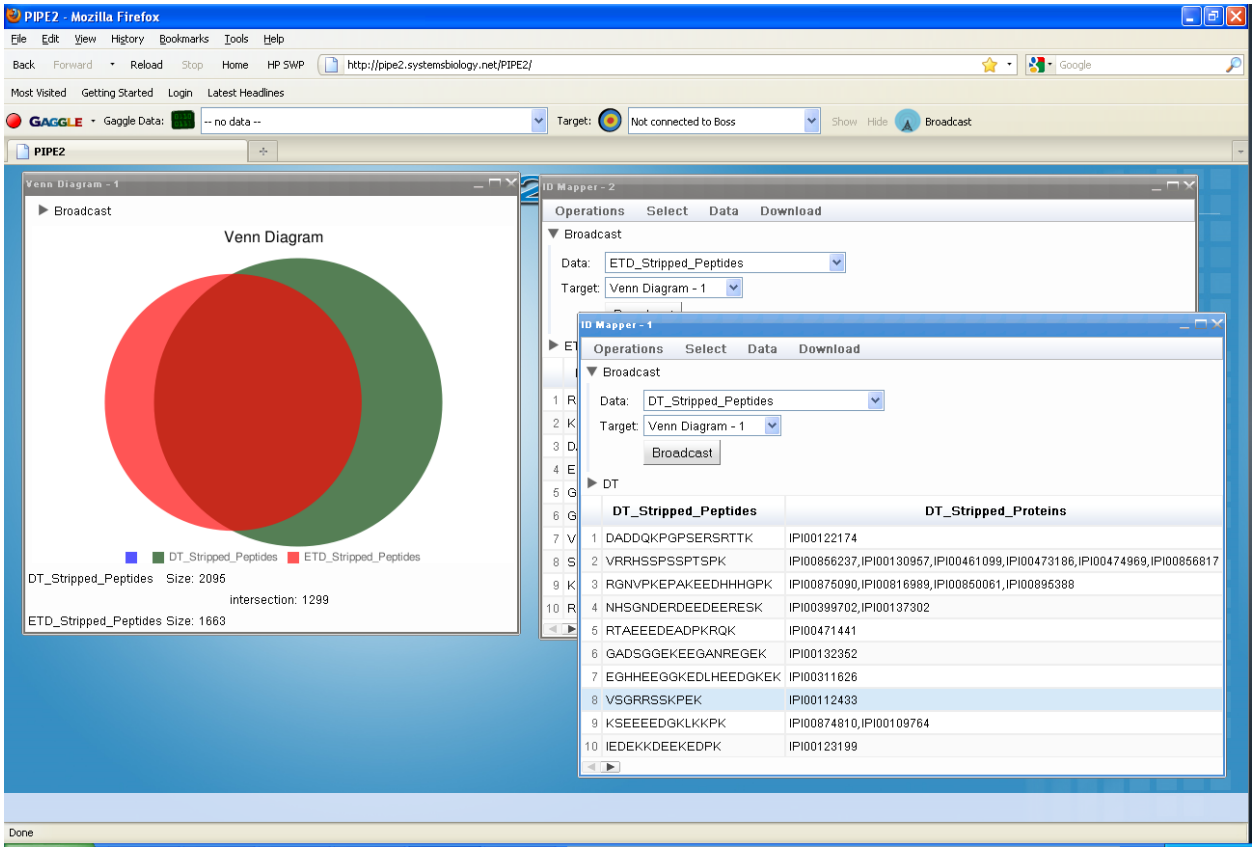
9. **Caution:** In String, select "Saccharomyces cerevisiae" as the organism and click "continue". On the page following that, String tries to map all of your input to identifiers it recognizes. In particular, at the bottom of the page, you'll notice that it also tried to map "alcohol catabolic process", "fructose import", "glucose catabolic process", and "glucose import". **Uncheck the mappings String attempted to make!** Then click "continue".

10. Explore the String network. In particular, look at edges that are in String and not in PIPE2. Click on them and investigate the evidence they provide for those edges. That type of information is not in PIPE2 yet… perhaps one day.

VIII. Venn Diagram Utility



Click IDMapper to open a new IDMapper, click “Browser” and select PIPE2_Demo_ETD_Stripped.tsv and choose ETD_Stripped_Peptides.
Click IDMapper to open another new IDMapper, click “Browser” and select PIPE2_Demo_DT_Stripped.tsv and choose DT_Stripped_Peptides.



Click Venn Diagram to open a Venn Diagram window, go back to IDMapper-1 then select ETD_Stripped_Peptides and select Venn Diagram-1. Click “broadcast”. Go back to IDMapper-2 then select DT_Stripped_Peptides and select Venn Diagram-1. Click “broadcast”. You will see the Venn Diagram with summary of those two sets overlap.

See if you can figure out how to download those peptides that are found only in the ETD_Stripped_peptides and not in the DT_Stripped_peptides list.

IX. Conclusion

No conclusive evidence for enrichment of any known protein complexes, however the co-occurrence of the 3 proteins FBA1, HXK1, and HXK2 in different annotation databases may warrant further experimental investigation into possible interactions.

Discovery and Validation Tools for Biomarker Research: ATAQS & TIQAM

Mi-Youn Brusniak

Day 5

October 29, 2010

Outline of Discussion

- Day 2: Discussion of Hypothesis Driven Target Analysis and Tools For Validation of Biomarkers
  - Introduction Hypothesis Driven Target Analysis SRM
  - Introduction to TIQAM-Digestor and TIQAM-viewer and Hands on Demo Using Tutorial
  - Introduction to ATAQS Hands on Demo Using Tutorial

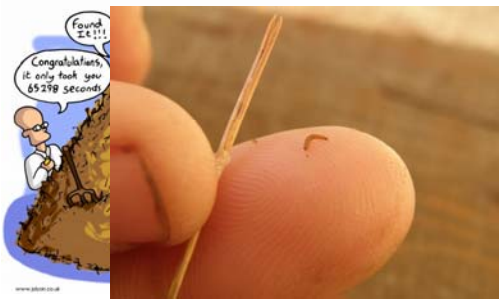
1

Outline of Discussion

- Day 2: Discussion of Hypothesis Driven Target Analysis and Tools For Validation of Biomarkers
  - Introduction Hypothesis Driven Target Analysis SRM
  - Introduction to TIQAM-Digestor and TIQAM-viewer and Hands on Demo Using Tutorial
  - Introduction to ATAQS Hands on Demo Using Tutorial

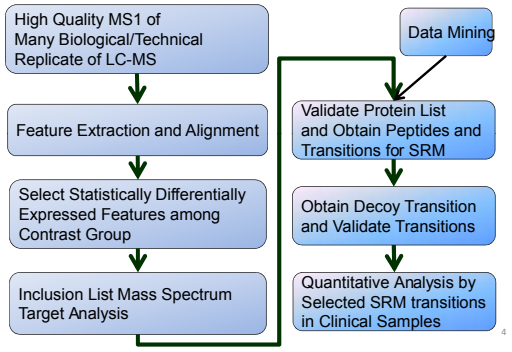
2

Challenges in plasma/tissue based biomarker discovery in proteomics



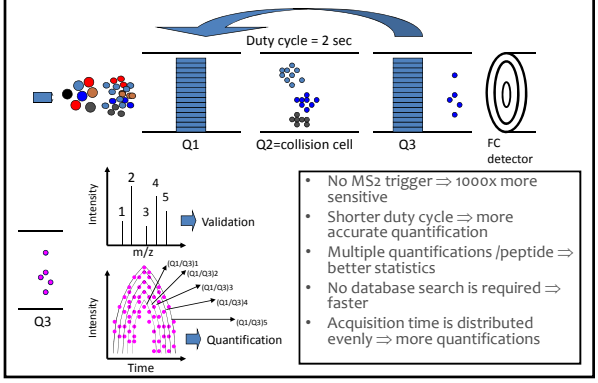
3

Corra & ATAQS Workflow

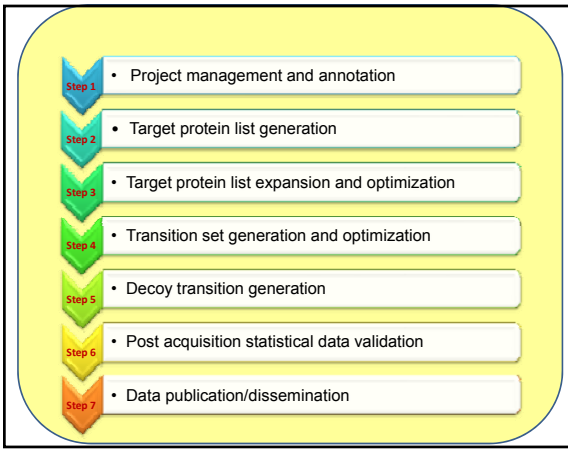


4

Fundamentals of SRM



- No MS2 trigger ⇒ 1000x more sensitive
- Shorter duty cycle ⇒ more accurate quantification
- Multiple quantifications /peptide ⇒ better statistics
- No database search is required ⇒ faster
- Acquisition time is distributed evenly ⇒ more quantifications



Outline of Discussion

- Day 2: Discussion of Hypothesis Driven Target Analysis and Tools For Validation of Biomarkers
  - Introduction Hypothesis Driven Target Analysis SRM
  - Introduction to TIQAM-Digestor and TIQAM-viewer and Hands on Demo Using Tutorial
  - Introduction to ATAQS Hands on Demo Using Tutorial

7

TIQAM: Publication

TIQAM
(Targeted Identification for Quantification Analysis by MRM)

Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring

Lange, V, Malmström,J, Didion,J, King,N, Johansson,B, Schäfer,J, Rameseder, J, Wong,C-H, Deutsch,E, Brusniak,M-Y, Bühlmann,P, Björck,L, Domon, B, Aebersold R. *MCP Papers in Press*. Published on April 13, 2008 as Manuscript M800032-MCP200

- http://tools.proteomecenter.org/TIQAM/TIQAM.html
- http://tools.proteomecenter.org/wiki
- http://groups.google.com/group/tiqam

8

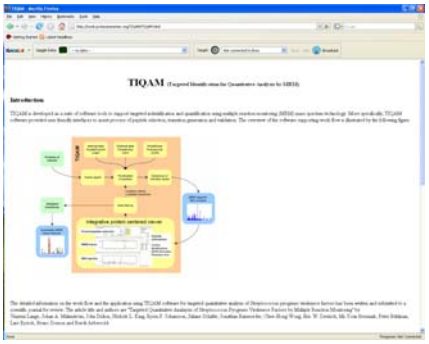
TIQAM:
Data Mining to Select Proteins of Interest

Bioinformatic Data Sets:

- Published biomarker list
- Medgene database hits for the query of the disease
- Microarray tissue profiling data of differential expression in particular tissues
- Microarray profiling for tissue specificity, protein-proteine interaction networks
- Enrichment of GO annotations

9

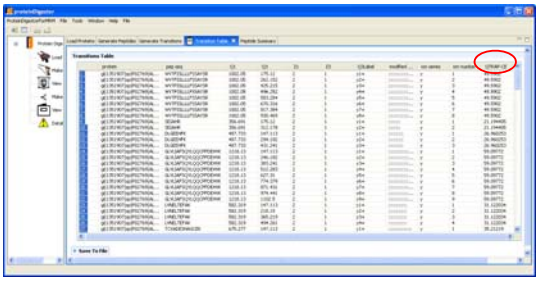
TIQAM: Work Flow



10

TIQAM: Digester

Candidate Transition Generator by TIQAM-digester



11

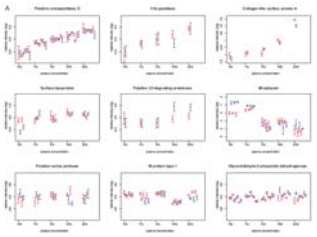
TIQAM: Viewer

Validator by TIQAM-viewer

12

TIQAM: Application

- Proteomic changes upon plasma exposure of S. pyogenes
- Identified a subset of virulence factors which is clearly induced upon contact with plasma and presumably is of particular importance during the early infection stage



13

Outline of Discussion

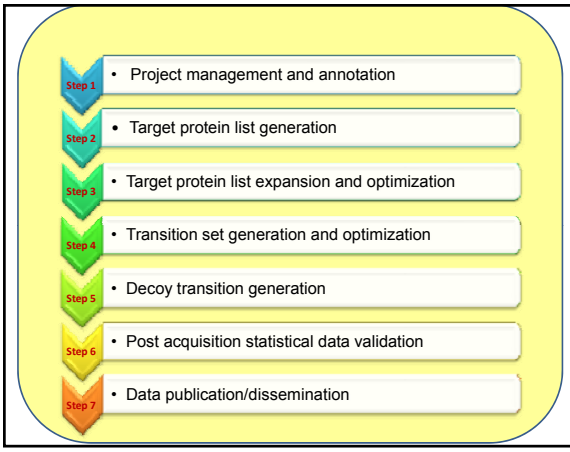
- Day 2: Discussion of Hypothesis Driven Target Analysis and Tools For Validation of Biomarkers
  - Introduction Hypothesis Driven Target Analysis SRM
  - Introduction to TIQAM-Digestor and TIQAM-viewer and Hands on Demo Using Tutorial
  - Introduction to ATAQS Hands on Demo Using Tutorial

14

ATAQS

- ATAQS (Automated and Targeted Analysis with Quantitative SRM) is a SRM (Selected Reaction Monitoring) pipeline software for high throughput SRM experiments in proteomics study.
- ATAQS provides a simple interface to generate and filter candidate peptide of research interests and to generate, filter and validate transition generation of peptides.
- TraML (Transition Markup Language) is proposed to proteomics community by PSI (proteomics standard initiatives) as a common data exchange format for validated transitions.
- ATAQS is a flexible pipeline that can start or end any point of the pipeline.

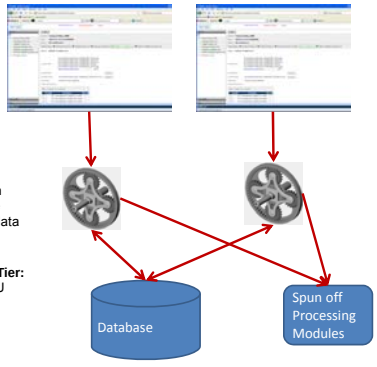
15



Presentation Tier: User interface ATAQS using Mozilla FireFox browser in any computer that can connect to ATAQS server.

Business Logic Tier: Servlet Coordinates the applications, by launching processing modules in processing system and interface with Database to store/retrieve data

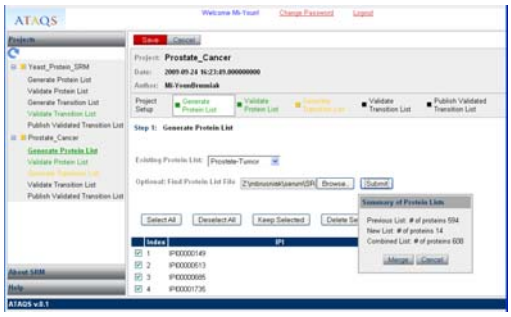
Data storage and Processing Tier: Data is stored/retrieved and CPU heavy processing modules are running in distributed CPU node system for high through put.



ATAQS

18

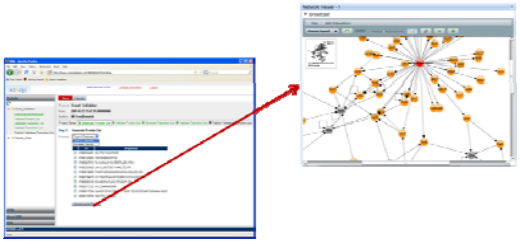
ATAQS



19

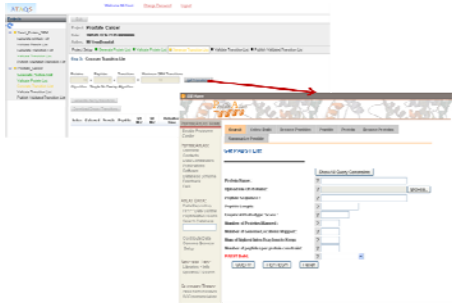
ATAQS

ATAQS provides automatic connection to PIPE2 web server.



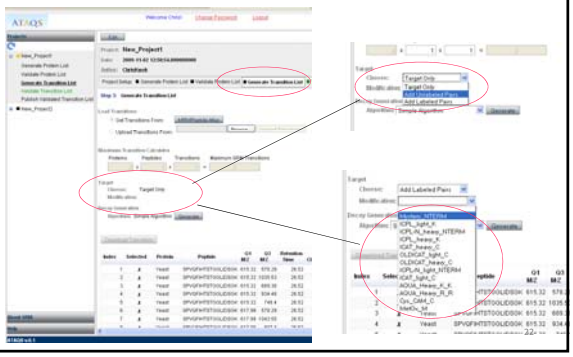
20

ATAQS

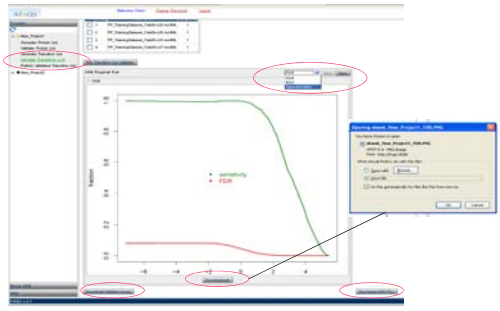


21

ATAQS

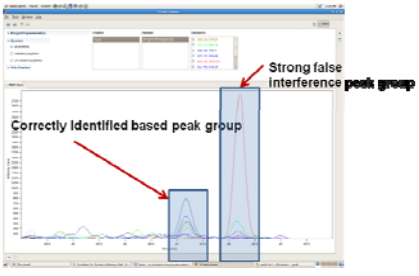


ATAQS



23

ATAQS



24

Summary

•LC-MS1 with Targeted MSMS and SRM technologies are a powerful method for identifying and quantifying proteins in complex samples

• Corra framework is very promising in large-scale protein profiling for discovery approach to quantitative proteomics studies.

•ATAQS is a SRM pipeline software for high throughput SRM experiments in proteomics study by providing generating/filtering/validation of protein, peptide and transition list.

• TIQAM is a suite of software which was developed for generating insilico transitions and visual SRM validation tools.

• SpectraComparer

Acknowledgments

ISB Seattle

Robert Moritz
Julian Watts
Kelly Cooke
Hollis Lau
Simon Letarte
David Campbell
Mark Christiansen
Sung-Tat Kwok
Hector Ramos
Vagisha Sharma
Nichol King

ETH Zurich

Ruedi Aebersold
Vinzenze Lange
Lukas Muller
Lukas Reiter
Paola Picotti
Alex Schmidt
Safia Thaminy

Funding

NHLBI, NCI, NIDDK,
Duchy of Luxembourg

Collaborators

Bernd Bodenmiller (Standford)
Olga Vitek (Purdue)

ATAQS v1.0 User's Guide

INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

Original tutorial with easy to follow screenshot can be downloaded from http://tools.proteomecenter.org/ATAQS/ATAQS.html

ATAQS: A Computational Software Tool for High Throughput Transition Optimization and Validation for Selected Reaction Monitoring Mass Spectrometry

Mi-Youn K. Brusniak[1], Sung-Tat Kwok[1], Mark Christiansen[1], David Campbell[1], Lukas Reiter[2], Paola Picotti[2], Ulrike Kusebauch[1], Hector Ramos[1], Eric W. Deutsch[1], Jingchun Chen[3], Robert L. Moritz[1], *Ruedi Aebersold[2,4,5]

[1]Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103, USA; [2]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland; [3]Information Warehouse, the Ohio State University Medical Center, OH; [4]Competence Center for Systems Physiology and Metabolic Disease, ETH Zurich, Zurich, Switzerland; [5]Faculty of Science, University of Zurich, Zurich, Switzerland;

Abstract

Since its inception, proteomics has essentially operated in a discovery mode with the goal of identifying and quantifying the maximal number of proteins in a sample. Increasingly, proteomic measurements are also supporting hypothesis-driven studies, in which a predetermined set of proteins is consistently detected and quantified in multiple samples. Selected reaction monitoring (SRM) (also referred to as multiple reaction monitoring, MRM) is a targeted mass spectrometric technique that supports the detection and quantification of specific proteins in complex samples at high sensitivity and reproducibility. In this manuscript, we describe ATAQS, an integrated software platform that supports all stages of targeted, SRM-based proteomics experiments, including protein and peptide target selection, SRM transition optimization and post acquisition data analysis. The new software will significantly facilitate the use of targeted proteomic techniques and therefore contribute to the generation of highly sensitive, reproducible and complete datasets that are particularly critical for the discovery and validation of biomarkers for hypothesis-driven studies in systems biology.

This documentation was prepared by Mi-Youn Brusniak (mbrusniak@systemsbiology.org).

1. Introduction

As a complement to the well-established discovery proteomic methods, targeted mass spectrometry based on SRM is becoming an important tool for the generation of reproducible, sensitive and quantitatively accurate data from biological samples. The method depends on the generation of target protein sets based on prior information and the one-time generation of validated mass spectrometric assays for each of the targeted proteins. The development of these assays depends on the optimal selection of peptides that represent the proteins on the target list and the optimal set of transitions for their detection in biological samples. Once developed, these assays can be continually applied across a multitude of studies.

The ATAQS pipeline and software provides a high throughput tool for organizing, generating and verifying transition lists and for the post acquisition analysis and dissemination of the data generated from applying the transition lists to studies of biological samples. ATAQS uses information from publicly accessible databases for the optimization of the protein and peptide target lists and for the optimization of a transition set. ATAQS is open source software that enables data-driven researchers to generate candidate protein lists and measure candidate proteins across a large number of biological samples, and allows algorithm-developing scientists to further develop the steps in the ATAQS pipeline. As needs arise, we plan to continuously expand on ATAQS functionalities (e.g., validation of quantification, support of SILAC type experiments, etc.).

We expect that ATAQS will find wide application as targeted proteomics increases in use to support hypothesis-driven research across all fields of life science. ATAQS is a single, user-friendly, informatics framework, that is simple to use and fully customizable, for the enabling of SRM-based proteomic workflows of any size, able to guide the user seamlessly from MS data generation, through data processing, visualization, and statistical analysis steps, to verify proteins of interest in biological samples. This is a user guide for ATAQS v1.0 software.

2. Login

Website: Ask administer in your institution which server the ATAQS is deployed to and ask ATAQS admin to add your account. For this guide, we will use ATAQS account.

The URL should be something like the following. http://moog.systemsbiology.net:8080/ATAQS. If you are using ATAQSDemo.systemsbiology.net for “read only” demo, you can type ATAQS in Username and Demo2010 in password and admin in Username and admin@isb2010 in password for Admin page demo in the following section.

3. Admin page

If you are logged in as admin, ATAQS leads to admin page, where you can configure for all users.

You can select **Users** in **Administers** panel, the main panel will display the current 'Users' list. When you click **"Add Users"** button, the **Add New Users** panel will be opened as shown in the left figure.

Administrator also can add available organisms. ATAQS has a step to connect publically available website PIPE2 (Protein Information and Property Explore) using the administrator defined organism. Thus, we advise administrator to see how the organism was named by PIPE2. Administrator can add organism by selecting **Organisms** in **Administrators** panel shown in the left figure.

Administrator also can add available mass spectrometry Instruments by selecting **Mass Spectrometry Instruments** option in **Administrators** panel. When you click "Add Mass Spectrometry Instrument" button, ATAQS connects to EBI website for http://www.ebi.ac.uk/ontology-lookup service to get Identifiers for all available mass spectrometry. Using the controlled vocabulary is necessary to generate resulted file to be exchangeable to community when user decided to publish their validated transitions at the end of ATAQS pipeline.

4. Project Setup Step

Creating and sharing project. After user login using their user account, ATAQS pipeline will show as shown in the left panel. Click "**New**" button in the top panel will ask the project name. You can type in Yeast_2000. Then it will go to "**Project Setup**" panel. User can select the organism and mass spectrometry for the project by using pull down options. There is "**Project Description**" text box that user can type overview of the experiment. You can select organism and mass instrument as shown in the figure and click "**Save**" button. ATAQS software is designed to serve institution where several collaborators are working with similar SRM experiments. Thus, ATAQS provide a way to share the projects with collaborators. When user select "**Share**" button, "**Share setting for Project:** *your named pro*ject" panel (circled in red) and list of users. You can select collaborators that you would like share the project. The current implementation of sharing project provides collaborators to access your project read-only mode. More specifically, collaborator cannot modify your steps. Create project by click "**New**" button and type *Yeast_2000*. Select Yeast in the Organism and select 4000 QTRAP for Mass spectrometry and you can write some description of the project. This project is an example of ATAQS paper using 100 heavy and light synthesized yeast peptides with 3 dilution series of three different background samples (glyco captured human plasma, C. elegans and Leptospira interrogans extracts). Then click "**Save**" button.

5. ATAQS Pipeline status indicator

ATAQS provides a quick way to help users to glace each project and steps status. In the left "**Projects**" Panel, it list all the projects that are you created and collaborators are sharing. The square button beside of each project indicate whether the projects are completed ("green") or not started ("black") or in the middle of the process ("yellow"). For example, left figure has yellow square box beside the project Yeast_Validator and there are total of five steps and the first four steps are completed but not the last step was done. Thus, the project status is yellow as not totally completed project. ATAQS lists the step in the right side panel where user can click to go to the step. In the panel, it also shows the status of the step. The step status is synchronized with the step status in "**Projects**" panel. The color of square buttons and step color indicates the each step status as shown in the left panel.

6. Generate Protein List

Select "**Generate Protein List**" step in the top panel. You will see "Step1: Generate Project List" step.

There are two ways to populate protein list in this step. If there are institution wide database, administrator can add the protein list in tsv format to ATAQS site so all the users are accessible to the same database. As example, this version of ATAQS contains three bioinformatically curated disease-specific protein candidate lists: (1) Prostate tumor containing 1055 proteins, (2) Type II diabetes containing 954 proteins and (3) Breast cancer-related human kinase signaling containing 32 proteins. You can select one of the three database or administrator installed institution specific database from the pull down menu of "**Existing Protein List**" option. Or you can load your own protein list by clicking "**Edit**" button and click "**Browser**" then select *ATAQS_YeastProteinList.csv* and click "**upload"** button. The summary of protein uploaded will be shown like the picture. In this step, if you had already some of protein, ATAQS will merge them with unique protein entry list.

7. Setting up connection to other website

Make sure you are using firefox (3.x versions) browser and installed Firegoose using Firegoose-0.8.208.xpi or higher from http://gaggle.systemsbiology.net/docs/geese/firegoose/install/ as indicated in ATAQS installation guide. Go to Firefox and select pull down menu of firegoose. Then select "Add Custom Website Handler". It will bring the panel shown in the left side. You will fill out the name and URL. This is the way you can receive back any data from public website. The current version of ATAQS uses three publically available website, PIPE2, MRMAtlas and TraML uploading website for MRMAtlas backend repository.

8. Investigating Protein Properties using PIPE2

Go back to ATAQS "**Generate Protein List**" step. You can investigate the uploaded protein list properties further using PIPE2. After investigating Protein list, you can expend or remove protein list.

Click "**Select All**" Button then click "**Send List to PIPE II**" button in the bottom of the panel. The list will be sent to PIPE2 website by firegoose. New tab in the firefox browser will be open and your protein list will be displaced as indicated in the right side figure. You can use PIPE2 tools (please refer PIPE2 tutorial guide for detailed PIPE2 functionality) to generate Protein-Protein interaction maps as shown in the figure. You can select subset of proteins from the network view and send back to ATAQS protein list as next page figures. For this exercise, we will broadcast back the same list.

In the PIPE2, IDMapper PIPElet, Select *fromFiregoose* in **Data** pull down menu and select "Firegoose" in **Target** pull down menu and click "**Broadcast**". Select "ATAQS" in the Firegoose menu in your Firefox browser and click "**Broadcas**t" button of the Firegoose. Your new protein list (in this example, the same protein list) will be back to ATAQS "**Validate Protein List**" panel shown below. Note that "**Generated Protein List**" step is in green and "**Validate Protein List**" step is in yellow.

When you click "**Save**" button. The "**Validate Protein List**" step turns to green to indicate that this step is also completed for this project.

9. Generating Transition List.

ATAQS "**Generate Transition List**" allows either uploading your optimized transition list or obtaining best observable peptides from MRMAtlas while considering user weighted penalty factor. In this manual, you will demonstrate both ways. First, click "**Edit**" button and click "**MRMPeptideAltas**" button as shown in the left figure. It will bring MRMPeptideAtlas page as shown in the next figure. Notice that your proteins are already filled in that page. Select *YeastPublic2010-02* in PABST (PeptideAtlas Best Transition) build and select 3 for number of peptide per proteins. Exam all the options in getting number of transitions and also options to excluding peptides with certain amino acid.When you click "Get Transition", it will show transitions. You can again broadcast back the transitions by clicking "Broadcast" button in Firegoose, those transition list will be back to your project in ATAQS as shown below figure.

As mentioned earlier, ATAQS can also just take transition list from users by Clicking "**Browser**" and select ATAQS_D2_Transition_TargetOnly.csv file. Then click "**upload**", the uploaded transition will replace the transitions we got from MRMPeptide Atlas. ATAQS allows latest transitions either from user upload or from MRM PeptideAtla to be the current transitions. As described in the paper, you need decoy to validate your transition detection. Thus, this step allows generating additional transitions to append to your current transition. If transitions are generated from MRM Peptide, you may need heavy peptide (e.g., AQUA peptide) pair transitions to measure both heavy and light transition in your biological sample. Or you optimized transition using heavy peptides, and then you need light pair to measure in your sample as well. Thus ATAQS allows several options to append transitions to your current transitions. Entire selection of those options are shown in the figure.

As described in the paper, ATAQS needs decoy transitions to score your measured transitions with your sample. You can easily extend ATAQS software to add additional decoy algorithms. Current version of ATAQS comes with two decoy generating algorithms described in the paper. In this example, the

uploaded yeast transitions have heavy and light transitions. Thus, we will use "**Target Only**" option in Target section and select "Simple No Overlap Algorithm" for "Decoy Generation" option. Then click "**Generate**" button. Notice the step goes to "**yellow**" state and the algorithm generating process were initiated in one of your institution computing node. ATAQS designed to separate computing intensive processed to be outside of servlet so ATAQS web application would not be locked.

In this example, when the process is finished, total 2000 transitions were generated. Then the transitions can be downloaded by click "**Download Transition**" button to download the transition in your desktop. The transition list can be used in measuring in your sample.

10. Validating Transition List.

In this example, we split 2000 transitions to four 500 transition and prepared 9 samples as described in the paper (3 dilutions and 3 kind biological samples). Thus total 36 (4 transition set per sample) LC-MS were run in 4000 QTrap MRM mode. It's daunting for users to go through and validate each transition to see whether they detected the peptide or not manually. Thus, the current ATAQS version has mProphet module to assign score to the peptide based on discriminate properties between decoy and target transitions. Since the transitions were split to four, ATAQS allows user to group those samples together. First, ATAQS can take either mzXML or mzML files. You select "**Add**" buttons to select all 36 mzXML files. Select files 5-8, 13-16, 21-24, 29-32, 37-40, 45-48, 53-56, 61-64, 69-72 numbered mzXML files. Click "**Max. Sample Set Count**" and type *9*. Then assign each four set of mzXML to each "Run ID" to indicate which samples belong to one. As mentioned in the paper, users can optimize their own transitions so ATAQS allows uploading transition list which matches their measured transitions. The transition list will be uploaded by clicking "**Add**" in transition file and select D2_TransitionList.csv. ATAQS allows computational biologists to extend any of algorithms to be part of ATAQS. In this version, we provide two algorithms. For this example, click "**Transition Group Algorithm**" in Algorithm section then click "**Run Validator**". This step takes a few minutes to finish up. Similar to generating transitions, mProphet module will be running in one of your institution distributing node. When "Validate Transition List" step is completed, the step will turn to green and ATAQS displays the graphical summery of the dataset. There will be a drop down menu to show "**ROC**", "**FDR**" and "**Separation Bar Chart**". You can download the scored transition to see which transition group has higher validated score. For example, IAWEALAVER_1 in downloaded file Yeast_2000_top_pg.xls means the peptide is detected in Run ID sample group 1.

11. Publishing Validated Transitions.

As an optional step, all optimized and validated transitions can be made available to the community, so that the data can be used for SRM-driven biological research such as biomarker validation. Or simply store your final transition list in standard format in your institution. ATAQS also introduces a new proposed file format called TraML (Transition Markup Language) as a common data exchange format for validated transition information as described in ATAQS paper. ATAQS helps create TraML format files for exchange of validated transition information and if the user chooses to publish their data, ATAQS provides an easy way to upload user-created TraML files to public SRM databases, such as MRMAtlas.

In the last step of ATAQS “Publish Validate Transition List”, you can fill out a few contact information for author of the generated TraML file. Simply click “Create TraML” will generate TraML file with your project name and author name. In this case, Yeast_2000_Mi-YounBrusniak.TraML will be generated and you can download and modify the file or simple save the file. You can browse the content of TraML using various xml viewers including Firefox browser shown below.

When you decide to publish your TraML file to MRMAltas website, you can upload the file and select “Publish” button. ATAQS will validate the TraML based on current TraML xml schema for well formed and then upload MRMAtlas designated website. When the files are successfully uploaded to MRMAltas, ATAQS will generate “**Event Notification**” Panel to indicate success as shown in the figure.

12. Conclusion

As a complement to the well-established discovery proteomic methods, targeted mass spectrometry based on SRM is becoming an important tool for the generation of reproducible, sensitive and quantitatively accurate data from biological samples. The method depends on the generation of target protein sets based on prior information and the one-time generation of validated mass spectrometric assays for each of the targeted proteins. The development of these assays depends on the optimal selection of peptides that represent the proteins on the target list and the optimal set of transitions for their detection in biological samples. Once developed, these assays can be continually applied across a multitude of studies.

The ATAQS pipeline and software provides a high throughput tool for organizing, generating and verifying transition lists and for the post acquisition analysis and dissemination of the data generated from applying the transition lists to studies of biological samples. ATAQS uses information from publicly accessible databases for the optimization of the protein and peptide target lists and for the optimization of a transition set. ATAQS is open source software that enables data-driven researchers to generate candidate protein lists and measure candidate proteins across a large number of biological samples, and allows algorithm-developing scientists to further develop the steps in the ATAQS pipeline. As needs arise, we plan to continuously expand on ATAQS functionalities (e.g., validation of quantification, support of SILAC type experiments, etc.).

We expect that ATAQS will find wide application as targeted proteomics increases in use to support hypothesis-driven research across all fields of life science.

TIQAM-Digestor tutorial

Original tutorial with easy to follow screenshot can be downloaded from http://tools.proteomecenter.org/TIQAM/TIQAM.html

1. Double Click Short cut of TIQAM-Digestor
2. **File >> New >> Create new PD project**
3. Type PD Project Name and Description as shown in the screen then click **"Finish"**
4. Click **"Select Fasta file holding proteins in Protein Loader"** Select StreptococcusPyogene. fasta in the "My Document/TIQAM-ClassDemoData"
5. Fructose-bisphosphate aldolase will be loaded as it is shown in the left figure. Click ClassDemo in **"Select PD Project"** tab, then click **"Load Proteins"** button
6. Change "pep mass >" query to be 600 as shown in the left figure and then click **"Make Petides"** button
7. Click **"Annotate Peptides"** tab and type in Name filed and select Type field and select File StreptococcusPyogene_pepXML.xml using **"Brower"** button shown in the left bottom figure. The clock **"Import Annotations"** button to see the right bottom figure.
8. Click "Proceed to Make Transitions" button it will show Generate Transitions page as you see in the left figure. Select +2 and +3 in z1 column while holding down shift key then click **"Make Transitions"**
9. It will show Transition Table, click "Save to File" tab then it will show **"save"** button (shown in the left figure) and save as StreptococcusPyogene_Digestor in "My Document/TIQAM-ClassDemoData".
10. Exercise
    Open the exported file StreptococcusPyogene.txt in Excel and order by CE (Collision Energy). Is there any correlation with Mass?
    Could you think of way to get better Retention time for schedule SRM?

TIQAM-Viewer tutorial

Original tutorial with easy to follow screenshot can be downloaded from http://tools.proteomecenter.org/TIQAM/TIQAM.html

1. Double Click short cut of TIQAM-Viewer.
2. **File >> Import**

   Select **"A new project"**

   Click "**Next"**
3. Fill out the Create Project wizard as shown in the left figure
4. Click "**Next"**
5. Fill out the Experiment wizard as shown in the left figure
6. Click **"Manually select files"** and select the following files located in the "My Document/TIQAM-ClassDemoData" : Transition File (StreptococcusPyogene.csv), mzXML (0223House4.mzXML, 0223House5.mzXML), pepXML (StreptococcusPyogene_pep.xml). Select Peptides only before click "**Next"** .
7. When it's imported, it will show the summary dataset shown in the right figure
8. Click **"refresh"** button to show your imported project and experiment.
9. Select **"all proteins"** in Queries section and you will have the screen shown below
10. Exercise.

    Go through the transitions and you can decide the level of validation whether the transition sets for a given peptide validate the peptide identification.

For example, the peptide AVQGAVEER of ALF_STRP1 Protein generates strong peptide (shown below figure). You can annotate this as strong category in Peptide Validation.

Which three transitions would you like to use for quantifying AVQGAVEER peptide? Please find your top three candidates from transitions table and write.

You may go through more peptides and proteins to see how you would like to annotate peptide validation and select best transitions.

PROTEOMICS INFORMATICS COURSE - READING LIST

Overview:

- **Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiol Genomics. 2008;33(1):18-25.**
- **Deutsch, EW, Mendoza, L, Shteynberg, D, Farrah, T, Lam, H, Tasman, N, Sun, Z, Nilsson, E, Pratt, B, Prazen, B, Eng, JK, Nesvizhskii, A, Aebersold, R, A Guided Tour of the Trans Proteomic Pipeline, 2010, Proteomics, 10, 6, 1150**
- Aebersold R and Mann M. Mass Spectrometry-based Proteomics. Nature 2003;422:198-207.
- Nesvizhskii AI, Vitek O, and Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007;4:787-797.
- Keller A, Eng J, Zhang N, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Molecular Systems Biology (2005) doi:10.1038/msb4100024.

Database Searching:

- **Steen H and Mann M. The ABCs (and XYZs) of peptide sequencing. Nature Reviews 2004;5:699-714.**
- Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics. 2005 Aug;5(13):3475-9.

PeptideProphet:

- **Keller A, Nesvizhskii AI, Kolker E., Aebersold R. Empirical Statistical Model to Evaluate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Anal. Chem. 2002;74:5383.**
- Keller A, Eng J, Zhang N, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Molecular Systems Biology (2005) doi:10.1038/msb4100024.
- Choi, H, Ghosh, D, and Nesvizhskii, AI. Statistical Validation of Peptide Identifications in Large-Scale Proteomics Using the Target-Decoy Database Search Strategy and Flexible Mixture Modeling, Journal of Proteome Research. 2008, 7, 1, 286-292.

ProteinProphet:

- **Nesvizhskii A.I. and Aebersold R. Interpretation of shotgun proteomic data. Molecular and Cellular Proteomics 2005;4:1419-40.**
- Nesvizhskii AI, Keller A, Kolker E, Aebersold R. Statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 2003;75:4646-4658.
- Nesvizhskii AI and Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discovery Today 2003;9:173-181.

ASAPRatio:

- **Li X-J, Zhang H, Ranish JA, Aebersold R. Automated Statistical Analysis of Protein Abundance Ratios from Data Generated by Stable-Isotope Dilution and Tandem Mass Spectrometry. Anal. Chem. 2003; 75:6648-6657.**

Qualscore:

- **Nesvizhskii AI, Eddes JS et al Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol. Cell. Proteomics. 2006 Apr;5(4):652-70.**

SpectraST:

- **Lam H et al Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 2007; 7(5): 655-67.**
- Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold A. Building consensus spectral libraries for peptide identification in proteomics. Nature Methods 2008; 5(10):851-910.

PeptideAtlas:

- **Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas Project. Nucleic Acids Research 2006;34:D655-D658.**
- Deutsch, EW, Lam, H., and Aebersold, R. PeptideAtlas: A Resource for Target Selection for Emerging Targeted Proteomics Workflows. EMBO Reports, 2008; 9:49.

Corra, TIQAM:

- **Brusniak MY , Bodenmiller B, Campbell C , et al. Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. BMC Bioinformatics 2008; 9:542.**
- Lange V, Malmstrom JA, Didion J, et al. Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. Molecular & Cellular Proteomics 2008;7(8): 1489.

PIPE, Gaggle, and Cytoscape:

- **Ramos H, Shannon P, Aebersold R. The Protein Information and Property Explorer: a rich-client web application for the management and functional exploration of proteomic data. Bioinformatics 2008; 24(18):2110-2111.**
- Shannon P, Reiss DJ, Bonneau R, and Baliga NS. Gaggle: An open-source software system for integrating bioinformatics software and data sources. BMC Bioinformatics 7: 176.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker . Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13: 2498-2504.

WEBSITE RESOURCES

- Contact us: info@proteomecenter.org
- Seattle Proteome Center – http://www.proteomecenter.org
- Course presentations – http://proteomecenter.org/course.php
- SPC Tools Wiki-- general information for and by users of SPC software including downloads: http://tools.proteomecenter.org/wiki/
- SPC Tools Discussion List-- question and answer list for SPC software users: http://groups.google.com/group/spctools-discuss
- SBEAMS web site – http://www.sbeams.org
- PeptideAtlas web site – http://www.peptideatlas.org
- Spectral libraries central web site – http://www.peptideatlas.org/speclib/
- PIPE - http://pipe.systemsbiology.net/
- Cytoscape web site – http://www.cytoscape.org
- The Australian Proteomics Computational Facility – http://www.apcf.edu.au