Applications for Peptide Mass Spectral Libraries in Proteomics

Paul Rudnick

Mass Spectrometry Data Center

National Institute of Standards and Technology

Gaithersburg, MD

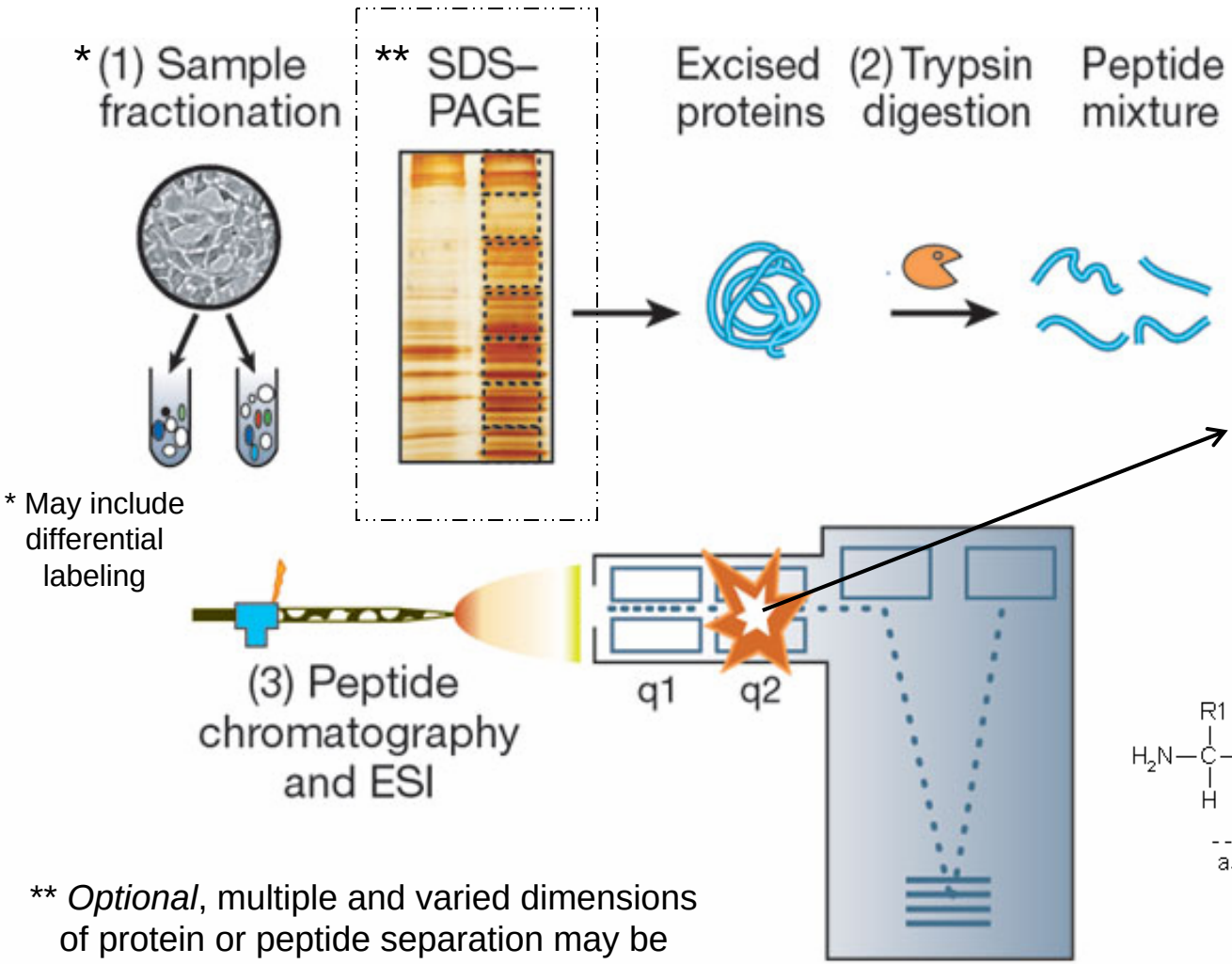
NIST
National Institute of Standards and Technology

NIST Chemical Reference Data Group

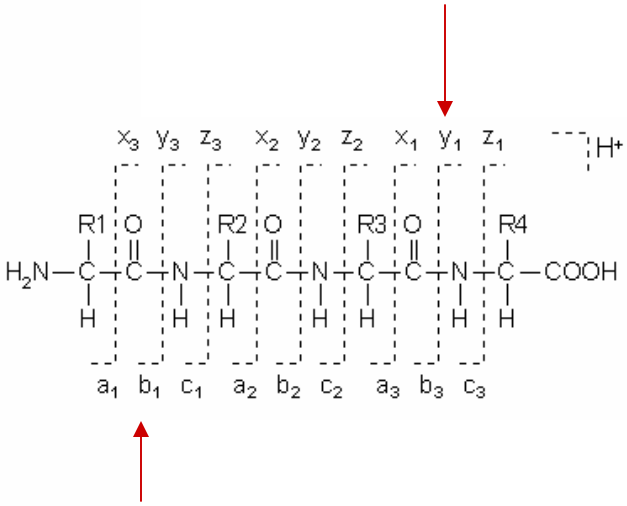
Overview

1. A typical “shotgun” proteomics workflow
2. Peptide identification by searching sequences (e.g. Sequest, MASCOT) – ‘current practice’
3. Libraries of peptide fragmentation spectra– a “new” alternative
   - Description
     - History – GC/MS, EI spectra
     - Construction of high-quality, consensus libraries
     - Structure/format of the library entries
   - Searching
     - NIST algorithms, Windows software
     - Other algorithms, publications
     - Compared to ‘current practice’
4. Integrating the two search methods (OMSSA/NISTMS)
5. Current status of peptide library projects and availability
6. What’s next

“Shotgun” Proteomics



backbone fragmentation from CID gives rise to <u>reproducible</u> ***b*** and ***y*** ions series spectra MS/MS



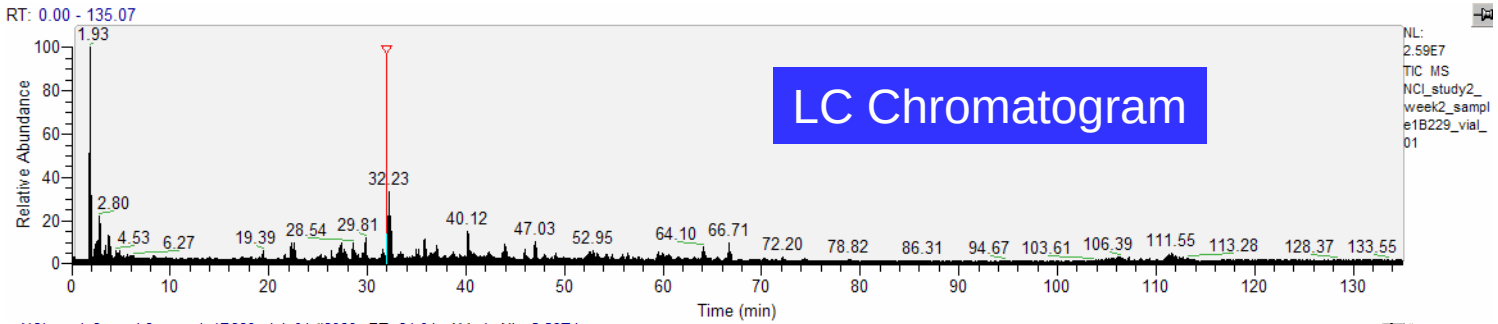
** *Optional*, multiple and varied dimensions of protein or peptide separation may be added to extend “depth” of analysis

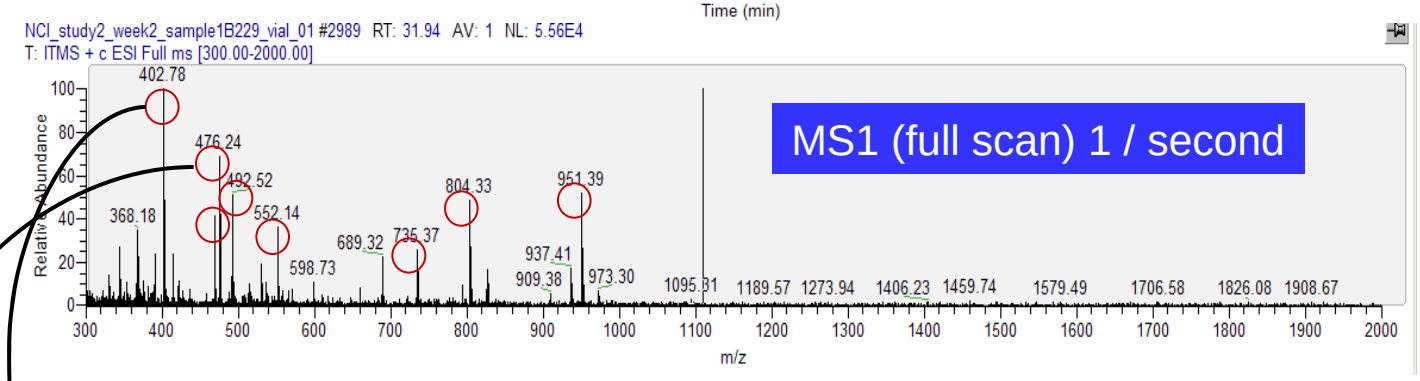
Aebersold & Mann, *Nature* **422**, 198-207 (13 March 2003)

Why do these experiments?

- Confirmation of gene products (protein sequencing)
- Differential expression studies (i.e.. global or targeted)
- Discovery of “biomarkers”
- Validation of candidate markers
- Discovery / mapping / validation of PTMs

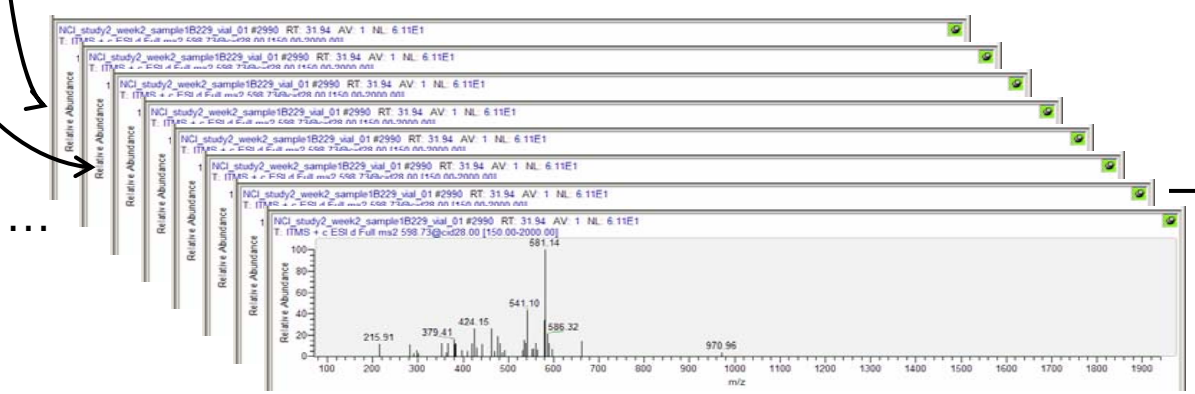
LC-ESI-MS/MS Data Acquisition





CID

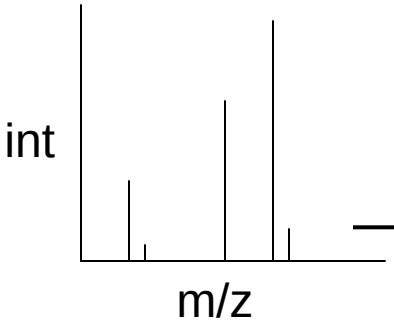
~8 MS/MS sampled from most intense MS1 ions (8-1)



…

Raw data passed to analysis

Data Analysis (Bioinformatics "Pipeline")



[~10,000 MS/MS spectra / 60 min. gradient, complex sample]

MS/MS spectra + monoisotopic precursor m/z extracted from raw data to ASCII text

extract

Peak list (e.g. dta, mgf, mzXML)

Sequest, MASCOT, X!Tandem, OMSSA, Spectrum Mill, others

searching / identification

Search Results – peptide identifications and raw scores

Reverse (decoy) database search results, PeptideProphet, distribution analysis

compute confidence

Discriminant scores, E-values, FDRs

User-defined threshold (e.g. 1%, 5%)

filter

Filtered peptide list with probabilities

Protein mapping, ProteinProphet, parsimony analysis

map proteins

Matched Protein – Peptide List

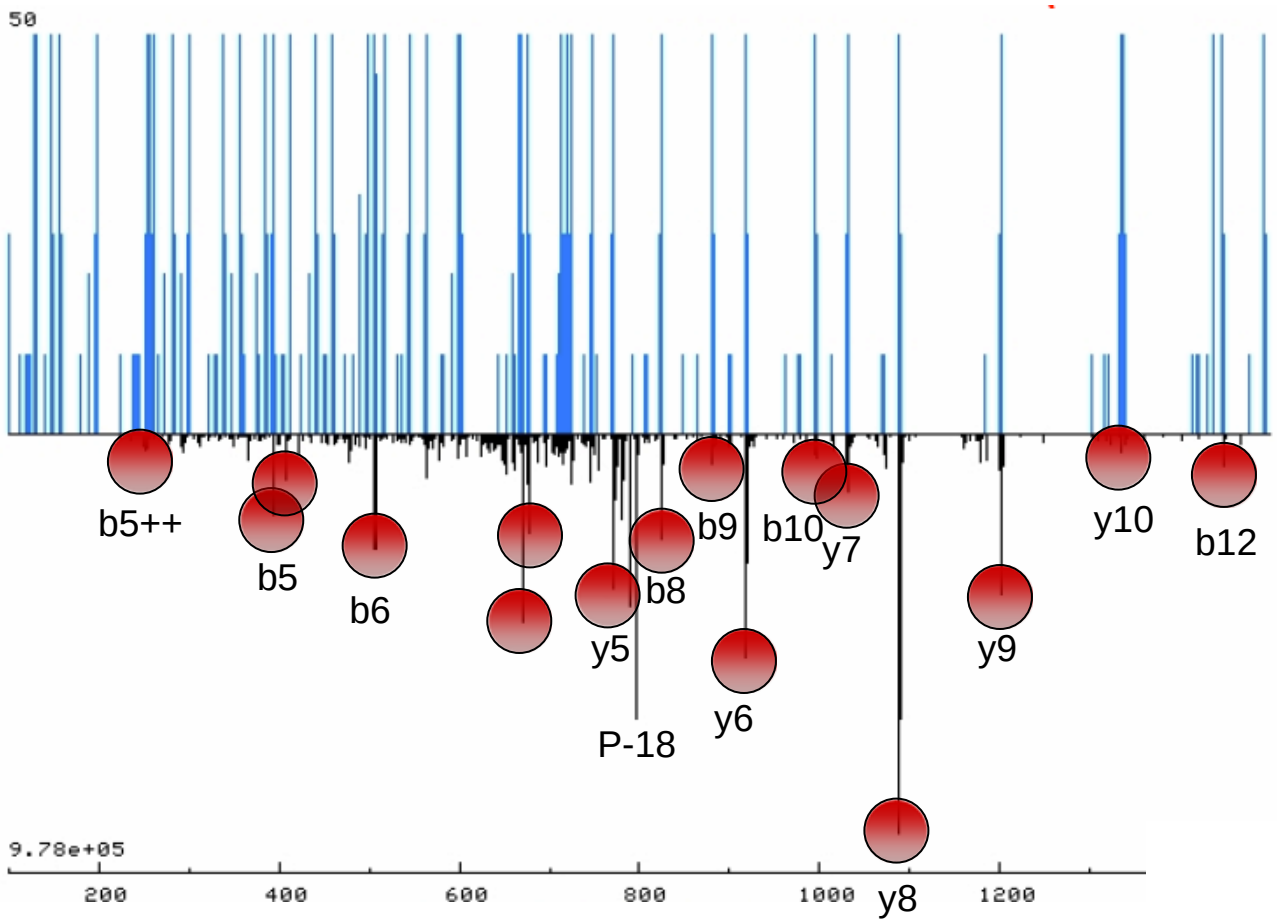
Peptide Identification – ‘Current Practice’

1. *In silico* digestion protein sequences
2. Predict sequence peaks (*b* and *y* ions)
3. Predict peaks due to minor isotopes and neutral losses
4. Normalize intensities
5. Find peaks in common within tolerance constraints (precursor & fragment)
6. Weight and score the correlation

Peptide Identification – ‘Current Practice’

Predicted spectrum

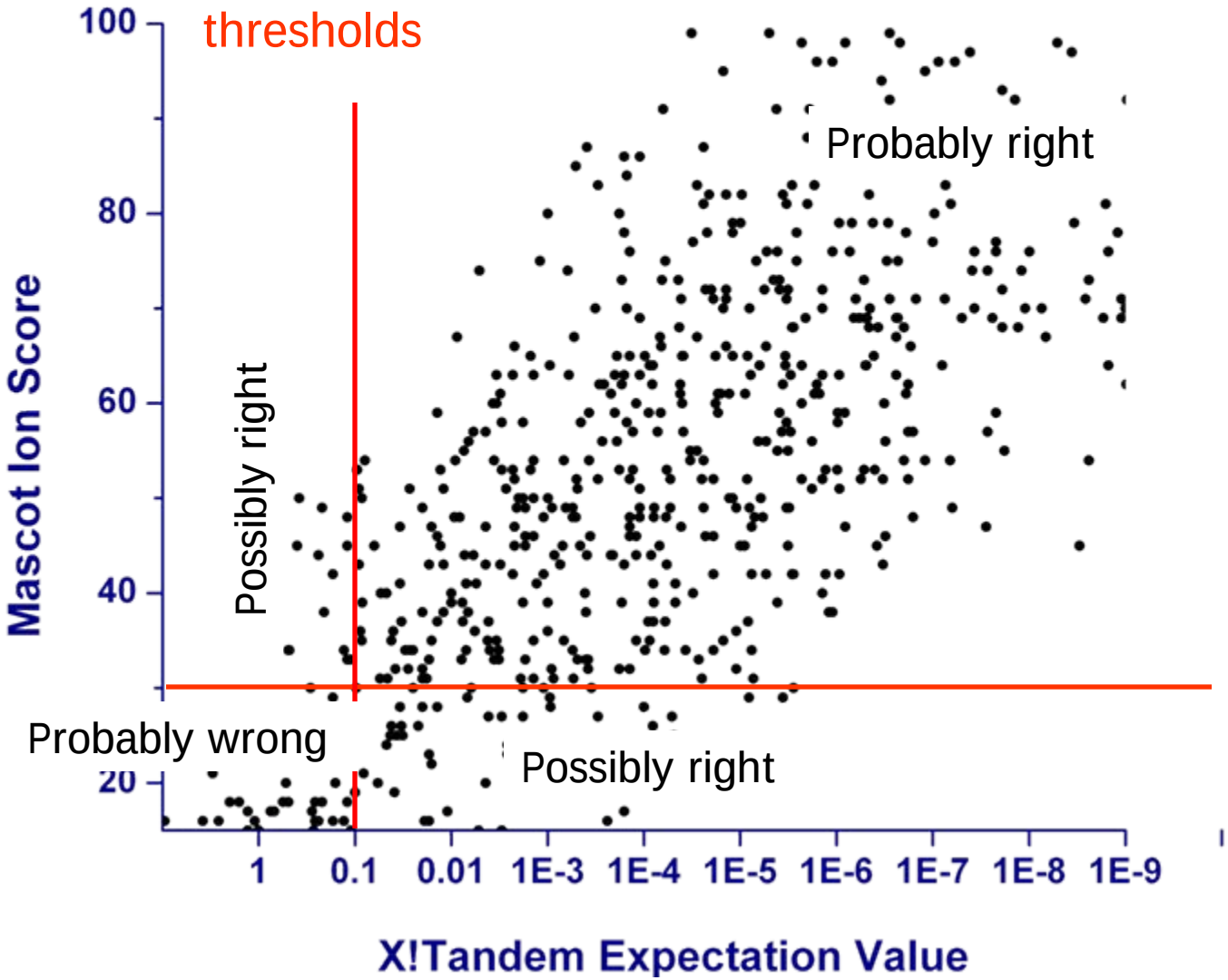
Measured spectrum



Good m/z matches. Not good predictions of relative intensities. Misses some major features.

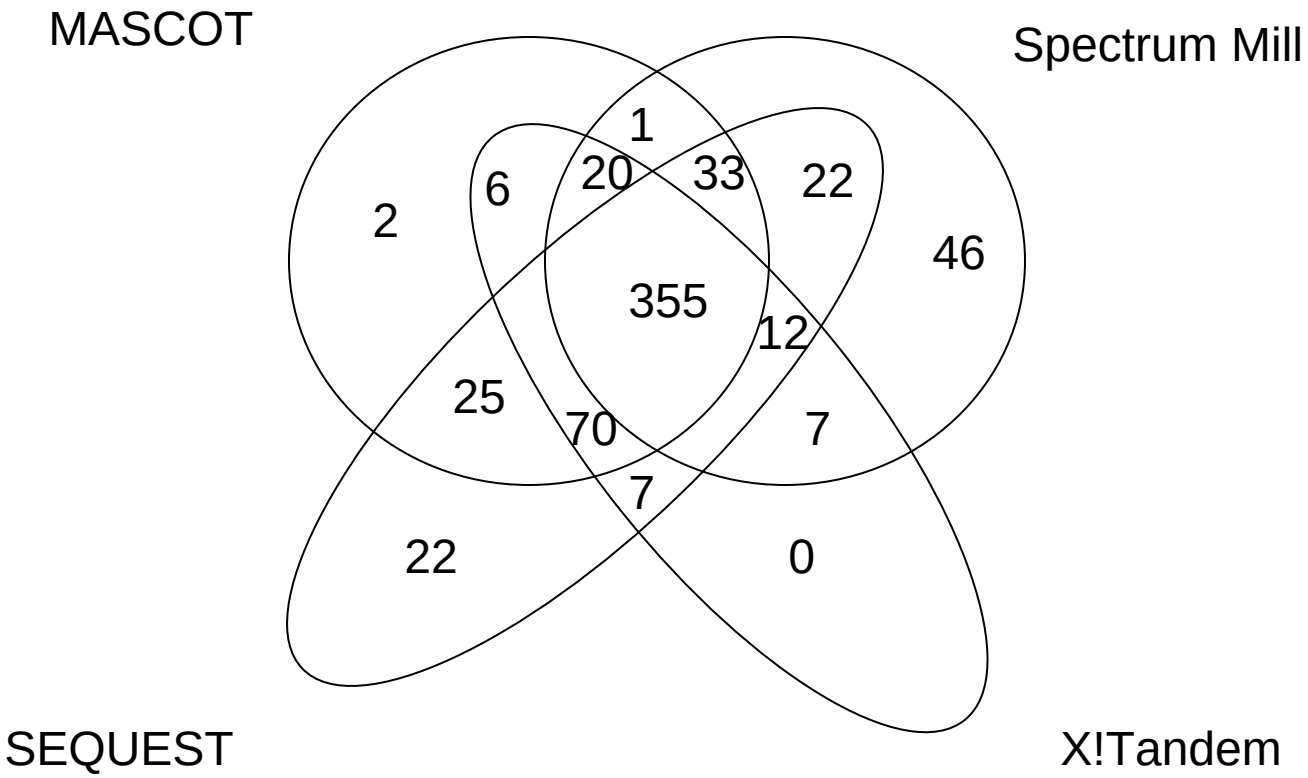
(Henry Lam, ISB)

Same Spectra - Different Scores



All Sequence Search Algorithms Report Different Matches

Kapp et al. *Proteomics*, 2005



Major algorithms are operating differently – 55% 4-way overlap

Drawbacks of ‘current practice’

- Unnatural use of peak intensities
- Little or no use of recurring, non-(b,y) fragmentation patterns, also higher charge states not accurate
- Scores for identical spectra vary between algorithms
- Time-consuming to compare results between algorithms
- Multiple commercial and ‘in-house’ data analysis systems create the appearance of variability – lack of ‘standardization’

MS Library Searching: a “New” Alternative for Peptide Ion Identification

- H.S. Hertz, R.A. Hites and K. Biemann *Anal. Chem.* **43** (1971)
- F.W. McLafferty, R.H. Hertel and R.D. Villwock *Org. Mass Spectrom.* **9** (1974)
- S. Sokolow, J. Karnofsky and P. Gustafson , The Finnigan Library Search Program. *Finnigan Application Report 2*, Finnigan Corp., San Jose, CA (March 1978).
- S.E. Stein, D.R. Scott J. Amer. Soc. Mass Spectrom., **5** (1994).

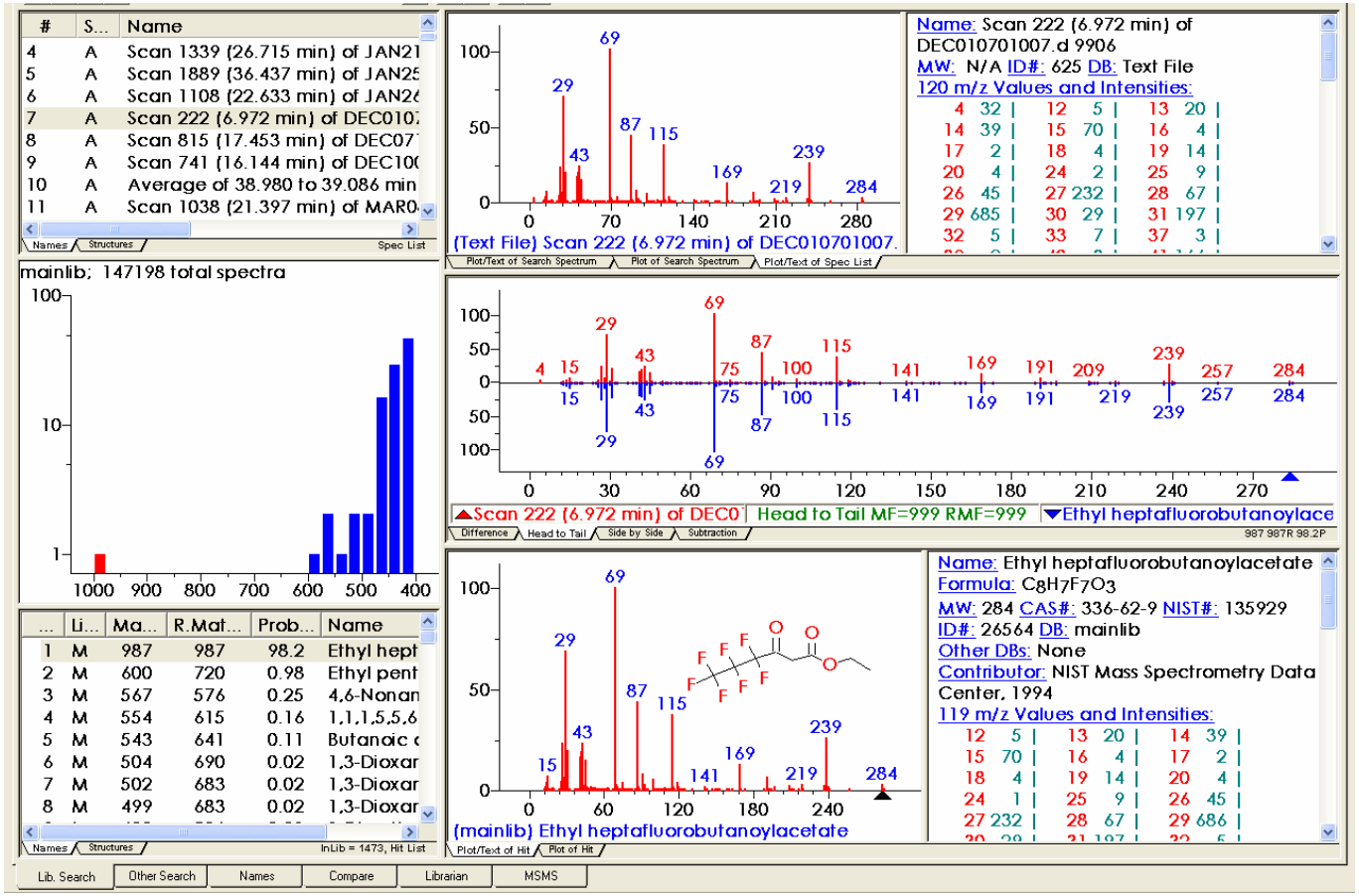
Traditional GC/MS Library Search

Uses a weighted 'dot product' score to quickly evaluate EI spectra – in use for >25 yrs.

$$\frac{\sum \sqrt{MR}}{\sqrt{\sum M \sum R}}$$

(M) Measured
(R) Reference

Cosine of angle between spectra expressed as vectors, weighted to improve performance



NIST/EPA/NIH - 163K Compounds

Compound Identification by GC/MS

But will it work for peptides?

- There are too many biologically relevant peptides! How will you purify and get spectra from them all?
  - Collect spectra from analyses of mixtures (i.e.. ‘naturally-occuring’ sources)
  - Collect spectra from public donations (thank you!)
  - Collect spectra from complex mixtures and synthetic peptides run at NIST
- Spectra are too variable
  - Not with modern instruments
  - Variability is function of S/N
- How do I know the spectra are correct?
  - Identified by up to 4 different algorithms
  - Stringent quality filters, S/N, fraction unexplained abundance
  - Many replicates used to generate consensus, minimizes run-to-run variability
- Aren’t searches limited to the peptides that are in the library (i.e. ones that have been seen before)?
  - Yes.
  - Breadth and depth of library only increased by processing new spectra

Reference Library Building*

- Extract identified spectra from sequence search
  - Multiple search engines
  - Instrument-class specific
  - Use reversed sequence library
- Create ‘consensus’ spectra
  - Two or more matching spectra, also save best
- Assign probability of being correct
  - Refine confidence
- Create searchable spectral library
  - Resolve conflicts, add annotation

* http://chemdata.nist.gov/mass-spc/ftp/mass-spc/PepLib.pdf

‘msp’- formatted library entry

Name: ALTDFERTYMK/2
MW: 923 ID#: 270 DB: yeast_consensus_final_true_lib
Mods: 1/10,M,Oxidation

Protein

Protein: gi|6322847|ref|NP_012920.1| N-terminally acetylated protein component of the large (60S) ribosomal subunit, nearly identical to Rpl14Bp and has similarity to rat L14 ribosomal protein; rpl14a csh5 double null mutant exhibits synthetic slow growth [Saccharomyces cerevisiae]; gi|300233|gb|AAC60550.1| orf YKL153 [Saccharomyces cerevisiae]; gi|320756|pir||S30133 ribosomal protein L14.e.A, cytosolic - yeast (Saccharomyces cerevisiae); gi|485979|emb|CAA81839.1| RPL14A [Saccharomyces cerevisiae]; gi|549656|sp|P36105|R14A_YEAST 60S ribosomal protein L14-A

SE scores

Comment: Spec=Consensus Pep=Tryptic Fullname=R.ALTDFERTYMK.F/2 Mods=0 Parent=461.736 Inst=it Mz_diff=0.184 Mz_exact=461.7356 Mz_av=462.011 Se=4^M13:sc=61.88/4.489,td=58.46/8.338,sr=16.68/3.719,sd=32.8/3.959,bs=23.8,b2=22.79,bd=69.63^X21:ex=0.0021/0.004377,td=6900/1.467e+004,sd=0/0,hs=30.9/0.1857,bs=0.00031,b2=0.00036,bd=45200^O21:ex=0.000222/0.02367,td=225000/2.262e+005,pr=4.48e-008/4.357e-006,bs=4.84e-005,b2=6.41e-005,bd=1.03e+006^P3:sc=24/0.5667,dc=15.1/0.4,ps=3.01/0.04667,bs=0

Sources

Sample=21/cbs00174_001_01_cam,1,2/cbs00174_001_02_cam,0,2/cbs00174_01_06_cam,0,2/cbs00174_01_07_cam,1,1/cbs00174_01_10_cam,0,2/cbs00175_02_01_cam,2,8/cbs00175_02_02_cam,1,2/cbs00175_02_05_cam,2,4/cbs00175_02_06_cam,0,6/silac_labelled,0,1/yeast_apaulovich_cam,1,7/yeast_comp12vs12sizefrac_cam,0,1/yeast_comp12vs12standscx_cam,0,1/yeast_gygi_cam,1,3/yeast_haynes_cam,0,5/yeast_lisa3_cam,12,18/yeast_lnt_cam,0,1/yeast_opd34_none,0,1/yeast_opd38_none,0,2/yeast_opd41_none,0,2/yeast_opd42_none,0,1 Nreps=21/107 Missing=0.0260/0.0370 Parent_med=461.94/0.10 Max2med_orig=1193.0/374.2 Dotfull=0.940/0.021 Dot_cons=0.972/0.022 Unassign_all=0.032 Unassigned=0.019 Dotbest=0.97 Flags=0,0,6 Naa=8 DUScorr=10/2/0.61 Dottheory=0.96 Pfin=3.5e+010 Probcorr=20 Tfratio=1.1e+008 Pfract=0.39 Specqual=0.0

Consensus statistics and measures

63 m/z Values and Intensities:

142.2	2.19	? 13/20 0.2
143.0	20.88	b2/-0.07 21/21 0.8
175.1	17.38	y1/-0.02 21/21 0.3
183.2	6.79	? 20/21 0.1
187.2	3.79	? 21/20 0.2
197.2	1.79	? 12/20 0.1
199.1	1.49	? 12/19 0.0
211.1	61.73	? 21/21 1.1
215.1	14.18	b5-44^2/0.48 21/21 0.7
227.6	2.29	b5-18^2/-0.02 17/20 0.3
244.1	2.89	? 19/20 0.2
256.1	46.75	b3/-0.06 21/21 1.2

[…]

Annotated spectrum
m/z, normalized intensities

steve.stein@nist.gov

NCBI Entrez Protein

My NCBI [Sign In] [Register]

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM | Books

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display GenPept Show 5 Send to

Range: from begin to end Features: CDD + Refresh

1: AAA39883. Reports p53...[gi:200203]

BLink, Conserved Domains, Links

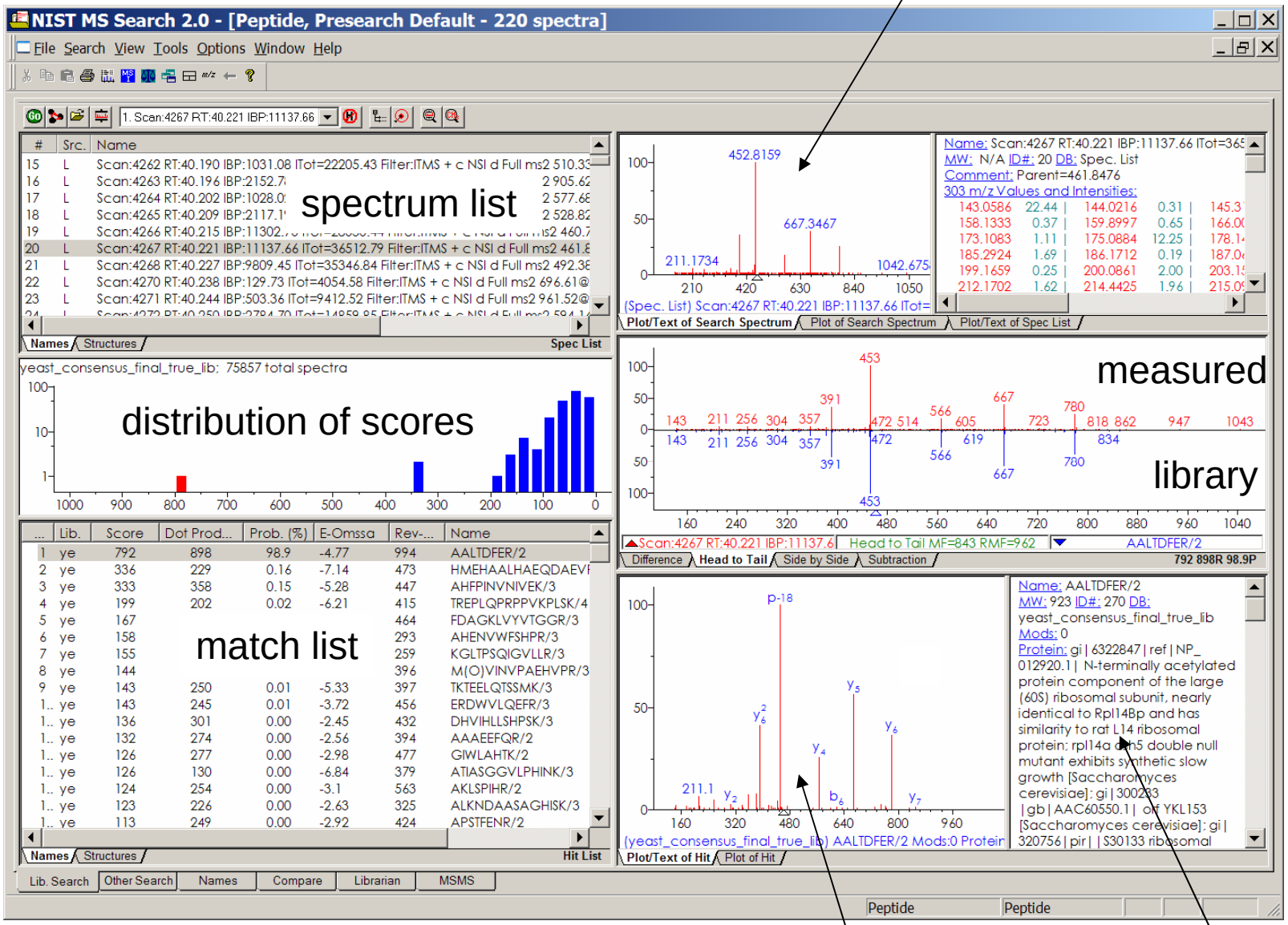
Features Sequence

```
LOCUS       AAA39883                 381 aa            linear   ROD 27-APR-1993
DEFINITION  p53.
ACCESSION   AAA39883
VERSION     AAA39883.1  GI:200203
DBSOURCE    locus MUSP53C accession M13874.1
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (residues 1 to 381)
  AUTHORS   Arai,N., Nomura,D., Yokota,K., Wolf,D., Brill,E., Shohat,O. and
            Rotter,V.
  TITLE     Immunologically distinct p53 molecules generated by alternative
            splicing
  JOURNAL   Mol. Cell. Biol. 6 (9), 3232-3239 (1986)
   PUBMED   3023970
REFERENCE   2  (sites)
  AUTHORS   Han,K.A. and Kulesz-Martin,M.F.
  TITLE     Alternatively spliced p53 RNA in transformed and normal cells of
            different tissue types
  JOURNAL   Nucleic Acids Res. 20 (8), 1979-1981 (1992)
   PUBMED   1579500
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..381
                     /organism="Mus musculus"
                     /db_xref="taxon:10090"
                     /map="11"
                     /cell_line="normal T-cell line L12"
                     /cell_type="Ab-MuLV transformed Meth A fibroblast"
     Protein         1..381
                     /name="p53"
     Region          8..31
                     /region_name="P53_TAD"
                     /note="P53 transactivation domain (TAD). The binding of
                     the p53 transactivation domain by regulatory proteins
                     regulates p53 transcription activation. This domain is
                     comprised of a single amphipathic alpha helix and contains
                     a highly conserved motif; pfam08563"
                     /db_xref="CDD:71989"
```

Not that different than other types of molecular biology data records

NIST MS Search 2.0 – Peptide Version (free demo)

measured spectrum



library spectrum

annotation

http://www.peptideatlas.org/speclib/

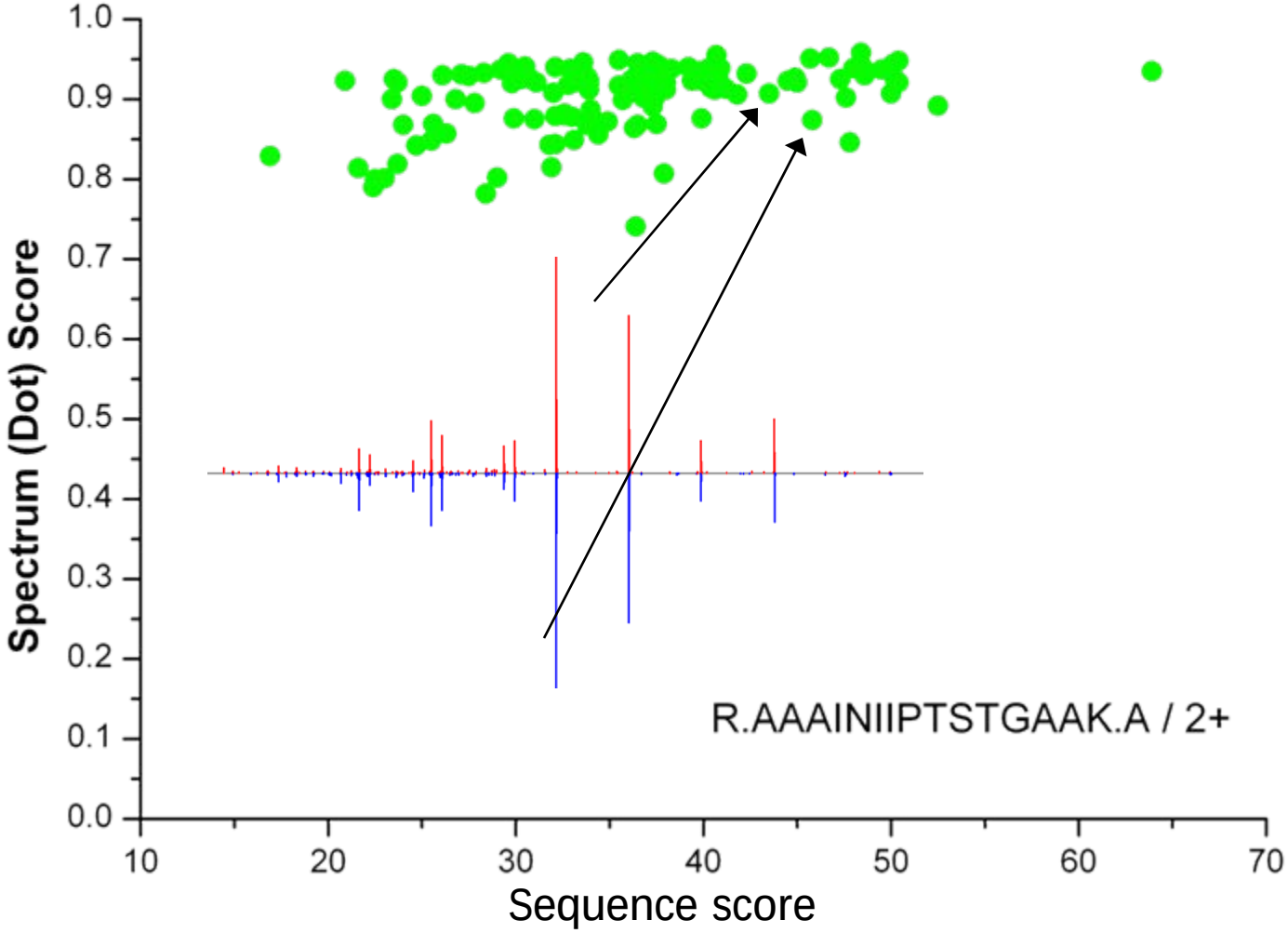
To download software & libraries

Library Searching for Peptides

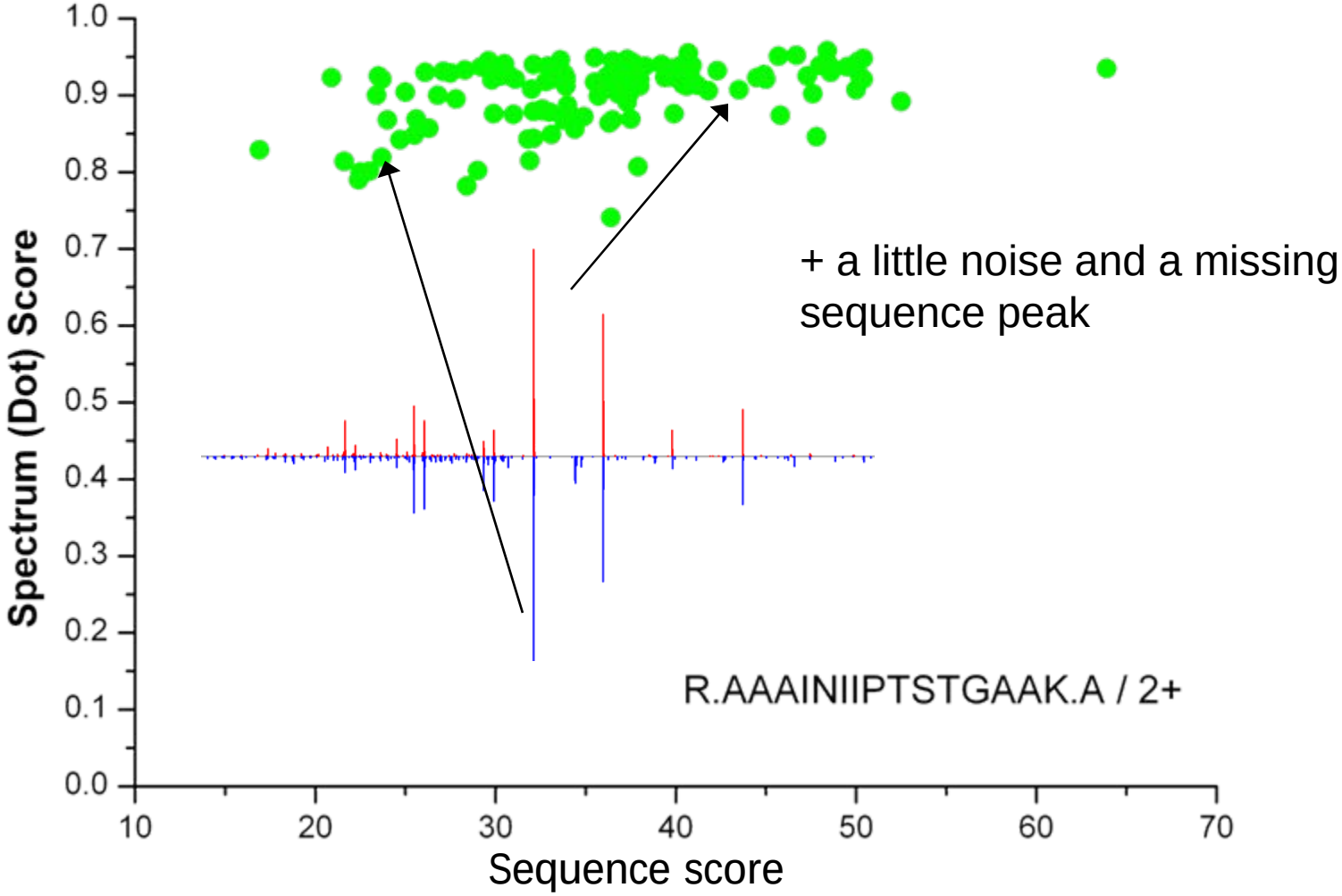
- LIBQUEST (Yates)
  - Yates et al, Anal. Chem., 1998, 70, 3557
- X!Hunter (Beavis)
  - Craig et al, J. Proteome Res., 2006, 5, 1843
- BiblioSpec (MacCoss)
  - Frewen et al., Anal. Chem. 2006, 78, 5678
- Spectral Comparison (Kearney)
  - Liu et al, Proteome Science 2007, 5:3
- SpectraST (Aebersold)
  - Lam et al., *Proteomics* 2007 6, 655-667
  - Lam et al., *in press* (SpectraST tool for building custom libs.)

- NIST Peptide Ion Fragmentation Library
  - June 2006 1st release (US-HUPO – March 2004)

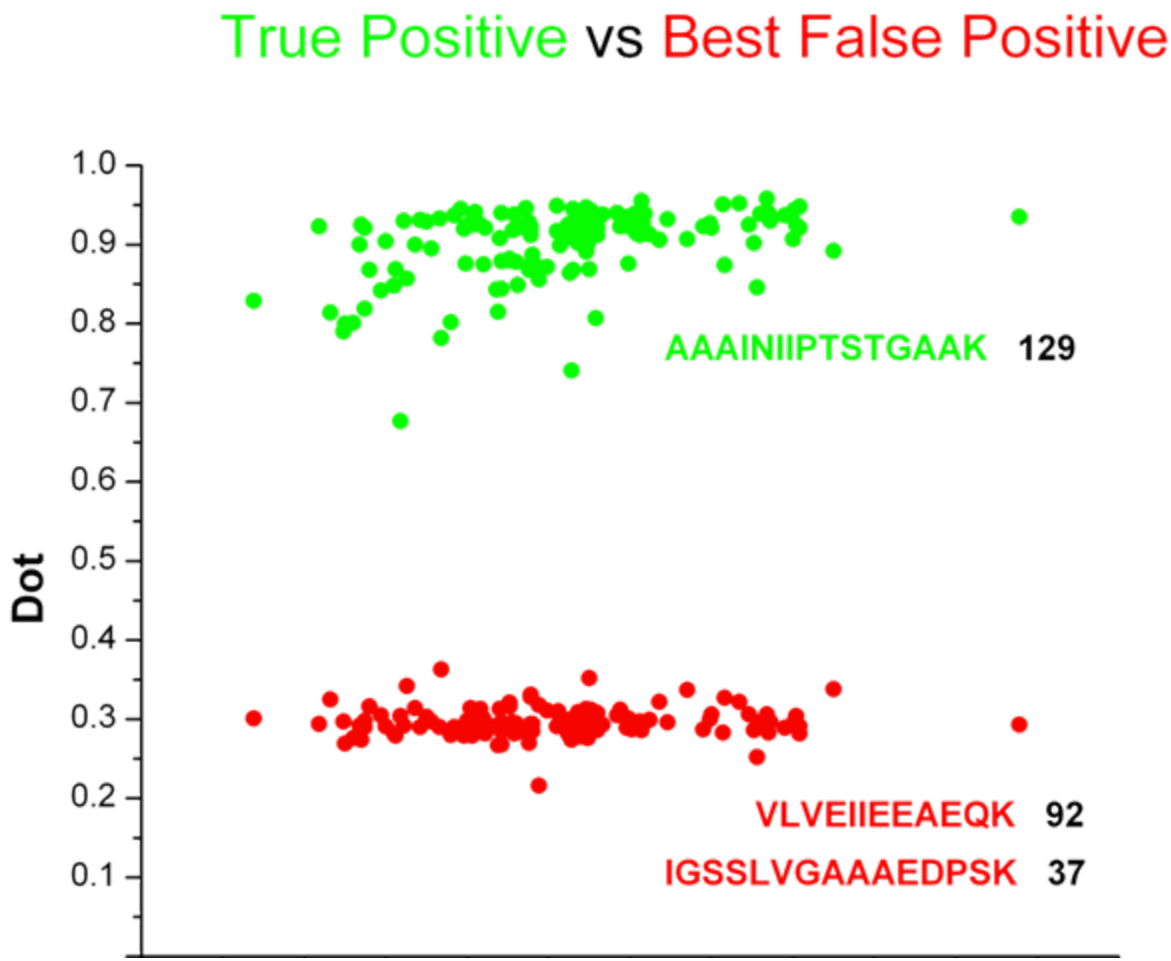
Spectrum/Sequence Scores Vary More than Spectrum/Spectrum Scores



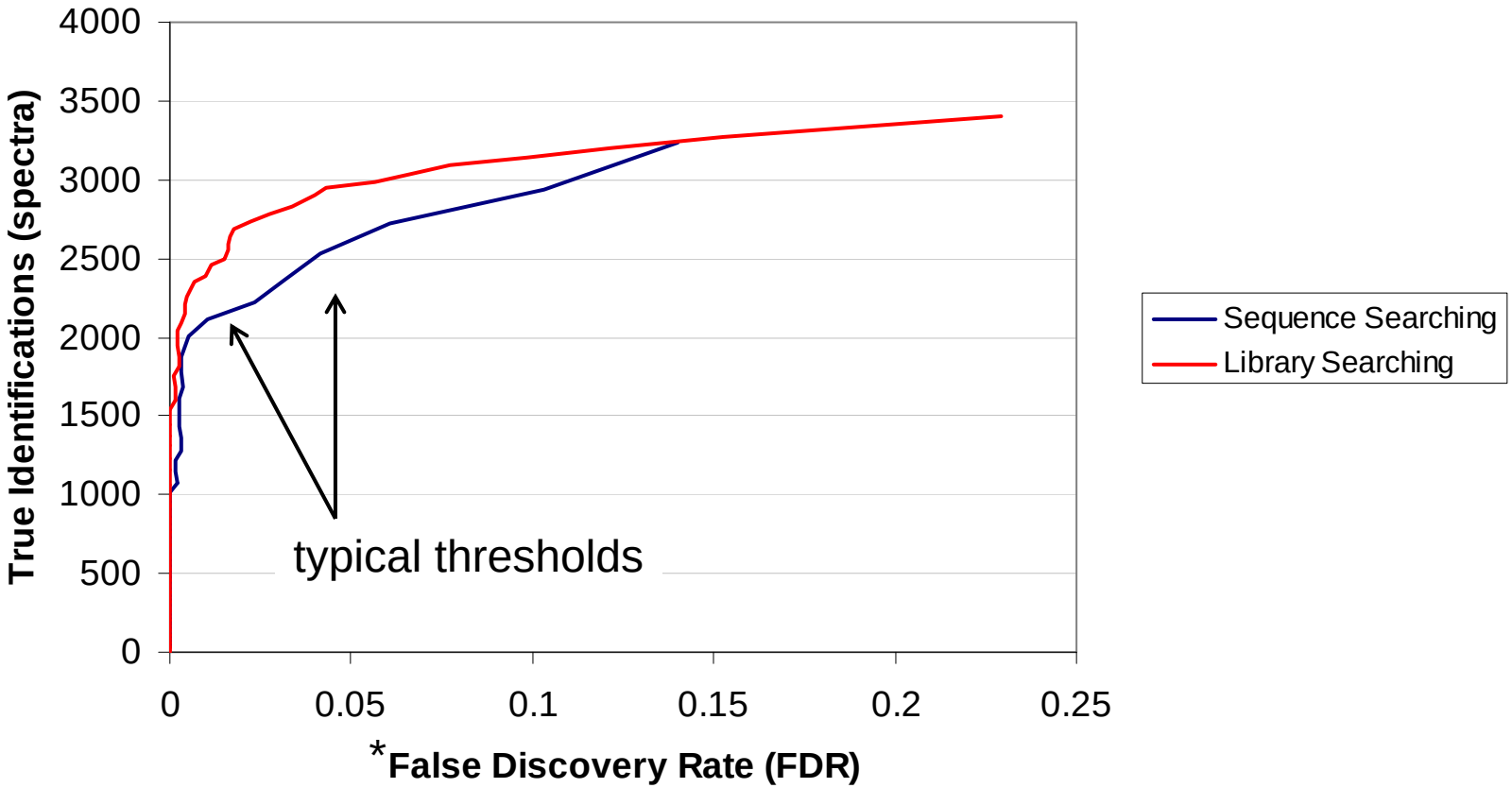
Spectrum/Sequence Scores Vary More than Spectrum/Spectrum Scores



Spectrum/Spectrum Score Separates True and False IDs



A Real Example: Spectral Library Searching Identifies More Peptides



Biological Source: yeast lysate
Separation: 1D RPLC
Instrument: Thermo LTQ (VUMC)
Spectra: 10,227 (not used in library construction)
Sequence Library: (Yeast SGD + reversed sequences)
Spectral Library: NIST Yeast Lib 2007 (78,450 entries)
Decoy Library: *Drosophila melanogaster* (65,503 entries)

*FDR=2F/(F+T) [integral]
'F' does **not** include 'homologous' or identical forward/decoy matches

Comparison of search space: NIST human lib vs. Human IPI sequence database (~50,000 proteins)

Criterion		Search space	
Sequence characteristics	Precursor m/z range	Sequence Search	Spectral Search
Tryptic, with no missed internal cleavage, +2/+3	300 – 2000	**3.5×10^5**	**1.1×10^4**
	800 – 803	**7.2×10^2**	**52**
At least semi-tryptic, with at most 1 missed internal cleavage, +2/+3	300 – 2000	**1.5×10^7**	**1.7×10^4**
	800 – 803	**3.6×10^4**	**77**

leads to a ~500X reduction in CPU time

(Henry Lam, ISB)

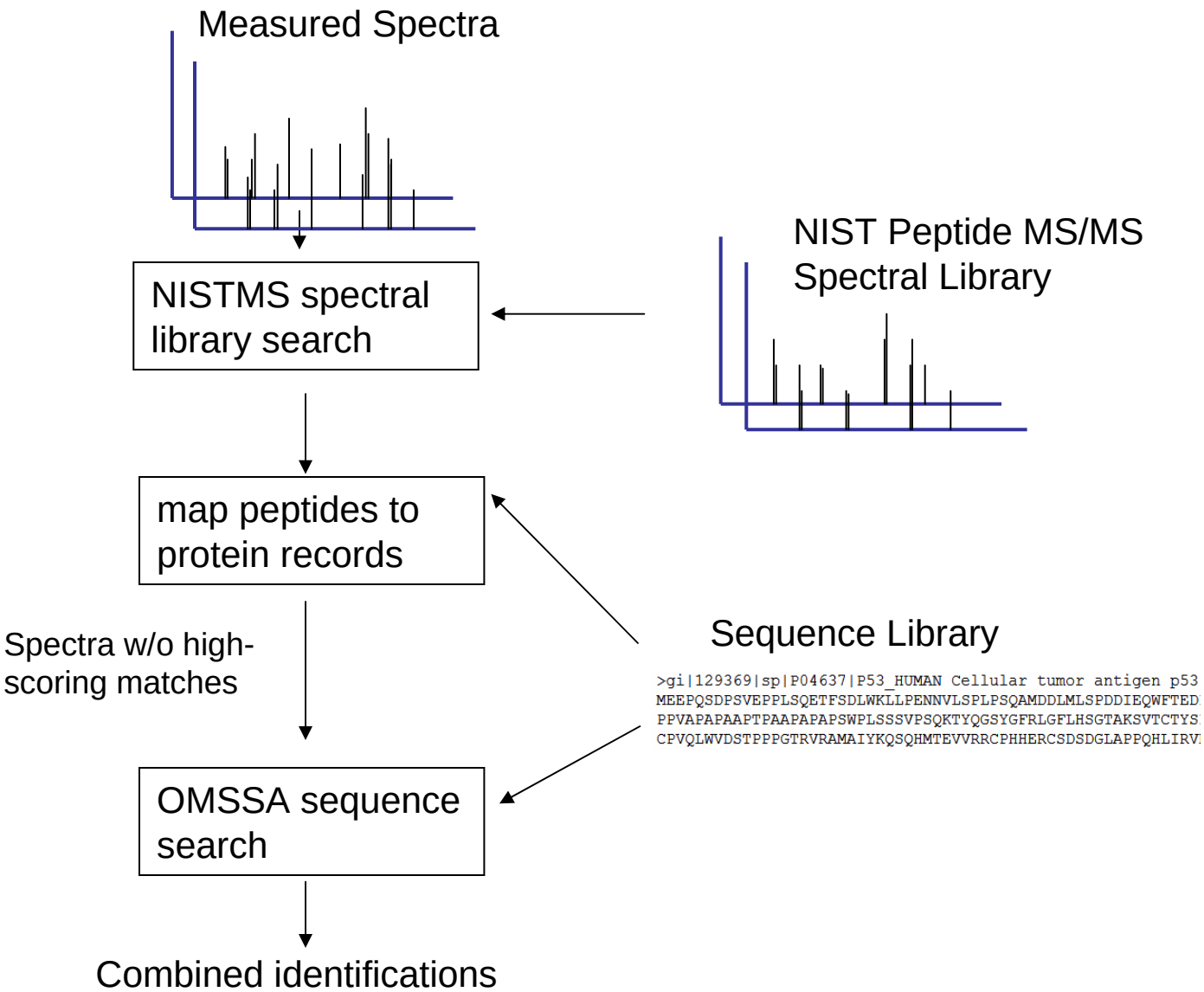
Advantages of Spectral Library Searching

1. *Higher confidence matches*
   - Each entry has been identified by up to 4 algorithms
   - Quality measures ‘built-in’
2. *Scoring is more precise*
   - Natural use of peak intensities
   - Makes full use all annotated peaks (e.g. recurring, non-sequence peaks)
3. *Scoring is more robust*
   - Small changes in spectra give small changes in score
   - Better separation of true and false matches
   - Reduces false positives and false negatives
4. *Faster*
   - Up to 500X faster, search space is limited to technologically-identifiable peptides (i.e.. *m/z* 300-2,000)
5. Bonus: Spectral library searching can also match ‘unidentified’ consensus spectra for later interpretation (e.g. SNPs, extensive known/unknown PTMs)

Integrating Sequence Searching with Spectral Library Searching

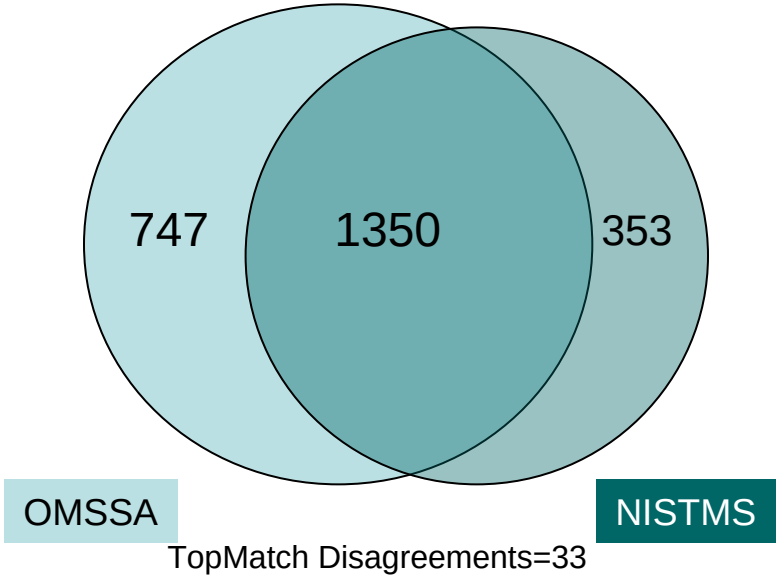
- Problem: Most missed IDs are because the peptide ions are not represented in the library
- Solution: Create a combined search program for matching sequences when no spec lib match is found
- Components
  - NISTMS, 'NIST MS Search 2.0' handling and scoring routines
  - OMSSA (Lewis Geer/NCBI), probability-based, sequence search engine
  - 'Pep2Prot' matches spec lib peptides to proteins for reporting
  - Wrapper script
  - Tools for handling/parsing output
  - Methods for calculating false discovery rate
  - Browser

OMSSA/NISTMS Workflow

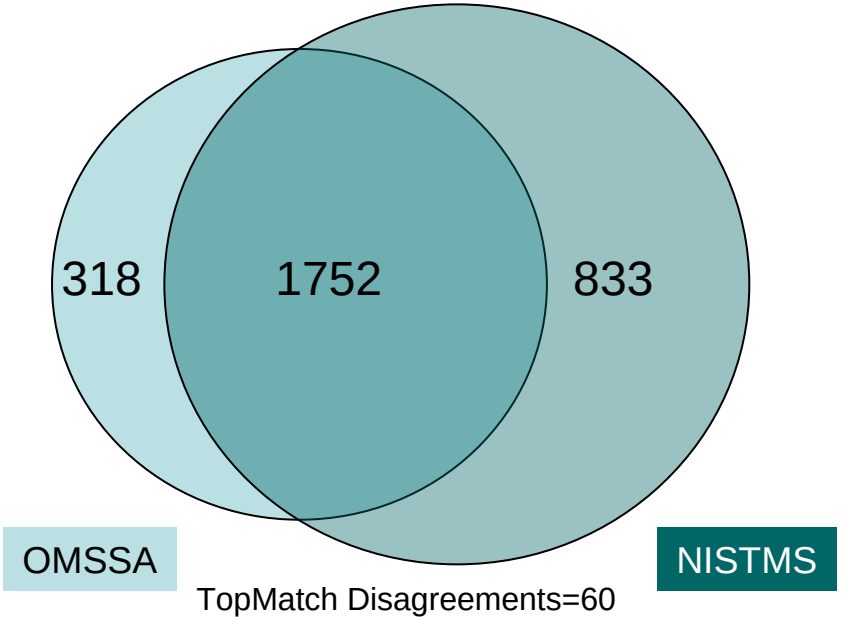


Separate & Combined Performance

NIST Yeast Lib. (06/06)



NIST Yeast Lib. (05/07)



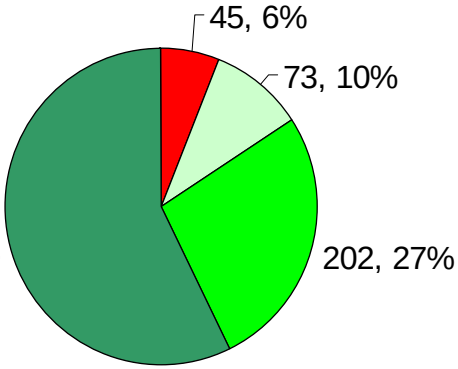
10,227 LTQ Spectra (1% FDR)	Spectra in Library	Spectral Counts				Unique Peptides	
		OMSSA	NISTMS	Union	Gain over OMSSA alone	OMSSA	NISTMS
NIST Yeast Peptide Fragmentation Lib. 06/06	34,426	2,130	1,736	2,450	+15%	1,677	1,375
NIST Yeast Peptide Fragmentation Lib. 07/07	78,450	2,130	2,645	2,903	+36%	1,677	1,974

Bonus OMSSA Matches

Green = ‘not in spec lib’

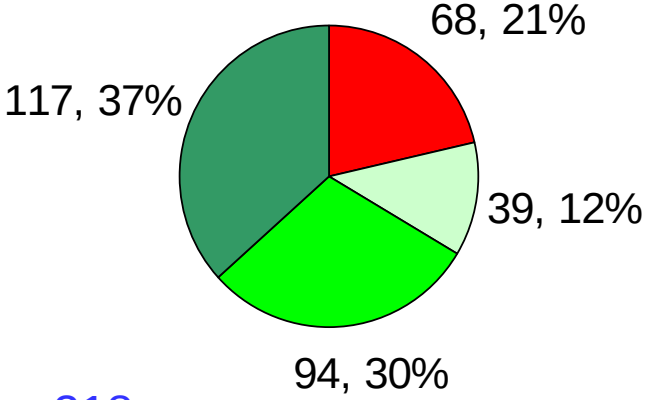
Red = below NISTMS threshold or no match

2006



747

2007



318

- In spec. lib. same charge/mods
- In spec. lib. same charge, diff. mods
- In spec. lib. same sequence, diff. charge
- Not in spec. lib.

OMSSA/NISTMS – ‘pipeline’

- Accepts common input formats (dta, mgf, pkl, etc.)
- NISTMS will run as a pre-search against any ‘.msp’- formatted NIST reference library
- All NISTMS scores used by MS Search 2.0 included in output
- OMSSA will run as an iteration with optional ‘E-value’ cutoffs for spectra and/or protein sequences to include
- Either an NISTMS or an OMSSA match is returned
- Peptide matches from NISTMS mapped to proteins
- Output is one OMSSA search result file (XML or ASN.1)
- Output will be viewable in the OMSSA Browser
- OMSSA available now- http://pubchem.ncbi.nlm.nih.gov/omssa/
- When finished (2007), will be freely available for testing (paul.rudnick@nist.gov).

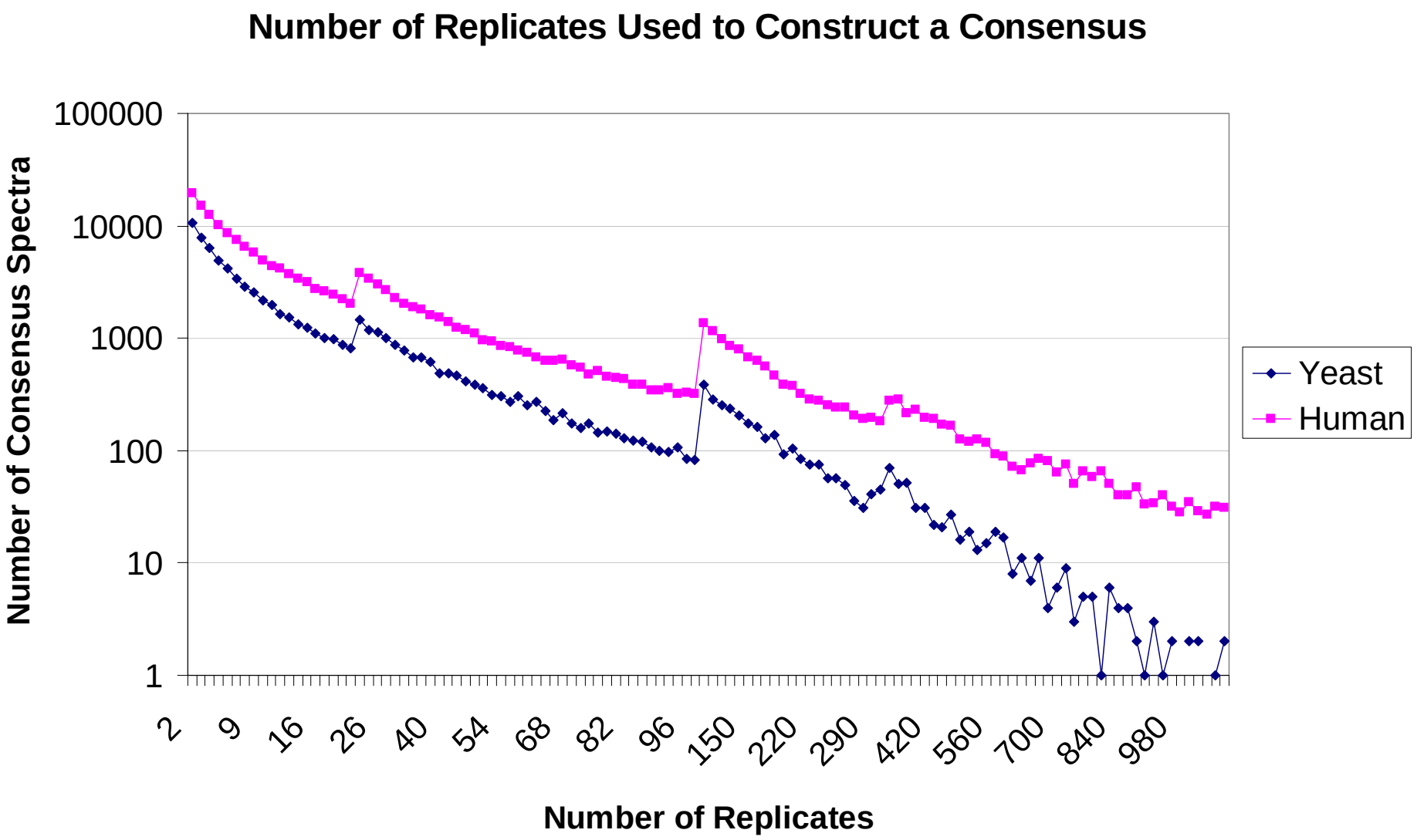
NIST Peptide Fragmentation Libraries: Status Update

Ion trap

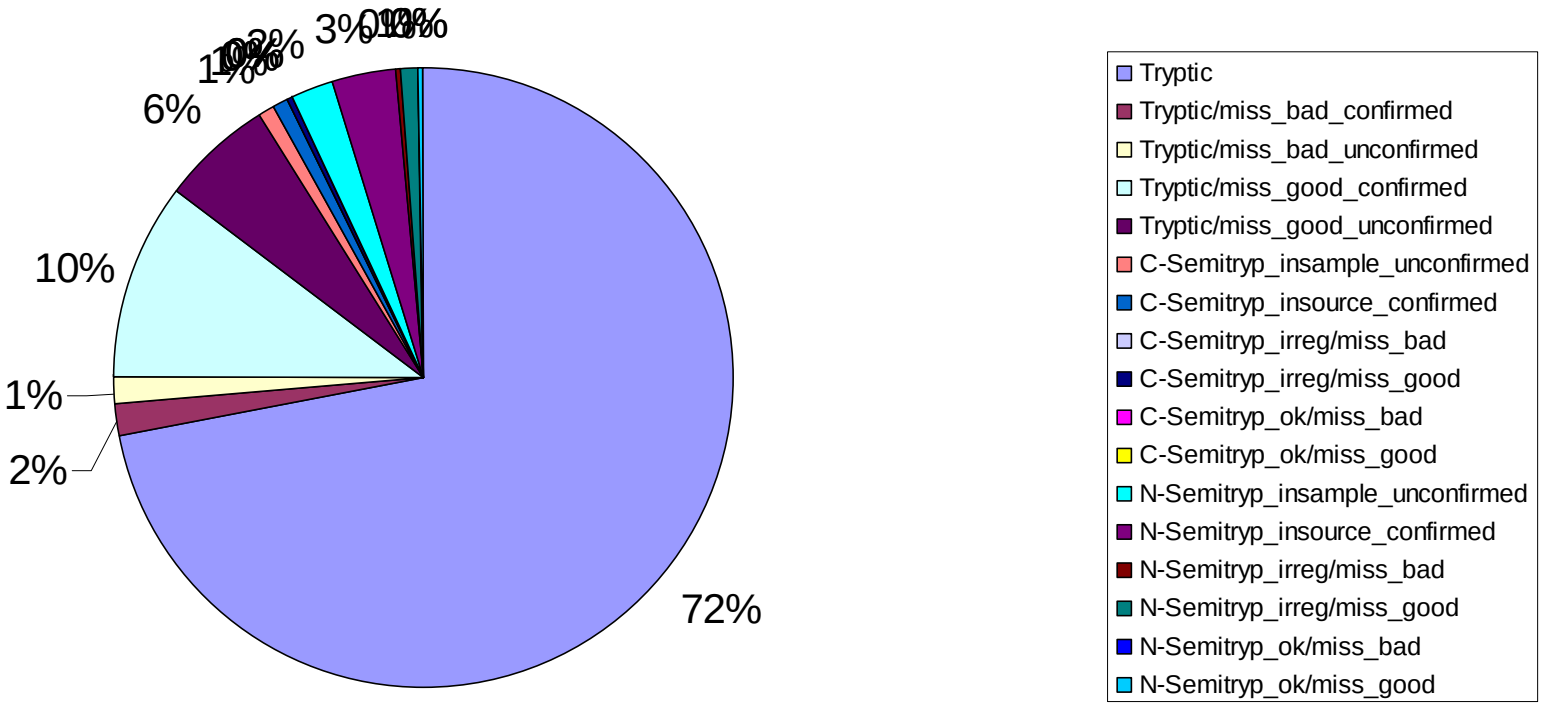
2007 Library	*Increase over '06*	*Approx. Coverage (simple inference)*	*Datasets used*	*Peptide Identifications*	*Consensus spectra*	*Different sequences*	*+1*	*+2*	*+3*	*+4*
Human	4.1x	13%	274	17,779,962	179,857	112,945	15,802	107,700	50,891	5,464
Yeast	2.2x	20%	59	3,535,532	76,044	46,819	5,569	46,125	23,206	1,144

Both new libraries set for release (09/07).

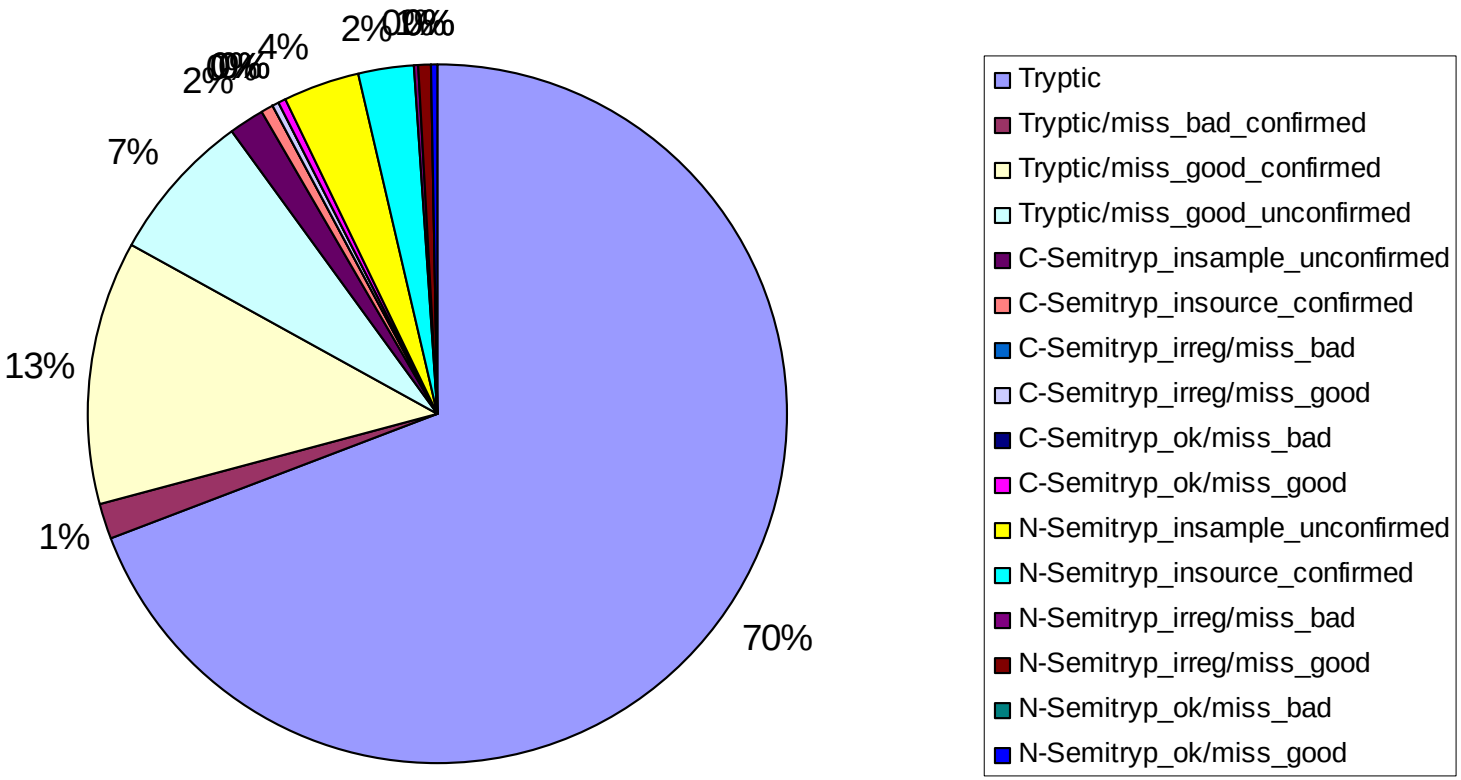
http://www.peptideatlas.org/speclib/


Number of Replicates Used to Construct a Consensus
Number of Consensus Spectra
100000
10000
1000
100
10
1
2
9
16
26
40
54
68
82
96
150
220
290
420
560
700
840
980
Number of Replicates
Yeast
Human

Human Peptides



Yeast Peptides



Next

- More data to improve releases (human > mouse > rat > yeast > fly > microbes / viruses – all instruments)
- Custom library building available by agreement
- More ‘batch’ software and integration with existing systems
- NIST’s focus is on library construction and maintenance
- Development of specialized applications (e.g. ‘real-time’ exclusion lists)

paul.rudnick@nist.gov

Acknowledgements

NIST

- Steve Stein
- Yuri Mirokhin
- Dmitrii Tchekhovskoi
- Jeri Roth
- Sara Yang
- Oleg Toropov
- Pedi Neta
- Quan-Long Pu
- Mark Wingerd
- Lisa Kilpatrick (HML)
- Qian Dong

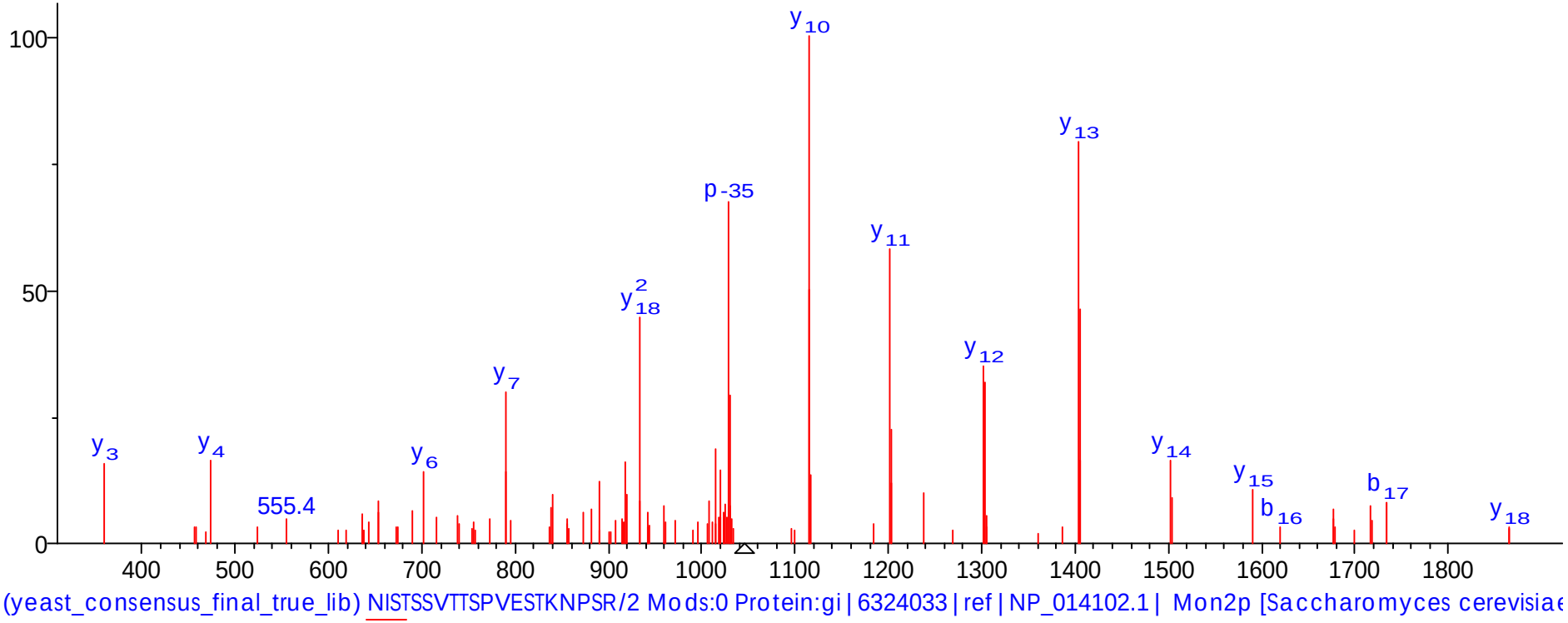
Elsewhere

- Henry Lam (ISB)
- Eric Deutsch (ISB)
- Jayson Falkner (ProteomeCommons.org)
- Ron Beavis (UBC)

- Lewis Geer (NCBI)
- Ming Xu (NCBI)

Thank YOU to everyone making public and private data donations!

End


100
50
0
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
y10
y13
p-35
y11
y2 18
y12
y7
y3
y4
y6
555.4
y14
y15
b17
b16
y18
(yeast_consensus_final_true_lib) NISTSSVTTSPVESTKNPSR/2 Mods:0 Protein:gi|6324033|ref|NP_014102.1| Mon2p [Saccharomyces cerevisiae