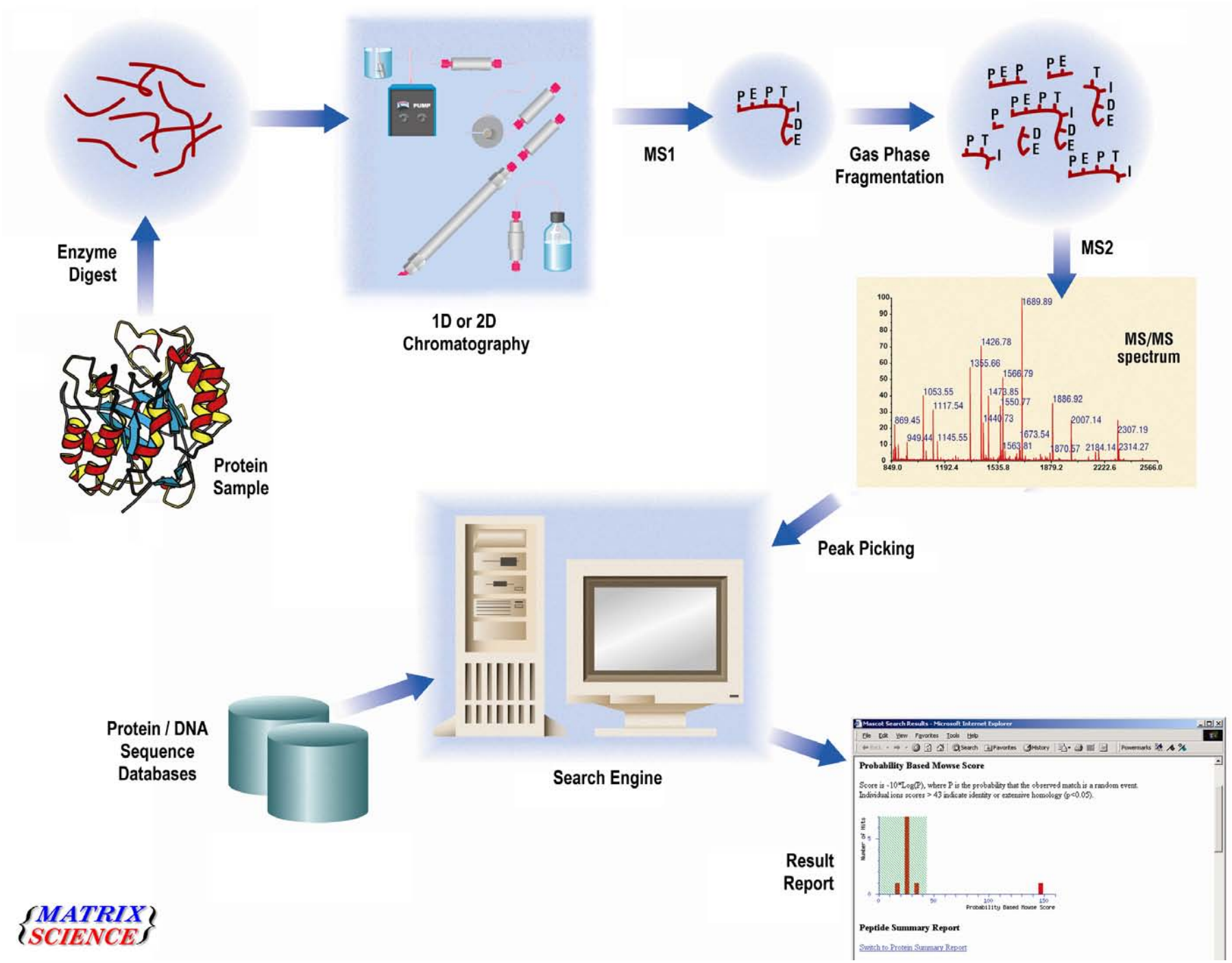


ABRF 2010:  
Introduction to Proteome Informatics Data Analysis

# Peptide Identification by Database Search

John Cottrell  
Matrix Science



Enzyme Digest

1D or 2D Chromatography

MS1

Gas Phase Fragmentation

MS2

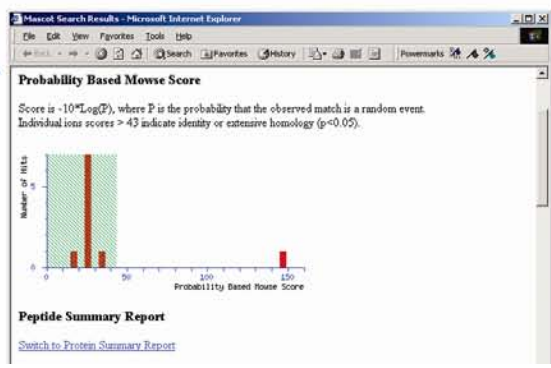
MS/MS spectrum

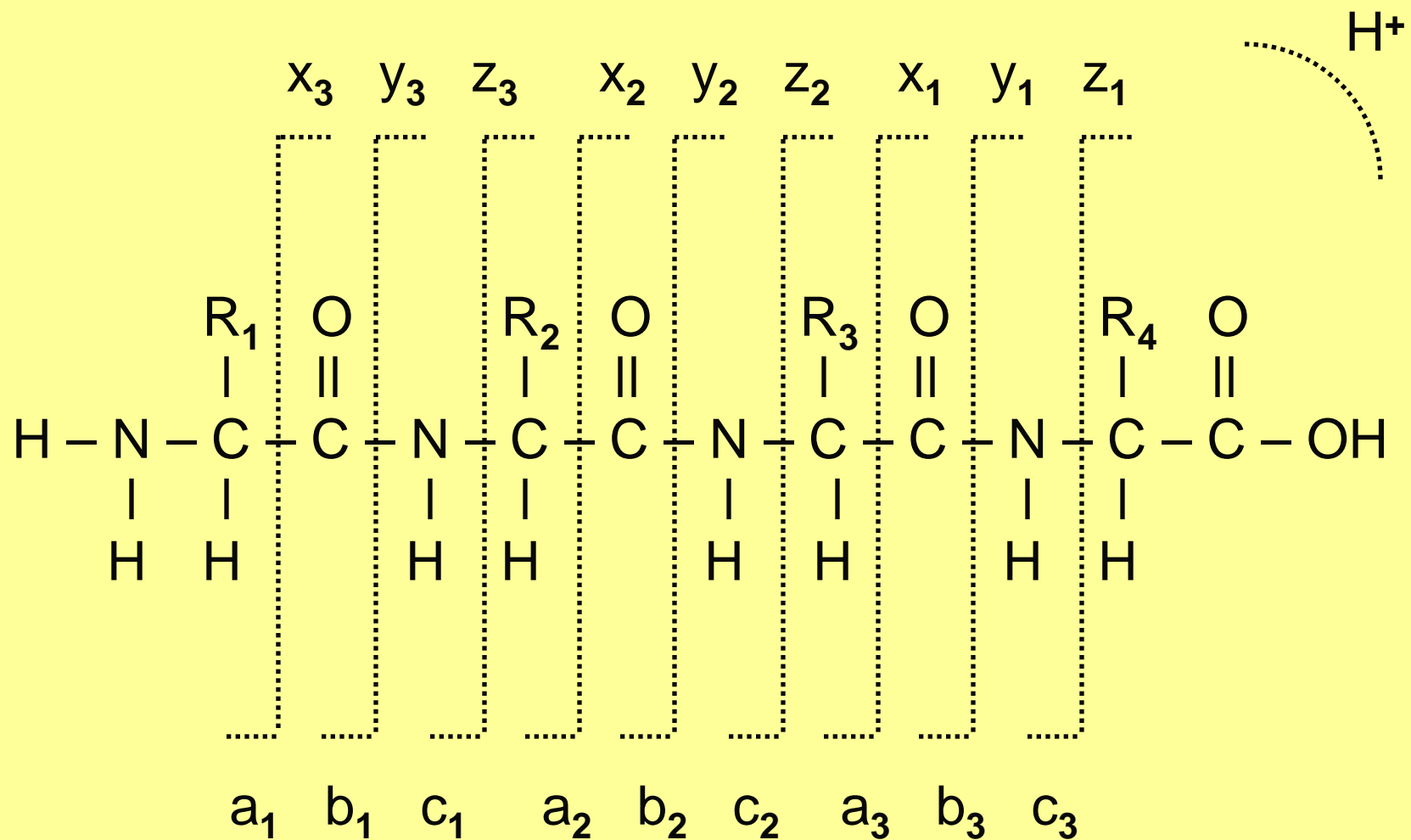
Peak Picking

Search Engine

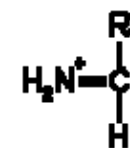
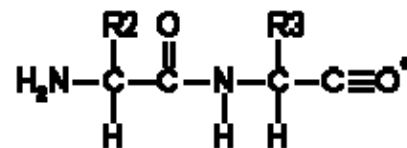
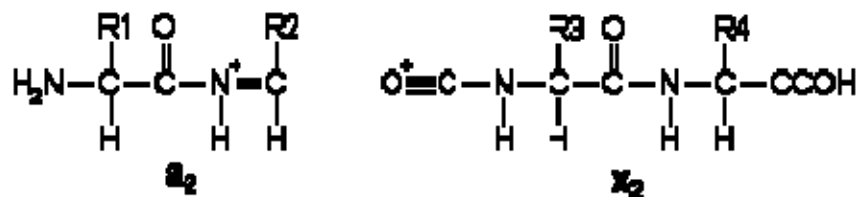
Protein / DNA Sequence Databases

Result Report



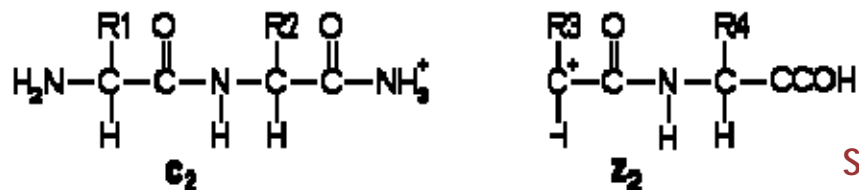
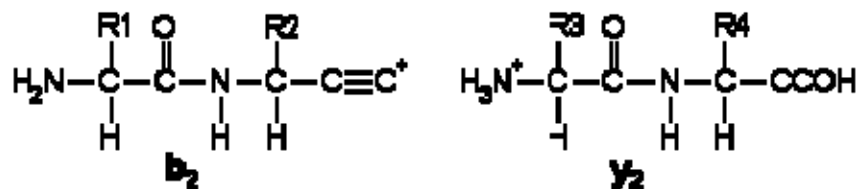


➤Roepstorff, P. and Fohlman, J. (1984). *Proposal for a common nomenclature for sequence ions in mass spectra of peptides*. Biomed Mass Spectrom 11, 601.

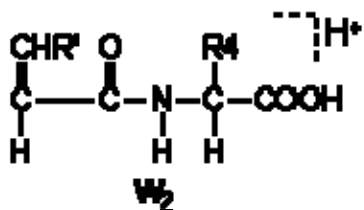
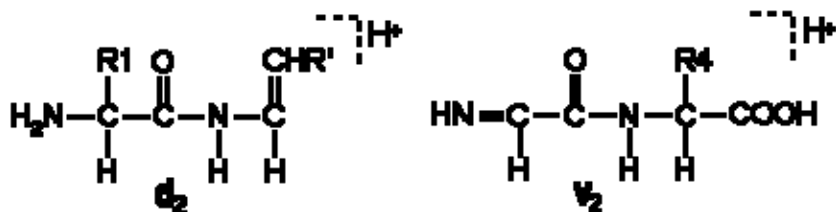


Internal

Immonium



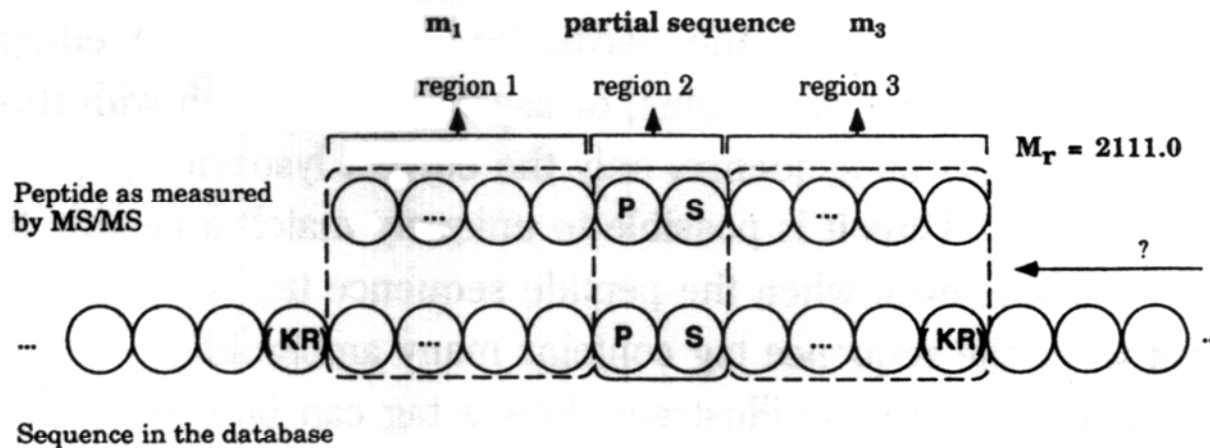
Sequence Ions



Satellite Ions

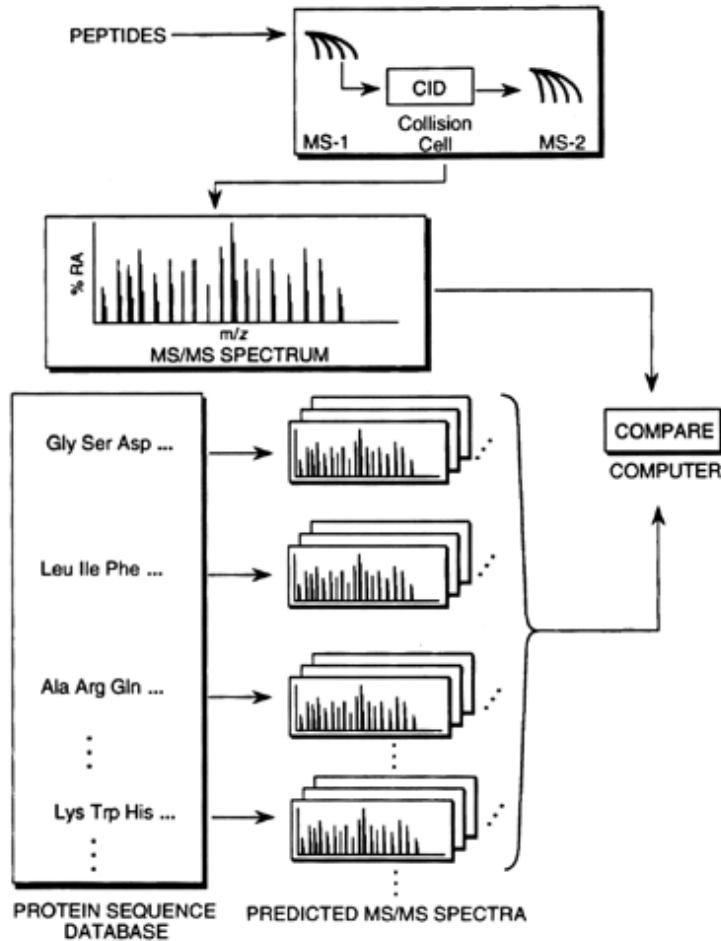
Ion Type	Neutral $M_r$
a	[N]+[M]-CHO
a*	a-NH <sub>3</sub>
a°	a-H <sub>2</sub> O
b	[N]+[M]-H
b*	b-NH <sub>3</sub>
b°	b-H <sub>2</sub> O
c	[N]+[M]+NH <sub>2</sub>
d	a - partial side chain
v	y - complete side chain
w	z - partial side chain
x	[C]+[M]+CO-H
y	[C]+[M]+H
y*	y-NH <sub>3</sub>
y°	y-H <sub>2</sub> O
z	[C]+[M]-NH <sub>2</sub>

➤ Papayannopoulos, IA, *The interpretation of collision-induced dissociation tandem mass spectra of peptides*. Mass Spectrom. Rev., 14(1) 49-73 (1995).



**Figure 1.** Principle of matching peptide sequence tags to a proposed sequence. The upper chain of amino acids represents the peptide sequence as measured by MS/MS (from Table 1 in this example), and the lower chain represents amino acids in the sequence database that the tag is compared to. Note that the partial sequence divides the peptide into three regions. The added mass  $m_1$  of the residues in region 1, together with the N-terminus, is a match criterion as is the added mass in region three,  $m_3$ . In region 2, the sequence is known. Furthermore, it can be required that the peptide obey the cleavage condition of the proteolytic enzyme, marked by KR for trypsin. The left pointing arrow indicates that both search directions may have to be considered.

➤Mann, M. and Wilm, M., *Error-tolerant identification of peptides in sequence databases by peptide sequence tags*. Anal. Chem. 66 4390-9 (1994).



**Figure 1.** Flow chart that depicts the algorithm for searching protein databases with tandem mass spectrometry data.

## SEQUEST

➤ Eng, J. K., McCormack, A. L. and Yates, J. R., 3rd., *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.* J. Am. Soc. Mass Spectrom. 5 976-89 (1994)

# Uninterpreted MS/MS Search Engines

InsPecT	<a href="http://proteomics.ucsd.edu/LiveSearch/">http://proteomics.ucsd.edu/LiveSearch/</a>
Mascot	<a href="http://www.matrixscience.com/search_form_select.html">http://www.matrixscience.com/search_form_select.html</a>
MS-Tag (Protein Prospector)	<a href="http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=mstagstandard">http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=mstagstandard</a>
Omssa	<a href="http://pubchem.ncbi.nlm.nih.gov/omssa/index.htm">http://pubchem.ncbi.nlm.nih.gov/omssa/index.htm</a>
PepFrag (Prowl)	<a href="http://prowl.rockefeller.edu/prowl/pepfrag.html">http://prowl.rockefeller.edu/prowl/pepfrag.html</a>
PepProbe	<a href="http://bart.scripps.edu/public/search/pep_probe/search.jsp">http://bart.scripps.edu/public/search/pep_probe/search.jsp</a>
Phenyx	<a href="http://phenyx.vital-it.ch/pwi/login/login.jsp">http://phenyx.vital-it.ch/pwi/login/login.jsp</a>
Popitam	<a href="http://www.expasy.ch/tools/popitam/">http://www.expasy.ch/tools/popitam/</a>
RAId_DbS	<a href="http://www.ncbi.nlm.nih.gov/CBBResearch/qmbp/RAId_DbS/index.html">http://www.ncbi.nlm.nih.gov/CBBResearch/qmbp/RAId_DbS/index.html</a>
Sonar (Knexus)	<a href="http://hs2.proteome.ca/prowl/knexus.html">http://hs2.proteome.ca/prowl/knexus.html</a>
X!Tandem (The GPM)	<a href="http://thegpm.org/TANDEM/index.html">http://thegpm.org/TANDEM/index.html</a>
XProteo	<a href="http://xproteo.com:2698/">http://xproteo.com:2698/</a>
Not on-line	ByOnic, Crux, MassMatrix, Myrimatch, Paragon, PepSplice, ProLuCID, ProBID, ProteinLynx GS, SIMS, Sequest, SpectrumMill, greylag, pFind, etc.

# Validation

**PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATIONS - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

http://www.mcponline.org/site/misc/ParisReport\_Final.xhtml

Sugar pubwww1 pubwww1 Status Currency Converter - ... MS Bugs & FAQ Twiki Family report 2009\_ASMS\_Fall\_Wor... MascotImproveReport...

**MCP MOLECULAR & CELLULAR PROTEOMICS**

QUICK SEARCH Author:  Keyword:  Year:  Vol:  Page:  **Go** [Advanced Search]

[Home](#) | [Current issue](#) | [Papers in Press](#) | [Archive](#) | [Reviews](#) | [HUPO Views](#) | [Editorials](#) | [Special Issues](#)

PUBLICATION GUIDELINES FOR THE ANALYSIS AND DOCUMENTATION OF PEPTIDE AND PROTEIN IDENTIFICATIONS

1. The following supporting information should be included with the manuscript:

- The method and/or program (including version number) used to create the "peak list" from the raw data and the parameters used in the creation of this peak list, particularly any which might affect the quality of the subsequent database search. Examples include whether smoothing was applied, any signal-to-noise criteria, whether charge states were calculated or peaks de-isotoped, etc. In cases where additional customized processing of the collections of peak lists has been performed, e.g. clustering or filtering, the method and/or program (including version number) should be referenced.
- The name and version of the program(s) used for database searching and the values of search parameters. Examples include precursor-ion mass tolerance, fragment-ion mass tolerance, modifications allowed for, any missed cleavages, protein cleavage chemistry, (if any), etc.
- The name and version of the sequence database(s) used. If a database was compiled in-house, a complete description of the source of the sequences is required. The number of entries actually searched from each database should be included. Authors should justify the use of a very small database or database that excludes common contaminants, since this may generate misleading assignments.
- Methods used to interpret MS/MS data, thresholds and values specific to judging certainty of identification, whether any statistical analysis was applied to validate the results, and a description of how applied.
- For large scale experiments, provide the results of any additional statistical analyses that indicate or establish a measure of identification certainty, or allow a determination of the false-positive rate, e.g., the results of randomized database searches or other computational approaches.

**Current Issue**  
March 2010, 9 (3)

**For Authors**

**For Subscribers**

**About the Journal**

**Editorial Board**

Done

# Validation

## Search a “decoy” database

- Decoy entries can be reversed or shuffled or randomised versions of target entries
- Decoy entries can be separate database or concatenated to target entries

## Gives a clear estimate of false discovery rate

- Elias, J. E. and Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, *Nature Methods* 4 207-214 (2007)

# Validation

Matrix Science - Mascot - MS/MS Ions Search - Microsoft Internet Explorer

Address: http://frill/mascot\_2\_2/cgi/search\_form.pl?FORMVER=2&SEARCH=MIS

MATRIX SCIENCE HOME: MASCOT:HELP Search Go

Mascot > MS/MS Ions Search Logged in as admin | Logout

### MASCOT MS/MS Ions Search

Your name	<input type="text"/>	Email	<input type="text"/>
Search title	<input type="text"/>		
Database	SwissProt		
Taxonomy	All entries		
Enzyme	Trypsin/P	Allow up to	2 missed cleavages
Fixed modifications	<ul style="list-style-type: none"><li>3-Nitrotyrosine (Y)</li><li>Acetyl (K)</li><li>Acetyl (N-term)</li><li>Acetyl (Protein N-term)</li><li>Amidated (C-term)</li></ul>	Variable modifications	<ul style="list-style-type: none"><li>3-Nitrotyrosine (Y)</li><li>Acetyl (K)</li><li>Acetyl (N-term)</li><li>Acetyl (Protein N-term)</li><li>Amidated (C-term)</li></ul>
Quantitation	None		
Peptide tol. ±	0.1 Da # <sup>13</sup> C 0	MS/MS tol. ±	0.6 Da
Peptide charge	1+	Monoisotopic	<input checked="" type="radio"/> Average <input type="radio"/>
Data file	<input type="text"/> Browse...	Precursor	<input type="text"/> m/z
Data format	Mascot generic	Error tolerant	<input type="checkbox"/>
Instrument	ESI-TRAP	Report top	AUTO hits
Decoy	<input checked="" type="checkbox"/>		

Start Search ... Reset Form

Copyright © 2006 Matrix Science Ltd. All Rights Reserved. Local intranet

Select Summary Report (A8 Sprot + integral decoy) - Microsoft Internet Explorer

Address [http://frill/mascot\\_2\\_2/cgi/master\\_results.pl?file=..%2Fdata%2F20061212%2FF001580.dat&REPTYPE=select&sigthreshold=0.05&REPORT=100&server\\_mudpit\\_switch=0.0000](http://frill/mascot_2_2/cgi/master_results.pl?file=..%2Fdata%2F20061212%2FF001580.dat&REPTYPE=select&sigthreshold=0.05&REPORT=100&server_mudpit_switch=0.0000) Go

[PARP1\\_HUMAN](#) (P09874) Poly [ADP-ribose] polymerase 1 (EC 2.4.2.30) (PARP-1) (ADPRT) (NAD(+)-ADP-rib  
[TBA1\\_DROME](#) (P06603) Tubulin alpha-1 chain  
[DDX21\\_HUMAN](#) (Q9NR30) Nucleolar RNA helicase 2 (EC 3.6.1.-) (Nucleolar RNA helicase II) (Nucleolar  
[COF1\\_HUMAN](#) (P23528) Cofilin-1 (Cofilin, non-muscle isoform) (18 kDa phosphoprotein) (p18)  
[SYEP\\_HUMAN](#) (P07814) Bifunctional aminoacyl-tRNA synthetase [Includes: Glutamyl-tRNA synthetase (E  
[1433E\\_BOVIN](#) (P62261) 14-3-3 protein epsilon (14-3-3E)  
[HSP70\\_MAIZE](#) (P11143) Heat shock 70 kDa protein  
[1433T\\_BOVIN](#) (Q3SZI4) 14-3-3 protein theta  
[PUR2\\_HUMAN](#) (P22102) Trifunctional purine biosynthetic protein adenosine-3 [Includes: Phosphoribos  
[HNRPL\\_HUMAN](#) (P14866) Heterogeneous nuclear ribonucleoprotein L (hnRNP L)  
[IMDH2\\_HUMAN](#) (P12268) Inosine-5'-monophosphate dehydrogenase 2 (EC 1.1.1.205) (IMP dehydrogenase 2)  
[PGK1\\_BOVIN](#) (Q3T0P6) Phosphoglycerate kinase 1 (EC 2.7.2.3)

	Sprot	Decoy	False discovery rate
Peptide matches above identity threshold	3290	8	0.24 %
Peptide matches above homology or identity threshold	6037	224	3.71 %

### Select Summary Report

Format As  Select Summary (protein hits)

Significance threshold p <  Max. number of hits

Standard scoring  MudPIT scoring  Ions score or expect cut-off  Show sub-sets

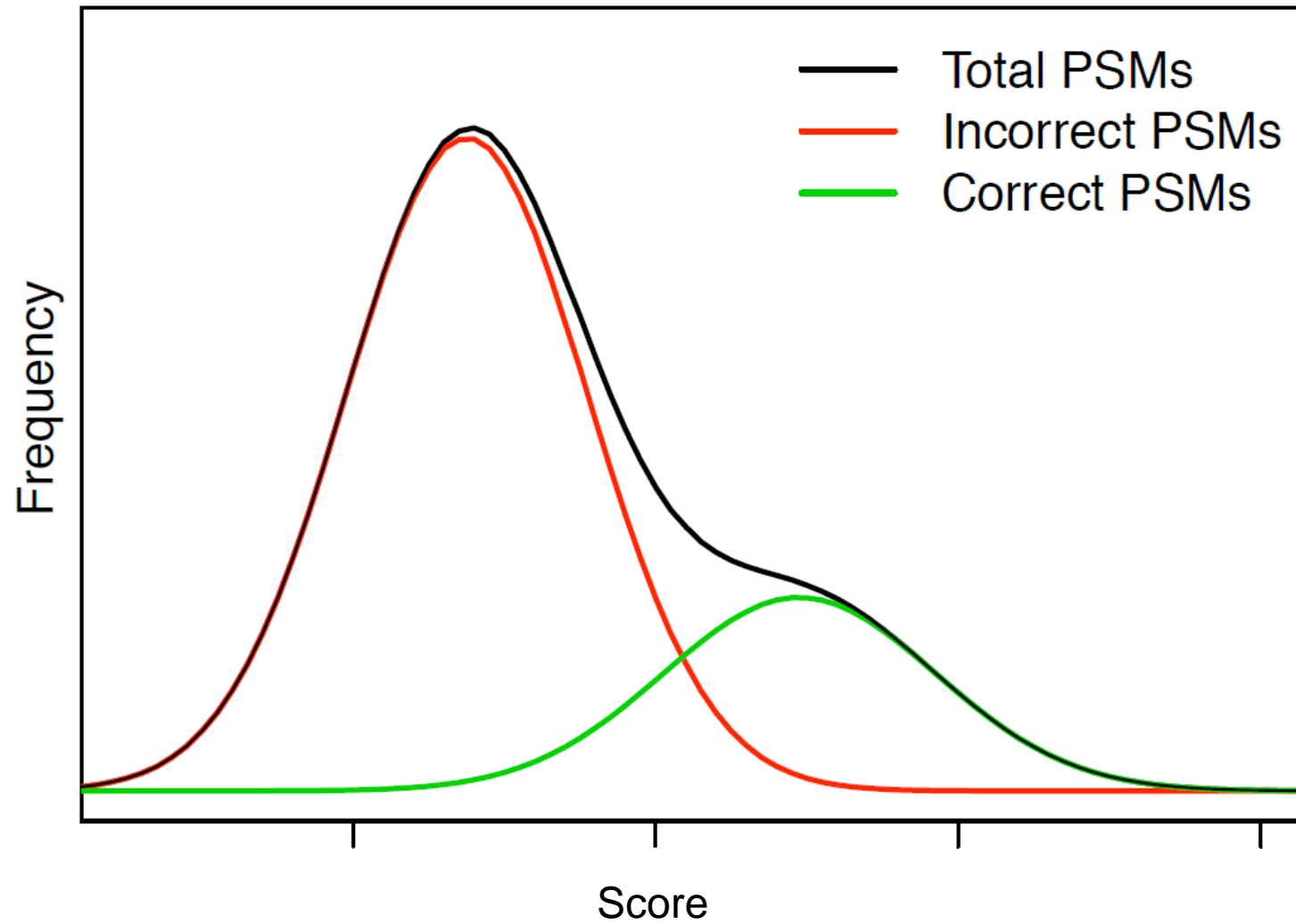
Show pop-ups  Suppress pop-ups  Sort unassigned  Require bold red

[Import results into MI](#)

1. [HS90B\\_HORSE](#) **Mass:** 83396 **Score:** 2231 **Queries matched:** 188 **emPAI:** 3.20  
 (Q9GKX8) Heat shock protein HSP 90-beta (HSP 84)

Local intranet

# Sensitivity optimisation



# Sensitivity optimisation

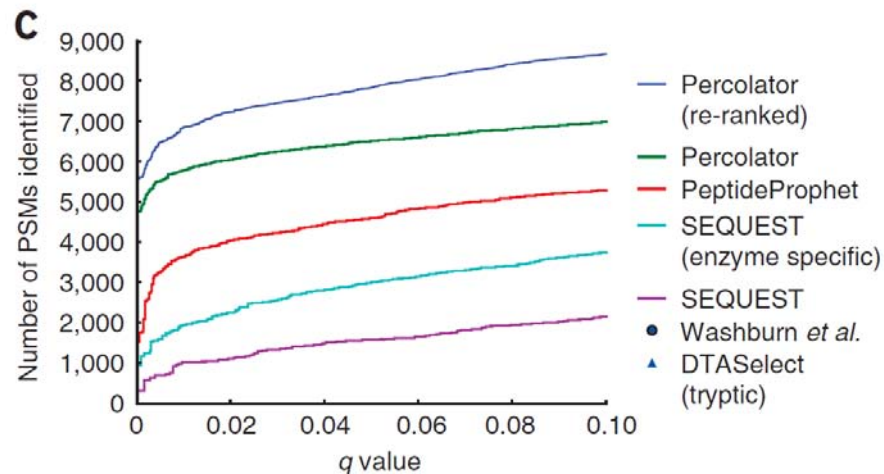


*Anal. Chem.* 2002, 74, 5383–5392

## Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search

Andrew Keller,<sup>\*,†</sup> Alexey I. Nesvizhskii,<sup>\*,†</sup> Eugene Kolker, and Ruedi Aebersold

*Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103*

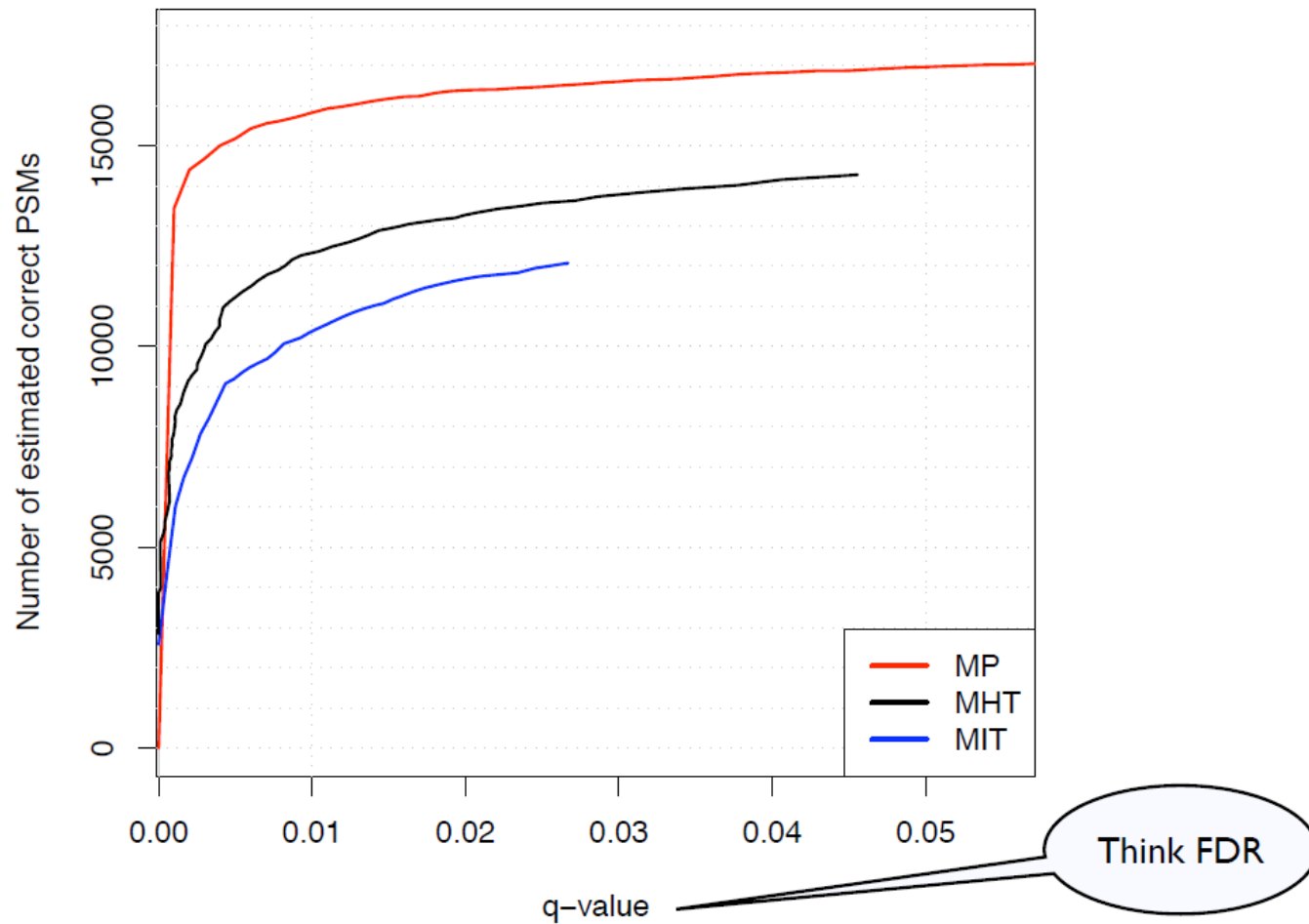


NATURE METHODS | VOL.4 NO.11 | NOVEMBER 2007 | 923

## Semi-supervised learning for peptide identification from shotgun proteomics datasets

Lukas Käll<sup>1</sup>, Jesse D Canterbury<sup>1</sup>, Jason Weston<sup>2</sup>, William Stafford Noble<sup>1,3</sup> & Michael J MacCoss<sup>1</sup>

# Sensitivity optimisation

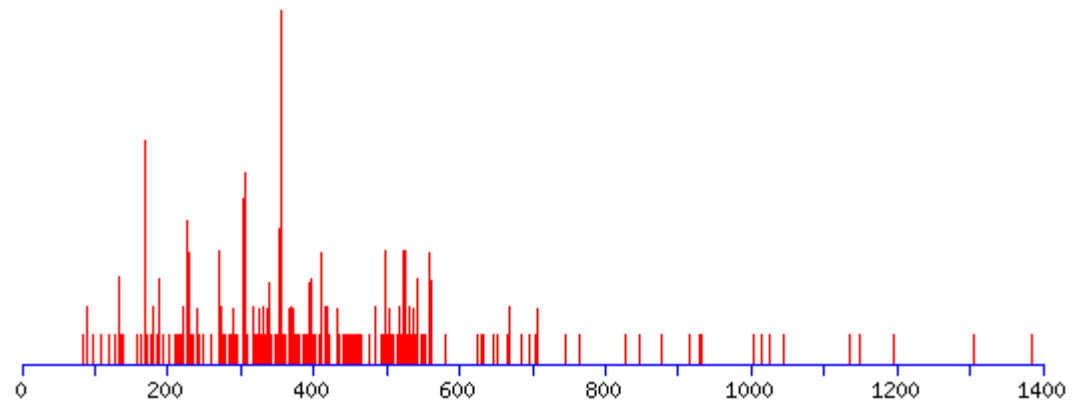


M. Brosch, L. Yu, T. Hubbard, J. Choudhary, *J Proteome Res* (2009).

# Practical Tips



# Data Quality



# Practical Tips

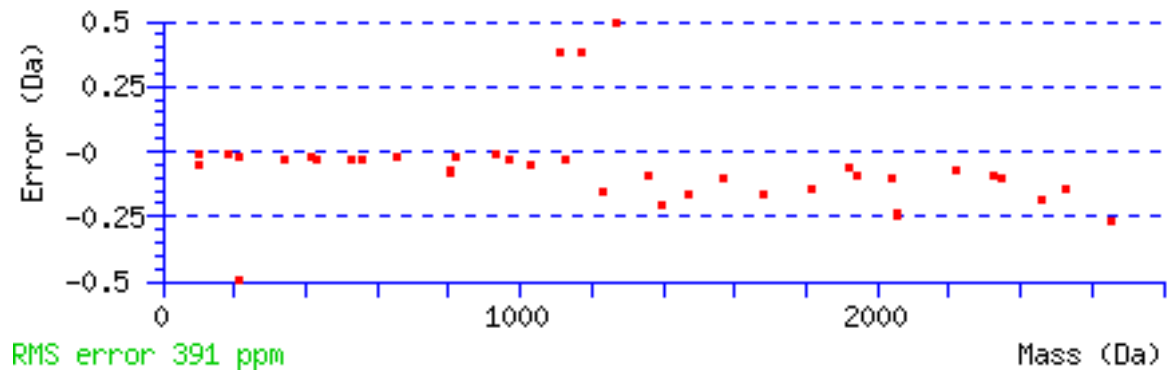
## Modifications

- Fixed / static modifications cost nothing
- Variable / differential modifications are very expensive
- Use minimum variable modifications
  - Maybe oxidation of M
  - Maybe pyro-glu
- User 'Error tolerant search' or 'Model refinement' to find post translational modifications

# Practical Tips

## Make a reasonable estimate of mass error

- Don't just guess, run a standard



# Practical Tips

## Which Database?

- Non-redundant protein database; there will be differences between the database sequence and the analyte

Swiss-Prot

- Large, comprehensive, non-identical database; explicit representation of every known peptide

NCBI nr, UniRef100

# Practical Tips

## Enzyme

- Loose trypsin (cleaves KP, RP)
- Semi-specific trypsin
- Only use “no enzyme” if strictly necessary
- Set missed cleavages by inspection of standards

# Further Reading

