iPRG2012: A Study on Detecting Modified Peptides in a Complex Mixture

John Cottrell[1] , Karl R. Clauser[2], Robert J Chalkley[3], Ruixiang Sun[4], Eugene Kapp[5], Matt Chambers[6], W. Hayes McDonald[6], Henry H. Lam[7], Nuno Bandeira[8], Eric Deutsch[9] and Thomas Neubert[10]

[1]Matrix Science, London, UK; [2]The Broad Institute of MIT and Harvard, Cambridge, MA; [3]University of California, San Francisco, CA; [4]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China; [5]Ludwig Institute for Cancer Research, Parkville, Victoria, Australia; [6]Vanderbilt University, Nashville, TN; [7]Hong Kong University of Science and Technology, Hong Kong, China; [8]University of California, San Diego, CA; [9]Institute for Systems Biology, Seattle, WA; [10]New York University School of Medicine, New York, NY

Proteome Informatics Research Group

A Proteome Informatics Challenge

Nature uses a wide variety of protein post-translational modifications to regulate protein structure and activity and tandem mass spectrometry has emerged as the most powerful analytical approach to detect these moieties. However, modified peptides present special challenges for characterization. First, they are generally present at sub-stoichiometric levels, meaning that without enrichment strategies samples are dominated by unmodified peptides, so finding the modified peptides may be a challenge. Secondly, the modifications may have unique fragmentation behaviors in collision-induced dissociation (CID), which may need to be considered by database search engines. Finally, if there are multiple residues within a given peptide that could bear a particular modification type, then it is necessary to identify fragment ions that frame either side of the modification site in order to be able to localize the exact site of modification within the peptide.

The Proteome Informatics Research Group (iPRG) created a collaborative data analysis study to enable proteomics laboratories to evaluate their ability to find a variety of post-translationally modified peptides within a complex peptide mixture background. The dataset consists of nearly twenty thousand high resolution and high mass accuracy tandem mass spectra. Within the sample there are peptides with a range of different natural and chemical modifications. This study enabled participants to evaluate their data analysis capabilities and approaches relative to others in analyzing a common data set, with a particular emphasis on their ability to detect and characterize peptides with modifications of potential biological significance.

Study Goals

1. Evaluate ability of participants to identify modified peptides in a complex mixture
2. Find out why result sets might differ between participants
3. Produce a benchmark dataset, along with an analysis resource

Study Design

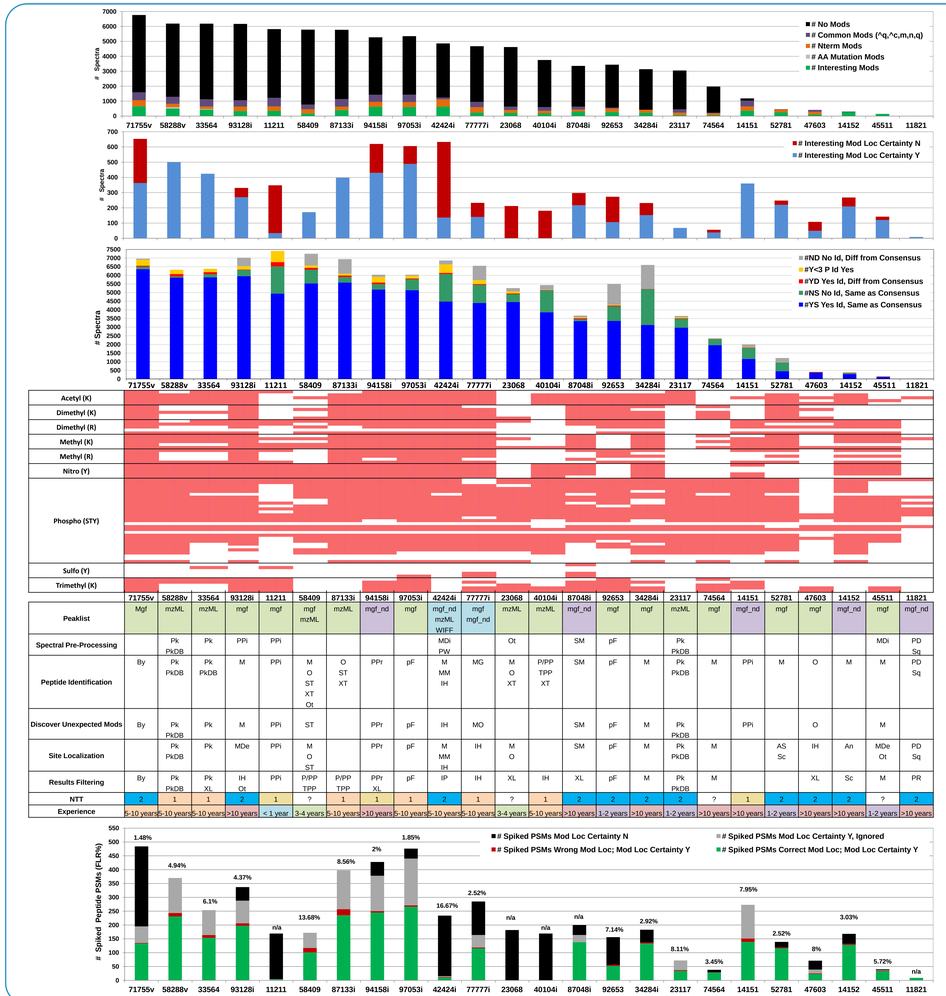
- Use a common, rich dataset
- Use a common sequence database
- Allow participants to use the bioinformatic tools and methods of their choosing
- Use a common reporting template
- Report results at an estimated 1% FDR (at the spectrum level)
- Ignore protein inference

Study Materials

- 5600 TripleTOF dataset (AB-SCIEX)
  - WIFF, mzML, dta, MGF (de-isotoped);– conversions by MS Data Converter 1.1.0
  - MGF (not de-isotoped – conversion by Mascot Distiller 2.4)
- 1 FASTA file (SwissProt *S. cerevisiae, human, + 1 bovine protein + trypsin* from Dec. 2011)
- 1 template (Excel)
- 1 on-line survey (Survey Monkey)

Study Instructions

1. Analyze the dataset
2. Report the peptide spectrum matches in the provided template
3. Report measures of reliability for PTM site assignments (optional)
4. Complete an on-line survey
5. Attach a 1-2 page description of your methodology



Total Spectra vs. Interesting Mods

There is a very wide range in the total number of spectra with identified peptides. Once one focuses only on the spectra containing modifications for which the ability to localize the modification to a particular residue, the range is much narrower. The 5 rightmost participants went so far as to report only spectra of modified peptides.

Room for Improvement in ID Certainty Thresholds

Identifications reported as: **Yes that matched the consensus**; **No, but still matching the consensus**; **Yes, but a different answer than the consensus**; **Yes, < 3 consensus**; **No, that disagreed with consensus**

Synthetic Peptide ID by Participant

Red corresponds to the presence of at least 1 PSM for a spiked synthetic peptide modified with the correct localization reported and the correct modification name. The localization certainty may have been reported as either Y or N. PSMs containing modification of residues other than s,t,y,k,r were excluded.

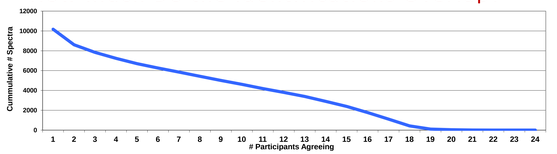
Table Key:

An = Andromeda/MaxQuant
MG = MS-GFDB
pF = pFind
Sc = Scaffold
AS = A-Score
MM = MyriMatch
Pk = PEAKS
SM = Spectrum Mill
By = Byonics
MO = MODa
PkDB = PEAKSDB
Sq = Sequest
IH = In-house software
O = OMSSA
Ppi =Protein Pilot
ST = SpectraST
IP = IDPicker
Ot = Other
PPr = Protein Prospector
TPP = TransProteomic Pipeline
M = Mascot
P/PP = Pep/Prot Prophet
PR = PhosphoRS
XL = Excel
Mde = Mascot Delta Score
PD = ProteomeDiscoverer
PW = ProteoWizard
XT = XTandem
Mdi = Mascot Distiller

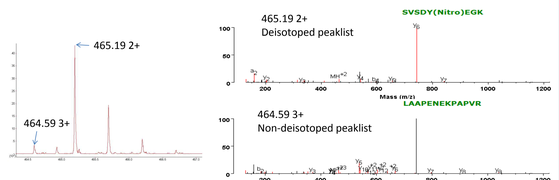
Peak Processing

- Two types of peak lists were supplied
  - **Deisotoped (AB SCIEX MS Data Converter)**
    - Cannot infer fragment charge state
    - Possibly lower chance of false fragment matches
  - **Non-deisotoped (Mascot Distiller)**
    - Can infer fragment charge state
    - Possibly higher chance of false fragment matches
- For 238 consensus spectra the peak lists had different specified charge state
  - 193 results only possible with deisotoped peak list
  - 45 results only possible with non-deisotoped peak list

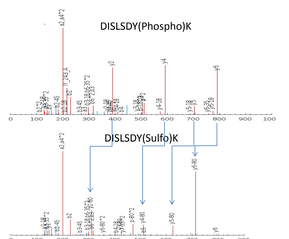
How Much Do the Identifications Overlap?



Mixed Spectra Exposed by Peak Processing



Phospho vs Sulfo



Observe modified fragment ions.

Observe 'unmodified' fragment ions. Spectrum looks essentially identical to unmodified peptide spectrum

Preliminary Conclusions

- Reasonable number of participants from around the globe, mainly experienced users but a few first-timers
- Large spread in number of spectra identified
- False negatives (NS) are generally much higher than false positives, so there is generally room for improvement
- Peak list was a significant factor on performance
- Varied performance in detecting PTMs
  - Most participants struggled with sulfation
  - Multiply phosphorylated harder to find than singly
- Most common errors in site assignment were:
  - Reporting sulfo(Y) as phospho(ST)
  - Mis-assignment of site/s in multiply phosphorylated peptides
- The iPRG2012 are in the process of preparing the data for publication. If you participated and would like to help out, contact the iPRG through anonymous.iPRG2012@gmail.com.

For more information on the iPRG and for copies of this poster and the talk please visit: http://www.abrf.org/iPRG after the meeting

Acknowledgements: The iPRG are grateful to *all of the participants*. We would also like to thank Jeremy Carver (UCSD) for serving as the "Anonymizer".

Who Participated:

24 submissions from 23 participants. 9 were iPRG members. Participation was international and covered a wide range of experience level.