iPRG2008 Study: Assessing the Quality and Consistency of Protein Reporting on a Common Data Set

W.S. Lane[1], A.I. Nesvizhskii[2], B. C. Searle[3], D.L. Tabb[4], J.A. Falkner[2], J.A. Kowalak[6], and S.L. Seymour[7]

[1]Harvard University, Cambridge, MA; [2]University of Michigan, Ann Arbor, MI; [3]Proteome Software, Inc., Portland, OR;

[4]Vanderbilt University Medical Center, Nashville, TN; [5]National Institute of Mental Health, Bethesda MD; [6]Applied Biosystems|MDS Sciex, Foster City, CA

Proteome Informatics Research Group

RG11

Abstract

A significant challenge in proteome informatics is accurate and concise reporting of protein identification data that result from mass spectrometry-based proteomic workflows. The Paris Guidelines represent the proteomics community's efforts to devise a standard methodology for reporting protein identification data. Conformance to these criteria is affected by the bioinformatics tools used, how one understands the issues, and interpretation of the reporting criteria. At ABRF2007 in Tampa, FL, the Bioinformatics Committee (BIC) of the ABRF sPRG presented a study assessing the consistency of protein reporting among a group of experts. In the present work, the newly formalized Proteome Informatics Research Group (iPRG) presents the results of a study that was designed to assess the quality and consistency of protein identification analysis across the broader ABRF community. The major goals of this study were to assess the present status of the field and to provide a benchmark for the quality of protein inference on a realistically complex dataset.

Introduction

- **Primary goal: What is the current state of protein reporting?**
  - Does the proteomics community still have a problem with reporting excessively long protein lists due to poor protein inference?
  - Is ambiguity among multiple non-differentiable accessions being properly reported in protein groups?
- **Secondary goal: Assess the consistency of protein identification analysis across users starting from the same mass spectra.**
- **Secondary goal: Establish a benchmark reference.**
  - Enable users of software to test their lab's methods.
  - Provide a frame of reference for future software development.

Methods

- **A mouse liver differential expression experiment with trypsin digestion, MMTS alkylation, and iTRAQ® reagent labeling was separated into 13 cation exchange fractions.**
- **The fractions were analyzed on a 3200 QTRAP® system producing 29 files and 41977 spectra.**
- **The data were provided in raw .wiff, MGF, DTA, mzData, and mzXML format. Use of the provided database (MGI, Dec 3, 2007 version, 53,826 proteins + 74 contaminants) or a derivative of it was required.**
- **Using the tools of their choosing, participants were asked to produce a protein list they would report to a journal, including supporting peptide IDs. An Excel template was required for submission.**
- **All accessions were combined into clusters of homolog proteins, and the Research Group (RG) consensus number of detectable isoforms per each cluster was determined. Clusters were assigned to classes based on the number of isoforms and degree of consensus:**
  - **Class 1 – RG consensus multi-detection clusters.**
  - **Class 2 – Debatable multi-detection clusters.**
  - **Class 3 – RG consensus single-detection clusters.**
  - **Class 4,5 – Non-consensus detections by RG and non-RG, respectively.**
- **Respondent submissions were graded vs. the expected number of isoforms per cluster for Class 1, 2, and 3 clusters (209 clusters, 258 detectable isoforms expected).**

Figure 1 – Results Overview

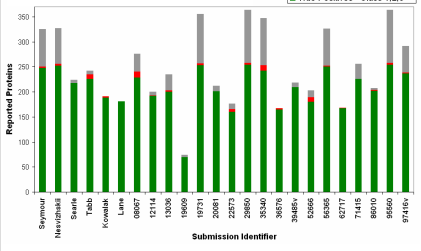
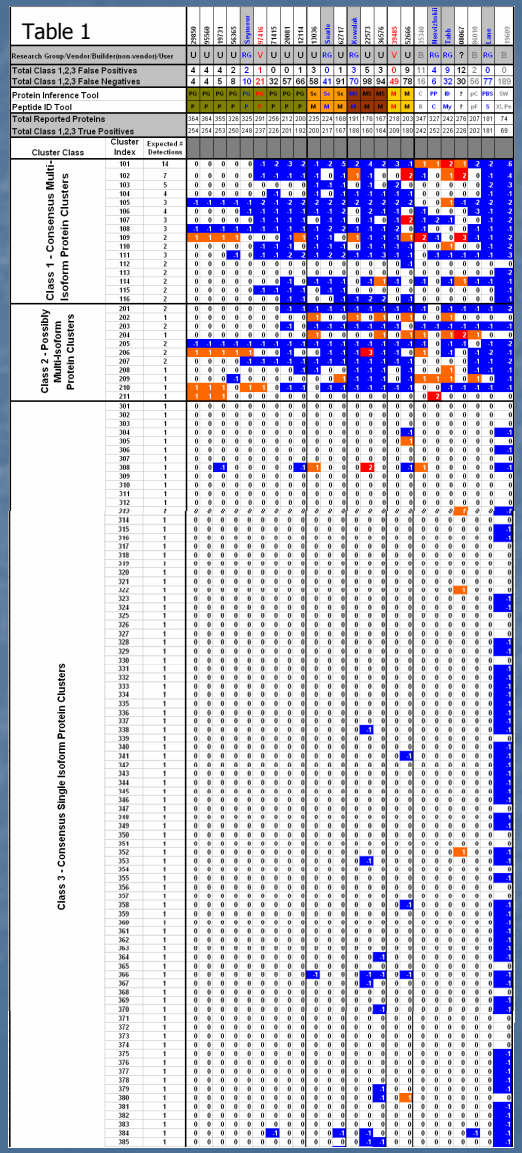


Table 1



Legend

Reporting the # of isoforms in column *Expected # Detections*

A false positive (one more than the expected # of detections).

A false negative (one fewer than the expected # of detections).

Peptide ID Tools

B ByOnic
M Mascot
My MyriMatch
P Paragon
Pe Peaks
pF pFind
S SEQUEST
XI XTandem

Protein Inference Tools

C ComByne
ID ID Picker
M Mascot
MS MassSieve
pC pCompare
PG Pro Group
PP Protein Prophet
PBS Proteomics Browser Suite
Sc Scaffold
SW Self-written scripts

Table 1.

- **Each column is one submission.**
- **Inference tools grouped & sorted by decreasing reported proteins.**

Table 1 (cont.)

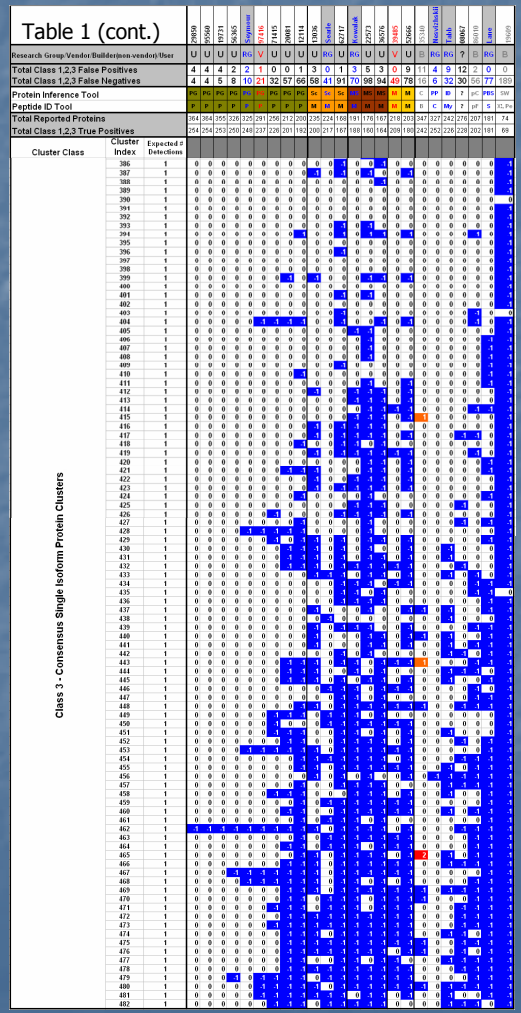


Figure 2 – Consistency of Results

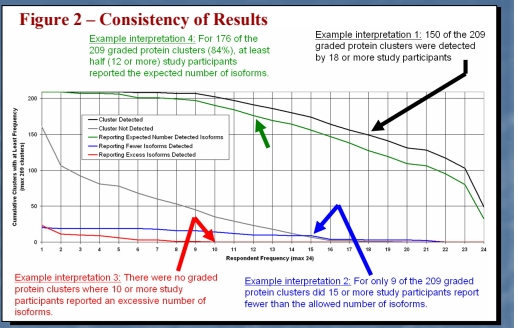


Figure 3 Result Comparison

- **Even with same inference tools, use of different search engines, versions, and thresholds settings result in different true/false positive tradeoffs.**

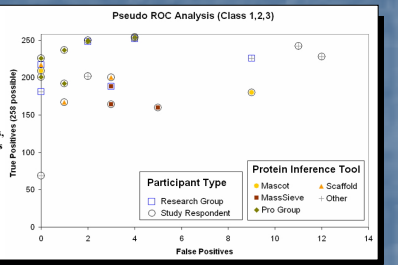
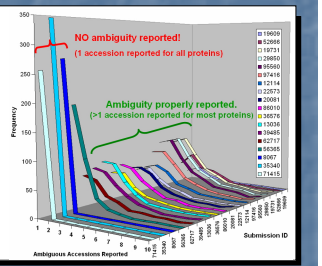


Figure 4 – Reporting Accession Ambiguity

- **Most respondents reported a reasonable level of accession ambiguity in groups.**
- **Three respondents inappropriately reported a single accession for every protein group.**



Conclusions

- **Virtually all respondents used some kind of protein inference tool.**
- **No gross inflation in number of reported proteins was seen in any submission.**
- **Some submissions did not properly report accession ambiguity (3/18).**
- **Not enough submissions to draw conclusions across the proteomics community.**
- **More submissions are needed! Please email iPRG2008@gmail.com if interested!**
- **For more information about this study visit: www.ABRF.org/iPRG**

Acknowledgments

The iPRG thanks Renee Robinson (Harvard University) for serving as our anonymizer.