ABRF iPRG 2008 Study: Characterization of Protein Inference Reporting from Proteomics Studies

Brian C. Searle[1]; David L. Tabb[2]; Alexey I. Nesvizhskii[3]; William S. Lane[4]; Jeffery A. Kowalak[5]; Jayson A. Falkner[3]; Sean L. Seymour[6]

[1]Proteome Software, Portland, OR; [2]Vanderbilt University Medical Center, Nashville, TN; [3]University of Michigan, Ann Arbor, MI; [4]Harvard University, Cambridge, MA; [5]National Institute of Mental Health, Bethesda, MD; [6]Applied Biosystems, Foster City, CA

Proteome Informatics Research Group

Abstract

Accurate and concise reporting of protein identification data that result from mass spectrometry-based proteomic workflows is a key bioinformatic challenge. One particular complication, referred to as the protein inference problem, results from the loss of peptide to protein mapping caused by the enzymatic digestion of protein mixtures. Indeed, the Paris Guidelines, which represent the proteomics community's efforts to devise a standard methodology for reporting protein identification data, puts particular emphasis on reporting of a minimal set of proteins that account for all of the identified peptides. In the present work, the newly formalized Proteome Informatics Research Group (iPRG) presents the results of a study to assess the quality and consistency of protein inference analysis across the ABRF community.

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?
- 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).

Protein Inference Terminology

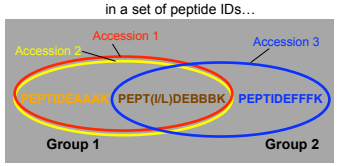
Cluster – A cluster of proteins is a collection of homologous proteins that could cite some common MS/MS fragmentation evidence.

Accession 1 AAKPEPTIDEAAAKAKPEPTIDEBBBKAKPEPTIDEDDDKAA
Accession 2 AAAAKPEPTIDEAAAKAKPEPTIDEBBBKAAAKPEPTIDEEEEK
Accession 3 AAAAAAAAAAAAAAKPEPTLDEBBBKAAAAKPEPTIDEFFFKA

- Common sequences not required – e.g. Ile/Leu difference
- Aggregation into a cluster is completely independent of how many of these isoforms can be detected in a given sample!

Group – One reported protein group implies the detection of a physical protein species. Because there may be ambiguity as to which accession best corresponds to this detected isoform, a group of accessions must be reported.

If we observe peptides A, B, and F in a set of peptide IDs…



…we report a protein ID list:

Protein	Accession	
1	Accession 1	Group 1
1	Accession 2	
2	Accession 3	Group 2

- Two isoforms can be detected in this cluster based on these IDs, thus two protein groups are reported in the results. Refer to this as a 'multi-detection' or 'multi-isoform' cluster.
- There is ambiguity in which accession is being detected in Group 1.

Figure 1: Results Overview

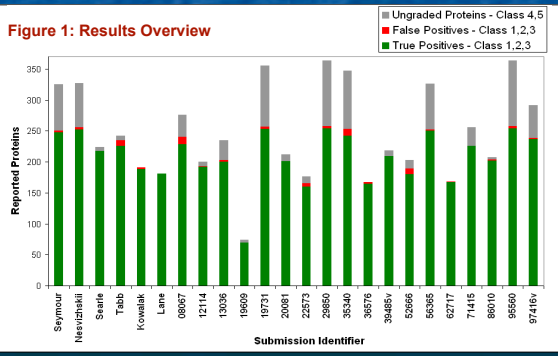


Figure 2 shows that in general most respondents reported the correct number of protein groups for each cluster. However, **Figure 3** shows that some rare inference errors were made resulting in over-inflated protein identification lists. The cells in **Figure 3** tally the number of respondents that identified the specified protein in each protein group.

Figure 2: Number of Protein Groups by Cluster

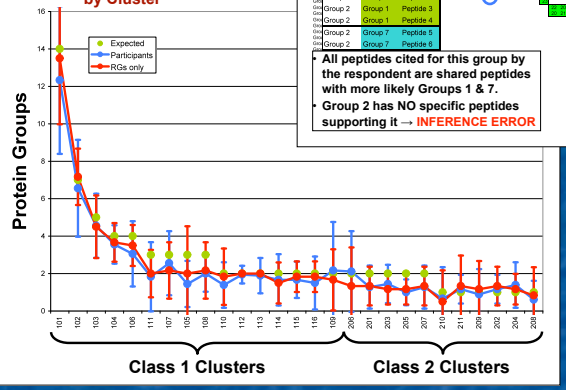


Figure 3: Cluster 101 (17 groups, 14 expected)

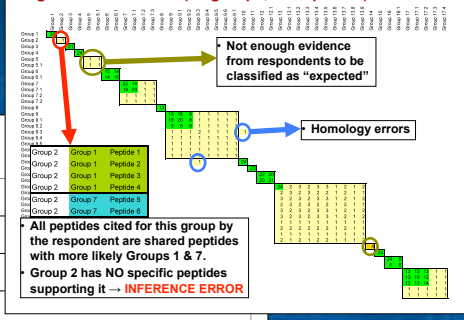
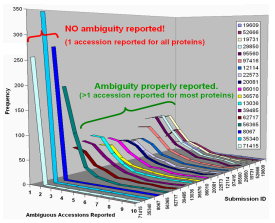


Figure 4: Reporting Protein Ambiguity

- **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, although in some cases mistakes were made.
- 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.