



Proteome Informatics
Research Group

iPRG2009 Study: Testing for Differences Between Complex Samples in Proteomics Datasets

Brian C. Searle¹; David L. Tabb²; Jayson A. Falkner³; Jeffery A. Kowalak⁴; Karen Meyer-Arendt⁵; Lennart Martens⁶; Manor Askenazi⁷; Paul A. Rudnick⁸; Sean L. Seymour⁹; William S. Lane¹⁰

¹Proteome Software, Portland, OR; ²Vanderbilt University Medical Center, Nashville, TN; ³University of Michigan, Ann Arbor, MI; ⁴National Institute of Mental Health, Bethesda, MD; ⁵University of Colorado, Boulder, CO; ⁶European Bioinformatics Institute, Cambridge, UK; ⁷Dana-Farber Cancer Institute, Boston, MA; ⁸National Institute of Standards in Technology, Gaithersburg, MD; ⁹Applied Biosystems, Foster City, CA; ¹⁰Harvard University, Cambridge, MA

Abstract

Determining significant differences between mass spectrometry datasets from biological samples is one of the major challenges for proteome informatics. Accurate and reproducible protein quantification in complex samples in the face of biological and technical variability has long been a desired goal for proteomics. The ability to apply difference testing is a first step towards that goal, and is routinely used in tasks such as biomarker discovery. In this work the ABRF Proteome Informatics Research Group (iPRG) presents the results of a collaborative study focusing on the determination of significantly different proteins between two complex samples. Datasets representing five technical replicates of each sample were provided to volunteer participants and their ability to evaluate reproducible differences was tested. A survey was used to determine the relative merits of various approaches to difference testing, whether sophisticated statistical methods are necessary, and if computer software must be augmented by scientific expertise. Results and survey responses were used to assess the present status of the field and to provide a benchmark for difference testing on a complex dataset.

Study Goals

Primary goal: Evaluate the effectiveness of current protein differentiation tools for mass spectrometry

- Are spectrum counting methods reliable enough to find differences between complex samples or are intensity-based quantitative methods required?
- Can participants accurately determine the confidence in the differences in the proteins they report?

Secondary goal: Establish a benchmark reference

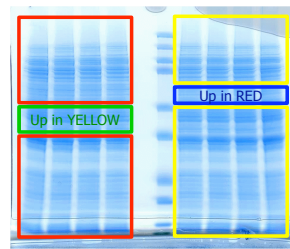
- Enable users of software to test their lab's methods.
- Provide a frame of reference for future software development.

Sample Distribution Methods

Only the red and yellow data files were distributed to participants. Samples blue and green were analyzed by the iPRG to determine the final answer key. Data files were disseminated in RAW, mzML, mzXML, MGF and DTA formats. The samples were distributed with a SwissProt *E. coli* database downloaded on 3/6/2008. The database was modified to contain bovine serum albumin, common contaminants and reverse database sequences as decoys. Participants were required to provide their results in an Excel template and fill out an online survey. 37 returned results from 30 laboratories were analyzed in a double-blind fashion.

Acknowledgments

We are enormously indebted to Dan Liebler and Amy Ham of Vanderbilt University for help developing the samples used in this study. In particular, we thank Kristin Cheek at the Ayers Institute for providing the *E. coli* samples and Salsisha Hill from the Mass Spectrometry Research Center for collecting the final MS/MS results. We are also grateful to Renee Robinson of Harvard University for acting as our anonymizer.



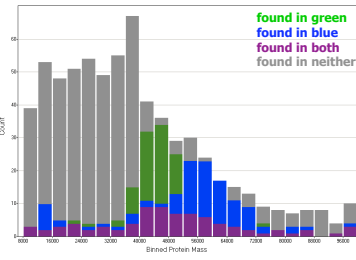
Creating the Samples

- 25 µg *E. coli* lysate was loaded onto each of 8 1D gel lanes and stained with Colloidal Coomassie Blue.
- The first four lanes were combined and designated "red" while the last four were designated "yellow".
- A region of proteins between 47 kDa and 65 kDa was excised from the yellow region of the gel and designated "blue".
- A region of proteins at between 37kDa and 47 kDa was excised from the red region and designated "green".
- Proteins were reduced, alkylated and in-gel digested.
- Proteins were spiked with bovine serum albumin (final concentration was 2 µg *E. coli* and 1.212 ng of bovine serum albumin).
- 5 technical replicates of red and yellow and 3 replicates of blue and green were acquired with an LTQ-Orbitrap.

Generating the Answer Key

Proteins in the changing regions of red and yellow should be enriched in blue and green. We assume that if a participant can identify a protein as enriched in red then it should be easily identifiable in blue. The same argument can be made for yellow and green.

Three technical replicates of both blue and green samples were analyzed by all 10 research group members. Proteins identified by 5 or more RG members were added to the answer key. Additionally, proteins identified by 2 or more RG members that were within the expected molecular weight cutoffs +/- 3kDa were also used. This resulted in 172 proteins present in blue, 174 proteins present in green, and 89 proteins present in both samples.



Results

Participants were asked to identify and rank proteins enriched in the red and yellow sample groups, which were then graded using the blue and green answer keys, respectively. Pseudo-receiver operating characteristic (ROC) curves depicting the total number of blue/green identifications versus total incorrect identifications were plotted in Figure 1 based on participant ranking. ROC curves that run closest to the upper left indicate best performance.

Two participants, marked with asterisks, sorted their results in alternate ways and declined the opportunity to re-submit new rankings. Their results have been re-sorted to improve their scoring but these data sets were not used to grade the overall methodologies.

Figure 1: Pseudo-ROC curves by method

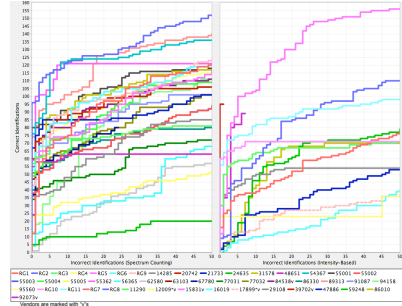
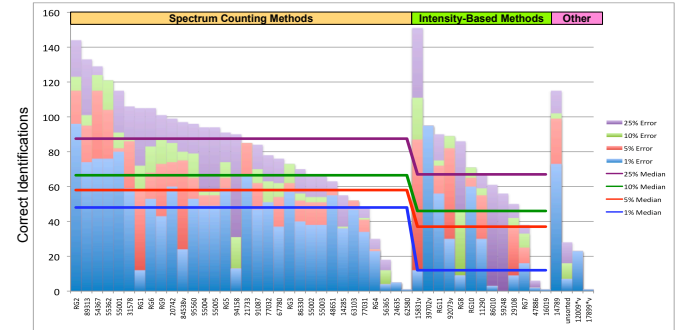


Figure 2: Results Overview



All participants were graded based on the number of correct blue/green identifications at a specified error rate. In Figure 2, correct identifications at four error rates (1%, 5%, 10% and 25%) are depicted.

Participants using either spectrum counting or intensity-based methods were able to succeed at this study and there was no clear best methodology. The best ranking participants at 1%, 5% and 10% used spectrum counting based methods while ultimately, the best ranking participant at 25% used an intensity-based method. Although it was possible to succeed using either method, the median spectrum counting user did 30% to 60% better than the median intensity-based user at all error rate levels. Average participants using intensity-based methods may have been less able to use their software to its fullest potential, suggesting that there is room for the field to grow in this area. As spectrum counting is a simpler method with fewer

pitfalls, average users tend to do better.

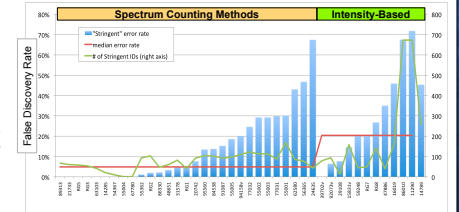
Unlike previous studies, (see www.abrf.org/iPRG) a distinct trend is that "expert" iPRG members did not necessarily perform better than other participants. Data from the survey clearly shows that general proteomics experience does not correlate with success in this study, but experience with label-free differentiation does ($p=0.097$).

Participants were requested to specify a "stringent" criteria level at which they would feel comfortable publishing results. The actual false discovery rate for stringent proteins by participant is shown in Figure 3. Although many participants specified stringent levels with less than 10% FDR, it is clear future work in determining error rates for biomarker discovery applications is necessary. Intensity-based users tended to report longer identification lists that were offset by higher error levels.

Conclusions

- Participants succeeded with both spectrum counting and intensity-based methods.
- More experience is required to succeed with intensity-based methods.
- Accurate estimation of error rates was difficult for many participants.
- Many of the protein differences in this study were much larger than in real biomarker discovery experiments. Further work is necessary to show if either method is capable of identifying changes in real world studies.
- The iPRG has produced a publicly available dataset useful to the field for testing.

Figure 3: Error Rate for "Stringent" Proteins



For more information, please visit www.abrf.org/iPRG