Meta-analysis for Protein Identification: A Case Study on Yeast Data

Roger Higdon,^{1,2,3} Winston Haynes,^{1,4} and Eugene Kolker^{1,2,3,5}

Abstract

Large amounts of mass spectrometry (MS) proteomics data are now publicly available; however, little attention has been given to how to best combine these data and assess the error rates for protein identification. The objective of this article is to show how variation in the type and amount of data included with each study impacts coverage of the yeast proteome and estimation of the false discovery rate (FDR). Our analysis of a subset of the publicly available yeast data showed that failure to reevaluate the FDR when combining protein IDs from different experiments resulted in an underestimation of the FDR by approximately threefold. A worst-case approximation of the FDR was only slightly larger than estimating the FDR by randomized database matches. The use of a weighted model to emphasize the most informative experimental data provided an increase in the number of IDs at a 1% FDR when compared to other meta-analysis approaches. Also, using an FDR higher than 1% results in a very high rate of false discoveries for IDs above the 1% threshold. Ideally, raw MS data will be made publicly available for complete and consistent reanalysis. In the circumstance that raw data is not available, determining a combined FDR on the basis of the worst-case estimation provides a reasonable approximation of the FDR. When combining experimental results, adding additional experiments results in diminishing and in some cases negative returns on protein identifications. It may be beneficial to include only those experiments generating the most unique identifications due to solid experimental design and sensitive instrumentation.

Introduction

THE DEMANDS OF SCIENCE to make experimental data publicly available have led to the stockpiling of increasing amounts of experimental data in on-line databases and repositories. This accessibility of data is motivated by a new paradigm in science of data-intensive discovery, which builds upon the previous paradigms of experimentation, theory and simulation (Gray, 2009). With this desire to uncover the patterns contained in these massive datasets, there is an associated need for novel approaches towards assembling and effectively analyzing these datasets.

For proteomics, databases such as the Open Proteomics Database, PeptideAtlas, Peptidome, and PRIDE serve as repositories for large numbers of proteomics experiments (Desiere et al., 2006; Prince et al., 2004; Slotta et al., 2009; Vizcaino et al., 2009). These services in some cases provide all the original mass spectrometry raw data and other summary data about protein identifications (IDs). Given the availability

of these resources, it is becoming worthwhile for researchers to perform meta-analyses on these datasets. Meta-analyses have become commonplace for clinical trials and biomedical research (DerSimonian and Laird, 1986; Farrer et al., 1997; Glass, 1976). There is also a growing literature on meta-analyses for microarray data (Choi et al., 2003; Moreau et al., 2003; Rhoades et al., 2002), but little has been said with respect to proteomics or protein identification. Currently, researchers often simply combine lists of relevant proteins without taking into account the impact this may have on the overall error rate of this list.

Publicly available raw mass spectrometry (MS) data allow researchers with ample resources to do complete reanalyses of proteomic studies while controlling for variability in analysis techniques between laboratories. In terms of statistical integrity and informational content, public access to the raw MS data is the ideal circumstance. Unfortunately, the sheer size of raw MS datasets and the computational demands of their analysis strain the resources of both the producers and consumers of these data. As an adverse side effect of these

¹Bioinformatics & High-throughput Analysis Laboratory and ²High-throughput Analysis Core, Seattle Children's Research Institute, Seattle, Washington.

³Predictive Analytics, Seattle Children's Hospital, Seattle, Washington.

⁴Hendrix College, Conway, Arkansas.

⁵Division of Biomedical & Health Informatics, Department of Medical Education & Biomedical Informatics, School of Medicine, University of Washington, Seattle, Washington.

demands, proteomics datasets are often reduced to summary information and lists of protein identifications. Difficulties arise when comparing summary information from different experiments due to the many different approaches that are applied to protein identification. These include different database search algorithms (Sequest, Mascot, X!Tandem, OMS-SA) (Eng et al., 1994; Fenyö and Beavis, 2003; Geer, et al., 2004; Perkins et al., 1999), postprocessing methods, and error rate estimation techniques. The comparability of summary data increases with the use of the false discovery rate (FDR) and the use of randomized or decoy database searches to estimate the FDR (Hather et al., 2010; Higdon et al., 2005, 2007; Higdon and Kolker, 2007).

If raw data are available, their reanalysis is preferred to a meta-analysis of summary data (Higdon et al., 2008). It has been also noted in the meta-analysis of microarrays that lack of standardization, bias in the availability of studies with publicly available raw data, and even just the sheer volume of data inhibits the reanalysis of raw data (Larsson et al., 2006). The issues are even more problematic in proteomics studies. In this work we focus on how differences in the amount of summary data available relative to raw data influence meta-analysis for protein identification with respect to proteome coverage and estimation of the FDR.

Methods

Meta-analysis approaches

The following list of possible circumstances of protein identification experimental data followed by reasonable approaches to estimate the FDRs in each circumstance is ordered in terms of the desirability of the situation:

- (0) Lists with no or noncomparable error rate estimates: no way to estimate FDR
- (1) Lists with comparable FDRs: upper bound on FDR
- (2) Lists with continuous FDR thresholds (q -value): set upper bound FDR threshold
- (3) Lists with randomized ID estimations for each FDR: estimate actual FDR
- (4) Lists with q -values and randomized IDs: set actual FDR threshold
- (5) Research raw data^a or common database search data: combined ID model
- (6) Reanalysis of raw data^a: multiple or optimal methods

To demonstrate the different approaches to meta-analysis of protein identification, large amounts of raw MS data on yeast were downloaded from different proteomics repositories. The objective was to compare numbers of protein identifications and the FDR estimation in circumstances 1–5 from the list above. Circumstance 0 was excluded because, by definition, it does not allow for meta-analysis. Circumstance 6 was not included because it focuses on comparisons of methods to increase the number of identifications for any dataset as opposed to meta-analysis of different datasets. Many methods have already been published toward that

objective including the use of multiple search engines, multistage searches, spectral libraries, sequence tags, and clustering of spectra (Deutsch et al., 2008; Everett et al., 2010; Gygi et al., 1999; Perkins et al., 1999).

To keep the focus on comparisons of the different data circumstances rather than differences in analysis methods used in different experiments, the yeast data were all reanalyzed using a common analysis. This common analysis eliminates the impact of differences between experiments and laboratories in search methods and sequence databases. The data were combined and FDRs estimated based on the limitations of each data circumstance.

Data download

A total of 11 datasets of yeast proteomics experiments in either RAW (Thermo) or mzXML format were downloaded from the Open Proteomics Database, Peptidome, and PeptideAtlas (Breci et al., 2005; Desiere et al., 2006; Maynard et al., 2004; Piening et al., 2006; Prince et al., 2004; Slotta et al., 2009; Swaney et al., 2008; Vizcaino et al., 2009). Details about the 11 individual experiments are given in Table 1. All experiments were examined to determine the specific experimental protocol that needed to be taken into consideration during data processing. These included whether variable or static mass modifications were necessary due to sample preparation treatments such as iodacetamide or labeling with ICAT (Gygi et al., 1999). Also, whether the data were obtained on a high or low mass accuracy instrument was taken into consideration.

Database search and protein ID for experiments

With all parameters and experimental designs in place, the MS data was run through SPIRE, the Systematic Protein Investigative Research Environment (<http://proteinspire.org>). For this analysis, SPIRE was set up to perform processing using X!Tandem against a protein database consisting of three parts. The first part was the National Center for Biotechnology Information *Saccharomyces cerevisiae* database, made of approximately 5,800 proteins (<http://www.ncbi.nlm.nih.gov>). The second was a database of common contaminants, and the third part was a randomly reshuffled version of the original *Saccharomyces cerevisiae* database. Raw MS data were searched using the assumption of tryptic cleavage of peptides with no more than one missed cleavage. Mass error was set to 2.5 Da for all searches. Experiments that used high mass accuracy instruments (e.g., LTQ Orbitrap) were adjusted for mass accuracy by postprocessing within the LIPS identification model (Higdon et al., 2004). Peptide identification probabilities and FDRs were estimated based on randomized databases matches using an isotonic regression (Hather et al., 2010; Higdon et al., 2007). Protein scores were generated based on the combination of the number unique peptide identifications (exceeding 90% certainty) and the maximum identification probability for that protein. Protein ID FDRs were calculated using the same isotonic regression approach (Hather et al., 2010; Higdon and Kolker, 2007).

Meta-analysis

Lists of proteins were combined across experiments at fixed FDRs (0.5, 1, 2, and 5%) for each experiment. The FDRs for the

^aUnderlying the use of raw data is the need for a minimum amount of meta-data, such as those required to set search parameters. Greater amounts of meta-data increase the utility of raw data by allowing further optimization, application of more advanced models or extension to quantitative analysis.

TABLE 1. SUMMARY OF INDIVIDUAL EXPERIMENTS

Source ^a	Primary author	Experiment name	Spectra ID	Protein IDs at FDR			
				0.5%	1%	2%	5%
PA	Swaney	Decision Tree Driven MS/MS	39,543	1,465	1,607	1,673	1,765
OPD	Prince	Orbitrap optimization	19,616	1,170	1,198	1,237	1,303
PA	Breci	Analysis of different prep methods	64,625	752	831	886	941
PD	Piening	QC metrics for LC-MS	121,701	738	775	834	932
PD	Maynard	Multidimensional LC-MS	5,5324	532	561	586	619
PA	Martin	Testing fraction on LCQ	5,509	435	444	456	471
OPD	Li	Gradient Experiment	4,870	351	369	390	441
OPD	Lu	Yeast grown in rich media	7,327	263	283	296	324
OPD	Lu	One-Carbon Metabolism	5,872	217	246	276	312
OPD	Lu	Growth in synthetic minimal media	3,449	146	177	192	218
OPD	Lu	Growth in YPD	6,516	131	145	187	208

^aPA, peptide atlas; PD, peptidome; OPD, open proteomics database.

combined list were calculated based on worst case error rate (circumstance 2) and based on the number of unique reshuffled protein sequence identifications (circumstance 4). The worst case occurs if all false positive identifications are unique and the worst case FDR can be estimated as follows:

$$FDR = \frac{N_{FP}}{N_T} = \frac{p \sum N_i}{\sum_{Unique} N_i} \quad (1)$$

where p is the experiment specific FDR and N_i is the number of protein identifications in the study at the given FDR threshold. The combined FDR is directly calculated by the number of unique reshuffled proteins IDs:

$$FDR = \frac{N_{RS}}{N_T} \quad (2)$$

where N_{RS} is the total number of unique reshuffled identifications across all experiments for given experiment FDR threshold and N_T is the total number of unique protein IDs across all the experiments. If continuous FDR thresholds (q -values) are available for each experiment (circumstances 2 and 4), then a threshold can be chosen for each experiment to achieve a desired overall FDR.

Combined analysis: unweighted and weighted

A combined analysis of all the data was conducted by generating protein summary statistics based on the peptide

identifications from all the individual proteomics experiments. Protein identification for the combined analysis was based on a weighted average of the number of unique peptide identifications (90% or greater probability) across all experiments and the maximum peptide probability across all peptide identifications (log of 1 minus the maximum probability). The weights were calculated based on a logistic regression model (Higdon and Kolker, 2007; Higdon et al., 2005). In addition, a weighted model was calculated by including indicator variables for each experiment of whether or not a protein was identified in that experiment. This gives variable weights to identifications made in different experiments and more weight to proteins identified in multiple experiments.

Results and Discussion

As can be seen in Table 1, the number of spectra and protein identifications vary greatly between different experiments. The single experiment with the most protein identifications identified was the Swaney et al. (2008) dataset, which identified 1,607 proteins at a 1% FDR. The Swaney dataset identified over nine times the number of proteins as the smallest single dataset. The large datasets represent comprehensive analyses of the yeast proteome, whereas some of the smaller datasets were generated for the purposes of testing instrumentation, experimental protocols, and data analysis methods.

An initial glance at circumstances 1 and 3 in Table 2 shows that simple union of experiments does increase the total

TABLE 2. SUMMARY OF META-ANALYSIS RESULTS FOR PROTEIN IDENTIFICATION ACROSS DIFFERENT CIRCUMSTANCES

FDR	1. LWC ^{bcd}	2. THWC ^c	3. LR ^{bcd}	4. TR ^c	5a. UM ^c	5b. WM ^c
0.5%	1,847 (1.4%)	1,244	1,847 (1.2%)	1,264	1,443	1,670
1%	1,977 (3.4%)	1,745	1,977 (2.8%)	1,778	1,723	1,895
2%	2,083 (7.0%)	1,877	2,083 (5.8%)	1,946	1,837	1,946
5%	2,244 (16.8%)	2,017	2,244 (14.4%)	2,062	1,954	2,068
2–5% ^a	NA	140 (45%)	NA	116 (54%)	117 (52%)	122 (53%)

^aThis category corresponds to ID made at 5% FDR but not 2% FDR with the corresponding FDR in parentheses.

^bNumbers in parentheses represent the recalculated FDR for combined lists.

^cLWC, lists worst case; THWC, threshold worst case; LR, lists randomized; TR, threshold randomized; UM, unweighted model; WM, weighted model.

^dLWC and LR yield identical protein lists. Only the method for calculating the FDR differs.

TABLE 3. WEIGHTS OF EXPERIMENTS IN COMBINED YEAST META-ANALYSIS PROTEIN ID MODEL

<i>Data-base Source</i>	<i>Primary author</i>	<i>Experiment name</i>	<i>Experiment weights</i>
PA	Swaney	Decision Tree Driven MS/MS	−1.02
OPD	Prince	Orbitrap optimization	−0.46
PA	Breci	Analysis of different prep methods	0.28
PD	Piening	QC metrics for LC-MS	0.05
PD	Maynard	Multidimensional LC-MS	0.51
PA	Martin	Testing fraction on LCQ	0.22
OPD	Li	Gradient Experiment	0.18
OPD	Lu	Yeast grown in rich media	0.23
OPD	Lu	One-Carbon Metabolism	−0.17
OPD	Lu	Growth in synthetic minimal media	0.83
OPD	Lu	Growth in YPD	0.22

The weighted combined model is based on a logistical regression model of protein summary scores versus whether or not a protein identification is a reshuffled or true sequence. The weights are parameter estimates for indicator variables of whether or not a protein was identified in that particular experiment. Positive weights imply a protein ID is more likely to be false (reshuffled) for that experiment, whereas negative values are less likely to be false. Therefore, protein IDs from experiments with negative values are more heavily weighted in the combined identification model.

number of protein identifications but at a cost of roughly tripling the FDR relative to the per experiment FDR. This clearly demonstrates that naively combining protein lists without thought for error rates dramatically increases the FDR. When the FDR threshold is decreased to control the FDR in the combined dataset (circumstances 2 and 4), the number of identifications is reduced and only exceeds the best single experiment by a relatively small amount ($\sim 10\%$). The threshold worst-case FDR estimate is actually quite close to the direct FDR estimate based on randomized database matches and yields nearly as many IDs despite being a conservative approach.

When data were analyzed using a combined model of all experiments (circumstance 5), the number of protein identifications at 1% FDR or more was actually slightly reduced when compared with the simple combining of protein lists (Table 2). This may be an artifact of equally weighting identifications from large experiments with smaller ones. Including weights for the individual experiments does give an increase in the number of protein identifications at a 1% FDR or below (Table 2). Weights were given to each exper-

iment as described earlier. The magnitude of the weights show that the first two experiments in Table 1 are the most beneficial for increasing the total number of protein identifications (see Table 3). These experiments had the largest number of protein identifications individually, and so might be expected to contribute the most. At a 5% FDR, the weighted model does not improve the number of protein identifications. This is likely due to the fact that between 2 and 5% FDR (Table 2) the error rate of additional protein IDs is very high (54%), and trying to add protein IDs by increasing the FDR at this stage yields an unacceptably high proportion of false positives.

Figure 1 shows the number of protein identifications by adding each experiment sequentially from the list in Table 1. The addition of each new experiment shows diminishing and even negative returns, although the number of identifications deduced by combining lists reaches its maximum or nearly so by the second experiment. The number of identifications from the weighted combined model flattens out after the second experiment. Adding additional experiments after a certain point appears to achieve little and may increase the noise,

FIG. 1. Comparison of the number of protein identifications at 0.5 and 1% FDR by sequentially combining the experiments in Table 1. This was done for the combined weighted model (red line) and the simple combination of protein lists (blue line).

particularly at low error rates. This is likely because smaller experiments add less new unique identification; however, small specialized experiments aimed at identifying rare proteins could still be beneficial.

Conclusions

The best case for meta-analysis is a reanalysis of the raw MS data. This is not only to maximize protein identifications, but also to ensure that common and comparable data analysis methods are used and that consistent sequence databases and protein identifiers are used (circumstance 5). However, reanalysis may not be feasible in many cases due to time limitations, computing power restrictions, lack of software, or inaccessibility of raw data.

Additionally, we conclude based on the results from this case study on *Saccharomyces cerevisiae* that, in the absence of raw data, several aspects of meta-analysis must be kept in mind. First, a reasonable estimation of the FDR can be found using the worst-case combination approach (circumstance 2), which yields nearly as many protein identifications at 1% FDR as a combined raw data analysis (1,745 vs. 1,850). Having randomized data or decoy protein identification and continuous FDR thresholds allows further refinement of protein identifications and estimations of the FDR. Second, the use of too many small experiments that add few unique protein identifications may only add noise to the final combined list. Experiments that are most beneficial to an increased understanding of the proteome are those that provide comprehensive or unique coverage of the proteome through the use of better instrumentation or experimental approaches. Third, failure to appropriately combine proteomic datasets can lead to FDRs that are greatly underestimated. This can be reasonably compensated for by weighting experiments in a combined analysis or with proper care taken to choose only the most solid experiments. Finally, although sometimes it seems that simply combining lists is unavoidable, recent initiatives in proteomics data capture and management make it clear that the future of proteomics will be raw data reanalysis with meta-data guidance. This study can serve as an interim approach to meta-analysis until then.

Acknowledgments

We thank Andrew Bauman, Gregory Hather, Simon Fortelny, Natali Kolker, Brenton Louie, Elizabeth Stewart, and Gerald van Belle for their critical reading. This work was funded by NIH 5R01 GM076680-02 (NIGMS) and UO1 DK072473 (NIDDK), NSF DBI-0544757, and NSF-07140 to E.K.

Author Disclosure Statement

The authors have declared no conflict of interest.

References

- Breci, L., Hattrup, E., Keeler, M., Letarte, J., Johnson, R., and Haynes, P.A. (2005). Comprehensive proteomics in yeast using chromatographic fractionation, gas phase fractionation, protein gel electrophoresis, and isoelectric focusing. *Proteomics* 5, 2018–2028.
- Choi, J.K., Yu U., Kim S., and Yoo O.J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19, i84–i90.
- DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* 7, 177–188.
- Desiere, F., Deutsch, E.W., Nesvizhskii, A.I., Mallick, P., King, N.L., Eng, J.K., et al. (2006). Integration of peptide sequences obtained by high-throughput mass spectrometry with the Human genome. *Genome Biol.* 6, R9.
- Deutsch, E.W., Lam, H., and Aebersold, R. (2008). Data analysis and tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* 33, 18–25.
- Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976–989.
- Everett, L., Bierl, C., and Master, S.R. (2010). Unbiased statistical analysis for multi-stage proteomic search strategies. *J. Proteome Res.* 9, 700–707.
- Farrer, L.A., Cupples, A., Haines, J.L., Hyman, B., Kukull, W.A., Mayeux, R., et al. (1997). Effects of age, sex, and ethnicity on the association between Apolipoprotein E genotype and Alzheimer disease. *JAMA* 278, 1349–1356.
- Fenyö, D., and Beavis, R.C. (2003). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* 75, 768–774.
- Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., et al. (2004). Open mass spectrometry search algorithm. *J. Proteomics Res.* 3, 958–864.
- Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educ. Res.* 5, 3–8.
- Gray, J. (2009). Jim Gray on eScience: a transformed scientific method. In *The Fourth Paradigm: Data-intensive Scientific Discovery*. T. Hey, S. Tansley, and K. Tolle, eds. (Microsoft Research, Redmond, WA), pp. xvii–xxxi.
- Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999.
- Hather, G., Higdon, R., Bauman, A., von Haller, P.D., and Kolker, E. (2010). Estimating false discovery rates for peptide and protein identification using randomized databases. *Proteomics* (in press).
- Higdon, R., and Kolker, E. (2007) A predictive model for identifying proteins by a single peptide match. *Bioinformatics* 23, 277–280.
- Higdon, R., Kolker, N., Picone, A., van Belle, G., and Kolker, E. (2004). LIP index for classification using MS/MS and Sequest search via logistic regression. *OMICS* 8, 357–369.
- Higdon, R., Hogan, J.M., Van Belle, G., and Kolker, E. (2005). Randomized sequence databases for tandem mass spectrometry peptide and protein identification. *OMICS* 9, 364–379.
- Higdon, R., Hogan, J.M., Kolker, N., van Belle, G., and Kolker, E. (2007). Experiment-specific estimation of peptide identification probabilities using a randomized database. *OMICS* 11, 351–365.
- Higdon, R., van Belle, G., and Kolker, E. (2008). A note on the false discovery rate and inconsistent comparison between experiments. *Bioinformatics* 24, 1225–1228.
- Larsson, O., Wennmalm, K., and Sandberg, R. (2006). Comparative microarray analysis. *OMICS* 10, 381–397.
- Maynard, D.M., Masudaa, J., Yang, X., Kowalak, J.A., and Markey, S.P. (2004). Characterizing complex peptide mixtures using a multi-dimensional liquid chromatography-mass

- spectrometry system: *Saccharomyces cerevisiae* as a model system. *J. Chromatogr. B* 810, 69–76.
- Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., and Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.* 19, 570–577.
- Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567.
- Piening, B.D., Wang, P., Bangur, C.S., Whiteaker, J., Zhang, H., Feng, L.C., et al. (2006). Quality control metrics for LC-MS feature detection tools demonstrated on *Saccharomyces cerevisiae* proteomic profiles. *J. Proteome Res.* 5, 1527–1534.
- Prince, J.T., Carlson, M.W., Wang, R., Lu, P., and Marcotte, E.M. (2004). The need for a public proteomics repository. *Nat. Biotechnol.* 22, 471–472.
- Rhoades, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D., Chinnaiyan, A.M. (2002). Meta-analysis of microarrays. *Cancer Res.* 62, 4427–4433.
- Slotta, D.J., Barrett, T., and Edgar, R. (2009). NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. *Nature Biotech.* 27, 600–601.
- Swaney, D.L., McAlister, G.C., and Coon, J.J. (2008). Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* 5, 959–964.
- Vizcaíno, J.A., Côté, R., Reisinger, F., Foster, J.M., Mueller, M., Rameseder, J., Hermjakob, H., et al. (2009). A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* 9, 4276–4283.

Address correspondence to:

Dr. Eugene Kolker
Seattle Children's Research Institute
1900 Ninth Avenue
Mail: M/S C9S-10
Seattle, WA 98101

E-mail: Eugene.Kolker@seattlechildrens.org