Published in final edited form as:

Proteomics. 2013 February ; 13(0): 493–503. doi:10.1002/pmic.201200269.

A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments

Melissa M. Matzke, Joseph N. Brown, Marina A. Gritsenko, Thomas O. Metz, Joel G. Pounds, Karin D. Rodland, Anil K. Shukla, Richard D. Smith, Katrina M. Waters, Jason E. McDermott, and Bobbie-Jo Webb-Robertson

Pacific Northwest National Laboratory

Abstract

Liquid chromatography coupled with mass spectrometry (LC-MS) is widely used to identify and quantify peptides in complex biological samples. In particular, label-free shotgun proteomics is highly effective for the identification of peptides and subsequently obtaining a global protein profile of a sample. As a result, this approach is widely used for discovery studies. Typically, the objective of these discovery studies is to identify proteins that are affected by some condition of interest (e.g., disease, exposure). However, for complex biological samples, label-free LC-MS proteomics experiments measure peptides and do not directly yield protein quantities. Thus, protein quantification must be inferred from one or more measured peptides. In recent years many computational approaches to compute relative protein quantification of label-free LC-MS data have been published. In this review, we examine the most commonly employed quantification approaches to compute relative protein abundance from peak intensity values, evaluate their individual merits, and discuss challenges in the use of the various computational approaches.

Keywords

label-free; peak intensity; protein quantification; relative

1 Introduction

High-throughput shotgun (discovery) proteomics experiments undoubtedly have had a significant impact in life science research. Protein abundance in a sample is typically measured by enzymatically digesting the proteins into peptides that are subsequently separated by liquid chromatography, ionized, and entered into a MS. Hypothesis-generating, discovery proteomics studies are optimized for peptide identification, typically using a small number of biological samples, but provide a large, yet incomplete, view of protein abundance [1]. However, the accurate quantitation of protein expression change in complex biological samples is of equal importance [1–4], and necessary for and biological modeling such as network and pathway analyses. Unique to label-free shotgun proteomics is the potential to measure peptides present in a sample and subsequently to determine the abundance of proteins identified across different samples [5].

Corresponding Author: Bobbie-Jo M. Webb-Robertson, P.O. BOX 999, J4-33, Richland, WA 99352, Phone: 509-375-2292, Fax: 509-372-4720, bj@pnl.gov.

The authors have declared no conflict of interest.

There are two approaches to label-free protein quantification: absolute and relative. The first approach, absolute protein quantification, has been traditionally determined using stable isotope –labeled peptides, however can be determined in a label-free manner with or without reference standards [3, 6–7]. The purpose of absolute protein quantification is to present protein expression changes in exact amounts or concentration. A comprehensive review of label-free absolute protein quantification methods, such as APEX [8], can be found in Arike et al. [6]. The second approach, relative protein quantification, is used to present protein expression changes relative to another sample (e.g., control) [2, 4]. The challenge to relative protein quantification lies in discerning signal from noise. There is a multitude of sources of variation which contribute to the noise in the LC-MS abundance data – for example, errors in sample preparation and processing, incorrect processing of the raw data, multiple charge states detected for a peptide, missed or incorrect peptide identifications, ‘shared’ peptides (i.e., peptides common to different proteins), protein isoforms that are indistinguishable on the basis of the detected peptides, post-translational modifications of proteins producing modified peptides that are often not identified, incomplete response across the pool of technical/biology replicates, or inconsistent response profiles among multiple peptides within a protein. Therefore, protein quantification must always be considered in light of a thorough understanding of the performance, and possible errors of all the previous data processing steps [9]. Moreover, the problem of relative protein quantification is worthy of discussion since there is no agreement about best practice, and often little is known concerning the relative contribution of the sources of variation noted above.

Two broad computational approaches are generally used to estimate relative protein abundance in label-free experiments: use of relative peptide peak intensities as a surrogate for abundance, and use of spectral properties such as counts [9]. The first, *spectral counting*, relies on counting the number of spectra that map to a given protein across multiple LC-MS analyses; the second, *peptide peak intensity*, uses the area under or the height of the precursor ion or fragment reporter ion peaks as a proxy for peptide abundance, and sometimes involves the introduction of labeled (e.g., with stable isotopes) versions of the same peptide for increased precision [2, 4, 10]. As a more direct measure of peptide abundance than spectral counts, peptide peak intensity approach is advantageous [11]. Specifically label-free protein quantification avoids the added cost and complexity of such labeling approaches, and aims to correlate the mass spectrometric signal of intact proteolytic peptides with the relative protein quantity directly [2]. In a typical quantitative proteomics experiment, although peptide abundance is inferred from ion intensity, relative protein abundance levels are calculated as a function of peptide abundance.

The focus of this review is on the process of inferring a single estimate of relative protein abundance from one or more measured peptide peak intensities (i.e., abundances). Several computational approaches to relative protein quantification in label-free LC-MS experiments are described. We examine commonly used relative protein quantification approaches for their individual merits and discuss challenges in the use of the various computational approaches.

2 Protein Quantification Methods

Prior to protein quantification using peptide peak intensity values several computational steps are performed, such as peptide identification, data quality assessment, normalization, and protein inference. All of these tasks have unique challenges and although essential to proteomics, they are not the focus of this review. In this review we focus exclusively on the use of peptide abundances resulting from label-free LC-MS proteomics experiments for the calculation of relative protein abundance. Due to the complexity and variability associated with label-free proteomics data there is no vetted approach to protein quantification as an

optimal approach. Thus, in this review we consider three categories of computational approaches commonly employed: (1) **additive**, in which peptide abundances are combined in an additive manner, (2) **reference**, in which a peptide is chosen to standardize all other peptides within a protein, and (3) **linear models**, in which peptide abundance values are modeled using least squares (Table 1). There are two additional protein quantification methods listed but not reviewed. The first given by Du et al. estimates protein abundance in temporal data using identified significant temporal patterns [12]. The second is a linear programming approach to protein quantification using shared peptide information [13].

2.1 Additive

This category of protein quantification approaches combines peptide abundances by simply summing, or standardizing the sum, to estimate relative protein abundance. This methodology is very simplistic, however is used in practice relatively frequently. For comparative purposes, we consider three additive approaches that have been described previously, 1) the sum of the top 3 most abundant peptides per protein (**SUM1**) [10]; 2) the average of the top 3 most abundant peptides per protein (**SUM2**) [14]; and, 3) the sum of all peptides divided by the protein length (**SUM3**) [10].

2.2. Reference

The reference approach to protein quantification identifies a single peptide within a protein, given a specific criteria, that is used as the scaling baseline for which all other peptides within the protein are compared [15]. The first approach (**REF1**) uses the peptide with the least amount of missing data as the reference peptide to which the remaining peptides are scaled. The protein abundance is the median of the scaled peptide abundances. The second approach (**REF2**) first centers and scales peptide abundances using the median and standard deviation, and then calculates the protein abundance as the median peptide abundance score.

Both approaches are available in the freeware DanteR (this software is an unpublished update to [15], <http://omics.pnl.gov/software>). As implemented, *f*-**REF1** and *f*-**REF2** filter the peptides before quantifying proteins (Supporting Information Figure S1 and S2). We allowed proteins identified by a single peptide to be included, however all other filtering options were left as default values. In addition, both quantification approaches were implemented without the additional peptide filters in Matlab® R2011a (*u*-**REF1**, *u*-**REF2**; script available in Supporting Information Table S1).

2.3 Linear Models

2.3.1 Additive model – main effects—A protein-level additive model accounting for the main effects of peptides and proteins is proposed by Karpievitch et al. (**LM1**) [16]. The **LM1** model is written as:

$$y_{ijkl} = \text{Prot}_i + \text{Pep}_{ij} + \text{Grp}_{ik} + \text{error}_{ijkl} \quad (1)$$

where y_{ijkl} is the intensity for protein i and peptide j in comparison group k for sample l . The **LM1** algorithm filters proteins for which no collection of peptides can produce an identifiable model. In addition, it uses a greedy search algorithm to select peptide sets for each remaining protein that produce optimal information content and filter out the rest. The **LM1** protein quantification approach is implemented in DanteR.

2.3.2 Additive model – main effects and interactions—Protein-level quantification using a linear model in which the interaction of peptides and comparison groups is modeled in addition to the main effects of each is proposed by Clough et al. (**LM2**) [17]. The **LM2** fixed-effects model is written as:

$$y_{ijkl} = \text{Pep}_i + \text{Grp}_j + (\text{Pep} * \text{Grp})_{ij} + S_k + \text{error}_{ijkl} \quad (2)$$

where y_{ijkl} is the intensity for peptide i in comparison group j for biological replicate k . Clough et al. assumes missing peptide values are of low abundance, and thus imputes with the average minimum observed intensity across all biological replicates. This imputation is performed for each peptide within a group. This is the only approach that directly performs imputation with a potential caveat that the variance structure of the data may be changed. The **LM2** protein quantification approach is implemented in the *MSstats* package [18] available for R software (<http://www.R-project.org>).

3 LC-MS Data

Two real world datasets are used to evaluate the protein quantification methods. These datasets represent typical LC-MS data that a user would present for protein quantification, including a highly complex dataset with considerable variability (human plasma) and a more controlled experiment with less variability (inbred mouse lung tissue). This review evaluates how well each method performs on these datasets from a user perspective. LC-MS materials and methods for processing these samples are detailed in Supporting Information.

3.1 Human plasma samples

Human plasma samples from adult volunteer subjects with normal glucose tolerance (NGT), impaired glucose tolerance (IGT), and type 2 diabetes mellitus (DIA) were collected as part of the Screening for Impaired Glucose Tolerance (SIGT) study [19] and were provided to Pacific Northwest National Laboratory via the National Institute of Diabetes and Digestive and Kidney Diseases Biosample Repository. Approval for the conduct of this programmatic research was obtained from the Institutional Review Board of Pacific Northwest National Laboratory. All samples were received frozen on dry ice. A hierarchical cluster analysis was performed on the SIGT participants to identify the two NGT ($n = 50$) and IGT ($n = 50$) individuals that best matched the DIA individuals ($n = 25$) in terms of age, race, sex, and blood donation site.

3.2 Mouse lung tissue samples

Lung tissue samples of 32 young male C57BL/6 mice, 8 in each of the 4 possible factor combinations of a 2-factor experiment; factor one included sham controls (SC) and exposure to lipopolysaccharide (LPS) and factor two included normal weight (NW) and diet induced obesity (OB). Obesity was based on the same diet described previously [20] and mice were exposed to LPS by nose-only inhalation exposure at the target concentration of 0.5 $\mu\text{g/L}$ or filtered air (sham control) for 1 hr/day for total of 4 days, over 2 weeks as follow: 1 day on exposure, 1 day off exposure, 1 day on exposure, 3 days off exposure, 1 day on exposure, 1 day off exposure, 1 day on exposure, sacrifice the day following their last exposure. On the morning following their last exposure, animals were sacrificed (overdosed with pentobarbital/bleeding/exsanguination).

3.3 Statistical Pre-processing

The abundance values for the final peptide identifications were processed in a series of steps using MatLab® R2012a. Peptide abundances were transformed to the $\log_{10}$ scale then processed to identify and remove contaminant peptides and proteins (e.g., porcine trypsin peptide fragments resulting from autocatalysis and peptides derived from spiked-in quality control proteins). In addition, peptides with an insufficient amount of data across the set of samples [21] and LC-MS runs that showed significant deviation from the standard behavior of all LC-MS analyses within an experiment [22] were removed. Peptides were normalized

across the technical replicates and averaged within each biological sample [20]. Missing data values were left as blank (not imputed) prior to processing with each protein quantification method to evaluate how well each approach deals with this factor of the data. The outcome of each step of pre-processing is outlined in Supporting Information Figures S3 and S4.

4 Results

The proteomics datasets were presented for quantification as a matrix of peptides by biological samples. The first dataset (human plasma) contained 119 samples measured across 14,254 peptides, of which 12,444 were unique. Thus there were 1,810 peptides that mapped to more than one protein (i.e., 'shared'). These 14,254 peptides mapped to 1,515 proteins. This data represents a typical highly complex human dataset allowing for highly homologous proteins or isoforms of a parent protein. The second dataset (mouse lung) consisted of 6,295 peptides, mapped to 1,679 proteins for 32 samples. In this case the 1,679 proteins are based on unique protein families as identified with protein prophet [23]. Thus, this data represents again a complex dataset, but with much of the challenge associated with shared peptides removed from consideration. The number of peptides associated with a protein ranged from 1 to 514 for the human data, and 1 to 113 for the mouse data (Table 2).

Each protein quantification method was run on the complete set of proteins for both datasets, including proteins identified by a single peptide. The number of proteins with biological replicate level abundance estimates is listed in Table 3. The **SUM1**, **SUM2**, **SUM3**, *u*-**REF1** and *u*-**REF2** methods calculate biological replicate level abundances for the complete set of proteins for both datasets. The *f*-**REF1** and *f*-**REF2** methods resulted in reduced lists of protein abundances for both human and mouse datasets. The filtered reference methods removed single peptides so that 87% and 62% of proteins were not estimated in the human and mouse datasets, respectively. The **LM1** method did not return biological replicate level protein abundance estimates, however the number of proteins was inferred from the filtered peptide list. The filtering algorithm used by the **LM1** method reduces the number of proteins for which it could estimate protein abundances. The **LM2** method calculated protein abundances for the complete set of mouse proteins, although had significant computational difficulties for the human proteins. As a result, the **LM2** method did not estimate abundances for 40% (618 out of 1515) of the human proteins. Of these proteins, 90% (558 out of 618) were identified by a single peptide with global response frequency of 1.6% (2 responses in a single group) to 100%. The balance of proteins without abundance estimates were identified by 2 to 16 peptides. Although we recognize the **LM2** algorithm will estimate protein abundance using a single peptide with 100% global response (it does so for the mouse lung tissue data), it is unclear to us at what point the response frequency to group sample size ratio prohibits **LM2** from estimating protein abundance.

The human dataset contained a single protein identified by 514 peptides. The **SUM** and **REF** methods estimated protein abundance following the specific algorithms; the **LM1** approach was unable to estimate protein abundance, likely due to the greedy algorithm used for optimization; and, the **LM2** approach used a random subset of peptides to fit a fixed effects linear model.

The **SUM1** approach, for which the top 3 peptide abundance values are summed, typically results in noisy within group protein abundance values. As a relative protein quantification approach this behavior is undesirable as it will lessen the sensitivity of subsequent comparative statistical analyses. Similarly, the **REF2** approach (modified z-score calculation) tends to result in highly variable within group protein abundance values. Conversely, the **SUM3** (top 3 peptide abundance values scaled by the protein length)

typically averages out noise, and subsequently any biological effect that may be present. These behaviors can be seen in Supporting Information Figure S5. The **SUM1**, **SUM3** and **REF2** approaches are removed from further consideration due to the lack of cluster performance. In addition, the **LM1** approach cannot be further considered since the biological replicate level protein abundance values were not returned. Therefore, the remainder of the review will focus on the **SUM2**, *filtered* and *unfiltered* **REF1**, and **LM2** relative protein quantification methods.

Behavior of the **SUM2**, *f*-**REF1** and *u*-**REF1**, and **LM2** protein quantification methods were explored in single peptide, low content (2, 3 or 4 peptides) and high content (15–20 peptides) proteins. In addition, extreme peptide response frequency (low response frequency, all peptides have less than 25% global response; high response frequency, all peptides have greater than 80% global response) were reviewed for single peptide and low content proteins; whereas, the high content proteins were reviewed in total due to the distribution of peptide response frequency across the increased number of peptides. The distribution of proteins based on content and peptide response frequency, listed in Table 4, shows that both datasets contain a range of “confidently identified” proteins. For example, there are 269 single peptide mouse proteins with no more than a 25% global response frequency. However, the low response frequency may be due to response in the affected group but no response, likely below the detection limit, in the control group (Figure 1). In this case, only the **LM2** quantification approach resulted in imputed abundance values for those groups without response, however the *u*-**REF1** and **LM2** approaches resulted in the same protein abundance values for the group with response. At first, this may appear to be a trivial result; however it provides the foundation for results from proteins with increased content (i.e., number of peptides) and response frequency. Now consider the example of a low content human protein CD248 with 2 peptides of high response (80%) frequency (Figure 2). The *f*-**REF1** and *u*-**REF1** approaches resulted in almost identical protein abundance values, whereas the **LM2** approach resulted in the same imputed value for all biological replicates. Although visual inspection supports the likely conclusion that there is no evidence of a statistically significant difference among the three diabetes group, the imputed outcome of the **LM2** approach is surprising.

The high content proteins containing 15–20 peptides did not exhibit extreme behavior in the global response frequency. The response frequency ranged from a median of 15% to 69%; that is, there were no proteins for which all peptides had a low or high response frequency. As an example of the behavior of the protein quantification approaches, consider the human protein 4F2. It was identified by 19 peptides with a response frequency range of 6% (7 responses out of 119 for a single peptide) to 100%, and a median global response frequency of 73% (Figure 3). The protein abundance estimates resulting from the **SUM2** (average of the 3 most abundant peptides) approach are strikingly variable. The *f*-**REF1** and *u*-**REF1** approaches track in a similar fashion due to the availability of peptide abundance values. That is, the *f*-**REF1** approach has sufficient amount of data that passes its filters, and thus is not influenced by the peptides with low response frequency. The *f*-**REF1** and *u*-**REF1** protein abundance estimates have a moderate amount of within group variability. Relative to the **SUM2** and both **REF1** approaches, the **LM2** approach results in minimal within group variability. This is likely due to the **LM2** linear model accounting for both factors *Peptide* and *Group*. However, the estimates appear to be highly influenced by the imputed data or low observed peptide abundance values.

5 Discussion

Peptide filtering is likely to have the greatest amount of influence on the any of the protein quantification methods reviewed. Filtering algorithms are employed to remove those

peptides for which we have lower confidence solely based on the frequency of response. The f -REF1 and f -REF2 methods require the user consider additional peptide filtering (minimum presence of at least one peptide, Grubbs' test for peptide outliers, etc) before the algorithm is executed. Worthy of note, none of the protein quantification methods reviewed incorporate spectra to peptide information in the algorithm. One attempt at the inclusion of spectra to peptide information in the estimation of relative protein abundance is the use of shared peptides [13]. However it is unclear if the results of this algorithm would be influenced by the degree of peptide filtering.

The influence of peptide filtering can easily be seen in a simple exercise by selecting a protein identified by a moderate number of peptides and incrementally filtering peptides based on global response frequency. High content mouse protein AMPL was identified by 18 peptides with global response frequency between 12.5% (4 out of 32, for which 2 responses occurred in RW-LPS and 2 responses in OB-LPS) and 100%. The LM2 relative abundance estimates were highly influenced by the amount of missing data resulting from the conservative filtering which allowed peptides with roughly 87% missing values to be mostly imputed with the average minimum observed intensity across all biological replicates. Figure 4A shows the protein abundance estimates accounting for all 18 peptides identified for the mouse protein AMPL. The LM2 relative abundance values are quantitatively less than those values calculated using the SUM2, f -REF1 and u -REF1 approaches, and thus likely due to the imputation the within group variability of LM2 estimates is smaller. As the peptide response frequency increases (Figure 4B, 10 peptides; Figure 4C, 5 peptides), and thus fewer values are imputed, the LM2 estimates become more like the estimates from SUM2, f -REF1 and u -REF1. The sum and reference-based methods are less likely to be influenced by the missing data since both the methods will use the most abundant peptides to compute protein abundance values. There remains the quandary of the confidence in the peptides with low response frequency – should these peptides be filtered, and what is the ratio of response to group size necessary for a peptide to be considered reliable?

Best practices for relative protein quantification methodology are not agreed upon, nor is there a single software package that does it all, and thus, the protein quantification method chosen may be a result of the user's capabilities. It is more than likely the user will not be an expert in the potentially multiple software packages required to move peptide abundance data through the pre-processing pipeline, protein quantification, comparative statistical analysis of groups, and biological modeling such as network and pathway analyses. It is this assumption we base our evaluation of the ease of use of the protein quantification methods. When considering the ease of use, we have defined six evaluation factors: 1) accessibility and set-up, 2) interface (GUI/script), 3) quality and intuitiveness of user help documentation, 4) importing data, 5) analysis run time, and 6) exporting data. To generate a usability profile, each protein quantification method was evaluated on a 5-point scale ranging from extremely challenging and requires expert user knowledge to extremely easy and a novice user would likely be comfortable in its use. The usability of each approach is compared visually by representing the individual characteristics on each of the six axes. Usability profiles for the evaluated protein quantification approaches are presented in Figure 5 as a radar plot, with the center point representing the a scored value of 1 (extremely challenging/suitable for a expert user) increasing in ease moving from the center to the edge representing a scored value of 5 (extremely easy/suitable for a novice user). The SUM, filtered and unfiltered REF, and LM2 algorithms were evaluated on a Dell system, Windows 7 Enterprise, Intel Xeon CPU W3503 @ 2.4GHz, with 12.0 GB RAM; the LM1 algorithm was evaluated on Dell system, Windows 7 Enterprise, Intel Core™2 Duo CPU E6550 @ 2.33 GHz, with 4.0 GB RAM.

The **SUM** and **LM2** approaches were the most easily accessed. The **SUM** approach depending on the size of the peptide abundance dataset may require software no more sophisticated than MS Excel. Although implemented in Matlab for this review, the scripting knowledge required was minimal. The time to run all three **SUM** approaches in MatLab was a matter of minutes. However, these approaches offer the least in terms of statistical rigor.

The **LM2** approach is easily downloaded, although R software must be installed separately, in the *MSstats* package. Importing and exporting the data does require very basic R script knowledge. The user documentation for the protein quantification function for the **LM2** approach is excellent. The **LM2** algorithm was cleanly executed in a moderate amount of time (run time was approximately 2 hours) for the mouse lung tissue dataset, whereas the execution of the algorithm for the human plasma dataset was troublesome. The `fitModels()` function failed and the error message returned did not provide an intuitive explanation for the failure. We believe the failure was due to the poor global response within the protein which is ultimately a filtering issue; but the user documentation does not provide guidance on this issue. The `fitModels()` function does return messages of progress, such as when peptides are missing data and values are to be imputed and when a randomly sampling of peptides is chosen rather than use an extremely large number of peptides to estimate protein abundance. However, overall, the **LM2** algorithm which is one of the best downloaded software, packaged in the *MSstats* library for R is well documented and relatively straightforward to use.

The DanteR software, which contain the **fREF** and **LM1** approaches, may prove difficult to install since a secondary software is required for the program to execute (Microsoft.NET Framework 2.0). For a novice user, the installation of this additional program may prove challenging. Subsequently, the **LM1** and **fREF** approaches are difficult to access due to the lack of informative user help documentation and poor import structure. In addition, the **LM1** method is accessed in two separate drop-down menus; however the differences between the two access points are unclear from the user help documentation and, most relevant to this review, biological replicate level protein abundance values are not output. The **LM1** approach, due to its implementation of a greedy search algorithm, is slow taking days to run the algorithm for each dataset evaluated; in contrast, the **fREF** approaches are complete in only a few minutes. Despite these issues, DanteR is the only software offering a Graphical User Interface.

The **u-REF** approach was inferred from the text in Polpitiya et al [15], and although implemented in MatLab, could have been scripted in another computing language such as R. There is no user help documentation to ensure the approach is implemented correctly, or to help with errors. Nevertheless, the **u-REF** approaches quickly return an estimated abundance value for each protein submitted to the algorithm, and thus caution is advised if peptides have been minimally filtered.

6 Concluding Remarks

Existing methods for protein quantitation are a good start, but significant challenges remain in dealing with the incomplete and noisy data returned from shotgun proteomics experiments. Software products available are still largely designed for expert users, or those proficient in *R*, *SAS*, or *MatLab*. In lieu of software, the methods for protein quantification that are available are statistically very simple (e.g., **SUM**). The use of real and complex biological dataset as would be input by the average researcher highlights the strength and weaknesses of each method from the perspective of the user.

We first note that protein quantification methods are highly sensitive to peptide filtering [16, 21]. Furthermore, this step needs to be incorporated into the protein quantification process with transparency. Second, metrics of variance at the peptide level (e.g., peptide identification confidence measures) are currently used for protein inference [24], but are not currently used for protein quantification [4]. These metrics could be highly valuable in the identification of the most conservative set of peptides to use for protein quantitation. Lastly, there remains a debate about missing data. It has been well described in proteomics that although some of the missing peptide identifications across samples is random error due to under sampling associated with the proteomics process, there are also data points missing due to biological effect (e.g., present in control and absent in treated individuals) [17, 21, 25]. The use of imputation in proteomics has advantages and disadvantages in protein quantification. Imputing with a constant will be reflected in the variability of the peptide and thus could have significant downstream effects on statistical analysis. However, avoiding imputation can greatly limit biological modeling (e.g., clustering, network inference).

Thus, accurate relative protein quantification is a non-trivial, multifaceted problem, with inherent limitations due to the nature of the measurements. In the context of the present state of proteomics, we believe that further improvements to protein quantification will be based upon much more robust peptide filtering, incorporation of peptide identification confidence metrics, and appropriate modeling of missing data. We also believe that these insights will serve to focus improvements in the measurement methods that will result in more confident peptide identifications, including modified peptides presently missed, as well as the development of approaches for more accurate quantitation. Most likely, these further advancement will require collaboration between multiple scientific disciplines, and will not be solved in isolation [4].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Computational work was supported by Laboratory Directed Research and Development at Pacific Northwest National Laboratory (PNNL) under the Signature Discovery Initiative (K.D.R, J.E.M). The human diabetes proteomics data was generated under National Institutes of Health grant DK071283 (R.D.S.) and the mouse lung LPS proteomics data was generated through National Institutes of Health grant U54-016015 (J.G.P.). Proteomics datasets originated from samples analyzed using capabilities developed under the support of the National Center for Research Resources (5P41RR018522-10) and the National Institute of General Medical Sciences (8 P41 GM103493-10) from the National Institutes of Health, and from the U.S. Department of Energy Office of Biological and Environmental Research (R.D.S). Proteomics data were collected and processed in the Environmental Molecular Sciences Laboratory (EMSL). EMSL is a national scientific user facility supported by the Department of Energy. All work was performed at PNNL, which is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under contract DE-AC06-76RL01830. We thank D. Daly for performing the hierarchical cluster analysis in support of the selection of matched samples for the human plasma dataset.

Abbreviations used in manuscript

SC	sham control
LPS	exposure to lipopolysaccharide
NW	normal weight
OB	diet induced obesity
SIGT	Screening for Impaired Glucose Tolerance

NGT	normal glucose tolerance
IGT	impaired glucose tolerance
DIA	type 2 diabetes mellitus

References

1. Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol.* 2005; 1:252–262. [PubMed: 16408053]
2. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem.* 2007; 389:1017–1031. [PubMed: 17668192]
3. Elliott MH, Smith DS, Parker CE, Borchers C. Current trends in quantitative proteomics. *J Mass Spectrom.* 2009; 44:1637–1660. [PubMed: 19957301]
4. Noble WS, MacCoss MJ. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput Biol.* 2012; 8:e1002296. [PubMed: 22291580]
5. Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol.* 2010; 28:710–721. [PubMed: 20622845]
6. Arike L, Valgepea K, Peil L, Nahku R, et al. Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*. *J Proteomics.* 2012
7. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics.* 2006; 5:144–156. [PubMed: 16219938]
8. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol.* 2007; 25:117–124. [PubMed: 17187058]
9. Martens L. Bioinformatics challenges in mass spectrometry-driven proteomics. *Methods in Molecular Biology.* 2011; 753:359–371. [PubMed: 21604135]
10. Ning K, Fermin D, Nesvizhskii AI. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res.* 2012; 11:2261–2271. [PubMed: 22329341]
11. Mallick P, Kuster B. Proteomics: a pragmatic perspective. *Nat Biotechnol.* 2010; 28:695–709. [PubMed: 20622844]
12. Du X, Callister SJ, Manes NP, Adkins JN, et al. A computational strategy to analyze label-free temporal bottom-up proteomics data. *J Proteome Res.* 2008; 7:2595–2604. [PubMed: 18442284]
13. Dost B, Bandeira N, Li X, Shen Z, et al. Accurate mass spectrometry based protein quantification via shared peptides. *J Comput Biol.* 2012; 19:337–348. [PubMed: 22414154]
14. Cheng FY, Blackburn K, Lin YM, Goshe MB, Williamson JD. Absolute protein quantification by LC/MS(E) for global analysis of salicylic acid-induced plant protein secretion responses. *J Proteome Res.* 2009; 8:82–93. [PubMed: 18998720]
15. Polpitiya A, Qian W, Jaitly N, Petyuk V, et al. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics.* 2008; 24:1556–1558. [PubMed: 18453552]
16. Karpievitch Y, Stanley J, Taverner T, Huang J, et al. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics.* 2009; 25:2028–2034. [PubMed: 19535538]
17. Clough T, Key M, Ott I, Ragg S, et al. Protein quantification in label-free LC-MS experiments. *Journal of Proteome Research.* 2009; 8:5275–5284. [PubMed: 19891509]
18. Clough, T.; Vitek, O. MSstats: Statistical protein quantification in label-free LC-MS experiments. 2011.
19. Phillips LS, Weintraub WS, Ziemer DC, Kolm P, et al. All pre-diabetes is not the same: metabolic and vascular risks of impaired fasting glucose at 100 versus 110 mg/dl: the Screening for Impaired Glucose Tolerance study 1 (SIGT 1). *Diabetes Care.* 2006; 29:1405–1407. [PubMed: 16732034]

20. Webb-Robertson BJ, Matzke MM, Jacobs JM, Pounds JG, Waters KM. A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors. *Proteomics*. 2011; 11:4736–4741. [PubMed: 22038874]
21. Webb-Robertson BJ, McCue LA, Waters KM, Matzke MM, et al. Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data. *J Proteome Res*. 2010; 9:5748–5756. [PubMed: 20831241]
22. Matzke MM, Waters KM, Metz TO, Jacobs JM, et al. Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics*. 2011; 27:2866–2872. [PubMed: 21852304]
23. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003; 75:4646–4658. [PubMed: 14632076]
24. Li YF, Arnold RJ, Li Y, Radivojac P, et al. A bayesian approach to protein inference problem in shotgun proteomics. *J Comput Biol*. 2009; 16:1183–1193. [PubMed: 19645593]
25. Wang X, Anderson GA, Smith RD, Dabney AR. A hybrid approach to protein differential expression in mass spectrometry-based proteomics. *Bioinformatics*. 2012; 28:1586–1591. [PubMed: 22522136]
26. Bukham Y, Dharsee M, Ewing R, Chu P, et al. Design and analysis of quantitative differential proteomics investigations using LC-MS technology. *Journal of Bioinformatics and Computational Biology*. 2008; 6:107–123. [PubMed: 18324749]
27. Daly D, Anderson K, Panisko E, Purvine S, et al. Mixed-effects statistical model for comparative LC-MS proteomics studies. *Journal of Proteome Research*. 2008; 7:1209–1217. [PubMed: 18251496]

Figure 1. Mouse protein AMPN protein abundance estimates

The mouse protein AMPN is identified by a single peptide. There are 4 out of 8 responses in the OB-SC group, and no response in the RW-SC, RW-LPS and OB-LPS groups. The **SUM2**, which is based on top 3 peptide abundance values, and **f-REF1**, which employs filters, do not return protein abundance estimates; the **u-REF1** returns protein abundance estimates for the OB-SC group only; and, the **LM2** approach returns protein abundance estimates for all samples for which the groups without a response are imputed.

Figure 2. Human protein CD248 protein abundance estimates

The human protein CD248 is identified by 2 peptides with more than 80% response frequency across the three groups. The **SUM2** approach, which is based on top 3 peptide abundance values, does not return protein abundance estimates; the **f-REF1** and **u-REF1** approaches return identical protein abundance estimates; whereas the **LM2** approach returns the same imputed value for all samples.

Figure 3. Human protein 4F2 protein abundance estimates

The human protein 4F2 is a high content protein identified by 19 peptides. The peptide abundance values and estimated protein abundance values are overlaid. The **SUM2** (average of the 3 most abundant peptides) protein abundance estimates are the most variable; the **f-REF1** and **u-REF1** approaches track in a similar fashion, and are less likely to be influenced by peptides with lower abundance values; and, the protein abundance estimates returned from the **LM2** approach, although have a relatively small amount of within group variability, appear to be highly influenced by either observed low abundance values or imputed abundance values.

Figure 4. Mouse protein AMPL protein abundance estimates

The high content mouse protein AMPL is identified by 18 peptides. Protein abundance estimates resulting from the use of (A) all 18 peptides identified, (B) 10 peptides (peptides with a low response frequency in any group filtered), and (C) 5 peptides (only peptides with $\geq 75\%$ global response frequency remain). The influence of the imputed values on the **LM2** estimated quantities can be seen by the movement of the estimated values toward the **SUM2** and **REF1** values. That is, the less impute data introduced by the LM2 approach results in protein abundance estimates, in quantity and within-group variability, which are more like the **SUM2** and **REF1** approaches.

Figure 5. Usability characteristics profiles of five relative protein quantification methods
Six basic tasks (accessibility, interface, user help, importing data, analysis run time, and exporting data) are rated for difficulty. A task is considered to be more difficult requiring expert knowledge if its rating is centered, and becomes less difficult requiring less expertise moving towards the edge of the axis.

Table 1

Computational approaches to relative protein quantification using peptide peak intensities resulting from label-free LC-MS proteomics experiments.

Method	Description of Method	Availability
Ning et al.[10]	SUM1 Sum the largest 3 peptide abundances per protein	<i>Filtered</i> approach implemented in DanteR software http://omics.pnl.gov/software <i>Unfiltered</i> approach implemented in MatLab Scripts available in Supporting Information Table S1
Cheng et al.[14]	SUM2 Average the largest 3 peptide abundances per protein	
Ning et al.[10]	SUM3 Sum over all peptides then divide by the protein's length	
Polpitiya et al.[15]	REF1 Uses reference peptide, REF2 1 REF1 : ratio all peptides to the peptide with the least amount of missing values; protein abundance is the median of the scaled peptide abundances 2 REF2 : median center and scale each peptides by standard deviation; protein abundance is the median peptide abundance	
Karpievitch et al.[16]	LM1 Fixed effects linear model, which requires model based filtering of peptides	DanteR software http://omics.pnl.gov/software
Clough et al.[17-18]	LM2 Fixed and mixed effects linear models ^{a)}	MSstats http://www.stat.purdue.edu/~tclough/MSstats/MSstats.html
Bukhman et al.[26]	Fixed effects linear model	R script http://ieco1.mie.utoronto.ca/~thodorus/Bukhman
Daly et al.[27]	Mixed effects linear model	MatLab script http://omics.pnl.gov/software
Du et al.[12]	Temporal pattern of protein is inferred from the temporal patterns of at least 2 peptides	
Dost et al.[13]	Use of shared peptides to calculate relative protein abundance using linear programming	
		C# script http://cseweb.ucsd.edu/~bdost/downloads.htm

Table 2

The distribution of the number of peptides per protein.

	Total Number of Proteins	Number of Peptides per Protein					
		1	2	3	4	5 – 15	>15
Human	1515	558	181	99	64	390	223
Mouse	1679	859	232	128	96	295	69

Table 3

The number of proteins with estimated abundance values returned from each method.

	Number of Peptides per Protein	Number of Proteins	SUM1	SUM2	SUM3	REF1 ^{a)}	REF2 ^{a)}	LM1 ^{b)}	LM2 ^{c)}
Human	3	776	776	776	776	508/776	508/776	598	752
	2	957	-	-	-	548/957	548/957	721	897
	1	1515	-	-	-	619/1515	619/1515	1062	897
Mouse	3	588	588	588	588	545/588	545/588	582	588
	2	820	-	-	-	699/820	699/820	797	820
	1	1679	-	-	-	1028/1679	1028/1679	1407	1679

^{a)}Default filtered results by DanteR/unfiltered results by Matlab script.

^{b)}The number of proteins in inferred from the filtered peptide list.

^{c)}Due to unresolvable computing issues, proteins identified by a single peptide were filtered out of the human dataset.

Table 4

The distribution of proteins based on content (single peptide, low content (n peptides = 2, 3, 4), or high content (n peptides = 15 to 20)) and peptide response frequency (low frequency, 25%; high frequency, 80%).

	Single Peptide		Low Content		High Content	
	Low Response Frequency	High Response Frequency	Low Response Frequency	High Response Frequency	Low Response Frequency	High Response Frequency
Human	383	44	113	4	15% to 69%, 77	
Mouse	269	140	19	16	41% to 92%, 44	