A maximum-likelihood approach to absolute protein quantification in mass spectrometry*

Chao Ji,[1] Yong Fuga Li,[2] Earl P. Bellinger,[1] Sujun Li,[1] Randy J. Arnold,[3]
Predrag Radivojac,[1] and Haixu Tang[1†]

1) Department of Computer Science and Informatics, Indiana University
2) Department of Bioengineering, Stanford University
3) Department of Chemistry, Indiana University

ABSTRACT

Label-free absolute protein quantification refers to a process of estimating protein abundances in complex biological samples based on the data acquired from a liquid chromatography mass spectrometry (LC-MS) analysis. Most approaches to label-free quantification rely on measuring peak areas from an extracted-ion chromatogram. However, because of the differences in physicochemical properties associated with different peptide ions, observed peak areas in a single experiment are determined not only by peptide abundances, but also the intrinsic biases of analytical platforms. Therefore, accurate modeling of these biases provides an opportunity to developing new computational methods for precise absolute protein quantification. In this work, we developed a new algorithm for absolute quantification of proteins. The approach is based on the concept of peptide response rate, which characterizes the peptide-specific signal detection bias in an LC-MS experiment. We argue that peptide response rate is an intrinsic and reproducible property of peptide ions that can be reliably predicted using non-linear regression and features extracted from the sequence of the parent protein. Protein abundances are estimated using a maximum likelihood model in which the observed peak areas of peptide ions are adjusted using predicted peptide response rates. We evaluate our approach on a large LC-MS dataset as well as simulated data and provide evidence that the accuracy of absolute protein quantification is improved when peptide-specific response rates are taken into account.

*The authors wish it to be known that, in their opinion, Chao Ji and Yong Fuga Li should be regarded as joint first authors. Correspondence should be addressed to Haixu Tang, hatang@indiana.edu and Predrag Radivojac, predrag@indiana.edu

BCB'15, September 09-12, 2015, Atlanta, GA, USA.

ACM 978-1-4503-3853-0/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2808719.2808750.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: General; J.3 [**Life and Medical Sciences**]: Biology and genetics

Keywords

Mass spectrometry, proteomics, protein quantification, estimation, artificial neural networks

1. INTRODUCTION

The advancement of techniques in quantitative proteomics has enabled broad applications of liquid chromatography tandem mass spectrometry (LC-MS/MS) to biomedical research ranging from functional profiling of cellular proteomes [3] to disease diagnosis and biomarker discovery [10, 27]. Quantitative proteomics is built upon a routine shotgun proteomics experiment, in which complex proteome samples are subject to proteolytic digestion followed by an LC-MS/MS analysis [1]. In addition to a list of identified proteins from the shotgun proteomics experiment, this approach provides information about the abundances of these proteins and thus can be used as a tool to monitor the changes of protein expression under different conditions [31, 20]; e.g., before and after viral infection [11] or among samples from healthy and diseased patients [9].

Two classes of methods have been developed for quantitative proteomics. On the one hand, isotope labeling techniques such as Isotope Coded Affinity Tags (ICAT) [15] and Stable Isotope Labeling by Amino acids in cell Culture (SILAC) [24] can be used to estimate the relative abundances of proteins in two samples by a single LC-MS analysis in which protein quantities from different samples can be distinguished based on specific isotopically labeled amino acids. On the other hand, label-free approaches are available for the direct comparison of protein abundances across multiple LC-MS analyses of different samples using peptide peak areas (precursor intensity) [22] or spectral counts [35, 5] attributable to the proteins of interest. In addition, there is a recently proposed protein quantification method [13] that measures peptide abundance based on the ion signals of multiple fragments of a given peptide in the fragmentation spectrum. Generally speaking, labeling techniques yield more accurate estimates of protein abundance ratios in two samples, but require extra steps in sample preparation.

Both labeling and label-free approaches primarily address relative protein quantification; i.e., comparing the abundance of the same protein in different samples. The determi-

nation of the absolute abundances of different proteins in the same sample; i.e., *absolute protein quantification* is useful for other important biological applications such as the mapping of protein expression patterns in a whole proteome [3]. However, in contrast to the same-protein-different-sample scenario, the quantitative measures such as spectral counts or peptide peak areas of different proteins in the same sample are not directly comparable. For example, two proteins with the same abundance may have distinct spectral counts because the peptides from one protein are more easily identified by LC-MS instruments than those from the other protein. A concept of *peptide detectability* was proposed to model the identification bias of peptides in a standard proteomics experiment [33]. As a result, the absolute protein abundances can be estimated from each protein's spectral counts corrected by the detectabilities of peptides in the protein [19, 34].

Similar to spectral counts, the detected peak areas belonging to different peptides are not directly comparable due to the detection bias inherent in the peptide sequence and need to be properly calibrated when used for absolute protein quantification. In this work we seek to quantitatively model the peptide signal detection bias in LC-MS experiment, which is referred to as *peptide response rate*. The response rate of a peptide ion is defined as the ratio between the observed total signal intensity of the peptide at a specific charge state (e.g. +1, +2, +3) and the actual abundance of the peptide in the sample. It is analogous to peptide detectability [33, 17], which models the identification biases of peptides in an LC-MS/MS experiment. A linear relationship between peptide response rate and peptide quantity has been previously demonstrated [4].

Peptide response rate is determined by three major factors: 1) the efficiency of trypsin digestion at its target sites, 2) the ionization efficiency of the peptide, and 3) the detection bias of a mass spectrometer with respect to the peptide ion. Note that we define the peptide response rate with respect to a specific peptide ion, i.e. the ions of the same peptide carrying different charges will receive different response rates. This is because the ionization efficiency and the detection bias differ for the same peptide at different charges [29]. We developed a non-linear regression model for the prediction of peptide response rate from peptide sequence and charge state. Finally, we present ALPINE, a novel Absolute and Label-free Protein quantification based on ion INtEnsity and predicted peptide response rate, and show that it performs favorably over existing quantification algorithms.

2. MATERIALS AND METHODS

2.1 Modeling peptide ion signals in LC-MS

Assuming the observed ion intensity of a peptide follows a linear relationship with the actual abundance of the peptide in the sample, we have

$$A_{ij} = r_{ij} \cdot q_i \cdot \epsilon_j^0, \tag{1}$$

where A_{ij} represents the observed peak area of peptide ion j from protein i, q_i represents the absolute abundance of the protein i, r_{ij} represents the peptide response rate of the peptide ion j from protein i, and ϵ_j^0 represents the error from peak area measurement of peptide j. Note that in this formulation we consider peptides with the same sequence but carrying different number of charges as different peptide ions. For simplicity, we only consider non-degenerate peptides (i.e. the peptides uniquely mapped to one protein in the sample) to model the protein abundance, so the abundance of peptide j depends only on that of its parent protein i. In reality the abundance of peptide j is lower than that of protein i because the enzyme digestion is not 100% efficient. However, as the digestion efficiency has been accounted for by the model for peptide response rate prediction, we assume that the abundance of peptide j is equal to that of its parent protein i. The peptide response rate r_{ij} is assumed to be independent of q_i as the ionization suppression effects and the detector saturation effects are ignored.

We assume ϵ_j^0 for each peptide j follows the *same* log-normal distribution $\log \mathcal{N}(0, \sigma^2)$, which leads to the quadratic relationship between mean and variance of peak area (See Figure 6) and thus we have

$$\log A_{ij} = \log(r_{ij} \cdot q_i) + \epsilon_j, \tag{2}$$

where $\epsilon_j = \log \epsilon_j^0$ follows a normal distribution $\mathcal{N}(0, \sigma^2)$. Note that the variance σ^2 is assumed to be independent of peptide j and the abundance of protein i.

The quantitative model described in Equations (1-2) can be used to estimate the abundance of the proteins in the sample, if the peptide response rates are known or can be predicted solely from the peptide sequence, its charge, and the sequence of its parent protein. Specifically, the abundance of protein i, q_i, can be estimated by maximizing the likelihood function as

$$L(q_i) = \prod_{j=1}^{n_i} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \cdot \exp\left\{ -\log^2\left(\frac{A_{ij}}{q_i \cdot r_{ij}}\right) / (2\sigma^2) \right\} \tag{3}$$

where n_i is the number of peptide ions confidently identified from protein i, and the likelihood function is the multiplication over all peptides j in the protein i and all charges. The maximum likelihood estimate of the abundance of protein i is then

$$\begin{aligned} \hat{q}_i &= \arg\max_{q_i} L(q_i) \\ &= \arg\min_{q_i} \sum_{j=1}^{n_i} \left(\log A_{ij} - \log(q_i \cdot r_{ij})\right)^2. \end{aligned}$$

This expression can be solved (see Appendix) to yield

$$\hat{q}_i = \left(\prod_{j=1}^{n_i} \frac{A_{ij}}{r_{ij}} \right)^{\frac{1}{n_i}}, \tag{4}$$

which is the geometric mean of the ratios between peak areas and peptide response rates over all identified peptide ions in the protein.

2.2 Learning peptide response rates

Equations (1) and (4) provide the relationship among the observed peptide ion intensities A_{ij}, the protein abundance q_i and the peptide response rates r_{ij}. If the peptide response rates were known, one can use Equation (4) to estimate the protein abundance based on the observed peptide ion intensities in an LC-MS experiment; on the other hand, Equation (1) can be used to construct a machine learning model to predict peptide response rates from a training dataset, in which the abundance of proteins in the sample are known.

However, the challenge is that, in practice, neither the peptide response rate nor the peptide abundance are known. To address this problem, we use an iterative learning algorithm that estimates peptide response rates and protein quantities simultaneously. Each iteration of the algorithm consists of three steps. In the first step, the per-protein average of peptide peak areas A_{ij} are used as the initial estimation of the abundance of protein i; i.e., $\hat{q}_i = \sum_{j=1}^{n_i} A_{ij}/n_i$; then the peptide response rate model is trained using $A_{ij}/\hat{q}_i$ as the target values. In the following two steps, protein abundances $\hat{q}_i$ are first estimated by using Equation (4), based on predicted peptide response rates o_{ij} from a regression model trained in the last step; and the target values of peptide response rates are subsequently updated using the estimated protein abundances. An overview of the algorithm is presented below. Note that the algorithm can be considered as having converged when the quantity estimation between consecutive iterations are sufficiently similar. In our experiments, we observed that the estimated protein quantities stay largely unchanged after the first three iterations, and hence the maximum number of iterations is set to three (see Results).

$\hat{q}_i \leftarrow \frac{\sum_{j=1}^{n_i} A_{ij}}{n_i}$;
$\hat{r}_{ij} \leftarrow \frac{A_{ij}}{\hat{q}_i}$;
while *not converge* **do**
  $ANN \leftarrow train(\log_{10} \hat{r}_{ij}, \mathbf{X})$;
  $o_{ij} \leftarrow ANN(\mathbf{x}_{ij})$;
  $o_{ij} \leftarrow 10^{o_{ij}}$;
  $\hat{q}_i \leftarrow \left(\prod_{j=1}^{n_i} \frac{A_{ij}}{o_{ij}}\right)^{1/n_i}$;
  $\hat{r}_{ij} \leftarrow \frac{A_{ij}}{\hat{q}_i}$;
end

Algorithm 1: Iterative learning of ion response rates. $\mathbf{X}$ represents the feature matrix (see Table 1) of all peptide ions in the training set, $\mathbf{x}_{ij}$ represents the feature vector for peptide j from protein i, and the o_{ij} represents the predicted peptide response rate by the neural network.

We used a two-layer Artificial Neural Network (ANN) to approximate the peptide response rate of a specific peptide ion, because it has been shown that this ANN is able to approximate an arbitrary bounded function with a sufficient number of hidden neurons [8]. We initialize the ANN with ten hidden neurons and we use sigmoidal and linear functions as the activation function of the hidden and output neurons to address the non-linear regression problem for learning peptide response rate.

The values of peptide response rates were log-transformed to be used as targets for ANN training, and the predicted values are transformed back to response rates by taking exponentiation; see Algorithm (1). We bootstrapped the whole set of proteins for training into 30 samples on which an ensemble of 30 feed-forward two-layer neural networks were trained. The predicted response rates averaged over 30 neural networks were used for updating protein abundances in Equation (4).

2.3 Features for response rate prediction

The peptide response rate is a result of multiple steps in an LC-MS experiment, including the proteolysis (e.g., trypsin digestion), peptide ionization, and the readout of the ion current signal by the mass spectrometer. Based on previous studies, we consider three types of peptide features in the response rate prediction. The first type of features are derived from the flanking amino acid residues in the parent protein around the cleavage sites, which facilitates modeling the digestion efficiency. Although the specific experimental protocol influences the digestion efficiency, systematic amino acid preference has been reported for trypsin and other proteases [28, 32]. The second type of features are the amino acid composition of the peptide, which is related to the ionization efficiency and detector response of the mass spectrometer. Ionization efficiencies and charge state distributions of peptides may be affected by multiple factors, such as the physicochemical properties of peptides, the properties of solvents and the ionization instrumentation [6]. Finally, we consider the charge state of the peptide ion as an independent feature, which affects both the ionization efficiency and the detector response. Table 1 summarizes the features used in our model. Obviously, the peptide response rate is determined not only by the features described above but also by the experimental instrument and protocol. However, because we build the predictive model of peptide response rate for each specific LC-MS experiment, where the experimental conditions and protocol are the same for all the peptides, the features used here determine the response rates among different peptides.

Description	# Features
Residues around N/C-terminal cleavage site	160
Normalized distance of N- and C-terminal cleavage sites to protein N-terminus	2
AA count of the peptide	20
Average AA location in the peptide	20
Number of residues in peptide sequence	1
Entropy of peptide sequence	1
Peptide precursor ion charge	1
Number of residues in protein sequence	1

Table 1: Features used for response rate learning. Each of the four residues around the N- or C-terminal cleavage sites (i.e. `X-K/R-X-X`) is represented by a length-20 vector of indicator variables. Note the effective number of features is 2 for `K/R`.

3. RESULTS

In this section we summarize the empirical performance evaluation of the proposed methodology. The protein quantity estimation by ALPINE was compared with three alternative methods that estimate protein quantity based on peptide precursor intensity; i.e., TopN, iBAQ and MeanInt. Briefly, TopN is a variant of Top3 [14, 26] that calculates the average intensity of its top 3 (or fewer, if there are not enough peptide hits for the given protein) best ionizing peptides; iBAQ calculates the sum of the extracted ion intensities of all identified peptides per protein and divides it by the number of theoretically observable peptides [30, 7]; and MeanInt [26] calculates the average precursor intensities of peptides within the same protein. In addition, we implemented a baseline approach (GeoMean) that calculates the geometric mean of peptide ion precursor intensity as the estimated protein quantity.

All evaluations were performed using simulated as well as real proteomics experiments. To our knowledge, there ex-

ists only one proteomics experiment with known quantities available for a subset of proteins; therefore, it was necessary to simulate a large number of experiments in order to more accurately establish the performance of each algorithm. As performance measures, we compute the Pearson correlation coefficient between the estimated and known abundance of the reference proteins, both in the linear and log scale. Similarly we compute the Euclidean distance between estimated and true quantities, again, both in linear and log space of quantities. Finally, we estimate the slope between the true and estimated quantities of the reference proteins as the measure of linearity of estimation. The algorithm with the slope closest to 1 would have the best performance according to this measure.

3.1 Evaluation on simulated data

Unlike most multicellular species, only a small number of peptides in unicellular organisms are shared by multiple proteins [18]. For example, only about 20% of the peptides in the *Schizosaccharomyces pombe* data are typically shared among different proteins, and 90% of proteins contain at least one unique peptides. Since shared peptides might contain important information for quantifying proteins in humans and other multicellular species, it is desirable to assess the accuracy of quantification algorithms on more complex data than just yeast. Further, since response rates are approximated in practice, it is important to see how these methods compare under various levels of noise. To this end, we numerically simulated a large number of LC-MS experiments under variable conditions and used them to benchmark ALPINE against other quantification approaches.

3.1.1 Generation of simulated data

Following the information on peptide identification in mass spectrometry given in the first table of [18], we assume that the number of peptides per protein and the number of proteins per peptide each follow shifted geometric distributions, and that the number of unique peptides per protein follows a geometric distribution. In order to generate realistic systems, we select the success rates of these distributions uniformly from the intervals $(0.05, 0.15)$, $(0.15, 0.5)$, and $(0.5, 1)$ respectively. We obtain protein quantities $\mathbf{q}$ by generating values from the log-uniform distribution $10^{\mathcal{U}(0,10)}$. We assume that standard peptide detectabilities $\mathbf{d}$ follow a beta distribution with equal shape parameters selected uniformly from the unit interval. The effective detectability of peptide j is then

$$d_j^{\text{eff}} = 1 - (1 - d_j)^{q_0^{-1} \sum_i \alpha_{ij} q_i},$$

where the standard quantity q_0 is the geometric mean of the true peptide quantities, and $\boldsymbol{\alpha}$ is a matrix of protein-peptide membership. We call a peptide identified if a random number in the unit interval is exceeded by that peptide's effective detectability. We assume that peptide response rates $\mathbf{r}$ are distributed log-normally with log-scale 0 and log-shape selected uniformly from $(1, 5)$. We sort response rates by standard detectabilities in order to enforce the correlation between them. Finally, we generate peptide peak areas $\mathbf{A}$ by multiplying response rates and peptide quantities with log-normal noise having shape parameter 0 and log-scale σ. The peak area for peptide j is then

$$A_j = \begin{cases} \epsilon \cdot r_j \sum_i q_i \alpha_{ij} & \text{if } d_j^{\text{eff}} > u \\ 0 & \text{otherwise} \end{cases}$$

with $\epsilon \sim log\mathcal{N}(0, \sigma)$ and $u \sim \mathcal{U}(0, 1)$. We generate between 100 and 1000 proteins according to these distributions and remove proteins with no identified peptides. The protein-peptide membership, peptide peak areas, and peptide response rates are then provided to each absolution quantification algorithm, and given this input we measure how accurately each approach is able to estimate the true protein abundances.

3.1.2 Evaluation results

We simulated LC-MS experiments on the Big Red II supercomputing cluster and obtained 20,288 proteomes for use in comparing the performance of ALPINE against iBAQ, GeoMean, TopN, and MeanInt. Overall, ALPINE outperforms all the other quantification methods in terms of several accuracy measures between the predicted and the true simulated absolute protein quantities, including linear correlation (Figure 1), logarithmic correlation (Figure 2), as well as the Euclidean distance between log-transformed predicted and log-transformed true quantities, and the displacement from 1.0 of the slope from log-transformed linear regression (data not shown). We further used the paired one-sided Wilcoxon signed-rank tests to evaluate the statistical significance of the improvement of ALPINE over the other methods. We observed that ALPINE significantly outperformed all other predictors in all of those measures ($p < 2.2 \times 10^{-16}$ for all). The only exception was logarithmic correlation against iBAQ, for which our method did not make a significant improvement. In this measure, all predictors had a very low average logarithmic correlation on simulated shared peptide proteomes of about 0.2 or less.

3.2 Evaluation on real proteomics data

Next, we used a public dataset from a large system-wide proteomic study of *S. pombe* to demonstrate the performance of learning peptide response rate and absolute protein quantification. The dataset consists of six technical replicates of 1D-LC-MS runs of unfractionated protein sample of yeast cell in quiescent states on the LTQ-Orbitrap platform [20]. Absolute abundances for 34 reference proteins were determined using spiked-in heavy reference peptides to translate the summed MS-intensities of all peptides to copies/cell for all identified proteins. The list of reference proteins and the associated peptides used in our evaluation (note that only a subset of reference proteins have at least one confidently identified peptide) can be found in reference [20]. The yeast dataset has been previously used for the performance comparison of different absolute protein quantification measures [2].

MASCOT [25] was used to search the tandem mass spectra against a target database consisting of 5143 *S. pombe* protein sequences appended with a decoy database with the same number of reversed protein sequences. The following search parameters were used: two missed cleavages were allowed; carbamidomethylation on cysteines was set as fixed modification; methionine oxidation was applied as a variable modification; mass tolerance for precursor and fragment ions were set to 10ppm and 0.6Da, respectively. We considered only peptide identifications with MASCOT ion scores

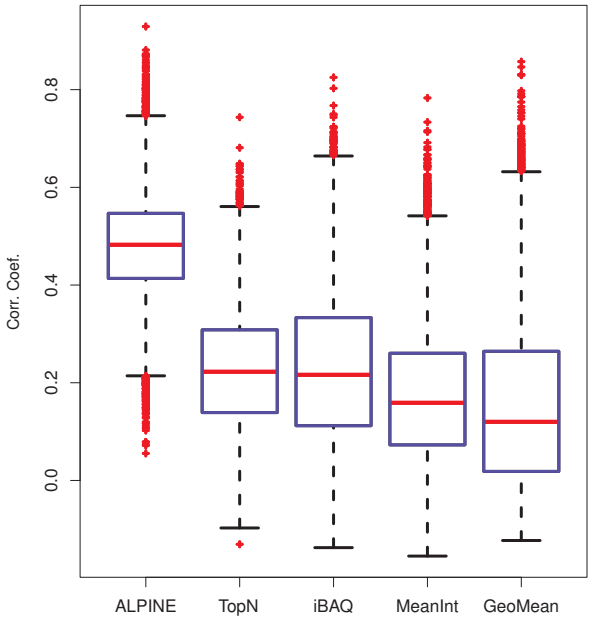


Figure 1: Boxplots of linear correlation coefficients between true protein quantities and predicted protein quantities from ALPINE, TopN, iBAQ, GeoMean, and MeanInt on more than 20,000 simulated LC-MS experiments.

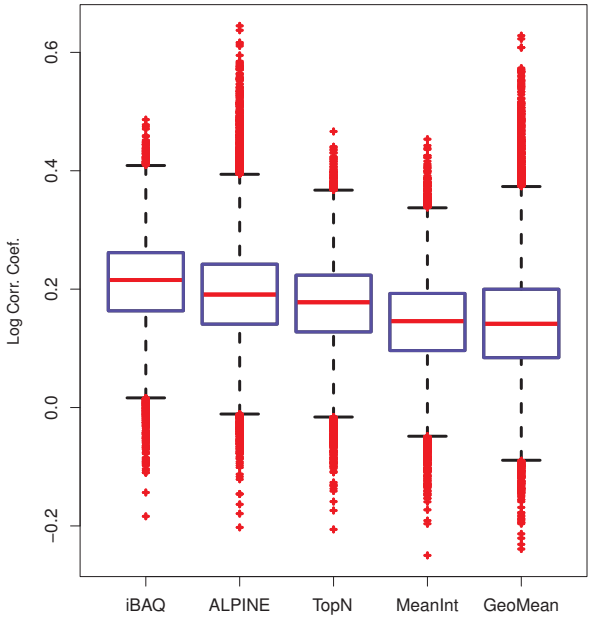


Figure 2: Boxplots of correlation coefficient between log-transformed true protein quantities and log-transformed predicted protein quantities from ALPINE, TopN, iBAQ, GeoMean, and MeanInt on more than 20,000 simulated LC-MS experiments.

		Exp. ID					
	Exp. ID	1	2	3	4	5	6
Peak area	1	1	0.936	0.885	0.886	0.900	0.928
	2		1	0.883	0.858	0.924	0.949
	3			1	0.940	0.888	0.882
	4				1	0.885	0.893
	5					1	0.968
	6						1
Response rate	1	1	0.909	0.839	0.838	0.877	0.871
	2		1	0.843	0.834	0.884	0.858
	3			1	0.887	0.803	0.812
	4				1	0.830	0.829
	5					1	0.899
	6						1

Table 2: Reproducibility of peptide peak area and estimated response rate. For each pair of replicated experiments, Pearson correlation coefficient were computed on peak areas and estimated response rates of peptides identified in both experiment i and j.

above 25.0, which corresponds to 0.2% False Discovery Rate (FDR). In addition, all peptide hits are required to be identified in at least two out of six replicate experiments, and no modified peptides were included. ProteinProphet [23] was used to assign a probability score to each protein identification. Precursor signal intensities (i.e. peak area) were calculated using MASIC [22] with default parameters. We used the list of proteins whose ProteinProphet probability scores were $\geq$0.75 (there are about 700-900 proteins identified in each of these datasets) to perform on-line training of peptide response rate in the iterative algorithm, and tested on the set of 20 reference proteins with known abundances [20].

We first investigated if the peptide response rates are sufficiently reproducible across technically replicated LC-MS experiments. To this end, we considered the set of non-degenerate peptides identified from all reference proteins for each of the six LC-MS runs in the *S. pombe* dataset [20]. Because the protein abundances should be the same across technical replicates (for unfractionated samples) and the quantity of each unique peptide is equal to the quantity of its parent protein, the observed peak areas of the same unique peptide identified in multiple runs are anticipated to be approximately the same, so assessing the reproducibility of true response rates and peptide peak area would be equivalent. This is consistent with our observation: as shown in Table 2, the correlation coefficient of the peak areas across replicated experiments are generally high ($\sim$0.9), which is comparable with other omics techniques, such as RNA-seq for transcriptomic analysis [21]. In addition to the high reproducibility of true response rates, the response rate of the same peptide ion that were estimated by the models trained separately on each replicate experiment are also reasonably similar (0.8 $\sim$ 0.9, Table 2), suggesting that response rate is likely to be an intrinsic property of peptides in LC-MS experiment.

We then proceeded to examine the degree to which response rates vary across different peptides in the same sample. Again, because these are unique peptides and are from reference proteins, the response rate of each peptide can be

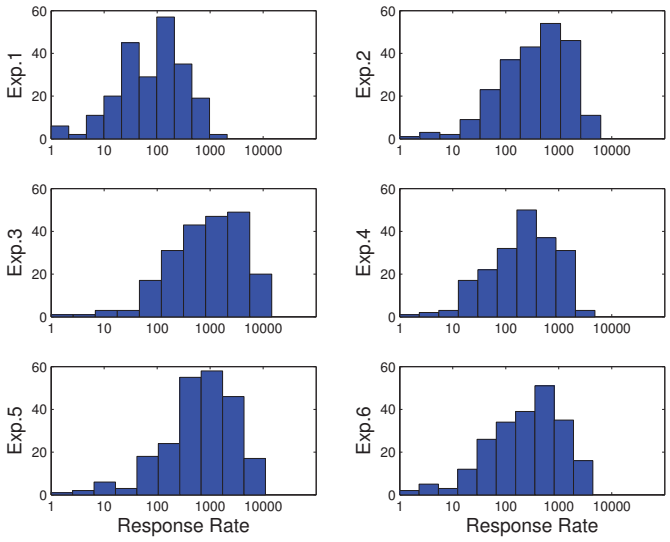


Figure 3: Distribution of peptide response rates. Peptide response rates are calculated for all peptides from the reference proteins. The values are normalized such that the minimum response rate equals 1.0.

calculated as the observed peak area divided by the known abundance of its parent protein. As indicated in Figure 3, the peptide response rates are widely distributed, spanning up to four orders of magnitude. This suggests that there is noticeable bias in signal detection for different peptides, and more importantly, because of its high variability, the response rate cannot be assumed to be constant across different peptides when estimating protein abundances based on the peak areas of identified peptides. Furthermore, as the response rate is highly reproducible for the same peptide, but highly variable for different peptides, we hypothesize that the response rate is an intrinsic property of a peptide and can be predicted from the peptide sequence by using machine learning methods. The predicted response rate can be incorporated to improve the quantification of proteins in complex proteome samples.

We ran and evaluated ALPINE independently for each of the six replicate experiments. Specifically, ALPINE was run for three iterations and the final protein quantity estimation was reported. First we examined how well the peptide response rates could be learned. Note that in each iteration of ALPINE algorithm the neural networks take the same feature matrix as input whereas the targets may vary. As shown in Figure 5A, the performance of neural network training (i.e. correlation coefficient between predicted and target response rates on the test set) improved noticeably in each successive iteration, indicating that the neural networks are supplied with more informative signals as target values. In addition, the improvement in peptide response rate learning translates to better similarity measure between estimated and true protein quantity (Figure 5B). This indicates that even the algorithm was initially supplied with supposedly inaccurate peptide response rate estimation as targets, the algorithm re-adjusts the target values of the response rate for neural network training by incorporating the updated protein abundance estimation, which eventually leads to improved estimation of both peptide response rate and protein quantity.

Table 3 summarizes the similarities between the estimated protein quantities by each method with respect to the true protein abundances based on different measures. ALPINE consistently outperforms the other methods for the correlation coefficient in the linear scale, and also gives generally best performance for the rank-based non-parametric correlation coeffients (Spearman's and Kendall's tau). Note that GeoMean computes protein abundances as the geometric mean of the observed precursor intensities of identified peptides, implicitly assuming all identified peptides have the same response rates. By taking into consideration the predicted peptide response rates, ALPINE gained noticeable improvement in the correlation coefficients over GeoMean.

An important and desirable property of the quantification method is the linearity between estimated and true protein quantity ([2]). Specifically, given the true quantity Q and estimated quantity $\hat{Q}$, Q can be expressed as $Q = K \cdot \hat{Q}$, where K is a constant scalar. By taking log-transformation of the both sides, it follows that $\log Q = \log \hat{Q} + \log K$. Clearly, the linearity is determined by the degree to which the slope of the regression line between $\log Q$ and $\log \hat{Q}$ is displaced from 1.0. As illustrated in Figure 4, the slope of ALPINE is closest to 1.0 in most cases, and both Q and $\hat{Q}$ span approximately three orders of magnitude, which indicates one can multiply the estimated quantity of a protein reported by ALPINE by a scalar, which can be easily estimated from a few well characterized proteins in the same sample, to approximate the true abundance of the protein.

It is interesting to observe that iBAQ has generally better correlation coefficient computed in log-scale while the slope of iBAQ is far from 1.0 (around 0.5). This implies a non-linear relationship between Q and $\hat{Q}_{ibaq}$: $Q = (\hat{Q})^{0.5}$. As shown in Figure 4, iBAQ consistently underestimates low-abundant proteins compared with high-abundant ones, which eventually caused the wider dynamic range of $\hat{Q}_{ibaq}$ and the non-linear relationship between Q and $\hat{Q}_{ibaq}$.

4. CONCLUSIONS

In this work, we formalize the definition of *peptide response rate*, which describes the precursor ion intensity detected in an LC-MS experiment for a peptide ion at a given abundance level. We show that in a typical LC-MS experiment, peptide response rates can differ significantly among different peptides, while still being highly reproducible across technical replicates. We show that peptide response rate is an intrinsic property of peptide ions (and their parent proteins) and can be reliably predicted using non-linear regression. Based on a simple quantitative model that relates peptide response rates and protein abundances, we develop an iterative algorithm for estimating them simultaneously. We employ a large-scale LC-MS dataset to demonstrate the effectiveness of this approach and show that it evaluates favorably against alternative protein quantification methods on both empirical and simulated data alike. To our knowledge, this approach represents the first attempt that explicitly incorporates peptide response rate when using peak area for proteomic quantification, and our results show that accurately predicted response rates can be used for calibrating the observed peak areas as well as significantly improving the accuracy of protein quantification.

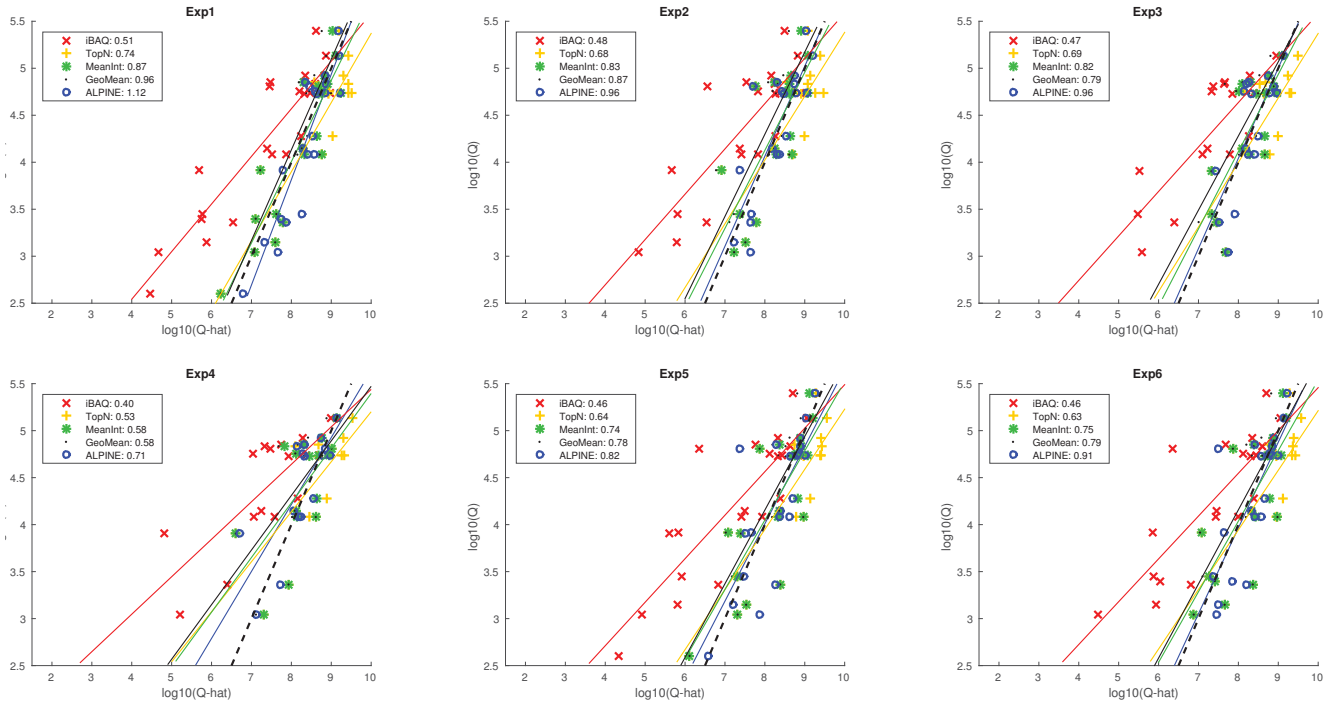
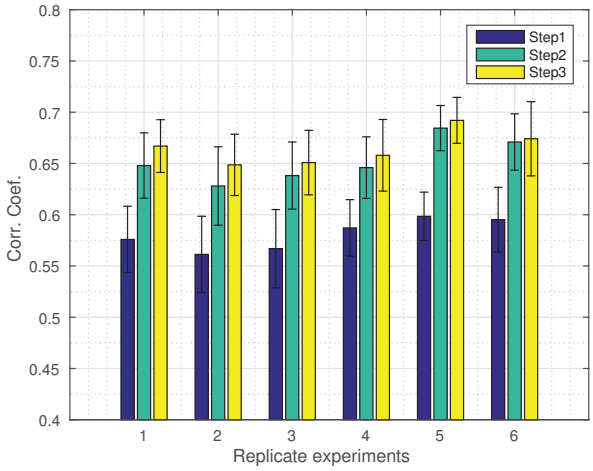


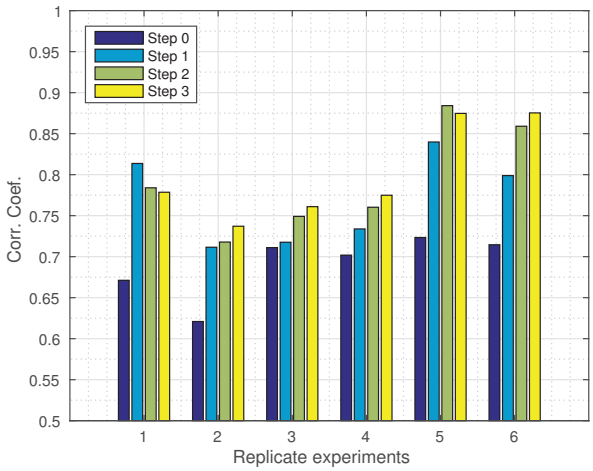
Figure 4: Scatter plot of the log-transformed true protein abundance versus the log-transformed estimated protein abundance for TopN, iBAQ, GeoMean, MeanInt, and ALPINE for the *S. pombe* dataset. For each method, the least-square linear regression was performed and the slope of the regression line is shown. The dashed line represents the diagonal.

	Pearson's corr. coef. (linear-scale)						Pearson's corr. coef. (log-scale)					
Method	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5	Exp.6	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5	Exp.6
ALPINE	**0.779**	**0.737**	**0.761**	**0.775**	**0.875**	**0.875**	**0.911**	0.819	0.768	**0.776**	0.785	0.781
iBAQ	0.571	0.504	0.698	0.708	0.648	0.626	0.910	**0.831**	**0.828**	0.773	**0.857**	**0.830**
TopN	0.557	0.483	0.734	0.757	0.621	0.609	0.903	0.778	0.792	0.704	0.823	0.788
MeanInt	0.671	0.621	0.711	0.702	0.723	0.715	0.903	0.769	0.756	0.647	0.808	0.770
GeoMean	0.604	0.572	0.564	0.527	0.553	0.539	0.879	0.757	0.698	0.612	0.765	0.711
APEX	0.369	0.416	0.756	0.772	0.468	0.434	0.531	0.482	0.449	0.499	0.626	0.534
	Spearman's rank corr. coef.						Kendall's tau					
ALPINE	**0.845**	**0.768**	0.683	**0.735**	0.767	**0.785**	**0.689**	**0.594**	0.519	**0.571**	**0.660**	**0.635**
iBAQ	0.821	0.731	**0.710**	0.658	**0.807**	0.780	0.651	0.559	**0.578**	0.521	0.651	0.603
TopN	0.810	0.709	0.683	0.679	0.779	0.724	0.632	0.524	0.489	0.521	0.593	0.519
MeanInt	0.834	0.652	0.683	0.597	0.738	0.740	0.670	0.477	0.489	0.471	0.565	0.561
GeoMean	0.799	0.637	0.626	0.535	0.680	0.633	0.641	0.441	0.459	0.403	0.507	0.466
APEX	0.570	0.581	0.607	0.607	0.674	0.606	0.418	0.463	0.439	0.439	0.496	0.473

Table 3: Similarity measures (Pearson's correlation coefficient (linear- and log-scale), Spearman's rank correlation coefficient, Kendall's tau rank correlation coefficient) are computed between true and estimated quantities of the reference proteins for ALPINE and alternative methods. Note that log-transformation would not change the relative order of values, and the rank correlation would be identical in linear- and log-scale.



(a)



(b)

Figure 5: Performance of iterative learning of peptide response rate and protein quantity estimation. At each iteration of ALPINE algorithm, each of the 30 bootstrapped samples was divided into a training (85%) and test(15%) set. A neural network was trained on the training set and response rates were predicted for the peptides in the test set. Correlation coefficient between predicted and target response rates were computed for the test set of each bootstrapped sample and for each iterative step (a). The error bars indicate the standard deviation among 30 bootstrapped samples. In addition, correlation coefficient was computed between estimated and true abundance of the reference proteins for each iterative step. Step 0 corresponds to the initial protein quantity estimation (i.e. average of peptide precursor intensity), and step 1 corresponds to the protein quantity estimation using Equation (4) after the first round of response rate learning.

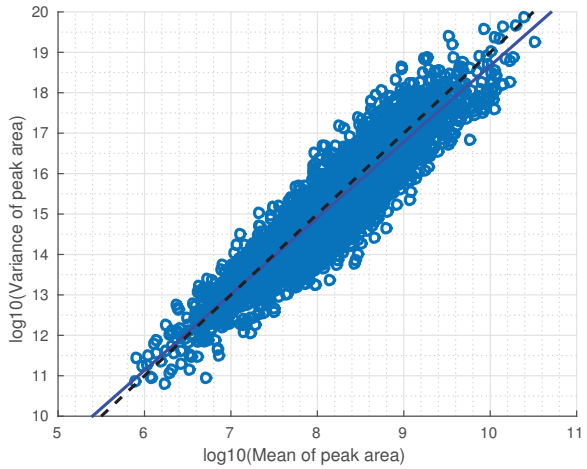


Figure 6: The quadratic relationship between mean and variance of peak area, indicating the error of peak intensity measurement follows a log-normal distribution. The mean and variance of the peak areas of the same peptide ion identified in all six replicate experiments are plotted. The correlation coefficient of log transformed means and variances is 0.9202. The slope of the linear regression line from the replicate experiments (blue line) is equal to 1.88 (close to 2.0), indicating the quadratic relationship (black dashed line).

Although peptide response rates are found to be highly reproducible across technical replicates, we observed the reproducibility of peptide response rates was relatively low across different platforms. For example, peptide ion intensities observed from replicate experiments on the same platforms (LTQ-FT and LTQ-Orbitrap) are found to be very reproducible (0.896 and 0.868, respectively), while the cross-platform reproducibility was observed to be 0.554. Therefore, we suggest that the predictive model for peptide response rates be trained simultaneously with protein quantification on the same dataset or different datasets from the same platform.

It should be noted that because shared peptides are rare in unicellular organisms, here we only considered non-degenerate peptides in the learning step. While we have seen that non-degenerate peptides are sufficient for the purpose of learning peptide response rates, shared peptides do contain important information for quantifying proteins in higher organisms such as human. A recent work has proposed a statistical model that explicitly incorporates information of shared peptides to improve the accuracy of protein quantification [12], and this information will be incorporated in future versions of the ALPINE model.

In addition to the properties of a peptide ion (such as its amino acid composition and charge state), post-translational modifications (PTMs) may alter the ionization efficiency of peptides in an LC-MS experiment. For instance, phosphorylation will make a peptide more likely to carry negative charges compared with the unmodified peptide, and thus it is likely that the modified and unmodified peptide have different response rates. Although it has been shown that

tyrosine phosphorylation may slightly (less than two-fold) alter the response rates of the peptides [16], a large-scale analysis is needed to elucidate the impact of PTMs on the response rate.

Finally, in this work, we assume peptide response rate is independent of protein abundances and ignore the effects of ionization suppression and detection saturation, while in LC-MS experiments the competition for ionization among co-eluting peptides (as well as other molecules) may lead to detection saturation. It should be noted that the training set has included all peptides in the sample (including co-eluting peptides), and therefore those properties that may affect the outcome of ionization competition were likely incorporated in the model for peptide response rate prediction, and the model is then applied to the peptides from the same sample, constituting the same background of competition. Nonetheless, the model may still benefit from the explicit modeling of the ionization competition when it is applied for cross-sample prediction (i.e. model trained on one dataset as the predictor of another dataset). This will be considered in the future work.

5. ACKNOWLEDGEMENTS

This work is supported by NIGMS/NIH (Grant no: R01 GM103725-04). This research was sustained in part by fellowship support from the National Physical Science Consortium and the National Institute of Standards and Technology. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at IU is also supported in part by Lilly Endowment, Inc.

6. REFERENCES

[1] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, **422**(6928):198–207, 2003.

[2] E. Ahrne, L. Molzahn, T. Glatter, and A. Schmidt. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics*, **13**(17):2567–2578, 2013.

[3] N. L. Anderson, N. G. Anderson, T. W. Pearson, C. H. Borchers, A. G. Paulovich, S. D. Patterson, M. Gillette, R. Aebersold, and S. A. Carr. A human proteome detection and quantitation project. *Mol. Cell. Proteomics*, **8**(5):883–886, 2009.

[4] S. Bark, W. Lu, and V. Hook. Linear and accurate quantitation of proenkephalin-derived peptides by isotopic labeling with internal standards and mass spectrometry. *Anal. Biochem.*, **389**(1):18–26, 2009.

[5] H. Choi, D. Fermin, and A. I. Nesvizhskii. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics*, **7**(12):2373–2385, 2008.

[6] R. B. Cole. Some tenets pertaining to electrospray ionization mass spectrometry. *J. Mass Spectrom.*, **35**(7):763–772, 2000.

[7] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**:1367–1372, 2008.

[8] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**(4):303–314, 1989.

[9] L. DeSouza, G. Diehl, M. J. Rodrigues, J. Guo, A. D. Romaschin, T. J. Colgan, and K. W. Siu. Search for cancer markers from endometrial tissues using differentially labeled tags itraq and cicat with multidimensional liquid chromatography and tandem mass spectrometry. *J. Proteome Res.*, **4**(2):377–386, 2005.

[10] E. P. Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Mol. Cell. Proteomics*, **3**(4):367–378, 2004.

[11] D. L. Diamond, J. M. Jacobs, B. Paeper, S. C. Proll, M. A. Gritsenko, R. L. J. Carithers, A. M. Larson, M. M. Yeh, D. G. Camp, R. D. Smith, and M. G. Katze. Proteomic profiling of human liver biopsies: Hepatitis c virusâĂŞinduced fibrosis and mitochondrial dysfunction. *Hepatology*, **46**(3):649–657, 2007.

[12] S. Gerster, T. Kwon, C. Ludwig, M. Matondo, C. Vogel, E. M. Marcotte, R. Aebersold, and P. Buhlmann. Statistical approach to protein quantification. *Mol. Cell. Proteomics*, **13**(2):666–677, 2014.

[13] L. C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, **11**(6):O111.016717, 2012.

[14] J. Grossmann, B. Roschitzki, C. Panse, C. Fortes, S. Barkow-Oesterreicher, D. Rutishauser, and R. Schlapbach. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *J. Proteomics*, **73**(9):1740–1746, 2010.

[15] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, **17**(10):994–999, 1999.

[16] L. L. Jin, J. Tong, A. Prakash, S. M. Peterman, J. R. St-Germain, P. Taylor, S. Trudel, and M. F. Moran. Measurement of protein phosphorylation stoichiometry by selected reaction monitoring mass spectrometry. *J. Proteome Res.*, **9**(5):2752–2761, 2010.

[17] Y. F. Li, R. J. Arnold, H. Tang, and P. Radivojac. The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *J. Proteome Res.*, **9**(12):6288–6297, 2010.

[18] Y. F. Li and P. Radivojac. Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics*, **3**(Suppl 16):S4, 2012.

[19] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, **25**(1):117–124, 2007.

[20] S. Marguerat, A. Schmidt, S. Codlin, W. Chen, R. Aebersold, and J. Bahler. Quantitative analysis of fission yeast transcriptomes and proteomes in

proliferating and quiescent cells. *Cell*, **3**(151):671–683, 2012.

[21] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517, 2008.

[22] M. E. Monroe, J. L. Shaw, D. S. Daly, J. N. Adkins, and R. D. Smith. MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Comput. Biol. Chem.*, **32**(3):215–217, 2008.

[23] A. I. Nesvizhskii, A. Keller, E. Kolker, and A. R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **17**(75):4646–4658, 2003.

[24] S. E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics*, **1**(5):376–386, 2002.

[25] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**(18):3551–3567, 1999.

[26] D. Qi, P. Brownridge, D. Xia, and K. Mackay. A software toolkit and interface for performing stable isotope labeling and top3 quantification using Progenesis LC-MS. *OMICS*, **16**:489–496, 2012.

[27] N. Rifai, M. A. Gillette, and S. A. Carr. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.*, **24**(8):971–983, 2006.

[28] J. Rodriguez, N. Gupta, R. Smith, and P. P. A. Does trypsin cut before proline? *J. Proteome Res.*, **7**(1):300–305, 2007.

[29] P. A. Rudnick, X. Wang, X. Yan, N. Sedransk, and S. E. Stein. Improved normalization of systematic biases affecting ion current measurements in label-free proteomics data. *Mol. Cell. Proteomics*, **13**(5):1341–1351, 2014.

[30] B. Schwanhausser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, **473**(7347):337–342, 2011.

[31] M. Selbach. Widespread changes in protein synthesis induced by micrornas. *Nature*, **455**(7209):58–63, 2008.

[32] J. A. Siepen, E. J. Keevil, D. Knight, and H. S. J. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J. Proteome Res.*, **6**(1):399–408, 2006.

[33] H. Tang, R. J. Arnold, P. Alves, Z. Xun, D. E. Clemmer, M. V. Novotny, J. P. Reilly, and P. Radivojac. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics*, **22**(14):481–488, 2006.

[34] C. Vogel and E. M. Marcotte. Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nat. Protoc.*, **3**(9):1444–1451, 2008.

[35] B. Zhang, N. C. VerBerkmoes, M. A. Langston, E. Uberbacher, R. L. Hettich, and N. F. Samatova. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.*, **5**(11):2909–2918, 2006.

APPENDIX

A. ALPINE DERIVATION

The likelihood of the i^{th} quantity q_i is

$$\mathcal{L}(q_i) = \prod_{j=1}^{n_i} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{ -\log^2\left(\frac{A_{ij}}{q_i \cdot r_{ij}}\right) / \left(2\sigma^2\right)\right\}.$$

We wish to find the quantity that maximizes this likelihood, which we may equivalently obtain by minimizing the negative log likelihood:

$$\hat{q}_i = \arg\min_{q_i} \left[-\ell\ell(q_i)\right]$$

where

$$\ell\ell(q_i) = \sum_{j=1}^{n_i} -\log^2\left(\frac{A_{ij}}{q_i \cdot r_{ij}}\right) / \left(2\sigma^2\right).$$

We can find critical points by differentiating the negative log likelihood with respect to the i^{th} quantity and setting it equal to zero as follows:

$$\begin{aligned}
\frac{\partial\left[-\ell\ell(q_i)\right]}{\partial q_i} &= \frac{\partial}{\partial q_i}\sum_{j}^{n_i} \log^2\left(\frac{A_{ij}}{q_i \cdot r_{ij}}\right) / \left(2\sigma^2\right) \\
&= -(q_i\sigma^2)^{-1}\sum_{j}^{n_i} \log\left(\frac{A_{ij}}{q_i \cdot r_{ij}}\right) \\
&= 0.
\end{aligned}$$

We may now multiply both sides of the by the i^{th} quantity and the variance, and then use the laws of logarithms to find

$$\sum_{j}^{n_i} \log\left(\frac{A_{ij}}{q_i \cdot r_{ij}}\right) = \log\left[\prod_{j}^{n_i}\left(\frac{A_{ij}}{q_i \cdot r_{ij}}\right)\right] = \log\left[\frac{\prod_{j}^{n_i} \frac{A_{ij}}{r_{ij}}}{q_i^{n_i}}\right] = 0$$

which we can then exponentiate to get

$$\frac{\prod_{j}^{n_i} \frac{A_{ij}}{r_{ij}}}{q_i^{n_i}} = 1$$

from which we arrive at the ALPINE solution:

$$\hat{q}_i = \prod_{j}^{n_i}\left(\frac{A_{ij}}{r_{ij}}\right)^{\frac{1}{n_i}}. \qquad \blacksquare$$