Statistics and Its Interface Volume 5 (2012) 99–107

# Protein quantitation using iTRAQ: Review on the sources of variations and analysis of nonrandom missingness

Ruiyan Luo* and Hongyu Zhao

As a technique that allows simultaneous quantitation of proteins in multiple samples, iTRAQ (isobaric Tags for Relative and Absolute Quantitation) has gained increased interest and applications in proteomics research. Despite its success, iTRAQ data present a number of statistical challenges even after the proteins and peptides are identified and the peak areas of the reported ions are estimated for peptide intensities. In this article, we review recent studies on the analysis of iTRAQ data, the computation problems involved and the nonrandom missingness in the iTRAQ data.

AMS 2000 subject classifications: 60K35.
Keywords and phrases: iTRAQ, ANOVA, Nonrandom missing, Bayesian hierarchical model, Mass spectrometry.

## 1. INTRODUCTION

One main objective of proteomics research is to detect and quantify all proteins present in a biological sample. Proteins that exhibit an increase or decrease in abundance between distinct proteomes (e.g., disease and nondisease or control and treatments) are potential biomarkers. Many different techniques have been developed to simultaneously compare protein levels across multiple samples. One method that has gained increased attention is iTRAQ [10, 14, 23, 30], a shotgun technique that uses Isobaric Tags for Relative and Absolute Quantitation. Compared to other methods such as 2DE [20], ICAT (isotope-coded affinity tags) [4], and DIGE (differential gel electrophoresis) [5, 21], iTRAQ offers improved quantitative reproducibility, higher sensitivity [32], and has broad applications in proteomics research [1, 2, 8, 13, 29, 33].

Using four or eight isobaric tags, iTRAQ can simultaneously analyze up to eight biological samples [3, 23]. The four reagents used in the 4-plex version of iTRAQ are named 114, 115, 116 and 117. The eight reagents include these four and four additional reagents named 113, 118, 119 and 121. Each reagent is composed of a peptide reactive group and an isobaric tag that consists of a reporter group and a balance group. The peptide-reactive group specifically reacts with primary amine groups of peptides. The reporter group gives strong signature ions in tandem mass spectrometry (MS/MS) and is used to determine the relative abundance of a peptide. The balance group keeps the overall mass of the isobaric tag constant. With this property, identical peptides labelled with different isobaric tags will not be distinguishable in mass spectrometry.

In the experimental workflow for iTRAQ, unlabelled protein samples are first trypsin-digested and labelled with different isobaric tags independently. These labelled peptides from different samples are then mixed together and separated by liquid chromatography. Identical peptides from different samples labelled with different isotopes are chromatographically indistinguishable and appear as a single precursor. The isolated peptides are finally run through MS/MS for further fragmentation and generate a collection of mass spectra. The property of isobaric tags allows otherwise identical peptides from different samples to be detected as a single peak by mass spectrometry and to produce a single set of sequencing ions in MS/MS. The ion signals produced from the reporter regions together with the normal fragment ions provide information on peptide identification and quantitation for different samples. Using softwares such as MASCOT (Matrix Science Inc., Boston, MA, USA), a protein database search can be performed on the fragmentation data to identify the labelled peptides and hence the corresponding proteins. The relative abundence of low molecular mass reporter ions generated from the isobaric tags can then be used to quantify the relative abundence of peptides and proteins across the samples studied.

The observed peptide intensities are approximated by the peak areas of the ions originating from the isobaric tags used to label different samples. Several factors can affect the observed peptide intensities, such as the expression level of the protein that generates the peptide, some peptide specific features relating to different efficiency in ionization and fragmentation, different amounts of samples loaded into different channels, differences in sensitivity to instrument detection, sample preparation and experimental variations. Hill et al. [7] described in detail these biological and experimental factors and incorporated them into an ANOVA model to evaluate differential protein expression from iTRAQ data that are generated by a single experiment or multiple experiments.

*Corresponding author.

One commonly encountered issue in iTRAQ data analysis is data missingness. Due to the nature of the technology, the overlap in identified proteins and peptides between replicate experiments is less than ideal, and many peptides are only observed for some samples in some spectra, leading to a large amount of missing data. For example, in a controlled study with 9 technical replicates described in [16], only 35.4% of the total 1,751 proteins were found in every experiment. Wang et al. [31] found that the total number of features identified in an experiment decreased over time by 49–73%. In a study of the effect of Caveolin-1 in three pairs of wild-type mice and knock-out Cav-1 mice, only about 1/3 of the proteins were identified in all three experiments, and only 1/4 peptides originating from these proteins were identified in all experiments [17]. These studies found that missingness does not occur at random. Instead, the probability that a protein/peptide is missing is related to its abundance. Less abundant peptides are harder to detect due to the data-dependent acquisition of the analysis process, hence more likely to be missing. This presents a nonignorable missing data problem. Ignoring the nonrandom missing pattern in statistical analysis may lead to significant bias in statistical inference and scientific conclusions.

To identify differentially expressed proteins across samples, one common approach is to calculate the ratio of the observed peptide intensities between two samples and to compare the calculated ratios against pre-specified upper and lower bounds. However, the criterion for threshold selection is subjective. For example, Seshi [27] considered iTRAQ ratios $>5/4$ or $<4/5$ as significant, whereas Salim et al. [26] used thresholds 1.20 and 0.83. These thresholds fail to consider the variability in data and are not statistically based. In this paper we review emerging new statistical approaches to quantitative proteomics that address the variations and missingness in iTRAQ data.

## 2. ANOVA ANALYSIS

Hill et al. [7] carefully studied the sources of variations in iTRAQ and applied ANOVA models to incorporate these variations in inferring differentially expressed proteins. They performed the normalization and quantification of differential protein expression with a single model fit to the observed peptide intensities obtained from the reporter ion peak areas from all observed tandem mass spectra. Their model relates differences in treatment to relative differences in protein expression, relates protein expression to peptide expression, and relates peptide expression to observed reporter ion peak areas. These relationships are captured using simple multiplicative expressions in the original scale, which is equivalent to a simple additive model in the logarithmic scale. The computional issues involved in the ANOVA model fitting for a medium or large size of global proteomics data sets were studied by Oberg et al. [19].

### 2.1 Model

Suppose that there are $K$ iTRAQ experiments and the proteome contains $I$ proteins. Let $j(i)$ indicate the $j$-th peptide derived from the $i$-th protein, $s$ index the biological sample obtained under a particular treated or control condition, and $l$ index the isobaric tag labeling the sample.

We use $y_{ijksln}$ to denote the log transformed value of the observed intensity for the $j$-th peptide derived from the $i$-th protein in the $s$-th biological sample, the $k$-th experiment, the $l$-th labeling reagent and the $n$-th MS/MS spectrum. Then the observed value is decomposed as

$$(1)\quad y_{i,j(i),k,s,l,n} = (\mu + b_k + v_{k,l}) + (p_i + f_{j(i)}) + (r_s + r_{i,s} + g_{j(i),s}) + h_{i,j(i),k,s,l,n},$$

where $\mu$ represents the grand mean, $b_k$ describes the effect due to a given iTRAQ experiment, $v_{k,l}$ describes the experimental effects of loading, mixing, and other sample handling effects, $p_i$ represents the protein effect, $f_{j(i)}$ corresponds to the peptide effect, $r_s$ denotes the sample effect, $r_{i,s}$ denotes the proteins differentially expressed between samples, and $g_{j(i),s}$ denotes the peptides differentially expressed between samples obtained under different conditions. The term $h_{i,j(i),k,s,l,n}$ represents the residual error for each observation that is not captured by the model. To ensure identifiability, one level of each predictor is referred to as the variable's "reference level". So the parameters in (1) (except $\mu$) represent the relative effect of the corresponding predictor, and the value of each parameter corresponding to the "reference level" is zero. For example, if the sample from the control condition is referred to as the "reference sample", then $r_s$ is the relative amount of total protein comparing the $s$-th sample to the reference sample, and $r_{i,s}$ denotes the relative amount of protein $i$ comparing the $s$-th sample to the reference sample (the primary parameter of interest). When $s$ indicates the reference sample, $r_{i,s} = r_s = 0$.

The terms in (1) are arranged into three groups describing the experimental effects, the protein and peptide effects, and the differences between samples (or the treatment effects). The first group $(\mu + b_k + v_{k,l})$ describing the experimental effects includes variations in the amount of samples loaded into iTRAQ channels, the labeling efficiency, the mixing of labelled samples, and so on. These effects would not exist in an ideal world of perfectly reproducible instruments, experiment procedures, and subjects. The second group $(p_i + f_{j(i)})$ describes the differential effects of protein $i$ and the $j$-th peptide derived from this protein. It has been observed that if a single purified protein is trypsinized and the results subjected to mass spectrometry, the reported peptide abundances may vary by the magnitude of two-to-three orders. The term $f_{j(i)}$ captures the variation of the expected amount of the $j$-th peptide to the expected amount of the $i$-th protein for subjects in the reference condition. The third group of effects $(r_s + r_{is} + g_{j(i),s})$ capture the interest of the research, from which we infer the differentially expressed

proteins and/or peptides between samples obtained under different treatment conditions. The term $g_{j(i),s}$ captures the effect of conditions at the peptide level. There are certainly biological conditions where a change to the levels of one or more peptides, but not the protein as a whole, will occur; for example a post-translational modification that involved a peptide substitution.

## 2.2 Model fitting

Parameters in models like (1) generally can be estimated using the standard method of least squares. However, the large size of global proteomics data sets may result in hundreds and thousands of parameters involved in the model (1), making it hard to estimate all of the parameters simultaneously using current software and computing facilities. Oberg et al. [19] described the following methods to partition the modeling process into a normalization portion (bias removal) and a differential expression portion.

### 2.2.1 Subsetting

This method partitions the global proteomics dataset into subsets by proteins and estimates the parameters separately for each identified protein. This will lead to biased estimates of parameters in the ANOVA model because model (1) involves the "experimental effects" ($b_k$, $v_{k,l}$) which would affect all proteins in an experiment. For example, a larger (or smaller) total amount of protein mixture loaded in an iTRAQ experiment will lead to all of the proteins in that experiment to have higher (or lower) intensities. Fitting model (1) separately for each protein will lead to different estimates of the global experimental effects for different proteins, which is unreasonable. So estimating the experimental effects for each protein individually rather than globally leads to incorrect normalization.

### 2.2.2 Stagewise regression

Denote the three groups of terms in the model (1) as groups I, II, and III, where group I corresponds to the experimental effects, group II corresponds to the protein and/or peptide effects, and group III corresponds to the differential expression portions of the model. The stagewise regression strategy fits the model to the entire data set in a stagewise fashion, that is, first group I, followed by group II, and then group III. Then it would be simple for each of the individual fits.

However, for the stagewise approach to give correct answers, it is necessary that the parameter estimates from the multiple stages are uncorrelated. In other words, to get unbiased estimates of parameters in the ANOVA model (1), it is necessary that the portions of the linear model design matrix corresponding to the multiple stages are orthogonal, which is not satisfied by MS data. For iTRAQ data, missingness is very common. Each global proteomics experiment detects different sets of proteins, resulting in an unbalanced data set for which the experimental and the protein/peptide parameters are correlated. Due to the imbalance in the proteomics data, groups I and II are not orthogonal. It has been found that the estimation bias in the stagewise estimation of group I can be extreme due to misssing data [31]. Wang et al. [31] proposed to compute the experimental effects only on the balanced subset of peptides that appear in all experiments as one approach to avoid this. To more efficiently use the data, [19] proposed to use all the data in an ANOVA model. Considering the imbalance in the data across multiple iTRAQ experiments, [19] proposed to estimate the group II effects together with the group I effects for correct estimation of group I terms. When the fraction of differentially expressed proteins is small, group III is nearly orthogonal to the group I and II model parts. Thus, estimating the differential expression terms in group III separately from the terms in groups I and II is likely to be reliable for most research studies. However, estimation of groups I and II simultaneously is still too large for current computational resources.

### 2.2.3 Iterative regression

Iterative regression is an alternative approach proposed in [19] to address the estimation of groups I and II simultaneously. The Gauss-Siedel algorithm [6] for instance, also known as backfitting, is one iterative technique that iterates over the stages, so that each stage is repeatedly re-fit given the solution to the previous stages. Specifically, the iterative regression for model fitting of (1) works as below. First, backfitting is used to iteratively solve for parameters in groups I and II (the experimental and protein/peptide terms). Second, the final result of the iterative fit is used to normalize the data by substracting out the systematic bias factors from the fits of groups I and II. The residuals are the normalized data. Third, these normalized values are used as inputs for estimating the differential expression effects in group III. In this analysis, the term $g_{j(i),s}$ in group III is removed assuming that there will be differential expression of certain proteins between the samples of interest but that any increase in protein expression will affect all of the peptides for that protein equally. With the peptide effects included in the normalization stages of the model fitting, the group III parameters are separable and can be estimated one protein at a time. Thus, the normalized data are used as inputs for the differential expression model, and the latter was fit separately for each of the identified proteins. In summary, the normalization terms ($b_k$, $v_{k,l}$, $p_i$, $f_{j(i)}$) are estimated globally, whereas the group III differential protein effects ($r_{is}$, $r_s$) are not. Fitting group III parameters on a protein-by-protein basis assumes that each protein has a different variance parameter, rather than a global variance parameter.

### 2.2.4 Mixed effets models

Treating some effects, such as $f_{j(i)}$, in the model (1) as random, is equivalent to assuming a prior distribution for

the corresponding parameters. This introduces additional global parameters, the hyperparameters in the prior distributions, to the mixed effects model. Similar computational issues are involved in this mixed effects model. It is computationally challenging to fit the entire model to all data simultaneously for large datasets. Fitting separate models for each protein is invalid with respect to the global parameters. Data imbalance leads to the orthogonality requirement in a stagewise approach unsatisfied for the linear model design matrix corresponding to the multiple stages. So parameters from groups I and II must be estimated together to correctly estimate the group effects. But the standard iterative regression methods available for fixed effects models are not applicable to mixed effects models, and a solution remains an open problem.

### 2.3 Differential protein expression

With the fitted model for (1), the log difference of expression levels for protein $i$ between the $s$-th sample and the reference sample (without loss of generality, let $s = 1$ for the reference sample), denoted by $\theta_{i,s}$, is estimated by

$$\hat{\theta}_{i,s} = \left( \hat{r}_{i,s} + \hat{r}_s + \frac{1}{J_i} \sum_{j=1}^{J_i} \hat{g}_{j(i),s} \right) - \left( \hat{r}_{i,1} + \hat{r}_1 + \frac{1}{J_i} \sum_{j=1}^{J_i} \hat{g}_{j(i),1} \right), \tag{2}$$

where $J_i$ is the number of peptides derived from protein $i$. The 95% confidence interval for $\theta_{i,s}$ is constructed under the assumption of the normality of $\hat{\theta}_{i,s}$ as given by

$$\hat{\theta}_{i,s} \pm 1.96 \times \hat{se}(\hat{\theta}_{s,i}).$$

Hill et al. [7] and Oberg et al. [19] studied the factors that could lead to variations in the observed peptide intensities and inferred differential protein expression via ANOVA analysis. The model (1) includes the experiment-to-experiment variation which increases with the introduction of additional experiments. Not all model elements are identifiable from one application to the next, and model (1) does not include all sources of error, either. For example, Keshamouni et al. [12] proposed an alternative ANOVA model for the analysis of data from a single iTRAQ experiment comparing a control and treated sample. Neither ANOVA model considers the missingness in iTRAQ data, potentially biasing their results.

## 3. NONRANDOM MISSINGNESS

Luo et al. [17] overcomes the limitations of ANOVA models through a Bayesian framework that incorporates the nonrandom missingness in iTRAQ data sets. Their model assumes that the measured peptide intensities are affected by both protein expression levels and peptide specific effects. The values of these two effects across multiple experiments are modeled as random effects. When a sample is labelled with multiple tags in a single experiment, the variations across different isobaric tags are also modelled as random effects. The nonrandom missingness of peptide data is modeled with a logistic regression which relates the missingness probability for a peptide with the expression level of the protein that produces this peptide. A Markov chain Monte Carlo method tailored for this model was developed for the inference of relative expression levels across different samples.

### 3.1 Model

We focus on describing the model for iTRAQ data from multiple experiments and the estimation of the relative expression levels of proteins. When the iTRAQ data is obtained from multiple experiments, [17] utilizes a Bayesian hierarchical model in the sense that the model has an observation component that models the observed peptide intensities as random effects whose conditional distribution depends on the expected protein expression levels and peptide effects, and a second (hierarchical) component that defines the distributions of these expected values.

In Luo et al. [17], the labelling effects are assumed to be removed by normalization methods such as quantile normalization. Assume that there are $S$ ($\geq 2$) biological samples studied in $K$ ($\geq 2$) experiments. Since multiple isobaric tags may label the same sample in one experiment, let $L_s \geq 1$ denote the number of tags labelling the $s$th sample. Then $\sum_s L_s = M$ is the number of isobaric tags used in one experiment, which is 4 when we use 4-plex isobaric reagents and 8 in the 8-plex version. Assume that there are $I$ proteins in the sample and $J_i$ peptides for the $i$th protein. For the $l$th label of the $s$th sample in the $k$th experiment, let $y_{kijsln}$ denote the log transformed value of measured observed intensity for the $j$th peptide of the $i$th protein from the $n$th spectrum. Note that $j$ should be more appropriately denoted as $j(i)$ to explicitly indicate that peptides are nested within proteins, and $l$ should be denoted as $l(s)$ to indicate the $l$th labelled tag of the $s$th sample. For notational simplification, we omit the parentheses. The measured intensity of a peptide depends on the protein expression level and the peptide effect. Let $x_{kisl}$ denote the log transformed expression level of the $i$th protein of the $s$th sample with the $l$th labelling tag in the $k$th experiment. Let $z_{kij}$ denote the log transformed peptide effect for the $j$th peptide of the $i$th protein in the $k$th experiment. Luo et al. [17] considered an additive model for $y_{kijsln}$ ($k = 1, \ldots, K$; $i = 1, \ldots, I$; $j = 1, \ldots, J_i$; $s = 1, \ldots, S$; $l = 1, \ldots, L_s$; $n = 1, \ldots, N_{kijsl}$):

$$y_{kijsln} = x_{kisl} + z_{kij} + \epsilon_{kijsln}, \tag{3}$$

which corresponds to a multiplicative model in the original scale. In (3), $\epsilon_{kijsln}$ is assumed to be independently normally distributed with mean 0 and variance $\sigma_\epsilon^2$: $\epsilon_{kijsln} \sim N(0, \sigma_\epsilon^2)$.

#### 3.1.1 Missing data mechanism

The statistical model for peptide missingness in [17] was motivated by the study on the dataset obtained from the study of the roles of Caveolae for postnatal cardiovascular function. In this research, three experiments were conducted where the protein profiles from two wild-type mice and two knock-out Cav-1 mice were analyzed by iTRAQ with four isobaric tags in each experiment. Luo et al. [17] studied the proportion of peptides observed in one experiment but missing in another experiment, and found that there was a negative correlation between the missing probability and peptide intensity. In other words, less abundant peptides are more likely to be missing since they are harder to detect due to the data-dependent acquisition of the analysis process. Observing that there was an approximate linear relationship between the peptide missing probability and the observed intensity at the logit scale, Luo et al. [17] modeled the missing probability through a simple logistic regression model:

$$logit(P(I_{kijsln} = 1|y_{kijsln}, a, b)) = a + b \times y_{kijsln}, \tag{4}$$

where $I_{kijsln} = 1$ indicates that the $j$th peptide of the $i$th protein is measured in the $k$th experiment, the $l$th replicate of the $s$th sample and the $n$th spectrum. Formula (4) implies that the logit of the probability of peptide missingness is linearly dependent on its intensity. It is expected that $b > 0$ because peptides with lower intensities are more likely to be missing.

#### 3.1.2 Priors

The Bayesian hierarchical framework in [17] takes into account the variabilities across experiments and samples, and assumes that $x_{kisl}$ and $z_{kij}$ are independently normally distributed across different experiments, i.e.:

$$x_{kisl} \sim N(x_{isl}, \sigma_x^2), \tag{5}$$

$$z_{kij} \sim N(z_{ij}, \sigma_z^2), \tag{6}$$

where $x_{isl}$ and $z_{ij}$ denote the protein and peptide effects averaged over multiple experiments, respectively. The protein expression levels in different replicates (labelled with different tags) of the same sample are also assumed to be normally distributed:

$$x_{isl} \sim N(x_{is}, \sigma_\delta^2), \tag{7}$$

where $x_{is}$ denotes the expression level of the $i$th protein in the $s$th sample. Assumptions (5)–(7) lead to an equivalent form of (3):

$$y_{kijsln} = x_{is} + z_{ij} + e_{isl}^t + e_{kisl}^x + e_{kij}^z + \epsilon_{kijsln}, \tag{8}$$

where $e_{kisl}^x \sim N(0, \sigma_x^2)$ and $e_{kij}^z \sim N(0, \sigma_z^2)$ denote the random effects across experiments, and $e_{isl}^t \sim N(0, \sigma_\delta^2)$ denotes the variation among multiple replicates of the same sample. When a sample is labelled with a unique isobaric tag in an experiment, there is no replicate variation component within a sample. Formula (8) is a mixed-effects model. To ensure the identifiability of the model, the restriction $x_{i1} = 0$ is added. Then $x_{is}$ denotes the expression level of the $i$th protein in the $s$th sample relative to the first sample.

The second level of priors are normal distributions for $x_{is}$ and $z_{ij}$:

$$x_{is} \sim N(0, \tau_x^2) \quad \text{for } s > 1, \tag{9}$$

$$z_{ij} \sim N(0, \tau_z^2). \tag{10}$$

The hierarchical model is finished by assuming inverse gamma distributions as priors for the hyperparameters of variance: $\sigma_x^{-2} \sim \text{Gamma}(\gamma_1, \gamma_2)$, $\sigma_z^{-2} \sim \text{Gamma}(\gamma_3, \gamma_4)$, $\sigma_\delta^{-2} \sim \text{Gamma}(\gamma_5, \gamma_6)$ and $\sigma_\epsilon^{-2} \sim \text{Gamma}(\gamma_7, \gamma_8)$, where $\gamma_1$ and $\gamma_2$ denote the shape and scale parameters of a gamma distribution, respectively, and assuming $a \sim N(0, \nu^2)$ and $b \sim N(0, \nu^2)$. The posterior distributions of relevant parameters are simulated by MCMC simulations and differentially expressed proteins are identified by analyzing the posterior distribution of $x_{is}$.

### 3.2 Comparison to ANOVA analysis

The most important difference between this Bayesian model in [17] and the ANOVA model proposed by Hill et al. [7] and Oberg et al. [19] is that [17] clearly modeled the nonignorable missingness in iTRAQ data. Oberg et al. [19] remarked at the end of their paper that using a censoring mechanism to fit the model would be a natural next step. Instead of censoring the data at an unknown threshold value, [17] modeled a higher probability of peptide missingness for lower peptide intensities. These two methods also differ in terms of variations included in the model. The experimental effect and the replicative effect (when multiple tags label a sample) are considered constants for all proteins in the ANOVA model. In contrast, [17] modeled them as random effects that were specific to peptides and (or) proteins. Furthermore, the ANOVA analysis involves additional effects such as the labelling effect and the interaction between labelling and experimental effect $g_{j(i),s}$, which are not modeled in [17]. Inclusion of the labelling effect is determined by the experiment design. When identical tags are used to label the same samples in multiple experiments, the labelling effect is not identifiable since it is confounded with the sampling effect. It is meaningful to include the labelling effect only when different tags are used to label the same samples in multiple experiments. For the interaction between labelling and experimental effect $g_{j(i),s}$, although it is theoretically appropriate to have it in the model, there exists large uncertainty in the estimate of $g_{j(i),s}$ due to the small number of replicates (or no replicates) for each sample.

The common assumption in both the Bayesian method and the ANOVA analysis is that all of the peptide-based observations accurately reflect the intact proteins. We ignore

the possibility of homologous genes resulting in two or more proteins that share identical and nonidentical peptides as well as the possibility of post-transcriptional modifications. Although (1) includes the interaction between peptide effects and treatment ($g_{j(i),s}$), it is removed in the analysis of [19]. This term is not included in [17] either. So both [17] and [19] assume that certain proteins will have differential expressions across samples under different treatments, but that any change in protein expression will affect all of the peptides for that protein equally.

## 3.3 Nonrandom missingness in mass spectrometry data

Targeting for mass spectrometry data, the model (proposed by Wang et al. [31]) described in this subsection is not tailored for iTRAQ data. But since iTRAQ data are obtained by running the isolated peptides through MS/MS, this probability model provides an alternative way of studying the missingness in iTRAQ. Wang et al. [31] proposed to first remove sources of systamatic variation between MS profiles via global normalization, and then to investigate the intensity-dependent missingness and to impute the missed peptide intensities.

### 3.3.1 Global normalization

In their global normalization, [31] assumed that the sample intensities are all related by a constant factor which is to be chosen. In order to avoid the possible bias due to the nonrandom missingness in mass spectrometry data, Wang et al. proposed to use the top $L$ ordered statistics (e.g., medians) of peptide intensities in each sample for rescaling, where $L$ is a user-specified parameter. Let $K$ ($K > 2$) be the number of MS profiles. Denote the observed intensities of the $k$-th profile as $Y^{(k)} = (y_1^{(k)}, y_2^{(k)}, \ldots, y_{n_k}^{(k)})$, where $n_k$ is the number of peptides identified in the $k$-th profile. For a given number $L$ ($L < \min(\{n_k\}_{k=1}^K)$), the population median is defined as

$$\mu_0 = \frac{1}{K}\sum_k \text{median}(y_1^{(k)}, y_2^{(k)}, \ldots, y_{n_k}^{(k)}),$$

and the scaling coefficient for normalization of the $k$-th profile is

$$\lambda^{(k)} = \frac{1}{\mu_0}\text{median}(y_{(1)}^{(k)}, y_{(2)}^{(k)}, \ldots, y_{(L)}^{(k)}). \tag{11}$$

### 3.3.2 Nonrandom missingness and imputation

To account for the nonrandom missingness, Wang et al. [31] proposed to impute the missed peptide intensity in one sample with the ratio of the observed intensity in another sample divided by a scale coefficient estimated from the intensities of other peptides observed in both samples. Suppose the minimum detectable level of the instrument is $d$. Let $x_j^{(k)}$ be the true abundance of the $j$-th peptide in the $k$-th profile corresponding to the observed value $y_j^{(k)}$. A peptide may or may not exist in a profile. Let $z_j^{(k)}$ be a latent variable indicating the presence of the $j$-th peptide in the $k$-th profile, with $z_j^{(k)} = 1$ if the $j$-th peptide exists in the $k$-th profile, and $z_j^{(k)} = 0$ otherwise. Then $x_j^{(k)} = 0$ if $z_j^{(k)} = 0$. Let $f_j^{(k)}$ be the density function of $x_i^{(k)}$ when $z_i^{(k)} = 1$, we have

$$x_j^{(k)} \sim I_0(\cdot)P(z_j^{(k)} = 0) + f_j^{(k)}(\cdot)P(z_j^{(k)} = 1), \tag{12}$$

where $I_0(\cdot)$ indicates a point-mass at zero. With (12), Wang et al. [31] assumed that the true abundance of a peptide has a mixture distribution. With probability $P(z_j^{(k)} = 0)$, the peptide does not exist in the $k$-th profile, and the abundance is zero. With probability $P(z_j^{(k)} = 1)$, the peptide exists, and the distribution of the abundance is described by $f_j^{(k)}$.

The missed value of the intensity level of the $j$-th peptide present in the $k$-th profile is imputed by the expected value $E(x_j^{(k)}|y_j^{(k)} = 0)$, which is calculated as

$$\begin{aligned} &E(x_j^{(k)}|y_j^{(k)} = 0) \\ &\quad = E(x_j^{(k)}|y_j^{(k)} = 0, z_j^{(k)} = 1)P_d(z_j^{(k)} = 1|y_j^{(k)} = 0) \\ &\quad = E(x_j^{(k)}|x_j^{(k)} < d, z_j^{(k)} = 1)P_d(z_j^{(k)} = 1|y_j^{(k)} = 0), \end{aligned} \tag{13}$$

where the first equality is due to the fact that $E(x_j^{(k)}|y_j^{(k)} = 0, z_j^{(k)} = 0) = 0$, and the second equality is due to the fact that when the $j$-th peptide exists in the $k$-th profile ($z_j^{(k)} = 1$), no signal detection ($y_j^{(k)} = 0$) is equivalent to low intensity ($x_j^{(k)} < d$). The term $E(x_j^{(k)}|x_j^{(k)} < d, z_j^{(k)} = 1)$ in (13) can be determined when $f_j^{(k)}$ and $d$ are specified, and $P_d(z_j^{(k)} = 1|y_j^{(k)} = 0)$, the probability that the $j$-th peptide exists in the $k$-th profile when no signal is detected, can be calculated as

$$\begin{aligned} P_d(z_j^{(k)} = 1|y_j^{(k)} = 0) &= \frac{P_d(z_j^{(k)} = 1, y_j^{(k)} = 0)}{P_d(y_j^{(k)} = 0)} \\ &= \frac{P_d(y_j^{(k)} = 0|z_j^{(k)} = 1)P(z_j^{(k)} = 1)}{\sum_{z=0}^1 P_d(y_j^{(k)} = 0|z_j^{(k)} = z)P(z_j^{(k)} = z)} \\ &= \frac{P_d(x_j^{(k)} < d|z_j^{(k)} = 1)P(z_j^{(k)} = 1)}{P_d(x_j^{(k)} < d|z_j^{(k)} = 1)P(z_j^{(k)} = 1) + P(z_j^{(k)} = 0)} \end{aligned} \tag{14}$$

where the third equality holds because $P_d(y_j^{(k)} = 0|z_j^{(k)} = 1) = P_d(x_j^{(k)} < d|z_j^{(k)} = 1)$ and $P_d(y_j^{(k)} = 0|z_j^{(k)} = 0) = 1$. The term $P_d(x_j^{(k)} < d|z_j^{(k)} = 1)$ in (14) can be obtained from the density function $f_j^{(k)}$ when the latter is specified, and

$$P(z_j^{(k)}=1)=\frac{P_d(x_j^{(k)}>d,z_j^{(k)}=1)}{P_d(x_j^{(k)}>d|z_j^{(k)}=1)}=\frac{P_d(y_j^{(k)}>0)}{P_d(x_j^{(k)}>d|z_j^{(k)}=1)}. \tag{15}$$

So when the conditional density $f_j^{(k)}$ and $d$ are specified, the missed peptide intensity can be imputed with (13)–(15).

The minimum instrument detectable level parameter $d$ is estimated by the background noise level in all MS raw profiles from the same instrument, denoted as $\hat{d}$. Then the detectable level of the $k$-th profile is $\tilde{d}^{(k)}=\hat{d}/\lambda^{(k)}$, where $\lambda^{(k)}$ is the normalization scale coefficient in (11). Wang et al. [31] assume that

$$\frac{x_j^{(k)}}{\lambda^{(k)}}\,\Big|\,(z_j^{(k)}=1)\sim N(\mu_j,\sigma_j^2)$$

independently for $k=1,2,\ldots,K$. This is equivalent to the assumption that the density function of $x_j^{(k)}$ when $z_j^{(k)}=1$, $f_j^{(k)}$, is $N(\lambda^{(k)}\mu_j,(\lambda^{(k)}\sigma_j)^2)$. In the special case that $\sigma_j\ll|\tilde{d}^{(k)}-\mu_j|$ and biological replications are available, Wang et al. [31] provided estimators for the missing probability $P_d(z_j^{(k)}=1|y_j^{(k)}=0)$ and the imputed value $E(X_j^{(k)}|y_j^{(k)}=0)$ as below:

$$P_d(z_j^{(k)}=1|y_j^{(k)}=0)=\begin{cases}\hat{P}(z_j^{(k)}=1), & \text{if } \hat{\mu}_j<\tilde{d}^{(k)},\\ 0, & \text{if } \hat{\mu}_j>\tilde{d}^{(k)};\end{cases}$$

$$E(X_j^{(k)}|y_j^{(k)}=0)=\begin{cases}\hat{\mu}_j\hat{P}(z_j^{(k)}=1), & \text{if } \hat{\mu}_j<\tilde{d}^{(k)},\\ 0, & \text{if } \hat{\mu}_j>\tilde{d}^{(k)};\end{cases}$$

where

$$\hat{\mu}_j=\frac{\sum_k y_j^{(k)}/\lambda^{(k)}}{\sum_k I(y_j^{(k)}>0)},$$

and

$$\hat{P}(z_j^{(k)}=1)=\frac{\sum_k I(y_j^{(k)}>0)}{\sum_k I(\hat{\mu}_j>\tilde{d}^{(k)})}.$$

The imputed data is used for further analysis such as estimation, clustering of proteins and differential protein identification.

The model proposed by Wang et al. [31] differs from the Bayesian model proposed by Luo et al. [17] in the following three ways. First, in [31], intensities lower than a certain level are censored and the censoring parameter is estimated based on the background noise levels; in [17], a logistic regression model is built to relate the missing probability with the potential true intensity. With the observation that less abundant peptides are more likely to be missing, the model based missing mechanism in [17] which links the probability of missing with peptide intensity is more reasonable than the censoring mechanism in [31]. Second, [31] conducts single value imputation and imputes the missed intensities with the expected values, while [17] conducts multiple imputation and simulates the posterior distributions of missed values. Third, [31] is not tailored for iTRAQ analysis and sources of variations should be removed when applying the idea in [31] to iTRAQ data. The strength of [31] lies in the smaller computation burden. When the density $f_j^{(k)}$ is specified, the missed peptide intensity can be easily imputed with the expected value obtained from formula (13).

## 4. CONCLUSION AND FUTURE DIRECTIONS

The protein and peptide identification from MS/MS data has been addressed by many researchers [9, 11, 15, 18, 22, 24, 25, 28]. In this article, we have focused on the quantitation of protein and peptide expression levels from iTRAQ data, which is a shotgun technique that uses isobaric tags to label peptides from different samples and analyzes the labelled peptides with tandem mass spectrometry. We have reviewed the studies on the sources of variations, the computational problems involved and the nonrandom missingness in the iTRAQ data. These studies are conducted after the protein database search for protein and peptide identification have been conducted from the collection of spectra, and the peak areas of the ions originating from the isobaric tags have been normalized for the estimation of peptide intensities. The uncertainties in the protein and peptide identification and the peak area evaluation are not considered. Furthermore, these studies assume that all of the peptide-based observations accurately reflect the intact proteins. It is possible that homologous genes can result in two or more proteins that share identical and nonidentical peptides. The possibility of post-transcriptional modifications is also ignored. The quantitation of protein would benefit from the improvement of protein identification and peak area evaluation from mass spectra.

As discussed above, due to the complex nature of iTRAQ data, it is very important to use sound experimental design and analysis strategies when using iTRAQ technology to detect and quantify the relative protein expression levels across samples, especially when multiple experiments are involved. Poor experimental design and analysis may confound signals with noises and lead to protein and peptide effects undistinguishable from systematic variations. To achieve the best power in sample comparisons, it is important to balance the treatment groups across experiments and to randomize the isobaric tags for samples, as much as possible, in the application of iTRAQ for comparative proteomic researches.

The nonrandom missingness in iTRAQ is modeled with a simple logistic regression in [17]. It is natural to consider more complex missingness models that include polynomial or local polynomial terms in the logistic regression, if the

latter better describe the relationship between the missing probability and the peptide intensity. These missingness models can also be built in the Bayesian hierarchical structure as in [17] to infer the relative expression levels of proteins across samples.

## ACKNOWLEDGEMENTS

The work was supported in part by NIH grants HV28286, DA018343, GM59507 and NSF grant DMS 0714817. The work was also supported in part by "Yale University Biomedical High Performance Computing Center" and NIH grant: RR19895, which funded the instrumentation.

*Received 05 May 2011*

## REFERENCES

[1] Boylan, K. L., Andersen, J. D., Anderson, L. B., Higgins, L. and Skubitz, A. P. (2010) Quantitative proteomic analysis by itraq for the identification of candidate biomarkers in ovarian cancer serum. *Proteome Science*, http://www.proteomesci.com/content/8/1/31.

[2] Casado-Vela, J., Martínez-Esteso, M. J., Rodriguez, E., Borrás, E., Elortza, F. and Bru-Martínez, R. (2010) iTRAQ-based quantitative analysis of protein mixtures with large fold change and dynamic range. *Proteomics*, 343–347.

[3] Choe, L., D'Ascenzo, M., Relkin, N. R., Pappin, D., Ross, P., Williamson, B., Guertin, S., Pribil, P. and Lee, K. H. (2007) 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics*, **7**, 3651–3660.

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

[5] Hamdan, M. and Righetti, P. G. (2002) Modern strategies for protein quantification in proteome analysis: Advantages and limitations. *Mass Spectrometry Reviews*, **21**, 287–302.

[6] Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. New York: Chapman and Hall. MR1082147

[7] Hill, E. G., Schwacke, J. H., Comte-Walters, S., Slate, E. H., Oberg, A. L., Eckel-Passow, J. E., Therneau, T. M. and Schey, K. L. (2008) A statistical model for iTRAQ data analysis. *Journal of Proteome Research*, **7**, 3091–3101.

[8] Hu, H.-D., Ye, F., Zhang, D.-Z., Hu, P., Ren, H. and Li, S.-L. (2010) iTRAQ quantitative analysis of multidrug resistance mechanisms in human gastric cancer cells. *Journal of Biomedicine and Biotechnology*, DOI: 10.1155/2010/571343.

[9] Kall, L., Canterbury, J., Weston, J., Noble, W. S. and MacCoss, M. J. (2007) A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nature Methods*, **4**, 923–925.

[10] Karp, N. A., Huber, W., Sadowski, P. G., Charles, P. D., Hester, S. V. and Lilley, K. S. (2010) Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular & Cellular Proteomics*, **9**, 1885–1897.

[11] Keller, A., Nesvizhskii, A., Kolker, E. and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.

[12] Keshamouni, V. G., Michailidis, G., Grasso, C. S., Anthwal, S., Strahler, J. R., Walker, A., Arenberg, D. A., Reddy, R. C., Akulapalli, S., Thannickal, V. J., Standiford, T. J., Andrews, P. C. and Omenn, G. S. (2006) Differential protein expression profiling by iTRAQ-2DLC-MS/MS of lung cancer cells undergoing epithelial-mesenchymal transition reveals a migratory/invasive phenotype. *Journal of Proteome Research*, 1143–1154.

[13] Kilner, J., Zhu, L., Ow, S. Y., Evans, C. and Corfe, B. M. (2011) Assessing the loss of information through application of the 'two-hit rule' in iTRAQ datasets. *Journal of Integrated Omics*, **1**, 124–134.

[14] Lau, E., Lam, M. P. Y., Siu, S. O., Kong, R. P. W., Chan, W. L., Zhou, Z., Huang, J., Lo, C. and Chu, I. K. (2011) Combinatorial use of offline scx and online RP–RP liquid chromatography for itraq-based quantitative proteomics application. *Molecular BioSystems*, DOI: 10.1039/C1MB05010A.

[15] Li, Q., MacCoss, M. J. and Stephens, M. (2010) A nested mixture model for protein identification using mass spectrometry. *Ann. Appl. Stat.*, **4**, 962–987. MR2758429

[16] Liu, H., Sadygov, R. G. and Yates, J. R. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, **76**, 4193–4201.

[17] Luo, R., Colangelo, C. M., Sessa, W. C. and Zhao, H. (2009) Bayesian analysis of iTRAQ data with nonrandom missingness: Identification of differentially expressed proteins. *Statistics in Bioscience*, DOI: 10.1007/s12561-009-9013-2.

[18] Nesvizhskii, A. I., Keller, A., Kolker, E. and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 4646–4653.

[19] Oberg, A., Mahoney, D., Eckel-Passow, J., Malone, C., Wolfinger, R., Hill, E., Cooper, L., Onuma, O., Spiro, C., Therneau, T. and Bergen, H. (2008) Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *Journal of Proteome Research*, **7**, 225–233.

[20] O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, **250**, 4007–4012.

[21] Patton, W. F. (2002) Detection technologies in proteome analysis. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, **771**, 3–31.

[22] Price, T. S., Lucitt, M. B., Wu, W., Austin, D. J., Pizarro, A., Yocum, A. K., Blair, I. A., FitzGerald, G. A. and Grosser, T. (2007) Ebp, a program for protein identification using multiple tandem mass spectrometry data sets. *Mol. Cell. Proteomics*, **6**, 527–536.

[23] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, **3**, 1154–1169.

[24] Sadygov, R., Liu, H. and Yates, J. (2004) Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.*, **76**, 1664–1671.

[25] Sadygov, R. and Yates, J. (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.*, **75**, 3792–3798.

[26] Salim, K., Kehoe, L., Minkoff, M. S., Bilsland, J. G., Munoz-Sanjuan, I. and Guest, P. C. (2006) Identification of differentiating neural progenitor cell markers using shotgun isobaric tagging mass spectrometry. *Stem Cells and Development*, **15**, 461–470.

[27] Seshi, B. (2006) An integrated approach to mapping the proteome of the human bone marrow stromal cell. *Proteomics*, **6**, 5169–5182.

[28] Shen, C., Wang, Z., Shankar, G., Zhang, X. and Li, L. (2008) A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. *Bioinformatics*, **24**, 202–208.

[29] Skorobogatko, Y. V., Deuso, J., Adolf-Bergfoyle, J., Nowak, M. G., Gong, Y., Lippa, C. F. and Vosseller, K. (2010) Human alzheimer's disease synapticO-GlcNAcsite mapping and

iTRAQ expression proteomics with ion trap mass spectrometry. *Amino Acids*, **40**, 765–779.

[30] Unwin, R. D., Griffiths, J. R. and Whetton, A. D. (2010) Simultaneous analysis of relative protein expression levels across multiple samples using iTRAQ isobaric tags with 2D nano LC-MS/MS. *Nature Protocols*, **5**, 1574–1582.

[31] Wang, P., Tang, H., Zhang, H., Whiteaker, J., Paulovich, A. G. and Mcintosh, M. (2006) Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pacific Symposium on Biocomputing*, **11**, 315–326.

[32] Wu, W. W., Wang, G., Baek, S. J. and Shen, R.-F. (2006) Comparative study of three proteomic quantitative methods, DIGE, cICAT, and iTRAQ, using 2D Gel- or LC-MALDI TOF/TOF. *Journal of Proteome Research*, **5**, 651–658.

[33] Ye, H., Hill, J., Kauffman, J. and Han, X. (2010) Qualitative and quantitative comparison of brand name and generic protein pharmaceuticals using isotope tags for relative and absolute quantification and matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Analytical Biochemistry*, **400**, 46–55.

Ruiyan Luo
Department of Mathematics and Statistics
Georgia State University
30 Pryor Street
Atlanta, GA 30303
USA
E-mail address: matrxl@langate.gsu.edu

Hongyu Zhao
Department of Epideomiology and Public Health
Yale University
300 George Street
Suite 503
New Haven, CT 06511
USA
E-mail address: hongyu.zhao@yale.edu