Accurate Mass Spectrometry Based Protein Quantification via Shared Peptides

BANU DOST,¹ NUNO BANDEIRA,¹ XIANGQIAN LI,² ZHOUXIN SHEN,²
STEVEN P. BRIGGS,² and VINEET BAFNA¹

ABSTRACT

In mass spectrometry-based protein quantification, peptides that are shared across different protein sequences are often discarded as being uninformative with respect to each of the parent proteins. We investigate the use of shared peptides which are ubiquitous (~50% of peptides) in mass spectrometric data-sets for accurate protein identification and quantification. Different from existing approaches, we show how shared peptides can help compute the relative amounts of the proteins that contain them. Also, proteins with no unique peptide in the sample can still be analyzed for relative abundance. Our article uses shared peptides in protein quantification and makes use of combinatorial optimization to reduce the error in relative abundance measurements. We describe the topological and numerical properties required for robust estimates, and use them to improve our estimates for ill-conditioned systems. Extensive simulations validate our approach even in the presence of experimental error. We apply our method to a model of *Arabidopsis thaliana* root knot nematode infection, and investigate the differential role of several protein family members in mediating host response to the pathogen. Supplementary Material is available at www.liebertonline.com/cmb.

Key words: ITRAQ, linear programming, mass spectrometry optimization, protein quantification, shared peptides.

1. INTRODUCTION

THE ANALYSIS OF THE PROTEOME USING MASS SPECTROMETRY TYPICALLY INVOLVES the separation of molecules (often, enzymatically digested peptides from expressed proteins) followed by tandem mass spectrometry (MS/MS) and mass measurement of each molecule, termed as the mass-spectrum. Together with mass, the spectrum also measures peak-intensity for each molecule. For any constituent peptide from a protein sequence, its spectral intensity is also a measurement of *abundance*, the amount of the expressed protein. However, the actual value is hard to interpret, as it depends upon a number of poorly understood factors, including instrument types, energetics of the process, and physico-chemical properties of the peptide itself (Jammes et al., 2005; Mallick et al., 2007). Consequently, it is often the *relative-abundance* of a

¹Department of Computer Science and Engineering and ²Cell & Developmental Biology, UC San Diego, San Diego, California.

peptide, measured as the ratio of intensities of a peptide across samples (conditions), that is investigated. By the same token, intensity values of different peptides are usually not comparable.

Using the relative abundance of a peptide as a proxy for the relative abundance of the parent protein is acceptable only when the peptide sequence is unique to the protein. By contrast, when a peptide is shared across proteins (e.g., proteins that share domains), its abundance (and relative abundance) depends upon contributions from multiple proteins. For this reason, shared peptides have been traditionally considered a challenging nuisance for protein identification and are typically disregarded in protein-level quantification analysis. However, this may significantly decrease the number of proteins for which abundance estimates can be obtained. While often unreported, a significant portion of the peptides and proteins (as much as 50%) are ignored. In our analysis of *Arabidopsis thaliana*, 4, 145 (48%) of the 8, 584 expressed proteins were not represented by a unique peptide and would normally be discarded.

Here, we demonstrate that shared peptides are a resource that adds value as illustrated by two simple examples in Figure 1. Consider a case with two proteins p_1, p_2 , and three constituent peptides s_1, s_2, s_3 , where s_1, s_2 are unique, and s_3 is shared. See Figure 1a for where a peptide is connected to a protein by an edge if the protein contains the peptide. For these examples, let the relative abundances (r_1, r_2, r_3) of the three peptides over two samples B and A be 16, 1, 4, respectively. The typical approach is to discard the shared peptide s_3 , and to assert that p_1 is $16\times$ over-expressed, while p_2 is unchanged. Formally, if Q_j^A, Q_j^B represent the actual abundance of protein j in samples A and B , respectively, then

$$\frac{Q_1^A}{Q_1^B} = r_1 = 16 \text{ and } \frac{Q_2^A}{Q_2^B} = r_2 = 1.$$

However, the ignored peptide s_3 also provides information because

$$r_3 = 4 \simeq \frac{Q_2^A + Q_1^A}{Q_2^B + Q_1^B} = \frac{Q_2^B + 16Q_1^B}{Q_2^B + Q_1^B} = \frac{16 + \frac{Q_2^B}{Q_1^B}}{1 + \frac{Q_2^B}{Q_1^B}}.$$

Thus, we additionally learn that $\frac{Q_2^B}{Q_1^B} = 4$, indicating that p_2 is $4\times$ more abundant than p_1 in sample B . Here, we have four unknowns from the two proteins, and four constraints, one from each of the peptide and one additional constraint that fixes the sum of the Q_j^B s. By using ratios (e.g., $R_j = Q_j^A / Q_j^B$), we reduce the number of unknowns and can solve to get the extra information. Note that the unit of measurement for Q_j^A, Q_j^B is immaterial. For this reason, we always reduce one degree of freedom, typically by adding the constraint $\sum_j Q_j^B = 100$, or solving for ratios, as we do here. As long as the number of constraints is not less than the number of unknowns, shared peptides enable the calculation of the relative abundances of *different* proteins in addition to the variations of the *same* proteins between different conditions.

FIG. 1. (a,b) Two examples illustrating our approach for protein quantification via shared peptides. **(c)** Protein-peptide bi-partite graph $G=(P \cup S, E)$ representing a mapping between m proteins and n peptides.

Figure 1b shows a more complex example with three proteins (p_1 – p_3) and five peptides (s_1 – s_5). Here, protein p_2 does not have any unique peptide, and would normally be predicted as non-existent even if it might be abundant in the samples. However, there are cases where it is evident that p_2 is existent. For instance, if $r_1 = 1, r_2 = r_3 = r_4 = 100$ and $r_5 = 1$ in the example, we can conclude that p_2 exists in the sample with a high relative abundance with respect to p_1 and p_3 . Formally, the corresponding system has $5 + 1$ constraints and six unknowns. Therefore, solving the system for Q_i^A, Q_i^B , we first identify if p_2 is abundant (i.e., $Q_2^A > 0$ and/or $Q_2^B > 0$). If so, we then compute its relative abundance $\frac{Q_2^A}{Q_2^B}$.

To summarize, shared peptides provide extra information in protein quantification. Note that by accurate protein quantification, we also obtain protein identifications that are more accurate than traditional maximum parsimony approaches. Under certain conditions, shared peptides allow us to (a) compute relative abundance of a protein even when it does not contain a unique peptide and (b) compute relative abundance values of two different proteins in a sample. To our knowledge, this is the first article to exploit shared peptides in this manner. However, the simple idea is confounded by the realities of missing data and error in experiments. Here, we lay out the theoretical foundations and practical considerations in determining when the shared peptide abundances can be used reliably. We show that the accuracy of protein quantification depends on the topological properties of the peptide-protein relationships as well as on the numerical properties of the experimentally determined intensity values; sometimes interesting cases cannot be resolved because of missing data, or numerical instability.

As an extension to our approach, we also consider some intrinsic properties of peptides. Informally, we define the *detectability* of a peptide as the probability that it will be detected via MS, when the parent protein is expressed. We propose an alternative formulation that accounts for the peptide detectabilities in addition to absolute and relative abundances of proteins when appropriate data is available.

Furthermore, we suggest two improvements to increase the number of systems (cases) that can be solved. First, we describe an algebraic technique based on singular value decomposition to make robust inferences for numerically ill-conditioned systems. Recent results have shown that detectability is indeed an intrinsic characteristic of peptides that can be computed in independent experiments and maintained for future use (Alves et al., 2007). We also point out that, by incorporating detectabilities as known variables in our formulation, it is possible to solve a much larger number of systems.

In Section 2, we describe the theoretical and empirical considerations for shared peptide analysis. In Section 3.1, we validate our approach with extensive simulations. We apply our methods to data from ITRAQ experiments comparing an *Arabidopsis thaliana* model of root-knot infection versus wild-type in Section 3.2. Finally, we elucidate the relative abundance among different members of a family in over 55 *Arabidopsis* protein families.

2. PROTEIN IDENTIFICATION AND QUANTIFICATION VIA SHARED PEPTIDES

We represent the protein quantification data using a bipartite graph $G = (P \cup S, E)$ where P is the set of proteins and S is the set of **detected** peptides. For all $p \in P, s \in S, (p, s) \in E$ if and only if peptide s is a substring of the protein sequence p . Note that different connected components of G do not influence each other, and we treat each component independently. Without loss of generality, assume that G is connected, and let $|P| = m$ and $|S| = n$. (Fig. 1c). Consider the case where only two samples are involved. In many experiments, the abundances are measured before and after a treatment, so we denote the samples as B , and A . We associate two variables (Q_j^B, Q_j^A) corresponding to the “before” and “after” abundance for each protein $p_j \in P$. As mentioned earlier, we also add the constraint $\sum_j Q_j^B = 100$.

Analogous to proteins, we associate values q_i^B, q_i^A, r_i with each peptide $s_i \in S, i = 1 : n$ where $r_i = q_i^A / q_i^B$ denotes the ratio of the peptide s_i abundance between samples. It is possible to generalize the representation for the data with more than two samples. While this abstraction hides many of the complexities of protein quantification via mass spectrometry, it is useful to present our approach which can be applied to many different quantification protocols, including labeled and label-free approaches.

Key to our computation are equations that connect all proteins p_j which contain a single peptide s_i . In the absence of experimental error, the abundance values must satisfy the following $n + 1$ constraints over $2m$ variables.

$$\sum_{(p_j, s_i) \in E} Q_j^A - r_i \times \sum_{(p_j, s_i) \in E} Q_j^B = 0 \quad \text{for all } s_i \in S$$

$$\sum_j Q_j^B = 100$$

With no errors, we can solve this equation uniquely as long as $n + 1 \geq 2m$. To incorporate errors, we consider a *linear-programming* formulation that minimizes the total error (Fig. 2).

Note that the ratios are not symmetric about 1, so we always choose a constraint where the ratio contribution is greater than 1. To simplify notation, we will also represent the LP formulation in a matrix form as

$$\min \sum_i |\varepsilon_i| \quad \text{where } \varepsilon = \mathcal{A}x - b, x \geq 0 \quad (1)$$

where x is vector of dimension $2m$, given by $x = [Q_1^B, \dots, Q_m^B, Q_1^A, \dots, Q_m^A]^T$, b is a $(n + 1)$ -dimensional vector described by $b = [100, 0, \dots, 0]^T$, and $\mathcal{A}$ is a $(n + 1) \times 2m$ matrix. While this LP is not in standard form, it can easily be transformed into one.

The formulation of the linear program is natural in that the LP seeks for protein abundances that optimally fit the observed peptide ratios. Nevertheless, it raises questions about our confidence in the estimates of Q_j . Note first that a low value for the objective does not necessarily result in robust estimates of Q_j . Consider an under-determined system, where $n + 1 < 2m$. By setting an arbitrary subset of $2m - (n + 1)$ variables to 0 and solving for the remaining, we obtain multiple solutions, each with 0 error. A simple illustration of this is found in the notion of *symmetric proteins*. Define proteins $p_1 \in P$ and $p_2 \in P$ as symmetric if and only if the set of incident peptides $S_1 = \{s | (p_1, s) \in E\}$ and $S_2 = \{s | (p_2, s) \in E\}$ are identical. Two symmetric proteins imply two identical columns in $\mathcal{A}$, which means that any linear combination of abundances for these two proteins will lead to an identical solution. Certainly, we can solve this problem as a special case: simply merge the two identical columns (proteins) into one, effectively reducing m . However, more complex dependencies might arise which are harder to detect.

Generalizing, when $\text{rank}(\mathcal{A}) < 2m$, we get multiple solutions with zero-error. If however, $\mathcal{A}$ has full column rank, then by parsimony arguments, the LP solution is likely to provide accurate estimates of protein abundance values. Even if the system is full-rank, it might be ill-conditioned, resulting in poor estimates. We define a *rank-threshold* function to characterize the solvability, a quantity that is closely

FIG. 2. Input, output, and computation summary of two LP formulations for protein quantification via shared peptides. **(a)** F_1 : A formulation that does not include peptide detectability. **(b)** F_2 : Using peptide detectabilities. We use $f_i = 1/d_i$ as the reciprocal of detectability to maintain linear constraints.

	Input	Output	Formulation
F_1	$r_i, \forall s_i \in S$	$Q_j^B, Q_j^A,$ $\forall p_j \in P$	$\min \sum_{i=1}^n \varepsilon_i $ $\text{s.t. } \sum_{p_j \in P} Q_j^B = 100$ $\varepsilon_i = \sum_{(p_j, s_i) \in E} Q_j^A - r_i \times \sum_{(p_j, s_i) \in E} Q_j^B \quad \forall s_i \in S, r_i \geq 1$ $\varepsilon_i = r_i \times \sum_{(p_j, s_i) \in E} Q_j^A - \sum_{(p_j, s_i) \in E} Q_j^B \quad \forall s_i \in S, r_i \leq 0$ $Q_j^B \geq 0, Q_j^A \geq 0 \quad \forall p_j \in P.$
F_2	$q_i^B, q_i^A, r_i,$ $\forall s_i \in S$	$Q_j^B, Q_j^A,$ $\forall p_j \in P$ $d_i, \forall s_i \in S$	$\min \sum_{i=1}^n \varepsilon_i^B + \varepsilon_i^A $ $\text{s.t. } \sum_{p_j \in P} Q_j^B = 100$ $\varepsilon_i^B = \sum_{(p_j, s_i) \in E} Q_j^B - q_i^B f_i, \quad \forall s_i \in S$ $\varepsilon_i^A = \sum_{(p_j, s_i) \in E} Q_j^A - q_i^A f_i, \quad \forall s_i \in S$ $Q_j^B \geq 0, Q_j^A \geq 0, \quad \forall p_j \in P.$

related to the condition number of the matrix. Start with the singular-value decomposition of $\mathcal{A}$. Using standard approaches, compute matrices U , V , and Σ such that

$$\mathcal{A} = V\Sigma U^T$$

Σ is a $(n + 1) \times 2m$ diagonal matrix with nonnegative real numbers $\sigma_1, \dots, \sigma_p$ on the diagonal. These p diagonal entries describe the singular values of $\mathcal{A}$ in decreasing order of magnitude, where $p \leq \min\{n + 1, 2m\}$. U is orthonormal with dimensionality $2m \times 2m$, and V is orthonormal with dimensionality $n + 1 \times n + 1$, respectively. The rank of $\mathcal{A}$ is given by the number of non-zero singular values. We define a related concept, *rank-threshold* of $\mathcal{A}$ as

$$\mathcal{R}(\mathcal{A}) = \min\{t \mid \sigma_j > 10^{-t} \forall j\}$$

$\mathcal{R}(\mathcal{A}) = t$ being low implies that all its singular values are large ($\geq 10^{-t}$), implying that estimates of protein abundance values should be robust. In our experiments, we will show that the rank-threshold is a good way to characterize the reliability of the final solution.

Robust estimates for ill-conditioned systems: This formalism allows us to distinguish high rank systems $\mathcal{A}$ for which we can estimate protein abundance reliably, but it also provides a handle into under-determined systems. Using our notation, we can describe an under-determined system as one in which $\mathcal{R}(\mathcal{A})$ is high. Specifically, $\mathcal{R}(\mathcal{A}) = \infty$ implies the case when some of the singular values are 0. For a rank threshold t , define the *rank_t* of $\mathcal{A}$ as

$$\text{rank}_t(\mathcal{A}) = \max\{j \mid \sigma_j > 10^{-t}\}$$

This “thresholded” rank allows us to get the true dimensionality of a system for which we could get robust results. For all j , let U_j denote the $2m \times j$ matrix formed by taking the first j columns of U (corresponding to the dominant singular values). Likewise, let V_j denote the matrix formed by the first j columns of V , and $\Sigma_j = \text{diag}[\sigma_1, \dots, \sigma_j]$. This implies that if $\text{rank}_t(\mathcal{A}) = k$, then

$$\mathcal{R}(\mathcal{A}U_k) = \mathcal{R}(V_k\Sigma_k) = t$$

We choose $B = \mathcal{A}U_k$, and solve the linear program for the k dimensional vector y

$$\min \sum_i |\varepsilon_i| \text{ where } \varepsilon = By - b, U_k y \geq 0 \quad (2)$$

Note that $\mathcal{R}(B) = t$ implying that the estimates of y are robust. The reason to keep y unconstrained, but impose $U_k y \geq 0$ is the following: The values y cannot be interpreted directly, but can be used to retrieve protein abundance values x by solving

$$x = U_k y$$

Our constraints ensure that the protein abundance values are non-negative.

2.1. Incorporating peptide detectability

Here we consider an alternative formulation that builds on different assumptions to improve robustness to measurement errors and potentially greatly increase the numbers of components that can be solved. Assuming one is able to estimate the *absolute* peptide abundances q_i^B and q_i^A (Bantscheff et al., 2007), this formulation allows one to relate the absolute peptide abundance with the total abundance of its parent proteins and thus make inferences about peptide *detectabilities* in addition to relative protein abundances.

We define the detectability of a peptide s_i as a quantity $d_i \in [0, 1]$ that relates peptide abundance to the abundances of its parent proteins. In the absence of experimental error, for each peptide $s_i \in S$,

$$\begin{aligned} q_i^B &= d_i \times \sum_{(p_j, s_i) \in E} Q_j^B \\ q_i^A &= d_i \times \sum_{(p_j, s_i) \in E} Q_j^A \end{aligned}$$

In dealing with errors, we use a linear programming formulation that is similar to F_1 , but with $2n + 1$ constraints and $2m + n$ variables (Fig. 2). We use $f_i = \frac{1}{d_i}$ as the reciprocal of detectability to maintain

linearity of equations. The previous discussion regarding reliable estimates of abundance values is unchanged from the previous section.

The ITRAQ data does not provide peptide abundance values that can be used directly for F_2 . However, recent developments indicate that the absolute peptide abundances can be experimentally estimated (Bantscheff et al., 2007). Also, recent results have shown that peptide detectabilities can be reliably estimated with very little variability across mass spectrometry runs (Alves et al., 2007; Mallick et al., 2007). This observation is especially important in F_2 . The knowledge of peptide detectabilities implies $2m$ variables instead of $2m + n$ and greatly increases the number of cases that can be solved.

In a recent personal communication, we were informed that a similar approach was considered by Xue et al. (2007). Instead of using detectabilities to infer absolute abundances, they used spectral counts of peptides as measures of absolute abundances, and an EM algorithm to distribute the spectral counts of shared peptides.

3. RESULTS

Data-set: We choose an *Arabidopsis thaliana* model of root-knot nematode infection. The root-knot nematodes are worm-like, microscopic plant-parasites that infect a multitude of plants, including all major crops, turf, and many ornamental plants. The diversity and extent of infection makes it economically significant to explore. The typical mode of infection is via the root. The female nematode lays its eggs at the root tip. The juveniles infect via the root tip, and move up. Inside, they manipulate the cellular machinery to create specialized *feeding* cells, which grow and multi-nucleate, but do not divide, eventually forming giant cells that provide nutrients to the parasite (Sasser et al., 1985). As the nematodes exploit the *Arabidopsis* cellular machinery to create the giant cell phenotype, an analysis of proteins that are differentially expressed in infected versus non-infected host cells can help elucidate the underlying mechanism (Jammes et al., 2005). As the *Arabidopsis* genome is sequenced, with extensive annotation on the known genes and pathways, it is an appropriate model for the host.

An ITRAQ method was used to collect protein abundance information. A brief overview of the method is given here (Wiese et al., 2007). The samples are enzymatically digested into short peptides. Peptides from different samples are *N*-terminally covalently labeled with tags of different mass, but then pooled and analyzed together using tandem mass spectrometry. Each spectrum contains both the fragment masses used to identify the peptide, and the intensities of the differential tags for abundance computation. In our terminology, for every peptide s_i , we read the intensities of the two tags as $q_i^A q_i^B$, and compute the ratio $r_i = q_i^A / q_i^B$, which approximates the ratio of the peptide abundance values in the two samples.

Our data-set is a collection of 118,426 spectra, encoding 27,728 peptides mapping onto 8,584 protein sequences. Each protein is mapped to at least one peptide and vice versa. The number of peptides mapping to a protein sequence varies considerably, ranging from 1 to 59. The distribution of the number of peptides per protein is shown in Figure 3.

Close to half of the proteins (4,145 out of 8,584) do not have a unique peptide. The number of unique peptides per protein range from 0 to 51. The distribution of the number of unique peptides per protein is shown in Figure 3. Likewise, there is tremendous spectral redundancy among peptides, with the number of spectra encoding a peptide ranging from 1 to 975. Close to half of the peptides (10,166) are shared by multiple protein sequences. We reduce the data by merging *symmetric* peptides, or peptides that belonged to an identical subset of proteins. The redundancy helps understand the measurement error, and the merging removes artificial dimensionality, giving a better measure of rank, and rank-threshold. Likewise, we also merge the symmetric proteins for reasons mentioned earlier.

After merging, we obtain a protein-peptide bi-partite graph $G = (P \cup S, E)$, where $|P| = 6,998$, $|S| = 8,069$, $|E| = 13,055$. G has 4119 connected components projecting onto 257 non-isomorphic topologies with size ranging from 2 to 127. In this study, we consider only the 1190 non-trivial components, with at least two proteins.

In addition to testing on this data, we also perform a series of controlled experiments by simulating data-sets based on the topologies of the *Arabidopsis* data-set.

Generation of simulation data: We start with 257 topologically distinct (non-isomorphic) components of the *Arabidopsis* data and generated 100 data-sets from each topology with different values (simulation data is available at <http://www.cse.ucsd.edu/~bdost/downloads/SimData.zip>) For each component, we do the following:

FIG. 3. Mapping of peptides to proteins in *Arabidopsis* root-knot nematode infection ITRAQ data. **(a)** Distribution of the number of peptides and unique peptides per protein. **(b)** Distribution of the number of proteins per peptide.

1. Sample protein amounts $\vec{Q}^B = [Q_1^B, \dots, Q_m^B]$ at random from the collection of ITRAQ tag intensities.
2. Generate ratio R_j by sampling from a log-normal $N(0, \sigma_R)$ distribution. σ_R is set to 0.7 which is the estimated standard deviation of the log peptide ratios in the *Arabidopsis* data. Compute $Q_j^A = R_j Q_j^B$.
3. For each peptide s_i , generate d_i uniformly from $(0, 1]$, and compute $q_i^A = d_i \sum_{(p_j, s_i) \in E} Q_j^A$, $q_i^B = d_i \sum_{(p_j, s_i) \in E} Q_j^B$. When detectabilities are not incorporated, choose $d_i = 1$.
4. Compute peptide ratios r_i , and perturb according to a log-normal $N(0, \sigma)$, over a range of values σ . Denote σ as *perturbation level*.

We consider the *system* of constraints for each data-set.

Once the data is generated, only the peptide ratios r_i are used as inputs. The linear programs are solved for $\vec{Q}^{B'} = [Q_1^{B'}, \dots, Q_m^{B'}]$, {and $\vec{Q}^{A'} = [Q_1^{A'}, \dots, Q_m^{A'}]$ using ILOG OPL Development Studio 6.1 (source code (C#) is available at <http://www.cse.ucsd.edu/~bdost/downloads/PQPLinearProg.zip>.) The reliability of the estimates is tested using three measures.

Validation statistics: Recall that the value of the optimized objective is a weak indicator of the quality of results. For the simulations, as the protein abundances are known, we can compute the error in the estimate as the protein-abundance-distance (PAD).

$$\text{PAD}(\vec{Q}^{B'}, \vec{Q}^B) = \frac{\|\vec{R}^B\|_1}{m} \quad (3)$$

where $\vec{R}^B = \left[\ln \frac{Q_j^{B'}}{Q_j^B} \right]$. While any norm can be used as a valid measure of distance, the choice of the 1-norm, averaged over the dimensions can be loosely interpreted as average fold difference between actual and estimated protein abundances. The true protein abundances are not available for the *Arabidopsis* ITRAQ data, so we compute an indirect measure LRD, defined as

$$\text{LRD}(\vec{r}, \vec{r}') = \frac{\|\vec{r} - \vec{r}'\|_1}{n} \quad (4)$$

where $\vec{r} = [\ln(r_1), \dots, \ln(r_n)]$ is the vector of experimental peptide log-ratios for the n peptides, and $\vec{r}'$ describe the peptide log ratios computed by using the estimated protein abundances. Intuitively, if the protein abundance estimates are accurate, the computed peptide log-ratios should match the experimental log-ratios. In a similar way, we compute the peptide log detectability distance LDD as the average 1-norm of the logs of detectabilities. We use PAD, LRD, and LDD to test performance on simulations.

3.1. Results of simulation

As the reliability of the estimates depend upon rank of $\mathcal{A}$, we loosely group each of 100×257 simulated systems into three categories according to $\text{rank}(\mathcal{A})$ for a fixed rank-threshold t , as follows:

TABLE 1. SIMULATION CATEGORY I SYSTEMS GROUPED ACCORDING TO THEIR RANK-THRESHOLDS

	$\mathcal{R}(\mathcal{A})$				
	1	2	4	8	16
F_1 —Category I	1074 (4.2%)	2514 (9.8%)	3044 (11.8%)	3980 (15.5%)	4388 (17.1%)
F_2 —Category I	663 (2.6%)	2251 (8.8%)	2955 (11.5%)	3104 (12.1%)	4989 (19.4%)

Category I : $(\text{rank}_t(\mathcal{A}) = 2m \leq n + 1)$ (Over-determined, full-rank systems).

Category II: $(\text{rank}_t(\mathcal{A}) < 2m \leq n + 1)$ (Ill-conditioned systems).

Category III : $(\text{rank}_t(\mathcal{A}) \leq n + 1 < 2m)$ (under-determined systems).

At rank-threshold 1, we obtain 1074, 3926, and 20700 systems in Categories I, II, and III for the F_1 simulation, and similar distributions for F_2 . As the rank-threshold is increased, some of the Category II systems move into Category I (Table 1). Additionally, the performance of under-determined systems is uniformly worse than the other two (data not shown). Therefore, we will focus on Category I evaluation using different rank-thresholds. For each category, and each validation statistic, we compute *cumulative-probability* as the fraction of systems in that category that achieve a certain distance or lower. The ideal case is when the cumulative probability is 1 at distance 0.

In the absence of noise, we achieve the ideal case, zero PAD and LRD, for all Category I systems at rank-threshold 4. As noise is introduced to data and less stringent rank-thresholds is used, we deviate from the ideal case. Figure 4a shows the cumulative probability distribution against PAD, and LRD at perturbation level 0.01. Out of 1074 Category I systems at rank-threshold 1, 75% have PAD error of less than 0.16 and an LRD error of less than 0.01. The performance degrades for higher rank-thresholds. Note that in all cases, the optimized objective is very close to 0 ($\sim 10^{-4}$). However, in the ill-conditioned and under-determined systems, multiple solutions will lead to a low-error solution, and an arbitrarily picked solution will have high PAD and LRD error. The performance also degrades with an increase in perturbation error. Figure 4b plots the cumulative ratio for Category I systems at rank-threshold 1 under increasing perturbation levels. Thus, while 93% of systems show LRD of at most 0.1 at perturbation 0.01, the number falls to 55% at perturbation 0.15.

Ill-conditioned systems: We identified a number of Category II systems where the rank-threshold was poor, but only because a small number of singular values were close to 0. For example, in a simulation with perturbation level 0.05, we observed 339 systems for which fewer than three singular values were at most 10^{-16} , and $\text{rank}_1(\mathcal{A}) \geq 2m - 3$ (remaining s.v. $\geq 10^{-1}$). Our results show that the revised LP, suggested for ill-conditioned systems in Section 2, indeed provides better estimates of protein abundance values of these

FIG. 4. Simulation results using F_1 formulation. Cumulative probability of PAD and LRD for Category I systems (a) perturbation level 0.01, but different rank-thresholds (b) at rank-threshold 1, but different perturbation levels. In all cases, we measure the fraction of systems that were estimated within a certain distance.

FIG. 5. Improved estimates of protein abundance values using SVD-based projection on Category II (over-determined but ill-conditioned systems).

systems (Fig. 5). For example, over 88% of the systems from the revised LP achieve an LRD of 0.25 or better, compared to 65% from the original formulation.

Peptide detectabilities: A similar behavior is observed during estimation of peptide detectabilities (Fig. 6a, b). The performance of PAD, and LRD surprisingly does not change with the addition of an extra n unknowns (also, n new constraints get added). We also plot the performance of detectability estimates. As expected, the performance is acceptable for low rank-threshold systems and low perturbations, but degrades for higher rank-thresholds, and higher perturbation levels. The detectability estimates are robust as well and degrade in a similar manner.

3.2. *Arabidopsis thaliana* ITRAQ data

We focus on the 1190 non-trivial systems from the ITRAQ data comparing infected samples to non-infected ones. ITRAQ data is not appropriate to get peptide abundance values reliably, so we only use the F_1 formulation on this data-set. The distribution of systems in different rank categories is described in Figure 7a. A total of 99 systems fall into Category I with the most stringent rank-threshold 1, covering 219 proteins and 357 peptides. In addition to relative protein abundances, we estimate abundance ratios across samples for four proteins which do not have a unique peptide.

As actual protein abundances are not known, we use LRD to evaluate the estimates. The LRD statistic of different categories is as expected with this group performing better than the other groups. Within the 99 systems, 79 have LRD smaller than 10^{-1} , and 55 have LRD smaller than 10^{-4} . The list of systems, constituent peptides, and protein abundance values are available at www.liebertonline.com/cmb as Supplementary Material. Here, we cherry-pick a few representative examples that point to the differential expression of individual sequences in response to the infection.

FIG. 6. Simulation results using F_2 formulation. Cumulative probability of PAD, LRD, and LDD for Category I systems (a) at different rank-thresholds (b) at rank-threshold 1, but different perturbation levels. In all cases, we measure the fraction of systems that were estimated within a certain distance.

FIG. 7. *Arabidopsis* root-knot nematode infection ITRAQ data. **(a)** Number of systems in Category I and II at different rank-thresholds. **(b)** Empirical cumulative probability distribution of LRD for Category I systems at different rank-thresholds.

Ca²⁺ ATPases: One of the systems is encoded by three proteins from the P-type Ca²⁺ ATPase superfamily involved in Ca²⁺ transport. The three members are the plasma-membrane bound AT5G57110 (ATPase 8), AT4G2990 (ATPase 10), and AT3G21189 (ATPase 9). Earlier reports have suggested that ATPase8,10 are co-expressed evenly over all vegetative tissues, while ATPase 9 is expressed almost exclusively in pollen (Marmagne et al., 2004). In our data, the three form a connected component with six peptides, and our analysis showed the relative wild-type expression of the three to be 34.3%, 57.5%, and 8.22%, respectively, confirming this observation. Further, we find that ATPase 8 is 3× over-expressed in the infected state.

Profilins: The Profilin family encodes proteins that regulate actin cytoskeleton formation. Five profilins are known. The two that are identified in our data (PRF-1,2) are constitutively expressed in all vegetative organs, and a regulatory element in their first intron is suspected to mediate this expression, differentiating them from the other Profilins (Jeong et al., 2006). In our data-set, the two are in a component with three peptides, one shared. Our analysis shows that Profilin-2 has only (13%) of the total abundance and is further reduced twofold upon infection.

Cinnamyl-alcohol dehydrogenases: A number of genes in *Arabidopsis* have been annotated as part of the CAD family, an assertion which has subsequently been challenged, pointing instead to the central role of two members (AtCAD4 and AtCAD5) in the CAD metabolic network. These two molecules have expression patterns consistent with lignification at stem tissues. Interestingly, expression was also observed in various non-lignifying zones (e.g., root caps) indicative of a possible role in plant defense (Kim et al., 2007). Our results have a single connected component with AtCAD4, 5, and 3 peptides. The analysis shows that both proteins are equally abundant, with AtCAD5 (AT4G34230) at 56% in non-infected cells. However, AtCAD5 is significantly (1.5×) over-expressed, while AtCAD4 (AT3G19450) is 2× under-expressed.

Phosphoglycerate kinases: Phosphoglycerate kinases have been previously shown to be differentially expressed during defense response of *Arabidopsis* (Jones et al., 2006). In our data, four proteins from this family are in a component with nine peptides as shown in Figure 8. In this example, along with the relative protein abundances, we also compute the abundance ratio across-sample for IPI00530695 even though it does not have a unique peptide. Our analysis suggests that both IPI00538665 and IPI00535490 are 1.5× over-expressed, but IPI00535490 is much more abundant in both samples. IPI00530695 is 3× under-expressed while the abundance of IPI00534991 does not change.

FIG. 8. PhosphoGlycerate kinase family members sharing peptides.

4. DISCUSSION

The extent of peptide sharing in proteomics is under-estimated or controlled by non-redundant databases that constrain identification of splice variants. Consequently, shared peptides and proteins with non-unique peptides are often discarded, corresponding to as much as 50% of the data, in our experience. As mass spectrometry based protein quantification becomes routine, shared peptide analysis will be increasingly important. Our results are the first to show that a careful analysis not only helps in recovering variations in abundance of the *same* proteins, but also helps quantify the relative levels of *different* proteins within each condition. These across-protein relative abundance computations can help elucidate these differential regulation of the proteins from a family. We investigate topological and numerical considerations in estimating reliability of our computations. We show that we can solve over-determined systems more confidently and measure the quality of estimates using a validation statistic. For the other systems, the estimates of which we cannot trust (i.e., under-determined systems or over-determined systems with low-quality estimates), shared peptides can still be useful to infer existence/non-existence of proteins.

The protein quantification problem considered here is a generalization of the ‘simpler’ protein identification problem. As we quantify different protein amounts relative to other proteins in a connected component, we are also making an assertion on the presence, or absence of specific proteins. However, there is one difference: when the system is under-determined, we cannot compute relative protein abundances, but we can still decide if a specific protein is expressed. Consider an under-determined system with two proteins p_1 and p_2 and two constituent peptides s_1 , and s_2 , where s_1 is unique and s_2 is shared. This system has infinite solutions with zero error and we cannot solve for relative abundances of p_1 and p_2 . The existence question is different. If p_2 is not expressed, only p_1 will contribute to s_1 and s_2 , and r_1 and r_2 are identical. A substantial difference between r_1 and r_2 implies that p_2 is indeed expressed (Fig. 9a). The prevalent parsimonious approaches often conclude that there is no evidence of p_2 expression in the absence of a unique peptide, leading to possible errors in downstream analysis.

A similar analysis also works for over-determined systems that give low-quality estimates. Consider the over-determined system shown in Figure 9b with two proteins and five peptides. We would expect the symmetric peptides to have the same ratios. If the observed ratios do not agree with the system configuration as in Figure 9b, this might be an indication of one or more proteins that do not exist in the protein databases. In the example, we expect $r_1 \sim r_2$ and $r_4 \sim r_5$, but they are substantially different. Assuming the difference is larger than expected by experimental variation of peptide abundance measurements, this observation clearly suggests the presence of a third protein p that shares s_2 and s_4 .

We agree, however, that existing experimental abundance measurements must show lower variance to make such inferences robust (Bantscheff et al., 2007). As mass spectrometers become more accurate, experimental variation will decrease and peptide quantification will improve and will directly increase the power of our methods. In a similar fashion, the estimation of peptide detectabilities is at an early stage of development (Alves et al., 2007; Mallick et al., 2007). Our results attest to the viability of using shared peptides for detectability computation, but also point to the importance of detectability values in extending the scope of shared peptide analysis. The model’s ability to automatically estimate peptide detectabilities may result in an ongoing cycle of self-refinement where different systems resulting from different experimental conditions may allow one to continuously expand the set of known detectabilities, which in turn would allow for the resolution of more complicated systems. In fact, we note that this progressive convergence towards an extensive database of peptide detectabilities may even allow one to learn more about

FIG. 9. (a) Protein identification in an under-determined system; p_2 must be present. (b) Protein prediction in an over-determined system; a protein p that shares peptides s_2 and s_4 should be present.

systems that were previously not solvable in a given experiment by adding information from different or even additional targeted experiments aimed at estimating the necessary detectabilities.

We use a linear programming formulation for optimization of relative protein abundances. Clearly, different algorithms can be used to optimize the error in estimation, including non-linear optimization and other machine learning approaches. We have experimented using a simulated annealing approach with a non-linear cost function that minimizes the absolute sum of differences between observed and expected peptide ratios. While such an approach is more time consuming, it provides better estimates for systems with unbalanced protein abundances, as linear programming formulation is biased towards the error terms associated with the more abundant proteins. Details of that study will be discussed elsewhere.

In this article, we describe a systematic simulation-based framework to compare and develop improved methods for shared peptide analysis, and we introduce novel evaluation methods for the estimations.

ACKNOWLEDGMENTS

The research was supported by the National Center for Research Resources of the National Institutes of Health (grant P-41-RR24851).

DISCLOSURE STATEMENT

No competing financial interest exist.

REFERENCES

- Alves, P., Arnold, R.J., Novotny, M.V., et al. 2007. Advancement in protein inference from shotgun proteomics using peptide detectability. *Pac. Symp. Biocomput.* 409–420.
- Bantscheff, M., Schirle, M., Sweetman, G., et al. 2007. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 389 1017–1031.
- Sasser, J.N. and Carter, C.C. 1985. *An Advanced Treatise on Meloidogyne: Volume I*. North Carolina State University Graphics.
- Jammes, F., Lecomte, P., de Almeida-Engler, J., et al. 2005. Genome-wide expression profiling of the host response to root-knot nematode infection in *Arabidopsis*. *Plant J.* 44, 447–458.
- Jeong, Y.M., Mun, J.H., Lee, I., et al. 2006. Distinct roles of the first introns on the expression of *Arabidopsis* profilin gene family members. *Plant Physiol.* 140, 196–209.
- Jones, A.M., Bennett, M.H. Mansfield, J.W. et al. Analysis of the defence phosphoproteome of *Arabidopsis thaliana* using differential mass tagging. *Proteomics* 6, 4155–4165.
- Kim, S.J., Kim, K.W., Cho, M.H., et al. 2007. Expression of cinnamyl alcohol dehydrogenases and their putative homologues during *Arabidopsis thaliana* growth and development: lessons for database annotations? *Phytochemistry* 68, 1957–1974.
- Mallick, P., Schirle, M., Chen, S.S., et al. 2007. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* 25, 125–131.
- Marmagne, A., Rouet, M.A., Ferro, M., et al. 2004. Identification of new intrinsic proteins in *Arabidopsis* plasma membrane proteome. *Mol. Cell Proteomics* 3, 675–691.
- Wiese, S., Reidegeld, K. A., Meyer, H. E., et al. 2007. Protein labeling by ITRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* 7, 340–350.
- Xue, X.F., Wu, S.F., Zhu, Y.P., et al. 2007. Improving label-free protein quantification methods using expectation maximization-like algorithm in shotgun proteomics. *Chinese J. Anal. Chem.* 35, 19–24. (In Chinese.)

Address correspondence to:

Dr. Banu Dost
Department of CSE
UC San Diego
San Diego, CA

E-mail: bdost@ucsd.edu