Published in final edited form as:

J Proteome Res. 2013 September 6; 12(9): 4111–4121. doi:10.1021/pr400438q.

IDPQuantify: Combining Precursor Intensity with Spectral Counts for Protein and Peptide Quantification

Yao-Yi Chen¹, Matthew C. Chambers¹, Ming Li², Amy-Joan L. Ham³, Jeffrey L. Turner⁴, Bing Zhang¹, and David L. Tabb^{1,*}

¹Department of Biomedical Informatics, Vanderbilt University Medical School, Nashville, TN 37232-8575

²Division of Cancer Biostatistics, Vanderbilt University Medical School, Nashville, TN 37232-6848

³Department of Pharmaceutical, Social and Administrative Sciences, Belmont University College of Pharmacy, Nashville, TN 37212-3757

⁴Protein Technologies and Assays, Sigma-Aldrich, 2909 Laclede Ave., Saint Louis, MO, 63103

Abstract

Differentiating and quantifying protein differences in complex samples produces significant challenges in sensitivity and specificity. Label-free quantification can draw from two different information sources: precursor intensities and spectral counts. Intensities are accurate for calculating protein relative abundance, but values are often missing due to peptides that are identified sporadically. Spectral counting can reliably reproduce difference lists, but differentiating peptides or quantifying all but the most concentrated protein changes is usually beyond its abilities. Here we developed new software, IDPQuantify, to align multiple replicates using principal component analysis, extract accurate precursor intensities from MS data, and combine intensities with spectral counts for significant gains in differentiation and quantification. We have applied IDPQuantify to three comparative proteomic datasets featuring gold standard protein differences spiked in complicated backgrounds. The software is able to associate peptides with peaks that are otherwise left unidentified to increase the efficiency of protein quantification, especially for low-abundance proteins. By combining intensities with spectral counts from IDPicker, it gains an average of 30% more true positive differences among top differential proteins. IDPQuantify quantifies protein relative abundance accurately in these test datasets to produce good correlations between known and measured concentrations.

Keywords

precursor ion intensity; principal component analysis; retention time mapping; protein differentiation; comparative proteomics; quantitative proteomics; spectral counting; CPTAC

Introduction

The differentiation and quantification of proteins between two or more biological samples is among the most important but challenging tasks in proteomics. Label-free shotgun proteomics has been widely used for identification and quantification of proteins in large studies due to its simple workflows and the possibility of comparison among multiple

*tel, (615) 936-0380; fax, (615) 343-8372; david.l.tabb@vanderbilt.edu.

Supporting Information Available: This material is available free of charge via the Internet at <http://pubs.acs.org>.

states^{1–5}. The two major strategies in label-free quantification either measure precursor ion intensities from mass spectra or count the tandem mass spectra associated with each protein. In the first strategy, the chromatographic peak height or area for each peptide can be compared to that in other experiments to yield relative quantitative information. The second strategy compares the number of spectra observed for proteins between two sets of experiments.

Intensity values can provide accurate protein relative quantification with a high resolution mass spectrometer and optimized liquid chromatography (LC)^{6–9}. In shotgun proteomics, proteins are first digested to peptides, which are then separated by chromatography and emitted by electrospray ionization. Peptides are fragmented when eluting from the LC column, and the peptides are identified from fragments measured in the resulting MS/MS spectra. In data-dependent acquisition, depending on the instrument's MS/MS sampling speed and the sample complexity, the mass spectrometer may not be able to sample new peptide ions as fast as they elute from the LC column. Eluting peptides could be within the detection limit but not selected for fragmentation due to more intense peptides.

Only 70% of identified peptides are shared between technical replicates¹⁰. A peptide might correspond to an MS/MS in one replicate but not in another replicate even though the peptide was present in both. Aligning retention times among sets of files enables one to map these sporadic identifications to precursor ion chromatograms in experiments that lack these identifications. Previous studies have shown several methods for label-free detection of peak intensity differences between samples with or without identification data^{5, 11–14}. However, the majority of differentiating peaks or regions do not map to identified peptides. Some of these methods are described only for pairwise alignment and do not easily generalize to alignment between more than two files. CRAWAD⁵ bins data into narrow *m/z* and time regions from which selected ion chromatograms are extracted. The advantage of CRAWAD over other peak alignment algorithms like Pepper and msInspect is that it is also applicable to lower resolution MS data. However, peak misalignments could result in associating intensity from unrelated ions, so false discovery rate estimation is needed to characterize the accuracy of this alignment. Chromatographic retention time reproducibility is one of the greatest challenges in measuring peptide-specific differences. Slight variations in flow rate, gradient slope, experimental temperature and column conditions contribute elution time shifts among repeated LC separations.

Spectral counting-based quantification has proved more reproducible and has a larger dynamic range than intensity-based quantification^{10, 15–17}. However, the accuracy of spectral counting-based quantification is affected by differential spectrum count response for every peptide and by saturation effects. Moreover, spectral counting-based quantification does not make full use of peak attributes such as height, area or volume. To make use of the complementary attributes of spectral counting and intensity-based quantification, previous studies have suggested several strategies for combining the two methods for protein quantification, such as normalized spectral index¹⁸ and ProPCA¹⁹.

Here, we propose a regularized iterative multiple correspondence analysis model²⁰ for mapping peptide retention time across multiple replicates based on sets of shared peptides. This retention time model enables extraction of chromatographic peak information across the whole set of replicates via QuaMeter^{5, 21–23}. A simulated “decoy” peptide technique enables false discovery rate estimation for the retrieved intensities. The broader pool of peak intensities improves discrimination of proteomic differences. We propose a new way to combine intensity and spectra counts for better protein differentiation using meta-analysis^{24, 25}. We will show that the new method resists errors inherent to each technique in candidate biomarker selection.

Materials and Methods

Data Processing Steps

MS/MS scans were converted to mzML by the msconvert tool of ProteoWizard²³. All protein databases contained both forward and reverse sequences for estimating protein and peptide identification errors. Peptides were identified with MyriMatch (version 1.6.79)²². MyriMatch applied a precursor tolerance of 10 ppm for Orbitrap data and configured to use a static mass shift of 57.0215 Da for alkylated cysteines and allow the variable modification of oxidation of methionine (+15.9949 Da) and formation of N-terminal pyroglutamate (−17.0265 Da). The search results were processed from pepXML format by IDPicker (version 3.0.506)²⁵ to yield a 1–2% peptide false discovery rate (FDR). Peptides passing these thresholds were considered legitimate. Proteins were assembled by IDPicker 3.0.506 from peptides with parsimony rules^{25, 26}. Proteins were filtered at the criteria described above to ensure less than 5% protein FDR. Significant differences in protein spectral counts between different groups were calculated by QuasiTel²⁴. Differences in protein intensity between groups were calculated by Student's t-test, reflecting that spectral counts are discrete while intensities are continuous. The resulting p-values were adjusted by Benjamini-Hochberg multiple testing correction. The differential proteins were compared with the gold standards to estimate the performance of the methods. The Myrimatch and idpQuantify configuration files appear in Supplementary material 3.

Data Sources

CPTAC study data—We used two datasets created by the Clinical Proteomic Technology Assessment for Cancer (CPTAC) program¹⁵: Study 5 and Study 6. In CPTAC Study 5, protein extract of a yeast lysate was analyzed by five different institutions. For each yeast lysate, six replicates were analyzed under Standard Operating Procedure control. In CPTAC Study 6, a yeast lysate was spiked with a mixture of 48 human proteins (Sigma-Aldrich UPS1) at five levels, each three-fold higher than the last. Each sample was analyzed in triplicate on seven independent instruments of four models (Thermo Fisher LTQ, LTQ-XL, LTQ-XL-Orbitrap, and LTQ-Orbitrap). We used only the Orbitrap data from the CPTAC dataset. Groups A, B, C, D, and E were yeast spiked with UPS-1 at 0.25, 0.74, 2.2, 6.7, and 20 fmol/μl respectively. Group Y included the yeast reference proteome without additional spikes. Data were processed using a FASTA database combining Yeast UniProt database and UPS proteins and common contaminant proteins, containing 6910 entries in total. Algorithm details were provided in the “Data Processing Step” section and supplementary material 3. The empirical protein FDR is 2.8% in OrbiP@65, 0% in OrbiW@56, 1% in OrbiO@65, 0.7% in Orbi@86. Peptides were filtered at maximum Q value of 2%. Proteins were filtered at 2 minimum distinct peptides, 2 minimum spectra per protein and further filtered at an average of 2 minimum distinct matches (including matches found by retention time mapping) per replicate before quantitative analysis. Best peaks for peptides were picked at peak width tolerance of 3 standard deviations between peak time and peak retention time, with average peak picking at an FDR of 5%. Peptide intensities were normalized by the total ion current of each run. Numbers of spectra, peptides and protein groups are shown in supplementary material 3.

NRD-Pfu data—In this study, a *Pyrococcus furiosus* (Pfu) lysate (Agilent Complex Proteomics Standard) was spiked with two different mixtures of 36 human Sigma UPS proteins. Each mixture included six cassettes of proteins, each containing six proteins. The concentration ratios for each cassette produced a range of narrow-range, defined (NRD) ratios (supplementary material 2). Cassettes of proteins were spiked in at 1:1 (30 pmoles : 30 pmoles), 4:3 (34.3 pmoles : 25.7 pmoles), 3:5 (22.5 pmoles : 37.5 pmoles), 2:1 (40 pmoles : 20 pmoles), 4:1 (48 pmoles : 12 pmoles), and 1:8 ratios (6.7 pmoles : 53.3 pmoles) in

Sample A and Sample B respectively. Two different instruments analyzed these samples. An Orbitrap captured spectra for six 95-minute LC gradients of each sample, and an Orbitrap Velos produced five technical replicates, using the same gradient. Data were processed using a FASTA database combining Pfu Uniprot database (08/20/2012), 36 UPS proteins and 71 common contaminant proteins, containing 2152 entries in total, and PSMs were filtered at a maximum Q value of 2%. Proteins were filtered at 2 minimum distinct peptides, 2 minimum spectra per protein and further filtered at an average of 2 minimum distinct matches (including matches found by retention time mapping) per replicate before quantitative analysis. The empirical protein FDR is 2.65%. Best peaks for peptides were picked at peak width tolerance of 3 with average peak picking FDR of 5%. Peptide intensities were normalized by the total ion current of each run.

UPS2 data—The dataset created by Ivanov et al.¹⁹ enabled comparison of IDPQuantify to the ProPCA technique introduced in their publication. The UPS2 standard set contains 48 human proteins with a dynamic range spanning 0.5–50,000 fmol. There were 38 LC MS/MS runs with each run containing one of 11 specified amounts of the UPS2 standard spanning over 2 orders of magnitude (Supplementary Table 2). Data were processed using a FASTA database combining 50 UPS2 proteins and 72 common contaminants. Peptides were filtered at a maximum Q value of 1% to forestall false protein identifications from this small protein database. Proteins were filtered at 2 minimum distinct peptides, 3 minimum spectra per protein. The empirical protein FDR is 0%. Best peak for each peptide were picked at peak width tolerance of 10 (since it is a simple protein mixture) with average peak picking FDR of 5%.

IDPQuantify—The workflow is shown in Figure 1. First, the MS/MS scan times for distinct matches of each identified peptide were extracted from mzML files by ProteoWizard. The scan time was then recorded in the database (idpDB) file generated by IDPicker. To increase the precursor intensity extraction sensitivity, a full union set of distinct peptide matches were combined from all included files. It used confidently identified distinct peptides within replicates as registration marks. All peptides for each run are compared to other retention times at which they were observed in other replicates with by regularized iterative multiple correspondence analysis (MCA)²⁷. MCA first calculates an indicator matrix of peptides and observed retention times, substituting initial values for missing retention times. Then it performs correspondence analysis to represent each experiment in a peptide and retention time coordinate space. This “map” is then adjusted to find the linear fit for each experiment to reduce the reconstruction error in observed retention times. This model can then be used to predict retention times for peptides that were not identified for a given experiment.

In this way, the alignment does not rely on picking a good reference run in advance to do retention time mapping. After peptide retention time mapping, a list of “decoy” peptides was generated with retention time and precursor m/z ratio randomized from two gamma distributions simulating the distribution of m/z and RT of identified peptides. The parameters of gamma distribution were estimated by the method of moments²⁸. The distribution of precursor m/z ratio and the corresponding gamma distributions were shown in Supplementary Figure 1. These “decoy” peptides were selected only if they had a precursor m/z ratio outside the range of m/z tolerance (10ppm in Orbitrap datasets) from any existing peptides with a randomized retention time. Sensitivity of this retention time mapping method was estimated by leaving out a subset of retention times for identified peptides. Iterative MCA estimated the missing scan times, and chromatogram extraction and peak picking followed. The software was able to recover 97–98% of these missing intensities through this retention time mapping technique. The error in retention time mapping was calculated by leaving out one of the 15 runs in CPTAC study 6 dataset and

estimating the retention time by MCA (Supplementary Table 1). As number of missing runs increases, the mean and median absolute biases for retention time mapping stayed around 10 seconds.

The list of all peptide *m/z* ratios and retention times was passed to QuaMeter²¹ to extract the chromatogram of each peptide within a tolerance window of retention time and *m/z* ratio. The best peak for each peptide was picked by CRAWDAD functions within ProteoWizard, and peak intensity and signal-to-noise ratio information was retrieved. If a best CRAWDAD peak is found for a decoy peptide, it was considered as a false hit and used to estimate the false discovery rate in the alignment process ($FDR = 2 * \text{number of detected decoy peptides} / \text{total number of detected peptides}$ for a given threshold of recovered peak retention time difference). The peak width tolerance parameter was used to limit the distance between the best CRAWDAD peak retention time and the aligned retention time to control the false discovery rate at 0.05. For example, if peak width tolerance was set at 4, the distance from the aligned retention time and the peak retention time is less than or equal to 4 times the standard deviation of the peak fitted by Gaussian distribution.

Differences in precursor charge and post-translational modification may lead to distinct matches for a given peptide. The intensities of distinct matches of peptides were normalized by total ion current (TIC) and summed up as the intensities of the corresponding peptides. The intensities of peptides were then summed up as the intensities of the corresponding proteins. Higher spectral counts are associated with higher protein or peptide abundance. By summing precursor intensities, proteins with more peptides or peptides with more distinct matches accumulate higher intensities, enabling improved discrimination. Proteins were filtered to only those with more than two distinct matches with intensities found per replicate. Internal comparisons showed that summing peptide intensities produced more robust estimation than averaging the observed intensities for each peptide. The more abundant a protein is in a sample, the more likely its peptides are to be identified, and the greater the intensity that can be extracted. The missing protein intensities were imputed as 0.

The protein intensities were compared between two groups by t-test. Log intensities employed 'e' as a base (ln). The protein spectral counting changes were tested by QuasiTel. The protein intensity ratio was calculated by the ratio of the average of the intensities across multiple replicates. When combining both intensity and spectral counting for protein differentiation, p-values from intensity and spectral counting were combined using Fisher's combined probability test²⁹:

$$\chi^2 = -2[\log_e(p\text{ value}_{intensity}) + \log_e(p\text{ value}_{spectral\ count})] \sim \chi^2 \text{ distribution with } 2k \text{ degrees of freedom}$$

Note that because spectral counts and intensities are correlated (Supplementary Figure 2), p-values resulting from this combination will be lower than under a model where the correlation structure is taken into account. The rankings by these metrics rather than the p-value themselves, however, were the basis upon which the algorithm results were evaluated.

Results

Peptide retention time mapping is an effective way to incorporate information shared between replicates. Figure 2 illustrates six technical replicates for a yeast lysate sample that were analyzed at three different institutions. Figure 2a shows that only 69% to 74% of distinct peptide identifications were identified in more than one replicate and only 21% to 33% of the peptides were found in all six replicates. Assuming zero intensity for a peptide simply because it was not identified in an experiment would lose valuable information.

Previous studies have shown that repeatability is positively correlated with precursor intensity¹⁵. Figure 2b confirms that finding via QuaMeter-extracted intensities; peptides identified in fewer replicates have lower intensities than those identified in more replicates.

Peptide retention time (RT) mapping and peak alignment by IDPQuantify reduced apparent differences among peak intensities. Figure 3 shows the correlation of the intensities of distinct peptide matches between replicates. In CPTAC study 6, Groups A, B, C, D, and E were yeast lysate spiked with UPS-1 human proteins at 0.25, 0.74, 2.2, 6.7, and 20 fmol/ μ l (1:3:9:27). Group Y was the reference yeast lysate. Theoretically, the correlation of UPS peptide intensities between replicates of the same group should be high, but the UPS peptides were present at low concentrations in samples A and B and were only sporadically identified. However, the left column of Figure 3 shows that missing values for peptide intensities led to diminished correlation between replicates. After retention time mapping (right column), the average correlation increased by 20%. In Group C, D, and E, because the abundance of spike-in proteins is high, the correlation values between replicates were close to 1 with or without RT mapping (Supplementary Figure 3). We also looked at changes in numbers of distinct peptides from UPS1 proteins by RT mapping across all 18 replicates (3 replicates every group, 6 groups- Y, A, B, C, D, E). The number of distinct peptide matches with extracted intensity increased considerably in A and B replicate groups but much less in the reference group Y (Supplementary Figure 4) since some peptides found in Y were shared between UPS and yeast proteins. The increase in total intensities of UPS peptides by retention time mapping was shown in supplementary Figure 5. These data show that IDPQuantify efficiently aligned peptides across multiple replicates with a controlled false discovery rate.

IDPQuantify efficiently extracted peptide intensities. High absence rates for peak intensities have been a problem in precursor intensity extraction. Previous studies have shown that as many as half of expected positions may not match to peaks in proteomics data³⁰. These absences can seriously impair the function of quantitative proteomic software. An “identification rate” can summarize the fraction of peptides matched to extracted peaks across a set of samples; if 1000 peptides are identified in aggregate for a data set that contains five replicates, the identification rate is the fraction of these 5000 potential occurrences that correspond to identified tandem mass spectra. Table 1 reports this rate for the data sets included in this analysis. The “matching rate” reflects the fraction of these associations that can be made when retention time mapping recovers additional peaks for these peptides. On average, the identification rate for these data sets was 43.7%, and retention time mapping boosts the matching rate to 82.3%. This rate of increase was greater than seen in MSInspect/AMT on the same UPS2 dataset¹⁹, where the identification rate was 43% and the matching rate was 52%.

Peptide retention time mapping by IDPQuantify greatly increased sensitivity and specificity, especially for proteins with low abundance. To test the performance of protein differentiation by peptide precursor intensity before and after RT mapping, we recorded the number of true positives in the top 50 differential proteins in the CPTAC study 6 dataset (48 proteins are true differences between these samples). The UPS1 proteins comprised an “answer key” in four pairwise comparisons: A vs. B, B vs. C, C vs. D, and D vs. E. We ranked the proteins by the p-values calculated by intensities only, by spectral counts only, or by Fisher’s method combining both techniques. We then counted the number of UPS1 proteins in the top 50 differential proteins with or without RT mapping. Figure 4a shows that in pairs of groups with relative low spike-in abundance (A vs. B and B vs. C) the true positive rate was highest for p-value combination in six of eight trials, and retention time mapping was essential to good performance for intensity-based methods. In pairs of groups with high spike-in abundance (C vs. D and D vs. E) retention time mapping was less

significant in effect, and the combined p-value method produced the best discrimination in seven of eight trials. In these high spike-in groups, most of the UPS1 peptides were already identified in every file, so RT mapping could not add much new information. In real biological and medical research, the proteins of interest are likely to be low abundance. Thus it is important to extract intensity through retention time mapping. ROC analysis and area under the curve (AUC) values confirmed these findings (Supplementary Figure 6). Retention time mapping increased AUC in 10 out of 12 cases in A vs. B, B vs. C and C vs. D comparisons by up to 12.8%. P-value combination produced the highest AUC in almost all cases. It increased the AUC by 3%–32% in separate analyses of spectral counting or intensity. A similar analysis was performed in the NRD-Pfu dataset (see Supplementary Figure 7).

IDPQuantify estimates protein relative abundance with high accuracy. Figure 5a is a boxplot of log intensity ratio of UPS1 proteins across 4 pairs with a fold change of 3 (the black horizontal line is $\log(3)$, the true ratio separating these samples). Each box indicates a comparison pair with 3 replicates in each group in one institute ($(A_1+A_2+A_3)/(B_1+B_2+B_3)$). The higher the concentration of UPS proteins, the narrower the distribution of estimated intensity ratios and the more accurate the quantification. In A vs. B, where the concentration was low, the ratio is widely distributed around the true value. In B vs. C, the ratio is closer to the true value with a narrower distribution and the median is very close to $\log(3)$ by all four institutes. In C vs. D, the ratio had a narrower distribution still. However, it is overestimated in Orbi@86. In D vs. E, the intensity ratios are narrowly distributed around the true value. Figure 5b shows the log intensity ratio of UPS proteins between Groups A and B in the NRD-pfu dataset. Each box indicates replicate k comparing A and B (A_k/B_k ; where k can be 1 to 6 in sextuplets, 1 to 5 in quintuplets). Cassette 1–6 were spiked in A:B at a ratio of 1, 1.33, 0.6, 2, 4, 0.13 respectively. The horizontal lines indicate the spike-in ratios of the estimated ratios with corresponding color. In quintuplets, the intensity ratios were accurately estimated by IDPQuantify. In sextuplets, the intensity values were overestimated in cassette 1–5 and were distributed around the true value in cassette 6. However, we see a clear trend of intensity ratio changes across the 6 cassettes. The correlation between the average estimated intensity ratio and the spike-in ratio is 0.9672 in Sextuplets, and 0.9536 in Quintuplets. Therefore, the relative protein intensities estimated by IDPQuantify are highly accurate.

IDPQuantify also quantifies peptide relative abundance with high accuracy. Figure 6 is a density plot of log UPS peptide intensity ratio, the true intensity ratio is indicated by the vertical line in the middle. The red curve shows Group A vs. B, where the spike-in abundance is low, and thus the distribution of peptide intensity ratio is divergent but centered on the true ratio in all four institutes. As the protein abundance increases, the distribution of peptide intensity ratio converges even more to the true ratio. Especially in Group D vs. E with the highest spike-in ratio, the peptide intensity ratio is very close to the true ratio and very narrowly distributed. Note that in Orbi@86, C vs. D is less convergent than B vs C and has a heavy tail above the true ratio; this observation is consistent with the protein ratio overestimation in Figure 5a.

IDPQuantify compares favorably to other quantification software-msInspect¹¹, MaxQuant⁹ and SING¹⁸. In the UPS2 dataset, we compared against two software pipelines: msInspect/AMT¹¹ followed by ProALT¹⁴ and ProPCA. For a fair comparison we used the same data set used by the papers that evaluated these pipelines¹⁹, which analyzed data from a mixture of UPS2 proteins with a dynamic range spanning 0.5–50,000 fmol. According to the paper, intensities should not be normalized due to differences in total protein abundance. We calculated the correlation between the logarithm of spiked-in protein abundance and intensity by IDPQuantify or spectral counts, then compared to the correlations reported by

the paper¹⁹. The authors extracted peptide intensities by msInspect/AMT¹¹, then used ProALT, an intensity-based method described by Jaffe et al.¹⁴ to do peptide-protein roll-up. They also used ProPCA, a multivariate analysis method to combine spectral counting and intensities. In Figure 7, the correlations of the real abundance and the relative abundance estimated by the four methods for individual UPS2 proteins were compared. We can see that 5 more UPS2 proteins were identified by our software. The abundances of five proteins were found only by our identification pipeline: NEDD8 (Q15843), SUMO-conjugating enzyme UBC9 (P63279), Creatine kinase M-type (P06732) and Lysozyme C (P61626), Microtubule-associated protein tau (P10636-8). They produced correlations of 0.9905, 0.9662, 0.9567, 0.7798 and 0.7712, respectively. IDPicker did not identify Ubiquitin-conjugating enzyme E2 (O00762) in the mixture. The correlation of spectral count with true concentrations was always lower than IDPQuantify or ProPCA, but comparable to ProALT. The correlation of IDPQuantify with true concentrations performed much better than ProALT in almost all cases, with an average improvement of 12.7%. In the first 8 proteins with highest abundance (50000 fmol), they both performed well, producing high correlations. In the second 8 proteins with abundance of 5000 fmol, ProPCA performed slightly better than IDPQuantify by an average of 6%. In proteins with abundance of 500 fmol, IDPQuantify performed slightly better than ProPCA by an average of 3%. The relationship of log spike-in concentration and log intensity of a representative protein of each group (50000 fmol, 5000 fmol, 500 fmol) was shown in upper panel of Figure 7. In all three groups, the estimated intensities are highly correlated with the real spike-in intensities. The correlations details and plots of all the UPS2 proteins were shown in Supplementary Table 2 and Supplementary Figure 8 in supplementary material.

In the CPTAC study 6 dataset, we compared the results from IDPQuantify, MaxQuant and SING. We used the O65 dataset which was tested by all the software comparatively in a previous study³¹. SING is software to implement normalized spectral index quantification. We compared the idpQuantify result with results of MaxQuant and SING from this paper³¹. Figure 8 showed the distribution of the log fold changes of the abundance of UPS proteins estimated by three methods in E/B comparison with 27 fold change, E/C, D/B comparisons with 9 fold change, and E/D, D/C, C/B comparisons with 3 fold change. On average, IDPQuantify estimated the fold change with higher accuracy than the other methods. MaxQuant overestimated the fold change in almost all cases. SING had high standard error in estimation especially in E/B, D/B and C/B since in Group B the protein concentration too low to produce good estimation with spectral counting method. The missing rate is lower in IDPQuantify than the other methods.

IDPQuantify outperforms ProALT (msInspect/AMT), MaxQuant, an alternative intensity-based quantification method and SING, a quantification method incorporating spectral counts and intensities. It performed comparably to the intensity-spectral counting combination method by ProPCA in a simple dataset. Since IDPQuantify intensities are more accurate, combining spectral counting and IDPQuantify intensities has potential for greatly improved protein and peptide quantification.

Conclusion

Protein differentiation and quantification is an important issue in shotgun proteomics. IDPQuantify, a new method for retention time mapping, peak alignment, peptide-protein roll-up, and combination of spectral counts with intensities, has been shown to substantially improve performance on these important tasks. IDPQuantify is also able to calculate relative protein abundance with high accuracy in datasets with gold standard proteins in a complex background, which are similar to real biological experiments. These results prove that it has

great potential to confidently determine expression differences in biological samples and biomarker studies.

This is the first application of the MCA method for peptide retention time mapping. This alignment of retention times does not require choosing a reference experiment, yielding quality alignments without the subjectivity of picking a best run in advance. The precursor intensity detection by CRAWDAD is more robust to RT run-to-run shifts. The best peaks picked by QuaMeter are assisted by but not dependent upon MS/MS acquisitions. Moreover, IDPQuantify allows the users to search for a customized list of peptides/analytes with expected or known retention time, m/z ratio and charge. It can also visualize chromatograms with both identified peptides and aligned analytes and customized analytes of interests. We proposed a new algorithm for estimating the false discovery rate of peak picking. Users can improve the sensitivity of peak picking by controlling false discovery rate. Although IDQuantify is best used with mass spectra collected in FTMS, it can be adapted to a variety of instrument types. Mass accuracy is important in the performance of CrawlDad chromatogram extraction; therefore, high-resolution mass data is required to reduce false alignment. IDPQuantify synergistically uses both spectral counting and intensity data and improves protein differentiation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by U01 CA152647 from the National Cancer Institute.

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003; 422(6928):198–207. [PubMed: 12634793]
2. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science*. 2006; 312(5771):212–7. [PubMed: 16614208]
3. Bantscheff M, Kuster B. Quantitative mass spectrometry in proteomics. *Anal Bioanal Chem*. 2012; 404(4):937–8. [PubMed: 22825679]
4. Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*. 2012; 404(4):939–65. [PubMed: 22772140]
5. Finney GL, Blackler AR, Hoopmann MR, Canterbury JD, Wu CC, MacCoss MJ. Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution muLC-MS data. *Anal Chem*. 2008; 80(4):961–71. [PubMed: 18189369]
6. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem*. 2003; 75(18):4818–26. [PubMed: 14674459]
7. Pan C, Kora G, McDonald WH, Tabb DL, VerBerkmoes NC, Hurst GB, Pelletier DA, Samatova NF, Hettich RL. ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Anal Chem*. 2006; 78(20):7121–31. [PubMed: 17037911]
8. Pan C, Kora G, Tabb DL, Pelletier DA, McDonald WH, Hurst GB, Hettich RL, Samatova NF. Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Anal Chem*. 2006; 78(20):7110–20. [PubMed: 17037910]
9. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008; 26(12):1367–72. [PubMed: 19029910]

10. Liu H, Sadygov RG, Yates JR. 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem.* 2004; 76(14):4193–201. [PubMed: 15253663]
11. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng J, Fang R, Lin C, Chen J, Goodlett D, Whiteaker J, Paulovich A, McIntosh M. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics.* 2006; 22(15):1902–9. [PubMed: 16766559]
12. Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics.* 2004; 3(10):984–97. [PubMed: 15269249]
13. May D, Fitzgibbon M, Liu Y, Holzman T, Eng J, Kemp CJ, Whiteaker J, Paulovich A, McIntosh M. A platform for accurate mass and time analyses of mass spectrometry data. *J Proteome Res.* 2007; 6(7):2685–94. [PubMed: 17559252]
14. Jaffe JD, Mani DR, Leptos KC, Church GM, Gillette MA, Carr SA. PEPpeR, a platform for experimental proteomic pattern recognition. *Mol Cell Proteomics.* 2006; 5(10):1927–41. [PubMed: 16857664]
15. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL, Carr SA, Clauser KR, Jaffe JD, Kowalski KA, Neubert TA, Regnier FE, Schilling B, Tegeler TJ, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Fisher SJ, Gibson BW, Kinsinger CR, Mesri M, Rodriguez H, Stein SE, Tempst P, Paulovich AG, Liebler DC, Spiegelman C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J Proteome Res.* 2010; 9(2):761–76. [PubMed: 19921851]
16. Zybailov B, Coleman MK, Florens L, Washburn MP. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem.* 2005; 77(19):6218–24. [PubMed: 16194081]
17. Chen YY, Dasari S, Ma ZQ, Vega-Montoto LJ, Li M, Tabb DL. Refining comparative proteomics by spectral counting to account for shared peptides and multiple search engines. *Anal Bioanal Chem.* 2012
18. Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, Koziol JA, Schnitzer JE. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol.* 2010; 28(1):83–9. [PubMed: 20010810]
19. Dicker L, Lin X, Ivanov AR. Increased power for the analysis of label-free LC-MS/MS proteomics data by combining spectral counts and peptide peak attributes. *Mol Cell Proteomics.* 2010; 9(12):2704–18. [PubMed: 20823122]
20. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics.* 2010; 2(4):433–459.
21. Ma ZQ, Polzin KO, Dasari S, Chambers MC, Schilling B, Gibson BW, Tran BQ, Vega-Montoto L, Liebler DC, Tabb DL. QuaMeter: Multi-Vendor Performance Metrics for LC-MS/MS Proteomics Instrumentation. *Anal Chem.* 2012
22. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res.* 2007; 6(2):654–61. [PubMed: 17269722]
23. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics.* 2008; 24(21):2534–6. [PubMed: 18606607]
24. Li M, Gray W, Zhang H, Chung CH, Billheimer D, Yarbrough WG, Liebler DC, Shyr Y, Slebos RJ. Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J Proteome Res.* 2010; 9(8):4295–305. [PubMed: 20586475]
25. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res.* 2009; 8(8):3872–81. [PubMed: 19522537]
26. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res.* 2007; 6(9):3549–57. [PubMed: 17676885]

27. Josse J, Chavent M, Lique B, Husson F. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *Journal of Classification*. 2012; 29(1):91–116.
28. Johnson, NLKS. *Continuous Univariate Distributions*. John Wiley & Sons; New York: 1994.
29. Fisher, RA. *Statistical Methods for Research Workers*. Oliver and Boyd; Edinburgh: 1932.
30. Karpievitch Y, Stanley J, Taverner T, Huang J, Adkins JN, Ansong C, Heffron F, Metz TO, Qian WJ, Yoon H, Smith RD, Dabney AR. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*. 2009; 25(16):2028–34. [PubMed: 19535538]
31. Trudgian DC, Ridlova G, Fischer R, Mackeen MM, Ternette N, Acuto O, Kessler BM, Thomas B. Comparative evaluation of label-free SING normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics*. 2011; 11(14):2790–7. [PubMed: 21656681]

Figure 1.

The IDPQuantify workflow extracts chromatograms for identified peptides to build intensity information for proteins. The mzML data model allows for identification through standard tools, with IDPicker generating sets of confidently identified spectra and assembling proteins. IDPQuantify uses QuasiTel to integrate chromatograms associated with each peptide and sums these peak areas for each protein group. Functions in the R language generate p-values from intensities and spectral counts, integrating them for combined analysis.

Figure 2a.

Figure 2b.

Figure 2.

Peptide repeatability in technical replicates of yeast lysates from CPTAC Study 5. (a) Distinct peptide matches shared among replicates. ~30% peptides are only found in one replicate. Most (67%–79%) peptides were not universal among all six replicates. (b) Boxplot of log intensity for the peptides shared by different numbers of replicates. The boxes represent the interquartile range, while the whiskers represent the full range of observed values. The midline in each box is the median. When a peptide was observed in multiple replicates, the figure records the median log intensity across replicates. Peptides that were observed in more replicates were more intense than those appearing in fewer replicates.

Figure 3. Peptide retention time mapping effectively increased apparent intensity correlation in CPTAC Study 6 replicates. The heatmap plots pairwise correlations between UPS peptide intensities in 12 replicates in each Group (A, B) before (left column) or after (right column) RT mapping. The 12 replicates consisted of data from 4 institutes with 3 replicates by each institute, for example, O65_B1 is the data from institute OrbiO@65 group B replicate 1. Peptide retention time mapping increased within and between institute correlation in CPTAC study 6.

Number of True Positives in Top 50 Differential Proteins in CPTAC study 6

Figure 4.

Peptide retention mapping and p-value combination increased the number of true positives while suppressing false positives in a fixed number of top differential proteins. In the CPTAC study 6 dataset, pairs of groups – A vs. B, B vs. C, C vs. D, D vs. E with 3 folds of change in spike-in UPS protein abundance were compared. Intensity with RT mapping (red) significantly increased the number of true positives in the top 50 differential proteins especially in groups with low spike-in abundance- A vs. B, B vs. C. In groups with high spike-in abundance C vs. D, D vs. E, because there were not many missing values in UPS protein intensity values, RT mapping did not significantly change the TP rate. P-value combination using Fisher's method performed the best among almost all cases.

Figure 5a.

Figure 5b.

Figure 5.

IDPQuantify accurately estimates protein intensity ratio. (a) Boxplot of log intensity ratios of UPS1 proteins in CPTAC study 6. The intensity of proteins in replicates were averaged and compared between two groups. The black horizontal line indicates the true spike-in ratio (standard) – $\log(3)$. With increased protein abundance, intensity ratio estimation accuracy increased, narrowing boxes vertically. There were two outliers below -1 and one above 1.6 which were not plotted. (b) Boxplot of log intensity ratios of UPS proteins in the NRD-pfu dataset. The horizontal line indicates the true spike-in ratio for each cassette with corresponding colors. Each box shows the intensity ratio of UPS proteins in a single replicate (e.g. Replicate 1 A vs. B). The intensity ratios were accurately distributed around the true value in quintuplets. The correlation between the average estimated intensity ratio and the spike-in ratio is 96.72% in Sextuplets, and 95.36% in Quintuplets.

Figure 6. Distribution of peptide intensity ratio in CPTAC study 6 dataset. The vertical black line indicates the true spike-in ratio: $\log(3)$. The intensity ratios of peptides were distributed around the true ratio. With increase in protein abundance, the distributions of peptides were more convergent to the true ratio. All UPS peptide ratios estimated by IDPQuantify were centered on the true spike-in ratio.

Figure 7a.

Figure 7b.

Figure 7.

Correlation between log spike-in protein abundance and a) Log intensity by IDPQuantify or ProALT; b) ProPCA score which combines intensity and spectral count c) Log spectral count. 25 UPS2 proteins were identified within 5 groups at abundance of 50000, 5000, 500, 5, 1 fmol, from left to right. Shown in 7b, the correlation of log intensity by IDPQuantify is substantially better than the alternative intensity-based method: ProALT, log spectral count and is comparable to ProPCA which combines intensity with spectral counting. Our pipeline quantified 5 more UPS2 proteins-NEDD8 (Q15843), SUMO-conjugating enzyme UBC9 (P63279), Creatine kinase M-type (P06732) and Lysozyme C (P61626), Microtubule-associated protein tau (P10636-8) - with correlation of 0.9905, 0.9662, 0.9567, 0.7798 and 0.7712 respectively. The correlation of spectral counts for lysozyme C(P61626) and Microtubule-associated protein tau(P10636-8) are lower than 0.5 and omitted from the figure. The scatter plot and fitted line of log spike-in concentration and log intensity of three proteins from three concentration groups -Ubiquitin (P62988), Small ubiquitin-related modifier 1 (P63165), Ribosylidihydronicotinamide dehydrogenase(P16083)- were shown in 7a.

Figure 8. Boxplot of log fold change of UPS proteins by idpQuantify, MaxQuant and SING in CPTAC study 6 O65 dataset

The real log fold changes in each comparison group were shown by blue horizontal lines. The missing MS1 intensity rate is 12.1% by MaxQuant, 31.8% by SING, and 4.9% by idpQuantify. The estimated ratios by idpQuantify were shown in Supplementary material 2.

Table 1

Matching rate and identification rate of peptides by dataset

Dataset	Matching Rate	Identification rate
CPTAC study 5 Orbi@86	0.8601	0.5249
CPTAC study 5 OrbiW@56	0.9283	0.5558
CPTAC study 5 OrbiP@65	0.9231	0.6080
CPTAC study 6 OrbiO@65	0.8390	0.3663
CPTAC study 6 OrbiP@65	0.7680	0.3204
CPTAC study 6 OrbiW@56	0.8662	0.3694
CPTAC study 6 Orbi@86	0.6749	0.3220
NRD-Pfu Sextuplets	0.9582	0.4801
NRD-Pfu Quintuplets	0.8934	0.5471
UPS2	0.5224	0.2787

* The matching rate was calculated by dividing the total number of idpQuantify matches to peptide ion LC-MS peaks by the product of the total number of samples and the total number of peptides identified by MS/MS spectra

* The identification rate was calculated by dividing the total number of identified peptide (matching to at least one spectra) by the product of the total number of samples and total number of peptides identified by MS/MS spectra