Label-Free Comparative Analysis of Proteomics Mixtures Using Chromatographic Alignment of High-Resolution μLC—MS Data

Gregory L. Finney,† Adele R. Blackler,‡ Michael R. Hoopmann,† Jesse D. Canterbury,† Christine C. Wu,‡ and Michael J. MacCoss*†

Department of Genome Sciences, University of Washington, Seattle, Washington 98195, and Department of Pharmacology, University of Colorado Health Sciences, Aurora, Colorado 80045

Label-free relative quantitative proteomics is a powerful tool for the survey of protein level changes between two biological samples. We have developed and applied an algorithm using chromatographic alignment of μLC—MS runs to improve the detection of differences between complex protein mixtures. We demonstrate the performance of our software by finding differences in E. coli protein abundance upon induction of the lac operon genes using isopropyl β-D-thiogalactopyranoside. The use of our alignment gave a 4-fold decrease in mean relative retention time error and a 6-fold increase in the number of statistically significant differences between samples. Using a conservative threshold, we have identified 5290 total μLC—MS regions that have a different abundance between these samples. Of the detected difference regions, only 23% were mapped to MS/MS peptide identifications. We detected 74 proteins that had a greater relative abundance in the induced sample and 21 with a greater abundance in the uninduced sample. We have developed an effective tool for the label-free detection of differences between samples and demonstrate an increased sensitivity following chromatographic alignment.

The comparative analysis of proteomic mixtures poses a complex analytical problem. Proteins are not well-suited to a one-size-fits-all approach for sample detection and quantification. They display a broad array of physicochemical properties and are expressed over a very large dynamic range—complicating the analysis of large numbers of intact proteins in parallel using a single technology. To overcome the complexities of handling proteins, proteomics methods routinely digest proteins to peptides prior to analysis.1,2 The peptides in the mixture are often separated by microcapillary chromatography (μC) and are emitted into the mass spectrometer “on-line” by electrospray ionization. Fragmentation spectra are acquired by data-dependent acquisition as peptides elute from the chromatography column and the resulting spectra are searched against a sequence database to identify the respective peptide sequences.3,4 This approach is undirected and largely unbiased because the data are acquired semirandomly, and there is no attempt to acquire data on predetermined molecular species. Fragmentation spectra, using data-dependent acquisition, are acquired for all precursor ions above a predetermined threshold—whether they are of interest or not. This technique provides a means of profiling the peptide contents in complex mixtures.

While the acquisition of tandem mass spectra (MS/MS spectra) using data-dependent acquisition is extremely powerful, the acquired fragmentation spectra constitute only a small fraction of the total information in a μLC—MS analysis. Furthermore, the selection of precursor ions by data-dependent acquisition is semirandom and leads to an irreproducible collection of product ion spectra in replicate analyses. Thus, a peptide could be within the detection limits of the mass spectrometer but unselected for fragmentation because the precursor is shadowed by other more abundant species. Likewise, because of this random sampling, only ~70% of the peptide identifications are shared between technical replicates,5 complicating comparisons between samples using MS/MS spectra acquired with data-dependent acquisition. Because of these complications, it is more appropriate to use the information in the MS scans to make comparisons between samples and to use the MS/MS spectra to annotate MS features of altered abundance with peptide identifications.

Over the past few years, several research groups have developed computational approaches for the label-free detection of differences between peptide mixtures using μLC—MS data. These methods can be divided into two classes of algorithms. The first class of algorithms uses peak detection to find features in either the mass spectra (i.e., peptide isotope distributions)6—14 or

The abundances of these features are then used to make quantitative comparisons between samples run with replicate μLC–MS runs using signals from the same m/z and retention time. The second class of algorithms does not rely on peak detection per se but instead uses statistics to identify regions of m/z and time that are significantly different between samples. Using this alternative approach, data are binned into narrow m/z and time regions and the mean intensity combined with the variance from replicate analyses is used to find bins with an altered abundance between samples. The second of the two approaches is particularly appealing for detecting differences between samples. Even if a feature cannot be detected in the μLC–MS analyses from any one of the samples using peak detection algorithms, regions of m/z and time can still be detected given enough replicates to obtain a statistically significant difference in abundance.

One of the greatest sources of error in any μLC–MS analysis is chromatographic retention time reproducibility. When using a computer algorithm that compares the mass spectrometer signal intensities to identify differences between samples, an analyte must appear at the same retention time, otherwise it will be treated as a different signal. Thus, regardless of the approach used to find regions of μLC–MS runs that have a difference in intensity, even minor variances in the chromatographic retention time reduce the ability to compare signals between analyses. Nonlinear corrections to run-to-run shifts in retention time can be applied to align each μLC–MS run from an experiment to a common chromatographic time scale. An established technique for determining nonlinear shifts between series of data is dynamic time warping (DTW), which determines an optimum mapping of time regions and the mean intensity MS run from an experiment to a common template. Here we report the use of an in-house-developed variant of the DTW algorithm to improve the detection of differentially expressed features in μLC–MS runs using the approach reported by Weiner et al. We have developed a suite of tools that applies DTW to align multiple μLC–MS analyses to a common template. These aligned data can then be used to find differences between samples using replicate μLC–MS analyses even in very complex mixtures. By using an LTQ-Orbitrap mass spectrometer, we have sufficient peak capacity and dynamic range for detecting differences between samples even with minimal chromatographic separation. The ability to align and compare retention times between μLC–MS runs is used not only to detect features of altered abundance but also to match peptide identifications obtained from low-resolution tandem MS/MS spectra to μLC–MS regions of differential abundance. This novel set of tools was applied toward the identification of membrane enriched proteins with altered abundance in Escherichia coli upon induction of the lac operon using the nonhydrolyzable alloactose molecular mimic isopropyl β-D-thiogalactopyranoside (IPTG). The detection of differences before and after chromatographic alignment is investigated in the context of a complex protein mixture, and we demonstrate that alignment is an essential component of differential mass spectrometry.

**Materials and Methods**

Sample Preparation. *E. coli* K12 strain MG1655 was cultured in LB media to mid-log phase, divided into two equal portions, and one sample was treated with 1 mM IPTG for 30 min to induce expression of the lac operon genes. Each sample was pellet and lysed at 1000 psi in a French press in PBS buffer (pH 7). Membrane fractions were obtained by spinning the sample at 10000g in a benchtop ultracentrifuge for 1 h at 4 °C. An aliquot of 500 μg equiv of membrane proteins was assayed using a modified Lowry assay with the RC/DC Protein Assay Kit (Biorad, Hercules, CA) and then resuspended in a solution of 0.1% Rapigest (Waters Corp., Milford, MA) in 50 mM ammonium bicarbonate. Proteins were reduced, alkylated, and digested with trypsin as described previously.

Microcapillary Liquid Chromatography Mass Spectrometry. Six technical μLC–MS replicates from each of the IPTG-induced and uninduced samples were analyzed with a μLC–MS similarity function. DTW has been applied successfully to μLC–MS data to determine retention time shifts to correct for chromatographic variance between replicate runs.

Here we report the use of an in-house-developed variant of the DTW algorithm to improve the detection of differentially expressed features in μLC–MS runs using the approach reported by Weiner et al. We have developed a suite of tools that applies DTW to align multiple μLC–MS analyses to a common template. These aligned data can then be used to find differences between samples using replicate μLC–MS analyses even in very complex mixtures. By using an LTQ-Orbitrap mass spectrometer, we have sufficient peak capacity and dynamic range for detecting differences between samples even with minimal chromatographic separation. The ability to align and compare retention times between μLC–MS runs is used not only to detect features of altered abundance but also to match peptide identifications obtained from low-resolution tandem MS/MS spectra to μLC–MS regions of differential abundance. This novel set of tools was applied toward the identification of membrane enriched proteins with altered abundance in *Escherichia coli* upon induction of the lac operon using the nonhydrolyzable alloactose molecular mimic isopropyl β-D-thiogalactopyranoside (IPTG). The detection of differences before and after chromatographic alignment is investigated in the context of a complex protein mixture, and we demonstrate that alignment is an essential component of differential mass spectrometry.
reversed-phase column coupled to an LTQ-Orbitrap mass spectrometer. A 5-μg aliquot of each sample was loaded onto a 75-μm i.d. column packed with 15 cm of Luna C18 (Phenomenex, Torrance, CA) material, using the divert valve to vary between a loading flow rate of ~2 μL/min to a running flow rate of ~500 nL/min as described in detail elsewhere. Peptides were eluted from the column using two buffer solutions: Buffer A was a mixture of 95% water, 5% acetonitrile, 0.1% formic acid, and buffer B was a mixture of 20% water, 80% acetonitrile, and 0.1% formic acid. The run began with 27 min of 95% buffer A during the loading period, followed by five data-dependent MS/MS scans at low resolution for a set of runs to discover changes in abundance. More specifically, given a given run \( T \), then at equivalent time points the warped run \( w \) as the vectors of \( m/z, \) retention time, RT) bins correspond to the presence of the same analyte across runs. To satisfy this condition, we aligned and warped \( \mu LC-MS \) runs in the retention time domain so that signals at a given \( m/z \) could be compared at the same retention time across a set of runs to discover changes in abundance. More specifically, if a given run \( A \) is to be aligned and warped against a template run \( T \), then at equivalent time points the warped run \( A_w \) and the template run should contain the same analyte elution order from the LC column. We define \( R_w, R_e \), and \( R_A \) as the vectors of retention times for scans, respectively, from scans in \( A, T, \) and \( A_e \). Our algorithm produces \( A_e \) from \( A \) to maximize the similarity

Figure 1. Strategy for computational label-free analysis of \( \mu LC-MS \) runs using CRAWDAD.
of scans in $A_n$ to the scans in $T$ at equivalent time points, assuming that the runs $A$ and $T$ contain a subset of their analytes in common, even though levels may differ. Unique to our alignment algorithm is the use of a two-step alignment that uses downsampled data to set alignment boundaries and then uses the full data to compute the final high-resolution path. This two-step process improves the speed of the alignment while maintaining the quality (see Supporting Information, SI, Figure 1).

**μLC—MS Run Chromatogram Alignment.** To minimize retention time variance from one analysis to another, a metric of scan similarity is used to map scans between analyses. We use the square of the normalized dot-product of two scans expressed as intensity vectors as a scoring function. A score matrix $S$ is generated, which defines the similarity between MS scans from run $A$ and the template run, $T$. Score values at each coordinate $(i, j)$, where $i$ is the time-normalized scan index in the align run, and $j$ is the index for the template, are populated by applying a similarity function to scans $A_i$ and $T_j$. A diagonal band constraint on the alignment search space is determined in the first step of the two-step alignment. To reduce having to search the entire score matrix to find the best alignment path, consecutive bins of 10 scans in the time dimension are averaged to produce a downsampling of the two runs being aligned by a factor of 10. CRAWDAD then performs an initial alignment on this downsampled data as described below. The maximum deviation in retention time of the align run from the template run in the downsampled alignment is found and then applied to the alignment of the full data set as a global alignment constraint on the score matrix (Sakoe-Chiba band) in the alignment of the full resolution data.

An alignment path through either the downsampled or full score matrix represents a continuous chain of relations between scans that indicates the chromatographic shift between two runs. It is composed of a set of pairs of monotonically increasing scan indices $(a, t)$ respectively indexing runs $(A, T)$. The path score $P$ is built iteratively by adding the scores from the cells of $S$ multiplied by a weight dependent upon whether the transition used signifies a stretch in $A$ relative to $T$, a shrink in $A$ relative to $T$, or an equal time progression in both runs (eq 1).

$$ P(i, j) = \max \left\{ \frac{\text{sum of weights in } P(i, j)}{\sum_{i=1}^{N} S(i, j)} \right\} $$

Weights giving an equal bias to diagonal transitions as well as stretch followed by a shrink are $w_1 = w_3 = 1$ and $w_2 = 2$. We have used a weight of 2.1 for diagonal transitions in this work to give a small bias to maintaining an equal progression of time points have used a weight of 2.1 for diagonal transitions in this work to give a small bias to maintaining an equal progression of time points.

Candidate paths to find the optimum path are constrained to begin on the edge bordered by $S(M - \epsilon, N)$, $S(M, N)$, and $S(M, N - \epsilon)$ where $M, N$ are the dimensions of the score matrix and $\epsilon$ is the size in scans of the window used to calculate the score matrix around the diagonal. DTW paths beginning from each of these points are calculated, and the path $P_{\text{max}}$ with the maximum score normalized by the sum of the weights (defined above) used along the path is chosen as the highest-scoring path.

**μLC—MS Run Warping.** The DTW path described above was used to transform run $A$ to run $A_n$, minimizing differences caused by chromatographic variation in the retention time domain with respect to template run $T$. As described previously by Prince et al., the scoring path is reduced to produce a one-to-one mapping of time points in $R_a$ to $R_n$. A bicubic spline function is fit using the curfit function from the DIERCKX curve-fitting package using the scan similarity score values from the score matrix as weights and a smoothing factor of 1/2 the number of scans. For every time point $t_i$ in the template run $R_n$, we obtain the corresponding time point in $R_a$, which is most similar, accounting for chromatographic differences. Because the warped time point is typically not a time point in the original observed scans in the unwarped run $A$, the spectra are linearly interpolated between the neighboring observed scans in $A$ (i.e., neighboring time points in $R_a$) to produce scan $A_{w_i}$ at time point $t_i$. At the end of the warping process, the set of aligned scans are stored in an SMAT file as XICs rather than spectra to simplify the analysis and to speed the detection of statistical differences in chromatographic peaks.

**Assessment of Alignment Quality.** Alignment quality was assessed using the retention time standard deviation for persistent peptide isotope distribution (PID) markers identified in all 12 runs using the program Harddlor. Briefly, peptide isotope distributions were detected with a signal-to-noise ratio of $\geq 3$ and a conservative correlation score of $\geq 0.99$. To remove redundancy and eliminate PIDs that did not persist chromatographically over time, only PIDs which were within 10 ppm in three or more consecutive scans were defined as a persistent PID. The criterion for assessing whether a persistent PID belonged to the same analyte between separate μLC—MS runs was that the $m/z$ for each member of a group of persistent PID markers is within a 10 ppm window and a retention time of 3 min. The relatively loose criterion for the time constraint was made possible by the high mass measurement accuracy of the LTQ-Orbitrap. Using the retention time values from the 12 analyses, the standard deviation (SD) was calculated for each marker before and after applying our alignment routine.

**Difference Region Discovery and Calculation of Relative Abundance.** Each XIC is treated independently to search for regions corresponding to chromatographic peaks where differences in mean abundance levels of the groups of replicates are statistically significant. A t-test (assuming independent variances between the groups) is assessed at every time point within an XIC against the null hypothesis that intensity values from technical replicates derived from the same sample are drawn from identical distributions. We define a region of a run as being a difference region (DR) if the $t$-test $p$-value persists below a threshold over a minimum chromatographic length. For the data described in the Results section below, a $p$-value of 0.005 was used and a width of 0.25 min; approximately half the length of a chromatographic peak in this data set. A receiver—operator characteristic (ROC) area under the curve was calculated to characterize the separation

between the maximum intensity within the difference regions between the induced and uninduced samples.

As the chromatographic alignment does not produce warped chromatograms that are perfectly in register, it is likely that the maximum value of a chromatographic peak from an individual replicate is not at the maximum mean value of an aligned set of replicates. The maximum values from each warped LC–MS replicate in a replicate group (a set of technical or biological replicates from a single sample) within a difference region are used to calculate a mean value for that group. The ratio of these replicate group means is defined as a difference region ratio.

**Mapping Difference Regions to Peptide Identifications.**

Fragmentation spectra acquired at low resolution in the LTQ linear ion trap were searched using SEQUEST against a database containing *E. coli* protein sequences (UniProt release 8.0) and common contaminants. Data were searched using a precursor ion mass tolerance of ±3 Da with no enzyme specificity. Search results from a shuffled decoy database were used to assign a q-value to each spectrum identification using the program Percolator, a semisupervised machine learning algorithm. Spectra matching to peptides with a q-value less than or equal to 0.005 were retained.

A difference region was annotated with a MS/MS peptide identification when passing the following criteria: (A) The warped retention time of the respective MS/MS spectrum must fall within the difference region and (B) the difference region m/z bin must lie within the precursor fragmentation window. The bin constraints were expanded to include potential isotope peaks expected for the respective charge state. Conflicts arising from multiple MS/MS identifications mapping to a single difference region were resolved by taking the MS/MS identification closest in the m/z of the difference region to the base isotope peak calculated for the peptide. Any additional conflicts were resolved by taking the peptide identification from an MS/MS spectrum acquired closest to the retention time of the maximum intensity value within the difference region.

**Calculation of Peptide and Protein Ion Current Ratios.**

Detected difference regions were mapped to an MS/MS spectrum when possible and then grouped by the respective peptide sequence returned from the protein database search. Abundance ratios for peptides and proteins are calculated by the mean of the respective difference region abundance ratios weighted by the square root of the mean intensity from the most intense replicate group:

$$ R_{\text{pep}} = \frac{\sum_{j \in \{\text{DR}\}} I_{j,1} \sqrt{\max (I_{j,1}, I_{j,2})}}{\sum_{j \in \{\text{DR}\}} \sqrt{\max (I_{j,1}, I_{j,2})}} $$

where $R_{\text{pep}}$ is an abundance ratio for a peptide, DR is the set of difference regions mapping to that peptide, and $I_{j,1}$ and $I_{j,2}$ are the intensities from replicate groups 1 and 2 of difference region $j$ used to calculate the difference region ratio as described above. The abundance ratio for a protein is calculated in a similar manner using the difference regions for peptides mapping to the respective protein locus. Results from difference regions are organized into a hierarchical list grouped by proteins, their constituent peptides, and the sample in which their relative abundance increased. While it is likely that the measured ion current ratios will reflect the appropriate mole ratios of the respective peptides between the samples, without validating the linear response for each analyte throughout the entire intensity range these data should be conservatively considered semiquantitative or simply “different” unless demonstrated otherwise.

**RESULTS**

As discussed above, a primary source of error in replicate LC–MS analyses is variation in chromatographic retention time. Unfortunately, even the most reproducible microcapillary liquid chromatography separation can have errors equal to or greater than the chromatographic peak width. Figure 2 illustrates the error routinely observed in these analyses by overlaying the base peak chromatograms from six replicate injections of a peptide mixture from the digestion of an identical *E. coli* membrane fraction. These data display a mean SD of 0.23 min, comparable to the width of a chromatographic peak, across the 12 different LC–MS analyses (6 induced and 6 control) for a defined group of 486 peptide isotope distributions detected in all runs using the program Hardkloër.

To minimize the error from the chromatographic retention time, we have implemented a modified form of dynamic time warping to align each individual run from the set of six *E. coli* IPTG-induced and six *E. coli* control (uninduced) runs from a membrane-enriched fraction to a common master template. The master template was chosen empirically from the IPTG-induced data series based upon the run with the chromatographic retention time closest to the mean of all 12 runs. As described above, to generate each pairwise alignment, a score matrix of the similarities between scans from the “align” and “template” runs was calculated using the square of the dot-product between scans binned at 0.1 m/z intervals. Figure 3 shows an example of the score matrix between two runs, and a path found by dynamic programming, which indicates the correction for chromatographic drift versus time between the two runs. By using the same master template for all runs, the set of 12 runs was aligned to a common retention time frame, so that comparisons of signal levels at a given set of m/z and RT values should indicate the same analyte between all replicates.

The quality of the alignment was determined as using the location of 486 peptide isotope distribution markers using Hardkloër as described in methods. The SD of the retention time for these markers before and after alignment is plotted versus mean retention time in Figure 4. The mean standard deviation across all markers was 0.23 min prior to alignment and was improved to 0.063 min following alignment. The efficacy of the alignment approach with data acquired from low-resolution instruments (e.g., a quadrupole ion trap) was estimated by binning the mass spectra using the difference regions for peptides mapping to the respective protein locus. Results from difference regions are organized into a hierarchical list grouped by proteins, their constituent peptides, and the sample in which their relative abundance increased. While it is likely that the measured ion current ratios will reflect the appropriate mole ratios of the respective peptides between the samples, without validating the linear response for each analyte throughout the entire intensity range these data should be conservatively considered semiquantitative or simply “different” unless demonstrated otherwise.

**REFERENCES**


---

from the same data set at an interval of 1 \( m/z \). A comparison of the distribution of SD values for the peptide isotope distribution markers from the 1 and 0.1 \( m/z \) binned amounts did not show a significant difference. In separate experiments, we have also confirmed that this alignment approach is compatible directly with data generated from a stand alone LTQ ion trap mass spectrometer (data not shown).

While the alignment of chromatograms is a challenging computational problem, the ultimate goal is to use chromatogram alignments to improve the detection of differences between \( \mu LC-MS \) analyses. The XICs in Figure 5 arise from the tryptic peptide LWSAEIPNLYR formed from the protein \( \beta \)-galactosidase. \( \beta \)-Galactosidase is the protein product of \( lacY \), one of the three genes present in the \( lac \) operon—the primary target of IPTG. The improvement in the data obtained from alignment was assessed using the signal-to-noise ratio of intensity bins before and after alignment across replicate \( \mu LC-MS \) runs. In this case, the signal is the mean intensity at each bin and the noise is the standard deviation from the six IPTG-induced replicates. Prior to alignment, the region from the induced \( \mu LC-MS \) runs shows a signal-to-noise ratio of 1.42 (RSD = 70.4%), while after alignment, the signal-to-noise ratio improves ~5-fold to 6.9 (RSD = 14.4%). As expected for \( \beta \)-galactosidase, the difference region corresponding to the peptide sequence LWSAEIPNLYR has a statistically significant increase in abundance in the IPTG-induced sample relative to the uninduced control. Overall, six peptides from \( \beta \)-galactosidase were detected as increasing in abundance.

As described above, we define regions of extracted ion chromatograms that differ significantly with a \( t \)-test as difference regions. In total, we detected 5920 difference regions from analysis of the aligned data, which is a significant improvement over the 947 detected from unaligned data using the same thresholds (Table 1). The complete output of the difference regions is available as part of the Supporting Information. Notably, 93% of the difference regions completely distinguish the IPTG-induced and uninduced samples with an ROC value of 1.0 (Supporting Information Figure 2). The log of the intensities from the difference regions of the induced and uninduced analyses are plotted against each other in Figure 6. Differences were detected over an intensity range that spans greater than 4 orders of magnitude. Difference regions were annotated by association with MS/MS spectral identifications as outlined above. Notably, only 22.8% of the difference regions found in the aligned data set could be mapped to MS/MS spectral IDs, indicating that the majority of our difference regions found would not be detectable by spectral counting. Even when allowing all MS/MS spectra to match on \( m/z \) and RT criteria alone (i.e., no score filter was applied to the database search results), only 37.5% of difference regions fell within \( \mu LC-MS \) regions selected for MS/MS fragmentation.

In total, 753 proteins were qualitatively identified using SEQUEST and postprocessing with Percolator (submitted Käll et al., 2007). These proteins were identified using conservative criteria (peptide spectrum match \( q \)-value <0.005, ≥2 peptides per protein) from the 12 reversed-phase runs. Using a less conservative \( q \)-value threshold of 0.01, and only requiring 1 peptide per protein, we identified 1010 unique proteins. Difference regions detected from the replicate MS scans were mapped to these identified peptides and grouped by proteins (Table 2). Proteins with <2 peptides with a change in the same direction were discarded. A total of 95 proteins were found with changing levels, with 74 increasing in abundance in the induced sample relative to the uninduced, and 21 decreasing in abundance (Table 2 and Table S2, SI).
Proteins with an altered abundance between the two samples were organized by Gene Ontology (GO) biological process and cell component classes using GoMiner. For subcellular localization purposes, additional annotations were used from UniProt, EcoCyc, and ePSORTdb. A division of changing proteins by subcellular localization is summarized in Figure 7. Notably, all 11 proteins changing in abundance levels that were annotated as localized to the periplasm were found to increase in the IPTG-induced samples. Biological function GO categories that were enriched for genes increasing in abundance with IPTG treatment (Fisher exact test p-value < 0.001) include glucose catabolism, tricarboxylic acid cycle, and oligopeptide transporter activity, while disaccharide transport was significantly enriched in the uninduced samples.


Figure 3. Scan similarity score matrix and chromatographic alignment path. A dynamic programming algorithm is used to find a high-scoring path through a scan similarity score matrix, which maps time points between two runs. This path is then used to "warp" one run to be in register with another run. (A) The dot-product score matrix from the alignment of an uninduced control (second run) run against an induced (third run) template run is displayed as a heat map. The high-scoring path found by dynamic programming is indicated by the overlaid white line. (B) A blowup of the region spanning from 45 to 51 min retention time in both runs, showing the path in more detail. (C) The Δ in retention time indicated by applying DTW with spline smoothing to the path displayed in panels A and B.
Figure 4. Improvement in chromatographic reproducibility using CRAWDAD. The mean RT of 486 individual peptide isotope distribution features present in all 12 runs is plotted on the x-axis against the standard deviation of the RT of the features on the y-axis. Features shown before alignment are shown in green, while features after alignment are in blue. The mean standard deviation of the RT before alignment was 0.23 min, which was improved to 0.063 min after alignment.

Figure 5. Chromatographic alignment revealing a differentially expressed feature. Pre- and post-alignment t-test detection of a difference region from the +2 ion of the LWSAEIPNLRY peptide of beta-galactosidase. (A) Replicate XICs at m/z 681.3–681.4 from the IPTG-induced μLC–MS runs are shown in black lines, the mean intensity in solid red, and ±1 SD in red dashes. (B) Replicate intensities, mean, and ±1 SD from panel A are shown after chromatographic alignment. (C) Aligned replicates from the induced and uninduced series are shown in red and blue, respectively. Mean intensity values are shown in solid lines and ±1 SD in dashed lines. A difference region corresponding to the LWSAEIPNLRY peptide, detected by t-test p-values ≤0.005 over a minimum length of 0.25 min, is shown as a shaded red region whose height is set to the −log of the minimum p-value.

High-resolution mass analyzers can resolve the individual isotope peaks of a multiply charged peptide. By binning at 0.1 m/z intervals, we are able to detect changes in abundance of the individual peaks of a peptide’s isotope distribution. The +2 ion of the TVINQVTYLPASEVTDVNR peptide from periplasmic oligopeptide binding protein (OppA) was detected with difference regions corresponding to the m/z values of its monoisotopic peak through M+5 isotope peak (Figure 8). The range of intensities between these isotope peaks span a 40-fold range in intensity, yet display a remarkably similar relative abundance. The ratios calculated by CRAWDAD for these individual isotope peaks between the induced and uninduced samples are extremely precise with an RSD of 3.7%. Similar changes and precision were found for the 3+ isotope distribution of this peptide (data not shown).

DISCUSSION

Shotgun proteomics has become a powerful tool for the characterization of complex protein mixtures. Peptides are now routinely identified and characterized using a combination of high-resolution separations, rapid and automated acquisition of tandem mass spectrometry data, and database searching algorithms (e.g., SEQUEST). Nevertheless, a large portion of the information acquired in these μLC–MS analyses goes unused; i.e. the MS scans themselves. These MS scans provide the basis for selecting precursor masses for MS/MS fragmentation in data-dependent acquisition experiments and may be used for identifying differences between samples. We present a method using statistical analysis of signals of μLC–MS technical replicates from two different samples, which corrects for chromatographic shifts to discover significant differences in analyte levels between complex protein mixtures.

Chromatographic Alignment for the Improved Detection of Difference Regions. We present the use of a generic chromatographic alignment routine that can handle data from any mass spectrometer and improves the detection of molecular species with altered abundance between samples. While the run-to-run chromatographic reproducibility of our data is similar to that found with splitless nanoflow HPLC systems, we have shown that even unaligned data with <1% chromatographic RSD complicates the detection of features of differential abundance between the two samples (Figure 6). Our alignment method operates on a scan or time point-based level, where a scan in one run is matched to a scan from another and scans must contain a subset of similar peaks to be accurately aligned. We demonstrate here that μLC–MS data from two E. coli cultures between an IPTG-induced and a control group were sufficient to align data using dynamic time warping. Other protein mixtures from our laboratory of various complexity and origin have also been alignable using similar methods (data not shown).

The decrease in the relative error in the signal intensity across replicates is a direct result of our alignment (Figure 5), giving
greater power to the statistical assessment of difference between two runs. Improvements in the signal-to-noise level of the replicate data (reciprocal of the relative standard deviation) not only improve the performance of statistical tests but also can be used as a metric of the efficacy of an alignment procedure itself. Our data suggest that chromatographic alignment can decrease the relative error in the intensity measurement by \(5\times\) fold and increase the detection of difference regions by \(\times6\) fold. Thus, using an efficient chromatographic alignment technique is an important and potentially essential step in the implementation of label-free quantitation software.

**Considerations in Using a Scan-Based Alignment Technique.** Our scan-based alignment algorithm assumes that analytes that coelute in a spectrum from one run also coelute in a spectrum in the other runs. This assumption is in contrast to feature-based alignment routines where individual detected peaks are aligned to one another. Scan-based alignments may be biased toward matching the more abundant of those components between time points if coelution is not consistent between scans in different runs, possibly introducing a small error. However, feature-based alignment methods may not consistently identify an analyte across all replicate runs, while a scan or signal-based technique does not rely on the thresholds or variance derived from feature detection. A further advantage of a scan-based alignment is that it is

**Table 2. Proteins Detected with Changing Abundance Levels**

<table>
<thead>
<tr>
<th>bin size, (m/z)</th>
<th>aligned</th>
<th>protein changes(^a)</th>
<th>elevated in induced</th>
<th>elevated in control</th>
<th>(\geq 3)</th>
<th>(\geq 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 Y</td>
<td>95</td>
<td>74</td>
<td>21</td>
<td>37</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>1.0 Y</td>
<td>72</td>
<td>57</td>
<td>15</td>
<td>32</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>0.1 N</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Number of proteins having two or more peptides detected with changes in abundance. \(^b\) Number of proteins with greater than or equal to three or five peptides detected with changes in abundance.

**Figure 6.** Distribution of difference region intensity values before (A) and after (B) alignment. Difference regions that were mapped (red) and not mapped (black) to MS/MS peptide identifications derived from data-dependent scanning are plotted with the log mean maximal intensity from the uninduced series on the x-axis and from the induced series on the y-axis. Lines indicating changes in abundance of 1.5-fold and 5-fold are overlaid on the plots.

**Figure 7.** Summary of subcellular localization of proteins with detected change in abundance. The subcellular localization of proteins with changes in abundance level was annotated using the Gene Ontology, SwissProt, and ePSORTdb databases. Bar plots associated with the pie chart segments indicate the number of proteins increasing in the induced or uninduced samples.
compatible with most mass spectrometers regardless of resolution and is generally tolerant of poor signal-to-noise data.

In these data, we have used a single “master template” to align all other μLC–MS runs. The computational time and space requirements of dynamic time warping are proportional to the power of the number of data sets aligned simultaneously—hence, we are practically limited to aligning runs in a pairwise fashion. While aligning all data in a pairwise fashion to a single template is the simplest approach for aligning large data sets, it performs quite well. A potential drawback to the use of a common master template is that the selection of the μLC–MS run to use as the master template may affect the overall quality of the alignment. CRAWDAD has been implemented in a modular fashion so that an alternative alignment technique could be used to preprocess the data for the detection of difference regions. A particularly elegant approach to align all analyses at once uses continuous profile models21 to align time series data and could be used with CRAWDAD in the future.

Dynamic Range in Label-Free Detection of Difference Regions. The LTQ-Orbitrap mass spectrometer has a large dynamic range making it an excellent instrument for differential proteomics measurements. This large dynamic range can be demonstrated by the large intensity range in which difference regions can be detected. The difference regions detected throughout the μLC–MS analyses span over 4 orders of magnitude, demonstrating an ability to sample peptides with altered abundance over a wide dynamic range with only a single dimension of chromatographic separation (Figure 5). Furthermore, difference regions were detected from six isotopic peaks of a peptide from the periplasmic oligopeptide-binding protein precursor of E. coli (oppA) with an impressive relative abundance precision and spanning almost 2 orders of magnitude within a single scan. While we have demonstrated a large dynamic range for the detection of difference ranges, we have not demonstrated that the response is quantitative over this wide range. Demonstration of the linear response of the LTQ-Orbitrap is beyond the scope of these experiments and will be reported elsewhere.

Considerations in Interpreting Quantitative Proteomics Results. While CRAWDAD can detect peptide and protein differences between samples, a potential caveat in quantitative or semiquantitative proteomics assays is that any difference between the two samples may be detected—including those differences arising from variances in sample preparation. Furthermore, even sample loading order can result in systematic biases in differential profiling experiments, and to minimize these biases, the sample order should be randomized. The over-representation of protein identifications from periplasmic, inner- and outer-membrane proteins indicates that our enrichment for membrane proteins was effective. However, the effect of a small variance in enrichment efficiency magnifies the observed degree of differences in fractions being depleted relative to the enrichment target and will account for some of the detected protein abundance differences. The use of biological and sample preparation replicates when comparing protein levels allows one to consider the variance in sample preparation and facilitates the discrimination between differences arising from sample preparation variances and those that are biologically relevant.25,56 While the focus of our experiments was to demonstrate the improvement in the difference region detection between samples following chromatographic alignment using an LTQ-Orbitrap mass spectrometer, future experiments that attempt to draw significant biological conclusions between conditions will have to incorporate biological replicates into the experimental design as opposed to just technical replicates as reported in this study.

Practical Advantages of Minimizing Our Dependence on Data-Dependent Acquisition. The majority of the difference regions found between the μLC–MS analyses were not mapped to MS/MS peptide identifications. Currently, the isolation and activation of precursor ions by data-dependent acquisition represents a subset of the total μLC–MS data acquired. While the development of very fast scanning tandem mass spectrometers has improved the data-dependent acquisition of MS/MS spectra

References

on low-abundance peptides in a mixture, these approaches still require extensive fractionation for comprehensive analysis. Methods based on MS/MS acquisition can miss significant changes in intensity derived from analytes that were not sampled for MS/MS fragmentation or not identifiable by database searching due to a range of reasons (i.e., a spectrum derived from a non-peptide molecule, an unannotated protein coding region of the genome, or an unanticipated post-translationally modified peptide). In contrast, using a method to detect differences from the MS scans themselves will only be limited by the peak capacity of μLC–MS and should be independent of the mass analyzer scan speed. Thus, while we can detect difference regions for which we have no MS/MS spectra, we are currently restricted to using MS/MS to determine the molecular identity of these regions. However, these detected peptides can be used to specifically target MS/MS data acquisition in future analyses and direct efforts for manual data analysis. We recently reported using differential analysis with CRAWDAD of an in vitro assay to detect specific modified peptide signals, thereby directing the interpretation of specific MS/MS spectra that were not identified by a database-search algorithm.

CONCLUSION

We have developed an algorithm for label-free comparative proteomics that combines dynamic time warping and differential mass spectrometry. We demonstrated the capabilities of CRAWDAD in the detection of differences resulting from IPTG in enriched membrane fractions from E. coli using a hybrid LTQ-Orbitrap mass spectrometer. Chromatographic alignment increased the detection of statistically significant differences between samples. The majority of the differences found between samples were not associated with MS/MS spectra. Information about obtaining the software and mass spectrometry data used in this article can be found at: http://proteome.gs.washington.edu/software/crawdad.

ACKNOWLEDGMENT

The authors thank Kevin Wheeler, Timothy Johnson, and Tonya Pekar from Thermo Fisher Scientific for providing time on an LTQ-Orbitrap to perform these experiments. We also acknowledge Barbara Frewen, Gennifer Merrihew, and Aaron Klammer for helpful discussions and critical reading of the manuscript. Financial support for this work was provided in part from National Institutes of Health grants R01-DK069386 (M.J.M.), P41-RR011823 (M.J.M.), R21-DA021744 (C.C.W.), and R21-HL083360 (C.C.W.). A.R.B. was supported by an NIH NRSA Predoctoral fellowship (F31-DA022825). G.L.F. and M.R.H. gratefully acknowledge training grant support from T32-HG00035.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review August 2, 2007. Accepted November 5, 2007.

AC701649E

(59) Xie, H. W.; Griffin, T. J. Proteome Res. 2006, 5, 1003–9.