Tools and Analyses for Differential Label-Free Proteomics Using Mass Spectrometry

Gregory L. Finney

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Michael MacCoss, Chair

Philip Green

Robert Synovec

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Tools and Analyses for Differential Label-Free Proteomics Using Mass Spectrometry

Gregory L. Finney

Chair of the Supervisory Committee:
Associate Professor Michael MacCoss
Department of Genome Sciences

The comparative measurement of protein abundance is a powerful method to detect changes in the biological dynamics of cells and tissues. Shotgun proteomics has proven to be a method where a wide range of proteins can be characterized in a single experiment by analysis of their peptide digests using micro-capillary liquid chromatography coupled to mass spectrometry (µLC-MS). We have the CRAWDAD software tool for label-free relative quantitation between samples using peptide signals in µLC-MS data associated with their identifications from MS/MS data.

CRAWDAD discovers features for quantification in µLC-MS data, lowers chromatographic and signal variance across replicates, and finds statistically significant changes in peptide abundance between samples. We have applied this tool toward model systems of induced expression of the *Lac* operator in *E. coli* to qualitatively assess the protein changes detected. A controlled set of changes using an *E. coli* protein digest spiked in at changing levels over a constant background of human proteins assesses the precision and accuracy of the detected changes in peptide and protein levels, and showed our results to be superior to label-free quantitation using spectral counting.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thank you to everyone who encouraged, guided, and supported me on this journey. My parents and family were an essential support and inspiration in my education. My thesis committee was an invaluable source of advice and challenging questions that helped in the development of this work. I would especially like to thank my thesis advisor, Mike MacCoss, for his inspiration, patience, and guidance on these projects.

I had the pleasure of joining and leaning from the dynamic lab he created. My labmates and colleagues Xianhua Yi, Michael Hoopmann, Jarrett Egertson, Jesse Canterbury, Ed Hsieh, Michael Bereman, Daniela Tomazela, Barbara Frewen, and everyone else who has passed through the lab. I'd especially like to thank Genn Merrihew and Veronika Glukhova for help with experimental work and reagents on projects. I'd also like to thank colleagues from the department such as Aaron Klammer, Lukas Kall, Oliver Serang, Katrina Claw, Chris Saunders, Sean McIlwain, Brian Giebel, and many others for good discussions and distractions. Finally, I hope that you enjoy this dissertation and find it useful.

1. Introduction

1.1. Overview

1.1.1. Proteomics and Mass Spectrometry

The comparative measurement of protein abundance is a powerful method to detect changes in the biological dynamics of cells and tissues. Detection of changes in the abundance of proteins between biological states is used for biological discovery and the identification and elucidation of disease states on a cellular or tissue level. Under the classical paradigm of quantitative protein biochemistry, analysis of the activity or structural properties is limited to a small number of proteins (Anderson and Anderson, 1998). In contrast, proteomics studies the protein complement of a cell type or tissue as a whole on a qualitative and quantitative basis (Wasinger et al., 1995). This thesis details the development and application of computational proteomics tools for differential analysis of protein samples without the use of chemical labels. These tools are particularly applicable to large amounts of data from the direct measurement of proteins through mass spectrometry (MS).Mass spectrometry involves the separation of proteins or protein fragments according to their physical properties of mass and charge. MS enables direct measurement of proteins without the creation of reagents specific to individual proteins, which contrasts with traditional biochemical tools such as enzymatic or immunologic approaches. The development of mass spectrometry technology for rapid acquisition of data, coupled with chromatographic separations, enables the identification of on the order of 1000 proteins in a single assay run. Recent developments in mass spectrometry proteomics approach the identification of the full complement of a moderately complex proteome such as yeast with a limited number of analyses (de Godoy et al., 2008).

The vast quantity of data generated presents two significant challenges among many. First, proteins in the sample must be separated before analysis to be quantified, typically by one or more dimensions of chromatography or electrophoresis. Second, the analysis of the vast amount of data requires computational tools to identify and quantify components, and to statistically validate detected differences in mass spectrometry signals which indicate relative changes in protein quantities. This thesis presents the development and use of computational and experimental tools for relative quantitation between protein samples without the use of labeling reagents, and the application of these tools to biological samples using multiple mass spectrometry techniques.

1.1.2. Overview of Shotgun Proteomics

Chemical analyses of intact proteins is extremely challenging, as the wide range of physiochemical properties including molecular weight, pI, hydrophobicity, and solubility make separations challenging. While recent advances in multidimensional separations (i.e. distinct separation technologies coupled in serial) have enabled the analysis of complex protein mixtures by MS (Kellie et al., 2012; Tran et al., 2011), the time for analysis is greatly increased due to the large number of fractions, and the technology has not widely disseminated to the proteomics community at this time.

The limitations of technologies for the separation and analysis of intact proteins has spurred the development of “shotgun proteomics” approaches (Figure 1-1). These shotgun approaches enzymatically digest a protein mixture to a more tractable peptide mixture to facilitate analysis. However, the large numbers of peptides in a single analysis cannot be analyzed simultaneously in the mass spectrometer, and are instead scanned over an *m/z* range during the course of a chromatographic separation. In current practice this is done by microcapillary liquid chromatography (μLC) using one or more dimensions of separation (Licklider et al., 2002) .

Although peptides are more easily separated than proteins, the vast complexity of a peptide digest requires high resolution separations. Furthermore, improved separations and shorter peak widths can decrease the dynamic range of the digested peptides in a single scan, thereby improving the detection of lower abundance peptides. Proteomics separations are performed in both one- and two-dimensional separations, with the 'last dimension' eluting into the mass spectrometer. The most common approach for single dimension chromatography of peptides is separation by hydrophobicity using reverse-phase techniques. . Peptides are loaded onto the column in aqueous solution with a low quantity of organic solvent, with most peptides preferentially partitioning with the solid binding material whose active surface is typically alkyl chains of length between C_4 and C_{18} bonded to a silica substrate. The organic content of the mobile phase is slowly increased from a minimal amount in a gradient. The rate of migration of an analyte from the column depends on its relative affinity for the mobile phase or the hydrophobic binding material. As the hydrophobicity of the solvent system increases, peptides of increasing hydrophobicity partition with the mobile phase rather than the alkyl chains on the chromatographic material, and migrate more quickly off of the column. A separation which is close to orthogonal, such as separating by charge, can be used prior to reverse phase separation to increase the peak capacity and resolution of the separations. Data presented in this thesis pertains only to one dimensional reverse phase separations.

The μLC separation is coupled to analysis in a mass spectrometer by nanoflow electrospray ionization, or nanospray (Wilm and Mann, 1996). A mild acid is added to ensure protonation of peptides, and a positive potential (~3kV) is applied at the base of the capillary, with the inlet to the analyzing MS instrument acting as an anode. Near the tip of the column, positively charged solutes apply pressure as they are attracted to the anode, forcing the liquid meniscus into a Taylor cone. This promotes production of small positively charged droplets which fly toward the negative electrode (Kebarle and Verkerk, 2009). As droplets containing peptides elute from the

column at low flow rates (~100-1000 nL/min), they rapidly desolvate. The positively charged peptides in the rapidly shrinking droplets cause explosive division of the droplets into smaller ones due to electrostatic repulsion. The shrinking and electrostatic explosion of droplets iterates, producing smaller droplets until solutes enter the gas phase (Fenn et al., 1989). Notably, the sensitivity and linear response of a peptide or protein detected by mass spectrometry is dependent upon its ionization efficiency in the electrospray process.

The ionized peptides can then be detected by a wide range of mass separators and analyzers. The typical mass analyzer used in proteomics is an ion trap, which uses radiofrequency and electric fields to trap ions, and selectively ejects them from the trap at specific mass/charge (*m/z)* ratios. Ions are detected by electron multipliers, and the signal level sampled over a range of *m/z* values is converted via an analog-to-digital converter for storage and computational analysis. Higher resolution (lower minimum difference in *m/z* necessary to distinguish signals) mass spectrometers such as the FT-ICR (Bogdanov and Smith, 2005) or the Orbitrap (Makarov et al., 2006) use a cyclotron detector where the orbital frequency of the ions is dependent upon their *m/z* and is converted from the frequency to the *m/z* domain by a Fourier transform.

However, unambiguous identification of peptides is not possible from the masses measured within the precision range of the instruments (typically +/- 1-3 ppm in superior instruments - Palagi et al., 2006), nor from their accurate masses alone (Clauser et al., 1999) and requires fragmentation of the peptides followed by detection of their characteristic fragments using tandem mass spectrometry. Peptide fragmentation is most commonly induced by collision induced dissociation (CID), in which ions collide with neutral gases to produce fragmentation along the peptide backbone. In tandem mass spectrometry, or MS/MS, precursor peptide ions are isolated over a narrow *m/z* range using a first stage of mass analysis, followed by their fragmentation and enumeration of the fragment *m/z* values and ion counts using a second stage of mass spectrometry. he complexity of the mixture requires a method capable of rapidly

selecting and fragmenting potential peptide analytes. A fast mass spectrometer such as the linear quadrupole ion trap can produce tens of thousands of MS/MS spectra per μLC-MS assay, requiring computational analysis downstream of the data acquisition.

An MS scan is produced by the mass analyzer reporting the ion abundance over a range of *m/z* values. A precursor, or MS1 scan, reports abundances for unfragmented ions, while an MS/MS, or MS2 scan, reports on the abundance of the fragment ions of a peptide. The sequential acquisition of MS1 scans of peptides produces a signal map of precursor ion *m/z* values, where the abundance at a given *m/z* and retention time (RT) is displayed using a heat map

The use of the precursor ion abundance is used to drive a semi-stochastic process for MS/MS known as data-dependent acquisition (DDA). As precursor scans are produced, abundant ions are noted and compared by order of abundance against an in-memory dynamic exclusion list of ions fragmented during a previous time window roughly corresponding to a chromatographic peak width. If the candidate ion was previously selected for fragmentation, the next most abundant ion is checked for selection. A candidate precursor not previously fragmented within the time window is fragmented in the trap as described above, and noted in the dynamic exclusion list. This cycle continues for up to a maximum of N precursor ions per MS1 scan, where N is typically 3-10.

Ions are isolated in the trap, and abundance values over an *m/z* range are measured as the instrument scans by selectively ejecting ions over an *m/z* range to the detector. This information is used by the onboard computer to select a narrow range (~2 *m/z*) of precursor ions which a waveform is used to excite in the trap. The excited ions are fragmented by collision-induced dissociation (CID), where inert gas molecules at low pressure (~1 mTorr) impart energy to the peptides by kinetic collisions, causing fragmentation along the amide bonds of the peptide backbone. Fragments are then selectively ejected from the trap over a mass range, and

detected as in MS1 scanning. Other approaches involve using inclusion lists of m/z and expected retention time for peptides of interest [inclusion list reference], or using exclusion information of peptides acquired in previous replicate runs to avoid over-sampling abundant peptides, thereby increasing identity coverage of the total peptide pool (Hoopmann et al., 2009).

However, peptides can often be obscured in MS1 scans due to background chemical noise. MS/MS spectra suffer less interference, and are more sensitive than MS1 spectra (Arnott et al., 2002) Continual acquisition of MS/MS spectra over an *m/z* region is an alternate mode of acquisition which can takes advantage of MS/MS both for identification and quantitation purposes. In contrast to DDA, it is termed DIA or data independent acquisition. Recent work compares this semi-directed approach with continual fragmentation of an *m/z* range during a chromatographic separation for identification purposes (Venable et al., 2004), and for quantitation.

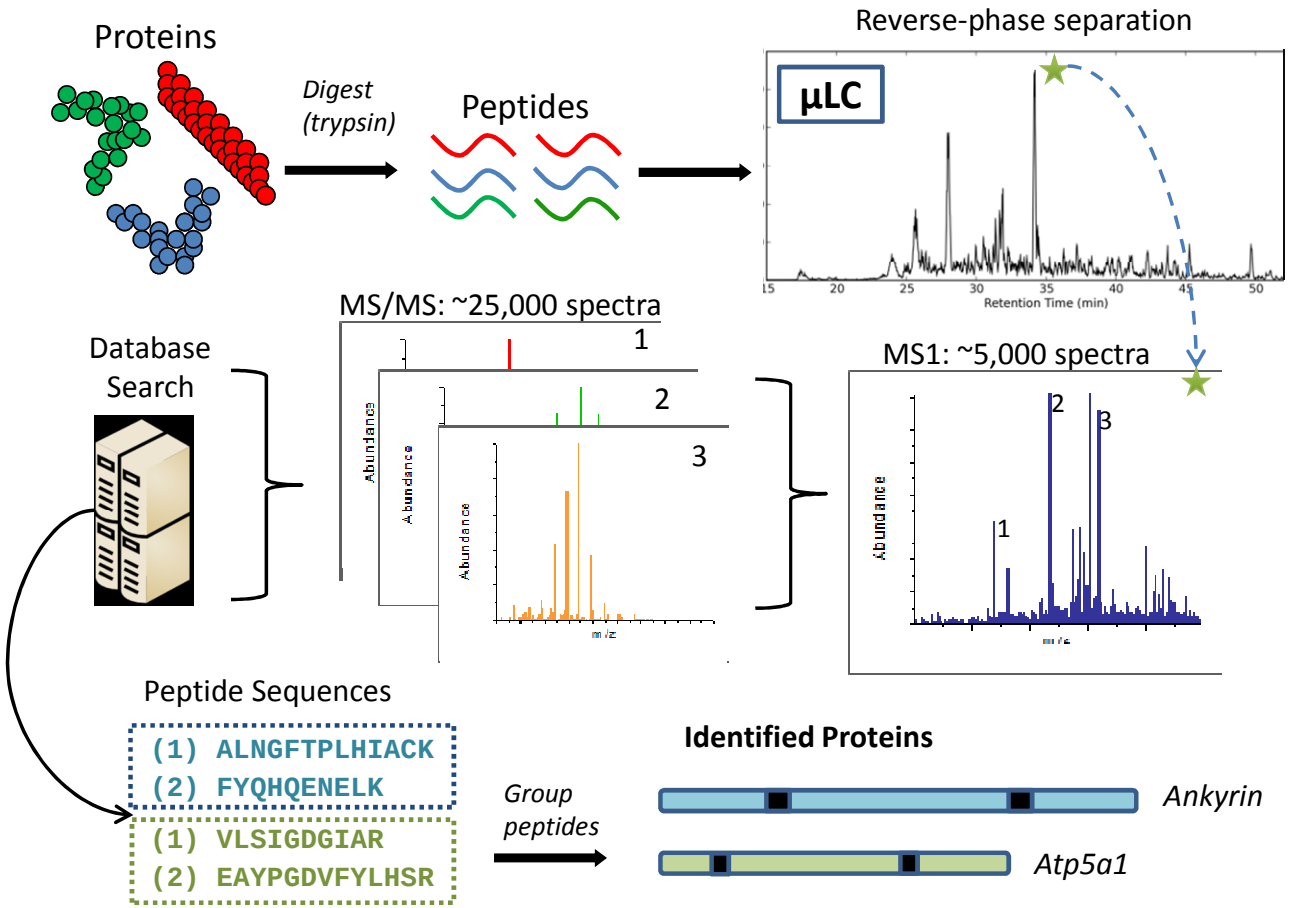

Figure 1-1: Overview of Shotgun Proteomics

(From top left) – Proteins are digested to peptides, loaded on a column, and separated by microcapillary liquid chromatography (µLC). Peptides are continuously ionized from the column into a mass spectrometer (MS), continually acquiring unfragmented MS scans (MS1). A subset of the abundant peaks are selected for tandem MS fragmentation (MS/MS). The MS/MS scans are searched against a protein database for matches to peptides, and detected peptides are assembled into proteins

1.2. Differential Proteomics by Mass Spectrometry

The technical challenges of quantitative proteome analysis are tremendous: the dynamic range of protein concentrations spans over seven orders of magnitude in tissue such as hepatocytes, and at least ten orders of magnitude in plasma (Anderson and Anderson, 2002). Furthermore, as discussed below, multiple sources of variance must be considered and accounted in an experimental approach. Comparisons between levels of peptides across µLC-MS shotgun proteomics experiments are complicated by multiple sources of noise and variance (Anderle et al., 2004): sample preparation and purification steps, chromatographic retention time drift (Finney et al., 2008), ionization suppression (Tang et al., 2004), and noise inherent to ion detection in the mass spectrometer (MacCoss et al., 2001).

While early approaches relied on ultracentrifugation or other bulk techniques to separate samples (Anderson and Anderson, 1998), molecular biologists popularized the use of electrophoretic separations followed by immunologic detection as exemplified by the Western blot (Renart et al., 1979). While this is a very sensitive technique, it is limited in its practicality for monitoring a large number of proteins, as a specific antibody reagent must be prepared for each protein of interest.

The two-dimensional electrophoretic gel, where proteins are separated first by their isoelectric point, followed by separation based on their molecular weight, was a seminal technologic development in the analysis of a complex protein mixture (O'Farrell, 1975). However, it is limited in its dynamic range and sensitivity (Gygi et al., 2000), and has limited utility to resolve membrane or highly hydrophobic proteins (Braun et al., 2007). Furthermore, analysis can be laborious, as each resolved protein ‘spot’ must be excised and separately analyzed for identification of the underlying protein. The aforementioned shotgun approach takes a large set

of proteins and produces a set of peptides with a narrower range of physiochemical properties which can be separated by liquid chromatography for analysis in a mass spectrometer.

Mass spectrometry is not directly quantitative: the ionization efficiency of peptides varies both inherently and due to ionization suppression by other co-eluting analytes (Annesley, 2003). Therefore, the level of ions striking the detector over a fixed time interval does not translate directly via a linear function to the peptide's abundance. This makes it impossible to take a single ion count of a peptide from a scan, or area under the curve of a chromatogram at fixed *m/z*, and absolutely quantify an analyte.

As the ion intensities of peptides are not directly quantitative, one must use known standards to achieve absolute quantitation. Absolute quantitation can be performed by preparing a curve using known concentrations of a standard to the response from the sample of interest. Addition of known unlabeled quantities of an analyte of interest to an experimental matrix can also be used to calibrate a response by using a standard addition. This was used to successfully quantify myoglobin in serum with to a high degree of precision (Mayr et al., 2006). However, the use of a labeled analogue standard can avoid possible matrix effects.

Isotopically-labeled analogues of the compound to be measured have nearly the same physiochemical properties as the original compounds, with the exception of a shift in mass detectable by the mass spectrometer. However, absolute quantitation of a whole proteome, or the fraction detectable by mass spectrometry, requires production of labeled internal peptide standards for each protein. Creating these standards and the characterization of standard curves for quantitation of their individual response functions is impractical for a large number of proteins.

In many biological questions, absolute quantitation is not necessary: the effects of a physiological, developmental, or other significant change from a base state can be elucidated by

the relative changes in protein levels. In this case, changes in proteomes are analyzed by determining the ratios of protein and peptide abundance between samples. Two general approaches exist for relative quantitation of proteomes: using ratios of the same peptide which are ‘labeled’ in different isotopic forms, and comparing the abundance of peptides in a label-free approach where one or more variants are analyzed.

Labeled strategies use ‘light’ and ‘heavy’ variants of the same molecule, where the heavy version has some atoms replaced with stable isotopes of a higher mass. Both variants have nearly identical responses on a chemical level with sample preparation techniques, chromatographic separation, and ionization into the gas phase, yet in the mass spectrometer they can be distinguished by the basic difference of mass. This property also allows the light and heavy versions to be run simultaneously

Strategies for the comparison of proteomes using light and heavy isotopic variants of peptides fall into two areas: i) metabolic labeling; ii) covalent modification of peptides or proteins with molecular tags. In metabolic labeling, an organism such as yeast or mouse (Wu et al., 2005; Kruger et al., 2008) is grown normally (light), or using foodstuff where ^{14}N (light) is replaced with ^{15}N (heavy). In the so-called “stable isotope labeling with amino acids in cell culture” (SILAC) approach, a single essential amino acid is supplied only in light- or heavy-labeled form (Ong et al., 2002). In tagging approaches, isotopic analogues of a molecular tag are attached covalently in separate reactions for each sample. Typically, chemically labile atoms of an amino acid side chain are covalently modified with a molecular tag, exemplified by the addition of heavy or light isotopic variants of a biotin derivative to the sulfhydrl group of cysteine side chains in the ICAT tagging technique (Zhou et al., 2002). Other variants of chemical tagging strategies are well described in the literature (Bantscheff et al., 2007). Labeled approaches minimize variation from differences in chromatography, ionization, and mass spectrometer sensitivity, but increase the sample complexity, thereby reducing the dynamic range of accurate measurements.

Furthermore, the labeling reactions themselves can vary in efficiency when applied to separate samples (Oberg and Vitek, 2009), and can be expensive and difficult.

In contrast, label-free approaches use direct peptide intensity measurements of the samples of interest for relative quantitation. As no isotopic difference in mass exists to distinguish the same peptide from multiple samples, each sample is run separately. Multiple µLC-MS replicates are used to gain statistical confidence in observed changes in peptide level between samples. A label-free approach has the advantage of a simpler experimental workflow without the expense of labeled reagents, and is not limited in the number of samples that can be directly compared as labeled approaches are. Two approaches are common: i) Quantification from the MS precursor ion intensity of a peptide, observed over its chromatographic peak, or ii) using the frequency with which the peptides of a protein are sampled from MS/MS fragmentation events as a proxy for protein intensity, known as spectral counting.

In the precursor intensity approach, features corresponding to a peptide are detected and quantified over their elution time as a chromatographic peak. This allows for the quantitation of multiple co-eluting peptides, as the entire *m/z* range selected by the mass spectrometer is scanned. Signals can be matched to the expected isotopic patterns of peptides (Hoopmann et al., 2007; Horn et al., 2000) or chromatographic peaks over *m/z* windows corresponding to a distinct isotope can be detected (Danielsson et al., 2002; Smith et al., 2006) for quantitation. Tandem MS/MS spectra are used for peptide identification of the peaks differing in intensity. An example shows two separate sets of biological replicate runs of a human heart sample (Kline et al., 2009) used in a label-free computational analysis using precursor intensities (Figure 1-3).

An indirect ‘spectral counting’ approach uses the number of MS/MS spectra identified as acquired from a given protein as a proxy for abundance, and uses statistical approaches to normalize for protein length and the total number of spectra acquired, and to estimate relative

abundances using sampling statistics (Liu et al., 2004). Spectral counts for all identified peptides derived from a protein are combined to a single protein level to maximize the power of statistical comparisons. The precursor intensity approach is technically more challenging, as it requires more sophisticated analysis of the data and it is more sensitive to experimental conditions such as chromatography and ionization efficiency. However, precursor intensity quantitation can give greater accuracy and dynamic range, particularly for low abundance peptides, and can give information on changes on the peptide level, rather than strictly the protein level. Below, I cover approaches for label-free quantitation using precursor ion intensity.

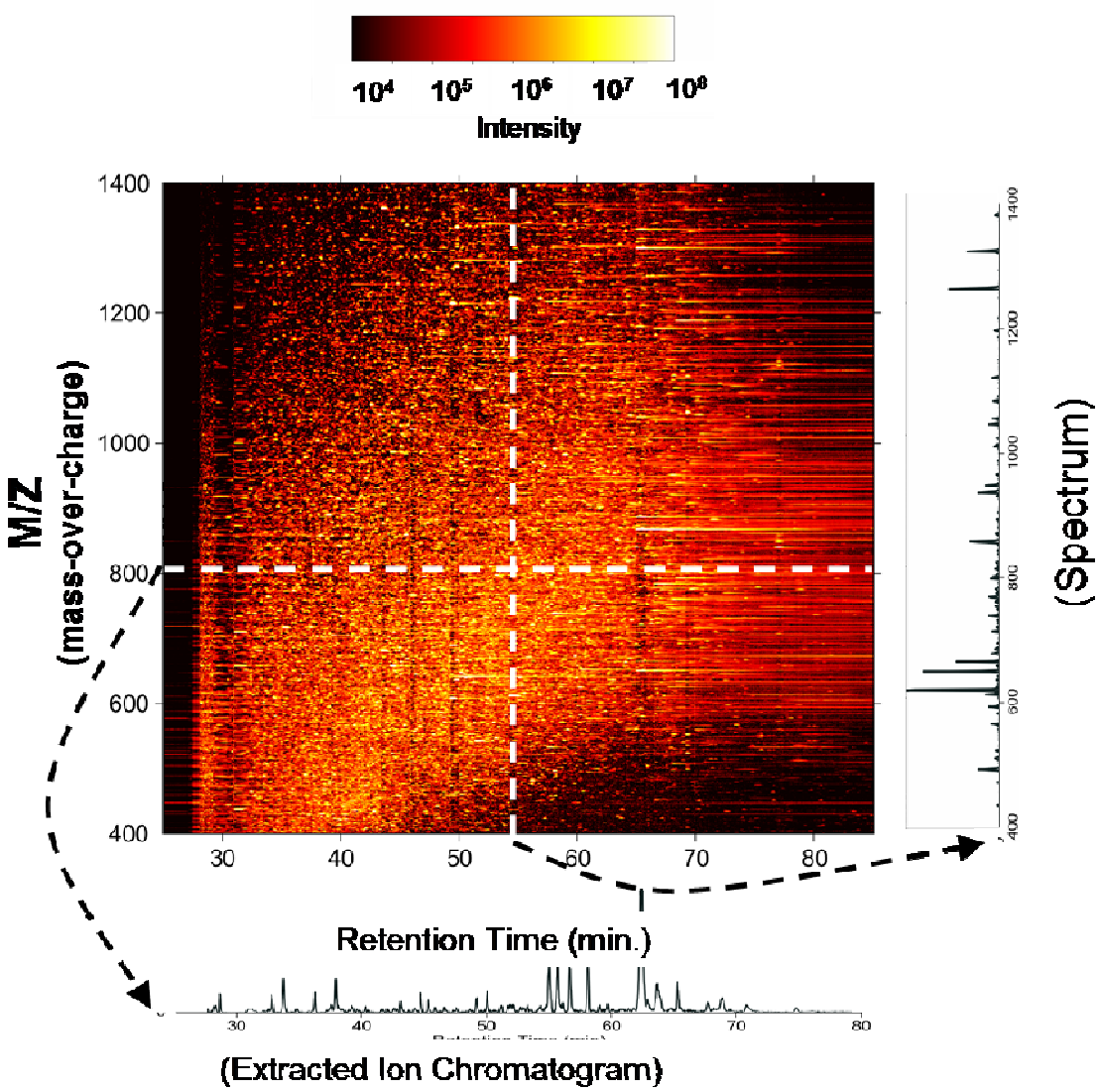

Figure 1-2: µLC-MS Data Obtained on a High-Resolution Mass Spectrometer.

A series of spectra created by the mass spectrometer scanning over an *m/z* range (A), gathered over time as compounds elute from the LC column. The abundance of analytes at a specific *m/z* is shown by an extracted ion chromatogram (XIC) (B) where *m/z* is fixed, and the ion abundance plotted over time.

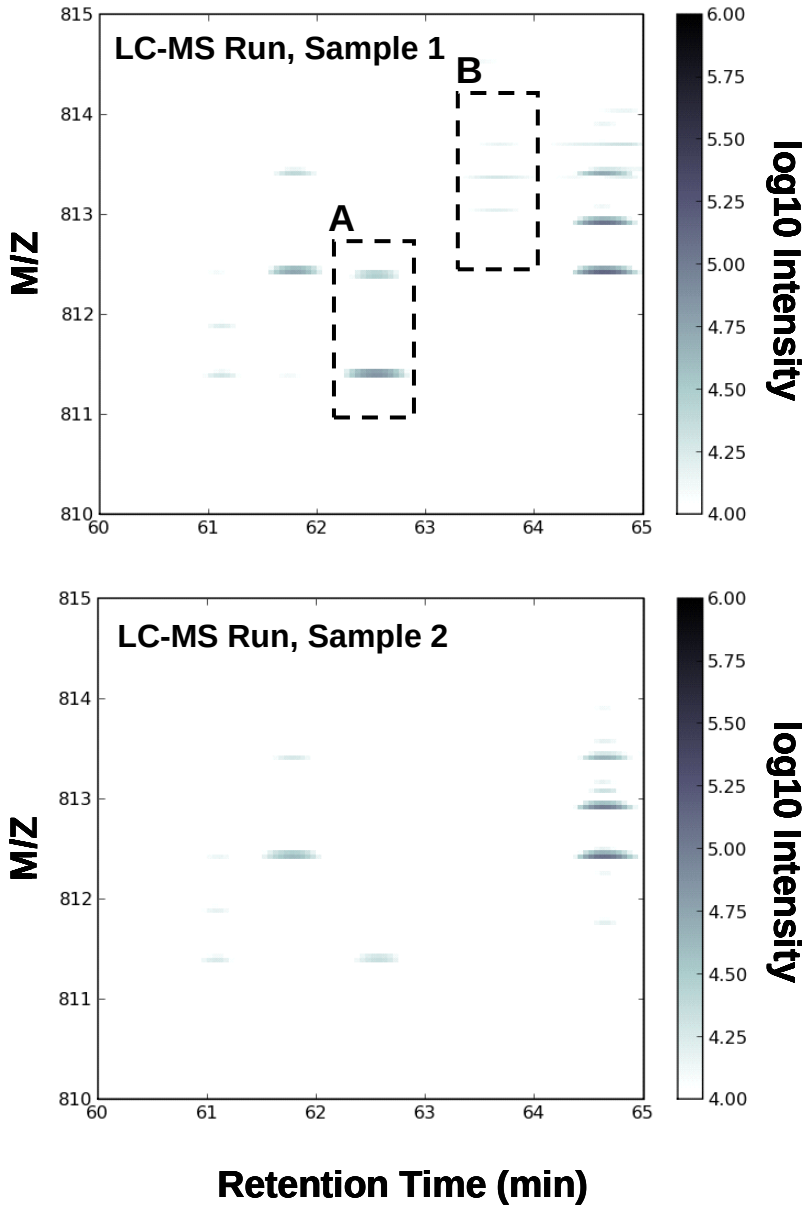

Figure 1-3: µLC-MS Signal Maps from Mean Abundances of Two Classes

A window in retention time and m/z of two µLC-MS runs from separate classes is shown above. Intensity of analytes are shown on $\log_{10}$ scale. Peptides changing in abundance between the two classes are outlined in boxes A, B.

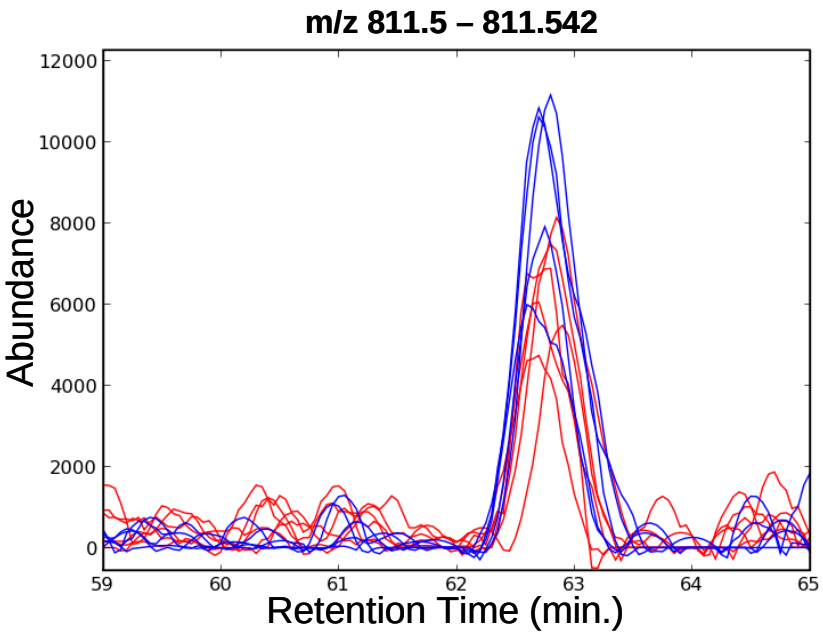

Figure 1-4: µLC-MS Feature Showing Difference in Abundance Between Two Classes

The peptide outlined in Figure 1-3, box B is plotted as a collection of XICs with five replicates from each of sample class 1 (blue) and 2 (red).

1.3. Computational Tools for Label-Free Relative Proteomic Quantitation

The vast quantity of data generated from shotgun proteomics experiments requires computational tools for identification and quantitation of the components. Early efforts were developed for identifying peptides from MS spectra by mass fingerprinting, wherein the precursor mass of a peptide is matched against a database of proteins. However, MS/MS can elucidate the primary structural information of peptides by fragmenting on the peptide backbone. This fragmentation was exploited by the correlation of MS/MS spectra against theoretical spectra derived from a database (Eng et al., 1994). The comparison of theoretical spectra from candidate peptides to the observed spectra has become the basis for most approaches to peptide identification. In cases where a database of predicted or actual proteins is unavailable, *de novo* sequencing of peptides has been developed, although the exact peptide sequence is often uncertain (Horn et al., 2000; Frank et al., 2007).

Software for quantifying µLC-MS analysis of peptides was not developed for almost a decade after the introduction of SEQUEST in 1993; It was first applied to the analysis of stable-isotope labeled proteomics samples with the introduction of the XPRESS software tool (Han et al., 2001), which quantifies the ratio between extracted ion chromatograms (XIC)s of both light and heavy analogues of the same peptide. The RELEX algorithm (MacCoss et al., 2003) uses a least-squares fit approach to filter out interferences from isobaric peptides, and to correct for the small chromatographic shifts from deuterated peptides (Zhang et al., 2001).

The development of readily available high-resolution µLC-MS instruments allowed the accurate detection of peptide isotope distributions (PID) from digested protein data. Only a subset of these features are selected for fragmentation by data dependent MS/MS, making it advantageous to perform relative quantitation based on unfragmented precursor scans.

Moreover, many differential features are either not selected for MS/MS identification by the mass spectrometer, or are not successfully identified as peptides (Finney et al., 2008). Multiple algorithms have been developed to perform relative quantitation on peptide isotope distributions (Mueller et al., 2008). While quantitation of isotope distributions is well-established, drawbacks include the reliance on high-resolution MS data, and cases where a peptide isotope distribution is not present in all replicates being analyzed (Andreev et al., 2007).

Over the last 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) (Li et al., 2005; Bellew et al., 2006; Jaffe et al., 2006; Fang et al., 2006; Andreev et al., 2007; Fischer et al., 2006; Lange et al., 2007; Kohlbacher et al., 2007; Katajamaa and Oresic, 2005; America et al., 2006) or the extracted ion chromatograms (Wang et al., 2003; Smith et al., 2006; America et al., 2006). 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 (Wiener et al., 2004; Meng et al., 2007; Yates et al., 2007; Listgarten et al., 2007; Finney et al., 2008). Using this alternative approach, data is binned into narrow *m/z* and time regions and the mean intensity combined with the variance from replicate analyses are 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.

We were motivated to develop an algorithm for label-free quantitation which would work in cases of low-resolution instruments, where isotopic peaks of peptides could not be distinguished, and also takes advantage of high-resolution instruments. Our approach focuses on using extracted ion chromatograms for detection of regions of *m/z* and retention time which differ in intensity between sample classes, using replicates for statistical power. Chromatographic alignment of all runs in an experiment to a common time frame is necessary to correct for chromatographic differences. The common workflow in quantitative proteomics of finding peptides, extracting intensities, and testing for significant changes is inverted, where differences in feature intensity are detected first, and then annotated with peptide identifications from MS/MS spectra. Fortuitously, this approach is also applicable to samples run on the aforementioned beam-type instruments, and from using the ion trap for continual acquisition of MS/MS data.

1.4. Thesis Overview

This thesis focuses on computational tools to quantify proteins using the ‘shotgun’ mass spectrometry approach, without the use of labeled internal standards or controls. Chapter 2 describes our initial development of experimental protocols and the CRAWDAD software toolkit for label-free quantitation. Briefly, proteins are enzymatically digested to smaller peptide fragments to facilitate separations and mass spectrometric analyses. Multiple biological or technical replicates are prepared at the same time, and run in a serial fashion on the same separation column and mass spectrometer. Software has been designed to minimize variance derived from chromatography and mass spectrometry, identify ‘features’ in the mass spectrometry data that correspond to peptides or other analytes, and find statistically significant differences in peptide abundance level between samples. Our goal is to produce an experimental and computational workflow that performs relative quantitation of proteins between

a number of biological samples, with a minimum of up-front reagent preparation and experimental processing.

A number of updates have been made to the CRAWDAD algorithm defined above. First, chromatographic peak detection was introduced, which has the benefit of improving alignments, and allowing for measurements using areas-under-the curve. Second, we have adapted techniques from the statistical community for estimating the false discovery rate (FDR) of our detected peptides called changing. We compare this to an empirically defined FDR from data with known static and changing peptides. Third, we outline simple heuristics to improve the speed of the retention time alignment.

The computational approach we have used for analysis of peptides in µLC-MS data is generic as it operates on the level of XICs, which can be viewed as time series data. Notably, our approach can be applied to the quantitation of continually acquired MS/MS data. Recent collaborations have applied this approach toward single channel chromatographic data from a quantitative study of yeast amino acids separated by capillary electrophoresis coupled with detection by laser-induced fluorescence (CE-LIF) (Cooper et al., 2010).

1.5. Dissertation Aims

The aims of this dissertation are:

1. To develop and characterize an algorithm for chromatographic time correction between µLC-MS replicate runs, so that features in the run can be compared semi-quantitatively (Chapter 2). This is characterized in a system comparing *E. coli* with IPTG induced vs. a control;
2. To refine the algorithms featured in aim 1 with chromatographic peak detection, and to correctly assign these peaks to groups representing single peptides. The accuracy and

precision of this approach is characterized in a system with an *E. coli* lysate spiked in at varying levels to human proteins from background (chapter 3).

3. To use technologies developed in aims 1 and 2 to determine the accuracy of p-values calculated using parametric and non-parametric techniques, and to assess the effect of biases such as run order and mitigation approaches (Chapter 4)

2. Algorithm for Label-Free Quantitation of Proteomics Samples

2.1. Introduction

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 (MacCoss and Yates, III, 2001; Wu and MacCoss, 2002). The peptides in the mixture are often separated by microcapillary chromatography (μLC) 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 (Eng et al., 1994). The data are acquired semi-randomly in an effort to maximize the total number of proteins identified by MS/MS spectra. 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. Alternate modes of MS/MS acquisition exist with using pre-targeted lists of compounds of interest and their expected *m/z* and RT range (Jaffe et al., 2008).

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 semi-random and leads to 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 approximately 70% of the peptide identifications are shared between technical replicates (Liu et al., 2004), complicating comparisons between samples using MS/MS spectra acquired with data-dependent acquisition. Furthermore, it can be difficult to quantitatively compare an MS/MS spectrum between two samples, as they will likely be acquired at different regions of the chromatographic peak. 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 last 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) (Li et al., 2005; Bellew et al., 2006; Jaffe et al., 2006; Fang et al., 2006; Andreev et al., 2007; Fischer et al., 2006; Lange et al., 2007; Kohlbacher et al., 2007; Katajamaa and Oresic, 2005)or extracted ion chromatograms (Wang et al., 2003; Smith et al., 2006; America et al., 2006). 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 (Wiener et al., 2004; Meng et al., 2007; Yates et al., 2007; Listgarten et al., 2007). Using this alternative approach, data is binned into narrow *m/z* and time regions and the mean intensity

combined with the variance from replicate analyses are 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 (Fraga et al., 2001; Johnson et al., 2003). 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. Non-linear 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 timescale. Dynamic time warping (DTW) is an established technique for determining non-linear shifts between time series of data. DTW determines an optimum mapping of timepoints across two data series that maximizes the value of a similarity function (Sankoff and Kruskal, 1983), and this mapping can be used to correct, or *align* one run's data to the timeframe of another. DTW has been applied successfully to µLC-MS data to determine retention time shifts to correct for chromatographic variance between replicate runs (Wang and Isenhour, 1987; Tomasi et al., 2004; Prakash et al., 2006a; Sadygov et al., 2006; Prince and Marcotte, 2006).

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. We have developed a suite of tools which 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 (Hu et al., 2005; Olsen et al., 2005; Yates et al., 2006) 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 to not only 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 towards the identification of membrane enriched proteins with altered abundance in E. coli upon induction of the lac operon using the nonhydrolyzable allolactose molecular mimic isopropyl-beta-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.

2.2. Materials and Methods

2.2.1. Sample Preparation

Escherichia 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 minutes to induce expression of the *lac* operon. Each sample was pelleted and lysed at 1000 psi in a French press in PBS buffer (pH 7). Membrane fractions were obtained by spinning the sample at 100,000xG in a bench-top ultracentrifuge for 1hr at 4℃. An aliquot of 500 µg equivalent 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 Corporation, Milford, MA) in 50 mM ammonium bicarbonate. Proteins were reduced, alkylated, and digested with trypsin as described previously (Klammer and MacCoss, 2006).

2.2.2. 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 reverse-phase column coupled to an LTQ-Orbitrap mass spectrometer. A 5 µg aliquot of each sample was loaded onto a 75 µm ID 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 (Klammer and MacCoss, 2006). 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 minutes of 95% buffer A during the loading of the sample onto the microcapillary column, followed by a 68 minute gradient of 5 to 35% buffer B, and a 5 minute gradient of 35 to 85% buffer B. The solvent composition was kept at 85% buffer B for 2 min. The column was then re-equilibrated with 95% buffer A for 18 minutes. Spectra were acquired using a cycle of 1 high

resolution MS scan (60,000 FWHM at 400 *m/z*) in the Orbitrap mass analyzer followed by 5 data-dependent MS/MS scans at low resolution in the LTQ linear ion trap.

2.2.3. Software Overview

We have created a suite of software tools, CRAWDAD (Chromatogram Retention time Alignment and Warping for Differential Analysis of Data), which aligns replicate LC/MS chromatograms in the retention time dimension, detects features of differential intensity between two biological samples, and incorporates peptide identifications from MS/MS spectrum identification software. The sequence of operations of the software is outlined in Figure 2-1, and the application of CRAWDAD for relative quantitation of E. coli membrane-enriched fractions is described below. Software was developed using Python 2.3 on Linux and Windows XP operating systems. The psyco library (Rigo, 2004) for just-in-time compilation of Python bytecode to machine code was used to speed execution, and the Numeric library was used to improve the storage and manipulation of large arrays. Each discrete step outlined in Figure 2-1 can be submitted as a job to a queuing engine (e.g. Sun Grid Engine) to speed execution using a computing cluster. An XML configuration file was used to define the sample groups, their respective µLC-MS data files, and the processing steps and relevant parameters. The matrix2png program is used in some data visualization steps (Pavlidis and Noble, 2003). Information about obtaining and using CRAWDAD for nonprofit use can be found at http://proteome.gs.washington.edu/software/CRAWDAD.

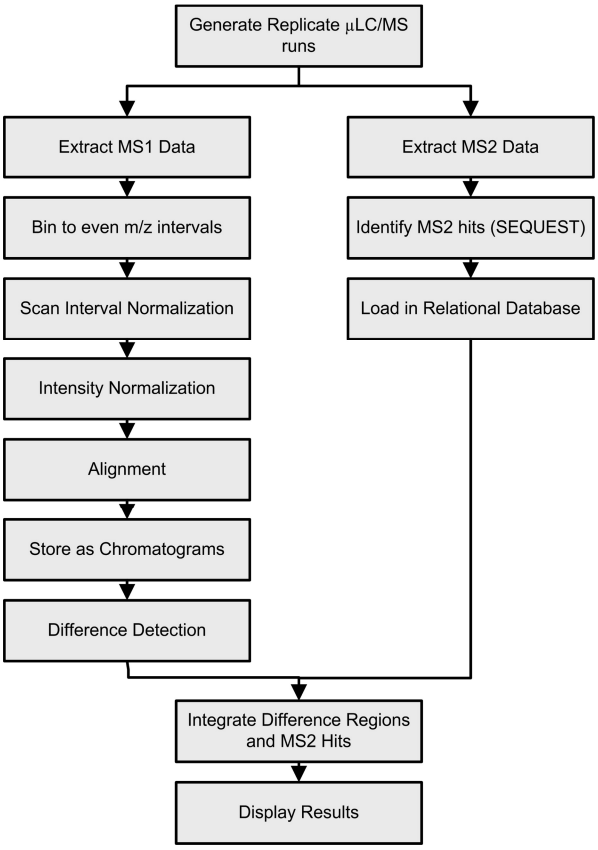

Figure 2-1: Strategy for computational label-free analysis of μLC−MS runs using CRAWDAD

2.2.4. Signal Extraction and Formatting

High resolution Orbitrap MS spectra were stored as MS1 files and low resolution ion trap MS/MS scans were stored as MS2 files (McDonald et al., 2004) using the MakeMS2 program (available freely from http://proteome.gs.washington.edu/software/makems2). The intensity values from the MS scans were binned to 0.1 *m/z* units, and the retention time intervals were re-sampled to a constant 0.01 minutes using linear interpolation between observed points (Myers et al., 1980). The data in the resulting (*m/z*, RT) bins were stored in the MSMAT binary file format to speed data retrieval (http://proteome.gs.washington.edu/software/CRAWDAD/). Runs from these analyses were truncated to include only the signal rich retention time regions between 20 to 95 minutes.

2.2.5. Signal Normalization and Smoothing

The total signal intensity of the target run was used to normalize the global intensity of all other runs from an experiment. Let α and β are the summed intensities across all scans for µLC-MS runs A and B respectively; run A was normalized respective to run B by multiplying all intensities from the (*m/z*, RT) bins by the ratio β/α. After intensity normalization against a common run, µLC-MS runs were smoothed in the RT domain by applying a 11-point second order Savitsky-Golay filter (Savitzky and Golay, 1964; Gorry, 1990) to the extracted ion chromatograms (XICs) from each *m/z* bin.

2.2.6. LC-MS Retention Time Alignment and Warping

We have implemented a derivation of dynamic time warping (DTW), for the alignment of µLC-MS analyses to a common template (Prince and Marcotte, 2006).The direct comparison of the intensity signals between two or more runs requires that signals at the same (*m/z*, RT) coordinates correspond to the presence of the same analyte across runs. To satisfy this condition, we aligned and warped µLC-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 µLC column. We define R_a, R_t, and R_w as the vectors of retention times for scans respectively from scans in A, T, and A_w. Our algorithm produces A_w from A to maximize the similarity of scans in A_w 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 down-sampled 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.

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 dot-product of two scans expressed as intensity vectors (Wan et al., 2002) 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(Sakoe and Chiba, 1978)) 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(eqn. 1).

$$P(i,\ j) = \max\begin{Bmatrix} P(i-1,\ j) + w_1 S(i,\ j) \\ P(i-1,\ j-1) + w_2 S(i,\ j) \\ P(i,\ j-1) + w_3 S(i,\ j) \end{Bmatrix} \quad (2\text{-}1)$$

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 between two runs. A local consecutive shift constraint of 5 shifts was used in addition to the weighting scheme above.

Candidate paths to find the optimum path are constrained to begin on the edge bordered by S(M-ε,N), S(M,N), and S(M,N-ε) where M,N are the dimensions of the score matrix, and ε 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_{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 $A_{w,}$ minimizing differences caused by chromatographic variation in the retention time domain with respect to template run T. As described previously by Prince et al. (Prince and Marcotte, 2006), the scoring path is reduced to produce a one-to-one mapping of timepoints in R_t to $R_{a.}$ A bicubic spline function is fit using the curfit function from the DIERCKX curve fitting package (Dierckx, 1993; Dierckx, 1982), using the scan similarity score values from the score matrix as weights and a smoothing factor of ½ the number of scans. For every time point t_i in the template run R_t, 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_{wi} at time point t_i. At the end of the warping process, the set of aligned scans are stored in an MSMAT file as extracted ion chromatograms (XICs) rather than spectra to simplify the analysis and to speed the detection of statistical differences in chromatographic peaks.

2.2.7. 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 Hardklör (Hoopmann et al., 2007). Briefly, peptide isotope distributions were detected with a signal-to-noise ratio ≥ 3 and conservative correlation score of ≥ 0.99. To remove redundancy and eliminate PIDs that did not persist chromatographically over time, only PIDs that were within 10 ppm in 3 or more consecutive scans were defined as a persistent PID. The criteria 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 minutes. 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.

2.2.8. 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 minutes; approximately half the length of a chromatographic peak in this dataset. A receiver-operator-characteristic (ROC) area under the curve(Metz, 1978) 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 µLC-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.

2.2.9. Mapping Difference Regions to Peptide Identifications

Fragmentation spectra acquired at low resolution in the LTQ linear ion trap were searched using SEQUEST (Eng et al., 1994) 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 (Storey and Tibshirani, 2003) to each spectrum identification using the program Percolator, a semi-supervised machine learning algorithm (Käll et al., 2007). Spectra matching to peptides with a q-value less than or equal to 0.005 were retained.

A difference region was annotated with an 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.

2.2.10. 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:

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

where R_{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 diff 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 by 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 semi-quantitative or simply “different” unless demonstrated otherwise.

2.3. Results

As discussed above, a primary source of error in replicate µLC-MS analyses is variation in chromatographic retention time (Sinha et al., 2004; Synovec et al., 2003). Unfortunately, even the most reproducible microcapillary liquid chromatography separation can have errors equal to or greater than the chromatographic peak width (Mason et al., 2005). Figure 2-2 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 standard deviation (SD) of 0.23 min, comparable to the width of a chromatographic peak, across the 12 different µLC-MS analyses (6 induced and 6 control) for a defined group of 482 peptide isotope distributions detected in all runs using the program Hardklör (Hoopmann et al., 2007).

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 6 E. coli IPTG-induced and 6 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 pair wise 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 2-3 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 peptide isotope distribution markers using Hardklör as described in methods. The standard deviation (SD) of the retention time for these markers before and after alignment is plotted versus retention time in Figure 2-4. The mean retention time 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 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 *m/z* 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 standalone 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 µLC-MS analyses. The extracted ion chromatograms (XIC) in Figure 2-5 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 of intensity bins before and after alignment across replicate µLC-MS runs. In this case, the signal is the mean intensity at each bin and the noise is the standard deviation from the 6 IPTG-induced replicates. Prior to alignment, the region from the induced µLC-MS runs show a signal-to-noise of 1.42 (RSD = 70.4%), while after alignment, the signal-to-noise improves approximately five-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 5,920 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 2-1). The complete output of the difference regions is available as part of supplementary materials. Notably, 93% of the difference regions completely distinguish the IPTG-induced and uninduced samples with an ROC value of 1.0. The $\log_{10}$ of the intensities from the difference regions of the induced and uninduced analyses are plotted against each other in Figure 2-6. Differences were detected over an intensity range which spans greater than 4 orders of magnitude. Difference regions were annotated by association with MS/MS spectral identifications as outlined above. Only 22.8% of the difference regions found in the aligned dataset 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 µLC-MS regions selected for MS/MS fragmentation.

In total, 753 proteins were qualitatively identified using SEQUEST and post-processing with Percolator (Kall et al., 2007). These proteins were identified using conservative criteria (peptide spectrum match q-value < 0.005, ≥ 2 peptides per protein) from the 12 reverse-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 2-2). 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 2-2).

Proteins with an altered abundance between the two samples were organized by Gene Ontology (GO) biological process and cell component classes using GoMiner (Zeeberg et al., 2003). For subcellular localization purposes, additional annotations were used from UniProt (The UniProt Consortium, 2007), EcoCyc (Keseler et al., 2005), and ePSORTdb (Rey et al., 2005). A division of changing proteins by sub-cellular localization is summarized in Figure 2-7. Notably, all 11 proteins changing in abundance levels which 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 w/ 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.

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 TVINQVTYLPIASEVTDVNR peptide from periplasmic oligo-peptide binding protein (OppA) was detected with difference regions corresponding to the *m/z* values of its monoisotopic peak through M+5 isotope peak (Figure 2-8). 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).

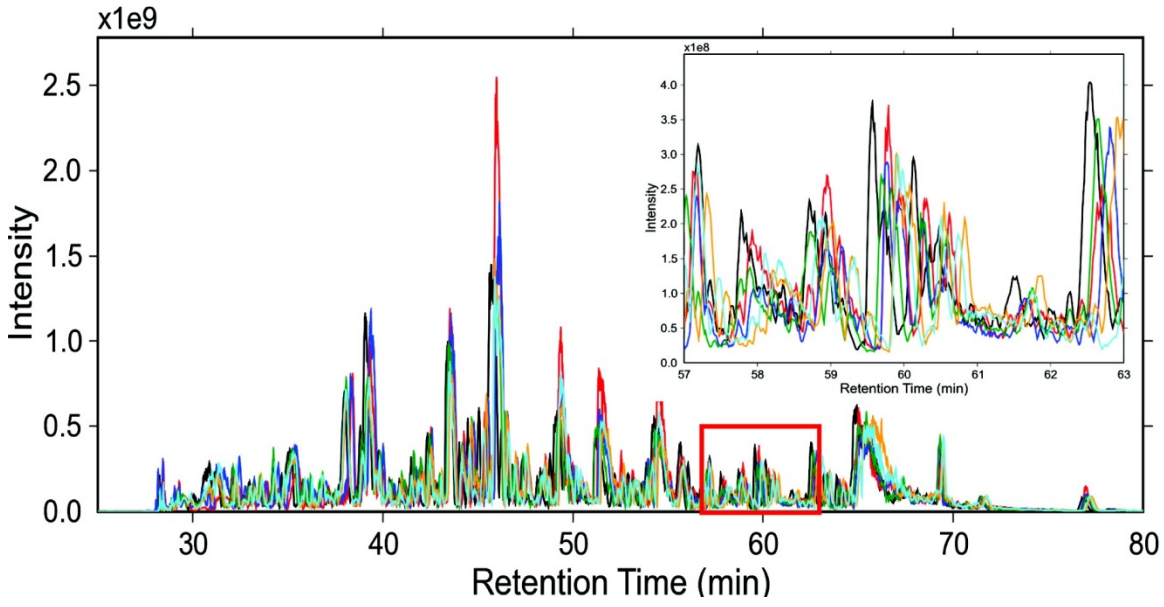

Figure 2-2: Chromatographic variability of technical replicate µLC−MS runs.

Run-to-run variation in chromatographic retention time results in peaks shifted over a range of retention times, making signal-based quantitation on a common time scale problematic. The variability is apparent in a base peak plot of six technical replicate µLC−MS runs from the IPTG-induced sample displayed over a RT range of 20−80 min. Individual replicates are shown in distinct colors. The inset plot displays the region of the unaligned replicate runs from 57 to 63 min.

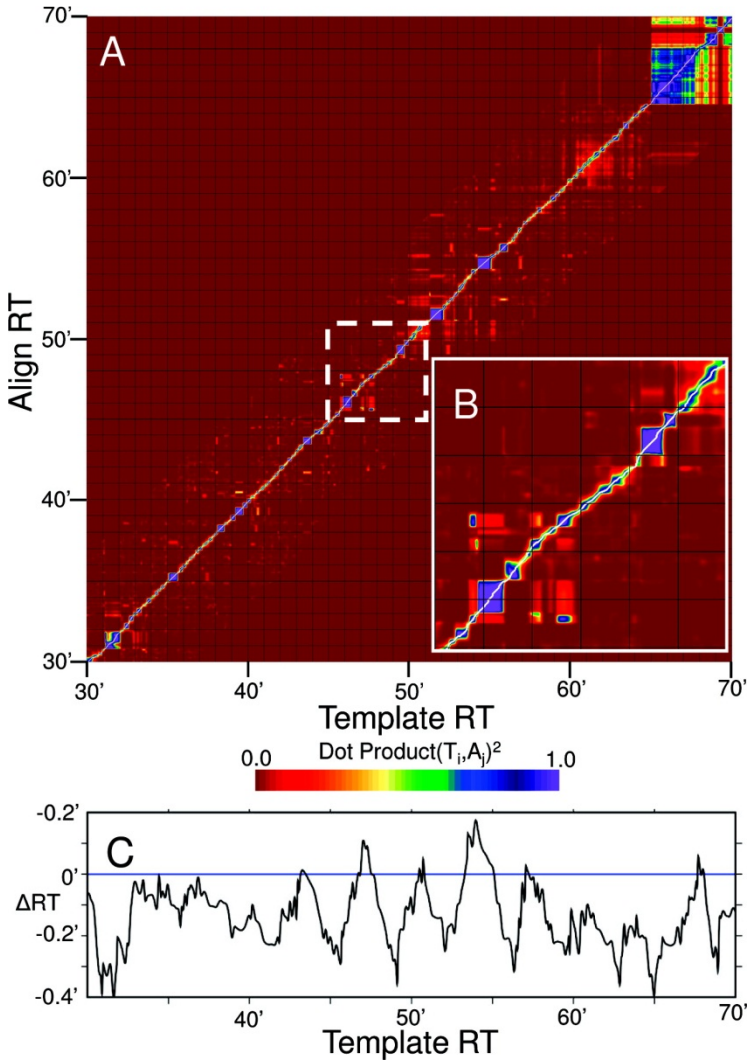

Figure 2-3: Scan similarity score matrix and chromatographic alignment path.

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.

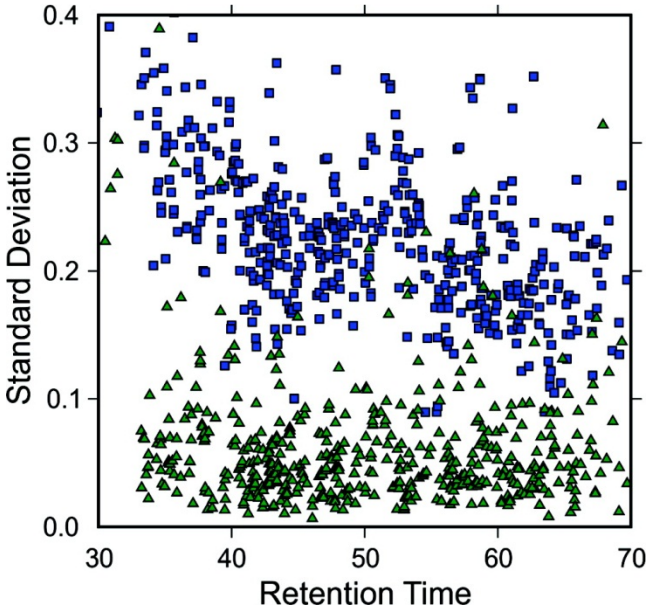

Figure 2-4: Improvement in chromatographic reproducibility using CRAWDAD.

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.

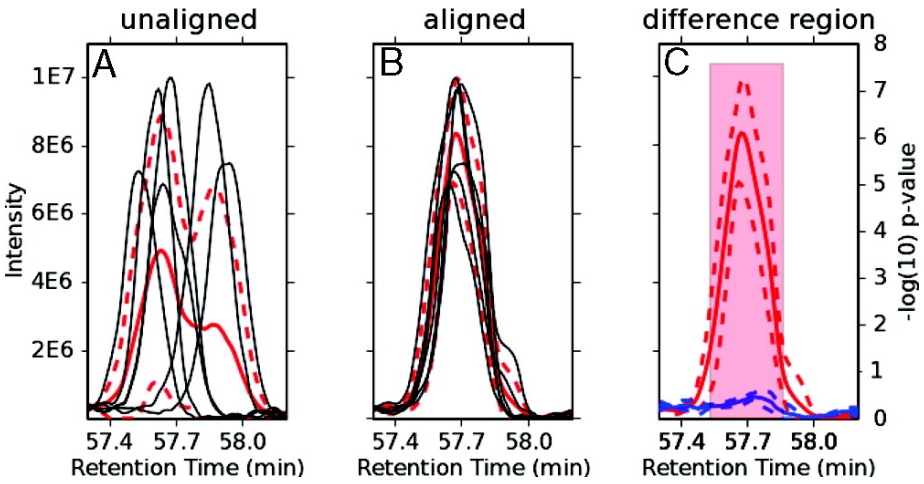

Figure 2-5: Chromatographic alignment revealing a differentially expressed feature.

Pre- and post-alignment *t*-test detection of a difference region from the +2 ion of the LWSAEIPNLYR 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 LWSAEIPNLYR 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.

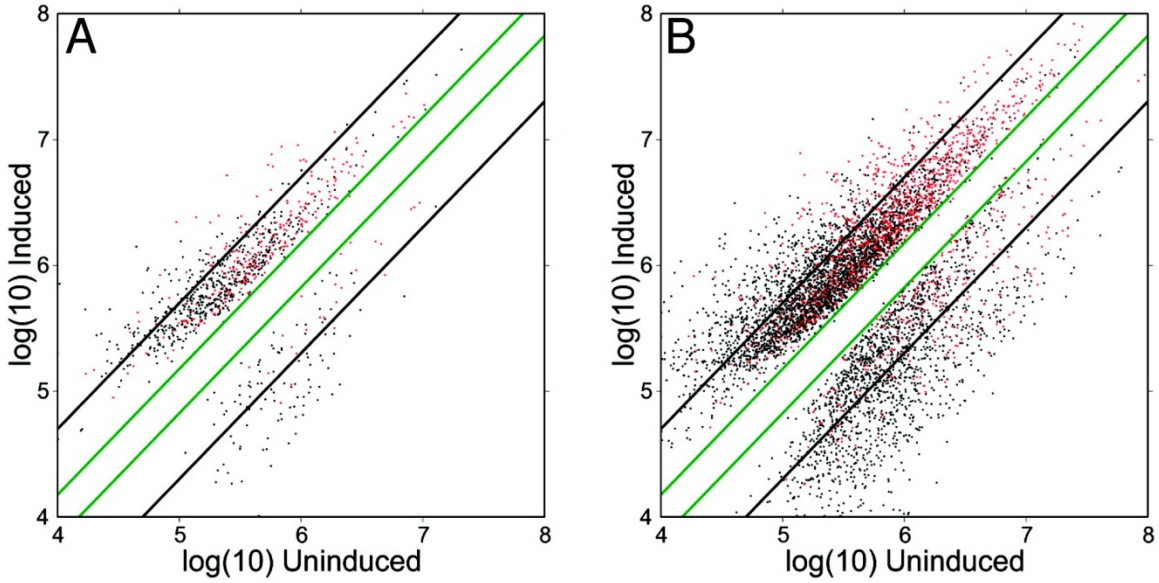

Figure 2-6: Distribution of difference region intensity values before and after alignment.

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. Panel

A shows difference regions detected before alignment, and panel B those detected after alignment. Lines indicating changes in abundance of 1.5-fold and 5-fold are overlaid on the plots.

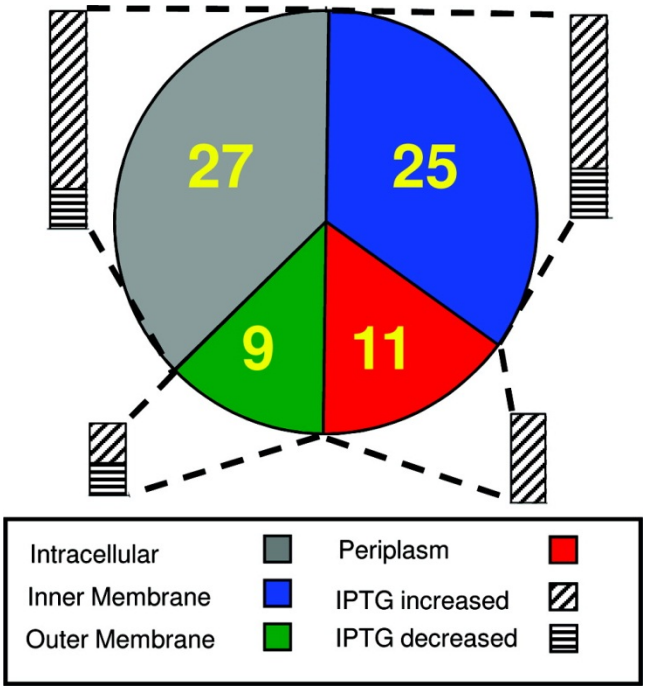

Figure 2-7: Summary of subcellular localization of differentially abundant proteins.

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.

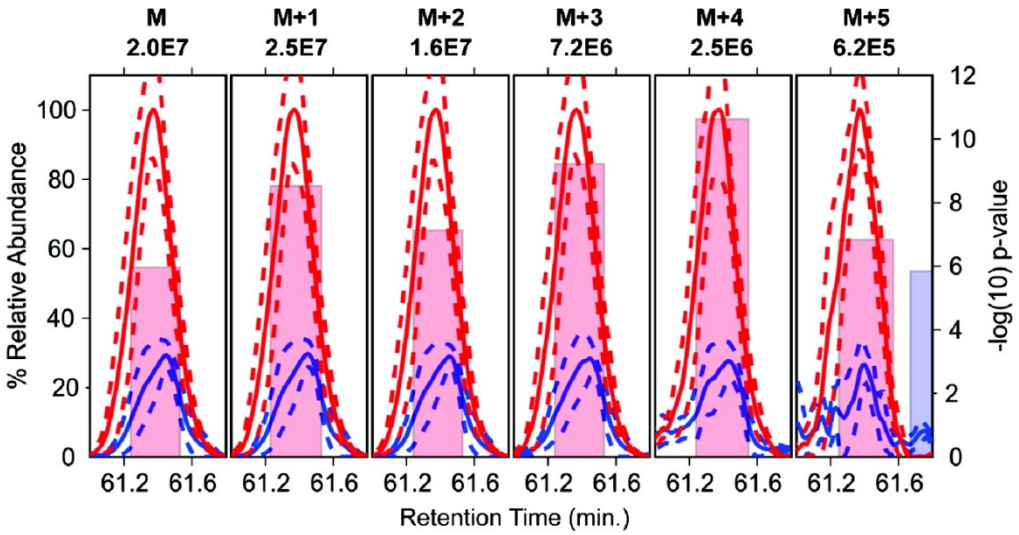

Figure 2-8: Consistent differential abundance ratios across isotopic peaks of a peptide

Reproducible difference ratios across a peptide isotope distribution isotopic peaks (M−M+5) of the +2 ion of the TVINQVTYLPIASEVTDVNR peptide from the periplasmic oligopeptide-binding protein precursor of *E. coli* (*OppA*) shown as ion chromatograms. Mean and ±1 SD intensities from the induced (red) and uninduced (blue) runs are shown with solid and dashed lines, respectively. The −log of the lowest *t*-test *p*-value differentiating each isotope mass chromatogram is shown as a red shaded bar. The mean of the abundance ratios between each data series for the isotopic peaks was 3.47 with a coefficient of variation of 3.7%.

Table 2-1: Summary of difference regions detected and mapped to MS/MS spectra

Bin Size	Aligned	Number of Difference Regions (DR) Elevated in Induced	Elevated in Control	Total # DR	# DR Mapped[2]	Mapped DR%
0.1 m/z	Y	4057	1863	5920	1350	22.8%
1.0 m/z	Y	1671	708	2379	733	30.8%
0.1 m/z	N	815	132	947	273	28.8%

Table 2-2: Proteins detected with changing abundance levels.

(1) Number of proteins having two or more peptides detected with changes in abundance. (2, 3) Number of proteins with greater than or equal to three or five peptides detected with changes in abundance

Bin Size	Aligned	Protein Changes[1]	Elevated in Induced	Elevated in Control	≥3 peptides[2]	≥5 peptides[3]
0.1 m/z	Y	95	74	21	37	18
1.0 m/z	Y	72	57	15	32	14
0.1 m/z	N	13	12	1	3	0

2.4. 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.

2.4.1. 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 (Mason et al., 2005), we have shown that even unaligned data with <1% chromatographic RSD complicates the detection of features of differential abundance between the two samples (Figure 2-6). 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 lab 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 2-5), 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 suggests that chromatographic alignment can decrease the relative error in the intensity measurement by approximately 5 fold and increase the detection of difference regions by >6 fold. Thus, using an efficient chromatographic alignment technique is an important and potentially essential step in the implementation of label free quantitation software.

2.4.2. Considerations in Using a Scan-Based Alignment Technique

Our scan based alignment algorithm assumes that analytes that co-elute in a spectrum from one run also co-elute 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 towards matching the more abundant of those components between timepoints if co-elution 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 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 pair-wise fashion. While aligning all data in a pair-wise 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 pre-process the data for the detection of difference regions. A particularly elegant approach to align all analyses at once uses continuous profile models (Listgarten et al., 2007) to align time series data and could be used with CRAWDAD in the future.

2.4.3. 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 which difference regions can be detected. The difference regions detected throughout the µLC-MS analyses span over four 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 2-5). Furthermore, the detection of difference regions from 6 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 two 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.

2.4.4. Considerations in Interpreting Quantitative Proteomics Results

While CRAWDAD can detect peptide and protein differences between samples, a potential caveat in quantitative or semi-quantitative 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 (Anderle et al., 2004; Prakash et al., 2006b). 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.

2.4.5. 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 on low abundance peptides in a mixture (Mayya et al., 2005; Blackler et al., 2006; Xie and Griffin, 2006), these approaches still require extensive fractionation for comprehensive analysis. Methods based on MS/MS acquisition can miss significant changes in intensity derived from analytes which 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, etc...). 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 (Eakin et al., 2007)

2.5. 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 our differential mass spectrometry software for non-profit use can be found at: http://proteome.gs.washington.edu/software/CRAWDAD

3. Differential MS using Peak Detection

3.1. Introduction

3.1.1. Quantifying Chromatographic Peaks in LC-MS Proteomics Data

The high complexity of proteomics mixtures demands a chemical separation step prior to characterization by mass spectrometry. The distinct peptide of a similar *m/z* cannot be resolved solely in the mass spectrometer at a given elution time. As described in Chapter 2, reverse-phase micro-capillary liquid chromatography (µLC) is used as a separation step with retention time (RT) as a proxy for hydrophobicity, and is coupled online to mass spectrometry for analysis. This is an effective means of analyzing the mixture in two dimensions of separation (RT, *m/z*). First, at any given moment, the mass spectrometer records the data eluting off of the column as a spectrum. Second, the effective concentration of any compound is increased as it elutes in a narrow time range. Third, µLC is easily and effectively coupled with electrospray ionization, giving an efficient on-line technique for analysis of the separated mixture. As the mass spectrometer separates analytes by *m/z*, a two-dimensional separation is performed with potentially high resolution in both dimensions, making the analysis of even a complex mixture tractable.

We have focused on the development of CRAWDAD as a generic algorithm for analytic quantitation by detecting peaks in chromatographic data rather than features from MS spectra. This allows us to handle both high and low-resolution mass spectrometry data, as well as alternate acquisition modes such as MS/MS data acquired continuously using a data-independent acquisition mode (Venable et al., 2004; Bern et al., 2010; Wong et al., 2009). We are not dependent on detecting features in spectra that are particular to peptides, allowing us to analyze any molecule detectable with the given chromatography and mass spectrometer

system. Furthermore, we can analyze purely chromatographic data from non µLC-MS data sets such as capillary electrophoresis experiments (Cooper et al., 2010). Chromatographic peaks, rather than ‘difference regions’ which may not correspond to a significant biological feature of the data (Finney et al., 2008), are the basic unit of quantitative comparison in the current version of CRAWDAD.

Replicate measurements are used for the discovery and quantitation of putative differences between biological conditions in our label-free approach. A *sample group* is defined as a set of replicates from a common origin, such as control or treatment set of samples from a common origin. A *peakgroup* represents a set of chromatographic peaks at identical *m/z* and similar retention time which originate from the same analyte, joined within and across sample groups in a single differential experiment. The previously described alignment algorithm is used to correct the retention time of peaks to a common timeframe so that they can be associated between replicate runs and samples to form peakgroups.

This chapter outlines the development and validation of such a toolkit, and its use on label-free mass spectrometry proteomics data (Figure 3-1). In addition, a chromatographic peak detection library was developed and incorporated into other algorithms (MacLean et al., 2010). The quantitative accuracy of detected changes is assessed using a simple proteome spiked in as a standard over the constant background of a more complex proteome. We use non-parametric statistical approaches to calculate p-values and q-values for differential changes in the *peakgroups* which reflect changes in MS1 ion abundance. The *peakgroups* are associated with peptides identified by MS/MS by the corrected retention time of the MS/MS acquisition event. Peptides that are differentially abundant between sample groups are used to summarize changes in protein level between samples.

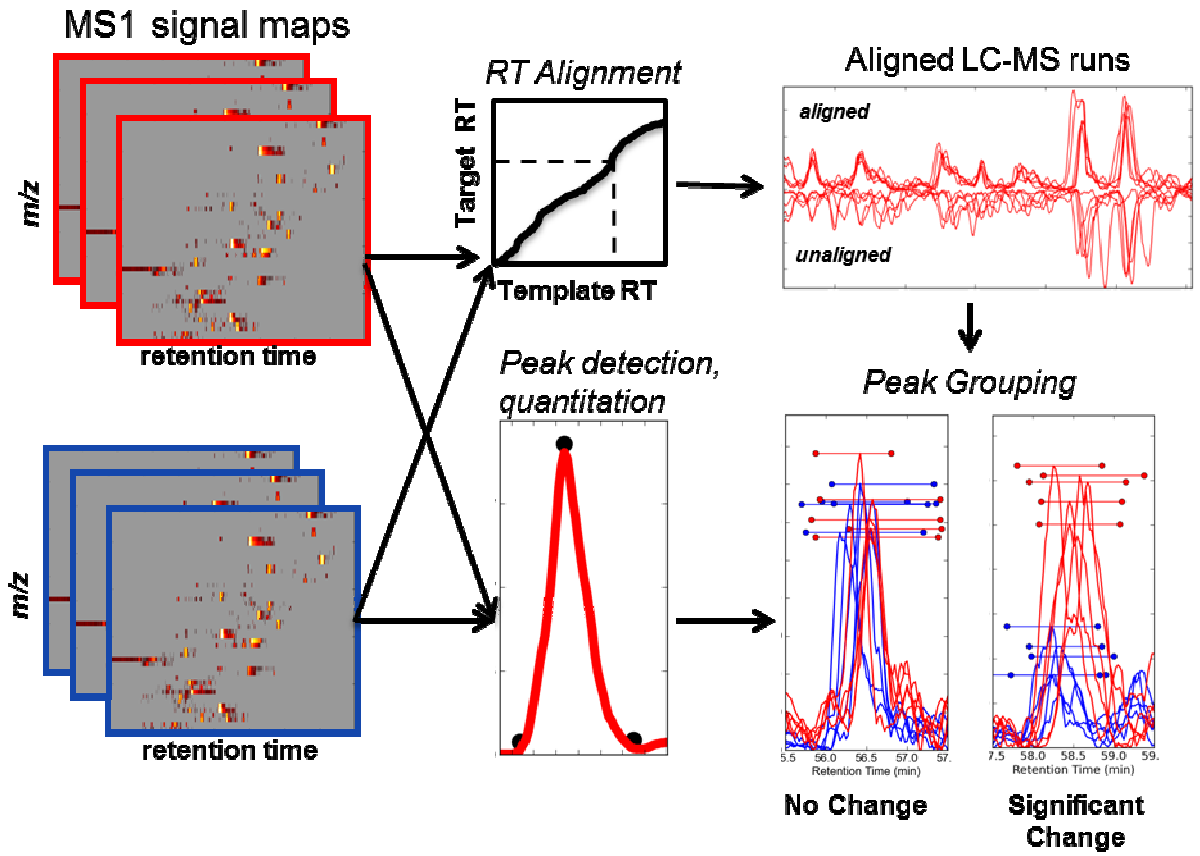

Figure 3-1: Overview of CRAWDAD using peak detection for relative quantitation

Left-to-right: Chromatographic peaks are detected in extracted ion chromatograms from MS1 signal maps of µLC-MS replicate runs from 2 or more groups of shotgun proteomics samples. Independent of peak detection,µLC-MS runs are aligned in the retention time dimension to a common template, giving a point-by-point mapping to correct the data to a common retention time frame. Detected peaks in the same XIC are grouped into peakgroups using the retention time mapping from the alignment. Peakgroups are mapped to peptide identifications from MS/MS data (not shown), and used to summarize quantitative information on a peptide and protein level

3.1.2. Overview of CRAWDAD Updated with Peak Detection

A brief summary of the steps CRAWDAD takes is summarized in Figure 3-2. MS1 spectra are binned, trimmed, and smoothed as described previously (2.2.4 above). Retention time alignment of MS1 signal map images is performed as reported previously (2.2.6 above) with

some improvements to speed and the addition of a novel scoring function using the Spearman rank correlation to compare spectra.

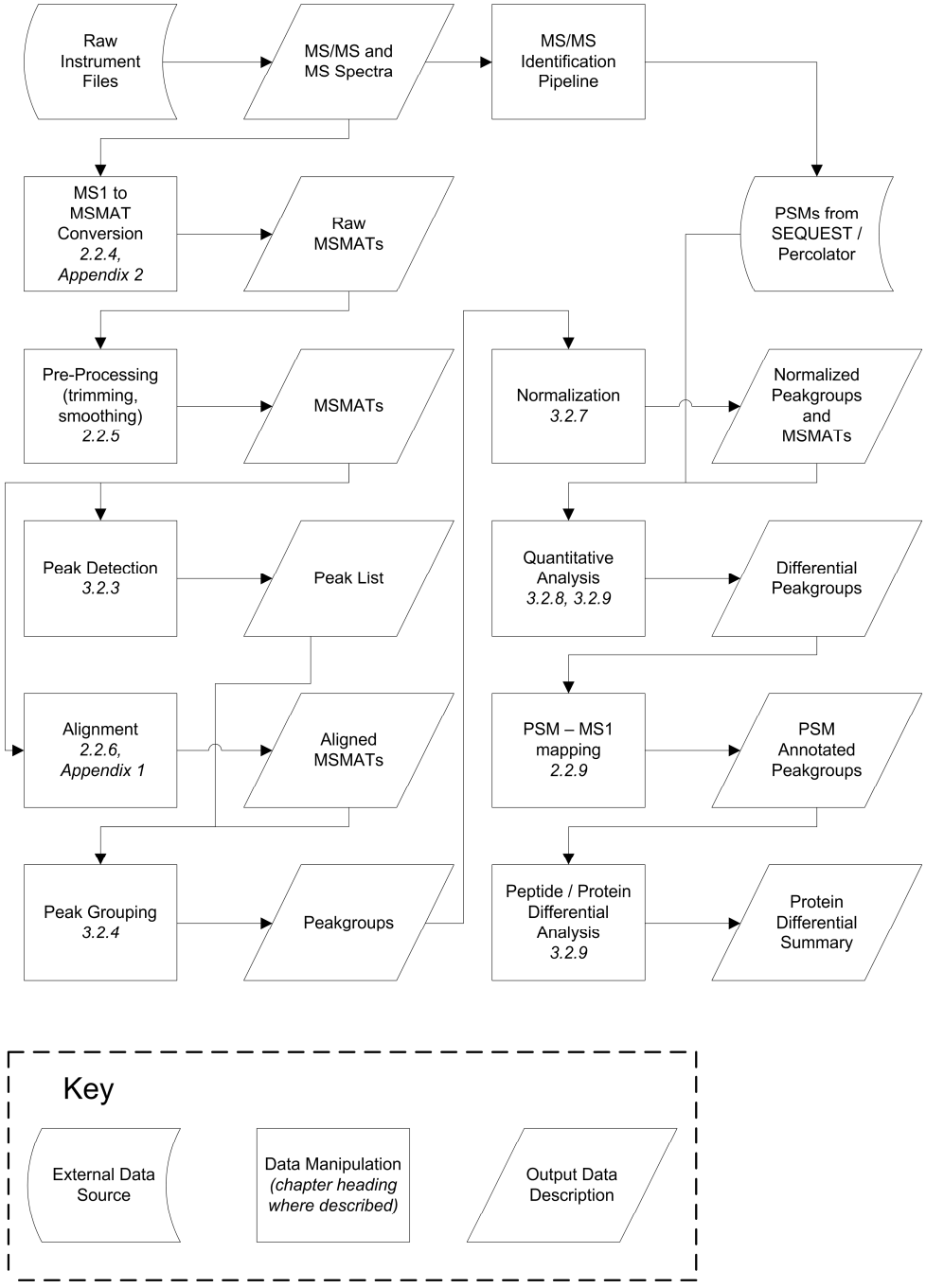

Figure 3-2: Overview of CRAWDAD Software.

CRAWDAD processing begins with raw instrument files as shown in the top left of the data flow diagram above. CRAWDAD finds significant differences in abundances in groups of precursor peaks representing peptides across sets of replicate LC-MS runs. After abundance normalization, these groups of peaks representing a common analyte, or *peakgroup*, are associated with MS/MS peptide identifications. These annotated differecnes in abundance are used to summarize changes in abundance on the peptide or protein level. Note that descriptions of each step can be found in the relevant chapter subheadings. Definitions: MSMAT – a binary MS1 file format (Appendix 2). PSM – Peptide-Spectrum match.

3.1.3. Chromatographic Peak Detection Methods

While much attention has been given in the biological mass spectrometry community to identifying features in spectra (Hoopmann et al., 2007; Kwon et al., 2008; Carvalho et al., 2009), fewer efforts have been applied towards chromatographic peak detection for proteomics. Some approaches reduce the complexity of chromatograms by multiplying adjacent values to remove or diminish noise spikes (Muddiman et al., 1997), retaining those signals over a pre-defined intensity level. Matched filtration, where the observed chromatogram is convolved with a function expected to match the roughly Gaussian shape of the peaks, has been shown to improve signal to noise of peaks in chromatographic data (Danielsson et al., 2002; Andreev et al., 2003). Wavelet methods, where the optimal size of a filter to be convolved with a chromatogram is determined along with the maximally responding position (suggesting the peak location), have also been applied to the detection of μLC-MS chromatographic peaks (Lange et al., 2006; Cappadona et al., 2008).

Accurate detection of peak boundaries is a crucial problem, as it determines the area-under-the-curve (AUC), which is used for quantifying a particular compound. Moreover, accurate peak boundaries are necessary for the proper detection and discrimination of background area and peak area. Given a fixed peak boundary, we compare two simple methods for the estimation of peak area and background area (3.2.3 below) from the accuracy of the results with simulated data (3.3.1 below).

A simple yet robust approach for peak boundary detection is to determine an appropriate signal-to-noise level and define peak boundaries by the regions where the signal increases above the cutoff (Horn et al., 2000). More complex approaches determine an appropriate function or combination of functions to model peaks by fitting the data to such (Li, 2002; Lan and Jorgenson, 2001). We adopt an approach where peaks are detected by the convolution of chromatograms with filters based on the 2nd derivative of a Gaussian peak (Danielsson et al., 2002) and boundaries are refined by use of a smoothed first derivative.

Signal noise arises in proteomics data from sources such as chemical noise, electronic noise, or interfering peptides. While biological variation is the largest source, it is not possible or relevant to reduce it by technical means so it is not considered further. Additive noise not only increases the variance, but can also result in the underestimation of the degree of change in peak area ratios between samples, as the ratio decreases if both numerator and denominator are increased.

3.1.4. Peak Detection and Cross-Group Peak Assignment

The initial algorithm for the detection of differential signals in CRAWDAD used a parametric t-test between the intensities from replicates of two sample classes at every binned (RT, *m/z*) point in the dataset to detect features that were significantly different between the two classes (Finney et al., 2008). However, this algorithm has two drawbacks. First, the criteria for determining that a region is significantly different was heuristic in nature, as p-values for differential expression were calculated at every timepoint, and ‘difference regions’ were defined as consecutive stretches of individually significant p-values. Examining each (RT, *m/z*) point greatly increased the number of statistical tests, potentially causing false positives from multiple testing. Second, many difference regions found using the original approach corresponded to

regions of extended baseline, portions of long tailing peaks, or to spurious differences caused by imperfect chromatographic alignment, such as misaligning the shoulders of a peak.

Limiting the differential comparisons between groups to the areas of chromatographic peaks both decreases the number of statistical tests, simplifying issues of multiple testing, and increases the analytical relevance of the detected features. We have developed a flexible approach for detecting chromatographic peaks, which we have used to quantify features in replicate runs across sample classes. Detected peaks are then joined across replicates, both within and across sample groups, to define *peakgroups*, which are the fundamental units of quantitative comparison in CRAWDAD.

A frequent problem in quantitative proteomics studies is that of cross-assignment (Andreev et al., 2007), where peaks which correspond to the same analyte must be mapped between runs, even when they may only be detected as peaks in a fraction of the number of runs. Two types of failure may occur: peaks may not be detected or may fall below a noise threshold. Multiple approaches are used to deal with the missing data: setting abundances to zero (Mueller et al., 2007), imputing the intensity from regions of *m/z* and retention time within the boundaries of the peak group in replicate runs which were missing peaks (Andreev et al., 2007), or limiting quantitation to those groups of peaks where all compounds were present. We have developed an approach which takes advantage of our retention time alignment to produce *peakgroups* which represent the quantity of an analyte in detected peaks, and infers the abundance over the retention time range consistent with sibling runs when a peak is not detected.

We examine the use of the retention time alignment to improve matching features between µLC-MS runs. There are multiple strategies for dealing with missing or incomplete peaks in a group of such. Assume two groups of three replicates, where group 1 is composed of runs {A,B,C} and group 2 is composed of runs {D,E,F}. The AUCs for any possibly detected peaks that comprise

the same peakgroup are represented by their lower case values (e.g. *a* for the AUC in run A). Assume that this compound was detected in all runs but E (Figure 3-3). In the *inferred* case, an AUC is calculated by integrating any values over a region corresponding to the peakgroup's retention boundary in the run with missing data. We refer to the inferred area as e_{inf}, and so we compare AUCs of {a,b,c} to {d,e_{inf},f} for detection of any differences between the replicate classes. In the *missing* case, we compare areas {a,b,c} to {d,f}. In the *exclude* case, the candidate peakgroup is dropped from consideration.

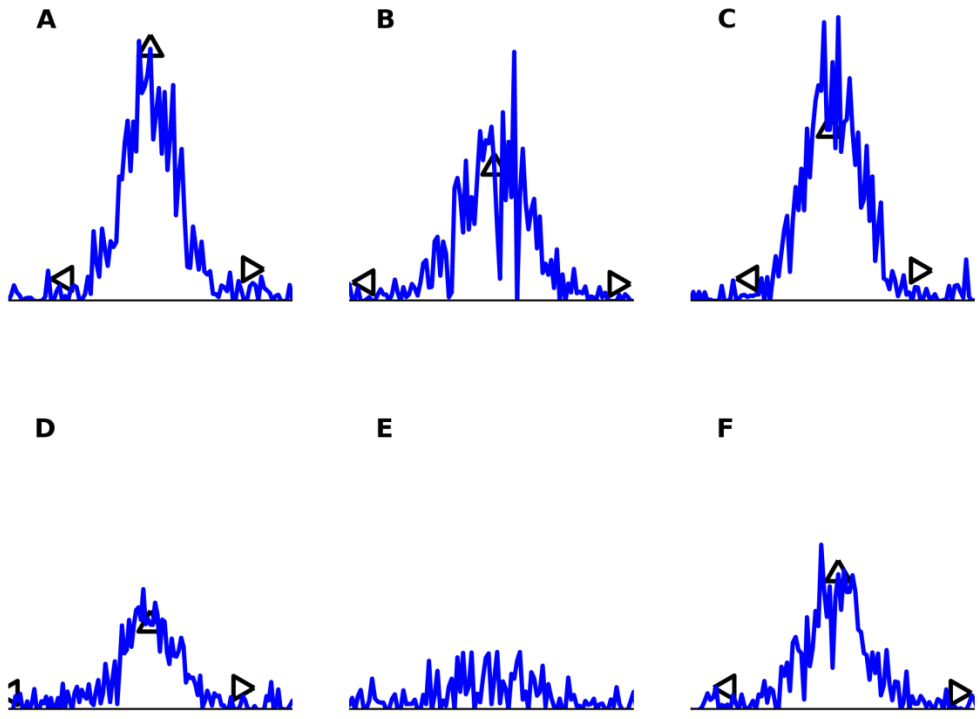

Figure 3-3: Example of six simulated peaks used for grouping.

Each panel represents a region representing a similar compound. Peak apexes are represented by triangles pointing up, and left and right extents by left- and right-pointing triangles. No peak was detected in run E. Two sample groups defined by {A,B,C} and {D,E,F} are compared in CRAWDAD, with the boundaries for the peakgroup defined from the peaks detected in all runs with the exception of E, where the area is integrated over the peakgroup boundary

3.1.5. LC-MS Intensity Normalization by Global Scaling

Normalization computationally removes biases and reduces variation within a group of replicates, giving a better estimate of the true group mean. Normalization also reduces the pooled variance as well, giving greater power to statistical tests such as the t-test or F-test. Both systematic and random errors can add noise to the measured abundance values in proteomics experiments. Multiple sources of analytical error accumulate, such as autosampler loading error, column degradation or partial clogging, and sample degradation while samples are waiting to be run. Abundance normalization estimates the systematic error between runs, and can be used to reduce the variation within a class of replicates, thereby improving the power of statistical tests (Oberg and Vitek, 2009).

Callister et al. compared global scaling with intensity-dependent linear regression, local regression normalization, and quantile normalization (Callister et al., 2006), but did not find a consistent advantage for any method across various data sets. More complex approaches estimate hidden structures/biases in the data using singular value decomposition (Karpievitch et al., 2009) on a per-peptide basis, but require that there be no missing data (e.g. peaks missing from some runs) for normalized peptides.

Internal standard approaches can be used, whether done by adding individual compounds to each sample, or by automated detection of analytes which are not changing between groups in a ‘pseudo-internal standard’ approach (Tabata et al., 2007). Normalization procedures are done by scaling intensities globally, or in an intensity dependent fashion, so that the standards are as close to equal as possible between sample groups. Caution must be applied with a limited number of standards, as matrix effects may vary in different samples.

Our previous method for the normalization of µLC-MS runs assumed that an equivalent quantity of digested protein is analyzed for each µLC-MS run (Finney et al., 2008). A default simple model of normalization is to match and adjust runs by the sum of the total ion current across all scans. This model has the flexibility of not relying on peak detection. Therefore, the total level of ion counts representing the peptides in a sample are compared across runs, and the signal level is globally scaled to match a template run. However, in our comparison of the *E. coli* spike-in series, the level of *E. coli* proteins varies between sample groups. We treat the human peptides as an invariant set of peptides, and apply normalization approaches which assume their abundances should be equal. Multiple approaches to global scaling using the invariant peptides in our mixture as internal standards are compared below.

3.1.6. Measurement of Quantitative Accuracy with a Large-Scale Spike-In of a Complex Mixture

The lack of internal standards in label-free quantitative approaches make numerical measurements of abundance only valid in a relative context, making it difficult to determine the accuracy of a quantitative measurement against a known true value. However, more accurate computational approaches will decrease the coefficient of variation of abundance (measured by background-subtracted AUC) of the same compound when measured across multiple replicates from the same sample class. Using this property, we compare the accuracy of different computational algorithms or parameters to assess a superior version. This is used below to assess both peak detection and grouping approaches, and the quality of alignment approaches. Poor or inconsistent peak boundary detection will introduce variability into the measured peak abundance. Errors in grouping peaks, e.g. erroneously matching between different peptides, would be expected to greatly increase the CV of a group of putatively identical peaks from identical samples. We identify which approaches are most accurate for accurately detecting the

level of change in our samples using the distribution and summary statistics of our estimates of peakgroup CVs for runs from a single ‘sample class’.

The accuracy of a quantitative proteomics experiment can be assessed by the use of spike-ins of analytes of known quantity into a background matrix. Typically, a few proteins or peptides are spiked in, and the measured fold-change in their relative abundance is compared to the known relative change. However, the use of a limited number of spike-ins does not accurately simulate the scale of protein-level changes expected in a biological experiment. Using a conservative estimate that 10% of proteins may show a significant level of change, and the existence of ~100,000 distinct protein species in a mammalian sample (including isoforms, splice variants, etc..), then we could expect on the order of 10,000 proteins changing in level in a complex background – although a large fraction of these will be undetectable due to issues of low abundance, dynamic range, differences in ionization efficiencies, and inaccessibility of some peptide regions to proteolytic digestion. It would be intractable and uneconomic to order or synthesize that number or significant fraction thereof of protein standards for quantitation, particularly for validating a quantitative method.

However, there are convenient sources of thousands of proteins: the proteomes of well characterized organisms with a limited protein complement such as *E. coli* or *S. cerevisiae.* In a spike-in experiment, it is important to ensure that spike-ins are not already contained in the background matrix, which has motivated our choice of *E. coli* as a spike-in proteome against a human background, as *E. coli* proteins have a low degree of homology against the human background matrix (354 of 1,600,525 fully tryptic peptides of length 7 or greater in the *E. coli* proteome). We are using an initial experiment of *E. coli* protein digest spiked into a human epithelial kidney (HEK) cell protein digest background with an eight-fold (8x) difference between

groups as an initial dataset to validate and test the accuracy of computational and statistical approaches as outlined below.

The assessment of peak area is a crucial problem, as it is the measure of abundance of the analyte in each replicate. When using a label-free method which analyzes many samples, we cannot obtain exact quantitation for an analyte by using a known standard. Let $AUC_{c,i,p}$ be the AUC from class *c*, replicate *i,* and peak *p*. The set of peak areas $AUC_{c,i,p}$, over all replicates *i* which are members of sample class *c,* are used for the summary statistics $\mu_{c,p}$ and $\sigma^2_{c,p}$ to characterize the peak's area and standard deviation in the replicates for a single sample group.

A peak area measurement is described by the model $AUC_p = \alpha_p + \epsilon_{exp} + \epsilon_{comp}$, where α_p is the true area, ϵ_{exp} is error derived from experimental apparatus, and ϵ_{comp} is error derived from the computational methods used to estimate peak area. Both $\epsilon_{exp}$and ϵ_{comp} are proportional to the signal level (Anderle et al., 2004), so we express the quality of the measurements as the coefficient of variance (CV) $\frac{\sigma_{c,p}}{\mu_{c,p}}$ which is inherently normalized by the mean peak abundance of the peakgroup.

We assume three main sources of computational error affect peak area and the composition of peakgroups. First, errors may occur in the characterization of the boundaries, area, and background noise of peaks. Second, if the alignment for the particular *m/z* being analyzed is not optimal, it will increase the probability of peaks being incorrectly grouped. Third, false positive or false negative peaks can result in adding noise rather than actual peak abundances to a group of replicate measurements. Improvements in our computational methods should result overall in a decrease in CV for peakgroups in an experiment. Therefore, we evaluate different algorithmic approaches or parameters – such as minimum peak area – by ranking the results according to the CVs generated from measurements within the same replicate class. Moreover, we assess

the quantitative accuracy of the approach for the measured spike-in ratios on the level of both the ratios of the peakgroups between groups and the final protein ratio between groups.

3.1.7. Multiple Testing and FDR Considerations

We use label-free shotgun quantitative proteomics to answer the following question: which peakgroups, peptides and proteins show a statistically significant change in abundance between two or more samples. This requires the use of a large number of statistical tests. Our approach also increases the number of tests beyond the number of detectable peptides, as we test all detected groups of µLC-MS peaks (peakgroups) to determine ones that can be mapped to peptide abundances. Statistical control of the specificity of peakgroups called significantly different using a two-tailed standard p-value cutoff of 0.05 results in a large number of false rejections of the null hypothesis when applied over multiple tests.

We examine multiple peakgroups for differential abundance, and take issues of multiple testing into account. For example, using a significance threshold of 0.05 when analyzing 20,000 peakgroups for differential abundance, the expectation value of the number of peakgroups being called different by chance would be 1,000. One approach to limit the number of expected false positives is to control the Family-Wise Error Rate (FWER), which is the probability that *at least one* type I error (false positive) will occur in a population, or *family,* of tests. For example, the Dunn-Bonferroni (Abdi, 2007) correction to control the family-wise significance level (α) calls for dividing α by the number of features being tested. In the example above, this would call for a feature-wise significance level of 0.05/20000 (2.5E-6) in order to control for one false change found – which will greatly reduce power in differential experiments.

However, we are not attempting to create a list of differentially expressed peptides or proteins with no false positives, but rather want to give users a list of differential features for

experimental follow up though likely to be significant. Under these conditions, we wish to control the expected number of false positive peakgroups which are called significantly different, or *discoverie*s. Let *R* be the set of features for which the null hypothesis has been rejected – or our assertions that a change in abundance exists. Within *R* exists *V,* the number of features for which the null is true (false positives), and *S*, the features for which the null hypothesis was correctly rejected (true positives), with *V*+*S*=*R.* Let the proportion of our assertions which are false, *V*/*R*, be defined as *Q*, and the expectation value E(*Q*) is defined as the false discovery rate (FDR) (Benjamini and Hochberg, 1995)

In standard experiments where we do not know which comparisons are null beforehand, we can calculate the q-value using statistical approaches based on the distribution of p-values (Storey and Tibshirani, 2003). P–values from truly null data should be consistent with data from a uniform distribution from 0 to 1, and detecting this proportion indicates the proportion of the data which is truly null. When some proportion of the data is known to be null, such as with the human peptides in our HEK/*E. coli* spike-in dataset, an empirical FDR can be easily calculated from the proportion of null features among those called significantly changing.

3.2. Methods

3.2.1. Sample Preparation and Processing

Human epithelial kidney (HEK) cell line (HEK293) samples were grown in DMEM supplemented with fetal bovine serum and antibiotics on 15cm plates at 37℃ to 80% confluency. Cells were separated from the plates by exposure to trypsin, lysed in a HEPES-based buffer, and protein levels quantified with a Bradford assay kit (Bio-Rad, Hercules, CA). Proteins were reduced, alkylated, and digested to peptides as described in 2.2.1 above.

Peptides were bound to a solid-phase Oasis MCX column (Waters, Billerica, MA) to remove salts, phospholipids, detergents, and neutrals. Peptides were eluted following the manufacturer's protocol with minor modifications. Briefly, the cartridge was conditioned using 1 ml methanol, 1 ml 10% ammonium hydroxide in H_2O, 2 ml methanol and finally 3 ml 0.1% formic acid in H_2O. The samples were then loaded onto the cartridge and washed with 1 ml 0.1% formic acid in H_2O and 1 ml of 0.1% formic acid in methanol. The peptides were eluted from the cartridge with 600 µl 10% ammonium hydroxide in methanol, collected in an eppendorf tube and evaporated using a SpeedVac (Labconco, Kansas City, MO) set to 50℃.

E. coli K12 MG1655 wild-type strain was grown in LB media to an OD_{600} of 1.0. The cells were spun at 10K rpm, 4℃ for 10 minutes and the supernata nt was discarded. Cells were resuspended in cold 50 mM ammonium bicarbonate at pH 7.8 and lysed via sonication. The lysate was then spun at 4K rpm, 4℃ for 10 minutes to r emove debris and at 14K rpm, 4℃ for 10 minutes to separate soluble and insoluble lysates. The soluble fraction was isolated, and protein concentration was measured using a BCA protein assay (Thermo-Fisher, Rockford, IL).

The HEK protein extract and the *E. coli* lysate were digested separately as follows: samples were solubilized with Rapigest (Waters Corporation, Milford, MA) to a final concentration of

0.1% and boiled for 5 min. The samples were then treated with 5 mM DTT at 60 ℃ for 30 minutes to reduce disulfide bonds. The free sulfhydryls were alkylated with treatment of 15 mM iodoacetamide at room temperature for 30 minutes. Trypsin was added to a final concentration of 1:50 (μg trypsin: μg protein) and the sample digested at 37 ℃ for 2 hours. The trypsin activity was halted and Rapigest was hydrolyzed with the addition of HCl to a final concentration of 200 mM and incubation at 37 ℃ for 30 minutes. The sampl es were centrifuged for 10 minutes at 20,000 × g and the supernatant saved.

A baseline spike-in level of *E. coli* peptide digest into a HEK background was prepared at a level of 4 μl of 1 μg/μl HEK digest, and 1μl of 0.1 μg/μl *E. coli* digest, for a 40:1 HEK/*E. coli* ratio by total protein amount. A sample with 8-fold increased *E. coli* level, labeled as the 8x sample, was prepared by adding 2 μl of 0.4 μg/μl to each sample. Samples were diluted with a 5% acetonitrile buffer to a total volume of 8 μl before analysis. The 1x and 8x samples were analyzed with 5 analytical replicates for each group.

A Waters nanoAcquity LC system was used for chromatographic separation. Peptides were separated at a flow rate of 250 nl/min over a homemade 35 cm long, 75 μm inner diameter, fused silica capillary column packed with Jupiter Proteo 90A C-12 resin (Phenomenex, Torrance, CA). The mobile phase consisted of buffer A (water, 5% acetonitrile, 0.1% formic acid) and buffer B (acetonitrile, 0.1% formic acid). The gradient used was: 9% buffer B to 36% buffer B for 180 minutes, followed by a 5 minute wash with 80% buffer B and a 15 minute re-equilibration at 9% buffer B. Buffer A comprises the remainder of the gradient at any given time

The LC was coupled to a Thermo Scientific LTQ-FT Ultra by electrospray ionization. The scan cycle used consisted of one full scan in the FTICR (400 – 1,400 *m/z*, 50,000 FWHM resolution at 400 *m/z*, profile mode) followed by five data dependent MS/MS scans of the five most intense

ions in the ion trap. Dynamic exclusion was used with a repeat count of one and an exclusion time of 30 seconds.

MS2 and MS1 files were extracted from the instrument-generated .RAW files using the in-house software MakeMS2.The monoisotopic mass of putative peptide spectra was determined with the Bullseye algorithm (Hsieh et al., 2009). MS/MS spectra from the combined *E. coli*/HEK mixtures were searched against a database composed of the 3.57 release of the human IPI database concatenated with the Genbank March 25 2009 release of the *E. coli* K 12 MG1655 sub-strain genome, and a short list of common laboratory contaminant proteins such as human keratin. A decoy database was generated by shuffling the above protein database. An in-house variant of the SEQUEST algorithm (Eng et al., 1994) was used for peptide-spectrum match (PSM) searches. Searches were performed using a precursor mass tolerance of 10ppm in *m/z* of the precursor mass range, with no enzyme specificity. Peptide-spectrum matches (PSMs) were assigned a q-value using the Percolator algorithm (v1.14) (Kall et al., 2007) using the decoy database for supervised training.

3.2.2. Pre-processing of MS1 data before alignment and differential detection

Each .RAW file generated from the LTQ-FT instrument was converted to an MSMAT file using the in-house written program MakeMS2. MSMAT files, representing the MS1 signals in a binary format, were initially stored as a series of scans. To remove regions of the run stemming from column equilibration, or from slowly eluting compounds at the wash phase, MS1 signals of runs were trimmed to retain signal between 25’ and 170’ of run time (in a 240’ run). The first run of each sample group (1x,8x) was excluded as they contained frequent signal dropouts in extracted ion chromatograms.

Signals were linearly re-interpolated in the chromatographic dimension to have an even time spacing of 0.025’ between scans. Spike-like signals were removed by a heuristic whereby any

signal in a binned XIC must persist for at least 9 scans over a threshold of 1 signal units. A 2nd order Savitzky-Golay smoother (Gorry, 1990) with a window size of 11 was used to smooth data in the chromatographic dimension. Retention time alignment templates and parameters are detailed in section 3.3.2 below.

3.2.3. Chromatographic Peak Detection using Gaussian 2nd and 1st Derivative Filters

Let C_i be an extracted ion chromatogram (XIC) *i* from precursor scans of a micro-capillary liquid chromatography – mass spectrometry (μLC-MS) run, or from any single-channel extract from chromatographic data. A Gaussian filter *G* is defined by sampling from the normal distribution at fixed points using a simplified Gaussian formula $e^{\frac{-x^2}{2\sigma}}$, where the normalization constant is excluded, and x is set to zero at the center of the filter. Let *w* be the width of the Gaussian filter in number of scans, calculated by $w = 4\,(\sigma + 0.5)$. Let $G_f{}'$ and $G_f{}''$ be filters derived from the first and second derivative of the Gaussian function, where the standard deviation is set to *f* in the number of scans, over the interval *w* defined above. Each chromatogram C_i in a run is convolved with $G_f{}'$ and $G_f{}''$ to produce $C_i(G')$ *and* $C_i(G'')$. For ease of use and visualization, we perform calculations on - $C_i(G'')$.

Peaks are found from local maxima in $-C_i(G'')$ between points where the convolved product of the filter and chromatogram crosses the x-axis (Figure 3-4). Preliminary boundaries are set to the local maxima in $C_i(G'')$. To avoid the inclusion of small shoulders or bumps in peaks, a minimum threshold for length – roughly half the width of a chromatographic peak – is applied to produce a list of preliminary peaks. For some applications, the shortened boundaries are sufficient. However, the peaks can be extended by analyzing the convolution product of the chromatogram with the first derivative filter. First, we extend the length of peaks by extending

each peak's boundary from both the leading or lagging end to the point where the first derivative crosses with a positive sign from left to right (Figure 3-4).

The preliminary peak-finding step described above creates a set of overlapping peaks, where conflicts are resolved using a greedy approach. The most abundant peak is first expanded to the new boundary as defined using the 1st derivative approach above and denoted as a 'seed peak'. The overlap *o* between the seed peak and any adjacent peaks it may have overlapped is calculated as a fraction of the length of the seed peak before expansion. If *o* is greater than or equal to a user-specified threshold (0.2) then the overlapped peak is subsumed into the seed peak and deleted. Any unextended peak which is fully encompassed within the extended peak is deleted. Unextended peaks which partially overlap with the extended peak have their boundaries adjusted to the borders of the extended peak. Conversely, a seed peak can only be extended up to the edge of another peak that has been already extended.

After peak detection is performed, peak areas are assessed by estimating the background level in chromatograms by one of two methods (Figure 3-5). In each method the peak area under the curve (AUC) is calculated using trapezoidal integration from point to point of the chromatogram above the estimated background level. Peaks may be optionally retained dependent upon the ratio of the AUC to the background level. The peak detection software has been developed as a C++ library for chromatographic data, and has been included in the Skyline software for analysis of targeted proteomics data (MacLean et al., 2010).

Any peak detection method which does not learn parameters from the data in either a supervised or un-supervised method can only be assessed in the context of a preset group of parameters (Zhang et al., 2009). Below, a variety of peak detection techniques is assessed not only with the simulated data as described above, but also with the impact of a number of peak

picking parameters on the quantitative accuracy of changes detected on the peptide and protein level.

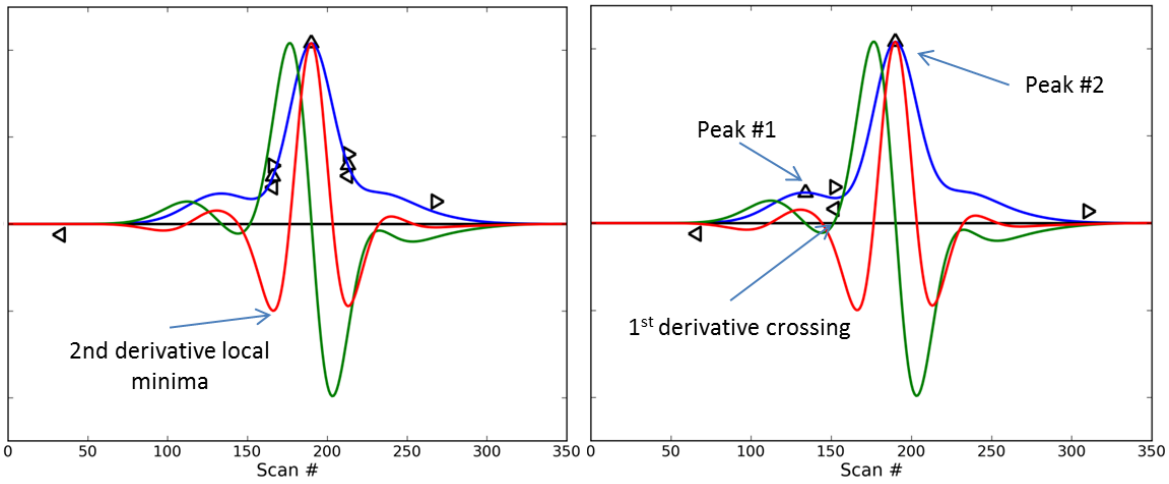

Figure 3-4: Peak boundaries detected using convolution with Gaussian derivative filters

Chromatographic traces are shown in blue, the convolution product with the second derivative filter in red, and with the first derivative filter in green. Each convolution product is scaled to have the same maxima as the original data series for display purposes. Left and right peak boundaries are respectively shown by triangles pointing left and right, and apices by triangles oriented upright. Peaks in the first panel are determined by the boundaries of the convolution with the 2nd derivative filter. The second panel is the result of extending the preliminary peaks by extension until the point prior to the 1st derivative crossing x-axis.

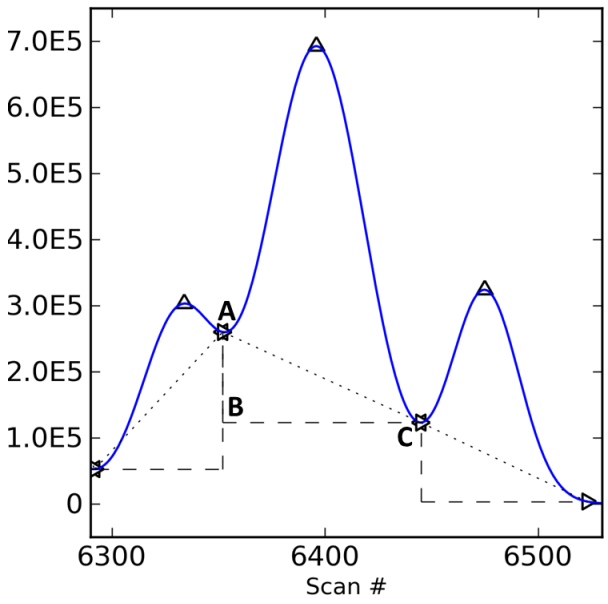

Figure 3-5: Background subtraction methods applied to simulated peaks.

(1) In the *'lower-edge'* background subtraction method, the peak background is assumed to extend from the lower of the two peak extents (C) to the equivalent point below the opposite extent (B), and peak area is calculated using a trapezoidal sum from peak edge to edge from the chromatogram (blue) to the line BC. (2) In the *'edge-to-edge'* method, the peak area is estimated using the trapezoidal rule with the background line from the chromatogram to the line AC. Any region where the chromatogram falls below the background line contributes zero area to the peak area calculation.

3.2.4. Peak Grouping Across Replicate Runs

We approach the problem of cross-assignment of MS/MS identification features between quantitative features in µLC-MS runs by the use of retention time alignment before matching detected peaks across replicates and classes. Peaks are matched only in equivalent XICs using a parameter where the ΔRT between them is within a distance set by default to be approximately half the chromatographic peak length. Rather than warping the entire chromatogram to match a template as in the initial version of CRAWDAD, we use the recorded retention time shifts from the CRAWDAD alignment to in-place translate the RT of the peak for grouping.

We define the set of µLC-MS runs as $M_{g,r}$ where $g \in (\ class1..classN)$ and $r \in (run1\ ..runN)$. A greedy algorithm for defining peakgroups is employed as follows:

1. For a given extracted ion chromatogram c_i from *m/z* bin i, we consider the set of peaks P across all runs $M_{g,r}$.These peaks are then sorted by height, and the most abundant peak is defined as a *seed peak* and is added to the peakgroup.

2. We iterate over all runs with the exception of the run to which the *seed peak* belongs. Within each run we add the nearest of the flanking peaks to the *seed peak* in corrected retention time. If either candidate peak has previously been added to a peakgroup, it is not considered. If no peak from a given run is found within the specified maximum RT range, the run is marked as missing from the peakgroup. Any peak, including the *seed peak*, which is added to a peakgroup is marked as being used.

3. The peakgroup is saved, and a consensus boundary is generated as described below.

4. If no unused peaks across all runs remain, we move on to step 1 for the next extracted ion chromatogram. Otherwise, the next highest peak that has not been selected as a member of a peakgroup is selected as the *seed peak*. Step 2 is repeated, until no matching peak is found from XIC c_i over all the runs.

A consensus boundary for each peakgroup is generated as follows: Let P_a be a synthetic peak constructed from the average of peaks in the group, P_c be the abundance-based centroid RT of P_a, and P_w be the average peak width of the constituent peaks in RT. Let P_r be the average ratio of the width of the left side to the right side, as measured from the apex to left/right boundaries, of all the peaks in the group. The left boundary G_l of the group is defined as $G_l = P_c - P_w P_r$, and the right boundary as $G_r = G_l + P_w$. When no peak can be added to a peakgroup for a given

run, the program has an option to integrate an AUC over the peakgroup boundaries as an *imputed peak*, and add the imputed area to those used in the peakgroup.

3.2.5. Simulation of Chromatographic Peaks and Estimating Quality of Peak Detection

Testing the peak detection library took two avenues of approach. First, we generate simulated peaks with added noise, and assess the accuracy of the peaks. Second, we examine the relative performance of different techniques and options in the peak detection library to find more accurate boundaries for peaks, which is assessed indirectly by a decrease in the coefficient of variance(CV) of groups of peaks.

Simulated Peak Generation

Below, we describe settings for peak locations and sizes as determined by the simulation software. Random values for noise, peak height, or location are drawn from the normal, lognormal, or uniform distributions. Limits on the random values can be specified, where a new random value is drawn if the limits are exceeded. For example, the function $normal(\mu = 0, \sigma^2 = 1, limits = (-0.5,0.5))$ draws a value from the normal distribution with mean 0 and standard deviation 1, constrained to lie within the interval [-0.5,0.5].

Peaks are generated by a stochastic model of peak size, shape, and position. Parameters are defined by the user including peak width (specified as full-width at half-max height (FHWM)) and peak height are drawn from distrLocation is determined from a model of intervals between peaks. Let peak apex locations be defined by the vector **L**. Intervals between peaks are drawn from a distribution, where $l_1 = 0 + D(\Theta)$ and $l_{i+1} = l_i + D(\Theta)$ where $D(\Theta)$ is a random variable drawn from the distribution D parameterized by Θ.

Likewise, peak widths defined by full-width at half-max (FWHM) and peak heights are drawn from distributions parameterized by the user. The peak profile points are based on values sampled from a Gaussian distribution centered at zero, with standard deviation as specified by the user. To avoid sampling over the infinite domain of the Gaussian function, we limit the size of the sampled data to integer values of x over a window of size $2 * floor(4\sigma + 0.5) + 1$ centered on 0.

Noise analogous to that added by the experimental process is also simulated. Experimental noise is added based on shot noise, noise derived from the variation in ionization efficiency or electrospray noise, and chemical noise (Blackler et al., 2006). Shot noise from ion traps is caused by sampling error from the number of ions being detected at the electron multiplier, and is proportional to the square root of the signal. Ionization efficiency is assumed to cause normally distributed noise, and is linearly proportional to the signal. Noise is present from the detectors and other electronics, but is of minimal impact compared to those forms mentioned above, and is modeled by a simple normal distribution.

Chemical noise is composed of contaminants, adducts of molecules (e.g. sodium complexed with an ionized peptide), and analytes present below the limit of detection. We obtained chemical noise at a high intensity level by sampling from ubiquitous Polydimethylcyclosiloxanes present in laboratory air (Schlosser and Volkmer-Engert, 2003). An extracted ion chromatogram from a typical µLC-MS run in our laboratory spanning the *m/z* range of 445-446 *m/z* was used as an optional additive noise component. A background low-frequency variation in the chemical noise chromatogram was subtracted by fitting a LOESS regression curve to the data, setting the mean of the interpolated LOESS fit data to zero, and subtracting that from the original data. The final chemical noise chromatogram had a mean signal of 1.87E+5 and a standard deviation of 4.08E+4, for a relative standard deviation of 21.7%.

Calculating Accuracy and Precision of Peak Library on Simulated Data

While many approaches for scoring the accuracy of a predictive method use measurements of specificity (TN/(TN+FP)) and sensitivity (TP/(TP+FN)), we do not have true negative cases where no peak exists – in this case we only know where peaks are, and in some simulation cases no 'empty spaces' are available. Instead, we use the F1 score (van Rijsbergen, 1979) from the information retrieval field, which is the harmonic mean of precision (TP / (TP+FP)) and recall (TP/(TP+FN); synonymous with sensitivity).

However, we do not only want to assess the accuracy of our methods in terms of accuracy of peak detection, but also the quality of the boundaries estimated for each peak, and any background subtraction method being used. Our main concern in a quantitative method is with the estimated area of the peak itself. Therefore, we use a peak quality score *q* (equation 3-1)

$$q = 1 - \frac{abs(A_{corr} - A_{sim})}{(A_{corr} + A_{sim})/2} \tag{3-1}$$

where A_{corr} is the correct area of a single peak, and A_{sim} is the area estimated from the peak detection library. The mean value of *q* over all true positive (correctly detected) peaks is used as the score *Q* for the quality of the peak areas detected by our library. As an extension of the F1 score, we define the F1Q score as a metric for all detected peaks, being the evenly weighted harmonic mean of precision, recall, and *Q* (equation 3-2)

$$F1Q = \frac{3}{\left(\frac{1}{precision}\right) + \left(\frac{1}{recall}\right) + \left(\frac{1}{Q}\right)} \tag{3-2}$$

Peak locations, heights, and widths are simulated using either the *NoOverlaps* or *Overlaps* peak models shown below (Table 3-1). A chromatogram is represented as a single vector of values *C* is simulated from an x value of zero to the apex of the terminal peak, plus three times the peak's

σ. The signal height C(*i*) at coordinate *i* is produced from summing over all simulated peaks with simulated abundance levels at point *i*. Noise can be added using the chemical noise described above, as well as the noise models listed in Table 3-2. We modeled three total types of noise: first, *NoNoise*, where the chromatograms were unaltered. Second, *Noise*, where the three noise types listed in Table 3-2 were added to a chromatogram, and *Noise+Chem,* where the chemical noise described above was added with the noise in the *Noise* model. The performance of the CRAWDAD peak library using the F1 and F1Q metrics were evaluated with chromatograms from the six combinations of peak models {*NoOverlaps,Overlaps*} and noise models {*NoNoise,Noise,Noise+Chem*} were evaluated in the results section.

Table 3-1: Peak models used in simulations

Subscripts: (a) - distribution and parameter used to generate random numbers for the corresponding peak parameter; (b) – if the value generated by the distribution in (a) exceeds these limits, a new value is generated; (c) – full-width at half-max height (in 1/100 of a minute) of the generated peak; (d) – height of the generated peak; (e) – interval between generated peaks in units of 1/100 of a minute.

NoOverlaps Peak Model		
Parameter	**Random Variable Distribution[a]**	**Limits[b]**
FWHM[c]	$\mathrm{normal}(\mu = 50, \sigma = 50/2.35)$	[20,80]
Peak Height[d]	$\mathrm{lognormal}(\mu = 0.75, \mathrm{scale} = 5e5)$	NA
Interval[e]	$\mathrm{uniform}(200{,}300)$	N
Overlaps Peak Model		
Parameter	**Random Variable Distribution**	**Limits**
FWHM	$\mathrm{normal}(\mu = 50, \sigma = 50/2.35)$	[20,80]
Peak Height	$\mathrm{lognormal}(\mu = 0.75, \mathrm{scale} = 5e5)$	NA
Interval	$\mathrm{lognormal}(\mu = 0.75, \mathrm{scale} = 150)$	[50,250]

Table 3-2: Noise models used for adding noise to simulated peaks

Subscripts: (a) – noise scaled linearly to the signal, modeling variation from electrospray efficiency; (b) – noise scaled to the square root of the signal, modeling the Poisson response from ions hitting the detector. (c) – modeling Johnson/Nyquist noise in detection electronics.

Noise Type	**Function**
LinearNoise[a]	$\mathrm{N_{linear}(i)} = 0.2\mathrm{C(i)} * \mathrm{normal}(\mu = 0, \sigma = 0.2)$
ShotNoise[b]	$\mathrm{N_{shot}(i)} = \mathrm{sqrt}\big(\mathrm{C(i)}\big) * \mathrm{normal}(\mu = 0, \sigma = 50)$
ElectronicNoise[c]	$\mathrm{N_{elec}(i)} = \mathrm{normal}(\mu = 0, \sigma = 100)$

3.2.6. Estimation of Retention Time Alignment Quality

The prior method described for CRAWDAD to assess alignment performance (Finney et al., 2008) by detecting persistent peptide isotope distributions (PPIDs) using Hardklör(v. 1.33) (Hoopmann et al., 2007) was used, with the following modifications: Peptide isotope distributions were detected with a S/N cutoff of 1, and a minimum correlation score of 0.9. PPIDs were created from those PIDs persisting over at least 7 scans. PPIDs were joined across runs using a greedy algorithm where:

1) Seed PPIDs were defined as all PPIDs in the first µLC-MS run in the set
2) From each other run, the set of PPIDs 10ppm and with 2' of RT from the seed PPID were created. Of that neighboring set, the nearest PPID in the *m/z* dimension is taken and used as a member of a PPID marker group.
3) Each member of the PPID marker groups is deleted from the entire list of PPIDs, such that it cannot be used again in another marker group. Step #1 is repeated until all PPIDs from the first run have been deleted.
4) All PPID marker groups are filtered to retain those containing a maximum retention time retention time distance between members of 2', and a maximum distance in *m/z* between members of 10ppm accuracy.

3.2.7. Normalization of Peak Abundances

Common methods for the normalization of µLC-MS runs assume that an equivalent quantity of digested protein is analyzed for each µLC-MS run. Therefore, the total level of ion counts representing a peptide is compared across runs, and the signal level is globally scaled to match a template run. However, in our comparison of the *E. coli* spike-in series, the level of *E. coli* proteins varies between groups. Therefore, we treat the human peptides as an invariant set of peptides, and apply normalization approaches which treat their results as equal. As in the

previous iteration of the CRAWDAD software (Finney et al., 2008), we used a global scaling approach where a factor is used to multiply the abundances in all peaks of a single run to normalize them to that of a template run (Figure 3-6).

Peakgroups assigned to the HEK dataset were assumed to be unchanging and used as a basis to normalize the data. We evaluated normalization methods using the detected fold-changes of AUC between matching peaks in any two runs. Abundances for a given run over all features are then adjusted to those in a template run by a global scaling factor. In all approaches, we took the set of peakgroups passing the above-mentioned filters, and derived correction ratios $R_{t,j}$ for correcting run *j* to template run *t* as follows. Given two µLC-MS runs *i* and *j*, let $\mathbf{AUC}_i$ be the set of peak abundances { $AUC_{i,1}$ … $AUC_{i,m}$ } for the M matching peaks between runs *i, j*, and in run *t* of the m peakgroups, and A_j be the matching set of abundances for run *j*, and $\mathbf{R_{i,j}}$ be the ratios { $AUC_{i,1}$ / $AUC_{j,1}$ … $AUC_{i,m}$ / $AUC_{j,m}$ } of the AUCs which are members of matching peakgroups between those runs. The *Ratio Median, Ratio Geometric Mean,* and *Ratio Arithmetic Mean* approaches all use the set of ratios $\mathbf{R_{i,j}}$ to determine a representative correction factor for adjusting peak abundances, and each uses the equivalently named statistic of $\mathbf{R_{i,j}}$ – e.g. the median for *Ratio Median.* A given run *t* is designated as a template, and for all runs $i \neq t$ abundance ratios $Ratio_{i,t}$ are calculated as shown above. All MS1 abundance values in run *i* are then multiplied by 1 / $Ratio_{i,t}$.

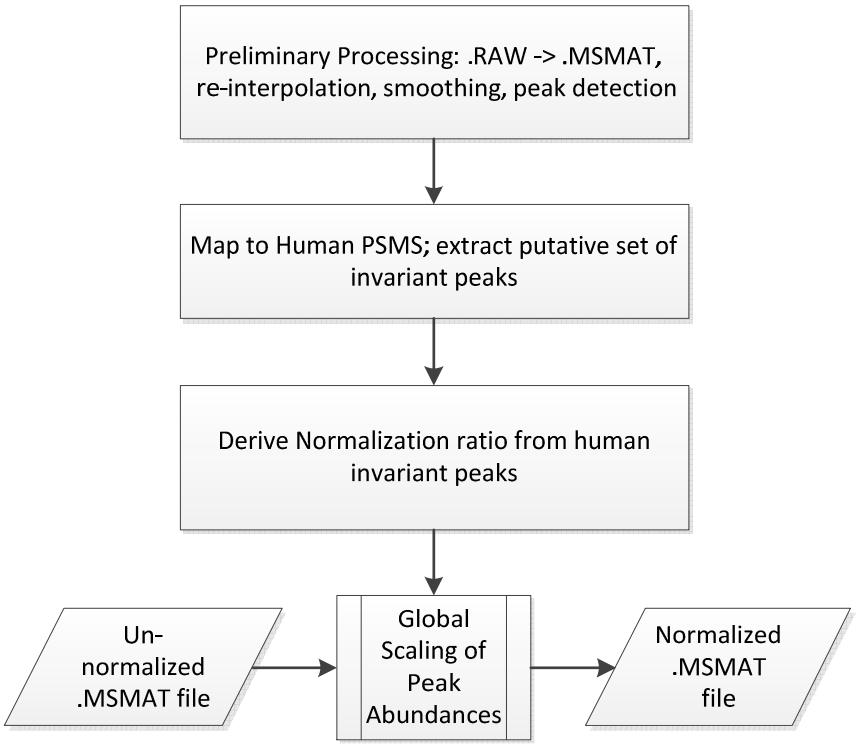

Figure 3-6: Normalization scheme using human / *E. coli* data set

µLC-MS peaks in chromatographic data are mapped to Human peptides. Summary statistics of the ratios for these human peptide peaks are compared between runs to derive a global scaling factor which is then applied to all peaks.

3.2.8. Estimation of *p*- and *q-values* for abundance differences between groups

We use non-parametric permutation tests to derive p-values with a null hypothesis of no difference in peak abundance between two sample groups. Two approaches are assessed below: First, an exact test performed by permuting the labels of peaks across two experimental groups. Second, we adapt a bootstrap method previously developed for microarray data (Storey and Tibshirani, 2003),in which we draw multiple permutations across all features in a bootstrap fashion. In both cases we calculate p-values from the rank of the observed test statistic against the set of test statistics created by permutation or sampling.

P-values for differential abundance between groups can be generated by two methods, in all cases using Welch's unpaired two-sample t-statistic, as follows: define $\mu_{i,g}$ and $\sigma^2_{i,g}$ as

respectively the sample mean and variance for the *i*th peakgroup from group *g.* We then calculate an unpaired t-statistic assuming equivalent variance for comparing two given groups 1,2 for peakgroup *i* (eqn. 3-3).

$$t_i = \frac{\mu_{i,2} - \mu_{i,1}}{\sqrt{\frac{(n_1 - 1)\sigma_{i,1}^2 + (n_2 - 1)\sigma_{i,2}^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (3\text{-}3)$$

Both non-parametric methods use permutation testing on the observed AUCs. In the *intra-group* case, we simply enumerate the absolute value of the test-statistic over all permutations of the labels creating a set of null t-statistics $\boldsymbol{T_i}$, and the p-value is calculated as $\frac{\#\{\boldsymbol{T_i} \geq |t_i|\}}{B}$ where B is the total number of permutations.

In the *inter-group* approach, following Storey et al., we calculate a distribution of t-statistics by drawing *B* randomized permutations of the labels, and creating a list $\boldsymbol{T}$ of null $|t|$ values for all *M* features over the *B* permutations. The default value of *B* is set to 100. The p-value is calculated as $\frac{\#\{\boldsymbol{T} \geq |t_i|\}}{B \cdot M}$, using the rank of $|t_i|$ over the set of null t-statistics created by *B* permutations of the *M* peakgroups.

P–values calculated by these methods are converted to q-values using either the method of Storey where p-value distributions are used to estimate the proportion of differential groups π_0 which are false, or using the methods in the Qvality software (Kall et al., 2009)

3.2.9. Calculation of Peptide and Protein Differential Abundance Ratios

Peptide abundance ratios are calculated from the mean of the ratios of individual peakgroups. These can optionally be weighted by the square root of the abundances of each peakgroup. Likewise, protein ratios are calculated from the weighted mean of all constituent peakgroups. Weights are calculated from the square root of the maximum abundance of an isotope across all

groups being compared (eqn. 3-4). Both arithmetic and geometric means are implemented, with the ratio calculated by the weighted geometric mean (eqn. 3-5) being the default.

$$w_i = \sqrt{\left(\max\left(I_{i,1}, I_{i,2}\right)\right)} \tag{3-4}$$

$$\mathrm{R}_{(\mathrm{pep,prot})} = \exp\left(\frac{\sum_{i=1}^{n} w_i \ln DR_i}{\sum_{i=1}^{n} w_i}\right) \tag{3-5}$$

3.2.10. Software Implementation Details

The computational core of the software was written in C++, compatible with both the Microsoft Visual Studio 2008 development platform on Windows 7 and the GNU gcc (4.1.2) compiler/linker on Linux (2.6.18 kernel). Programs and scripts written in Python (2.5.2) using the Numpy numerical library (1.0.4) were used extensively for reporting of results and automating processing of runs. The R programming language (2.10.1) was used extensively for processing the protein-level summary data and for comparison with spectral counts. Plots were generated with using the Python library matplotlib (0.91.2) and the R statistical analysis platform.

3.3. Results

3.3.1. Peak Detection on Simulated Peaks

Simulated peaks were generated with varying stochastic peak and noise models to test the efficacy of the peak detection library against data where peak locations and boundaries in the presence of noise were known. In all cases, 2000 peaks were drawn, using the same seed in a random number generator to allow comparisons across different methods of adding noise. Peaks were generated randomly using two models with parameters listed below: *NoOverlaps*, and *Overlaps*. Three simulated sets of noise were then added to the two peak series generated above: no added noise(*NoNoise*), added detector and ionization noise (*Noise*), and added

detector, ionization, and chemical noise(*Noise*+*Chem*)(as described in section 3.2.5 above). Peaks detected on chromatograms generated from the *Overlaps*/*Noise* peak/noise generation model are shown in Figure 3-7, where on visual inspection peaks were successfully detected with well-defined boundaries. The same peak detection parameters applied to the *Overlaps*/*NoNoise* and *Overlaps*/*Chem*+*Noise* were also successful, with some false positives emerging (Figure 3-8).

Each of the six simulated chromatograms was run using CRAWDAD with fixed and varying peak detection parameters as described in Table 3-3. The number of true and false peaks detected at each parameter setting was scored by F1, the harmonic mean of precision and recall, and by F1Q, which incorporates q, a setting of peak area quality.

The relationship of the scores F1,F1Q to the parameters outlined in Table 3-3 are plotted in Figure 3-9 for the Noise/Overlaps and Noise+Chem/Overlaps data series.

Three main trends in the relationship of peak finding parameters to the quality measures were observed (Table 3-3, Figure 3-9): First, filter sizes that were larger than the peak FWHM performed better than smaller ones, as seen by a filter of size 50 having lower scores; Second, lower peak area cutoffs gave good results with data with the Noise/Overlaps chromatogram, but the increase in noise observed in the Noise+Chem/Overlaps data series caused the optimal cutoff to be higher; Third, the ‘lower-edge’ background subtraction method gave more accurate peak areas as seen by its superior performance with the F1Q metric, and marginally superior precision/recall by the differences in the F1 curves.

Examples of adjacent peak groups and boundaries of their individual peaks are shown in Figure 3-10. Peaks are grouped using the retention time mapping from the alignment process (Figure 3-11). While the aligned data minimizes the error in retention time, the peak shape becomes

distorted for quantitation purposes. However, we use the uncorrected runs, but linearly shifted by the retention time correction used at the peak apex for grouping peaks together.

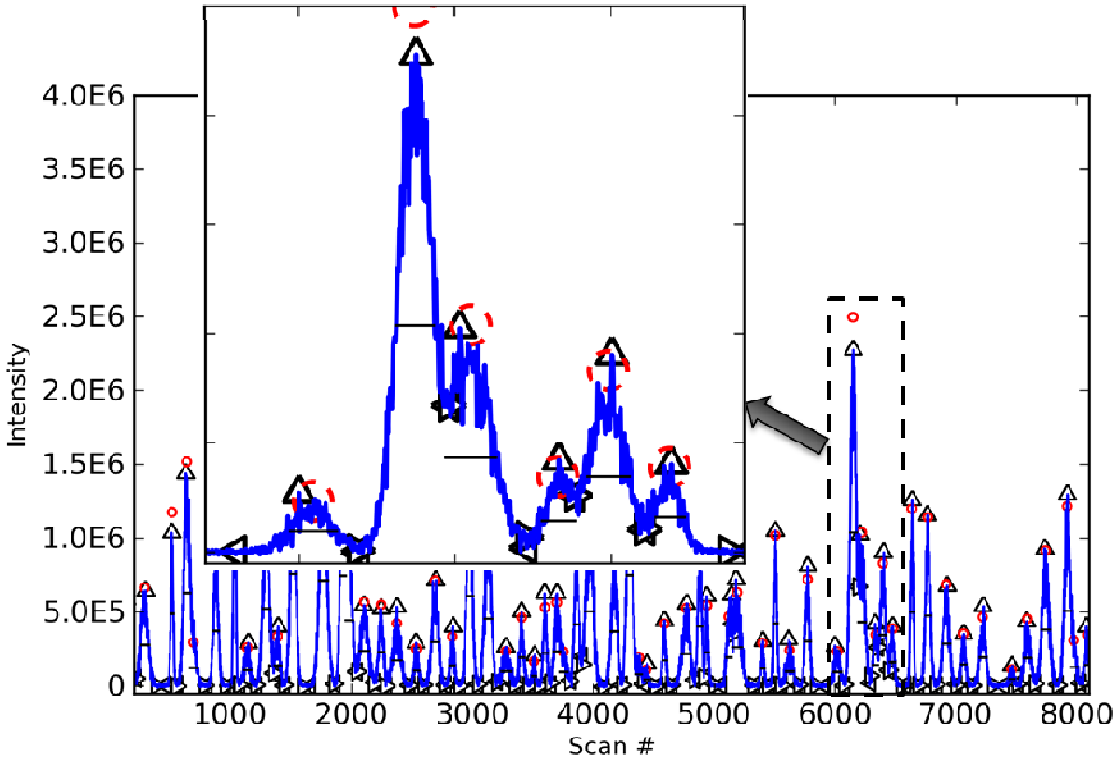

Figure 3-7: Example of peak detection on simulated peaks

Peaks were simulated and detected using CRAWDAD. Peaks and a chromatogram (blue) were simulated with the 'OverlapsNoise' setting, and peak apexes labeled with red circles. Detected apexes are labeled with upwards pointing triangles, and left/right boundaries with left/right pointing triangles respectively.

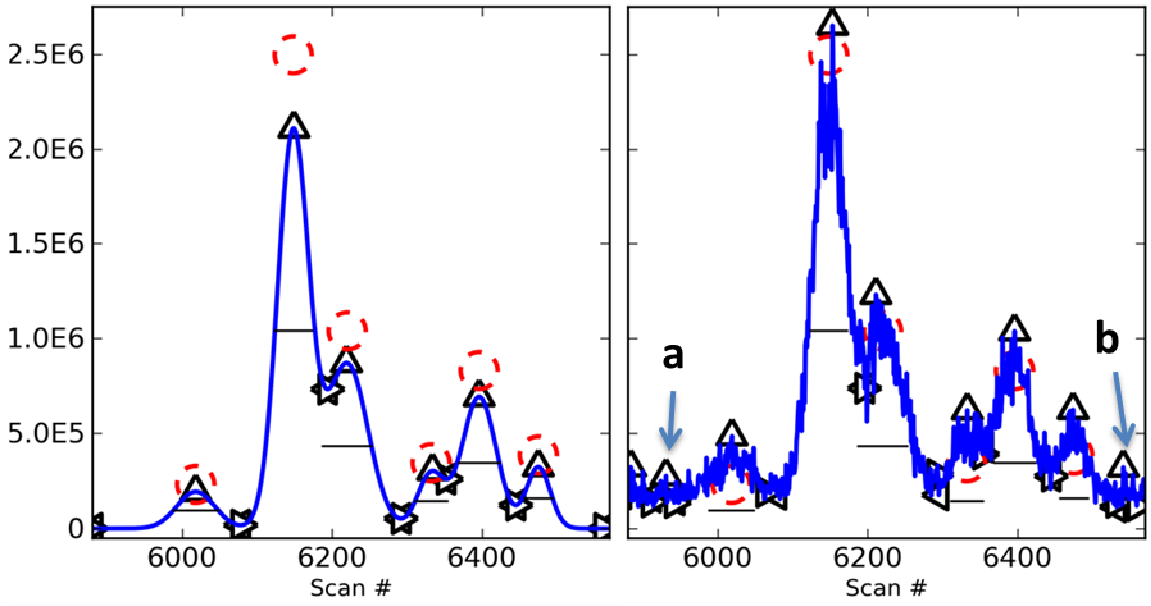

Figure 3-8: Illustration of peak detection with and without added noise

Inset of Figure 3-7 plotted with no noise added(left) and chemical noise added(right). Note examples of false positive peaks (a,b) detected with the addition of noise.

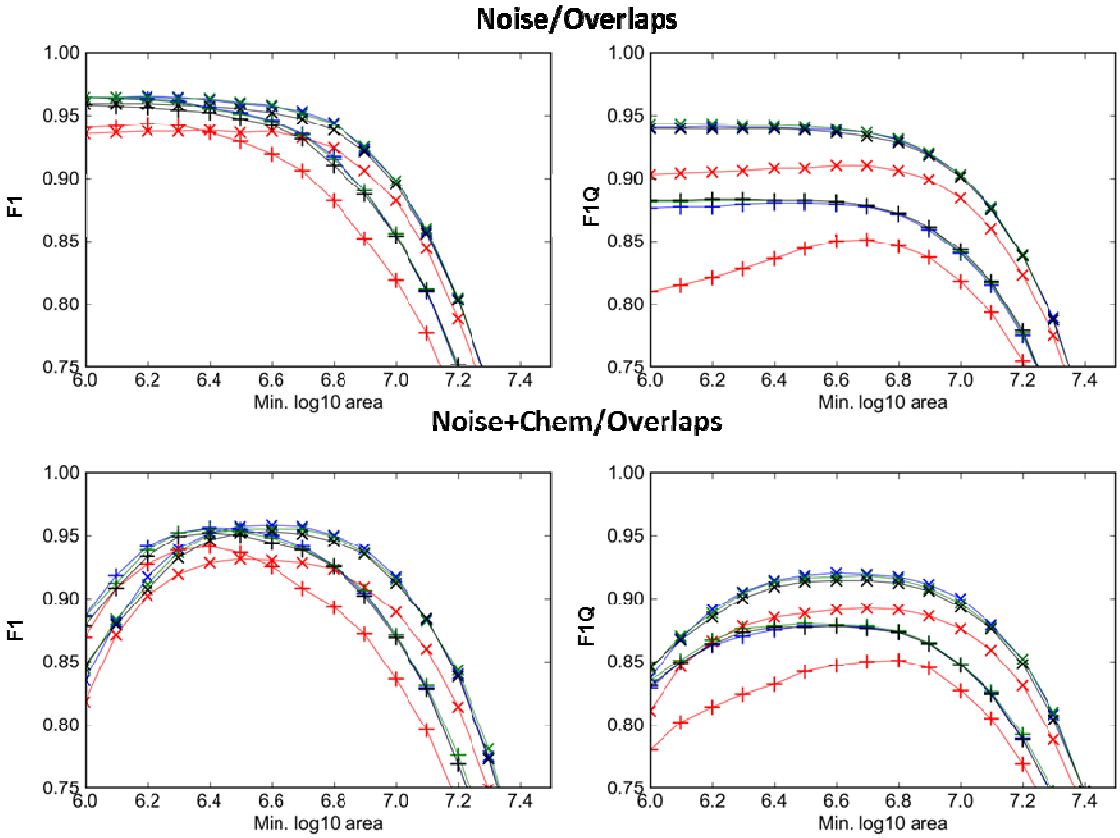

Figure 3-9: Peak detection scores while varying CRAWDAD parameters

F1 and F1Q (scores for peak quality) plotted vs. minimum $\log_{10}$ peak area with varying peak-finding parameter sets applied to the *Noise/Overlaps* and *Noise+Chem/Overlaps* simulated data sets. The colors (red,blue,green,black) correspond to peak FWHM filter sizes of (50,100,150,200). Diagonal crosses: 'lower-edge' background subtraction method (Figure 3-5). Crosses: 'edge-to-edge' background subtraction method. F1: harmonic mean of precision and recall over all simulated and detected peaks. F1Q: harmonic mean of precision, recall, and Q. Q is a measure of the accuracy of the peak area measurement ranging from 0 to 1, averaged over all detected true positive peaks.

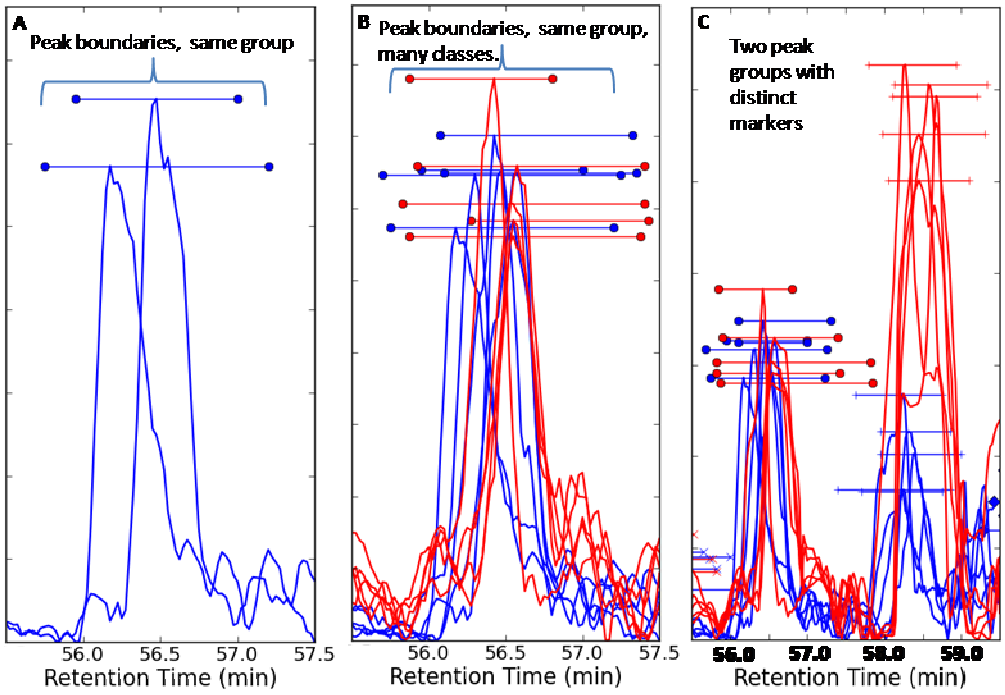

Figure 3-10: Illustration of peak grouping with varying boundaries

A: Replicate runs are shown, with their boundaries displayed by horizontal lines which intersect their respective apex, and the peak boundaries shown by markers at the ends of the lines. Each line is colored with the same coloring as its constituent sample group, and small geometric markers are at the end of each line to delineate its peakgroup. B: The boundaries of each peak in a peakgroup are shown overlaid on each other. C: Multiple groups are delineated by distinct marker shapes at the ends of the peak boundary lines (circles, crosses).

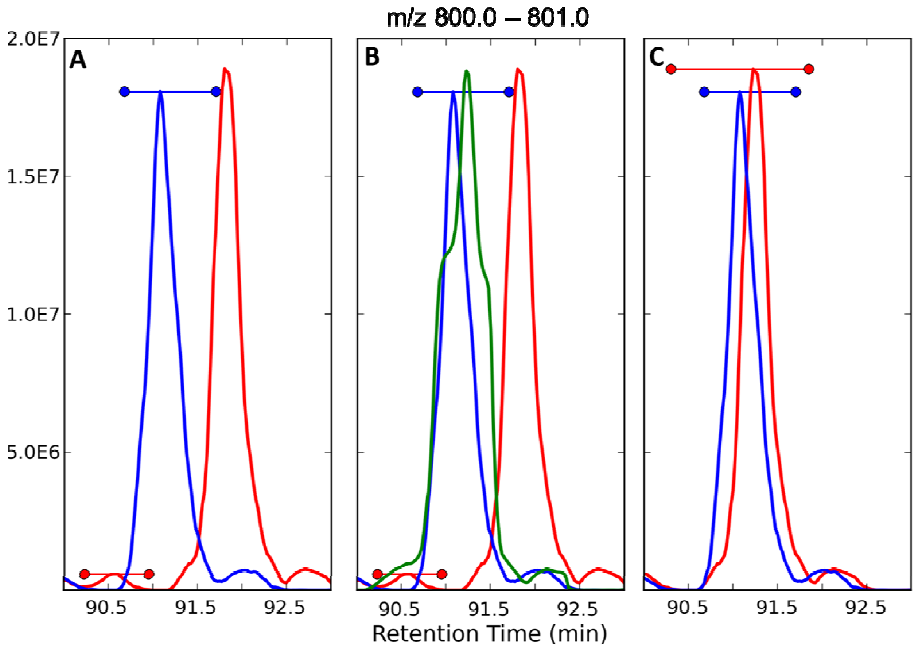

Figure 3-11: Using retention time alignment for grouping peaks

Retention time alignment allows grouping of peaks together, A: Blue and red runs are analytical replicates derived from the same 1X E. coli spike-in sample. Peaks were detected and grouped before alignment, showing a mismatch between peaks. B: The red run has been aligned to the blue run, producing the warped green run. C: The alignment correction discovered for the apex point of the red peak is used to correctly group the unaligned peaks together

Table 3-3: Parameters used for CRAWDAD peak detection

Fixed parameters are listed in the top sub-table, while all combinations of the varying parameters listed below were used to assess the quality of the peak detection. Subscripts: (a) – the minimum length of the peak in scans as detected from the minimum values of convolution with the Gaussian 2nd derivative; (b) the minimum final length of the peak in scans; (c) the maximum final length of the peak in scans; (d) the peak extents are retracted until each is 0.01 of the height between the edge and the apex; (e) the peak is extended using the 1st derivative of the filter as described in 3.2.3; (f) minimum peak area-under-the-curve (g) size of the Gaussian filters used in peak detection in FWHM units; (h) peak background subtraction methods evaluated.

Fixed Parameters	
Name	**Value**
min_peak_g2d_len[a]	*10*
min_peak_len[b]	*20*
max_peak_len[c]	*200*
ratchet_back_to_frac_maxval[d]	*0.01*
crawpeak_extend_method[e]	*1d_extend*
Varying Parameters	
Name	**Value Range**
min_peak_area[f]	$10^x : x \in \{6, 6.1, 6.2 \ldots 8\}$
crawpeak_fwhm[g]	*{50,100,150,200}*
peak_background_estimate[h]	*{'lower-edge','edge-to-edge'}*

Table 3-4: Summary of optimal settings for sets of simulated peak data as shown by maximum F1 and F1Q scores

F1 is the harmonic mean of precision and recall, while F1Q is the harmonic mean of precision, recall, and the Q score for peak area (details in Methods). Subscripts: (a) – Noise added to the data series (see methods for details): N – No, Y – Yes, spray and shot noise simulated, Y,Chem – as in 'Y', with the addition of chemical noise derived from polysiloxane ions. (b) Simulated peaks created with overlaps – Yes/No. (c) F1 max: maximum score for F1 across the peak detection parameter space d) peak detection parameters for the maximum F1 value in c. (d) - FWHM - full-width half max of the Gaussian 2nd and 1st derivative filter size; BGS - background subtraction method: 'peak-to-edge' indicates that peak area is determined with y-coordinates set at the lowest of the two edge values, while 'normal' indicates that a line is interpolated from each peak edge; AreaCutoff – minimum peak area cutoff to be accepted as a peak. (e) maximum value of F1Q found across parameters. (f) parameters for maximum F1Q value in (e), as explained for (d).

Noise[a]	Overlaps[b]	F1 max[c]	F1 max params[d]	F1Q max[e]	F1Q max params[f]
N	N	0.990	FWHM:200 BGS:edge-to-edge AreaCutoff:1E+5	0.991	FWHM:50 BGS:lower-edge AreaCutoff:1E+5
Y	N	1.000	FWHM:200 BGS:edge-to-edge AreaCutoff:1E+5	0.989	FWHM:200 BGS:lower-edge AreaCutoff:1E+5
Y, incl. Chem	N	0.996	FWHM:200 BGS:edge-to-edge AreaCutoff:5E+6	0.954	FWHM:100 BGS:edge-to-edge AreaCutoff:5E+6
N	Y	0.965	FWHM:50 BGS:lower-edge AreaCutoff:1E+5	0.932	FWHM:100 BGS:lower-edge AreaCutoff:1E+5
Y	Y	0.965	FWHM:100 BGS:lower-edge AreaCutoff:1E+5	0.943	FWHM:150 BGS:lower-edge AreaCutoff:1E+5
Y, incl.	Y	0.958	FWHM:100	0.920	FWHM:100

Chem	BGS:lower-edge AreaCutoff:5E+6	BGS:lower-edge AreaCutoff:5E+6

3.3.2. Retention Time Alignment of *E. coli* / Human spike-in Dataset

The 1x vs. 8x *E. coli* / Human spike-in dataset described above was processed for chromatographic alignment with CRAWDAD as described previously (Section 2.2.6). In the first step, an optimal alignment template was chosen for the group, with a reduction in the mean standard deviation of retention time markers from 0.442' to 0.146' (Table 3-5). Second, the use of the Spearman rank correlation reduced the deviation to 0.134', and use of a heuristic to remove spikes in chromatograms resulted in a reduction to 0.1329' (Table 3-5, Figure 3-12). The median speed of alignments was 486 seconds on a 2.33 GHz Intel Xeon processor with 8GB of RAM, but this can be decreased to 16.6 seconds with minimal loss of quality by binning the data in *m/z* space (Appendix 2).

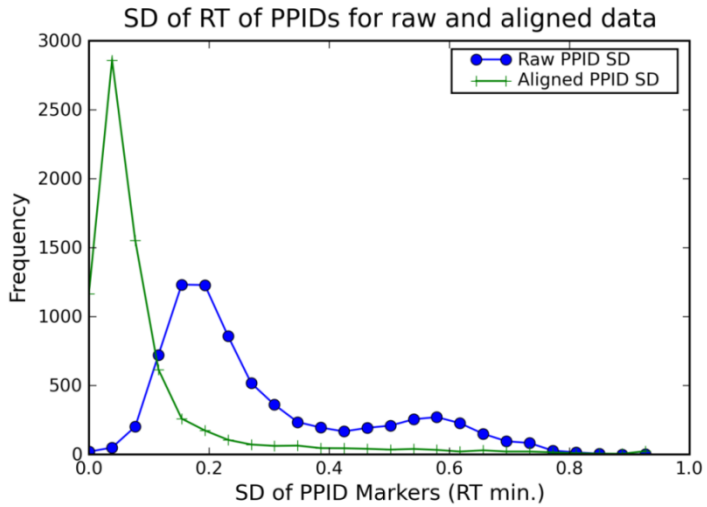

Figure 3-12: Retention time errors before and after alignment

Distribution of standard deviation of persistent peptide isotope distribution markers before (Raw - Blue) and after (Aligned - Green). Markers were detected from PPIDs found across all data sets with alignment parameters as shown in the final row of Table 3-5. The mean SD of retention time decreased from

Table 3-5: Alignment quality compared between template runs

Alignment quality measured by standard deviations of persistent peptide isotopic distributions (PPID) dependent upon choice of alignment template run. The best PPID marker standard deviation for each grouping is underlined. Fields: Template Run – µLC-MS run used as alignment template. Marker SD – mean standard deviation in retention time of PPID markers across all groups. Marker SD (1X) – Marker SD limited to runs from the 1X spike-in group. Marker SD (8X) – Marker SD limited to runs from the 8X spike-in group.

Template Run	Marker SD	Marker SD (1X group)	Marker SD (8X group)
Unaligned	0.4424	0.2399	0.3008
1X02	**0.1461**	**0.1141**	0.1089
1X03	0.1493	0.1146	0.1179
1X04	0.1485	0.1151	0.1165
1X05	0.1501	0.1166	0.1195
8X02	0.1467	0.1185	**0.1062**
8X03	0.1488	0.1174	0.1083
8X04	0.1475	0.1234	0.1097
8X05	0.1488	0.1239	0.1068

Table 3-6: Effect of further alignment refinements

Effect of further alignment refinements after template choice on quality. The best PPID marker standard deviation for each grouping is underlined. Fields: Template Run – LC-MS run used as alignment template. Score Method – (DP^2 – dot product squared, R_s _spearman rank correlation). Other Parameters – Spike(N) – LC-MS signals were limited to those that persisted over N scans. Marker SD - mean standard deviation in retention time of PPID markers across all groups.

Template Run	MS1 Spectral Comparison Method	Other Parameters	Marker SD
1X02	DP^2	NA	0.1461
1X02	R_s	NA	0.1344
1X02	R_s	Spike(6)	0.1333

3.3.3. Peptide and Protein Identification

Peptides and Proteins were identified using the Bullseye/SEQUEST/Percolator pipeline as outlined in *Methods*. Identified proteins were collapsed to parsimonious non-redundant protein groups based on the identified peptides (Zhang et al., 2007).

Table 3-7: Summary of proteins and peptides identified from *E. coli*/human spike-in mixture. Five µLC-MS runs were acquired from each of the 1X and 8X spike-in groups. Proteins were parsimoniously assembled to non-redundant groups, and numbers of groups were reported based on requiring a minimum of one or two peptides per protein group. Number of acquired spectra are shown as well as the number and percent fraction having a q-value of less than or equal to 0.01%

Sample Group	Total Protein Groups (≥2 pep; ≥1 pep)	#*E. coli* Protein Groups (≥2 pep; ≥1 pep)	#Human Protein Groups (≥2 Pep; ≥1 pep)	#*E. coli* Peptides	#Human Peptides	#Spectra (total; #qval ≤ 0.01%)
1X	1062; 1529	82; 158	880; 1471	381	2180	78717; 12327 (15.7%)
8X	860; 1491	315; 506	545; 985	1622	1314	93524; 16836 (18.0%)
All runs	1246; 2073	316; 507	930; 1566	1633	2346	172241; 29163 (16.9%)

3.3.4. Abundance Normalization of LC-MS runs

Normalization by total TIC was attempted using with a run in the 1x spike-in series as a template. While the mean normalization factor for the 1x runs to the template was 0.98, the mean normalization factor for the 8x series to the template was 0.51, indicating that the total ion current in the 8x spike-in series was approximately least two-fold higher. Applying this would spuriously increase the signals in the 1x series containing less total protein.

The empirical distribution of detected protein fold changes was bimodal, with two peaks corresponding to the 1x and 8x spike-in series (Figure 3-13). To determine global scaling

normalization ratios, we filtered the data by limiting our data considered to areas under the curve from the peakgroups that were mapped with a high degree of confidence to peptides present solely in the human sample. As we have confidence that the PSMs have a conservative FDR rate of 0.01, we estimate an error rate for the mappings we observed that the distribution of annotated peakgroups by the $\log_2$ abundance fold-change reached a minimum at approximately 1.5 in $\log_2$ space, and this was used as a criteria to determine if the ion current ratio was derived from a peakgroup belonging to a HEK peptide ($\log_2$ ratio < 1.5) or *E. coli* ($\log_2$ ratio > 1.5) (Figure 3-6).

We detected 7755 peakgroups with peaks present in all groups across all replicates. 4042 of these were annotated with HEK peptides for the normalization data set, and 3715 had a $\log_2$ ratio between the 1x and 8x groups in the interval [-1,1], suggesting they were truly from the HEK dataset. These were filtered to retain those with an RSD of abundance (across all replicates) <= 90% quantile (RSD value 0.425), leaving 3342 peakgroups used to assess normalization algorithms.

We use the median and mean RSD as metrics of normalization efficacy for the different approaches (Table 3-8). Moreover, the mean and inter-quartile range (IQR) of the t-statistic should not increase in its deviation (for the non-changing HEK peptides, the mean t-statistic should close to zero). The *geometric* approach yielded the best improvement in the mean RSD, along with the mean of the ratios, and had a lower deviation from zero in the mean t-statistic (Table 3-8).

The noise or error in the ion abundances can be seen to have a multiplicative effect, and are less biased after the simple normalization (Figure 3-14). However, one run (1x02) did not respond well to any of the normalization approaches, as it appeared to contain multiple outliers of peptide ratio measurements.

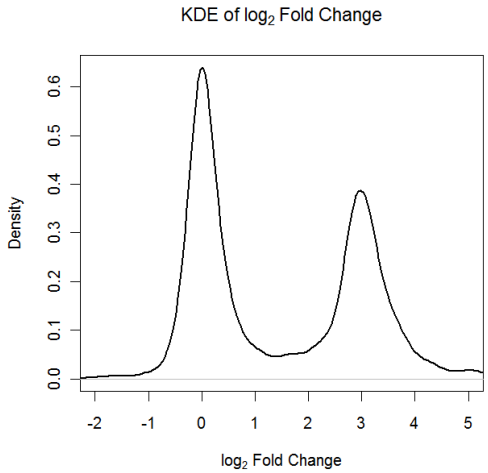

Figure 3-13: Estimate of empirical distribution of $\log_2$ ratio of abundance between 8x and 1x spike-in series.

Total number of peakgroup ratios is 7755. Density was estimated using a Gaussian kernel density estimate.

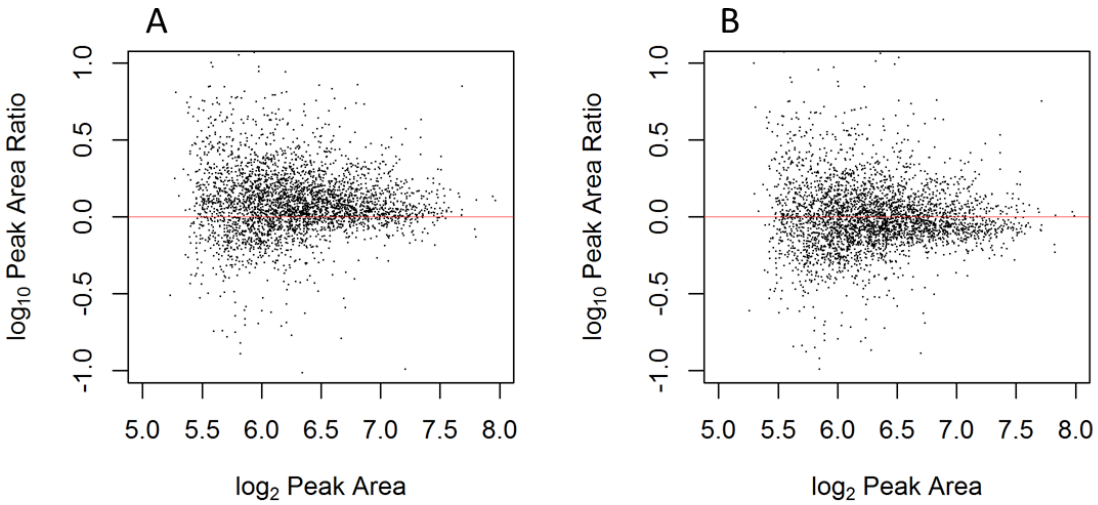

Figure 3-14: Effect on peak abundance ratios of global scaling of peak abundances

Comparison of peak area ratios before (A) and after (B) normalization by global scaling of the geometric mean of the ratios between matching peaks of the 2nd and 3rd run in the 1x spike-in series. Log_{10} mean peak area in the two runs (abscissa) is plotted against the $\log_2$ ratio of peaks in the pair of runs (ordinate)

Table 3-8: Summary of peakgroup CV and statistics before and after normalization procedures.

Methods used: *Ratio Median*(a), *Ratio Geometric Mean*(b), *Ratio Arithmetic Mean*(c), and the ratio of the summed AUCs of peakgroups(d)

Normalization Type	Median CV	Mean CV	Mean |t|	T statistic IQR
None	0.134	0.158	0.07912	-1.227 - 1.428
Median[a]	0.128	0.151	-0.1917	-1.66 - 1.277
Geo. Mean[b]	0.124	0.147	-0.137	-1.66 - 1.44
Mean[c]	0.124	0.147	-0.255	-1.79 - 1.34
Peakgroup sums[d]	0.128	0.151	-0.057	-1.52 - 1.37

Table 3-9: Normalzation global scaling ratios calculated by total TIC vs. geometric mean of constant peakgroups

Adjustment ratios for individual runs against the 1x03 run, calculated by the geometric mean of ratios, or the total TIC of the µLC-MS run. a) the 1x02 run was not normalized, as it contained many ratio outliers. b) 1x03 was used as the 'template' for normalization, and no correction ratio was applied. The mean adjustment for the 1x and 8x series of runs using the TIC approach were 0.9833 and 0.5125 respectively.

Method/Run	8x02	8x03	8x04	8x05	1x02	1x03	1x04	1x05
Geo. Mean	0.99	1.03	0.93	0.91	1.00[a]	1.00[b]	0.94	0.91
Total TIC	0.53	0.54	0.49	0.49	1.09	1.00	0.95	0.91

3.3.5. Evaluating the Effects of CRAWDAD Parameters on Accuracy and Sensitivity

In order to determine the effects of CRAWDAD peak finding and differential detection parameters on the sensitivity and accuracy of the detected changes, we ran the 1x vs. 8x spike-in dataset using the combinations of parameters outlined in (Table 3-3). Results from each run were scored by four metrics – number of differential peakgroups detected (num_drs); median coefficient of variation (CV) of detected peakgroups (cv_median); number of peptides detected changing (*num_peps*); and the sum of a protein accuracy score *prot_raw* over all detected proteins. The *prot_raw* accuracy score is defined in eqn 3-5, where #prots is the total number of detected proteins called as changing, $ratio_{exp}$ is the expected ratio of that change, and $ratio_{obs}$ is the observed ratio of that change.

$$\sum_{i}^{\#\,prots} 1 - \left|\log_2(ratio_{\exp}) - \log_2(ratio_{obs})\right| \quad (3\text{-}5)$$

An initial analysis using principal component analysis on the scores and parameters with mean set to zero and unit variance showed six distinct clusters, which were defined by the min_peak_area and craw_max_qvalue parameters. Density estimates of the distribution of the above mentioned metrics over all parameterized runs, with the given parameter held constant, were produced for a set of parameters (Figure 3-15).

The importance of a parameter can be indicated by the degree to which it splits the distributions of scores. Minimum peak area was shown to have the greatest influence on the scores. The median CV of detected peakgroups increased with a lower peak threshold, indicating that the population of peaks had higher variance overall. Likewise, the number of peptides and peakgroups detected increased. The protein score, which benefits from both an increase in detecting the number of proteins and the accuracy of measured ratios, did increase from the lower peak detection threshold, even though the median CV of detected peaks was higher. While the peak filter FWHM size showed an effect on results with the simulated data, it was less pronounced in this case. Likewise, comparing a permutation t-test vs. a parametric one had minimal effects on the final results.

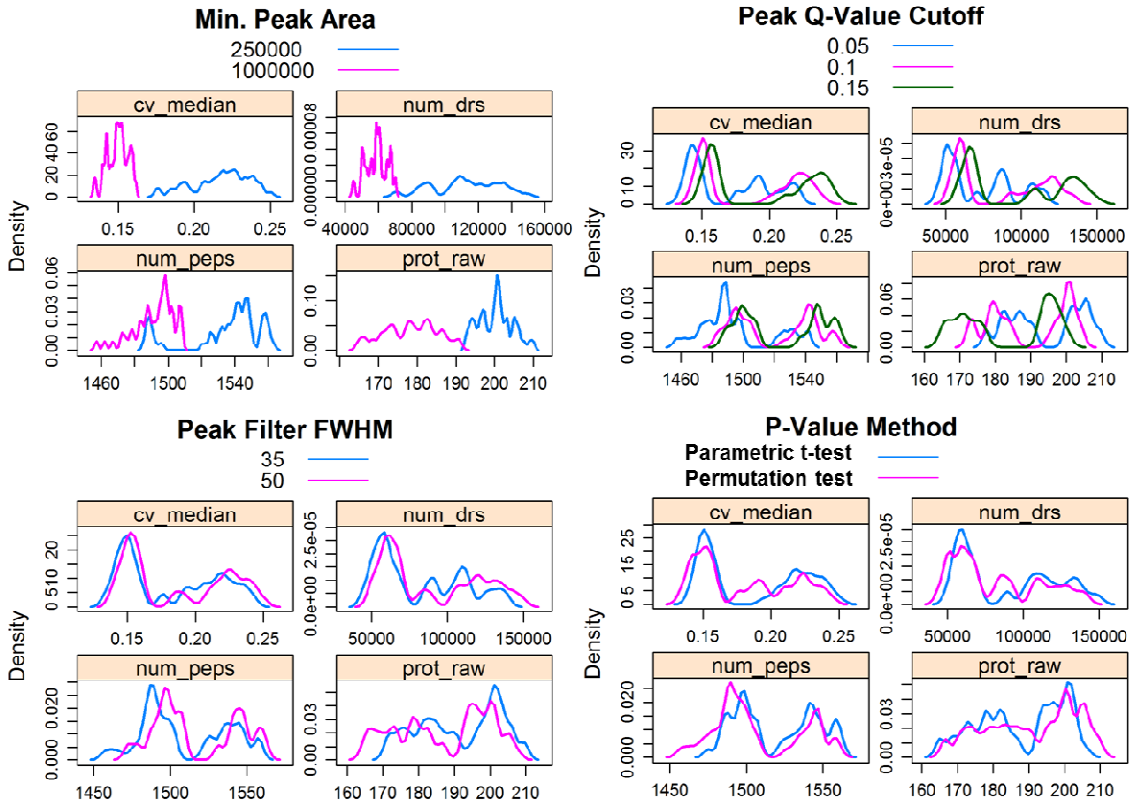

Figure 3-15: CRAWDAD scores conditioned by configuration parameters defined in Table 3-10.

CRAWDAD was run with 96 distinct settings, and scores for the median CV of peakgroups (*cv_median*), number of peakgroups (*num_drs)*, number of peptides (*num_peps*), and protein accuracy score (*prot_raw*) were calculated for each set of parameters. Density plots of the scores from all parameter sets conditioned by the titled parameter are shown in each sub-plot above. For example, in the top left panel, the blue line shows the distribution of scores for the 48 of 96 parameter sets where min_peak_area was held to 1,000,000, while the magenta line shows scores for the remaining 48 parameter sets where min_peak_area was held at 250,000.

Table 3-10: Parameters examined for scores in CRAWDAD comparisons

CRAWDAD was run on the 1x/8x HEK/*E. coli* data set, with the parameters below varying. Parameters: (a) – minimum peak area in units quantified under the curve. (b) – maximum distance for members of a peakgroup in RT (minutes) from the seed peak (c) – size of the Gaussian filter at full-width half-max used for peak detection (d) – minimum length of a peak in scans found using the 2nd derivative Gaussian filter (e) – maximum q-value for peakgroups found to be differential (f) – method for calculating p-value of peakgroups being differential – (permutation, t-test)

Varying Parameters	
Name	**Value Range**
min_peak_area[a]	*{2.5E5, 1E6}*
crawpeak_FWHM[c]	*{35,50}*
craw_max_qvalue[e]	*{0.05,0.10,0.15}*
pvalue_method[f]	*{permutation, t-test}*
cpg_seed_peak_delta[b]	*{0.5,1.0}*
min_peak_g2d_len[d]	*{0,15}*

3.3.6. Comparison of CRAWDAD results to Spectral Counting

We are primarily interested in using MS1 ion abundance as a proxy for peptide or protein abundance. However, precursor ion intensity alone is not sufficient to quantify identified peptides, as MS/MS spectra are necessary to identify even a highly accurate mass. Mass spectrometers routinely acquire MS/MS spectra in a data-dependent fashion, designed to sequence as many of the high-abundance peaks as possible. This however, results in oversampling of peptides proportional to their abundance. As a side effect of this oversampling, spectral counts can be used as a proxy for abundance. Spectral counting has become commonly used as a method for relative quantitation, as they are relatively robust against poor chromatography or electrospray, and easier to work with computationally. Moreover, spectral counting may undersample from highly abundant peptides, causing a bias toward lower ratios. We compare the measured abundance ratios from CRAWDAD to those determined by spectral counting below.

Although NSAF (Zhang et al., 2010) and $S_I N$ (Griffin et al., 2010) are considered more accurate forms of reporting quantitative changes from spectral counting data, we use R_{SC} (Old et al.,

2005), as it incorporates pseudo-counts which allow us to obtain some quantitative information from proteins with 0 spectral counts in one of the categories being compared. After Old et al., R_{SC} is calculated where iterating over all proteins(eqn. 3-6)., where n_1 and n_2 are spectral counts for the protein in Samples 1 and 2, respectively; t_1 and t_2 are total numbers of spectra over all proteins in the two samples; and f is a correction factor or pseudocount set to 0.5 in this case.

$$R_{SC} = \log_2\left(\frac{n_2 + f}{n_1 + f}\right) + \log_2\left(\frac{t_1 - n_1 + f}{t_2 - n_2 + f}\right) \quad (3\text{-}6)$$

Spectral count $\log_2$ ratios were calculated for those *E. coli* proteins with MS/MS spectra mapped to precursor peak AUCs by CRAWDAD with spectral counts from both categories using NSAF, R_{SC}, and the arithmetic ratio (Figure 3-16). CRAWDAD abundance ratios were calculated by the square root-weighted geometric mean of the feature ratios. The expected $\log_2$ ratio for the 8x spike-in is 3, and the protein ratios derived from peak AUCs were more accurate than those from spectral counting. Moreover, the three spectral counting methods gave comparable results to each other, showed greater variance, and were slightly biased towards an underestimation of the detected fold-change.

We then compared the measured fold change detected by CRAWDAD to that from spectral counting using the R_{sc} method, thereby comparing the detected $\log_2$ ratio for both Human and *E. coli* proteins using spectral counting or AUC ratio as a function of summed protein AUC signal in

Figure 3-17. The CRAWDAD ratios (left panel) fit well to the expected ratios, with greater variance at lower protein signals. Spectral counting ratios were both less accurate and precise relative to CRAWDAD measured ratios. The frequent ‘bands’ in the spectral counting data reflect ratios from low numbers of counts.

The relationship between the ratios measured by two techniques for the same proteins was examined (Figure 3-18), and only a weak significant correlation ($r^2 < 0.1$) was found for the Human proteins, while the *E. coli* spike-in proteins showed no significant correlation (Table 3-11), indicating that the errors in differential protein ratio measurement were independent of the measure used.

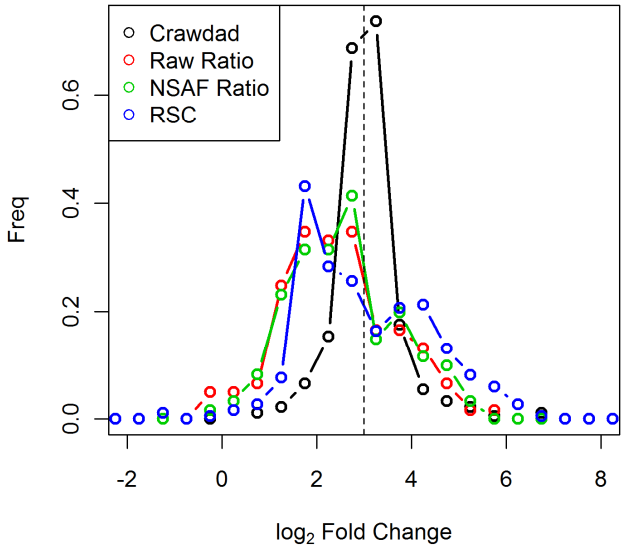

Figure 3-16: Histogram of protein $\log_2$ fold-changes between 8x and 1x samples as detected by CRAWDAD and by the spectral counting methods raw ratio, NSAF, and RSC.

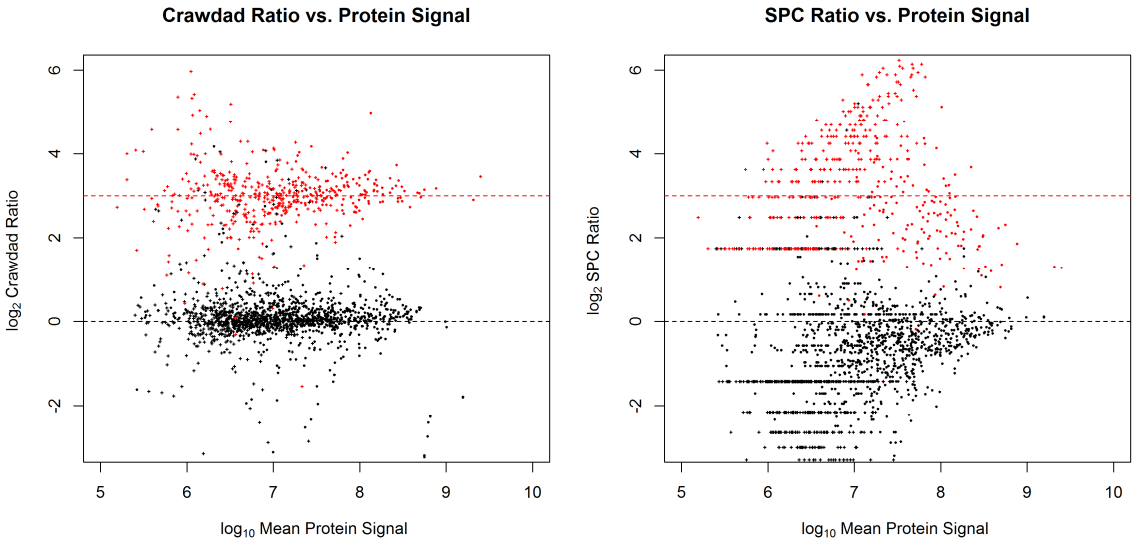

Figure 3-17: Comparison of protein abundance ratios between CRAWDAD and SpC

Measured protein abundance ratios for proteins derived from *E. coli* spike-in (red) and Human control (black) using chromatographic AUC calculated with CRAWDAD vs. spectral counting. Log_2 protein abundance ratio is plotted against mean summed protein AUC from both groups: ($ProtAUC_{1X}$ + $ProtAUC_{8X}$) / 2 where $ProtAUC_c$ is the sum of AUCs for peakgroups comprising the relevant protein from class *c*. Expected log_2 ratios of 0(human) and 3(*E. coli*) are shown as horizontal dashed lines. Left panel: Protein abundance ratios calculated with CRAWDAD. Right panel: Protein abundance ratios calculated with the RsC spectral count ratio.

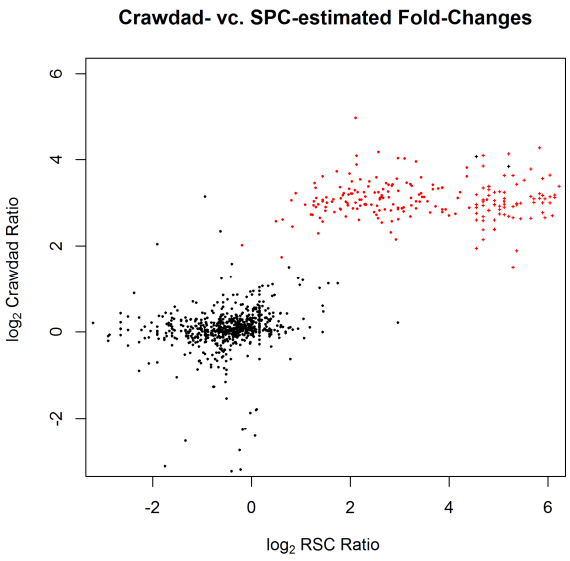

Figure 3-18: Comparison of errors in proteins abundance between CRAWDAD and SpC

Relation of measured abundance ratios for proteins from Human control set (black) vs. E. coli spike-in (red). Log_2 protein abundance ratio measured by RSC is plotted against log_2 protein abundance ratio measured by CRAWDAD.

Table 3-11: Correlation between protein fold-changes detected by CRAWDAD precursor intensity averaging and fold-change from spectral counting.

Organism	SPC Limit	# Proteins	(SPC vs. CRAWDAD) Fold-change correlation r^2	p – value
E. coli	>= 1	477	0.0048	0.130
E. coli	>= 10	221	0.0006	0.72
Human	>= 1	2865	0.0600	0
Human	>= 10	1217	0.0529	4.44e-16

3.4. Discussion and Conclusions

3.4.1. Peak Detection on Simulated Peaks

The use of a matched filtration or convolution based method for peak detection is sensitive to the size of the filter. Our simulation studies on data indicate a preferred Gaussian filter size of approximately twice the size of the peak: the best results in this comparison were found with a full width at half-max (FWHM) filter size of 100 using the F1Q score (Table 3-4) on simulated peaks ranging from 20-80 units of FWHM and a mean FWHM of 50. The conclusions regarding specific score values should be considered to be limited to the simulated data set.

The method for edge detection of peaks had a strong effect when considering the F1Q score, simply seen by the score always being higher for any given combination of FWHM filter size and peak area cutoff (Figure 3-9). Peaks modeled with overlaps consistently performed better using the ‘lower-edge’ background subtraction method – where the boundary background level is set at constant level of the lowest of the two peak edges - both in terms of solely precision and recall (F1 score), and when including peak area accuracy (F1Q score) (Table 3-4). However, on peaks simulated without overlaps, the ‘edge-to-edge’ method performed better with the F1 score, while results were mixed using the F1Q score (Table 3-4). These results indicate that the ‘lower-edge’ background subtraction was superior in cases of both high noise and overlapping peaks, which is likely to be characteristic of high-complexity proteomics data. Notably, the use of lower resolution instruments such as ion traps would benefit from more careful background subtraction (Eilers, 2003), due to the greater number of co-eluting compounds detected within the resolution of the instrument.

The use of a score incorporating both detection in terms of precision and recall (F1) and peak area quality is useful to look at the tradeoffs of different approaches. It should be noted that in

our application, each component (precision, recall, and *q*) is weighted equally in the harmonic mean, but it may be useful to apply these criteria that weight the components differently. For example, if quantitative accuracy were the greatest concern the weight of *q* could be increased, while where false positives were to be avoided precision could be weighted higher.

The use of area cutoffs as a criterion for peak calling is also dependent upon noise, where a higher peak area cutoff was useful when a larger amount of noise was added such as with the Noise+Chem set, even when background was being subtracted. However, the optimal set of thresholds with regards to recall, precision, and peak area quality determined from simulated data may not carry over to actual analytical data. The behavior of known standards spiked in to a complex matrix, or a well characterized set of targeted selected reaction monitoring (SRM) data, could serve as a template to calibrate the most effective parameters on a particular instrument for samples with complexity similar to the test data being considered.

Chromatographic peaks are generally not symmetric (Li, 2002), and the use of a symmetric filter in their detection may be less sensitive than the performance shown here with peaks modeled with only symmetric Gaussian functions. The robustness of our approach could be assessed with data simulating a wider range of peak shapes, created by a randomly weighted (per peak) mix of Gaussian and exponential functions (Schulz-Trieglaff et al., 2009).

3.4.2. Qualitative Characterization of the *E. coli* / Human Spike-In Dataset

The complexity of an analytical sample composed of two unfractionated proteomes should be quite challenging, but the combination of using a high-resolution instrument such as the LTQ-FT, and a long separation (3 hours), make it possible to see that a large number of proteins were detectable over the 8 analyses (2073 distinct protein groups from both species - Table 3-7).

The detected number of human peptides decreased by 39% from 2180 to 1314 between the 1x to 8x groups, even though the fraction by mass of the human peptides in the mixture decreased by only 15% - from 97.5% by mass (40:1 ratio) to 83.3% by mass (5:1 ratio). This indicates that although the *E. coli* peptide mixture was at a much lower baseline level compared to the human unchanging background, it is likely that the lower complexity of the *E. coli* proteome causes that individual peptide peaks to be at a higher intensity level. The increase in the *E. coli* peptide abundance may cause either ionization suppression which could reduce the likelihood of identifying a co-eluting peptide at a fixed level, or compete with human peptides for acquisition by the mass spectrometer.

3.4.3. Evaluation of CRAWDAD Parameters

A limited exploration of the parameter space for CRAWDAD settings for both feature/peak detection and the statistical thresholds for determining significant differences showed that peak area cutoffs had the largest effect, followed by q-value cutoff (Section 3.3.5). A superior approach to determining optimal parameters in both feature/peak detection and statistical thresholds for differential analysis would be to control for protein or peptide-level empirical FDR, and to search the parameter space more exhaustively using a single objective function. The engineering of the CRAWDAD software and processes for evaluating the results are not currently modular enough to support efficient evaluation of a large number of different settings (e.g. by saving intermediate work and only varying downstream parameters (Figure 3-2)), but it is possible to search larger sets of parameters for scoring.

3.4.4. Comparison of Results between CRAWDAD and Spectral Counting

Both quantitation by integrating peptide ion intensity over time as the area-under-the-curve (AUC) of chromatographic peaks (CRAWDAD) or spectral counting (number of peptide detection events from MS/MS spectra) are based on counting: in the former case, the effect of

the number of ions on a detector, and in the latter the number of discrete MS/MS peptide identification events. The number of counts from MS/MS spectral counting is far lower, as a large population of ions (on the order of 20,000 in the data shown above) must be sampled to give a single MS/MS spectrum. This results in the lower precision seen with spectral counting as compared to the CRAWDAD quantitation (Figure 3-17).

Relative quantitation on the peptide level is also possible with AUC-based methods, while the lower precision of spectral counting due to a limited number of counts would be exacerbated. This enabled the use of CRAWDAD in the detection of changes in peptides that contain post-translational modifications, both in an increase in the stoichiometry of the modified form, and in a synchronized decrease in the unmodified form (Eakin et al., 2007).

4. Non-Parametric P-Values for Differential Label-Free LC-MS Proteomics Experiments

4.1. Introduction:

Proteomics advances using LC-MS has made it possible to study the simultaneous differential abundance of thousands of proteins. A large number of statistical tests are used to determine which features from LC-MS data, whether proteins, peptides, or individual isotopic peaks, are significantly changing in pairwise comparisons. The calculation of accurate p-values is essential both on a per-feature basis, and as a necessary step to calculate the expected false discovery rate (FDR) at a given significance level.

The statistical analysis of label-free LC-MS mass spectrometry data is challenging from many perspectives. First, the distribution of the abundance of each protein is not well understood. Few replicates, whether biological or analytical, are available to detect if the measurements are normally distributed. Second, the high number of features being compared in an experiment (approximately 10^4 to 10^5 if considering peptides or isotopic peaks) forces one to consider multiple testing issues when choosing a statistical cutoff score. Below, we assess the accuracy of p-values for differential abundance calculated using both parametric and non-parametric tests on differential proteomics data.

Non-parametric permutation tests are used to derive p-values without assumptions about the underlying distribution. The set of null test statistics is created by permuting the group labels of each replicate. The rank of the observed test statistic in a set of test statistics modeling a null distribution is taken as the p-value. Two methods of generating a null distribution with our quantitative proteomics data are assessed below: First, we permute the group assignment of

replicates across two experimental groups; Second, we use a bootstrap method developed for differential gene expression data (Storey and Tibshirani, 2003) drawing multiple null permutations from all features using a bootstrap. We can directly assess the accuracy of calculated p-values by analyzing data known to not truly show any differences between groups to be compared, and comparing the distribution of calculated p-values against the uniform distribution over [0,1] expected by chance.

We generate null quantitative LC-MS data with multiple LC-MS replicates of the same sample, which should only show differences by random chance. We compare these non-parametric p-values with those from a parametric *t*-test on null LC-MS data, under the assumption that p-values from data which conform to the null hypothesis (no difference in abundance between groups) will have a uniform distribution over the interval (0, 1).

The structure of experimental design needs to be considered when defining the groups to be compared. We have observed an effect where the signal intensity of LC-MS runs declines in those run later in an experiment (Schulz-Trieglaff et al., 2009), which may be due to column degradation or clogging. If two biological classes are run in sequential order, this drop-off in intensities from the first half of runs to the second half may create a bias showing false differences in mean abundances between the two biological classes. By randomizing or alternating the run order, we can diminish the bias that the drop in intensity has on the measured difference of means between the two groups (Oberg and Vitek, 2009).

Naively using a p-value cutoff on a large differential data set does not give the intuitive behavior of controlling the probability that any type 1 error (false positive) will occur in the data set. For example, using a significance threshold of 0.05 when analyzing 20,000 peaks for differential abundance, the expectation value of the number of peaks being called different by chance would be 1,000. One approach to limit the number of expected false positives is to control the

Family-Wise Error Rate (FWER), which is the probability that *at least one* type I error (false positive) will occur in a population, or *family,* of tests. The Bonferroni (Abdi, 2007) correction to control the family-wise error rate(alpha) calls for dividing the significance level for a single test by the number of features being tested. In the example above, this would call for a feature-wise significance level of 0.05/20000 (2.5E-6) – which will greatly reduce power in these experiments, and is overly conservative for our purposes(Benjamini and Hochberg, 1995).

We are not interested in a conservative error-free list of differential features with low power, as many features would be lost. Rather, we would like to define the expected number of false rejections of the null as a proportion of our data set. While p-values are used to control the false positive rate, or the proportion of truly null tests which are called significant, it can be more useful to control the proportion of false positives in those tests which are called significant, known as the FDR (Benjamini and Hochberg, 1995).

4.2. Methods:

Sample Preparation

The soluble fraction of mixed state *C. elegans* homogenate was prepared as described previously (Hoopmann et al., 2009) and stored at −80 ℃. The lysate was denatured using 0.1% RapiGest SF (Waters Corporation, Milford, MA) in 50 mM ammonium bicarbonate pH 7.8, vortexed, and boiled at 100 ℃ for 5 min. After cool ing, the lysate was reduced with 5 mM DTT, alkylated with 15 mM IAA and digested to peptides using trypsin (Promega, Madison, WI) at a substrate to enzyme ratio of 100:1 for four hours at 37 ℃ with shaking. The lysate was then treated with 200 mM HCl to remove Rapigest from the sample.

Mass Spectrometry

Mass spectrometry experiments were obtained from 14 runs sampling a *C. elegans* total protein digest on an LTQ mass spectrometer coupled to an Agilent 1100 liquid chromatography system. The LC-MS apparatus and chromatography gradient were run as described previously (2.2 above). Data was transformed into CRAWDAD MSMAT file format (Finney et al., 2008) using bin sizes of 1 m/z wide in mass range and 0.025' in retention time using MakeMS2.

Data Analysis

Two schemes were used to divide the analytical replicates into sample classes for pairwise differential comparison of LC-MS peak areas: *alternating*, comparing runs with odd run number to even run number and *first-last,* comparing the first seven to the last seven runs. This allows us to analyze the number of false positive differences found dependent upon the analytical effects sample run order.

Retention time alignment, LC-MS peak detection and grouping, chromatogram smoothing, and intensity normalization were performed as described previously (3.2 above). LC-MS runs were

trimmed to 25' to 90' in length, and peaks were required to have a minimum area-under-the-curve (AUC) of 5.0E5, and a minimum length of 0.5'. The background level was estimated by the *edge-to-lower* method (3.2.3 above). Peaks detected within a 0.5' window in at least 5 of seven runs in at least one sample class were joined using a greedy algorithm to form *peakgroups* (3.2.4 above) which were used for differential analysis of peak area between sample classes.

Peakgroup features from the LC-MS runs were compared using two different groupings to check the effects of run order: P-values for differential abundance between groups were generated by three methods: *parametric*, *inter-group,* and *intra-group*. In each method, Welch's unpaired two-sample t-statistic was calculated as follows: define $\bar{x}_{i,g}$ and $\sigma^2_{i,g}$ as the sample mean and variance for the *i*th peakgroup from group *g.* We then calculate an unpaired t-statistic assuming equivalent variance for comparing two given groups 1,2 for peakgroup *i* as:

$$t_i = \frac{\bar{x}_{i,1} - \bar{x}_{i,2}}{\sqrt{\frac{(n_1-1)\sigma^2_{i,1} + (n_2-1)\sigma^2_{i,2}}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

In the *parametric* case, p-values are calculated using Student's *t*-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. Both non-parametric methods use permutation testing on the observed peak areas. In the *intra-group* case, we simply enumerate the absolute value of the test-statistic over all permutations of the labels creating a set of null t-statistics $\boldsymbol{T_i}$, and the p-value is calculated as $\frac{1+\#\{\boldsymbol{T_i} > |t_i|\}}{B}$ where *B* is the total number of permutations.

In the *inter-group* approach, as developed by Storey (Storey and Tibshirani, 2003), we calculate a distribution of t-statistics by drawing *B* randomized permutations of the labels, and creating a list $\boldsymbol{T}$ of null $|t|$ for all *M* features over the *B* permutations. The default value of *B* is set to 100.

The p-value is calculated as $\frac{\#\{\boldsymbol{T} \geq |t_i|\}}{B \cdot M}$, using the rank of the observed t-statistic t_i over the set of null t-statistics created by *B* permutations of the *M* peakgroups. Below, we compare results from these three methods (*inter-group, intra-group*, parametric t-test) of calculating p-values for putatively null data.

4.3. Results and Discussion

We detected 33269 peakgroups in the 1000 m/z bins with the *first-last* sample group scheme, and 32729 peakgroups using the *alternating* scheme. While the runs used in each grouping were the same, the small difference in the number of peakgroups could be accounted for by the criteria that at least 5/7 of one of the two groups have peaks represented for it to be a valid peakgroup. We analyzed the validity of p-values calculated by the parametric and two non-parametric (*inter-group, intra-group*) schemes described above by checking their distribution against the property that p-values from data which is truly null should have a uniform distribution over the interval (0,1).

We calculated p-values as described above, and compared them against their rank normalized over the range $[\frac{1}{M}, 1]$ in Figure 4-1. When the runs were run in the *alternating* order, both permutation methods had p-value distributions close to uniform, although low p-values were slightly under-represented. All methods of calculating the p-value were slightly conservative by these approaches, with the p-value distributions from the two non-parametric methods are visually similar, indicating near-identical behavior. In the *first-last* arrangement, all methods significantly differ from the uniform p-value distribution, and appear to mimic the pattern of an experiment with a set of truly differing samples between the groups (Figure 4-2).

Notably, there is a bias for decreasing abundance as the number of the run increases. By comparing the total ion current of runs, or all signals extracted in a run, we can see that the number of detected ions drops off with the later runs in the experiment (Figure 4-3). This effect is an example of an external bias that can be minimized by accounting for experiment run order, and has led to our use of alternating run order between sample classes(2 classes) or randomizing the order of each group in a batch of replicates (>2 classes). The spurious increase

in ‘significant’ p-values from these data without true differences in the *first-last* sample class scheme (Figure 4-2, panel B) indicates that the decrease in overall TIC from the early runs to the later runs can cause a spurious increase in the number of changes detected between the groups.

The calculation of a p-value by permuting all class labels of replicates, as in the *intra-group* is well established in statistics, but suffers from low p-value resolution and a high minimum p-value when a limited number of samples are compared (Knijnenburg et al., 2009). The number of unique absolute values of the t-statistic calculable by permutation is $\frac{\binom{2N}{N}}{2}$ for a single feature comparing two replicates with groups of size ***N***. For example, 1716 permutations are for the groups above with two groups of 7, and only 126 for two groups of 5. The limited range of distinct p-values lowers the power of a test implicitly by setting a minimum p-value of 1 / #permutations. Furthermore, the limited resolution of p-values and complicates the calculation of FDR and q-values (Guo and Pan, 2005).

However, calculating a p-value using the *inter-group* method, where information is borrowed across features detected in the LC-MS runs, removes the limitation of a low number of discrete p-values. This prompted the comparison of both permutation-based methods, as in this case there are sufficient replicates with the intra-group method for fine-grained p-values with 1/1716 (5.8E-4) as a minimum p-value. The p-values for the alternate-ordered experiment in panel A of Figure 4-1 show a pattern consistent with being properly calculated from a null distribution, with the exception of the lower-ranked p-values being slightly conservative. The p-value results from the parametric test are indicative of a lack of power, which is expected when applied to data which does not fit the assumption of normality.

Notably, both permutation methods produce p-value distributions that are highly similar, and a Kolmogorov-Smirnov (KS) test for similarity between the *inter-group* and *intra-group* distributions accepts the null hypothesis that they are sampled from equivalent distributions (*first-last* p-value: 0.9624; *alternating* p-value: 0.9973). In summary we have shown that the permutation test defined in (Storey and Tibshirani, 2003) performs similarly to the classical permutation test on null mass spectrometry data, suggesting it is appropriate for use in differential mass spectrometry experiments.

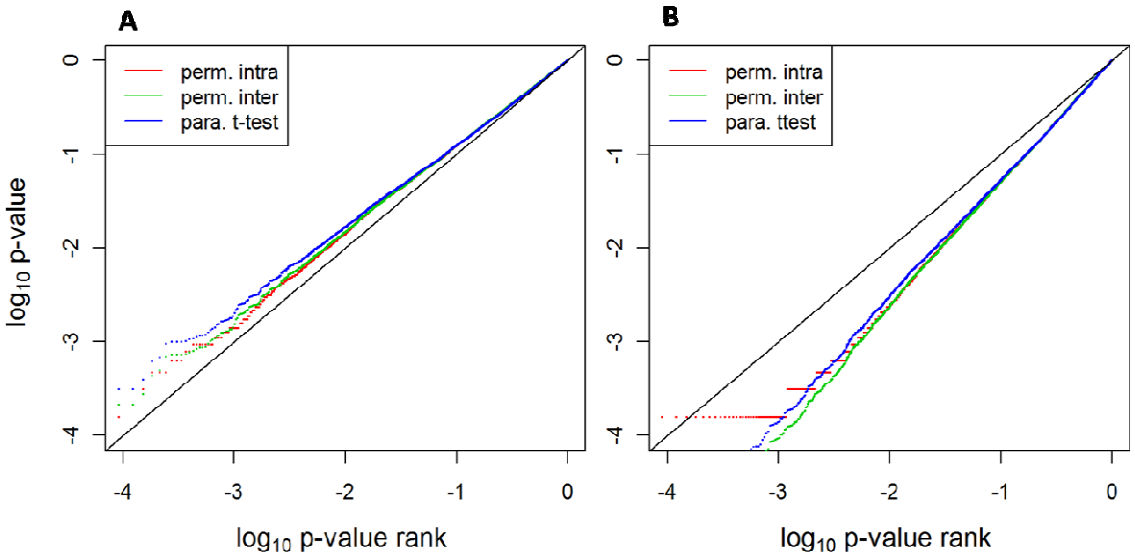

Figure 4-1: P-values calculated by parametric and non-parametric methods on null data

P-values calculated from the three methods described above are plotted against their $\log_{10}$ rank (scaled to 0 to 1). A) p-values from null LC-MS runs divided into groups by the *alternating* arrangement. B) p-values from null LC-MS runs divided into groups by the first and second half sequentially (*first-last*). In both plots, the p-value vs. rank for a uniform distribution are shown as a black line.

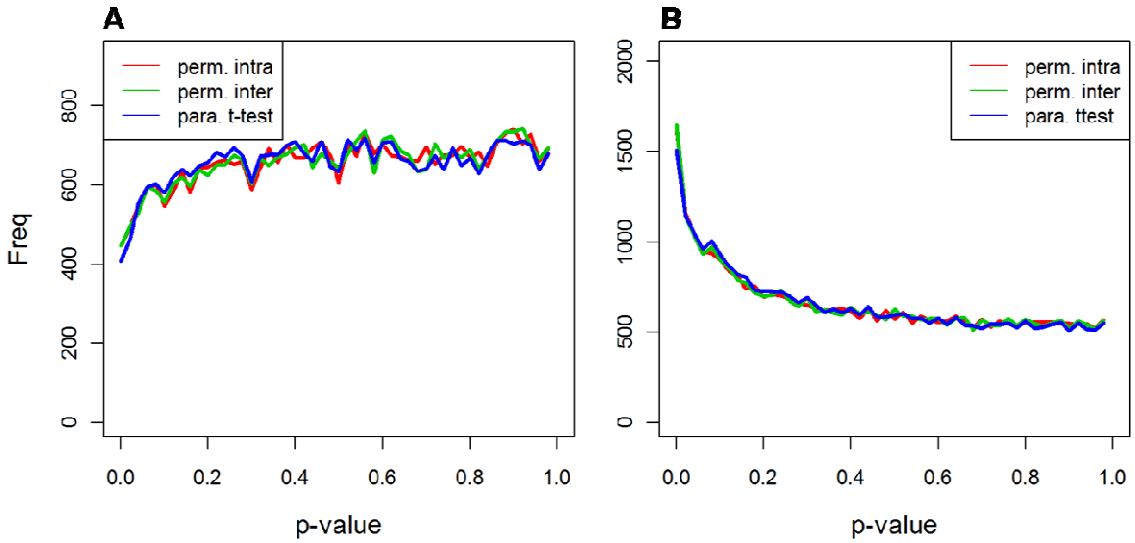

Figure 4-2: Histograms of p-values calculated by the non-parametric and parametric techniques

A) p-values from null LC-MS runs divided into groups by the alternating arrangement. B) p-values from null LC-MS runs divided into groups by the first and second half sequentially (first-last).

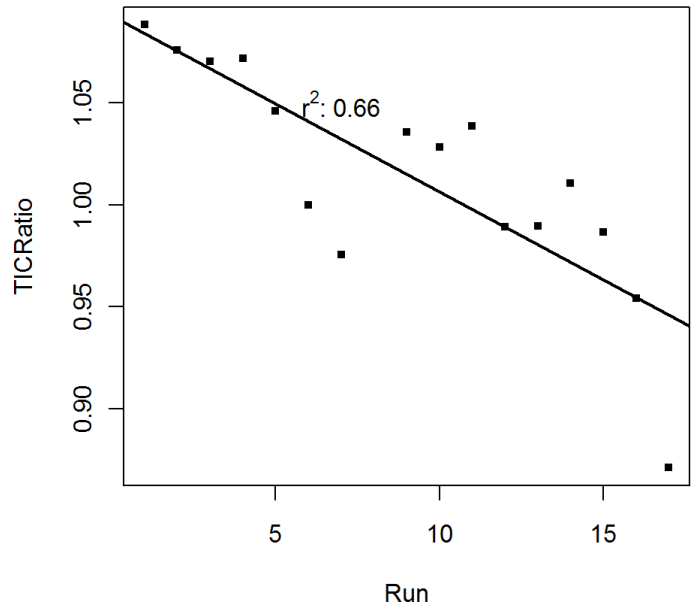

Figure 4-3: Total ion current (TIC) as a function of run order

TIC is shown as a function of the ratio of the current to that in run #6. The sum of TIC across all scans was compared between runs, giving a coefficient of determination (r2) of 0.66

Appendix 1: Efficient DTW Retention Time Alignment by Decimation of LC-MS Data

Introduction:

Retention time (RT) alignment of two LC-MS data runs using a dynamic time warping approach is time consuming. The algorithm is $O(MN^2)$ with regards to the length of the runs in *N* scans, and linear in respect to the *M* of m/z bins. However, it is possible to decrease the time taken by downsampling the data to a reduced size, thereby reducing M or N. We explore the tradeoffs in speed and alignment accuracy for reducing the dimensionality of the data in both retention time and m/z space.

Diminishing the number of scan comparisons made should greatly decrease the time spent in running the algorithm. However, this will also reduce the point-to-point resolution of the retention time mapping points, increasing error due to the coarseness of the output. Many prior alignment algorithms have shown positive results from aligning the total ion current (sum of abundances over all m/z at a given timepoint) of runs (Listgarten et al., 2007) – indicating that reducing the resolution of runs in m/z space can also reduce the run time significantly while retaining some alignment accuracy.

If alignments performed on data which is reduced in dimensionality provide an alignment approximately as good as the original data, then the dynamic time warping path from the reduced data set can be used to limit the possible search space in ΔRT when comparing the data at full resolution and size. We implement a simple procedure to constrain the retention time search space for the dynamic time warping algorithm presented in Chapter 2 and compare the time speedup to any loss of accuracy.

More specifically, we determine the maximum retention time shift in the data of reduced resolution, and use this to constrain the search space in ΔRT in our DTW approach. This idea has been extended further in the FastDTW project, which has been applied to multiple data types (Salvador and Chan, 2007).

Methods:

Data is decimated by converting the stored series of MS1 scans to down-sampled scans in either retention time or m/z dimensions. In the retention time case, non-sliding adjacent windows of observed scans spanning size N are averaged, where for every m/z bin m, a new scan is created whose values are the arithmetic mean of $\{ S_i , S_{i+1} \ldots S_{i+(N-1)} \}$, where scan i occurs every N scans. The retention time of the interpolated scans is set to the midpoint of the integrated range. This results in a lesser number of scans with the same m/z resolution.

Data is decimated in the m/z dimension by increasing the size of the bins used in the observed data, and using the same binning procedure as outlined in chapter 2, which briefly is to sum abundance values over windows of size M in m/z. Alignment was performed as described in Chapter 3, and alignment quality assessed by the median value in retention time deviation of persistent peptide isotope distributions as described in 2.2.7.

The alignment of LC-MS runs from the HEK/*E. coli* dataset referenced above (Chapter 3) were compared while varying m/z binning in size over a range of {1,2.5,5,10..100,250}. Runs were also compared using retention time binning of {1,5,10,11,12..20,25..45} scan sizes. Alignment quality was scored by comparing the improvement in mean alignment standard deviation from the non-decimated data to the decimated version as $\frac{\Delta RT_{decimated}}{\Delta RT_{original}}$, where $\Delta RT_{decimated}$is the change in mean retention time of the PPID markers after alignment for the decimated data.

Alignments were run using CRAWDAD on a custom built server with Intel Xeon® 5140 processors running at 2.33 GHz and 8GB of RAM.

Results and Discussion:

Alignment quality was maintained when using *m/z* decimation up to using bins of 100 *m/z* in size, but decreased in a linear fashion using decimation in the retention time dimension (Figure A2-1). The initial undecimated run completed in 2397 seconds, while runs using a decimation from bins of size 1/24 *m/z* to 100 *m/z* completed in 16.6 seconds, with a relative alignment quality of 93% of the original. Note that these alignment times are independent of the time taken to decimate the data, Based on these results, it would be recommended to use CRAWDAD while decimating the data at 100 *m/z* before alignment, or to explore the retention time decimation amounts in the neighborhood of 100 *m/z.*

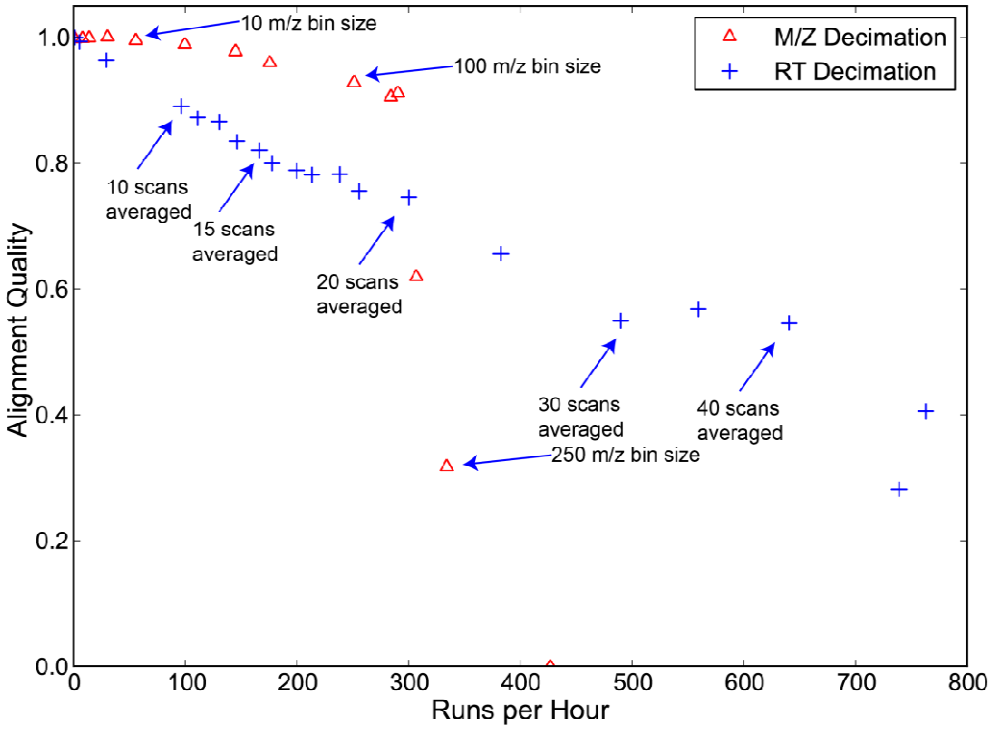

Figure A2-1: Alignment quality as a function of speed dependent on decimation

Alignment quality is measured as a fraction of the undecimated alignment quality, ranging from 0.0 (no improvement over unaligned) to 1.0 (same alignment improvement as undecimated alignment). Alignment quality is plotted against alignment speed (alignments per hour). Runs were decimated in retention time decimation (crosses), measured in windows of scans over a range of 2-50 scans. The relationship was also plotted as an effect of *m/z* decimation (red triangles), with bins of 1 to 250 *m/z*.

Appendix 2: MSMAT file format

Overview

The use of multiple data files for the analysis of a large set of LC-MS technical replcates in differential proteomics experiments creates constraints where not all data files can be held in memory. A binary file format was chosen to store binned LC-MS data for CRAWDAD to facilitate quick extraction of extracted ion chromatograms for statistical analysis. Existing file formats such as mzXML are based on storing scans - hence the development of a file format which stores data as extractetd ion chromtograms (XIC) in a binary format to afford rapid access. This had been designed for data which is binned in the m/z dimension -- hence, data is stored as a binary matrix. It can be stored in XIC row-order (default), or scan row-order, depending on what type of extraction needs to be done quickly.

Organization

MSMAT files consist of an identifying plaintext line, followed by a header which encodes information about binned m/z, retention time values, and other metadata about the LC-MS run. Scans or XICs follow -- as data is stored as an equal number of m/z bins, or retention times, the basic unit of data is the same length, removing the need for an index -- the offset of any scan or XIC can be quickly calculated.

Byte Encoding

Binary values are currently encoded in little-endian format. Future support will work for big-endian or network- byte orders.

Header

MSMAT files begin with a plaintext line of '_MSMAT_V#_' where # is a version number. Lines in the header are terminated with the '\n' character (UNIX newline). A series of header fields follows, stored as:

FIELD_NAME:BYTE_LEN,BINARY_DATA\n

Where FIELD_NAME is a plaintext label for a field (fields defined below), BINARY_DATA is binary data, and BYTE_LEN the length of the binary data in bytes, stored as ASCII digits. The field is terminated with a newline character.

The header is terminated with a '_END_MSMAT_HEADER_' terminated with a newline character. Binary data encoding scans or XICs follows until the end of the file.

Header Data Types

MSMAT header fields can encode strings, arrays of floats, associative arrays of floats, or associate arrays of strings to floats

STRING: text encoded as an ASCII string -- not null terminated

FLOAT: a single floating point value encoded as an IEEE-754 32-bit value – (i.e. typical ‘float’ value in C).BYTE_LEN (see above) will be 4.

FLOAT_ARRAY: an array of 32-bit floating point values, stored contiguously.

FLOAT_FLOAT_MAP: an array of 32-bit floa-ting point keys, immediately followed by an equally sized array of 32-bit floating point values

STR_FLOAT_MAP: an array of strings serving as keys, followed by an equally sized array of floating point values. As the strings are not of a predefined length we need a more complicated encoding scheme as follows:

total_bytes,string_bytes<1>str1<0>str2<0>float_bytes<1>float1float2

total_bytes -- the complete size of this field, including all separator characters

string_bytes -- The length in bytes of the string field, terminated with a character of value one (i.e. (char)1 in C)

float_bytes -- The length in bytes of the float field, terminated with a (char)1 character. str1, str2 -- byte strings terminated with a null character float1, float2 -- floating point values, not terminated due to fixed size

Example:

```
100,60<1>abc<0>def<0>ghi<0>30<1>flt1flt2flt3
```

This maps keys to values as follows { “abc”,”def”,”ghi” } map to their equivalently-indexed counterparts in {flt1,flt2,flt3} where the fltNs are the machine single-precision floating-point representation of the values, and <N> refers to a byte with value N. [does this end with a null or specially valued byte?

Header Fields

The ordering of header fields is not defined. New fields can be introduced -- the existing parser passes over field labels that it does not recognize until the '_END_MSMAT_HEADER_' token is encountered

- array_type : Defines whether the binned LC-MS intensity values are stored as scans or XICs. Values are 'scans' or 'chroms'. Type: STRING
- bin_size : Defines the (evenly spaced) m/z bin size. Type: FLOAT
- mzs : Lists the binned m/z values. Type: FLOAT_ARRAY
- rts : Lists the retention time value of scans. Type: FLOAT_ARRAY
- ort_wrt_map : An associative array of original (or unmodified) retention times ('ort') vs. warped retention times ('wrt'). Type: FLOAT_FLOAT_MAP
- audit_trail : An XML-format audit trail of actions performed on the MSMAT file. Currently in development -- a previous incarnation used YAML
- mzstr_id_map : reserved for future use. Type: STR_FLOAT_MAP

Comments

While the MSMAT header format does not have the ease of use or full self-documenting nature of XML, it still retains extensibility. Also, a field could be stored as XML itself -- an example application would be for an audit trail on actions perfomed on the file.

The use of floats to store the mzs data may be problematic in the future – The increment between adjacent single-precision floating point values has a precision of ~ 0.06 ppm at a value

of 1000, which limits the accuracy of m/z expressed with floats. It would be better to add a double precision type.

The precision of floats at 1000.0 m/z was calculated with the following C99 program:

```
#include "math.h"
#include <stdio.h>
int main() {
 float base = 1000.0f;
 float next = nextafterf(base,base+1);
 printf("%30.30f\n",base);
 printf("%30.30f\n",next);
 return 0;
}
```

Implementation

Reading / Writing code has been implemented in the C++ and Java programming language.

List of References

Abdi,H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In Encyclopedia of Measurement and Statistics, N.J.Salkind, ed. (Thousand Oaks: Sage).

America,A.H., Cordewener,J.H., van Geffen,M.H., Lommen,A., Vissers,J.P., Bino,R.J., and Hall,R.D. (2006). Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. Proteomics. *6*, 641-653.

Anderle,M., Roy,S., Lin,H., Becker,C., and Joho,K. (2004). Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. Bioinformatics *20*, 3575-3582.

Anderson,N.L. and Anderson,N.G. (1998). Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis *19*, 1853-1861.

Anderson,N.L. and Anderson,N.G. (2002). The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics *1*, 845-867.

Andreev,V.P., Li,L., Cao,L., Gu,Y., Rejtar,T., Wu,S.L., and Karger,B.L. (2007). A new algorithm using cross-assignment for label-free quantitation with LC-LTQ-FT MS. J. Proteome Res. *6*, 2186-2194.

Andreev,V.P., Rejtar,T., Chen,H.S., Moskovets,E.V., Ivanov,A.R., and Karger,B.L. (2003). A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. Anal Chem *75*, 6314-6326.

Annesley,T.M. (2003). Ion suppression in mass spectrometry. Clin Chem *49*, 1041-1044.

Arnott,D., Kishiyama,A., Luis,E.A., Ludlum,S.G., Marsters,J.C., and Stults,J.T. (2002). Selective detection of membrane proteins without antibodies: a mass spectrometric version of the western blot. Mol Cell Proteomics *1*, 148-156.

Bantscheff,M., Schirle,M., Sweetman,G., Rick,J., and Kuster,B. (2007). Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem *389*, 1017-1031.

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., and McIntosh,M. (2006). A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. Bioinformatics. *22*, 1902-1909.

Benjamini,Y. and Hochberg,Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289-300.

Bern,M., Finney,G., Hoopmann,M.R., Merrihew,G., Toth,M.J., and MacCoss,M.J. (2010). Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. Anal Chem *82*, 833-841.

Blackler,A.R., Klammer,A.A., MacCoss,M.J., and Wu,C.C. (2006). Quantitative comparison of proteomic data quality between a 2D and 3D quadrupole ion trap. Anal Chem *78*, 1337-1344.

Bogdanov,B. and Smith,R.D. (2005). Proteomics by FTICR mass spectrometry: top down and bottom up. Mass Spectrom Rev *24*, 168-200.

Braun,R.J., Kinkl,N., Beer,M., and Ueffing,M. (2007). Two-dimensional electrophoresis of membrane proteins. Anal Bioanal Chem *389*, 1033-1045.

Callister,S.J., Barry,R.C., Adkins,J.N., Johnson,E.T., Qian,W.J., Webb-Robertson,B.J.M., Smith,R.D., and Lipton,M.S. (2006). Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. J. Proteome Res. *5*, 277-286.

Cappadona,S., Levander,F., Jansson,M., James,P., Cerutti,S., and Pattini,L. (2008). Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry. Anal Chem *80*, 4960-4968.

Carvalho,P.C., Xu,T., Han,X., Cociorva,D., Barbosa,V.C., and Yates,J.R. (2009). YADA: a tool for taking the most out of high-resolution spectra. Bioinformatics *25*, 2734-2736.

Clauser,K.R., Baker,P., and Burlingame,A.L. (1999). Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem *71*, 2871-2882.

Cooper,S.J., Finney,G.L., Brown,S.L., Nelson,S.K., Hesselberth,J., MacCoss,M.J., and Fields,S. (2010). High-throughput profiling of amino acids in strains of the Saccharomyces cerevisiae deletion collection. Genome Res *20*, 1288-1296.

Danielsson,R., Bylund,D., and Markides,K.E. (2002). Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography-mass spectrometry. Analytica Chimica Acta *454*, 167-184.

de Godoy,L.M.F., Olsen,J.V., Cox,J., Nielsen,M.L., Hubner,N.C., Frohlich,F., Walther,T.C., and Mann,M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251-1254.

Dierckx,P. (1993). Curve and Surface Fitting with Splines. Oxford University Press).

Dierckx,P. (1982). A fast algorithm for smoothing data on a rectangular grid while using spline functions. SIAM. J. Numer. Anal. *19*, 1286-1304.

Eakin,C.M., MacCoss,M.J., Finney,G.L., and Klevit,R.E. (2007). Estrogen receptor alpha is a putative substrate for the BRCA1 ubiquitin ligase. Proc. Natl. Acad. Sci. U. S. A. *104*, 5794-5799.

Eilers,P.H.C. (2003). A perfect smoother. Anal Chem *75*, 3631-3636.

Eng,J.K., McCormack,A.L., and Yates III,J.R. (1994a). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976-989.

Fang,R., Elias,D.A., Monroe,M.E., Shen,Y., McIntosh,M., Wang,P., Goddard,C.D., Callister,S.J., Moore,R.J., Gorby,Y.A., Adkins,J.N., Fredrickson,J.K., Lipton,M.S., and Smith,R.D. (2006). Differential label-free quantitative proteomic analysis of Shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach. Mol. Cell Proteomics. *5*, 714-725.

Fenn,J.B., Mann,M., Meng,C.K., Wong,S.F., and Whitehouse,C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science *246*, 64-71.

Finney,G.L., Blackler,A.R., Hoopmann,M.R., Canterbury,J.D., Wu,C.C., and MacCoss,M.J. (2008). Label-Free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution μLC-MS Data. Anal Chem *80*, 961-971.

Fischer,B., Grossmann,J., Roth,V., Gruissem,W., Baginsky,S., and Buhmann,J.M. (2006). Semi-supervised LC/MS alignment for differential proteomics. Bioinformatics (Oxford, England) *22*, e132-e140.

Fraga,C.G., Prazen,B.J., and Synovec,R.E. (2001). Objective data alignment and chemometric analysis of comprehensive two-dimensional separations with run-to-run peak shifting on both dimensions. Anal. Chem. *73*, 5833-5840.

Frank,A.M., Savitski,M.M., Nielsen,M.L., Zubarev,R.A., and Pevzner,P.A. (2007). De novo peptide sequencing and identification with precision mass spectrometry. J. Proteome Res. *6*, 114-123.

Gorry,P.A. (1990). General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. Anal. Chem. *62*, 570-573.

Griffin,N.M., Yu,J., Long,F., Oh,P., Shore,S., Li,Y., Koziol,J.A., and Schnitzer,J.E. (2010). Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol *28*, 83-89.

Guo,X. and Pan,W. (2005). Using weighted permutation scores to detect differential gene expression with microarray data. J Bioinform Comput Biol *3*, 989-1006.

Gygi,S.P., Corthals,G.L., Zhang,Y., Rochon,Y., and Aebersold,R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. PNAS *97*, 9390-9395.

Han,D.K., Eng,J., Zhou,H., and Aebersold,R. (2001). Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotech *19*, 946-951.

Hoopmann,M.R., Finney,G.L., and MacCoss,M.J. (2007). High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Anal. Chem. *79*, 5620-5632.

Hoopmann,M.R., Merrihew,G.E., von Haller,P.D., and MacCoss,M.J. (2009). Post analysis data acquisition for the iterative MS/MS sampling of proteomics mixtures. J. Proteome Res. *8*, 1870-1875.

Horn,D.M., Zubarev,R.A., and McLafferty,F.W. (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. J Am. Soc. Mass Spectrom *11*, 320-332.

Hsieh,E.J., Hoopmann,M.R., MacLean,B., and MacCoss,M.J. (2009). Comparison of database search strategies for high precursor mass accuracy MS/MS data. J. Proteome Res. *9*, 1138-1143.

Hu,Q., Noll,R.J., Li,H., Makarov,A., Hardman,M., and Graham,C.R. (2005). The Orbitrap: a new mass spectrometer. J. Mass Spectrom. *40*, 430-443.

Jaffe,J.D., Mani,D.R., Leptos,K.C., Church,G.M., Gillette,M.A., and Carr,S.A. (2006). PEPPeR, a platform for experimental proteomic pattern recognition. Mol. Cell Proteomics. *5*, 1927-1941.

Jaffe,J.D., Keshishian,H., Chang,B., Addona,T.A., Gillette,M.A., and Carr,S.A. (2008). Accurate inclusion mass screening: a bridge from unbiased discovery to targeted assay development for biomarker verification. Mol. Cell Proteomics *7*, 1952-1962.

Johnson,K.J., Wright,B.W., Jarman,K.H., and Synovec,R.E. (2003). High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. Journal of Chromatography A *996*, 141-155.

Kall,L., Canterbury,J.D., Weston,J., Noble,W.S., and MacCoss,M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods. *4*, 923-925.

Kall,L., Storey,J.D., and Noble,W.S. (2009). QVALITY: non-parametric estimation of q-values and posterior error probabilities. Bioinformatics (Oxford, England) *25*, 964-966.

Karpievitch,Y.V., Taverner,T., Adkins,J.N., Callister,S.J., Anderson,G.A., Smith,R.D., and Dabney,A.R. (2009). Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. Bioinformatics (Oxford, England) *25*, 2573-2580.

Katajamaa,M. and Oresic,M. (2005). Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics. *6:179.*, 179.

Kebarle,P. and Verkerk,U.H. (2009). Electrospray: from ions in solution to ions in the gas phase, what we know now. Mass Spectrom Rev *28*, 898-917.

Kellie,J.F., Catherman,A.D., Durbin,K.R., Tran,J.C., Tipton,J.D., Norris,J.L., Witkowski,C.E., Thomas,P.M., and Kelleher,N.L. (2012). Robust analysis of the yeast proteome under 50 KDa by molecular-mass-based fractionation and top-down mass spectrometry. Anal Chem *84*, 209-215.

Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M., and Karp,P.D. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res. *33*, D334-D337.

Klammer,A.A. and MacCoss,M.J. (2006). Effects of modified digestion schemes on the identification of proteins from complex mixtures. J. Proteome Res. *5*, 695-700.

Kline,K.G., Finney,G.L., and Wu,C.C. (2009). Quantitative strategies to fuel the merger of discovery and hypothesis-driven shotgun proteomics. Briefings in Functional Genomics & Proteomics *8*, 114-125.

Knijnenburg,T.A., Wessels,L.F.A., Reinders,M.J.T., and Shmulevich,I. (2009). Fewer permutations, more accurate p-values. Bioinformatics *25*, i161-i168.

Kohlbacher,O., Reinert,K., Gropl,C., Lange,E., Pfeifer,N., Schulz-Trieglaff,O., and Sturm,M. (2007). TOPP--the OpenMS proteomics pipeline. Bioinformatics. *23*, e191-e197.

Kruger,M., Moser,M., Ussar,S., Thievessen,I., Luber,C.A., Forner,F., Schmidt,S., Zanivan,S., Fässler,R., and Mann,M. (2008). SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. Cell *134*, 353-364.

Kwon,D., Vannucci,M., Song,J.J., Jeong,J., and Pfeiffer,R.M. (2008). A novel wavelet-based thresholding method for the pre-processing of mass spectrometry data that accounts for heterogeneous noise. Proteomics *8*, 3019-3029.

Lan,K. and Jorgenson,J.W. (2001). A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks. Journal of Chromatography A *915*, 1-13.

Lange,E., Gropl,C., Schulz-Trieglaff,O., Leinenbach,A., Huber,C., and Reinert,K. (2007). A geometric approach for the alignment of liquid chromatography-mass spectrometry data. Bioinformatics. *23*, i273-i281.

Lange,E., Gröpl,C., Reinert,K., Kohlbacher,O., and Hildebrandt,A. (2006). High-accuracy peak picking of proteomics data using wavelet techniques. Pac Symp Biocomput 243-254.

Li,J. (2002). Comparison of the capability of peak functions in describing real chromatographic peaks. Journal of Chromatography A *952*, 63-70.

Li,X.J., Yi,E.C., Kemp,C.J., Zhang,H., and Aebersold,R. (2005). A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. Mol. Cell. Proteomics *4*, 1328-1340.

Licklider,L.J., Thoreen,C.C., Peng,J., and Gygi,S.P. (2002). Automation of nanoscale microcapillary liquid chromatography/tandem mass spectrometry with a vented column. Anal. Chem. *74*, 3076-3083.

Listgarten,J., Neal,R.M., Roweis,S.T., Wong,P., and Emili,A. (2007). Difference detection in LC-MS data for protein biomarker discovery. Bioinformatics *23*, e198-e204.

Liu,H., Sadygov,R.G., and Yates,J.R., III (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. *76*, 4193-4201.

MacCoss,M.J., Toth,M.J., and Matthews,D.E. (2001). Evaluation and optimization of ion-current ratio measurements by selected-ion-monitoring mass spectrometry. Anal. Chem. *73*, 2976-2984.

MacCoss,M.J., Wu,C.C., Liu,H., Sadygov,R., and Yates,J.R., III (2003). A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. Anal. Chem. *75*, 6912-6921.

MacCoss,M.J. and Yates,J.R., III (2001). Proteomics: analytical tools and techniques. Curr. Opin. Clin. Nutr. Metab Care *4*, 369-375.

MacLean,B., Tomazela,D.M., Shulman,N., Chambers,M., Finney,G.L., Frewen,B., Kern,R., Tabb,D.L., Liebler,D.C., and MacCoss,M.J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics (Oxford, England) *26*, 966-968.

Makarov,A., Denisov,E., Kholomeev,A., Balschun,W., Lange,O., Strupat,K., and Horning,S. (2006). Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal. Chem. *78*, 2113-2120.

Mason,C.J., Johnson,K.L., and Muddiman,D.C. (2005). Reproducibility of retention time using a splitless nanoLC coupled to an ESI-FTICR mass spectrometer. J. Biomol. Tech. *16*, 414-422.

Mayr,B.M., Kohlbacher,O., Reinert,K., Sturm,M., GrÃ¶pl,C., Lange,E., Klein,C., and Huber,C.G. (2006). Absolute myoglobin quantitation in serum by combining two-dimensional liquid chromatography-electrospray ionization mass spectrometry and novel data analysis algorithms. J. Proteome Res. *5*, 414-421.

Mayya,V., Rezaul,K., Cong,Y.S., and Han,D. (2005). Systematic comparison of a two-dimensional ion trap and a three-dimensional ion trap mass spectrometer in proteomics. Mol. Cell. Proteomics *4*, 214-223.

McDonald,W.H., Tabb,D.L., Sadygov,R.G., MacCoss,M.J., Venable,J., Graumann,J., Johnson,J.R., Cociorva,D., and Yates,J.R., III (2004). MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Commun. Mass Spectrom. *18*, 2162-2168.

Meng,F.Y., Wiener,M.C., Sachs,J.R., Burns,C., Verma,P., Paweletz,C.P., Mazur,M.T., Deyanova,E.G., Yates,N.A., and Hendrickson,R.C. (2007). Quantitative analysis of complex peptide mixtures using FTMS and differential mass spectrometry. J Am Soc Mass Spectrom *18*, 226-233.

Metz,C.E. (1978). Basic principles of roc analysis. Semin. Nucl. Med. *8*, 283-298.

Muddiman,D.C., Huang,B.M., Anderson,G.A., Rockwood,A., Hofstadler,S.A., Weir-Lipton,M.S., Proctor,A., Wu,Q., and Smith,R.D. (1997). Application of sequential paired covariance to liquid chromatography-mass spectrometry data enhancements in both the signal-to-noise ratio and the resolution of analyte peaks in the chromatogram. Journal of Chromatography A *771*, 1-7.

Mueller,L.N., Brusniak,M.Y., Mani,D.R., and Aebersold,R. (2008). An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J. Proteome Res. *7*, 51-61.

Mueller,L.N., Rinner,O., Schmidt,A., Letarte,S., Bodenmiller,B., Brusniak,M.Y., Vitek,O., Aebersold,R., and Mueller,M. (2007). SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. Proteomics *7*, 3470-3480.

Myers,C., Rabiner,L., and Rosenberg,A. (1980). Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE T. Acoust. Speech. *28*, 623-636.

O'Farrell,P.H. (1975). High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. *250*, 4007-4021.

Oberg,A. and Vitek,O. (2009). Statistical design of quantitative mass spectrometry-based proteomic profiling experiments. J. Proteome Res. *8*, 2144-2156.

Old,W.M., Meyer-Arendt,K., Aveline-Wolf,L., Pierce,K.G., Mendoza,A., Sevinsky,J.R., Resing,K.A., and Ahn,N.G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol. Cell. Proteomics *4*, 1487-1502.

Olsen,J.V., de Godoy,L.M., Li,G., Macek,B., Mortensen,P., Pesch,R., Makarov,A., Lange,O., Horning,S., and Mann,M. (2005). Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol. Cell Proteomics. *4*, 2010-2021.

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

Palagi,P.M., Hernandez,P., Walther,D., and Appel,R.D. (2006). Proteome informatics I: bioinformatics tools for processing experimental data. Proteomics *6*, 5435-5444.

Pavlidis,P. and Noble,W.S. (2003). Matrix2png: a utility for visualizing matrix data. Bioinformatics. *19*, 295-296.

Prakash,A., Mallick,P., Whiteaker,J., Zhang,H.D., Paulovich,A., Flory,M., Lee,H., Aebersold,R., and Schwikowski,B. (2006a). Signal maps for mass spectrometry-based comparative proteomics. Mol. Cell. Proteomics *5*, 423-432.

Prakash,A., Sutton,J., Richmond,T., Piening,B., Whiteaker,J., Zhang,H., Paulovich,A., Watts,J., Martin,D., Goodlett,D., Aebersold,R., Schwikowski,B., and Bonilla,L. (2006b). Assessing reproducibility of mass spectrometry experiments for biomarker discovery in human plasma. Mol. Cell. Proteomics *5*, S280.

Prince,J.T. and Marcotte,E.M. (2006). Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal Chem *78*, 6140-6152.

Renart,J., Reiser,J., and Stark,G.R. (1979). Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. PNAS *76*, 3116-3120.

Rey,S., Acab,M., Gardy,J.L., Laird,M.R., deFays,K., Lambert,C., and Brinkman,F.S. (2005). PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res. *33*, D164-D168.

Rigo, A. Representation-based just-in-time specialization and the psyco prototype for python. 15-26. 2004. ACM Press. PEPM '04: Proceedings of the 2004 ACM SIGPLAN symposium on Partial evaluation and semantics-based program manipulation.

Sadygov,R.G., Maroto,F.M., and Huhmer,A.F. (2006). ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. Anal. Chem. *78*, 8207-8217.

Salvador, S and Chan, P. Toward accurate dynamic time warping in linear time and space (2006). Intelligent Data Analysis *11*, 561-580.

Sakoe,H. and Chiba,S. (1978). Dynamic-programming algorithm optimization for spoken word recognition. IEEE T. Acoust. Speech. *26*, 43-49.

Sankoff,D. and Kruskal,J. (1983). Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison. (New York: *Addison Wesley*).

Savitzky,A. and Golay,M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. *36*, 1627-1639.

Schlosser,A. and Volkmer-Engert,R. (2003). Volatile polydimethylcyclosiloxanes in the ambient laboratory air identified as source of extreme background signals in nanoelectrospray mass spectrometry. Journal of Mass Spectrometry *38*, 523-525.

Schulz-Trieglaff,O., Machtejevas,E., Reinert,K., Schlüter,H., Thiemann,J., and Unger,K. (2009). Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments. BioData Min *2*, 4.

Sinha,A.E., Hope,J.L., Prazen,B.J., Fraga,C.G., Nilsson,E.J., and Synovec,R.E. (2004). Multivariate selectivity as a metric for evaluating comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry subjected to chemometric peak deconvolution. Journal of Chromatography A *1056*, 145-154.

Smith,C.A., Want,E.J., O'Maille,G., Abagyan,R., and Siuzdak,G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem *78*, 779-787.

Storey,J.D. and Tibshirani,R. (2003). Statistical significance for genomewide studies. PNAS *100*, 9440-9445.

Synovec,R.E., Prazen,B.J., Johnson,K.J., Fraga,C.G., and Bruckner,C.A. (2003). Chemometric analysis of comprehensive two-dimensional separations. Adv. Chromatogr. *42*, 1-42.

Tabata,T., Sato,T., Kuromitsu,J., and Oda,Y. (2007). Pseudo internal standard approach for label-free quantitative proteomics. Anal. Chem. ..

Tang,K., Page,J.S., and Smith,R.D. (2004). Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. J Am Soc Mass Spectrom *15*, 1416-1423.

The UniProt Consortium (2007). The Universal Protein Resource (UniProt). Nucleic Acids Res. *35*, D193-D197.

Tomasi,G., van den Berg,F., and Andersson,C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. Journal of Chemometrics *18*, 231-241.

Tran,J.C., Zamdborg,L., Ahlf,D.R., Lee,J.E., Catherman,A.D., Durbin,K.R., Tipton,J.D., Vellaichamy,A., Kellie,J.F., Li,M., Wu,C., Sweet,S.M.M., Early,B.P., Siuti,N., LeDuc,R.D., Compton,P.D., Thomas,P.M., and Kelleher,N.L. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature *480*, 254-258.

van Rijsbergen,C.V. (1979). Information Retrieval. (London; Boston: Butterworth).

Venable,J.D., Dong,M.Q., Wohlschlegel,J., Dillin,A., and Yates,J.R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat. Methods *1*, 39-45.

Wan,K.X., Vidavsky,I., and Gross,M.L. (2002). Comparing similar spectra: From similarity index to spectral contrast angle. J Am Soc Mass Spectrom *13*, 85-88.

Wang,C.P. and Isenhour,T.L. (1987). Time-warping algorithm applied to chromatographic peak matching gas-chromatography fourier-transform infrared mass-spectrometry. Anal Chem *59*, 649-654.

Wang,W., Zhou,H., Lin,H., Roy,S., Shaler,T.A., Hill,L.R., Norton,S., Kumar,P., Anderle,M., and Becker,C.H. (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal. Chem. *75*, 4818-4826.

Wasinger,V.C., Cordwell,S.J., Cerpa-Poljak,A., Yan,J.X., Gooley,A.A., Wilkins,M.R., Duncan,M.W., Harris,R., Williams,K.L., and Humphery-Smith,I. (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis *16*, 1090-1094.

Wiener,M.C., Sachs,J.R., Deyanova,E.G., and Yates,N.A. (2004). Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. Anal. Chem. *76*, 6085-6096.

Wilm,M. and Mann,M. (1996). Analytical properties of the nanoelectrospray ion source. Anal. Chem. *68*, 1-8.

Wong,J., Schwahn,A., and Downard,K. (2009). ETISEQ - an algorithm for automated elution time ion sequencing of concurrently fragmented peptides for mass spectrometry-based proteomics. Bmc Bioinformatics *10*, 244.

Wu,C.C., Dong,M.Q., and MacCoss,M.J. (2005). Quantitative proteomic analysis of mammalian organisms using metabolically labeled tissues. In Quantitative Proteomics by Mass Spectrometry, S.Sechi, ed. (Totowa, NJ: Humana Press, Inc.).

Wu,C.C. and MacCoss,M.J. (2002). Shotgun proteomics: tools for the analysis of complex biological systems. Curr. Opin. Mol. Ther. *4*, 242-250.

Xie,H.W. and Griffin,T.J. (2006). Trade-off between high sensitivity and increased potential for false positive peptide sequence matches using a two-dimensional linear ion trap for tandem mass spectrometry-based proteomics. J. Proteome Res. *5*, 1003-1009.

Yates,J.R., Cociorva,D., Liao,L., and Zabrouskov,V. (2006). Performance of a linear ion trap-Orbitrap hybrid for peptide analysis. Anal. Chem. *78*, 493-500.

Yates,N.A., Deyanova,E.G., Geissler,W., Wiener,M.C., Sachs,J.R., Wong,K.K., Thornberry,N.A., Roy,R.S., Settlage,R.E., and Hendrickson,R.C. (2007). Identification of peptidase substrates in human plasma by FTMS based differential mass spectrometry. Int J Mass Spectrom. *259*, 174-183.

Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S., Bussey,K.J., Riss,J., Barrett,J.C., and Weinstein,J.N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. *4*, R28.

Zhang,B., Chambers,M.C., and Tabb,D.L. (2007). Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J. Proteome Res. *6*, 3549-3557.

Zhang,J., Gonzalez,E., Hestilow,T., Haskins,W., and Huang,Y. (2009). Review of peak detection algorithms in liquid-chromatography-mass spectrometry. Curr. Genomics *10*, 388-401.

Zhang,R., Sioma,C.S., Wang,S., and Regnier,F.E. (2001). Fractionation of isotopically labeled peptides in quantitative proteomics. Anal. Chem. *73*, 5142-5149.

Zhang,Y., Wen,Z., Washburn,M.P., and Florens,L. (2010). Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. Anal Chem *82*, 2272-2281.

Zhou,H., Ranish,J.A., Watts,J.D., and Aebersold,R. (2002). Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. Nat. Biotechnol. *20*, 512-515.

Vita

Greg Finney was born in Honolulu, Hawaii, and grew up in Rome, Italy and Northern California. He earned a B.S. in Biochemistry and Molecular Biology at U.C. Santa Cruz, and spent some time in the biotechnology industry nearby, as well as brief stints as a software engineer. After enjoying a bit of computer science coursework at Chico State, he came up to beautiful grey Seattle to enjoy the science, seafood, and outdoors. He graduated with a Doctor of Philosophy in Genome Sciences in 2012 from the University of Washington.