Efficient Algorithms for Liquid Chromatography Coupled Mass Spectrometry Based Protein Quantification

Zia Khan

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Advisers: Mona Singh and Leonid Kruglyak

September 2011

Abstract

Identification of genes and genetic pathways affected in disease and disease treatment is one of the driving aims of studies that conduct comprehensive, quantitative surveys of proteins across many experimental samples and replicates. The prevailing tool for conducting such surveys is a class of experimental techniques and instrumentation known as liquid chromatography coupled tandem mass spectrometry (LC-MS/MS). LC-MS/MS generates large data sets in the form of thousands to millions of mass spectra in a single experiment. Converting these spectra into interpretable quantitative measurements of proteins, their peptide fragments, or enrichment and depletion of their post-translational modifications presents a substantial computational challenge. This thesis describes a new application of space partitioning data structures and a series of algorithms that leverage the fast geometric queries supported by these data structures to significantly improve the speed and quality LC-MS/MS data analysis. In addition, this thesis develops a collection of methods, implemented in an open source software system called PVIEW (http://compbio.cs.princeton.edu/pview), that use the output of these algorithms to enable accurate quantification of proteins, protein fragments, and post-translational modifications. These methods are evaluated with respect to their quantitative accuracy and computational efficiency on a wide range of experimental data sets spanning several experimental methodologies and source protein samples.

Acknowledgments

I want to thank my parents and my cousin Selma for their unyeilding support. I want to thank my brother, even though he doesn't talk to me right now and hasn't for over a year. He's been successful in all the ways I feel like a failure. I want to thank all of the mentors I've had along the way: Steve Bunes, Jerry & Leslie Nykel, Javier Lopez, John Rawlins, Jimmy Burnette, Susan Henry, Deborah Gordon, Tucker Balch, Frank Dellaert, Harpreet Swahney.... I know I'm forgetting a few people. Sorry :-). I would like to thank my advisors Mona Singh and Leonid Kruglyak for the tremendous freedom that they gave me to pursue my ideas and interests. I want to thank my committee members Ben Garcia, Hilary Coller, and Saeed Tavazoie for taking the time to help me out with graduating. Anyway, I look forward to my graduation present to myself for finishing this thing: a new motorcycle. As always, it took longer than I expected. ZK was supported by the first year Princton University Department of Computer Science graduate fellowship (year 1), a teaching assitantship in the Department of Computer Science (academic year 2), NIH grant GM076275 (summer, year 2), the Quantitative and Computational Biology Program NIH grant T32 HG003284 (years 3 , and 4), and NIH Center Grant P50 GM 071508, PI: D. Botstein (year 5) . Portions of this thesis were published in the following manuscript: Khan, Z., Bloom, J. S., Garcia, B. A., Singh, M., and Kruglyak, L. (2009). Protein quantification across hundreds of experimental conditions. *Proceedings of the National Academy of Sciences*, **106**(37), 15544–15548.

Contents

Chapter 1

Mass Spectrometry

Mass spectrometry is a flexible, quantitative, and data-intensive technology for identification and measurement of biomolecules ranging from small metabolites to proteins. Algorithms play a critical role in the analysis and interpretation of the data these instruments produce. Therefore, through applications to mass spectrometry, algorithms play a critical role in biological discovery. The size and complexity of the data these instruments produce presents a tremendous opportunity for new applications of techniques developed in computer science. This thesis explores one such application for the quantitative measurement of proteins.

This thesis focuses on proteins because they are work-horse molecules of the cell. They play an important role in virtually every biological process in an organism. By quantitatively measuring all proteins present in an experimentally perturbed sample relative to a control sample, a biologist can determine which specific proteins are involved in a biological process of interest. This type of analysis can be used to help dissect how the process works. Therefore, the methods developed in this thesis have wide application to the biological sciences.

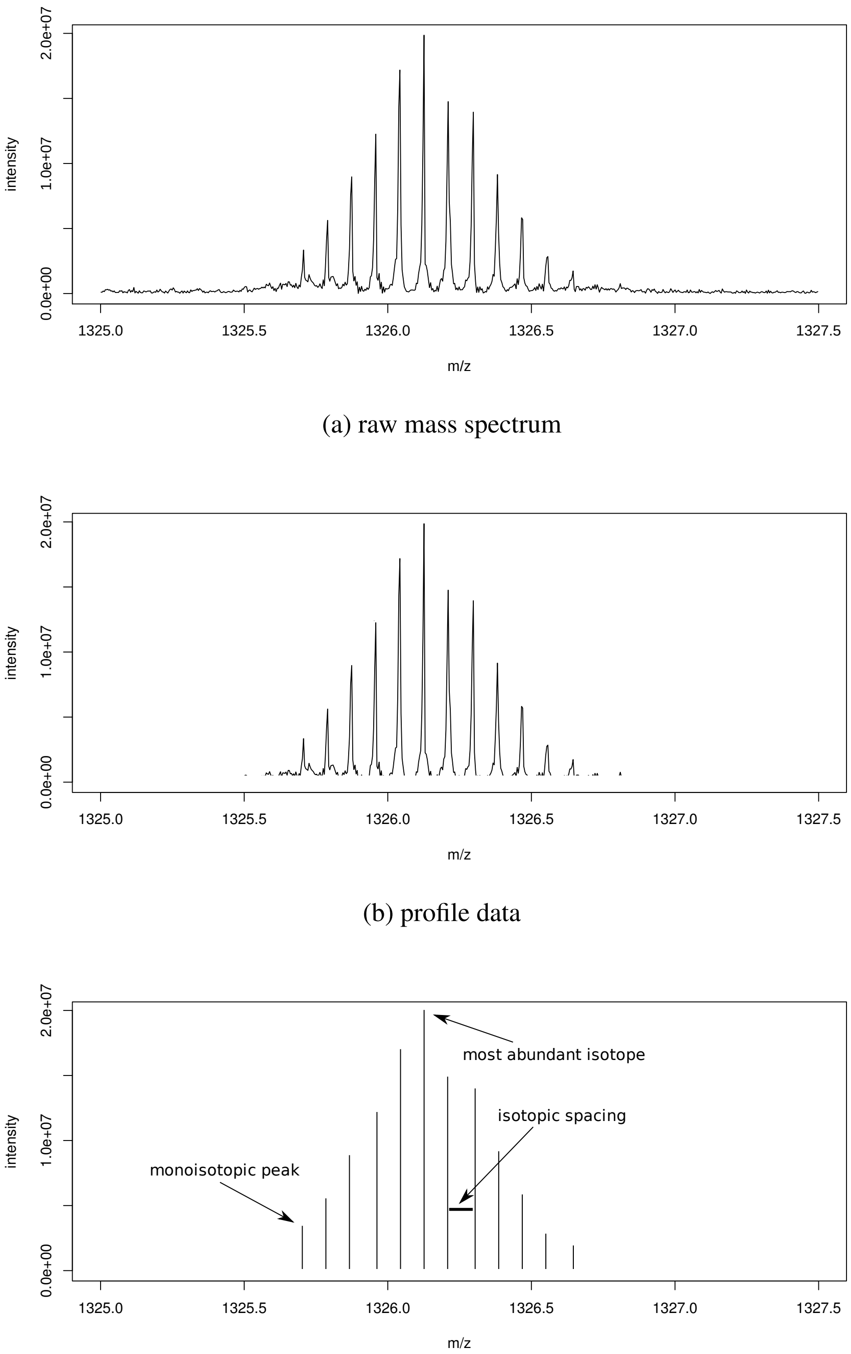
This chapter provides the technical background necessary for understanding the main contributions of this thesis. Subsequent chapters provide details of the main contributions and experimental validation.

1.1 Interpreting Mass Spectra

A mass spectrometer is a sensitive instrument for measuring the m/z or mass-to-charge ratio of ionized molecules. If one knows the charge of a molecule, one can determine the mass of a molecule. Furthermore, the intensity of the molecule's signal is correlated with the molecule's abundance.

The output of a mass spectrometer is a mass spectrum. A raw mass spectrum is a digital version of a continuous analog signal in the instrument (see Figure 1.1a). This signal can be stored as a vector of I intensity values sampled at very fine increments of m/z. The unit on this m/z axis is the Dalton. A raw mass spectrum is subject to a few post-processing steps. First, the mass spectrum is base lined to obtain what is typically called a mass spectrum collected in profile mode (see Figure 1.1b). Only values that are above a certain baseline are stored. Second, peaks are detected in this spectrum by an algorithm that finds convex regions of intensity values across the m/z range. This process of detecting peaks is called centroiding. A centroided mass spectrum consists of a list of two values, m/z, mass-to-charge and I intensity, where these convex regions, or peaks, have been located.

A mass spectrum provides the information necessary to accurately determine the mass of a molecule, and the mass might be sufficient to determine that the molecule exists in the sample analyzed by the instrument. The first step in determining the mass is to determine the charge of an ionized molecule. This is accomplished, for most molecules of biological interest, by examining the spacing between peaks caused by naturally occurring ^{13}C isotopes. In the example in Figure 1.1c, the spacing between isotopes is $\frac{1}{12} \approx 0.08333$ indicating the molecule measured has charge $z = +12$. The second step is to locate monoisotopic $(m/z)^*$. This is the left-most peak in Figure 1.1c.



(a) raw mass spectrum

(b) profile data

(c) centroid data

Figure 1.1: (a) A raw mass spectrum is a vector of intensity I values at finely sampled increments of mass-to-charge m/z. (b) A profile mass spectrum only stores m/z and intensity values above a baseline intensity. (c) A centroided mass spectrum will consist of a list of m/z and I values where peaks have been identified in a profile spectrum.

This peak reflects the signal from the entirely ^{12}C containing version of the molecule. Given the charge, one can compute the mass by multiplying by charge and subtracting out the weight of the protons added to ionized molecule

$$m = (m/z)^* \cdot z - 1.00727646677 \cdot z$$

where 1.007276446677amu is the mass of a proton. Note that the monoisotopic mass might be difficult to determine as its intensity diminishes as the mass of the molecule increases, but, in practice, most experiments are designed to avoid this problem.

An observed charge and mass-to-charge ratio may be sufficient to determine the identity of the molecule analyzed using a mass spectrometer. However, as a biological sample becomes more complex the chance that two molecules in the sample will have an indistinguishable m/z increases rapidly. Consequently, mass spectrometry is often coupled to liquid chromatography.

1.2 Liquid Chromatography Coupled Mass Spectrometry

Liquid chromatography (LC) refers to a class of techniques for separating complex mixtures where the sample being separated has been solublized in a liquid. The most common type of liquid chromatography is reverse phase chromatography. In this type of LC, the sample is dissolved in a polar mobile phase and forced, using high pressure pumps, through a column containing an immobile non-polar stationary phase. Modern LC systems employ high pressure pumps and instrumentation capable of varying the composition of the liquid mobile phase enabling separation of molecules in highly complex mixtures.

With the addition of LC, a mass spectrometer is configured to collect spectra as quickly as possible. Consequently, peaks in a single mass spectrum acquire an additional value. Each peak now not only has an m/z and an intensity I, but also a time t which is the time point at which the mass spectrum was collected (see Figure 1.2).

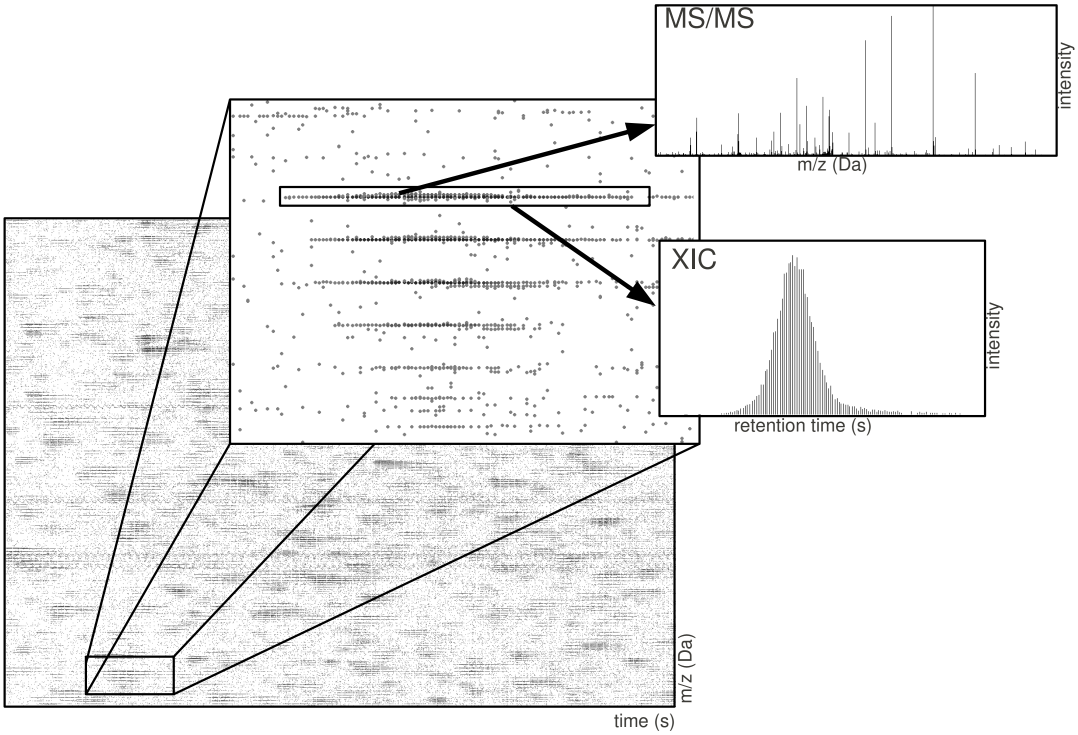


Figure 1.2: A representative LC-MS/MS data set. The data set consists of millions of peaks, each with a specific time, mass-to-charge ratio (m/z), and intensity. The intensity of a peak is indicated in gray scale (darker gray corresponds to higher intensity). The first inset shows several extracted ion chromatograms (XICs) corresponding to the naturally occurring isotopes of a molecule. The area under an XIC provides a measure of the relative abundance of a peptide. In this example, one peak in the XIC has an MS/MS fragmentation spectrum from which the identity or amino acid sequence of the peptide can be determined.

As molecules elute out, or leave, the LC column, they create characteristic patterns known as eXtracted ion chromatograms (XIC) (see Figure 1.2 inset XIC). These XICs can be labeled monoisotopic or isotopic just as peaks in a single mass spectrum can be labeled monoisotopic and isotopic. The spacing between the XICs can be used to determine the charge, and subsequently, the mass of the molecule. The position in time of these XICs reflects the length of time for which the molecule was retained by the stationary phase of the LC column. Further, area under an XIC is correlated with the abundance of the molecule in the sample.

The mass computed from a set of XICs may be insufficient to determine the identity of a molecule, especially if the space of possible candidates is large. To address this problem, molecules can be further fragmented by isolating ions in a small m/z window and fragmenting these ions using several methods. The resulting fragments are measured a second time in the instrument to obtain a fragmentation mass spectrum (see Figure 1.2 inset MS/MS). An instrument capable of this selection and fragmentation is called a tandem mass spectrometer. When such an instrument initiates an event where ions are isolated and fragmented, it records a precursor m/z value and a time t. In summary, an LC-MS/MS data set will consists of two types of items:

1. $(m/z,t)$ peaks (usually in the tens to hundreds of millions) each with an intensity I attribute

2. $(m/z,t)$ peaks with a fragmentation spectrum in the form of a list of m/z and intensity I pairs.

1.3 Thesis Summary

The technical contributions of this thesis are summarized by chapter below. In addition to these contributions, this thesis also contributes an open source software system called PVIEW (available for download at http://compbio.cs.princeton.edu/pview). Using the algorithms described in this thesis, PIVEW enables quantification of proteins, protein fragments, and post translational modifications with significantly greater depth and accuracy than previous approaches. PVIEW also demonstrates that the algorithms described in this thesis significantly improves the speed of data analysis in practice.

- **Chapter 2**

  This chapter describes my primary contribution to the field of LC-MS/MS data analysis: the introduction of space partitioning data structures to enable rapid extraction of XICs in very large LC-MS/MS data sets.

- **Chapter 3**

  XICs by themselves contain a limited amount of useful information. When coupled with protein identifications and peptide sequence information, the quantitative data in XICs can be linked to the biology of the sample under study. This chapter describes a collection of methods for database search of peptide fragment ions, false discovery rate estimation, grouping of peptides belonging to multiple proteins, and post-translational modification site localization that were implemented in a software system called PVIEW. PVIEW is the first software system that seamlessly integrates these algorithms. This software system is the second contribution of this thesis. More details on the source code of the system are provided in Appendix A.

- **Chapter** 4

  This chapter describes two major methodologies for quantifying peptides and proteins: label-free quantification and isotope labeled quantification. Leveraging the ideas developed in previous chapters, I introduce new algorithms for quantifying peptides and proteins using both of these methodologies. The algorithms developed here are the third contribution of this thesis. The algorithms developed in this chapter are evaluated on a wide range of previously published data sets.

- **Chapter 5**

  This chapter describes a new approach for analyzing ^{15}N labeled mass spectra.

This chapter introduces the fourth contribution of this thesis, a new method for isotope labeled quantification using mass spectra where shifts are dependent on the nitrogen composition of peptides. The approach is evaluated on a large data set comparing stationary and log-phase growing *E. coli*.

- **Chapter 6**

  In this last chapter, I summarize the main contributions of this thesis. From the perspective of data analysis, I outline some of the strengths and weaknesses of LC-MS/MS in the context of protein and peptide quantification and provide some guidance for its application and use in experimental work. I also outline some promising directions for future work.

Chapter 2

Extracted Ion Chromatograms

This chapter describes the key contribution of this thesis: a new application of space partitioning data structures. It shows how the orthogonal range queries supported by one particular space partitioning data structure, the kd-tree, can be used to develop a series of algorithms for processing LC-MS/MS data sets. The algorithms abstract away peaks leaving only XICs for further computational processing. The algorithms apply generally to the many types of samples and molecules that can be analyzed using a liquid chromatography coupled tandem mass spectrometer. These molecules include metabolites, proteins, lipids, and posttranslationally modified peptides. The subsequent steps for processing the output of these algorithms will differ based on the sample analyzed, the experiment performed, and the goals of the study.

2.1 Orthogonal Range Queries for Analyzing LC-MS/MS Data

Central to all of the algorithms in this chapter is a data structure D that indexes objects based on two of their dimensions: time (t) and mass-to-charge ratio (m/z). The data structure supports the following interface:

> **RangeQuery**(D, t_1, t_2, m_1, m_2) Use D to conduct an orthogonal range query, returning K objects (t, m/z), from a large set of N total objects, that are within a rectangular region defined by $t_1 < t < t_2$ and $m_1 < m/z < m_2$.

Orthogonal range queries have been the subject of intense study in computational geometry (de Berg *et al.* (2008)). Consequently, there are many data structures that support this interface, each with its own space, speed, simplicity, and efficiency trade offs. Examples include kd-trees, range trees, quad-trees, binary space partitioning trees, and all of their many variants (Samet (2005)).

Algorithm 2.1 Draw($w, h, D, t_1 t_2, m_1, m_2$) - Display $(t, m/z)$ within the range $t_1 < t < t_2$ and $m_1 < m/z < m_2$ on an image of width w and height h.

```
$Q$= RangeQuery($D, t_1 t_2, m_1, m_2$)
a = $w/(t_1 - t_2)$
b = $h/(m_2 - m_2)$
for each $(t, m/z)$ in $Q$, DrawPeak(a($t - t_1$), b($m/z - m_1$))
```

The immediate usefulness of an orthogonal range query in data visualization of LC-MS/MS data is demonstrated by Algorithm 2.1. Instead of iterating through all of the points to see if they appear in a region of given width and height on a computer display, a range query that uses a data structure to rapidly accelerate the process of finding points that should be displayed. While incredibly simple, this algorithm is important. It plays a key role in allowing an experimenter to assess the data quality.

2.1.1 kd-trees

Data structures that support range queries employ two main algorithmic techniques. Theoretically optimal data structures that achieve $O(\log N + K)$ query time use a combination of range searching and fractional cascading. Approaches used in practice such as kd-trees, use recursive space partitioning (Samet (2005)).

The kd-tree is the most frequently used of these data structures for orthogonal range queries because it occupies $O(N)$ linear space with respect to N, the number of data points ((Bentley, 1975)). The worst case bound $O(\sqrt{N}+K)$ is pessimistic as it assumes an unlikely distribution of the data points (de Berg *et al.* (2008)).

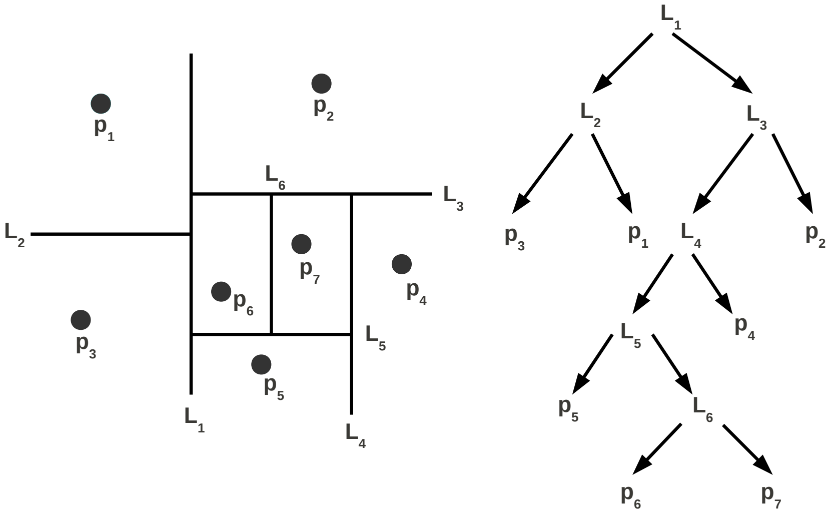


Figure 2.1: kd-tree built on points p_1 to p_7. Splitting planes are designated by lines L_1 to L_6. A kd-tree is constructed top-down by partitioning points based on the introduction of a splitting plane.

kd-trees leverage the algorithmic technique of space partitioning by recursively partition space using axis parallel splitting planes. The planes recursively divide the space along the x-axis (t time) and y-axis (m/z mass-to-charge ratio) into halves (see Figure 2.1). Orthogonal range queries are conducted by intersecting the half spaces defined by the splitting planes with the query to rapidly narrow down regions containing points that satisfy the query. Selecting the axes on which to split the data and position these splitting planes depends significantly on the types of queries conducted and the distribution of points (Maneewongvatana and Mount (1999))

Unfortunately, the kd-tree has a significant problem in practice, as the number of points N might be in the millions or hundred of millions. Consequently, the hidden constant factor in front of the N becomes problematic as data set size increases. The inplace kd-tree addresses this problem by carefully arranging the data points stored in a single array (Katz (2005); Brönnimann *et al.* (2004)). This data structure requires only $12N$ bytes if 4-byte single precision floating point values are used for peaks. The method has one notable downside. It is restricted to a single splitting plane strategy. The splitting plane is selected by alternating dimensions and choosing the median point for the split.

2.2 Removing Noise in LC-MS/MS Data

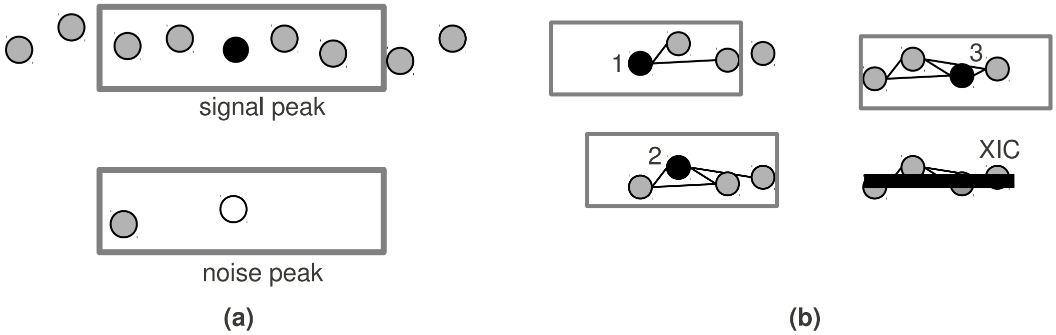


Figure 2.2: (a) A planar orthogonal range query determines whether or not a peak is labeled as signal or as noise. A peak is labeled signal if the query returns a threshold number of peaks. (b) After filtering, another set of planar orthogonal range queries are used to connect signal peaks in an undirected graph. XICs correspond to connected components in this undirected graph.

The algorithm described in this section performs the first step in processing an LC-MS/MS data set: removing peaks that are caused by noise (see Figure 2.2a). Given a kd-tree on all of the data points, the algorithm iterates through each peak and performs an orthogonal range query with specified range in time and range in m/z around the peak. If the number of peaks returned by the query exceeds a threshold and the peak is above a nominal absolute intensity threshold, the peak is labeled as signal. Otherwise, the peak is labeled as noise. After this process, the peaks labeled as noise are removed from further processing. This process can be described concisely in Algorithm 2.2.

Algorithm 2.2 RemoveNoise(P) - Given a set of peaks P, returns a set of filtered peaks F. The parameters Δt and Δm set the width and height of the orthogonal range query. R thresholds the number of peaks and M sets the minimum intensity.

```
for each $(t, m/z, I)$ in P
    Q = RangeQuery(D, $t - \Delta t$, $t + \Delta t$, $m/z - \Delta m$, $m/z + \Delta m$)
    if(size(Q) > $R$ and $I > M$) add p to F
return F
```

2.3 eXtracted Ion Chromatograms

This section details an algorithm that uses orthogonal range queries and an undirected graph structure to find XICs in a set of filtered peaks. The first step of this algorithm makes each peak a node in the graph. Then, the algorithm constructs a kd-tree on all of the filtered points. Next, the algorithm iterates though each signal peak and connects the current node to any signal peak nodes returned in a query with specified width in time dimension and height in m/z. Last, the algorithm finds XICs by computing the connected components of the constructed graph. An individual connected component corresponds to an XIC (see Figure 2.2b). Algorithm 2.3 describes this procedure concisely.

In practice, a graph on all of the points requires a substantial amount of memory. Rather than construct the graph explicitly, an algorithm that finds connected components by depth first search can construct this graph implicitly by finding and traverse

Algorithm 2.3 FindXICs(F) - Given a set of filtered peaks F, returns XICs in the form of connected components of the graph G. The parameters Δt and Δm set the width and height of the orthogonal range query.

```
for each $p = (m/z, t, I)$ in F
    Q = RangeQuery(D, $t - \Delta t$, $t + \Delta t$, $m/z - \Delta m$, $m/z + \Delta m$)
    For each returned point $q$ in Q
        Add an edge between $q$ and $p$ in a graph G.
return FindConnectedComponents(G)
```

edges by performing orthogonal range queries. Because this approach uses orthogonal range queries to traverse edges, it runs in $O(N + \sqrt{N}E)$ time where E is the number edges between peaks. Note that the size of the range queries assure that the number of edges is not very large (i.e. $E \ll N(N-1)/2$) so the algorithm runs quickly. Furthermore, chromatograms sometimes contain multiple peaks for the same time value. When this occurs, the most intense peak is kept at that time value.

Once an XIC is found, it can be augmented with two additional values.

1. m/z, the intensity weighted average of all of the m/z values of the peaks

2. t time of the most intense peak in the XIC

3. q the computed area under the chromatogram value, a value that summarizes the quantitative data in an XIC.

Because XICs now have an associated m/z and time t, they too can be indexed in a kd-tree and queried using orthogonal range queries. The next section leverages this observation.

2.4 XIC Type and Charge

The algorithm in this section augments XICs with two additional pieces of useful information.

1. A label that designates the XIC is of type monoisotopic, isotopic, or unknown.

2. An integer $(+1, +2, \ldots)$ that designates the charge of a monoisotopic XIC.

This process is detailed concisely in Algorithm 2.4. First, the algorithm examines, for a range of charge values, positions where an isotopic XIC might be present. These positions are computed using the mass shift that is observed by the presence of a single ^{13}C isotope in a molecule. If there are any XICs present in these positions the XIC is labeled as isotopic. If no XICs are found, then values greater than the m/z of the current XIC are examined in order of decreasing charge. If an XIC is found, then the XIC is labeled as monoisotopic and the current charge value z is assigned to the XIC.

Algorithm 2.4 TypeAndCharge($m/z, t_1, t_2$**)** Determines the type (monoisotopic, isotopic, or known) of an XIC with a given m/z, a minimum peak time value t_1 and a maximum peak time value t_2.

```
for charge $z$ = maximum charge down to $+1$
    $m = m/z - 1.003355/z$
    Q = RangeQuery($D, t_1, t_2, m - \varepsilon, m + \varepsilon$)
    if Q not empty then type is isotopic return
// Examine $^{13}$C shifts at higher $m/z$
for charge $z$ = maximum charge down to $+1$
    $m = m/z + 1.003355/z$
    Q = RangeQuery($D, t_1, t_2, m - \varepsilon, m + \varepsilon$)
    if Q not empty then return type is monoisotopic and charge is $z$
return type is unknown
```

The algorithm presented here makes several assumptions about sample analyzed. The sample must contain a sufficient fraction of ^{13}C isotopes to allow accurate detection of monoistopic and isotopic XICs. Furthermore, monoisotopic XIC must correspond to the XIC detected at the smallest m/z value. As the mass of the molecules analyzed increases or the compounds contain an unusual atomic isotopes, variations on the approach presented here must compute exactly or approximate the isotope distribution of a compound (Rockwood *et al.* (1995)).

2.5 Summary of Previous Work

Existing approaches for finding XICs in LC-MS/MS data sets fall into two categories: image processing based approaches and single spectrum processing. Because image processing typically requires computation on individual pixels, these methods require significant computation time and are subject to loss of precision (Jaffe *et al.* (2006); Kohlbacher *et al.* (2007); Bellew *et al.* (2006); Palagi *et al.* (2005)). To address the limitations of image-based methods, more recent approaches have relied on algorithms that process an individual spectrum at one time point at a time and combine those results across time points in a second step (Park *et al.* (2008); Finney *et al.* (2008); Mueller *et al.* (2007); Cox and Mann (2008); Noy and Fasulo (2007)). The algorithms presented in this chapter skip this individual spectrum processing step, processing the data entirely in two dimensions: time and m/z.

Chapter 3

Protein Mass Spectrometry

This chapter describes methods that address the problem of assigning an amino acid sequence of a peptide to an XIC. The methods were implemented in an open source software system called PVIEW (http://compbio.cs.princeton.edu/pview). It is the first software system, as of date, that seamlessly integrates algorithms for quantification of extracted ion chromatographic peaks, database searching of MS/MS spectra, statistical validation, protein grouping, and posttranslational modification (PTM) site localization with positional scoring. Each of these algorithms is described briefly in this chapter. The algorithms implemented in PVIEW assume that experimenters apply to following widely used experimental methodology used for protein mass spectrometry:

1. Proteins are extracted from a tissue or cell culture.

2. The proteins are enzymatically digested into peptide fragments and possibly enriched for posttranslational modifications.

3. A liquid chromatography coupled tandem mass spectrometer is used to analyze the digested sample to generate an LC-MS/MS data set.

Additional details on the source code of PVIEW are provided in Appendix A.

3.1 Filtering Fragmentation Spectra

Prior searching a database of predicted fragment masses and spectra, PVIEW filters fragmentation spectra using two filtering algorithms:

1. **Isotope filtering:** Initially all peaks are labeled as signal. Peaks are scanned by decreasing intensity. If the peak is labeled as noise, it is skipped. Peaks that are within a $[-\varepsilon, 2+\varepsilon]$ window in the m/z of the current peak and of lower intensity than the current peak are labeled as noise. Only peaks labeled as signal are kept for the next phase of spectrum filtering.

2. **Peak filtering:** Initially all peaks are labeled as noise. Peaks are scanned by decreasing intensity. If there are no more than K peaks of higher intensity in a $[-W, +W]$ window around the current peak, then the peak is marked as signal. Peaks are scanned until there are F peaks (the default parameters are $K = 8$, $F = 100$, and $W = 100$).

Filtering has been shown to significantly improve the quality of search results obtained by several different algorithms for database search (Renard *et al.* (2009))

3.2 Peptide Database Search

The algorithms implemented in PVIEW use the precursor mass, derived from the XIC's m/z and charge, and a fragmentation spectrum to search a database of predicted fragment masses based on the genome sequence of the species under study and the enzyme used for digestion (Geer *et al.* (2004); Craig and Beavis (2003); Eng *et al.* (1994)). Using a statistically significant match to the database, the algorithms assign an XIC an amino acid sequence along with additional information about the identity of proteins and protein isoforms from which the sequence originated. Through enumeration, the methods localize the amino position of a posttranslational modification, and assign this position a score and significance value.

Database search relies on the predictable fragmentation pattern of peptides in tandem mass spectrometer. Collision induced dissociation (CID) is the most prevalent technology for fragmenting peptides. As illustrated in Figure 3.1, CID produces two types of fragment ions: b-ions and y-ions. The mass of these fragment ions correspond to the masses of subsequences from the underlying peptide. Using a theoretical fragmentation model and a database of peptide sequences, a database search algorithm matches these observed fragment ions with predicted fragment ions. A high scoring match allows a peptide sequence to be assigned to a particular fragmentation spectrum.

PVIEW uses the fragmentation model from the Open Mass Spectrometry Search Algorithm (Geer *et al.* (2004)), and the m/z and charge of the XIC associated with a fragmentation spectrum to compute the peptide's mass. It scores peptide spectrum matches according to

$$S = \frac{M}{L} \sum_{m=1}^{M} \log I_m$$

where I_m is the intensity of the mth matched peak, M is the number of matched peaks, and L is the number of peaks in the theoretical spectrum.

3.3 Combinatorics of PTM Search

Since the addition and removal of posttranslational modifications (PTMs) is a dynamic process in the cell, algorithms that localize and identify these modifications while matching theoretical and experimental fragmentation spectra rely on enumeration. PTMs transform database search into a combinatorial search problem. In two stages of database search, PTMs expand the search space:

1. **Precursor Mass Query**: Using the intact, or precursor mass, of peptides ob-

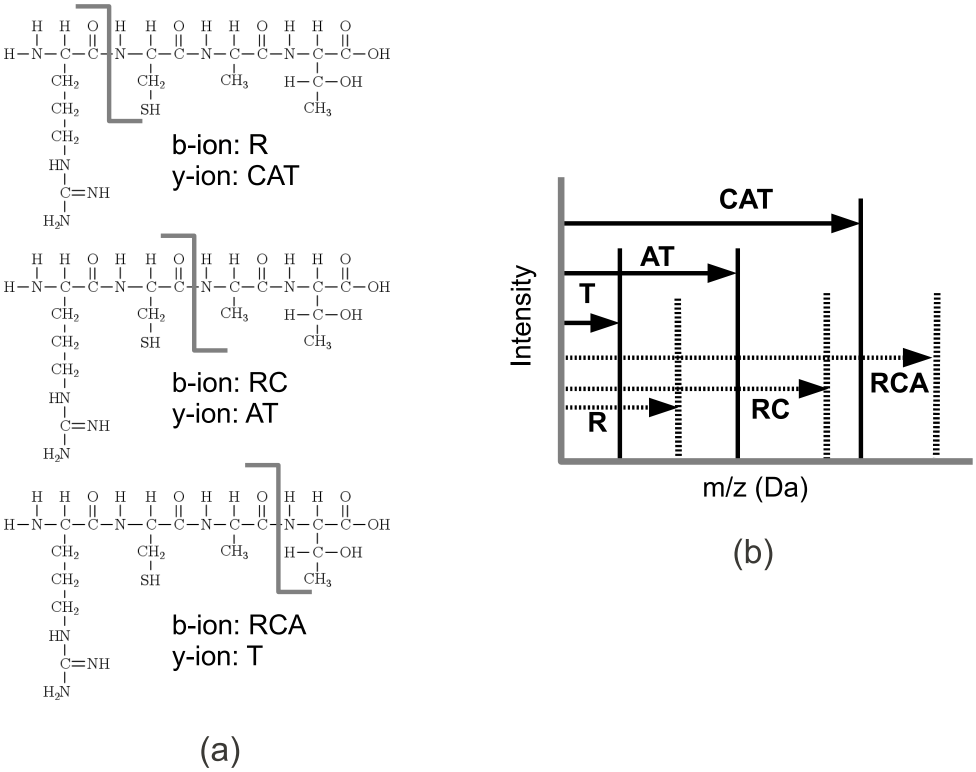


Figure 3.1: (a) Collision induced disassociation (CID) is a mass spectrometry technology that breaks peptides at their peptide bonds to produce fragment b-ions and y-ions. (b) These b and y ions produce peaks in a fragmentation spectrum corresponding to the mass-to-charge ratios of these fragments.

tained from XICs, database search narrows a set of candidate unmodified peptides by subtracting out the mass of the modification. If we include M types of modifications where K is the maximum number of allowed modifications, the ways of selecting K objects from a menu of M types of objects with repetition

$$\begin{pmatrix} M+K-1 \\ K \end{pmatrix}$$

is the number peptide queries that must be conducted. For small values of M and K, brute-force enumeration is reasonably efficient, but the search space grows quite rapidly as these values increase.

2. **Site Enumeration**: Once unmodified peptides are found they can be analyzed for amino acids that can be modified. If there are A modifiable peptides then the search space is $O(M^A)$ where M is the number of modifications.

A number of strategies can be employed to limit the search space for PTMs. The simplest approaches examine the sequences returned by the precursor mass query for modifiable amino acids. The most effective approach is to rely on amino acid sequence tags to find a series of fragment ions that correspond to known amino acid sequences (Mann and Wilm (1994); Bandeira *et al.* (2007)).

3.4 FDR and q-values

PVIEW implements the methods of Käll *et al.* (2008) to assign statistical significance to search results with three modifications: (1) PVIEW uses a concatenated reverse decoy database (Elias and Gygi (2007)). In order to avoid underestimation of the false discovery rate (FDR), PVIEW corrects for the imbalance between target and decoy scores that results from the use of a concatenated database. It does so by multiplying the ratio of accepted decoy scores over accepted target scores by the ratio of the total target scores over the total number of decoy scores. (2) PVIEW uses the modification to the concatenated reverse decoy database introduced by Cox and Mann (2008). Prior to reversing the amino acid sequence of a protein during reverse decoy database construction, PVIEW swaps the amino acid before the cleavage site with its predecessor to the decoy spectrums dependance on the enzyme used. (3) Because charge +2 and charge > +2 spectra use different theoretical fragmentation models, the FDR is computed separately for these two groups of fragmentation spectra.

3.5 Algorithm For Protein Grouping

For any one statistically significant peptide to spectrum database search result, the peptide assigned to the spectrum will originate from one or more proteins (or protein isoforms) in an amino acid sequence database. Based on this information, protein grouping attempts to discover the minimum set of proteins supported by the statistically significant search results (Nesvizhskii and Aebersold (2005)). PVIEW implements a novel approach to protein grouping. The algorithm is detailed below:

1. Create a node in an undirected graph for each entry (protein or protein isoform) in the sequence database.

2. For each protein node i in the graph, find the peptide P_i supported by a peptide to spectrum assignment that originates from fewest number of proteins. Record this number A_i on the graph node.

3. Scan through each node by increasing A_i. For each peptide P_i examine the nodes j that correspond to proteins from which the peptide originates. Add an edge between nodes $i \neq j$ if $A_j \leq A_i$, node j does not have a peptide originates from fewer proteins.

4. Compute connected components in the resulting graph. Each protein is assigned a component number.

5. For each connected component, examine the peptides assigned to each node in that component. Look for a peptide that originates from proteins in that component only. If such a peptide exists, the protein group is labeled as conclusive and reported by PVIEW, otherwise it is marked a non-conclusive and not reported.

From the algorithm above, the connected component numbers correspond to unique identifiers for protein groups. For any peptide supported by a statistically significant database search result, the component numbers of the proteins the peptide originates from are used to determine the protein groups to which the peptide belongs. If a peptide belongs to two or more protein groups, it is labeled as a shared peptide. These shared peptides are referred to as “razor” peptides in previous work (Nesvizhskii and Aebersold (2005); Cox and Mann (2008)).

3.6 PTM Site Localization and Scoring

PVIEW localizes PTM sites by creating an theoretical spectrum consisting of only peaks from b-ion and y-ions that provide evidence for the PTM at a specific amino acid position. These PTM specific theoretical peaks are matched and scored using the same score S as a peptide spectrum match. PVIEW’s database search with a concatenated reverse decoy database search yields two populations of PTM site scores: decoy positional scores and target positional scores. PVIEW conducts statistical validation of these scores using the methods from Käll *et al.* (2008), with the same modifications made for statistical validation of peptide database search results.

3.7 Summary of Previous Work

Software for analyzing high-resolution proteomic data obtained by liquid chromatography coupled to mass spectrometry (LC-MS) has historically consisted of separate tools for visualization, quantification, database search, and statistical validation. Recently, software pipelines for aggregating these tools have been developed (Keller *et al.* (2005); Kohlbacher *et al.* (2007); Cox and Mann (2008); Park *et al.* (2008)). While bringing these separate components together in a single release has been highly useful, such aggregation has not been without its drawbacks. Software installation becomes complex, data processing bottlenecks arise from unanticipated interactions between components, and the number of configurable parameters becomes daunting for users. In contrast, our system uses algorithms that rely a small number of parameters.

All methods for peptide identification rely on the mostly predictable fragmentation of peptides at peptide bonds. The earliest approaches attempted to perform automatic interpretation of mass spectra (Bartels (1990)) and were followed by approaches that use the database of manually curated mass spectra (Stein and Scott (1994)). The mid-90s also saw the introduction of the dominant strategy for peptide identification that is still used today. This strategy relies on the availability of a database of peptides predicted from an organism's DNA sequence (Eng *et al.* (1994); Mann and Wilm (1994)). Database search, especially in the presence of PTMs remains an active topic of research, but most algorithms continue to rely on variations on the same theme of database search and theoretical models of how peptides fragment.

Chapter 4

Protein and Peptide Quantification

This chapter describes two methodologies for quantifying peptides and proteins once XICs have been assigned a peptide sequence and protein ID using the algorithms in Chapter 3. The methodologies are distinguished by whether or not a label composed of atomic isotopes of are incorporated into peptides. Incorporation of isotopes render the peptides, in one of 2 or more different samples, heavier in mass allowing the signal from two samples to be separated in mass spectra. In contrast, label-free methodologies do not rely on such label incorporation.

4.1 Label-Free Quantification

Given a set of XICs and their assigned peptide sequence and protein identification, we now have enough information to perform quantitative analysis of an LC-MS/MS data set. This is accomplished by using the charge and amino acid sequence assigned to an XIC to cross reference the same XICs across samples or experimental conditions. If both the charge and the amino acid sequence are the same, the area under the XICs can be compared across data sets and experiments. The basic idea behind XIC-based quantification is illustrated in Figure 4.1. This approach is called label-free to distinguish it from a quantitative methodology involving stable isotope labels. Isotope labeled quantification is described in Section 4.2.

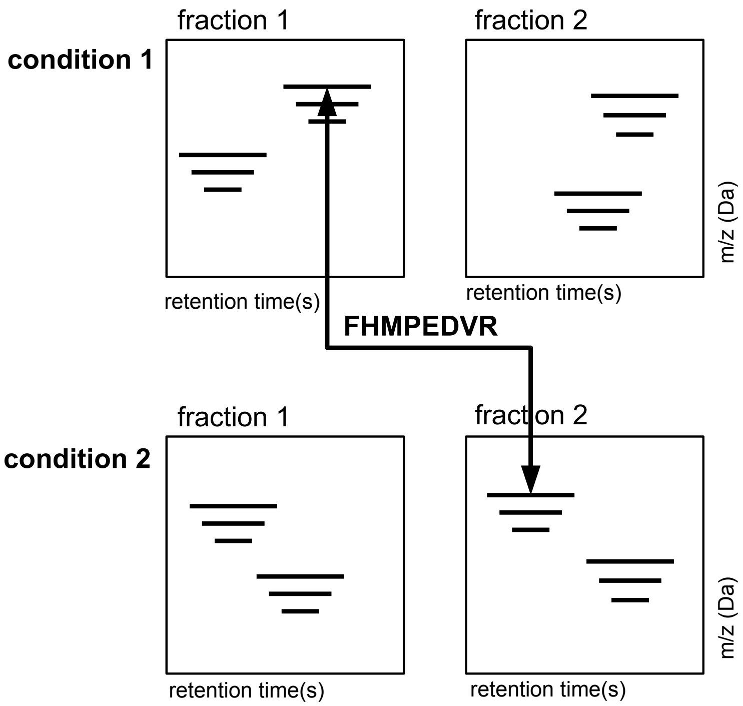


Figure 4.1: Here, two experimental conditions separated into two fractions are compared by XIC-based quantification. In this label-free quantification methodology, the peptide sequence FHMPEDVR is used to cross reference monoisotopic XICs, shown by straight lines, across the two experimental conditions and across two different fractions. The areas under these XICs provide two quantitative measurements of the same peptide in two different experimental conditions.

4.1.1 Time Alignment

Due to noise and the limited sampling speed of an LC-MS/MS instrument, XICs may not be assigned a fragmentation spectrum or assigned a peptide sequence and protein identification. This makes it impossible to cross reference the XICs based on peptide sequence and charge. One approach to overcome this problem is to align and group XICs along the time dimension. Non-linear variation in time of these XICs due to slight differences in chromatography, along with measurement error in m/z, complicates this process. A solution requires multi-run alignment step in order to correct for these sources of variation. As the number of runs increases, assuring correctness of the global alignment becomes significantly more difficult Foss *et al.* (2007).

Before grouping the XICs in multiple LC-MS/MS instrument runs, each run must be aligned to a reference run via simple translation. This alignment compensates for any differences in when the samples began to elute out of the LC column. Our approach iterates through each XIC in the current run and finds the nearest XIC in the reference run that is within a rectangular window with specified width in time and height in m/z. For this nearest XIC in the reference run, our approach also computes the reciprocal nearest XIC in current run. If this reciprocal XIC is the same as the current XIC, the difference in time between the current XIC and the nearest reference XIC is stored in a list. After all of the XICs in the current run are processed, the median of these differences in time is used to translate each XIC in the current run to the reference run.

The translational alignment algorithm is shown more precisely in Algorithm 4.2. The algorithm translates all runs to the reference run to adjust for differences in when data collection was started in both data sets. It then labels XICs with identifiers that indicate which data set they belong to and combines all of the XICs into one merged data set.

Algorithm 4.1 AlignAndGroup(S) - Here in a set of instrument runs *S*, each run C are aligned to a reference run R. Then XICs are grouped across these runs.

```
for each pair (R, C) in S where R is a reference run
   drt = GetTranslation(C, R)
   Translate each XIC in C by drt in the time dimension
for each Si in S
   Mark all of the XICs in Si as originating from run i
   Add these labeled XICs to the set Z.
return GroupXICs(Z)
```

Aligning and grouping requires that GetTranslation() in Algorithm 4.2 use reciprocal nearest neighbor queries to determine corresponding XICs. The median difference in time between these XICs is used to align a run to translation to a reference run R.

Algorithm 4.2 GetTranslation(C, R) – Computes the median translational difference between run C and run R by finding the nearest corresponding XIC in time between these runs.

```
for each XIC x in C
   Q = RangeQuery(R, x.mz - dmzA, x.mz + dmzA, x.rt - drtA, x.rt + drtA)
   find nearest XIC b in Q
   Q = RangeQuery(C, b.mz - dmzA, b.mz + dmzA, b.rt - drtA, b.rt + drtA)
   find nearest r in Q
   if x = r save drt = x.rt – b.rt
return median of drt values
```

Once each run is aligned to the reference run via translation, XICs are grouped across these runs. First, each XIC belonging to a run is labeled using the LC-MS/MS run's identifier. Then, each XIC from that run is combined into a larger set of labeled XICs from all runs. Last, we use the same technique applied to find XICs from peaks to group XICs across runs. Each XIC starts out as a node in a graph. Planar orthogonal range queries between the start and end of an XIC in time and XIC width in m/z are used to connect XICs in a graph. Connected components in this graph correspond to grouped XICs (see Figure 4.2). The planar orthogonal range queries used in this step automatically compensates for any nonlinear differences in the positions of XICs. Further, the range queries can be expanded far beyond the start and end of the XICs to compensate for more severe differences in XIC time.

The algorithm for grouping XICs is shown more precisely in Algorithm 4.3. Planar orthogonal range queries construct a graph connecting XICs in a merged data set Z. As illustrated in Figure 4.2, the range queries use the start and end time of the XIC and a user specified parameter width in the m/z dimension of an XIC. Connected components in this graph correspond to XICs that have been grouped across runs. Instead of using nonlinear parametric alignment, the range queries automatically account for variance the position of XICs.

Algorithm 4.3 GroupXICs(Z) - Groups XICs in Z across time aligned data sets.

```
T = BuildIndex(Z)
for each x in Z
    Q = RangeQuery(T, x.start, x.end, p.mz – dwidth, p.mz + dwidth)
    For each returned point q in Q
        Add an edge between q and p in a graph G.
return FindConnectedComponents(G)
```

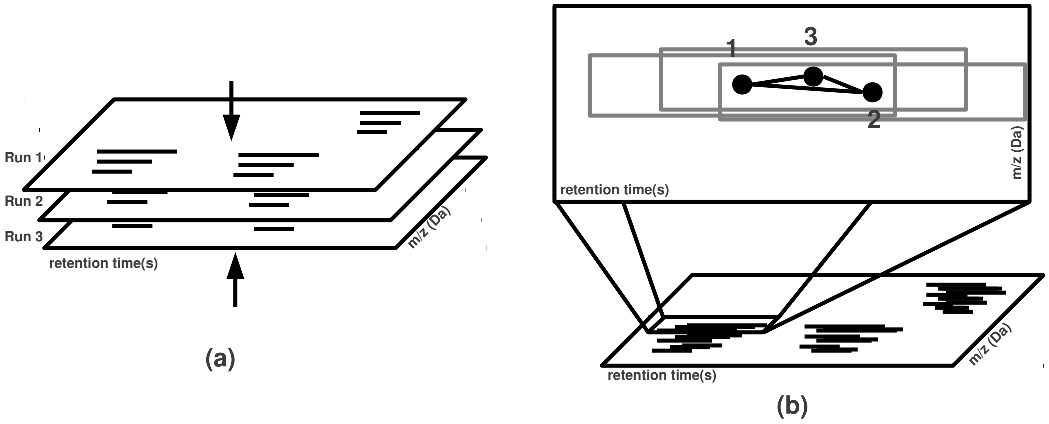


Figure 4.2: (a) XICs are labeled according to the run from which they originate and combined into a single run. (b) Planar orthogonal range queries are used to connect labeled XIC centers in an undirected graph. The range queries compensate for any non-linear differences in time between LC-MS runs. Connected components in the graph correspond to grouped XICs of the same peptide.

Once these XICs are grouped across runs, relative abundance values are computed and an amino acid sequence and protein identity is assigned to the group. Relative abundance values across runs are computed as the areas under each XIC in the group. These relative protein abundance values are normalized using median of medians normalization to adjust for differences in overall run intensity Callister *et al.* (2006). Then, the entire XIC group is assigned an amino acid sequence for a particular tryptic peptide by database search. Database search is conducted using all or a subset of fragmentation spectra within an XIC group.

Once the XIC group is assigned an amino acid sequence and protein identity, relative abundance data is available for a known tryptic peptide from a known protein. Because individual proteins are digested into several tryptic fragments, there may be many different XIC groups for a single protein, but in most applications, only the relative abundance of a protein is required. Because each tryptic fragment from a protein ionizes with varying efficiency, the quantitative values in each XIC group cannot be combined by simple averaging. Instead, a representative XIC group is selected for each protein. Specifically, we select the XIC group with highest signal to noise ratio. We note that other criteria, tailored to specific applications, can be used to select a representative XIC group.

4.1.2 Validation

In order to validate the approach for label-free quantification by time alignment presented in this chapter, we used three data sets. The first data set is a spike-in data set used in a previous study by Mueller *et al.* (2007). The data set contains six non-human proteins added at six different known concentrations to a background sample of bulk human serum. Three replicates of each dilution were collected using an FT-LTQ Thermo Electron mass spectrometer and an Agilent 1100 chromatographic separation system. In total, this data set consists of 18 LC-MS/MS instrument runs and is 15 Gigabytes (GB) in size. The second data set, described by Foss et al., measured total unfractionated cellular proteins from in 107 genotyped segregants from a cross between two parental strains of yeast (BY4716 and RM11-1a) Foss *et al.* (2007). Four replicate LC-MS/MS instrument runs were carried out for each segregant. The data also includes 10 replicates each of parent strain. It also includes 2 replicates of each of 6 gas phase fractions from each parent strain. In total, this data set includes 472 LC-MS/MS instrument runs and is 42 GB in size. The data set was generated using a Thermo Electron Corp. LTQ-FT mass spectrometer and a Michrom Bioresources Paradigm MS4B MDLC Nanoflow liquid chromatography system. We obtained the data from the authors. The third data set was from Jaffe *et al.* (2006). Here only proteins were added to a sample at two known concentrations. The sample was run 25 times for each concentration.

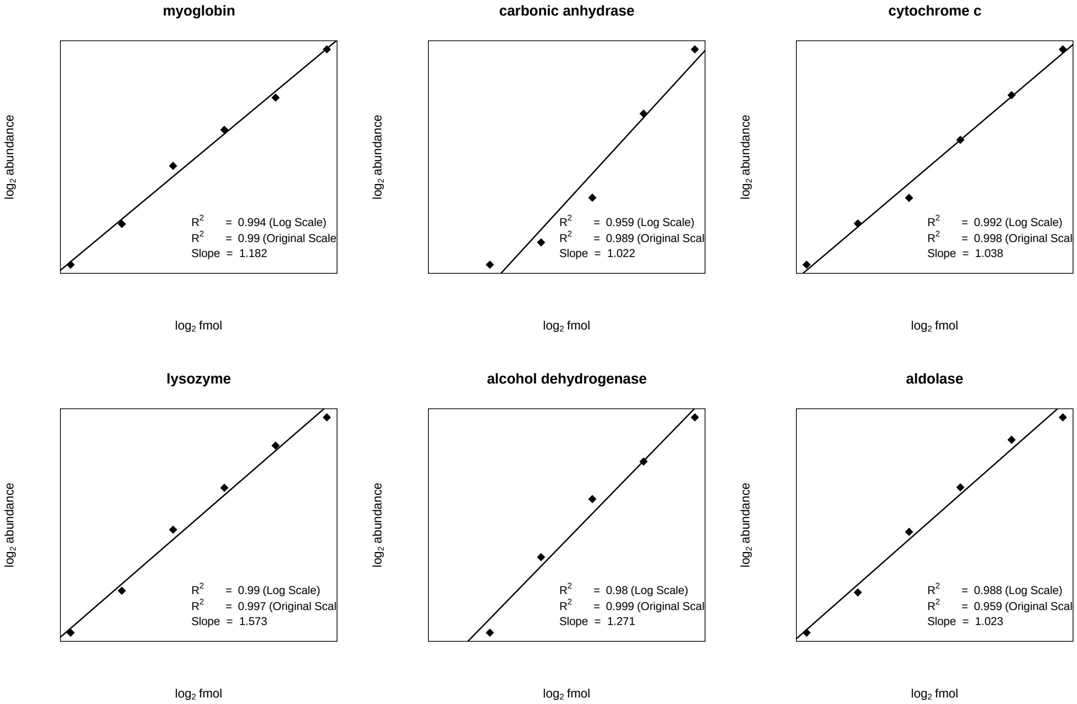
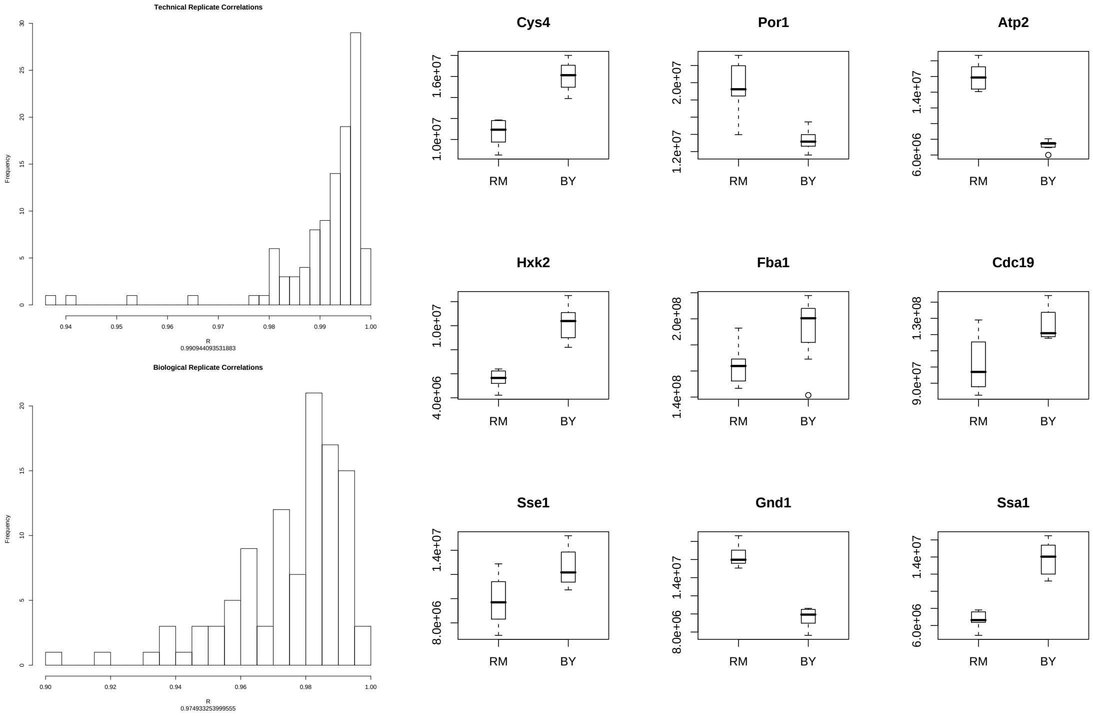


Figure 4.3: Comparison of known protein concentration with relative abundance measured by mass spectrometry. Measured protein abundance is plotted on a log–log scale against known femtomole concentration for each of six nonhuman proteins spiked in to human serum by Mueller *et al.* (2007) The lines show best fit by regression, and the corresponding log–log scale slopes and correlation values (R^2) for both the log–log scale and the original scale are shown below each plot.

Analysis of the quantitative output of the label-free time alignment algorithm on each of these three data sets confirmed that the methods produced accurate quantitative measurements. The results of the spike-in data set are shown in Figure 4.3. The quantitative output of our approach tracks the dilutions accurately across the replicates and varying spike-in concentrations. For the large data set, we examined the distribution of Pearson correlations for technical and biological replicates (see Figure 4.4a). The large mean of this distribution indicated the data sets were correctly aligned as the quantification values were highly reproducible. Furthermore, comparisons to quantitative Westerns performed for the BY4716 and RM11-1a samples confirmed that the measured quantitative differences were accurate (see Figure 4.4b). Last, for the data set from Jaffe *et al.* (2006), the ratios expected based on the known concentrations across the two sets were accurately determined by the algorithms (see Figure 4.5).



(a)

(b)

Figure 4.4: (a) distribution of Pearson correlations between protein abundance values measured between technical and biological replicates in the large Foss et al data set. (b) Boxplots show the distribution of proteins assayed by Westerns in the original Foss et al study. GND1, which was present in the original study, was eliminated because we found that the measured tryptic fragment contained a polymorphism.

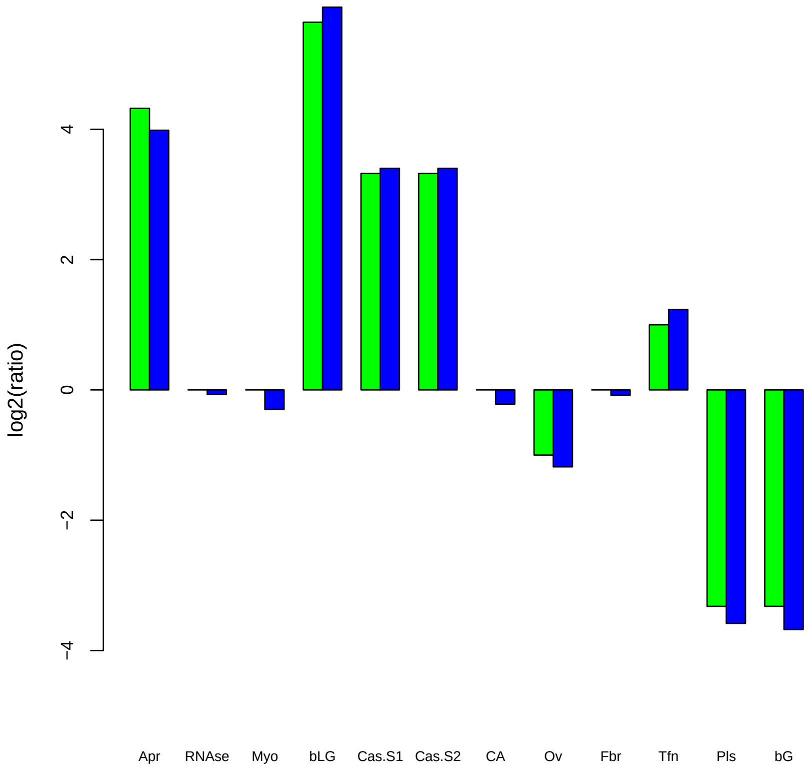


Figure 4.5: The ratios of proteins at two known concentrations of 13 different known proteins designed by abbreviations on the x-axis in the plot. The ratios were determined from two sets, for the two concentrations, of 25 technical replicates where accurately determined by the algorithms presented in this chapter. The green bars designate the observed $\log_2$ ratios and the blue bars designate the actual $\log_2$ ratios based on the concentrations.

4.1.3 Summary of Previous Work

Existing methods do not adequately address the computational challenges of XIC detection and multi-run alignment and are limited to a small number of runs. Several of these methods rely on image processing to collect quantitative data from LC-MS/MS instrument runs Jaffe *et al.* (2006); Kohlbacher *et al.* (2007); Bellew *et al.* (2006); Palagi *et al.* (2005). Because image processing typically requires computation on all individual pixels, these methods require a large amount of computation per LC-MS/MS run. Further, they require expensive image correlation computations to non-linearly align LC-MS/MS instrument runs Listgarten and Emili (2005). Our algorithmic techniques avoid pixel-level image representations.

To address the limitations of image-based methods, more recent approaches have relied on complex algorithms that process individual spectra one time point at a time and combine those results Park *et al.* (2008); Finney *et al.* (2008); Mueller *et al.* (2007). We skip this individual run processing step, representing the data in two dimensions: time and m/z. Our algorithm is much simpler because it exploits structure in the data across both dimensions. Furthermore, existing methods rely on iterative optimization techniques to align runs. These iterative techniques become increasingly less reliable and computationally more expensive as problem dimensionality increases with the introduction of additional runs Foss *et al.* (2007). In this work, we present a technique that is not susceptible to these dimensionality problems and demonstrably scales to hundreds of LC-MS/MS instrument runs.

4.2 Isotope Labeled Quantification

This section describes the algorithms presented in this thesis can be used for a second quantification methodology called isotope labeled quantification. In this experimental approach, one of two samples is labeled with a heavier isotope tag Ong *et al.* (2002). Because both the unlabeled and isotope labeled tryptic fragments are subject to the same chromatographic conditions they elute out of a liquid chromatography column at approximately the same time. Isotope labeled quantification reduces to finding pairs of XICs within a single run at the same time that are spaced according to the charge of the and the weight difference of the heavier isotope tag. The ratio of the areas of these XICs measures the relative abundance of the corresponding fragment between the experimental samples.

Instead of aligning and grouping XICs across runs, isotope labeled data is handled by grouping XICs within a run. This is accomplished by reciprocal planar orthogonal range queries between XICs. XICs are processed in increasing m/z order, starting with XICs corresponding to the lighter variant of the peptide. XICs that have already been paired are skipped. Based on the charge of the XIC, the expected isotopic spacing of the heavier variant is computed. An orthogonal range query between the start and end of the current XIC and m/z window is queried for a putative paired XIC. The longest of the returned XICs is used to conduct a reciprocal planar orthogonal range query between the start and end of the putative paired XIC. If the current XIC is returned by this reciprocal query, the two XICs are paired.

Algorithm 4.4 IsotopePairs(X) – Finds isotope pairs in a given XIC set X

```
T = BuildIndex(X)
Sort-by-increasing-m/z(X)
for each x in X
   if x has been paired then skip
   Q = RangeQuery(T, x.start, x.end, x.mz + labelshift - tol , x.mz + labelshift - tol)
   find largest XIC x2 in Q
    Q2 = RangeQuery(T, x2.start, x2.end, x2.mz - labelshift - tol , x2.mz + labelshift – tol)
   find largest XIC r in Q2
   if r = x then save isotope pair (x, x2)
```

This approach is described more precisely in Algorithm 4.4 in which XICs are grouped within a single run to find light and heavy isotope pairs. The algorithm iterates through XICs by increasing value in the m/z dimension, pairing light XICs with XICs caused by the heavier species first. By building a spatial index on the XICs, the algorithm assures the same XICs are returned by reciprocal planar orthogonal range queries using the given label shift in the m/z dimension.

Fragmentation spectra assigned to the paired XICs are used to determine the amino acid sequence and protein identity associated with the pair. Because proteins are digested into several tryptic fragments, a data set may contain several paired XICs per protein. Unlike label-free data, the abundance ratios computed by these pairs are comparable. The median ratio from all of these pairs can be used to compute a robust estimate of the abundance ratio for an individual protein.

4.2.1 Validation

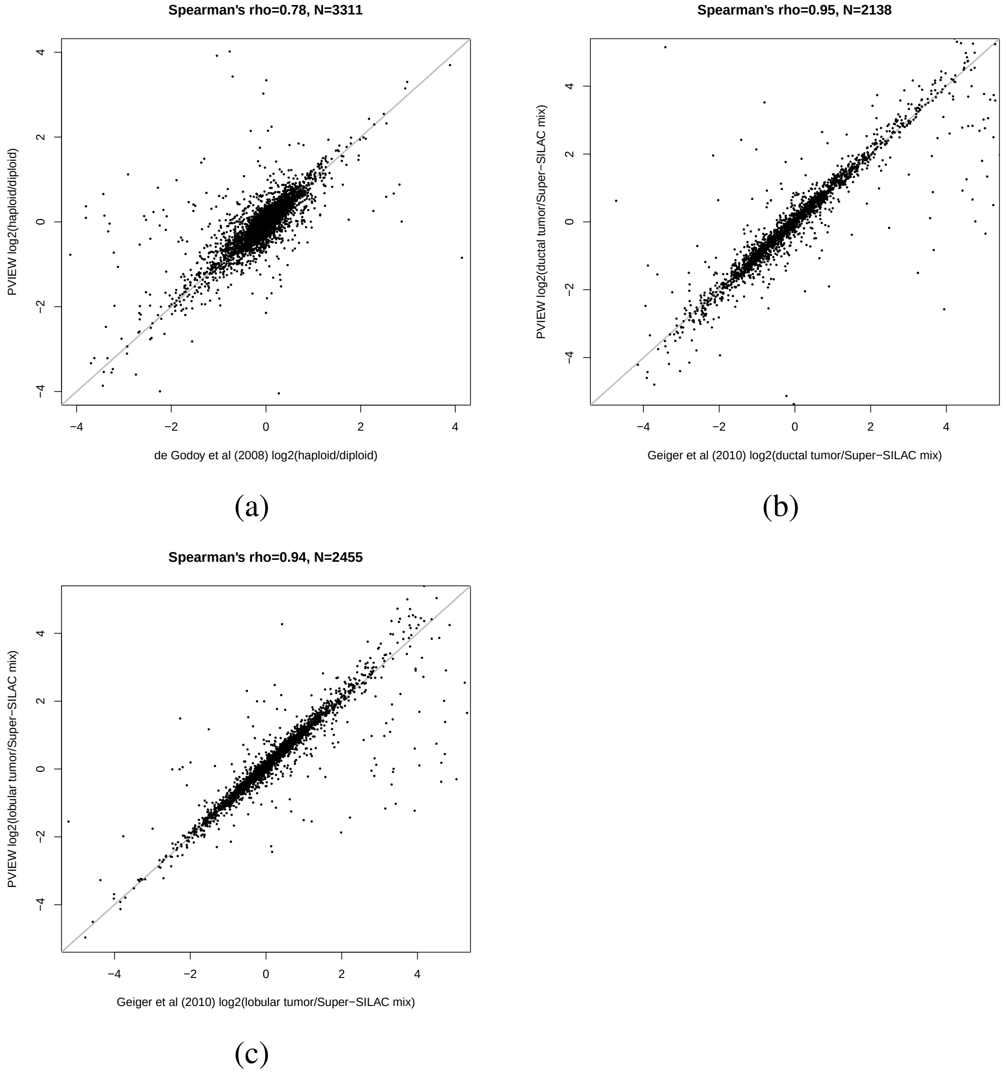
The method presented in this chapter was validated using two publicly available data sets that quantified yeast and human samples (Table 4.1). These data were selected because they were generated under experimental conditions in which the expression levels of a large number of proteins varied over a wide range. We observed high correlations between output reported by our method and an independent analysis of the data for a yeast sample (Spearman's correlation 0.78, Fig. 4.6a) and two human samples (Spearman's correlation 0.95 and 0.94, Fig. 4.6b and Fig. 4.6c) using a software pipeline known as Maxquant Cox and Mann (2008). We found that our method quantified more protein groups at a false discovery rate of 1% for both yeast (3,966 our method vs. 3,632 Maxquant) and human samples (4,687 our method vs. 4,336 Maxquant). Furthermore, correlation of technical replicates of one of the human samples suggested that our method produced more accurate quantitative output for large expression differences

(Fig. 4.7).

We also validated our method on data sets enriched for PTMs (Table 4.1). Since PTM site occupancies can vary independently of protein abundances, there may be little correspondence in the quantification of different modified peptides from the same protein. Therefore, we tested the correlation between quantifications of charge +2 and charge greater than +2 spectra that were independently assigned to the same peptide and PTM site by database search (Fig. 4.8). This test for the correlation of charge isoforms is a rigorous demonstration of the accuracy of database search due to the vastly different fragmentation ion spectra produced by collision-induced dissociation of ions of different charge states. High correlations (Spearman's correlation 0.85 and 0.87) confirmed the accuracy of our methods for PTM quantification and site localization.

Data Set	**Description**	**PTM Enriched**	**Size**	**Precursor Mass Window**	**Processing Time**
de Godoy *et al.* (2008)	A very large data set that compares proteins from haploid yeast to diploid yeast. The authors performed extensive gradient and gel fraction of the sample.	No	225 instrument runs, 32GB	±3.5 ppm	18 minutes 21 seconds
Geiger *et al.* (2010)	This data set compares a lobular and ductal cancer sample to a mix of isotope labeled cancer cell lines. The authors collected 3 replicates of each cancer sample.	No	36 instrument runs, 10GB	4 ppm	10 minutes 56 seconds
Daub *et al.* (2008)	The authors combined kinase selective affinity purification with TiO2 phosphopeptide enrichment to compare isotope labeled samples between S and M phase arrested HeLa cells.	Yes	149 instrument runs, 12GB	±5 ppm	10 minutes 49 seconds
Van Hoof *et al.* (2009)	The authors compared a TiO2 phosphpeptide enriched sample from differentiated human embryonic stem cells (hESCs) 30 minutes, 60 minutes, and 240 minutes after induction with BMP4 to isotope labeled undifferentiated hESCs.	Yes	72 instrument runs, 41GB	±8 ppm	17 minutes 23 seconds

Table 4.1: All data sets were processed with a database search FDR 1% and PTM positional localization FDR 10%. MS/MS tolerance was set to ±0.5 Da. Carboxyadmidomethylation of cysteine was the only fixed modification. Up to 4 modified amino acids were allowed per fragment. Methionine oxidation and serine, threonine, and tyrosine phosphorylation (including neutral losses of H3PO4 for serine and threonine) were used as variable modifications. Up to 1 missed trypsin cleavage was allowed. Running times were collected using 8 CPU threads on a 2-CPU 2.93GHz Xeon X5570 CPU with 68GB of RAM (8 physical CPU cores total). Original .RAW files were converted to centroided mzXML files prior to running PVIEW.

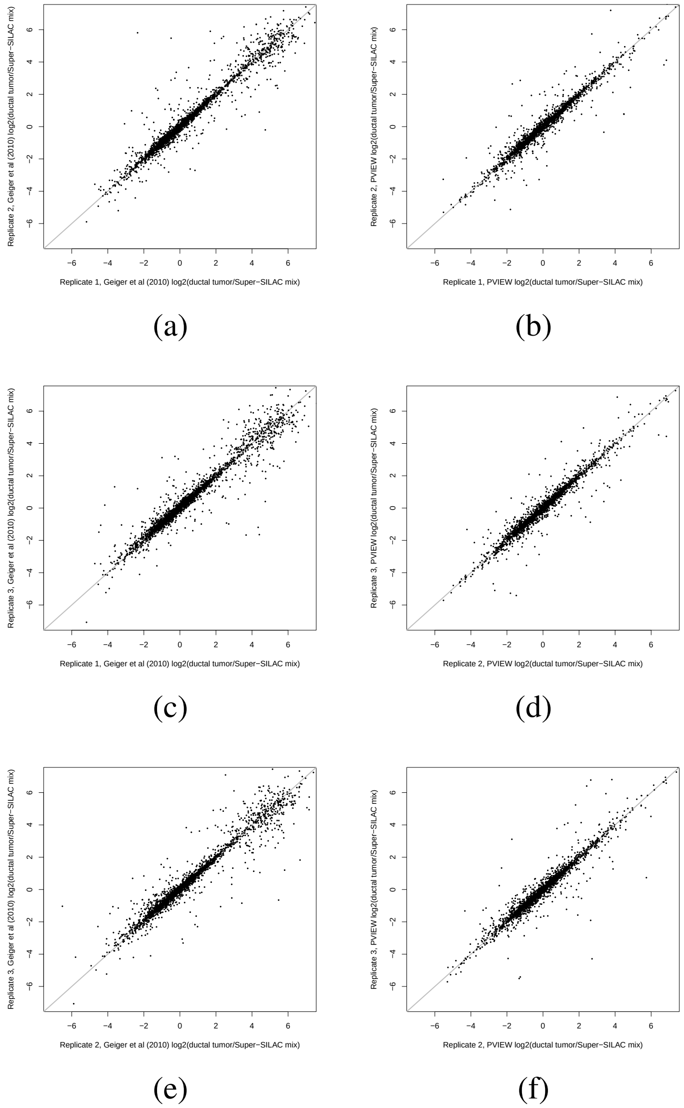


(a)

(b)

(c)

Figure 4.6: (a) Correlation between protein ratios reported in de Godoy et al (2008) using Maxquant for a large data set quantifying the proteomes of haploid and diploid yeast with ratios computed on the same data set using our approach, (PVIEW). Correlation between ratios reported in Geiger et al (2010) with those computed by PVIEW for a data set quantifying (b) a ductal cancer sample and (c) lobular sample using a Super-SILAC mix of isotope labeled cell lines. Perfect correlation is designated by the diagonal gray line.

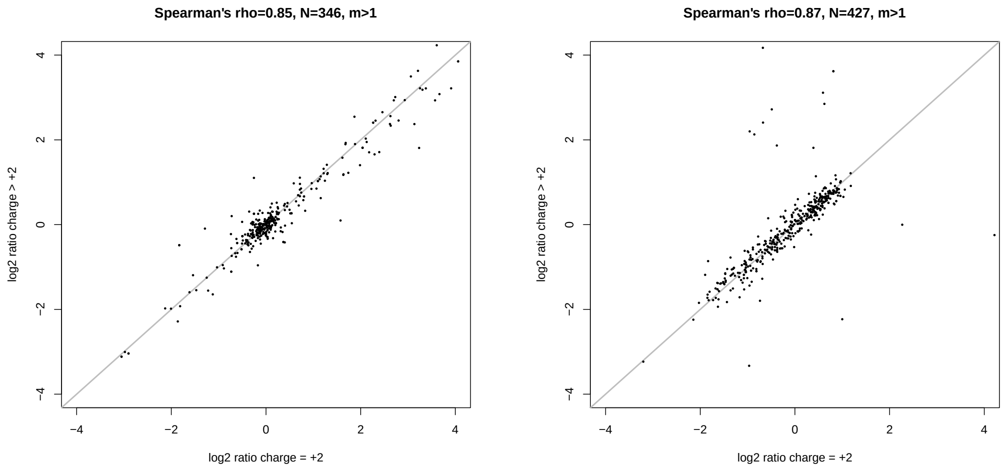


(a) (b)

(c) (d)

(e) (f)

Figure 4.7: (a), (c), and (e) correlation between protein abundance ratios of 3 technical replicates reported in Geiger et al (2010) using Maxquant for a ductal cancer sample compared to a Super-SILAC mix of isotope labeled cell lines. (b), (d), and (f) correlations between the 3 replicates for the same data set as reported by our approach (PVIEW). Note larger variance in expression differences of 4-fold or greater as reported by Maxquant. Perfect correlation is designated by the diagonal gray line.



(a) (b)

Figure 4.8: Correlations of $\log_2$ ratios (labeled over unlabeled) between posttranslationally modified peptides where both the charge +2 and charge > +2 isoforms where both quantified by our approach (PVIEW) from (a) Daub et al (2008) and (b) Van Hoof et al (2009). From the Daub et al (2008) data set PVIEW quantified 5,343 unique unmodified peptides and 4,174 unique modified peptides. 343 of these modified peptides had both the charge +2 and charge > + 2 isoforms quantified. From the Van Hoof et al (2008) data set, PVIEW quantified 17,668 unique unmodified peptides and 10,817 modified peptides. 427 of these modified peptides had both the charge +2 and charge > +2 isoforms quantified. Perfect correlation is designated by the diagonal gray line.

Chapter 5

^{15}N Peptide Quantification

This chapter introduces new methods for quantification using an isotope label in which the shifts between XICs depend on the nitrogen composition of the peptide. Experimental methods for proteome-wide quantification using liquid chromatography-coupled mass spectrometry rely heavily on algorithms that analyze mass spectra (Mortensen *et al.* (2010); Park *et al.* (2008); Cox and Mann (2008); Khan *et al.* (2009); Keller *et al.* (2005)). Computational analysis typically uses the following framework. First, a peptide identification algorithm assigns peptide sequences to fragmentation spectra by searching a database of theoretical peptides obtained by an in silico digest of an organism's proteome. Second, a quantification algorithm, independent of the peptide database, coordinates intensity measurements in mass spectra with peptide identifications.

This framework has been shown to produce robust and accurate relative quantification results using an experimental methodology where isotope-labeled amino acids are incorporated metabolically by an organism, typically in one of two experimental conditions (Ong *et al.* (2002); Cox and Mann (2008)). Because both samples are combined prior to protein extraction, the samples are subjected to the same extraction, sample handling, digestion, chromatography, and ionization conditions. This eliminates much

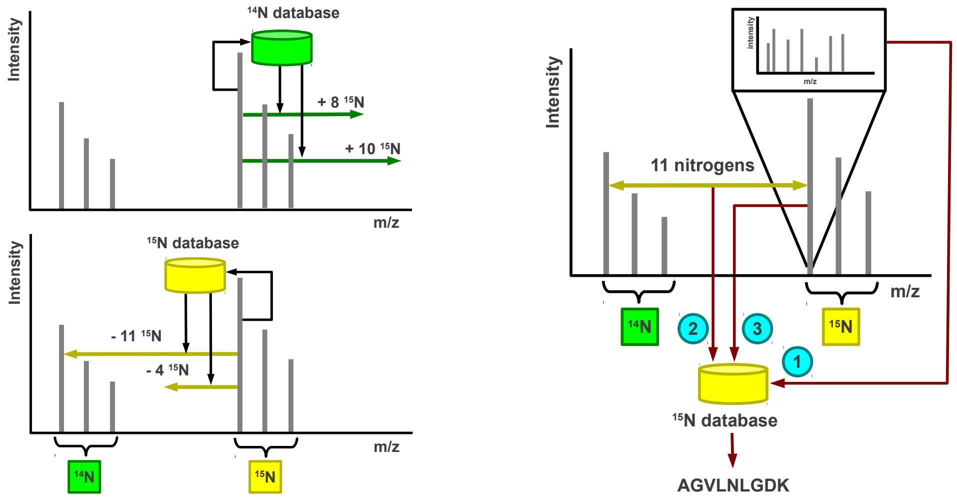


Figure 5.1: (a) Our integrative approach uses mass measurements to search a 14N unlabeled database and a 15N labeled database of peptides obtained by in silico digest of an organism's proteome. The number of nitrogens in each of the returned peptides is used to examine a limited number of mass differences designated by arrows. On finding an unambiguous 14N-15N pair, the algorithm labels each member of the pair as originating from either the unlabeled or the 15N labeled sample. (b) During peptide identification, a fragmentation spectrum (1) associated with a member of the pair is searched against an unlabeled database or 15N labeled database of peptides based on the assigned label. The nitrogen composition of a peptide (2), in addition to the mono-isotopic mass (3), is used to limit the search space of peptides scored against the spectrum. The intensities of each member of the 14N-15N pair are used to derive a peptide ratio.

of the technical variation between individual samples and enables robust measures of protein expression differences. Furthermore, this methodology allows for multiple intensity measurements to contribute to a peptide's quantification value, as several mass spectra are acquired over the duration of its chromatographic elution. Consequently, the final relative abundance measurements, in the form of ratios of the areas of paired extracted ion chromatograms (XICs) from unlabeled and labeled peptides, are highly accurate. To assure that only labeled amino acids are incorporated into proteins, amino acid auxotrophs are typically employed in conjunction with labeled amino acids, rendering this strategy unavailable for prototrophic microorganisms, such as wild-type bacteria and yeast.

15N labeling, through the metabolic incorporation of a sole labeled nitrogen source, provides an alternative strategy for metabolic labeling of virtually all the nitrogens in expressed proteins in prototrophs and auxotrophs (Oda *et al.* (1999)). 15N labeling has two notable advantages to amino acid labels: all peptides can be used for quantification regardless of the endoprotease used to generate peptide fragments, and the label is not subject to the complications of metabolic conversion of amino acids (Park *et al.* (2009)). Despite these substantial advantages of 15N labeling, labeled amino acids have been used in favor of 15N labeling for proteome-wide quantification (de Godoy *et al.* (2008)). This is due to the fact that amino acid labels typically produce a small number of distinct mass differences between pairs of XICs, whereas 15N labels induce mass differences in spectra that depend on the length and composition of a peptide, making the downstream analysis significantly more challenging.

5.1 An Integrative Algorithm

To address the problem of length and composition dependent mass differences, we developed an algorithm that implements a computational analysis framework in which

XIC pairing is integrated with the database search component of peptide identification; we integrated these two components such that each informs and provides additional constraints for the other. Our algorithm uses the accurate mass information from an XIC to search entries in unlabeled and 15N-labeled peptide databases and uses the returned nitrogen counts to examine a limited number of 15N mass differences and pair XICs (Fig. 5.1a). Because absolute nitrogen counts are used, this step is resilient to complex isotope patterns of partially labeled peptides, if they are present. On unambiguously pairing XICs, the algorithm designates each XIC as originating from either the unlabeled or the 15N labeled sample. This process is repeated for all XICs until no more XICs can be paired. During peptide identification, the algorithm iterates through paired XICs that have an associated fragmentation spectrum. Based on the label assigned to an XIC, the algorithm searches the fragmentation spectrum against an unlabeled database or a database of 15N labeled peptides. In addition to the mass of the peptide, this search uses the discovered constraints on the nitrogen composition of a peptide from the pairing phase to obtain an identification (Fig. 5.1b). Last, the peptide expression ratios are obtained from the areas of paired XICs.

5.2 Validation

In order to validate this algorithm and its quantitative output, we performed a label-swap experiment using E. coli samples comparing stationary to exponential phase cells. We selected these two conditions because the transition to stationary phase is well characterized and the expression levels of a large number of proteins during this transition are known to differ over a wide range (Han and Lee (2006)). To conduct the label-swap experiment, we used two biological replicates. In the first replicate, cells in stationary phase were cultured in 15N labeled media and cells in exponential phase were cultured in unlabeled 14N media. In the second replicate, cells were independently cultured with

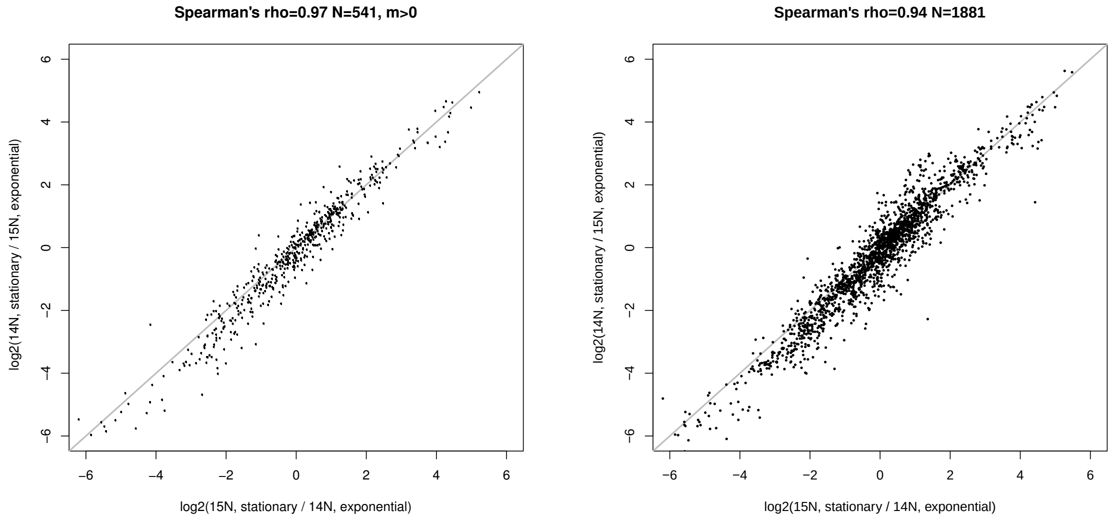


Figure 5.2: (a) Correlation of $\log_2$ protein ratios computed from an unfractionated, label-swap experiment in which two biological replicates where prepared comparing E. coli cells growing in stationary phase to exponential phase. (b) Correlation of $\log_2$ peptide ratios computed from an unfractionated, label-swap experiment in which two biological replicates where prepared comparing E. coli cells growing in stationary phase to exponential phase.

the labeled media swapped between stationary phase and exponential phase cells. LC-MS analysis of whole-proteome samples enabled detection and quantification of 701 proteins and 2,910 peptides of which 541 proteins and 1,881 peptides were quantified in both biological replicates at 1% peptide level false discovery rate. High correlations across protein and peptide ratios from the two biological label-swap replicates indicated the analysis was highly reproducible (Fig. 5.2). In order to extend the detection range of LC-MS to proteins with low abundance, we subjected one of the two label-swap samples to additional fractionation steps. First, the sample was partitioned based on protein solubility into two fractions: high solubility and low solubility. Next, peptides derived from each of these two protein level fractions were further fractionated using strong cation exchange (SCX) chromatography. These additional fractionation steps enabled quantification of 2,039 distinct proteins, 10,165 distinct peptides, and 27,037 quantitative measurements at a 1% peptide level false discovery rate.

To evaluate the algorithm on this deeper survey of the proteome, we examined the correlations between the ratios measured for different peptides from the same protein (Fig. 5.3). High correlations indicate that the algorithm could accurately identify and quantify distinct peptides from the same protein to measure the fold-change of the whole protein. In addition, we correlated ratios between charge +2 and charge greater than +2 isoforms of the same peptide (Fig. 5.3). Even though the differing charge states produce very different fragmentation spectra, the peptide ratios from the two charge states are highly correlated. Next, we examined the intra-experiment variability of the ratios reported by the algorithm. From the ratios of distinct peptides from each protein, we calculated a protein level median coefficient of variation (CV) of ~20% (Fig. 5.4). We also calculated a peptide level median CV of ~12% from ratios of peptides observed in multiple fractions or with differing charge states (Fig. 5.4). These low CVs at both the protein level and peptide level indicate high quantification accuracy.

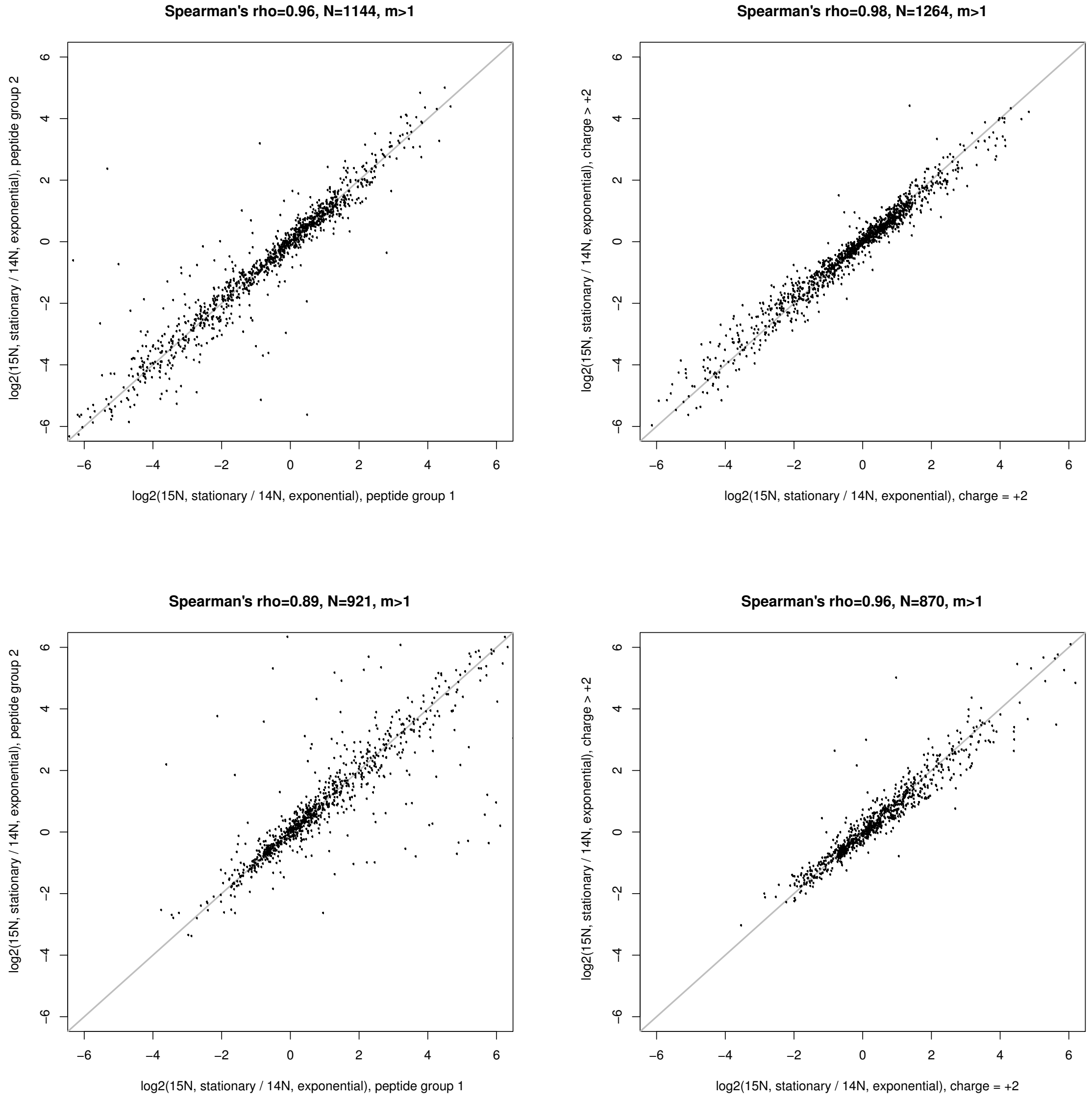
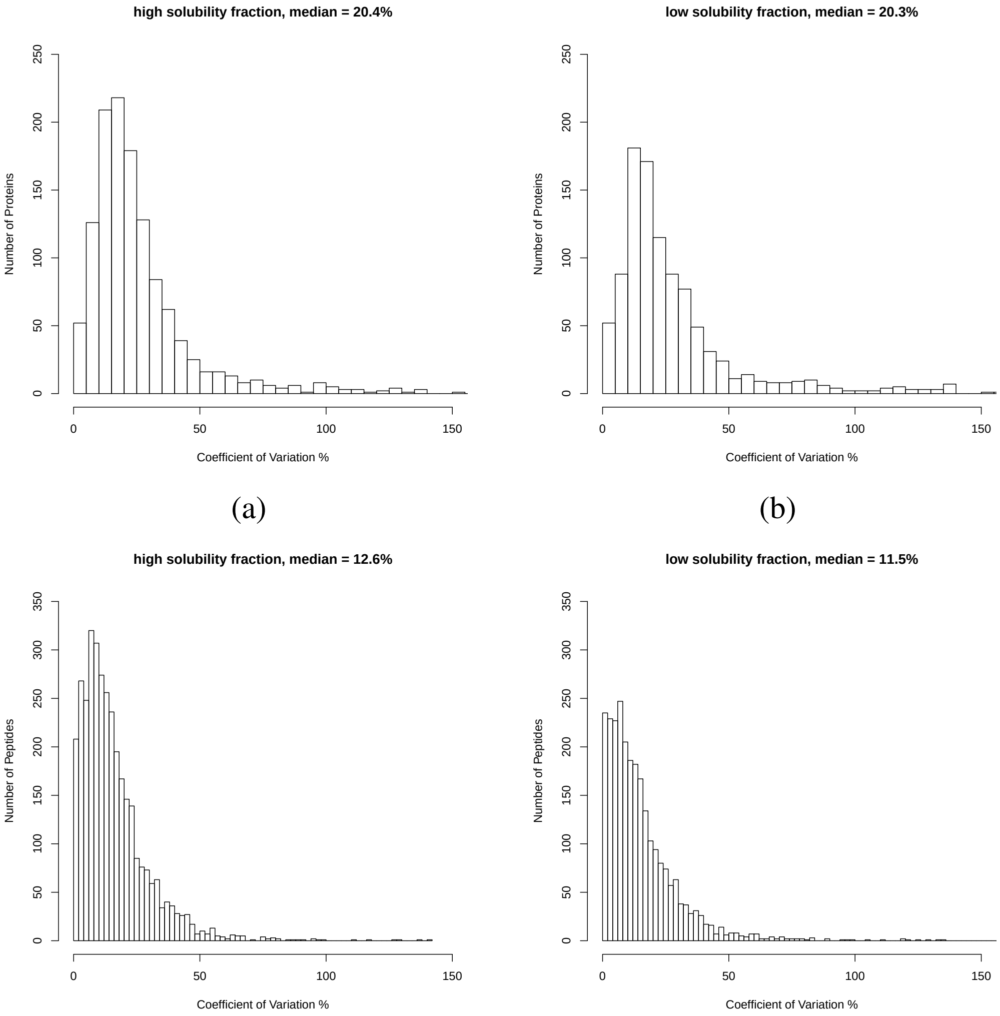


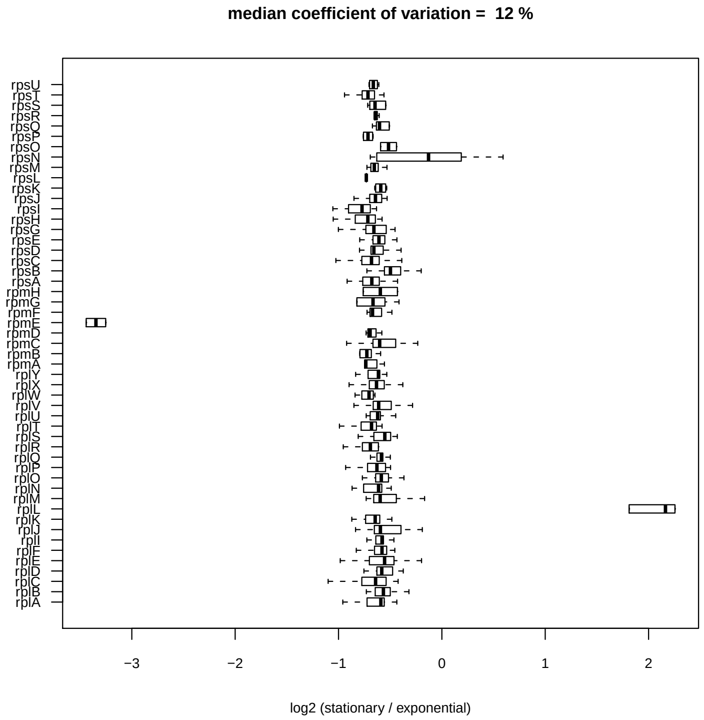
Figure 5.3: (a) Correlation of $\log_2$ ratios between distinct peptides separated into two groups, all from the same protein, from the SCX fractionated high solubility protein fraction comparing stationary to exponential phase. (b) Correlation of log2 stationary vs. exponential ratios comparing charge isoforms of the same peptide from the same set of SCX fractions. (c) Correlation of $\log_2$ ratios between distinct peptides from the same protein in an SCX fractionated low solubility protein fraction comparing stationary to exponential phase. (d) Correlation of $\log_2$ stationary vs. exponential ratios comparing charge isoforms of the same peptide from the low solubility SCX fractions.



(a) (b)

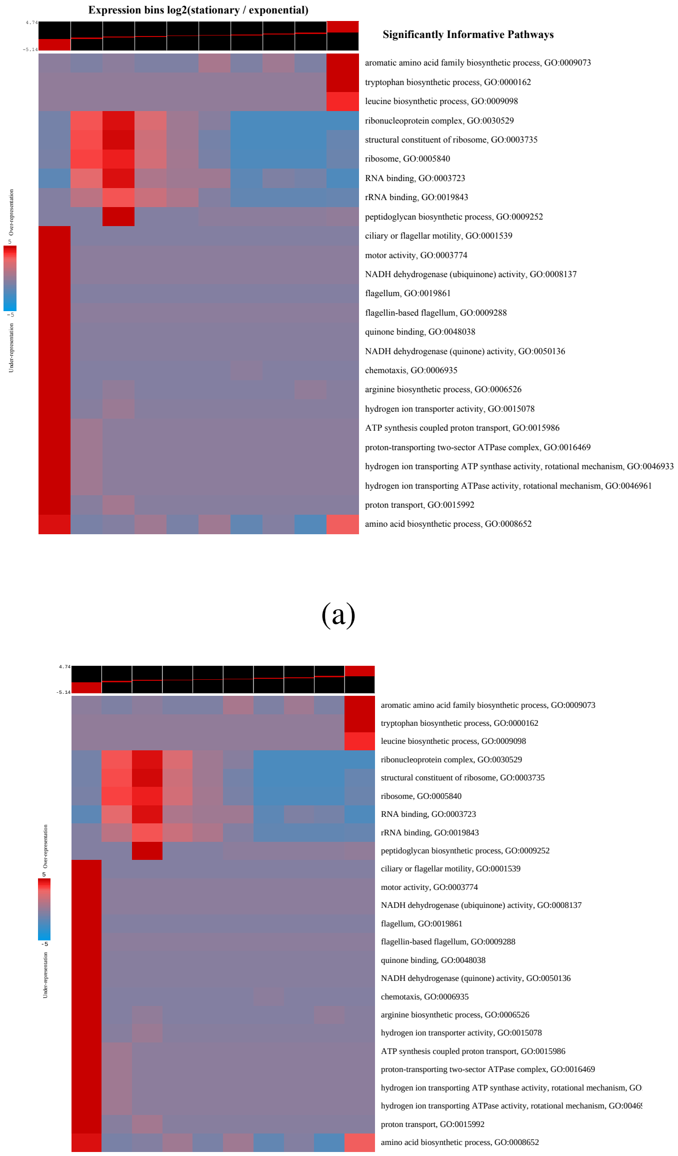
(c) (d)

Figure 5.4: (a) and (b) show histograms of the intra-experiment coefficient of variation (CV) estimated using pooled ratios of distinct peptides from each protein for the high solubility and lows solubility protein fractions respectively. (c) and (d) show histograms of the CV estimated using pooled ratios of peptides observed in multiple SCX fractions or with differing charge states for both the high solubility and low solubility protein fractions.



(a)

Figure 5.5: (a) Box and whisker plot of the log2 ratios of stationary over exponential phase for 51 ribosomal proteins from the low solubility protein fraction. Boxes designate the quartiles. The thick center line designates the median. Thin lines illustrate the range.



(a)

(b)

Figure 5.6: A graphical representation of the pathway analysis method of Goodarzi et al. (4) applied to protein expression ratios (a) and transcript ratios (b) comparing stationary to exponential phase of growth in E. coli computed by our integrative algorithm. In this representation, rows correspond to significantly informative pathways and columns correspond to 10 equally-populated expression bins of log2 ratios (stationary over exponential). Colors indicate pathway over- or under-representation levels across the expression bins. Red designates (in log10) over-representation of a particular pathway in any expression bin, whereas, blue designates under-representation.

Next, we investigated whether the quantitative output captured expression level changes of proteins known to differ between exponential and stationary phase. Globally decreased rates of translation and ribosome biosynthesis characterize stationary phase growth (Aviv *et al.* (1996)). In strong agreement with expectation, we observed that 49 out of 51 quantified ribosomal proteins were down-regulated by the same amount on transition into stationary phase (Fig 5.5). Another known feature of stationary phase is that cells adapt to higher cell density and nutrient limitation by inducing the expression of many stress proteins, including starvation-induced DNA protection protein, Dps, and osmotically inducible proteins, OsmC and OsmY (Han and Lee (2006)). This is orchestrated by several regulators, including the RNA polymerase sigma factor, RpoS (Lange and Hengge-Aronis (1991)). We observed elevated expression levels of each of these proteins in stationary phase cells, in agreement with previous results.

To examine whether our measured protein expression differences on a global level accurately characterize the differences between exponential phase and stationary phase cells, we conducted a pathway analysis that bins expression differences based on their magnitudes and directions and determines which biological processes are over- or under-represented in these bins (Fig. 2b) (Goodarzi *et al.* (2009)). Our results demonstrated that, in addition to cessation of ribosome biosynthesis (Aviv *et al.* (1996)), we captured many other adaptive physiological changes across major biological processes that are known features of stationary phase in E. coli. These include repression of flagellum biosynthesis and motility associated processes (Patten *et al.* (2004)) and down-regulation of oxidative metabolism proteins (Yoon *et al.* (2003)).

Further, we assessed the overall agreement between changes in protein abundances and changes in transcript abundances by simultaneously conducting transcriptomic profiling on the same samples using two-channel microarrays. While the magnitudes of protein expression ratios differed from the magnitudes of gene expression ratios, we observed a statistically significant overlap between the directionality of the changes (Pearson's chi-square p-value of 2.25E-25 and 5.89E-15 for native and denatured preps, respectively). Pathway analysis separately conducted on the transcripts that displayed abundance differences revealed over-represented biological processes that are in high correspondence with those obtained using the protein data (Fig. 5.6a, c.f., Fig 5.6b). This underscores that, while particular transcript-level and protein-level regulation may diverge, overall trends are conserved, and it provides additional evidence that our computational framework provides an accurate global read-out of underlying biological differences. Further investigation of specific discrepancies between transcriptional and protein level differences remain an interesting avenue for future work.

5.3 Conclusions

In summary, we have introduced an integrated framework for mass spectrometry-based protein identification and quantification using 15N-labeled samples. We show that this approach addresses the challenges of composition and length dependent mass differences by leveraging the additional constraints gained when identification is allowed to inform quantification and vice-versa. Integrative approaches to compound quantification and identification, such as ours, may have broad applicability to mass spectrometry data analysis.

Chapter 6

Conclusions

The primary contribution of this thesis is a new application of space partitioning data structures. This new application is based on the idea that peaks in mass spectra collected by LC-MS/MS are a set of 2.5d points, (m/z, time) plus an attribute intensity, allowing the data to be treated as points on a plane. This leads to a natural application of techniques from computational geometry, in particular, the orthogonal range query, and series of simple algorithms that enable the rapid extraction of quantitative measurements in the form of eXtracted ion chromatograms (XICs).

Compared to previous approaches for processing mass spectra, which process a single spectrum at a time or rely on images as the primary data structure, this thesis shows that this new approach significantly improves the speed of computational analysis. Using a wide range of LC-MS/MS data sets spanning several experimental methodologies and source protein samples, this thesis also shows these methods significantly improve the quantitative accuracy of protein expression measurements obtained from LC-MS/MS mass spectra.

Another contribution of this thesis is a collection methods, implemented in an open source software system called PVIEW. PVIEW uses the output of the algorithms developed in this thesis to enable accurate quantification of proteins, protein fragments, and posttranslational modifications.

The work in this thesis is far from finished, and below are descriptions of several avenues for future work.

1. **Develop methods that analyze the shapes of XICs.** The shapes of chromatographic peaks in XICs reveal aspects of the quality of the liquid chromatography. Problems are revealed by three types or artifacts: (1) fronting where the peak shape acquires a long tail at early time points (2) tailing where later time points acquire a long tail, and (3) sputtering and splitting where an XIC breaks apart into smaller XICs. Computational methods might be able to address, in part, several of these common problems to improve quantification despite the presence of these artifacts.

2. **Develop a theoretically justified method for determining the statistical significance of a match between a mass spectrum and a peptide sequence.** Mass spectrometry currently lacks a theoretically justified null model for peptide spectrum matches. It still remains unclear what biases that the current dominant approach for evaluating the significance of peptide spectrum matches introduce or if these approaches can somehow be justified by theory (Käll *et al.* (2008); Elias and Gygi (2007)).

3. **Develop methods that correctly capture ambiguity in the interpretation of mass spectra.** Mass spectra have inherent ambiguities that become increasingly important to consider for problems that involve the localization of an amino acid substition or position of a posttranslational modification. While scoring methods such as Beausoleil *et al.* (2006) have been developed, they are difficult to interpret and fail to convey the exact ambiguity present in mass spectrum peptide match. A method that summarizes ambiguity graphically, in conjunction to a score, might assure correct interpretation of mass spectrometry data by non-experts.

4. **Apply the techniques developed here to find XICs during acquisition of mass spectra over an LC-MS instrument run. Use those results to target fragmentation.** The methods developed here can be applied to a dynamically growing space partitioning data structure such as a kd-tree as mass spectra are acquired over time. If XICs can be found dynamically, fragmentation can be targeted such that a spectrum is acquired when the chromatogram is near its peak and the ion is most abundant. Efficiently re-balancing kd-trees and related space partitioning data structures in the general case remains an open problem, but theoretical guarantees on re-balancing might be possible if points are added only along one dimension as in the case of LC-MS data (Samet (2005)).

6.1 Future Outlook

LC-MS is likely to remain the dominant technology for quantitative measurement of proteins and their posttranslational modifications for years to come. Advances in mass analyzers will likely increase the rate of acquisition of mass spectra over chromatographic elution, enabling fine sampling of XICs and highly accurate quantification. Advances in chromatography will enable the separation and analysis of increasingly complex mixtures of proteins. Improvements in ionization methods may one day assure detection of all proteolytic peptides from every single protein present in a sample, enabling, in a single experiment, the simultaneous measurement of the tremendous diversity of proteins that can be produce by a genome. These avenues for technological advancement will enable the rapid acquisition of data sets terabytes in size comprised of millions or billions of mass spectra. The algorithms developed in this thesis have the potential to meet the increasing computational demands of these data. Therefore, they have the potential to play a key role in biological discovery for years to come.

Appendix A

PVIEW Source Code

All of the PVIEW source code can be downloaded freely under the BSD open source license from the following web site:

http://compbio.cs.princeton.edu/pview

PVIEW was written in C++ and based on the cross-platform GUI toolkit Qt (http://qt.nokia.com/products/) and the Expat XML parser for loading mass spectra from a data file. PVIEW has been tested on Linux, MacOS, and Windows for both 32-bit and 64-bit platforms. The source code is structured so that the algorithmic components describe in this thesis can be developed independently of the GUI.

Below is a description of each of the source files from the PVIEW:

`2dtree.hpp` Implementation of a kd tree ($k = 2$) with an interface for orthogonal range queries based on the in-place construction ideas presented in Brönnimann *et al.* (2004).

`lcms.[h,c]pp` Provides an implementation of the algorithms for finding XICs presented in Chapter 2. Includes algorithms for the quantification strategies presented in Chapter 4. Also includes an implementation of the quantification strategy for ^{15}N labeled mass spectra presented in Chapter 5.

`ms2.[h,c]pp` Includes algorithms presented in Chapter 3 for identifying peptides by database search of fragmentation spectra.

`fdr.[h,c]pp` Implements an approach by Käll *et al.* (2008) for estimating the statistical significance of a peptide spectrum match.

`analysis.[h,c]pp` Simple algorithms that collect quantitative and peptide identification data for output to a user. Also provides an organizational framework for the analysis of large LC-MS data sets spanning many instrument runs, experiments, and replicates.

`msconfig.hpp` Contains, in a single location, configuration parameters for all algorithms in PVIEW.

`centroider.hpp` Includes an algorithm for converting profile mode spectra to centroid mode spectra as described in Figure 1.1.

`util.hpp` Several utility functions, not provided by any standard libraries, used throughout PVIEW.

`xml.[h,c]pp` Provides an interface for loading mass spectra and configuration data from XML files. Depends on the Expat XML parser library.

`MSWin.[h,c]pp` Implements a GUI for visualization of LC-MS data. Depends on the Qt GUI library.

`MSMSWin.[h,c]pp` GUI component for visualizing fragmentation spectra and any peptide spectrum database search results. Depends on the Qt GUI library.

`dialogs.[h,c]pp` GUI components for configuring and interacting with PVIEW components. Depends on the Qt GUI library.

`reports.R` Provides an R statistical programming language script for producing informative plots that summarize the quality of a particular data set.

Bibliography

Aviv, M., Giladi, H., Oppenheim, A., and Glaser, G. (1996). Analysis of the shut-off of ribosomal RNA promoters in Escherichia coli upon entering the stationary phase of growth. *FEMS Microbiol Lett*, **140**, 71–76.

Bandeira, N., Tsur, D., Frank, A., and Pevzner, P. A. (2007). Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences*, **104**(15), 6140–6145.

Bartels, C. (1990). Fast algorithm for peptide sequencing by mass spectroscopy. *Biological Mass Spectrometry*, **19**, 363–368.

Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotech*, **24**(10), 1285–1292.

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**(15), 1902–1909.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, **18**(9), 509–517.

Brönnimann, H., Chan, T. M., and Chen, E. (2004). Towards in-place geometric algorithms and data structures. In *Proceedings of the 20th Annual ACM Symposium on Computational Geometry (SoCG)*, pages 239–246, Brooklyn, NY. ACM Press.

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

Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, **26**, 1367 – 1372.

Craig, R. and Beavis, R. C. (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry*, **17**(20), 2310–2316.

Daub, H., Olsen, J., Bairlein, M., Gnad, F., Oppermann, F., Korner, R., Greff, Z., Keri, G., Stemmann, O., and Mann, M. (2008). Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Molecular Cell*, **31**(3), 438–448.

de Berg, M., Cheong, O., van Kreveld, M., and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin.

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**(7217), 1251–1254.

Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, **4**(3), 207–214.

Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). An approach to correlate ms/ms data to amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

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. *Analytical Chemistry*, **80**(4), 961–971.

Foss, E., Radulovic, D., Shaffer, S., Ruderfer, D., Bedalov, A., Goodlett, D., and Kruglyak, L. (2007). Genetic basis of proteome variation in yeast. *Nature Genetics*, **39**, 1369–1375.

Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004). Open mass spectrometry search algorithm. *Journal of Proteome Research*, **3**(5), 958–964.

Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., and Mann, M. (2010). Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature Methods*, **7**(5), 383–385.

Goodarzi, H., Elemento, O., and Tavazoie, S. (2009). Revealing global regulatory perturbations across human cancers. *Molecular Cell*, **36**(5), 900 – 911.

Han, M. and Lee, S. (2006). The Escherichia coli proteome: past, present, and future prospects. *Microbiol Mol Biol Rev*, **70**, 362–439.

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**(10), 1927–1941.

Käll, L., MacCoss, J. D. S. M. J., and Noble, W. S. (2008). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research*, **7**, 29–34.

Katz, I. (2005). *Space-efficient geometric algorithms and data structures*. Master's thesis, Polytechnic University, Brooklyn, NY.

Keller, A., Eng, J., Zhang, N., Li, X., and Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology*, **1**(2005.0017).

Khan, Z., Bloom, J. S., Garcia, B. A., Singh, M., and Kruglyak, L. (2009). Protein quantification across hundreds of experimental conditions. *Proceedings of the National Academy of Sciences*, **106**(37), 15544–15548.

Kohlbacher, O., Reinert, K., Gropl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., and Sturm, M. (2007). TOPP - the OpenMS proteomics pipeline. *Bioinformatics*, **23**(2), 191–197.

Lange, R. and Hengge-Aronis, R. (1991). Identification of a central regulator of stationary-phase gene expression in Escherichia coli. *Mol Microbiol*, **5**, 49–59.

Listgarten, J. and Emili, A. (2005). Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry. *Mol. Cell Proteomics*, **4**(4), 419–434.

Maneewongvatana, S. and Mount, D. M. (1999). It's okay to be skinny, if your friends are fat. In *Center for Geometric Computing 4th Annual Workshop on Computational Geometry*.

Mann, M. and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**(24), 4390–4399.

Mortensen, P., Gouw, J. W., Olsen, J. V., Ong, S.-E., Rigbolt, K. T. G., Bunkenborg, J., Cox, J., Foster, L. J., Heck, A. J. R., Blagoev, B., Andersen, J. S., and Mann, M. (2010). MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *Journal of Proteome Research*, **9**(1), 393–403.

Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, S., Brusniak, M. Y., Vitek, O., Aebersold, R., and Muller, M. (2007). SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*, **7**(19), 3470–3480.

Nesvizhskii, A. I. and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics*, **4**(10), 1419–40.

Noy, K. and Fasulo, D. (2007). Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics*, **23**(19), 2528–2535.

Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences*, **96**(12), 6591–6596.

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**(5), 376–386.

Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S., Burgess, J., Zimmermann-Ivol, C. G., Sanchez, J.-C., Binz, P.-A., Hochstrasser, D. F., and Appel, R. D. (2005).

MSight: An image analysis software for liquid chromatography-mass spectrometry. *Proteomics*, **5**(9), 2381–2384.

Park, S. K., Venable, J. D., Xu, T., and Yates, J. R. (2008). A quantitative analysis software tool for mass spectrometry-based proteomics. *Nature Methods*, **5**(4), 319–322.

Park, S. K., Liao, L., Kim, J. Y., and Yates, J. R. (2009). A computational approach to correct arginine-to-proline conversion in quantitative proteomics. *Nature Methods*, **6**(3), 184–185.

Patten, C., Kirchhof, M., Schertzberg, M., Morton, R., and Schellhorn, H. (2004). Microarray analysis of RpoS-mediated gene expression in Escherichia coli K-12. *Mol Genet Genomics*, **272**, 580–591.

Renard, B. Y., Kirchner, M., Monigatti, F., Ivanov, A. R., Rappsilber, J., Winter, D., Steen, J. A. J., Hamprecht, F. A., and Steen, H. (2009). When less can yield more - computational preprocessing of MS/MS spectra for peptide identification. *Proteomics*, **9**(21), 4978–4984.

Rockwood, A. L., Orden, S. L. V., and Smith, R. D. (1995). Rapid calculation of isotope distributions. *Analytical Chemistry*, **67**(15), 2699–2704.

Samet, H. (2005). *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Stein, S. E. and Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, **5**(9), 859 – 866.

Van Hoof, D., Munoz, J., Braam, S. R., Pinkse, M. W., Linding, R., Heck, A. J., Mummery, C. L., and Krijgsveld, J. (2009). Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell stem cell*, **5**(2), 214–226.

Yoon, S., Han, M., Lee, S., Jeong, K., , and Yoo, J. (2003). Combined transcriptome and proteome analysis of Escherichia coli during high cell density culture. *Biotechnol Bioeng*, **81**, 753–767.