STATISTICAL METHODS FOR THE ANALYSIS OF MASS SPECTROMETRY-BASED PROTEOMICS DATA

A Dissertation

by

XUAN WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012

Major Subject: Statistics

STATISTICAL METHODS FOR THE ANALYSIS OF MASS SPECTROMETRY-BASED PROTEOMICS DATA

A Dissertation

by

XUAN WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Alan R. Dabney
Committee Members,	Michael T. Longnecker
	Huiyan Sang
	Joseph Sturino
Head of Department,	Simon J. Sheather

May 2012

Major Subject: Statistics

ABSTRACT

Statistical Methods for the Analysis of Mass Spectrometry-based Proteomics Data.

(May 2012)

Xuan Wang, B.S., University of Science and Technology of China, Anhui, China;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. Alan R. Dabney

Proteomics serves an important role at the systems-level in understanding of biological functioning. Mass spectrometry proteomics has become the tool of choice for identifying and quantifying the proteome of an organism. In the most widely used bottom-up approach to MS-based high-throughput quantitative proteomics, complex mixtures of proteins are first subjected to enzymatic cleavage, the resulting peptide products are separated based on chemical or physical properties and then analyzed using a mass spectrometer. The three fundamental challenges in the analysis of bottom-up MS-based proteomics are as follows: (i) Identifying the proteins that are present in a sample, (ii) Aligning different samples on elution (retention) time, mass, peak area (intensity) and etc, (iii) Quantifying the abundance levels of the identified proteins after alignment. Each of these challenges requires knowledge of the biological and technological context that give rise to the observed data, as well as the application of sound statistical principles for estimation and inference. In this dissertation, we present a set of statistical methods in bottom-up proteomics towards protein identification, alignment and quantification.

We describe a fully Bayesian hierarchical modeling approach to peptide and protein identification on the basis of MS/MS fragmentation patterns in a unified framework. Our major contribution is to allow for dependence among the list of top candidate PSMs, which we accomplish with a Bayesian multiple component mixture

model incorporating decoy search results and joint estimation of the accuracy of a list of peptide identifications for each MS/MS fragmentation spectrum. We also propose an objective criteria for the evaluation of the False Discovery Rate (FDR) associated with a list of identifications at both peptide level, which results in more accurate FDR estimates than existing methods like PeptideProphet.

Several alignment algorithms have been developed using different warping functions. However, all the existing alignment approaches suffer from a useful metric for scoring an alignment between two data sets and hence lack a quantitative score for how good an alignment is. Our alignment approach uses "Anchor points" found to align all the individual scan in the target sample and provides a framework to quantitify the alignment, that is, assigning a p-value to a set of aligned LC-MS runs to assess the correctness of alignment. After alignment using our algorithm, the p-values from Wilcoxon signed-rank test on elution (retention) time, M/Z, peak area successfully turn into non-significant values.

Quantitative mass spectrometry-based proteomics involves statistical inference on protein abundance, based on the intensities of each protein's associated spectral peaks. However, typical mass spectrometry-based proteomics data sets have substantial proportions of missing observations, due at least in part to censoring of low intensities. This complicates intensity-based differential expression analysis. We outline a statistical method for protein differential expression, based on a simple Binomial likelihood. By modeling peak intensities as binary, in terms of "presence / absence", we enable the selection of proteins not typically amendable to quantitative analysis; e.g., "one-state" proteins that are present in one condition but absent in another. In addition, we present an analysis protocol that combines quantitative and presence / absence analysis of a given data set in a principled way, resulting in a single list of selected proteins with a single associated FDR.

To Jason Tang, Xincheng Tang, Shuhuan Wang and Hanming Wang

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to my advisor, Dr. Alan R. Dabney, for his direction, suggestion, encouragement, patience and continuous support toward my professional development. He guided me to think about statistical research problems, how to do research and how to write a paper step by step.

I am very grateful to all my committee members, Dr. Michael Longnecker, Dr. Huiyan Sang, Dr. Joseph Sturino, for their guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculties and staff for making my time at Texas A&M University a great experience.

Finally, I feel lucky to have my husband, my son and my parents be my side, thank you for your unconditional love.

This research is supported by the. This work was sponsored by a subcontract from PNNL and by the NIH R25-CA-90301 training grant at TAMU. Additional support was provided by KAUST-IAMCS Innovation grant, by NIH grant DK070146 and by the National Institute of Allergy and Infectious Diseases (NIH/DHHS through interagency agreement Y1-AI-4894-01).

NOMENCLATURE

AMT	Accurate mass and time
ANOVA	Analysis of variance
CDF	Cumulative density function
FDR	False discovery rate
GEV	Generalized extreme value
LC-MS	Liquid chromatographyCMass spectrometry
M/Z	Mass over charge ratio
NET	Normalized elution time
NMC	Number of missed cleavage sites
NTE	Number of tryptic ends
PEP	Posterior error probability
PM	Potential matches
PMF	Probability mass function
PNNL	Pacific northwest national laboratory
PSM	Peptide spectrum match
RBF	Radial basis function
RT	Retention Time
SVM	Support vector machines

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

1. INTRODUCTION

1.1 General Background

Proteins Proteins are the the main component of physiological metabolic pathways of cells. Proteomics serves a important role in a systems-level understanding of biological systems since it is the large-scale study of proteins, particularly their structures and functions. Mass spectrometry proteomics has become the tool of choice for qualitative and quantitative study of the proteome of an organism. Fundamental challenges in Mass spectrometry based proteomics include (i) Identification of the peptide / proteins that are present in a sample, (ii) Aligning different samples on Mass, elution time, intensity and etc. and (iii) Quantifying the abundance levels of the identified proteins after alignment. We note that protein identification and quantitation are complementary exercises. Unidentified proteins cannot be quantified, and the confidence with which a protein was identified should perhaps be incorporated into that protein abundance estimate. All of these challenges require understanding of the biological and technological perspective as well as the development of novel statistical inference methodology.

The difficulty of protein level identification is generally caused by widespread missingness, peptide degeneracy and misidentification. The limitation of existing alignment algorithms include the lack of automated framework or quantitative statistical assessment. Large-scale missingness due in part to low abundance expression contributes to the complexity of intensity-based protein quantitation. We describe a fully Bayesian hierarchical modeling approach to peptide and protein identification with False Discovery Rate constructed in a unified framework. Across different experiment samples with identified peptide/protein, “Anchor Points” are defined and

This dissertation follows the style of *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

used to align each identified peptide automatically and a p-value is assigned to a set of aligned LC-MS runs to quantify the alignment performance. A statistical method for protein differential expression is outlined on converted presence/absence data based on a simple Binomial likelihood and a hybrid protocol is proposed to combines quantitative and presence / absence analysis and result in a single list of selected proteins with a single associated FDR.

1.2 LC-MS Proteomics

LC-MS refers to liquid-chromatography mass-spectrometry. Liquid chromatography is a technique that could be used in protein differential expression studies by separating peptides into multiple MS scans. This enables higher-resolution analysis of the resulting mass spectra. Mass spectrometry is a tool for measuring mass-to-charge ratios (M/Z) of ions.

The key components of a mass spectrometer are the ion source, mass analyzer, and ion detector (Figure 1.1). The ion source is responsible for assigning charge to each molecule. Mass analyzer measures the mass-to-charge(M/Z) ratio of each ion. The detector captures the ions and measures the intensity of each ion species. In terms of a mass spectrum, the mass analyzer is responsible for the m/z information on the x-axis and the detector is responsible for the peak intensity information on the y-axis. In recent years tremendous improvement in instrument performance and computational tools are used. Several MS methods for interrogating the proteome have been developed: Surface Enhanced Laser Desorption Ionization (SELDI) [1], Matrix Assisted Laser Desorption Ionization (MALDI) [2] coupled with time-of-flight (TOF) or other instruments, and gas chromatography MS (GC-MS) or liquid chromatography MS (LC-MS). Since GC-MS and LC-MS allow for online separation of complex samples and thus they are much more widely used in high-throughput quantitative proteomics.

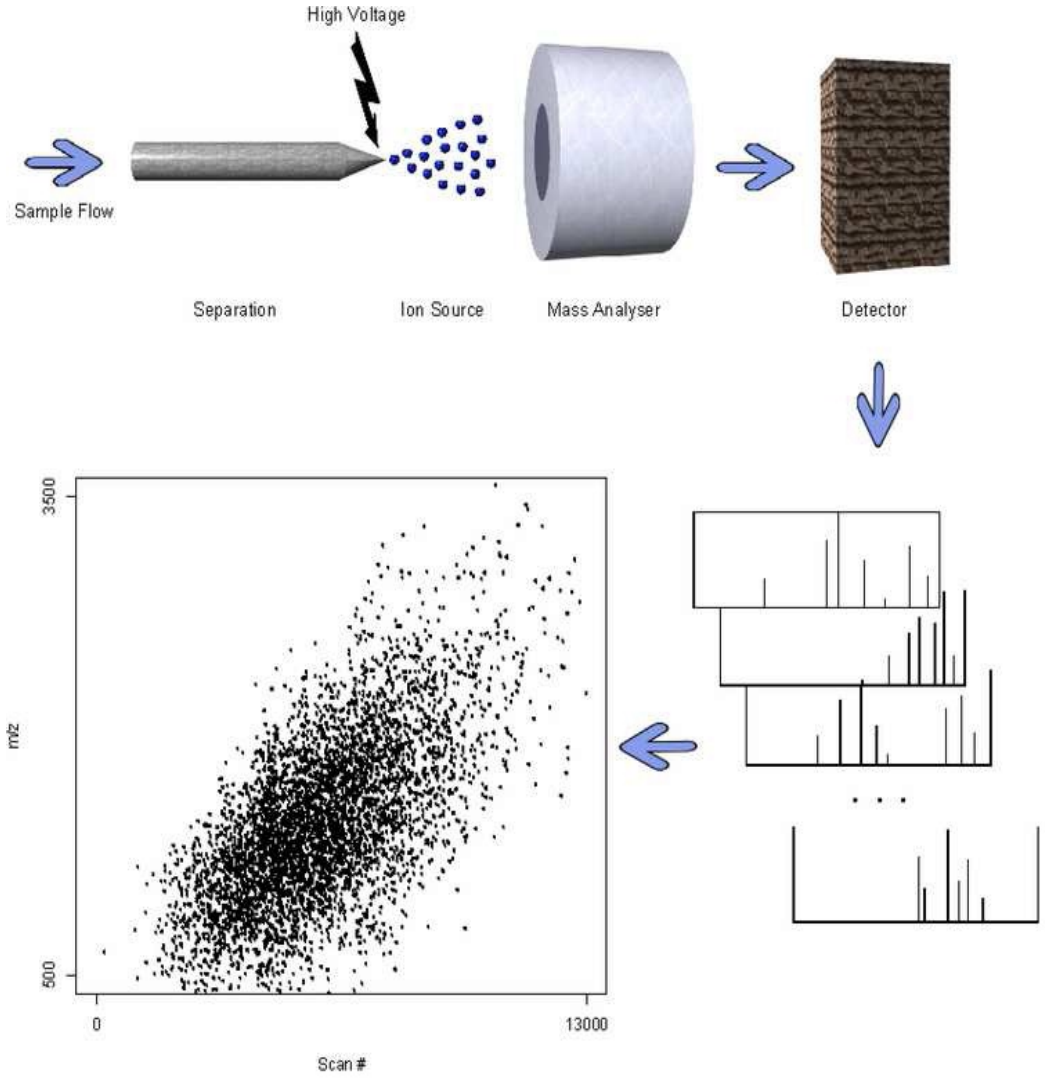

Fig. 1.1. Mass spectrometry. The mass spectrometer consists of an ion source, responsible for ionizing peptides, the mass analyzer and the detector, responsible for recording m/z values and intensities, respectively, for each ion species. Each MS scan results in a mass spectrum, and a single sample may be subjected to thousands of scans.

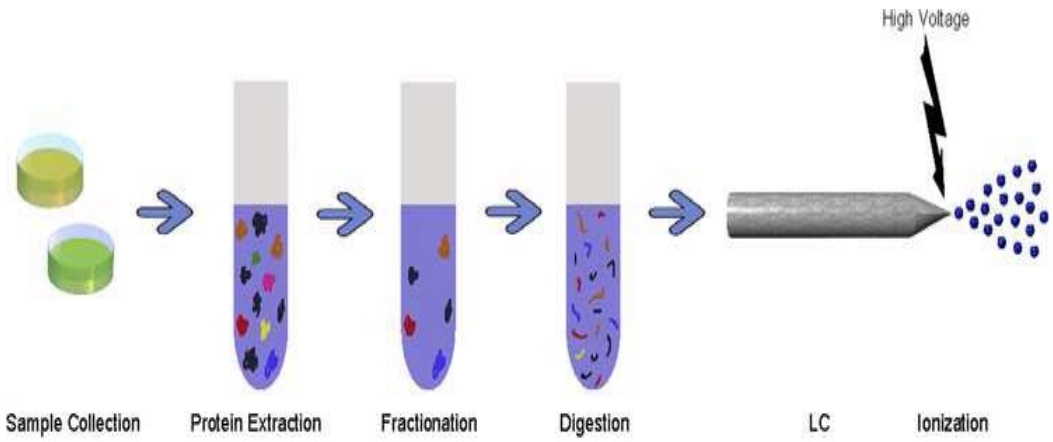

Fig. 1.2. Sample preparation. Complex biological samples are first processed to extract proteins. Proteins are typically fractionated to eliminate high-abundance proteins or other proteins that are not of interest. The remaining proteins are then digested into peptides, which are commonly introduced to a liquid chromatography column for separation. Upon eluting from the LC column, peptides are ionized.

Here we focus on the most widely-used bottom-up approach to quantitative MS-based proteomics, LC-MS, which has become the tool of choice for identifying and quantifying the proteome of an organism. A LC-MS-based proteomic experiment requires several steps of sample preparation (Figure 1.2), including (i) cell lysis to break cells apart and protein extraction, (ii) protein separation to spread out the collection of protein into more homogenous groups, i.e. remove contaminants and proteins that are not of interest, especially high abundance house-keeping proteins that are not usually indicative of the disease being studied, (iii) protein digestion to break intact proteins into more manageable peptide components. Once this is complete, peptides are further separated into a more homogeneous mixture to be ionized and introduced into the mass spectrometer. In tandem mass spectrometry (denoted by MS/MS), several of the most intense (high abundance) peaks from a parent MS (MS1) scan are automatically selected and the corresponding ions are subjected to further fragmentation and scanning. This process is repeated until all candidate peaks of a parent scan are exhausted [3], [4]. This results in a fragmentation pattern for each selected peptide, providing detailed information on the chemical makeup of the peptide.

MS/MS is preceded by LC separation and can more accurately be denoted by LC-MS/MS. High-resolution LC-MS instruments (e.g., FTICR) are very fast and can achieve mass measurements that are sufficiently accurate for identification purposes by comparing the fragmentation patterns to fragmentation spectra in a database, using software. Alignment on the scans is prerequisite for downstream quantitative analysis since day-to-day and run-to-run variation in the complex experimental equipment can create systematic biases. With identified peptide/protein information as well as aligned elution time, m/z and peak intensity, it is possible to proceed to the quantification of the abundance level of those proteins present. Each step contributes to the overall variation in the estimation and inference of the MS based proteomics.

1.3 Peptide/Protein Identification

To facilitate protein identification, proteins are usually separated, cleaved/digested chemically or enzymatically into fragments. Digestion overcomes many of the challenges associated with the complex structural characteristics of proteins, as the resulting peptide fragments are more tractable chemically, and their reduced size, compared to proteins, makes them more amenable to MS analysis.

The first step in protein identification is the identification of the constituent peptides. Multiple distinct peptides can have very similar or identical molecular masses and thus produce a single intense peak in the initial MS (MS1) spectrum, making it difficult to identify the overlapping peptides. The use of separation techniques not only increases the overall dynamic range of measurements (i.e., the range of relative peptide abundances) but also greatly reduces the cases of coincident peptide masses simultaneously introduced into the mass spectrometer.

In tandem mass spectrometry (denoted by MS/MS), a parent ion possibly corresponding to a separated peptide is selected in MS1 for further fragmentation in MS2. Resulting fragmentation spectra are compared to fragmentation spectra in a database, using software like SEQUEST [5], Mascot [6] or X!Tandem [7], see Figure 1.3. PeptideProphet [8] is a widely-used for peptide identification by modeling a collection of database match scores as a mixture of a correct-match distribution and an incorrect-match distribution. The confidence of each match is assessed by its estimated posterior probability of having come from the correct-match distribution, conditional on its observed score. Improvements have been made to PeptideProphet to avoid fixed coefficients in computation of discriminant search score and utilization of only one top scoring peptide assignment per spectrum [9].

Protein identification can be carried out by rolling up peptide-level identification confidence levels to the protein level, a process that is associated with a host of issues and complexities [8]. The goal of the identification process is generally to identify as many proteins as possible, while controlling the number of false identifications at

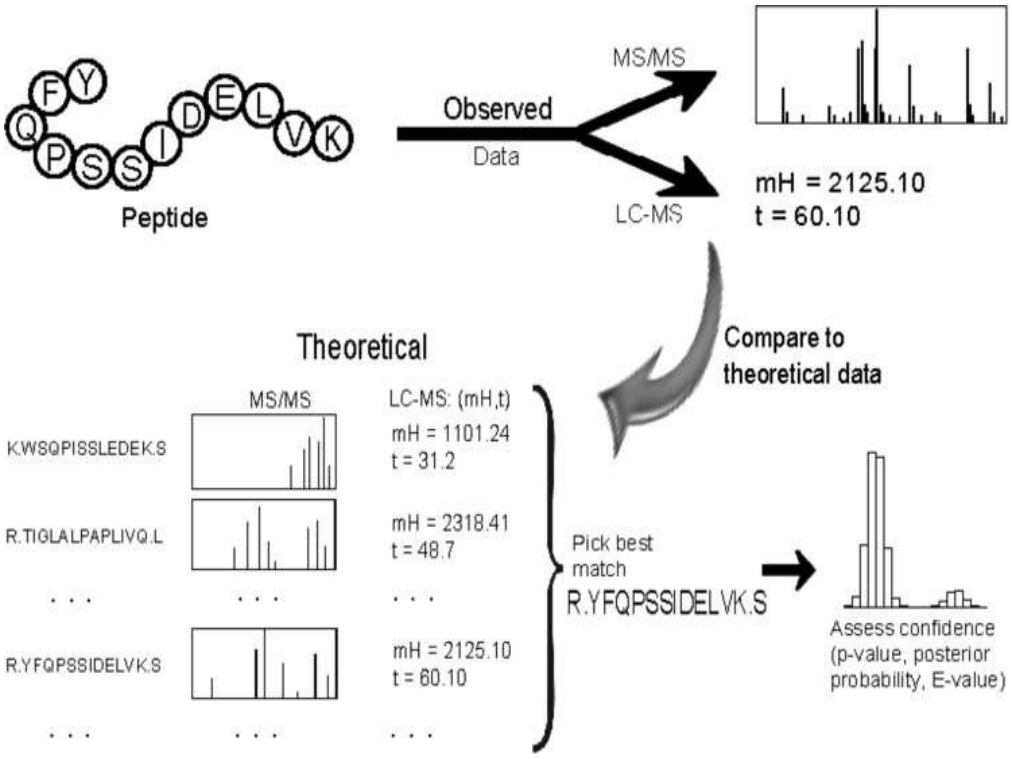

Fig. 1.3. Peptide/protein identification. Peptide and protein identification is most commonly accomplished by matching observed spectral measurements to theoretical or previously-observed measurements in a database. In LC-MS/MS, measurements consist of fragmentation spectra, whereas mass and elution time alone are used in high-resolution LC-MS. Once a best match is found, one of the following methods for assessing confidence in the match is employed: decoy databases, empirical Bayes, or expectation values.

a tolerable level. There are a myriad of options for the exact identification method used, including (i) the choice of a statistic for scoring the similarity between an observed spectral pattern and a database entry [7], [6], and (ii) the choice of how to model the null distribution of the similarity metric [10], [11].

In each of the above approaches, there is a statistical problem of assessing confidence in database matches. This is typically dealt with in one of two ways. The first involves modeling a collection of database match scores as a mixture of a correct-match distribution and an incorrect-match distribution. The confidence of each match is assessed by its estimated posterior probability of having come from the correct-match distribution, conditional on its observed score [12];

The second approach to assessing identification confidence involves the use of a so-called "decoy database. A decoy database is created by scrambling the search database so that any matches to the decoy database can be assumed to be false [13], [12]. The distribution of decoy matches is then used as the null distribution for the observed scores for matches to the search database, and p-values are computed as simple proportions of decoy matches as strong or stronger than the observed matches from the search database. A hybrid approach that combines mixture models with decoy database search can also be used [13].

There are several limitations of current methods. First, many current methods are designed to evaluate the top 1 ranked PSM returned by a database searching application; this discard potential correct match that does not rank the first but among the top several highest match score. Second, recent published work [14], [15], [9] has extended the analysis from the top-ranked peptide per spectrum to a list of candidate PSMs per spectrum with independent assumption which is against the underlying truth that at most one PSM being correct. In Section 2, we describe a fully Bayesian hierarchical modeling approach to peptide and protein identification on the basis of MS/MS fragmentation patterns in a unified framework. Our major contribution is to allow for dependence among the list of top candidate PSMs, which we

accomplish with a Bayesian multiple component mixture model incorporating decoy search results and joint estimation of the accuracy of a list of peptide identifications for each MS/MS fragmentation spectrum. Peptide and protein network structure is modeled in the latent stages of the hierarchical model. We also implement a novel approach to the normalization of database searching scores to scores obtained from decoy databases, which is demonstrated to greatly improve the peptide identification performance. Finally, we propose an objective criteria for the evaluation of the FDR associated with a list of identifications at both peptide level and protein level. Using this criteria, our method is found to result in more accurate FDR estimates than existing methods like Peptide Prophet [8].

1.4 Alignment

Peptides could also be identified on the basis of extremely accurate mass measurements and LC elution times as the output of high-resolution LC-MS instruments. When analyzing two independent samples, peptides elution times are affected by shifts relative to instrumentation effects and it is common to observe systematic differences in the elution times of similar samples on difference columns. However, the LC-MS data have added dimension of m/z and intensity information, which makes it not sufficient to provide alignment for individual peptides by only mapping the retention time coordinates between two LC-MS samples. The goal of alignment is then to match corresponding peptide features in terms of elution time, m/z and peak intensity (see Figure 1.4) from different experiment samples so that the downstream quantitation could be effectively employed.

A time warping method based on raw spectrum for alignment of LC-MS data was introduced by Bylund and others [16], which is a modification of the original correlated optimized warping algorithm [17]. Wang and others, implemented a dynamic time warping algorithm allowing every Retention Time (RT) to be moved. Jaitly, et al. [18] introduced a non-linear alignment technique that uses a dynamic time

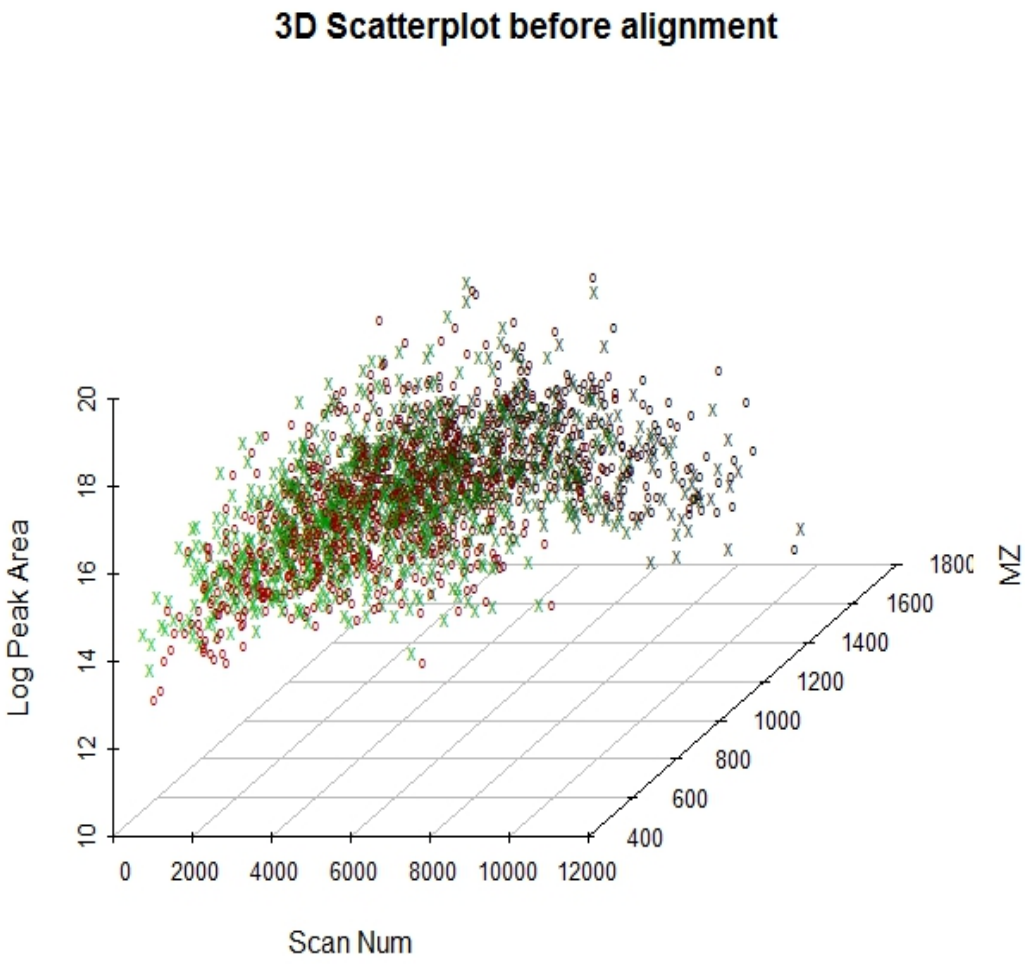

Fig. 1.4. Two sample of scans before alignment, red dots represent sample 1 and green dots represent sample 2. X axis is ScanNum (equivalent to elution time), Y axis is M/Z and Z axis is Log (Peak intensity).

warping approach at a feature level. Radulovic and others [19] performed alignment based on (m/z,RT) values of detected features by dividing the m/z domain into several intervals and fitting different piece-wise linear time warping functions for each m/z interval and then applying a "wobble" function to peaks and allow peaks to move. Palmblad, et al. [20] applied a genetic algorithm to establish an alignment warping function by using peptide elution times from MASCOT output to define anchor points between two datasets.

All of these alignment techniques still suffer from either not taking m/z and peak intensity information into account or manually inappropriate division of m/z domain or incorrect parameterization of warping function as well as lacking a useful metric for scoring an alignment between two datasets. To score alignments, a ground truth is required to assess the accuracy of an alignment by establishing links between datasets via database searches to find the same peptide present in two datastes. (Simply matching two features based on mass and elution time alone is not very supportive).

Our approach in Section 3 uses "Anchor points" found between two samples to align all the individual scan in the second sample and provides a framework to quantitify the alignment, that is, assigning a p-value to a set of aligned LC-MS runs to assess correctness of alignment. In our method, for different experiments, we have the elution time, mass over charge (m/z) value, peak intensities, peptide information for each of the thousands of scans in the SEQUEST search output. A feature is treated as an "anchor point" if it corresponds to very high confidence identification to the same peptide in all samples (in addition to meeting other quality standards). The anchors can be relied upon with very high confidence as being paired across samples. As such, they can be used as the basis of an alignment algorithm, as well as for assessing the performance of an alignment algorithm.

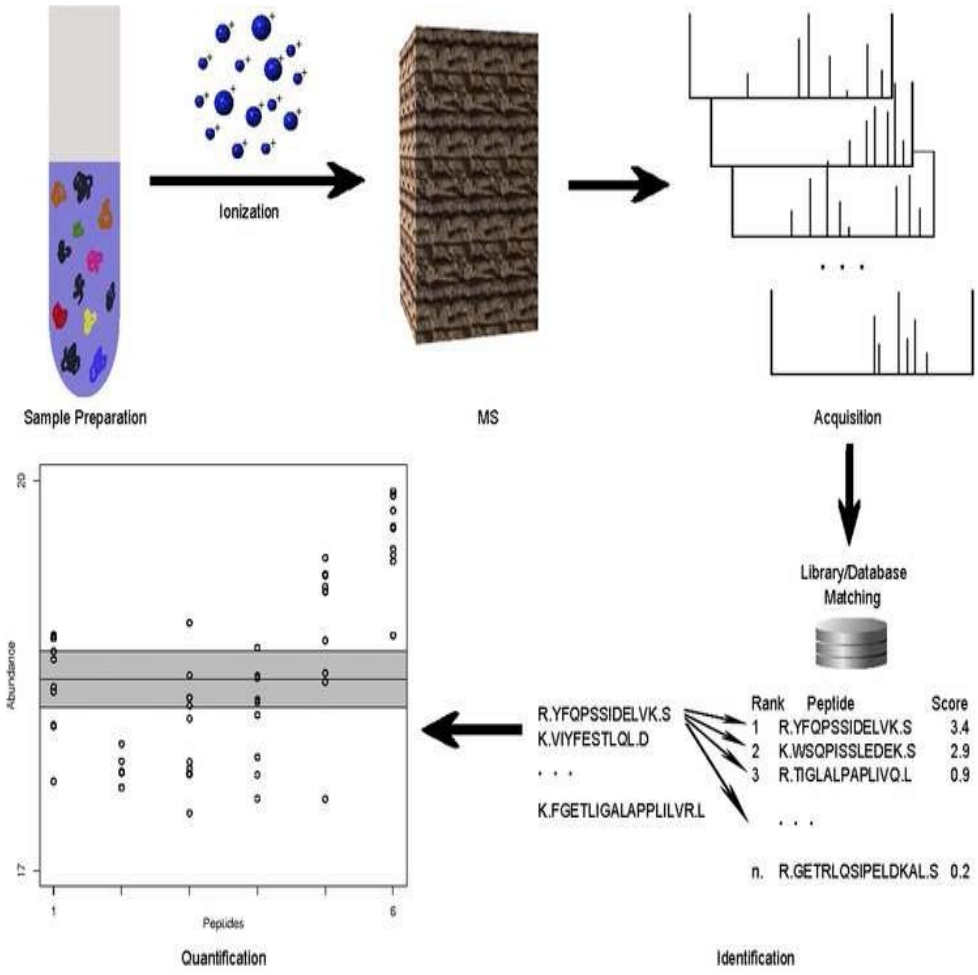

Fig. 1.5. Overview of LC-MS-based proteomics. Proteins are extracted from biological sam- ples, then digested and ionized prior to introduction to the mass spectrometer. Each MS scan results in a mass spectrum, measuring m/z values and peak intensities. Based on observed spectral information, database searching is typically employed to identify the pep- tides most likely responsible for high-abundance peaks. Finally, peptide information is rolled up to the protein level, and protein abundance is quantified using either peak intensities or spectral counts

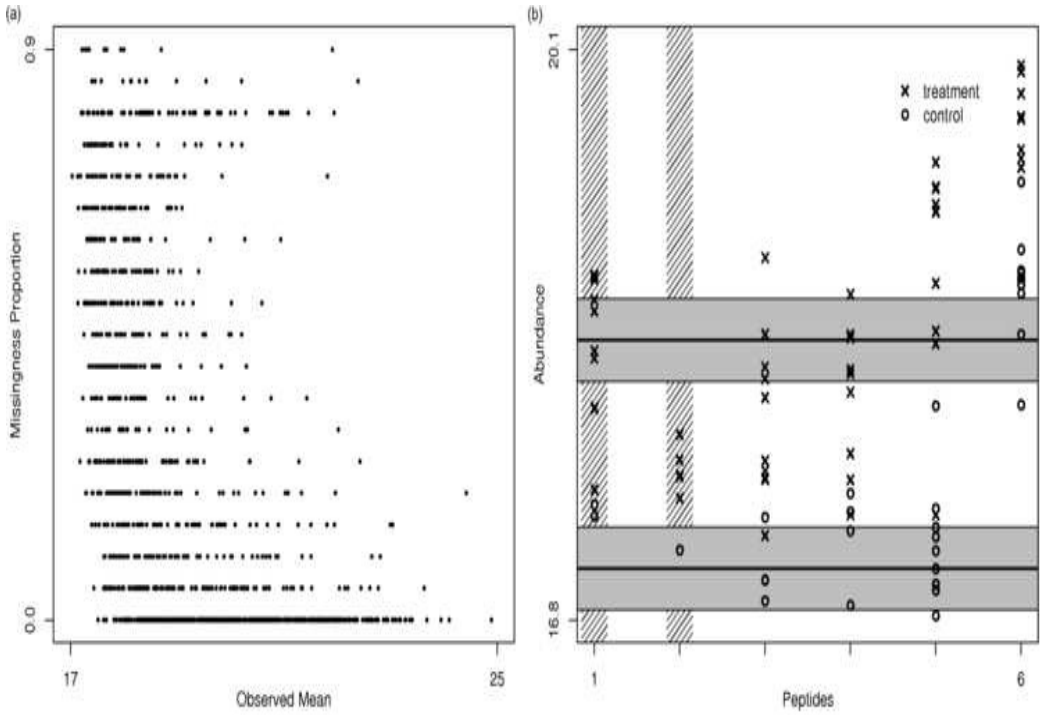

Fig. 1.6. Protein quantitation. The left panel shows the proportion of missing values in an example data set as a function of the mean of the observed intensities for each pep- tide. There is a strong inverse relationship between these, suggesting that many missing intensities have been censored. The right panel shows an example protein found to be differentially expressed in a two-class human study. The protein had 6 peptides that were identified, although two were filtered out due to too many missing values (peptides 1 and 2, as indicated by the vertical shaded lines). Estimated protein abundances and confidence intervals are constructed from the peptide-level intensities by a censored likelihood model [21].

1.5 Protein Quantitation

Quantitative proteomics is concerned with quantifying and comparing protein abundances in different conditions (Figure 1.5). Once a list has been constructed of the proteins believed to be present in the sample and the peak intensity is aligned, the next task is to quantify the abundance of the proteins. Protein abundance information is contained in the set of peaks that correspond to the protein's component peptides. Peak height or area is a function of the number of ions detected for a particular peptide, and is related to peptide abundance [22]. Regardless of the specific technology used to quantify peptide abundances, statistical models are required to roll peptide level abundance estimates up to the protein level. Intensity based procedures for differential protein expression are naturally constructed in the context of regression or ANOVA, or as a "rollup" problem [23].

However, intensity-based procedures are challenged by the presence of widespread missing intensities, which are prevalent in MS-based proteomic data. In fact, it is common to have 20% -40% of all attempted intensity measures missing. Abundance measurements are missed if, for example, a peptide was identified in some samples but not in others, see Figure 1.6. This can happen partially due to the low abundances of present peptides, which is essentially a censoring mechanism [24]. With standard regression or ANOVA procedures, peptides with missing values must either be removed from the analysis, or their missing values must be imputed. There will typically be very few peptides with no missing values, so filtering peptides in this way results in a much less informative data set. The simple imputation routines are not appropriate [25] since the vast majority of missing values are the results of censoring of absent or low-abundance peptides. This complicates intensity-based quantitation, as simple solutions will tend to be biased. For example, analysis of only the observed intensities will tend to overestimate abundances and underestimate variances. Simple imputation routines like row-means or k-nearest-neighbors suffer from similar limitations. Parametric imputation and other specialized method-

ology can be employed to enable intensity-based inference with lessened information loss [21]. However, some information loss is inevitable. In particular, "one-state" (or nearly so) peptides, those for which there are many observed intensities in one comparison group but few in another comparison group, are of great biological interest but not amenable to an intensity-based analysis and filtered out in intensity-based analysis. Statistical models are needed to address these issues, as well as to handle the peptide-to-protein rollup [21], [26], see Figure 1.5.

In Section 4, we propose a "presence / absence" analysis, in which peak intensities are digitized into binary measurements depending on whether a peak was observed or not. Data collected in our laboratory does not necessarily have MS/MS fragmentation data associated with it, instead being obtained according to the Accurate Mass and Time (AMT) tag pipeline [27]. We also present a hybrid analysis protocol that consists of two stages: (i) intensity-based analysis, and (ii) a presence / absence analysis. The results of each are merged to create a single collection of "interesting" proteins, to which we use novel methodology to apply a single FDR. For the proposed hybrid analysis protocol, we demonstrate the following: (i) Resulting FDR estimates are conservative, (ii) One-state proteins are consistently selected as differentially expressed, and (iii) The number of differentially expressed proteins selected at a specified FDR exceeds that either intensity-based or presence / absence analysis alone.

2. A BAYESIAN HIERARCHICAL MODEL FOR PEPTIDE / PROTEIN IDENTIFICATION BY LC-MS/MS.

2.1 Introduction

A fundamental challenge in quantitative mass spectrometry (MS)-based proteomics is the identification of peptides and proteins that are present in a sample. This is typically carried out by comparing observed features to entries in a database of theoretical or previously-identified peptides (figure on page 7). In tandem mass spectrometry (denoted by MS/MS or MS^n), fragmentation spectra are obtained for each subset of observed high-intensity peaks and compared to fragmentation spectra in a database, using software like SEQUEST [5], Mascot [6], or X!Tandem [7]. Alternatively, high-resolution MS instruments can be used to obtain extremely accurate mass and time (AMT) measurements, and these can be compared to AMT measurements in a database [28]. In either case, a statistical assessment of the level of confidence for each identification is desired; for the purposes of this section, we focus on peptide-spectrum matches (PSMs) and MS/MS. Protein identification can be carried out by rolling up peptide-level identification confidence levels to the protein level [8], or by the simultaneous modeling of peptides and proteins using hierarchical models [29]. The goal of the identification process is generally to identify as many features as possible, while controlling the number of false identifications at a tolerable level. There are a myriad of options for the exact identification method used, including (i) The choice of a statistic for scoring the similarity between an observed spectral pattern and a database entry, and (ii) The choice of how to model the null distribution of the similarity metric [10], [11].

There are several limitations of current methods. First, many current methods are designed to evaluate the top 1 ranked PSM returned by a database searching

application. Recent published work [14], [15], [9] have extended the analysis from the top 1 ranked peptide per spectrum to a list of candidate PSMs per spectrum. But existing approaches have made the assumption that multiple PSMs in the list of candidate matches are independent. In complex samples, it is possible that multiple peptides may give rise to very similar fragmentation patterns. However, it is reasonable to expect that there is only one correct database entry that matches a specific fragmentation pattern, in which case the PSMs in a list of top candidate matches will not be independent. Our experiments suggest that the independence assumption may lead to underestimated false discovery rates among identified peptides, see figure on p. 36 in section 2.3.

In this section, we describe a fully Bayesian hierarchical modeling approach to peptide and protein identification on the basis of MS/MS fragmentation patterns in a unified framework. Our major contribution is to allow for dependence among the list of top candidate PSMs, which we accomplish with a Bayesian multiple component mixture model incorporating decoy search results and joint estimation of the accuracy of a list of peptide identifications for each MS/MS fragmentation spectrum. Peptide and protein network structure is modeled in the latent stages of the hierarchical model. Our model can incorporate arbitrary collections of discriminant features for quantifying match quality; examples include scores from different search applications like XCorr and Sp from SEQUEST, hyperscore and E-value from X!tandem, and other auxiliary discriminant information. We also implement a novel approach to the normalization of database searching scores by utilizing scores obtained from decoy databases, which is demonstrated to greatly reduce the dependency among discriminant features, see figure on page 21. Finally, we propose an objective criteria for the evaluation of the FDR associated with a list of identifications at peptide level. Using this criteria, our method is found to result in more accurate FDR estimates than existing methods like PeptideProphet [11].

2.2 Methods

2.2.1 Experiments

The data we used came from 5 quality-control LC-MS/MS runs of Shewanella oneidensis, prepared by and run at the Pacific Northwest National Laboratory. For each sample, we have SEQUEST output for use in the peptide and protein identification process.

2.2.2 Peptide and Protein Identification by Database Search

The identification of peptide assignments to MS/MS spectra is primarily based on database search scores computed by different search engines together with various peptide-specific properties. Most database search approaches employ a score function to measure the similarity between peptide MS/MS spectra and theoretical spectra constructed for each peptide in the searched protein sequence database. Different search engines such as SEQUEST [5], X!tandem [7] and Mascot [6] adopt different scoring systems. For example, SEQUEST computes a correlation score between a normalized MS/MS spectrum and a unit-intensity fragmentation model and corrects it by an estimation of the background. X! Tandem defines a score function based on the shared peak count approach and calculates an "expectation value" for each peptide assignment. Mascot computes a probability-based score called the ion score (often referred to simply as Mascot score) using the Mowse scoring algorithm. Other than database search scores, some additional measurements might also contain useful discrimination information, such as the difference (dT) between the observed and predicted elution time (such as the "Normalized Elution Time" (NET) [30]), the difference between the measured and calculated peptide mass (dM), the fractional difference between current and second best Xcorr δ_{C_n}, the Number of Tryptic Ends (NTE), the Number of Missed Cleavage Sites (NMC), the peptide degeneracy

measurement "multi-protein". All these features are potentially informative for distinguishing between correct and incorrect PSMs. Ideally, incorporating all available discriminant features should greatly improve tandem mass spectrum identification.

In this section, we consider a target-decoy search strategy, which has been used successfully in peptide and protein identification analysis [10]. Decoy databases are usually created by reversing or randomly shuffling the target peptide sequences. The distributions of some PSM attributes (e.g., peptide length, elution time, charge state, database search score) from a decoy database search are assumed to be the same as those of false identifications from target database search. And incorrect PSMs from decoy sequences are assumed to be equally as likely as those from target sequences. Based on these assumptions, target-decoy search strategies have been successfully used to distinguish correct identifications from incorrect ones [14], [31] and estimate confidence levels for peptide assignments [32].

In some cases, database search scores should be transformed prior to analysis. For example, consider the SEQUEST primary score, XCorr, a measure of the correlation between observed and theoretical fragmentation spectra. XCorr is highly dependent on peptide length and precursor ion charge state (see, e.g., [11] and [33]). In particular, long peptides tend to have higher XCorr values than short peptides, due simply to more frequent random matching of observed and theoretical spectral features. Precursor ions with different charge states also have different probabilities of having random hits, yielding shifts in the distributions of XCorr scores. To alleviate this problem, PeptideProphet normalizes XCorr based on a deterministic transformation, which is a function of peptide length. The transformed XCorr is denoted as XCorr$'$. Due to the existence of charge state effects in the transformed XCorr, PeptideProphet models each charge state separately.

We propose a novel approach to the normalization of discriminant scores, utilizing the scores observed for matching to the decoy database. As an example, assuming the top R ranked decoy XCorr scores for each spectrum are available, we normal-

ize the target XCorr scores by subtracting the average XCorr scores for the top R ranked decoy PSMs. In what follows, we let XCorr* denote the decoy-normalized version of XCorr. The underlying assumption of this transformation is that peptide length dependence and charge state dependence in decoy XCorr scores can be used to approximate the corresponding dependencies in target XCorr scores.

This transformation is easy to implement yet effective to reduce the sequence length dependence and charge state dependence. Figure 2.1 clearly suggests that the distribution of the normalized XCorr score shows much weaker dependence on peptide length and charge state. In addition, XCorr* is by nature a measurement of relative XCorr score defined for the top R ranked target PSMs. XCorr* alone offers comparable discrimination information with both XCorr$^{'}$ and δ_{C_n}. Because of these desired properties, we employ XCorr* as a primary SEQUEST searching score in the subsequent analysis.

When multiple (transformed) discriminant features are available, one common strategy is to combine (a subset of) them into a single discriminant score for simplification purposes, which can greatly relieve the complexity of the subsequent mixture distribution specification. There are many dimension reduction and classification tools to choose from, including linear discriminant analysis (LDA), principle component analysis (PCA), logistic regression and support vector machines (SVM). For example, PeptideProphet employs LDA to derive a scaler score "fval" from a number of database search scores. All other information together with "fval" is modeled in an unsupervised or semi-supervised fashion at subsequent steps [11], [31], [13]. Percolator uses a semi-supervised machine learning method that iteratively trains a SVM classifier containing all discriminant features [14], where each PSM is assigned a decision score.

In this section, we employ a similar approach as Percolator, using a radial basis function (RBF) as the kernel function to train a semi-supervised SVM in a dynamic fashion. The top 1 ranked target and all decoy discriminant features are used as

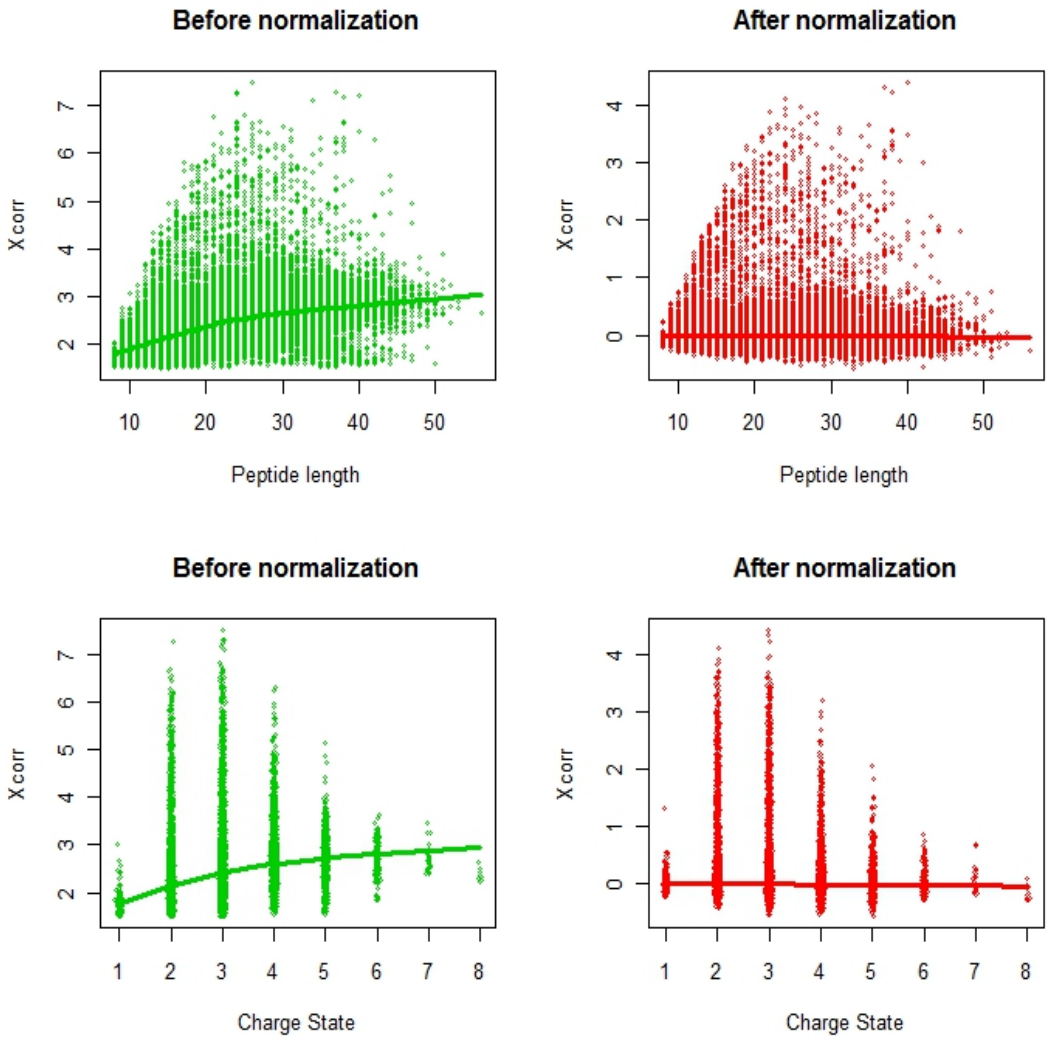

Fig. 2.1. Scatter plot of normalized Xcorr vs peptide length and charge state. The top left panel (green) is the scatter plot of Xcorr vs Peptide length before normalization and reveals a positive correlation by the fitted lowess curve. The top right panel (red) is the scatter plot of Xcorr vs Peptide length after normalization. The lowess curve fitted is essentially flat, indicating a much weaker dependency on peptide length. The bottom left panel (green) is the scatter plot of Xcorr vs charge state before normalization and reveals a positive correlation by the fitted lowess curve. The bottom right panel (red) is the scatter plot of Xcorr vs charge state after normalization with the fitted lowess curve relatively flat, indicating a much weaker dependency on charge state after normalization.

the training dataset, and the combined discriminant score is then calculated for all of the top R target PSMs using the SVM function derived by the training dataset. Figure 2.2 shows the distributions of target and decoy SVM scores. The bi-modality of the distribution of target scores clearly suggests that target PSMs are comprised of a mixture of correct and incorrect PSMs, suggesting the use of mixture modeling. Also, the distribution of scores assigned to decoy PSMs is very similar to the distribution for incorrect target scores. We note that SVM is primarily used as for scoring purposes, rather than as an ultimate classifier or validation tool in the identification process. There are many other scoring approaches for classification tools that might achieve better combined discriminant scores. It is also possible to apply SVM to database search scores only and then specify a multivariate mixture distribution for the database search score combined with other discriminant information. However, it is beyond the scope of this section to discuss and compare all of these possibilities.

2.2.3 Model

Many existing approaches to peptide identification are designed to model the top 1 scoring PSM for each spectrum only (ranked according to a certain scoring criterion). The combined discriminant score (or multiple discriminant scores) can be modeled as a parametric or semi-parametric mixture distribution with two components representing correct and incorrect identifications, respectively. In a Bayesian framework, the posterior probability that a PSM belongs to the correct-match distribution can be used as a statistical measure of confidence.

In this section, we present a model that accommodates a list of potential matches (PM) for each spectrum. We might, for instance, wish to use the top 10 ranked PSMs returned by a particular searching application, or a list of unique top 1 ranked PSMs based on multiple search algorithms. In practice, it is expected that many correct peptide assignments are ranked slightly lower than the top 1 ranked PSM based on an algorithm-derived score. Our goal is to take advantage of the information for all

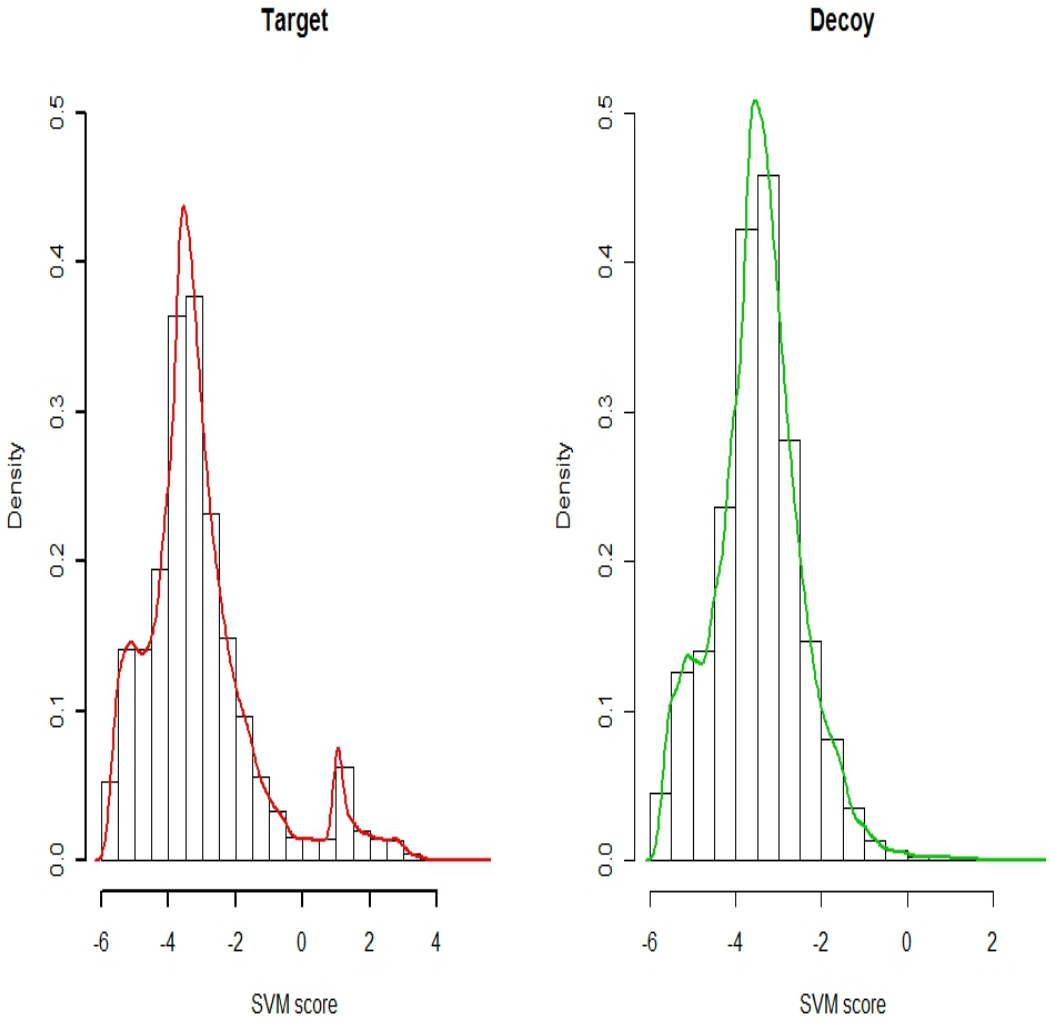

Fig. 2.2. SVM scores. The histograms and density curves of target and decoy SVM scores. The left panel with green curves is the distribution of the decoy SVM score and the right panel with red curves is the distribution of the target SVM score. The decoy histogram and density curve have similar shape to the incorrect SVM score in target PSMs.

available discriminant features and peptide protein grouping information to bump correct PSMs up the ranking list. This would enable us to discover more correct PSMs than approaches based on single best PSM scores alone.

In this approach, database search algorithms such as SEQUEST are employed to filter out a list of candidate PSMs and calculate their corresponding discriminant features. The subsequent statistical analysis serves as a second search step to find the most likely match within the candidate list based on their estimated confidence levels. Although several existing approaches also consider multiple peptide assignments per spectrum [14] and [15], they typically impose independence assumption in the correctness for PSMs in the same candidate list. However, this assumption may not hold in practice and could render higher false discovery rate. For example, given that one candidate PSM has a significantly higher chance to be correct, the probability that other candidates are also correct should be low.

We relax the independence assumption in our approach. A multiple component mixture model is proposed in the first stage to jointly infer the correctness of every candidate PSM for each spectrum based on the combined discriminant score. Since the prior probabilities of being correct matches are connected to the unobserved presence/absence of peptides and proteins. We employ a Bayesian hierarchical modeling approach that models peptide and protein network information in the latent layers. Some common complexities in peptide/protein identifications are also addressed in our model.

The First Stage: a Multi-component Mixture Model for Discriminant Scores

We introduce some notations first. Assume we have K experimental spectra, each of which is assigned a PM list with R_k PSMs, e.g., the top R ranked PSMs from SEQUEST. Let s_k^r denote the combined discriminant score of the rth match for spectrum k, $r = 1, \cdots, R_k$, $k = 1, ..., K$. Let $s_{k,1:R_k} = (s_{k,1}, s_{k,2}, ..., s_{k,R_k})$ denote a vector

containing scores for those R matches, $s_{1:R_k \backslash r} = (s_{k,1}, s_{k,2}, ..., s_{r-1}, s_{r+1}, ..., s_{k,R_k})$, denote a vector containing scores for all candidates except for the rth match.

We introduce a $R_k + 1$ dimensional component label vector Z_k, where the rth element of Z_k, $Z_{k,r} = (Z_k)_r$, is a binary indicator. We assume there is at most one correct PSM in the PM list for each spectrum. $Z_{k,R_k+1} = 1$ indicates none of the R_k matches in the PM list being correct. For r from 1 to R_k, $Z_{k,r} = 1$ indicates the rth PSM is the only correct match in the PM list for spectrum k. Define $p_k = (p_{k,1}, ..., p_{k,R_k+1})'$, where $p_{k,r} = Pr(Z_{k,r} = 1)$. Then $\sum_{r=1}^{R_k+1} p_{k,r} = 1$. The imposition of this restriction connects the probabilities of being correct for multiple PSMs assigned to the same spectrum. It enables us to compute the peptide probabilities for one candidates while taking into account information from other matches in the same PM list.

Let f_1 denote the density function for score with correct-match and f_0 denote the density function for score with incorrect-match. Conditional on the label vector Z_k, discriminant scores are assumed to be independently following the distribution shown below,

$$Pr(s_k|Z_k) = \prod_{r=1}^{R_k} f_1(s_{k,r})^{I(Z_{k,r}=1)} f_0(s_{k,r})^{I(Z_{k,r}=0)}$$

Integrating out the latent label vector Z, it yields a mixture distribution with R_k+1 components for target discriminant scores.

$$\begin{aligned} s_{k,1:R_k}^{(target)} &\sim \sum_{r=1}^{R_k} I(Z_{k,r} = 1) f_1(s_{k,r}^{(target)}) f_0(s_{k,\{1:R_k\}\backslash r}^{(target)}) + I(Z_{k,R_k+1} = 1) f_0(s_{k,1:R_k}^{(target)}) \\ &\sim \sum_{r=1}^{R_k} p_{k,r} f_1(s_{k,r}^{(target)}) f_0(s_{k,1:R_k\backslash r}^{(target)}) + p_{k,0} f_0(s_{k,\{1:R_k\}}^{(target)}) \end{aligned}$$

Recall that distribution of decoy scores provides a satisfying approximation to the distribution of scores from incorrect target matches. Incorporating decoy scores

in the model could obtain improved robustness in the estimation of f_0 and hence better discrimination for target PSMs. The model is:

$$s_{k,1:R}{}^{(incorrect)} \quad \sim \quad f_0(s_{k,1:R}^{(decoy)}) \tag{2.1}$$

The choice of functional forms for f_0 and f_1 will depend on the specific method used to derive the combined database search score. In our case, we used Xcorr, Rank S_p, NTE, "multi-protein" and charge state as the covariates for the combined SVM score, which greatly simplifies the task of mixture distribution specification. We model f_0 using the generalized extreme value (GEV) distribution with location parameter μ_0, scale parameter σ_0 and shape parameter ξ_0. The GEV distribution is the limit distribution of properly normalized maxima of a sequence of independent and identically distributed random variables. It is commonly used as an approximation to model the maxima of long (finite) sequences of random variables. In our case, the combined scores for PSMs in the PM list are among the highest of the entire database, making GEV distribution a natural choice for f_0. The diagnostic analysis also showed that the GEV distribution can describe the shape of combined discriminant score well, see Figure 2.3. Similarly, empirical exploration study suggests using a GEV density to model f_1 rather than a normal density, with location paramter μ_1, scale parameter σ_1 and shape parameter σ_0 to be estimated, see Figure 2.4.

Latent Stages: Prior Models for Peptide and Protein Network

In tandem MS/MS dataset, peptide level and protein level information are naturally connected in a hierarchical structure. Figure on page 29 briefly illustrates the work flow of the LC-MS/MS experiment and the peptide and protein network information. As is shown in this figure, each protein can generate a list of peptides and some peptides can be generated by multiple proteins. If peptide j is correctly identified and unique to protein i, then protein i must be present in the sample,

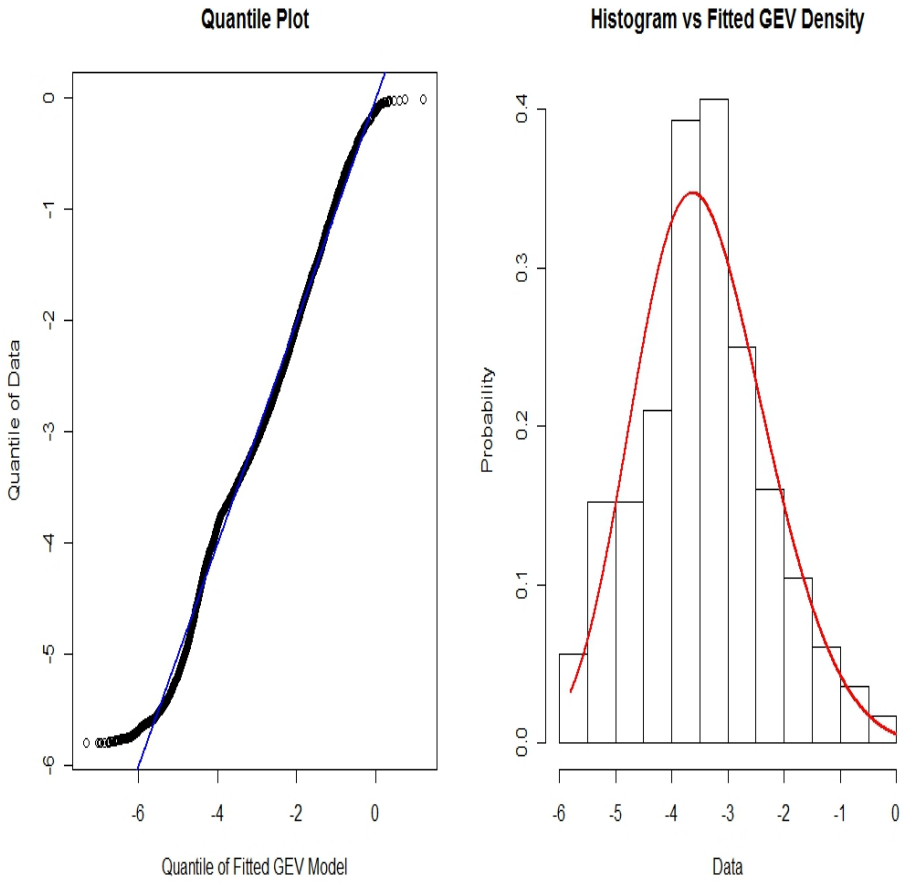

Fig. 2.3. Diagnostic of GEV fit on f_0, the density of incorrect matching scores. Left panel is quantile plot, the blue line is diagonal line. Right panel is density curve vs histogram.

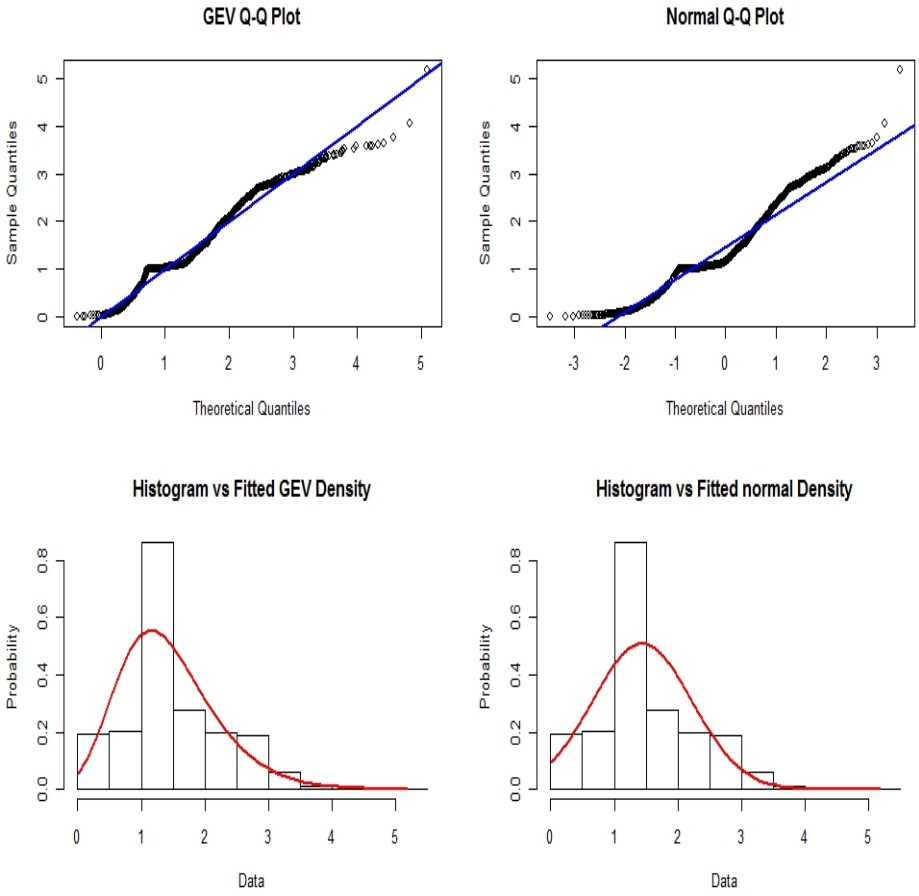

Fig. 2.4. Diagnostic of GEV fit on f_1, the density of correct matching scores. Upper Left panel is GEV quantile plot, the blue line is diagonal line. Upper right panel is Normal quantile plot, the blue line is fitted QQ-line. Bottom left panel is GEV density curve vs histogram and bottom right panel is normal density curve vs histogram

which in turn implies higher chance of detecting the sibling peptides of peptide j in the experiment.

Many conventional approaches follow a two step procedure to assess peptide and protein confidence levels, in which peptide confidence levels are obtained from a likelihood model for discriminant score(s) in the first step and protein confidence levels are estimated based on the peptide-level results. Apparently, by borrowing strength from peptide/protein grouping information, an integrated analysis can produce more accurate assessment for peptide and protein identification [29].

There are two major complexities that need to be addressed in modeling peptide and protein network. The first complexity is called "degeneracy" problem [11]. If a peptide can be generated by multiple proteins, then it adds ambiguity into protein identification since we only know at least one of these proteins must be present. Few work has been done to address this problem. Another common complexity is that multiple spectra with different scores can be assigned to the same peptide, again, adding ambiguity in assessing confidence levels for peptides with multiple spectra. This complexity primarily arise from two different situations: i) tandem MS/MS technique can generate repeated fragment ion spectra [34], which are mostly likely to be assigned to the same peptide; ii) a false identified peptide can be assigned to multiple spectra due to random matches. Many conventional approaches only keep the maximum score for analysis, leading to two potential limits. First, it ignores information from other spectra that may help with peptide identification. For example, if multiple PSMs associated with the same peptide are due to repeated spectra, larger number of repeated spectra should imply higher peptide confidence level. Second, if multiple PSMs associated with the same peptide are due to random matches, ignoring information from lower scores can possibly yield overestimated peptide confidence level since a false identified peptide can be assigned with a high score by chance. Apparently, modeling multiple PSMs to the same peptide would require distinguishing those two situations. Frank etal. [34] propose a clustering approach

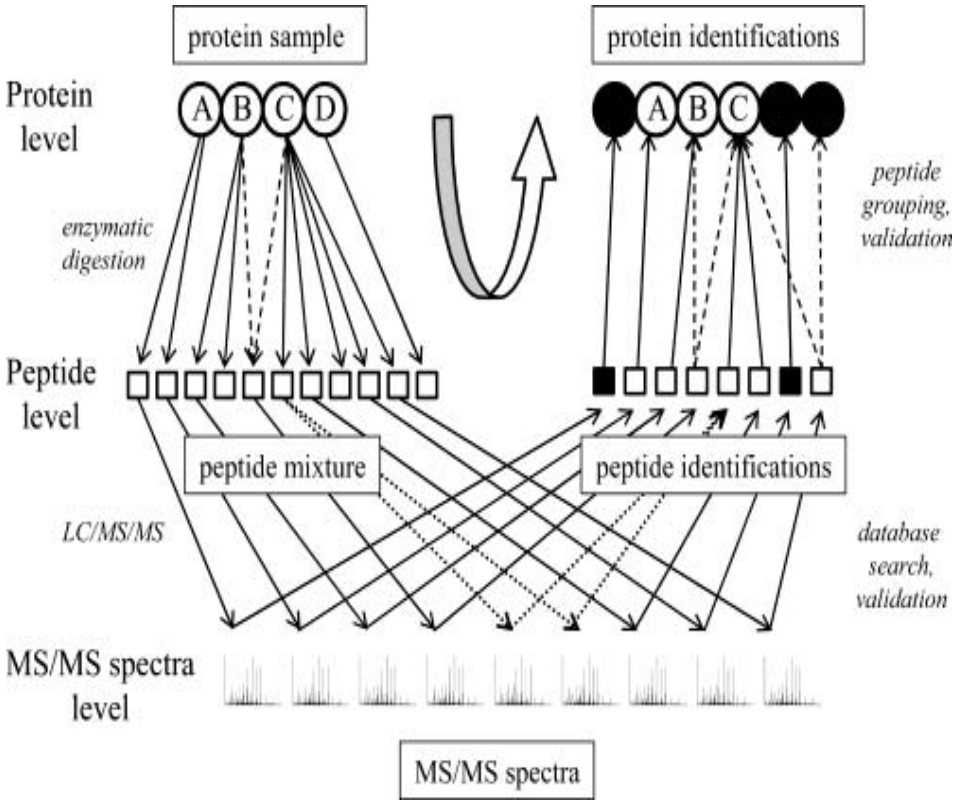

Fig. 2.5. Simplified outline of the experimental steps and work flow of the data in a typical high-throughput MS-based analysis of complex protein mixtures. Each sample protein (open circle) is cleaved into smaller peptides (open squares), which can be unique to that protein or shared with other sample proteins (indicated by dashed arrows). Peptides are then ionized and selected ions fragmented to produce MS/MS spectra. Some peptides are selected for fragmentation multiple times (dotted arrows) while some are not selected even once. Each acquired MS/MS spectrum is searched against a sequence database and assigned a best matching peptide, which may be correct (open square) or incorrect (black square). Database search results are then manually or statistically validated. The list of identified peptides is used to infer which proteins are present in the original sample (open circles) and which are false identifications (black circles) corresponding to incorrect peptide assignments. The process of inferring protein identities is complicated by the presence of degenerate peptides corresponding to more than a single entry in the protein sequence database (dashed arrows) [8]

to identify redundant spectra and replace each cluster with a single representative spectrum. We believe this clustering type of approaches followed by database search applications can greatly reduce the number of repeated spectra. However, it may not be able to find all redundant spectra.

Bayesian hierarchical model is a convenient choice that allows us to incorporate different levels of information in a unified framework. We describe models with multiple latent layers to describe peptide and protein network. Models at latent stage is connected to the mixture model in the first stage through the prior model for $p_{Z|Y}$. In this section, we describe a model based approach to address the two major complexities discussed above.

We first introduce peptide level indicator and protein level indicator. Let X_i denote a binary indicator such that $X_i = 1$ if protein i is present in the sample and detected, and 0 otherwise. Let Y_j be a binary indicator such that $Y_j = 1$ if peptide j is present in the digested sample and detected, and 0 otherwise. Let $pep(k, r)$ denote the peptide index number of the rth PSM for spectrum k. Recall Z is a $R_k + 1$ dimensional component label vector Z_k, where the rth element of Z_k indicating whether the rth match is correct for $r \leq R_k$. Conditional on peptide indicator Y, PSM indicators Z and protein indicators X are independent. Based on peptide and protein network information, we build three models in the latent stages, $Pr(Z|Y)$, $Pr(Y|X)$ and $Pr(X)$ respectively.

The Second Stage: $\mathbf{Pr(Z|Y)}$

Conditional on a peptide is absent, it is reasonable to assume that all PSMs that are matched to this peptide are incorrect. If a peptide is correctly detected and only one PSM is matched to it, we assume that PSM is correct. If a peptide is correctly detected but have multiple hits, it is possible that some hits are due to

random matching. Those assumptions become the major guideline for us to model the conditional probability, $Pr(Z|Y)$.

$$Pr(Z|Y) = Pr(Z_{k,r} = 1|Y_{pep(k,r)}) = \begin{array}{ll} \tau_{k,r} & Y_{pep(k,r)=1} \\ 0 & Y_{pep(k,r)=0} \end{array} \tag{2.2}$$

where $\tau_{k,r}$ is the conditional probability that the rth match for spectrum k is correct given peptide $pep(k,r)$ is present. As we discussed above, if multiple PSMs assigned to the same peptide originate from repeated spectra, we expect their conditional probabilities of being correct are consistently close to 1. If PSMs are assigned to the same peptide by random chance, we expect their conditional probabilities of being correct take a much smaller value than 1. We use DM, the difference between observed Mass and theoretical Mass as a major covariate to distinguish between repeated spectra and random matches. When a PSM is unique to a peptide, DM = 0. We expect its conditional probability to be 1 or close to 1. In other cases, we expect that the larger the absolute deviation is, the more likely the PSM is assigned due to random matching. A logistic regression or Probit model using DM as covariate information is desired to model the conditional probability $\tau_{k,r}$. For simplicity, we pick a threshold C for $\mathrm{DM}_{k,r}$ and assume a prior model for $\tau_{k,r}$ as follows:

$$\tau_{k,r} = \begin{array}{ll} 1 & |\mathrm{DM}_{k,r}| \leq C \\ \tau & |\mathrm{DM}_{k,r}| > C \end{array} \tag{2.3}$$

Based on exploratory data analysis on previously observed NET combined with expertise's suggestion, we choose $C = 0.5$ as a threshold. τ is an unknown parameter. We assign a prior distribution on this parameter and estimate it using MCMC approach (see Appendix for details).

The Third Stage: $\mathbf{Pr}(\mathbf{Y}|\mathbf{X})$

If a protein is absent, it is reasonable to assume that none of its constituent peptides can be correctly detected. If a protein is present, it is common that only

a subset of its constituent peptides can be correctly identified. In the case of 'degeneracy', a peptide can be generated by different proteins. Again, we follow those information to specify the conditional probability model for peptide indicators Y given protein indicators X.

We let $\pi_{i,j} = Pr(Y_j = 1|X_i = 1)$ denote the probability that peptide j is correctly identified conditional on its parent protein i being present. Let C_j denote the set of proteins that could potentially generate peptide j. Notice that the probability that peptide j is present in the digested sample is equal to the probability that at least one protein in C_j generates it.

We have

$$Pr(Y_j|X) = 1 - \prod_{i \in C_j} (1 - \pi_{i,j})^{X_i} \tag{2.4}$$

This conditional probability $\pi_{i,j}$ might depend on certain peptide sequence specific information, such as amino acid content, charge, hydrophobicity and polarity (see [29]). Ideally, incorporating those covariates that contain information about the observability of peptides can help us accurately estimate peptide and protein probabilities. Again, a logistic regression or Probit model can be a natural choice to incorporate those relevant explanatory variables in a prior model for $\pi_{i,j}$. Due to the lack of those measurements in our data, we simply assume $\pi_{i,j} = \pi$, a constant unknown parameter which has prior information assigned before MCMC steps.

The Fourth Stage: $\mathbf{Pr}(\mathbf{X})$

Finally, we assume the prior model for the presence of protein i is a Bernoulli distribution, i.e., $P(X_i = 1) = q_i$. Here q_i is the prior probability that protein i is present. Again, prior knowledge on the presence/absence status of proteins can be naturally incorporated in the prior model for q_i. For example, if organism level information is available, we can further specify another layer of model in the

hierarchical framework that reflects protein/organism grouping information. Again, due to the lack of this information in our study, we simply assume $q_i = q$, a constant unknown parameter that has prior information assigned before MCMC steps.

2.2.4 Bayesian Implementation and Bayesian False Discovery Rate

We employ MCMC methods to do the model fitting. We begin with prior specifications for the parameters. Recall that we have unknown parameters (μ_0, σ_0, ξ_0) in the GEV distribution f_0, (μ_1, σ_1, ξ_1) in the GEV distribution f_1, τ, π and q in the latent stage models. Vague normal priors are assigned to the location parameter μ_0, μ_1, shape parameter ξ_0, ξ_1, and an inverse gamma priors are assigned to σ_0, σ_1. We also assign $Beta(1, 1)$, i.e., uniform distribution on $[0, 1]$, to the parameters π, τ and q.

Posterior inference for the model parameters is completed using Gibbs sampling [35] with Metropolis-Hastings updating [36]. With the above hierarchical model and prior setup, all the conditional distributions turn out to be of known standard except for the GEV parameters in f_0 and f_1. $(\mu_0, \sigma_0, \xi_0), (\mu_1, \sigma_1, \xi_1)$ are then updated using Metropolis steps. Typically, random walk Metropolis with normal proposals is adopted. The detailed MCMC steps are given in Appendix.

In particular, we collect posterior samples of the latent peptide indicators Y_j. The posterior correctness probability for each peptide is estimated by the posterior sample proportion of Y_j. The correctness probability for each protein is obtained in a similar way. The associated posterior error probability (PEP, also referred to as local false discovery rate) is defined as the probability of being incorrect, simply calculated by subtracting posterior probability from one. For a given probability threshold p_c, peptides with PEPs lower than p_c are decided to be positive identifications. The Bayesian FDR (see [37] and [32]) associated with the above decision is estimated by the average PEPs that are below p_c.

$$FDR(p_c) = \frac{\sum_{PEP<p_c} PEP}{\#\{PEP < p_c\}}$$

A similar approach is easy to be applied to protein posterior PEPs. The Bayesian FDR could be used to assess the protein level identifications by our approach.

2.3 Results

Our identification approach is applied to the SEQUEST output on the Shewanella data described in section 2.1. The independence assumption among PSMs causes underestimated FDR, i.e. over-estimated number of identification at a specific FDR cutoff, which confronts with the rule of conservative estimation. Figure 2.6 shows that, compared with the true FDR lower bound, there are more identified peptides at the same Bayesian FDR estimated under PSMs independence assumption (red curve), which obeys the conservative rule for FDR estimation. Meanwhile, the Bayesian FDR estimated under PSMs dependence assumption (green curve) shows its conservative character.

The number of PSMs incorporated into the model–R has an impact on peptide identification. Figure 2.7 shows the positive trend between the number of identified peptides versus R–the number of top PSMs candidates utilized in the model, at a specific FDR cutoff 0.05. The same trend maintains at different FDR cutoffs.

Our approach that models the top 10 PSMs, which identifies more peptide features than PeptideProphet under most of the circumstances except for estimated FDR below 0.02, see Figure 2.8.

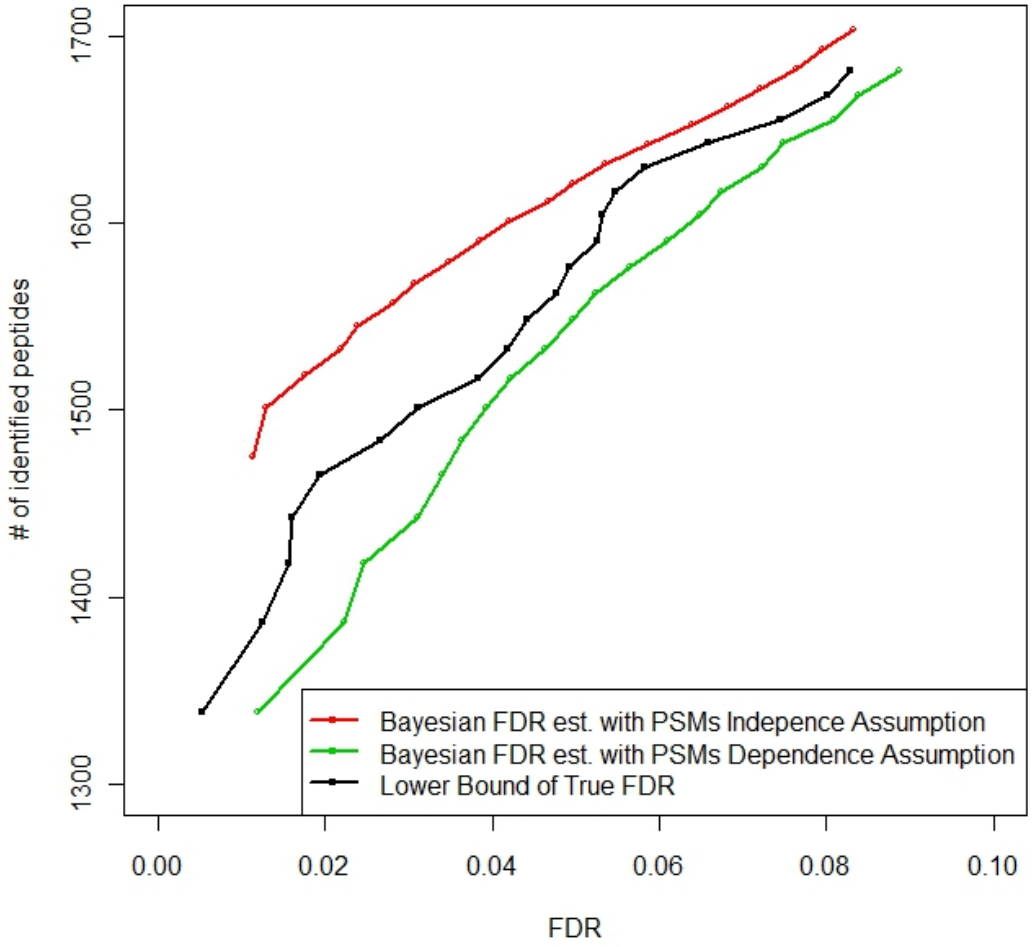

Fig. 2.6. The effect of independence assumption towards FDR estimation. Red curve is estimated Bayesian FDR under independence assumption of PSMs on the same peptide, green curve is estimated Bayesian FDR under dependence PSMs assumption of PSMs on the same peptide, black curve is the true FDR lower bound.

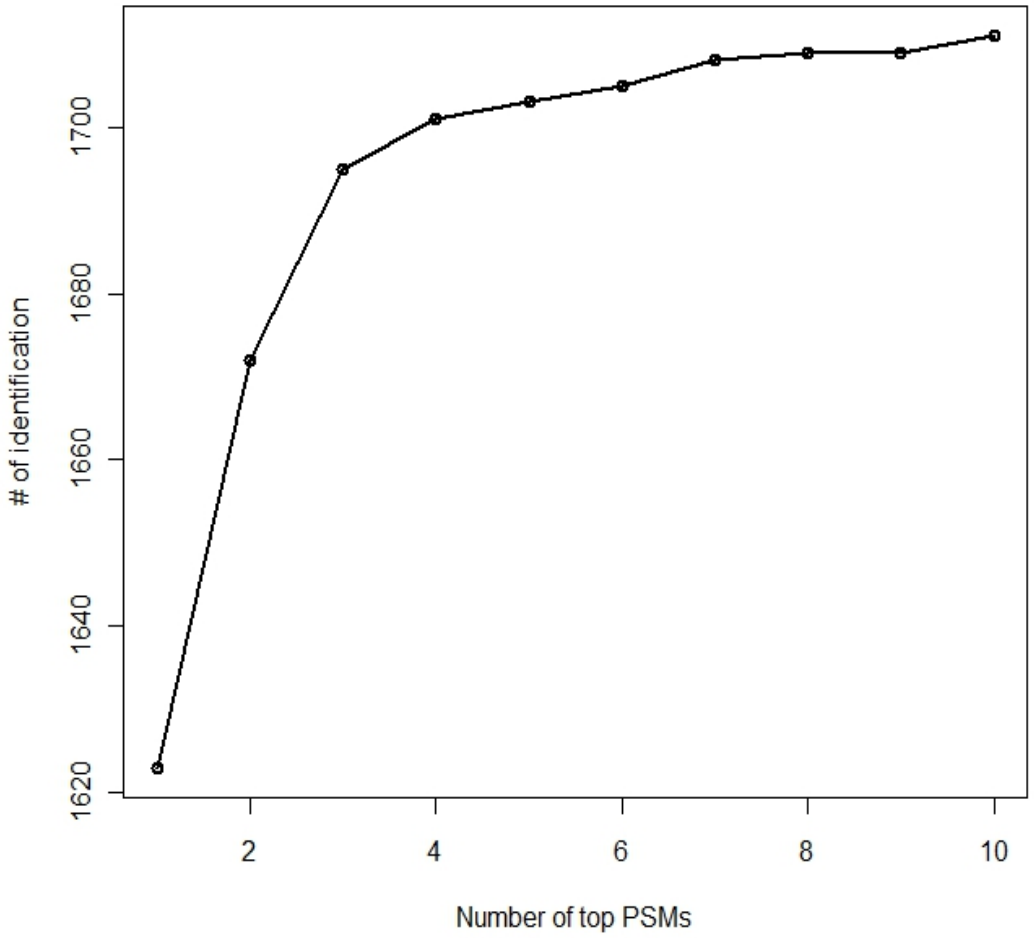

Fig. 2.7. The number of peptide identification vs the number of PSMs used in the model at 0.05 FDR cutoff.

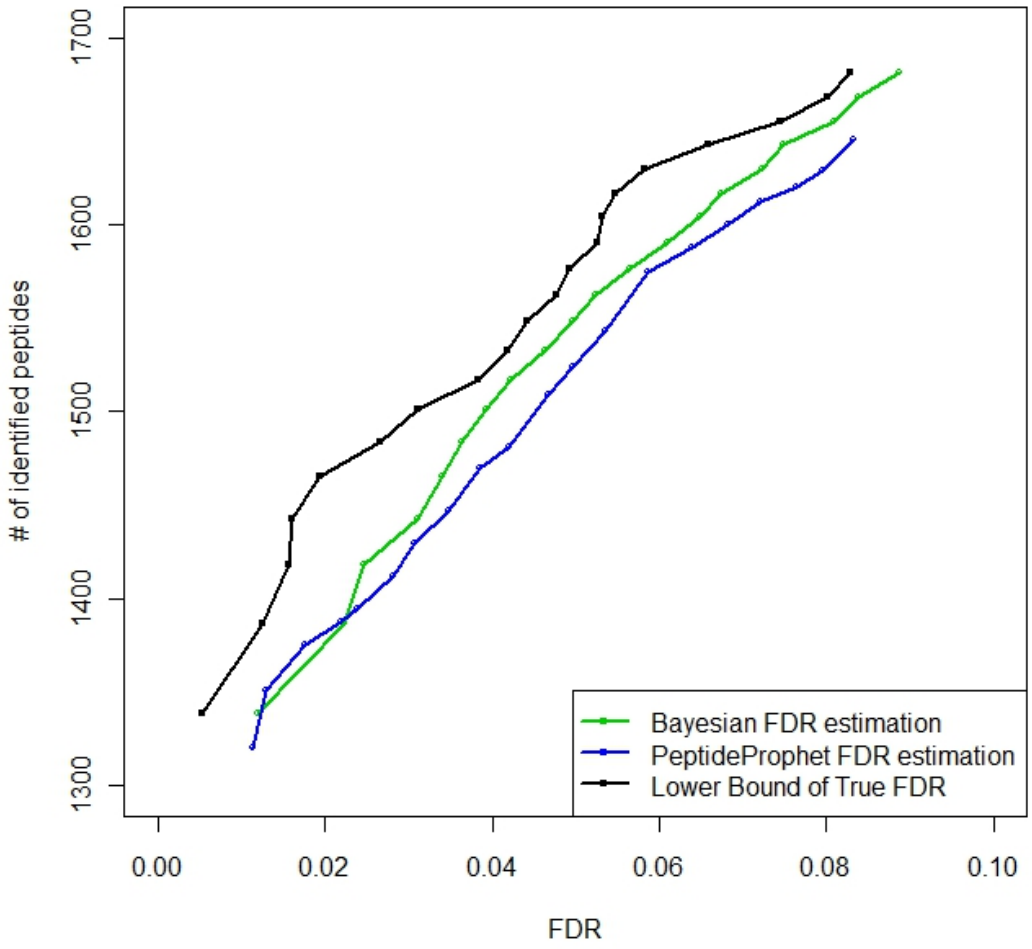

Fig. 2.8. The number of peptide identification vs estimated FDR using top 10 PSMs candidates. The green curve is generated by our approach and the blue curve is given by PeptideProphet. The black curve is associated with the true FDR lower bound

2.4 Discussion

In our framework, protein level posterior probability could be estimated one layer after obtaining peptide level posterior probability estimation. Hence protein level inference such as PEP, Bayesian FDR and the comparison with empirical lower bound FDR could be done in a similar way. However, due to the lack of result from other protein level identification algorithms on this Shewanella dataset, no comparison is carried on in the same fashion as comparison at peptide level with PeptideProphet.

Our model could also be applied to the AMT tag approach, where we would have both MS/MS data as well as hi-resolution mass and elution time measurements. The MS data with AMT tag could also be searched against AMT database with matching scores returned. We need to normalize matching scores and model the distribution of the correct and incorrect normalized matching scores as well as other informative covariates. Four layers need to be constructed for score—AMT match, AMT match—peptide, peptide—protein and protein. In order to run MCMC, prior information also need to be determined for the parameters in each layer as well as in the model of matching scores. After the simulation, posterior probability at peptide level and protein level could be inference by sample proportion and then produce Bayesian FDR estimation.

3. A STATISTICAL APPROACH TO THE ALIGNMENT OF LC-MS/MS SAMPLES

3.1 Introduction

3.1.1 Background

In LC-MS, each sample may have thousands of scans, each containing a mass spectrum. The mass spectrum for a single MS scan can be summarized by a plot of M/Z values versus peak area values. These data contains signals that are specific to individual peptide. As a first step towards quantifying these peptides, features need to be distinguished from background noise. One simple method is to employ a filter on the peak's signal-to-noise ratio relative to its local background. Each peptide gives an envelope of peaks due to a peptide's constituent amino acids. The presence of a peptide can be characterized by the M/Z value corresponding to the peak arising from the most common isotope, referred to the mono-isotopic mass.

3.1.2 Existing Alignment Methods

The goal of alignment is to match corresponding peptide features in the Scan Number vs M/Z plot (see figure on p. 49 in section 3.3.2) from different experiments. A time warping method based on raw spectrum for alignment of LC-MS data was introduced by Bylund and others [16], which is a modification of the original correlated optimized warping algorithm [17]. Wang and others [38] implemented a dynamic time warping algorithm allowing every Retention Time (RT) point to be moved. However, the LC-MS data have added dimension of mass spectral information, so only mapping the retention time coordinates between two LC-MS files is not

sufficient to provide alignment for individual peptide. Radulovic and others [19] performed alignment based on (M/Z,RT) values of detected features. Their method first divides the M/Z domain into several intervals and fitted different piece-wise linear time warping functions for each M/Z interval. After the time warping, they applied a "wobble" function to peaks and allow peaks to move (1-2% of total scan range) in order to match with the nearest adjacent peaks in another file. Their method relies on the (M/Z,RT) values of detected peptide features, it fails to take advantage of other information in the raw image. Wang and others [38] proposed an alignment algorithm, PETAL, for LC-MS data. It uses both the raw spectrum data and the information of the detected peak features for peptide alignment.

In Section 3, two Shewanella data sets are obtained from Pacific Northwest National Laboratory (PNNL) and were analyzed by SEQUEST on different days. SEQUEST correlates uninterpreted tandem mass spectra of peptides with amino acid sequences from protein and nucleotide databases, which determines the amino acid sequence and thus the protein(s) and organism(s) that correspond to the mass spectrum being analyzed. Based on the SEQUEST output files, each sample has thousands of scans, and M/Z, peak areas and peptide identification information associated. It's obvious that there's some systematic error between the the two data sets before alignment.

In this study, we first applied some filter criteria to choose data points matched in both samples with high confidence, which are called "Anchor points". We then use these "Anchor points" in sample one as the baseline and modify the data points in sample two, to make the "Anchor points" between the two samples aligned together, which means after alignment the "Anchor points" in both samples show up at the same locations. The alignment algorithm is then generated to all the other data points in sample two. Finally, statistical measurements of the performance of alignment are given on sample level and regional level.

In future study, we hope this alignment method can be applied to several samples of one organism, and as a guide to justify the points with same peptide information in different samples.

3.2 Methods

3.2.1 Anchor Points

In our method, for different experiments, we have the raw data analyzed by SEQUEST. Based on the SEQUEST output files, each sample has thousands of scans, M/Z, peak areas and peptide information. A sample record of data is given in Table 3.1. Define a point with high probability that its peptide shows up in SEQUEST output files as "Anchor points".

Table 3.1
A sample record of SEQUEST output.

ScanNum	MZ	PeakArea	NTE	PassFilt	SignalToNoiseRatio
8239	826.4	4.98E+06	2	1	25.1
ChargeState	Xcorr	DelCN	RankXc	Reference	Peptide
3	7.9171	0	1	SO_2336	K.LA...GYVHA

In LC-MS, we need to distinguish the peptide features from the background noise, the first step for doing this is MS peak detection. We employ a simple filter routine on the Signal-to-Noise Ratio of a peak relative to its local background. In our approach, in order to find peptides that exist in both samples with high confidence, three filtering criteria are applied. The first criterion is PassFilt equaling to 1, the second criterion is Number of Tryptic Ends (NTE) equaling to 2 and the third criterion is Signal-to-Noise Ratio being greater than 10. PassFilt is a score that does not come from by SEQUEST but is calculated from syn-fht summary generator using

Xcorr, δ_{C_n}, RankXc and the number of tryptic termini (NTT), where NTT is the number of termini that conforms to the expected cleavage behavior of trypsin (i.e. C-terminal to R and K). Note that K-P and R-P do not qualify as tryptic cleavages because of the proline rule. However, the protein N-terminus and protein C-terminus do count as tryptic cleavage sites. Values can be $0, 1, 2$ with 2 = fully tryptic; 1 = partially tryptic; 0 = Non tryptic. Any point associated with the same peptide in both samples satisfied these three criteria is called "Anchor point", see Figure 3.1.

3.2.2 Alignment Algorithm

"Anchor points" found from both samples differ on "Scan Number", which represent elution (retention) time, on which we need to find some algorithm to make these points in the two samples aligned, as well as on M/Z and Peak Area. Since the range of Peak Area is not as the same magnitude as Scan Number or M/Z, a logarithm transformation is applied. Let S_i be the Scan Number for peak i, M_i be M/Z for peak i, and P_i be the log(peak Area) for peak i. The data is normalized to zero-to-one range by dividing normalization factors of Scan Number, M/Z, and log(Peak Area), denoted as R_X, R_M and R_P, which are the maxima of Scan Number, M/Z, and log(Peak Area) across the two samples. The reason why we do normalization is that due to the different scales of Scan Number, M/Z and log(Peak Area), the distances can not be equally measured, for example, the scan number is very large compared to the M/Z. The normalization could transform the three dimensional measurement under the same scale and thus give equal weighting to all of them. The alignment algorithm uses both the raw spectrum analyzed by SEQUEST and the detected peak features for peptide alignment. Although in the two samples, the peaks with same identified peptide information appear in different scan numbers due to systematic bias, we assume that the scan numbers, M/Z and log(Peak Area) should not differ much within the paired peaks.

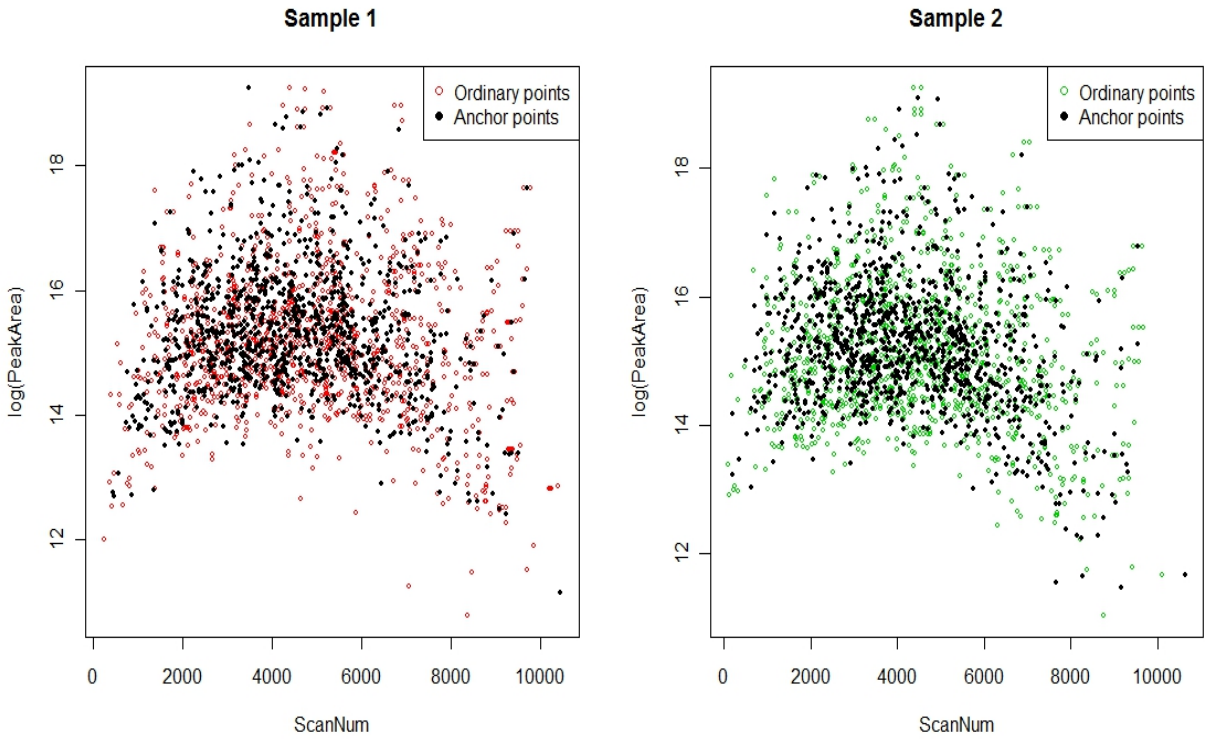

Fig. 3.1. Plot of anchor points embedded in both samples. Left panel is for sample one and right panel for sample two.

With the pool of "Anchor points" found between sample one and sample two, we are able to locate the K nearest anchor points for peak i in sample two, where the distance is defined by Euclidean metric considering the three dimensions of normalized Scan Number, M/Z and log(PeakArea). With the defined metric, let D_{ij} be the distance between peak i and an arbitrary "Anchor point" j in the sample two, $D_{ij} = (((S_i - S_j)/R_S)^2 + (M_i - M_j)/R_M)^2 + (P_i - P_j)/R_P)^2$.

Since each "Anchor point" in sample two is paired with an "Anchor point" in sample one, let Δ_S be be the averaged scan number differences across the K nearest "Anchor point" pairs; let Δ_M be the averaged M/Z differences across peak i's K nearest "Anchor points" pairs; and let Δ_P be the averaged log(peak Area) differences. Then we use the differences to modify peak i in sample two by adding $(\Delta_S, \Delta_M, \Delta_P)$ to (S_i, M_i, P_i). $\Delta_S = \overline{S}_{i.1} - \overline{S}_{i.2}$, $\Delta_M = \overline{M}_{i.1} - \overline{M}_{i.2}$, $\Delta_P = \overline{P}_{i.1} - \overline{P}_{i.2}$, where $\overline{S}_{i.1}$ is the average over Scan Number of K "Anchor points" in sample one.

3.3 Real Data Example

3.3.1 Visualization of Alignment

In the Shewanella datasets, $R_S = 10606$, $R_M = 1519.48$, $R_P = 10.08476$. The data is firstly normalized with R_S, R_M and R_P and then searched for "Anchor points". There are systematic bias between the differences on Scan number, M/Z and log(Peak Area) of "Anchor points" from both samples before alignment. After alignment, the differences should be randomly distributed around 0. We draw histograms to compare the distance of "Anchor points" between the two samples before and after alignment. Figures 3.2 and 3.3 are the histogram of Scan Number and M/Z. The histogram of log (Peak Area) is shown in Figure 3.4 on page 48. The histograms show that, after alignment, the differences on Scan Number, M/Z and log (Peak Area) between the two samples are mostly around 0.

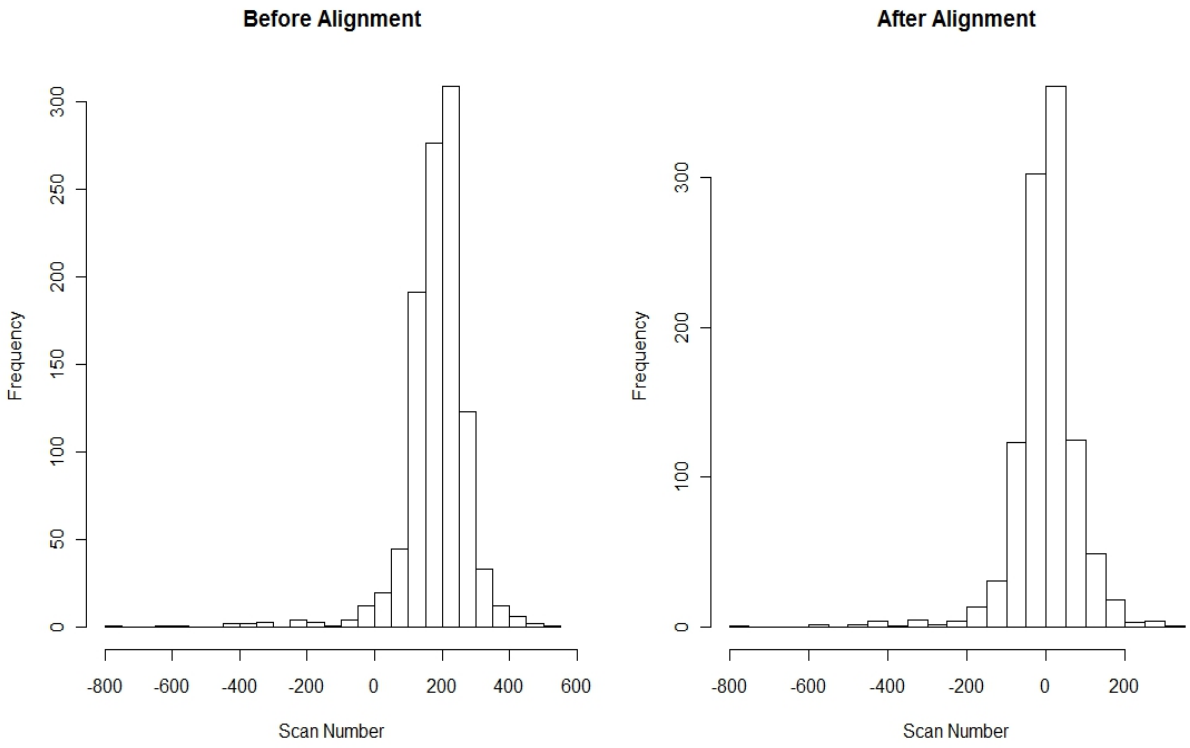

Fig. 3.2. Histograms of Scan Number of anchor points. Left panel is for before alignment and right panel for after alignment.

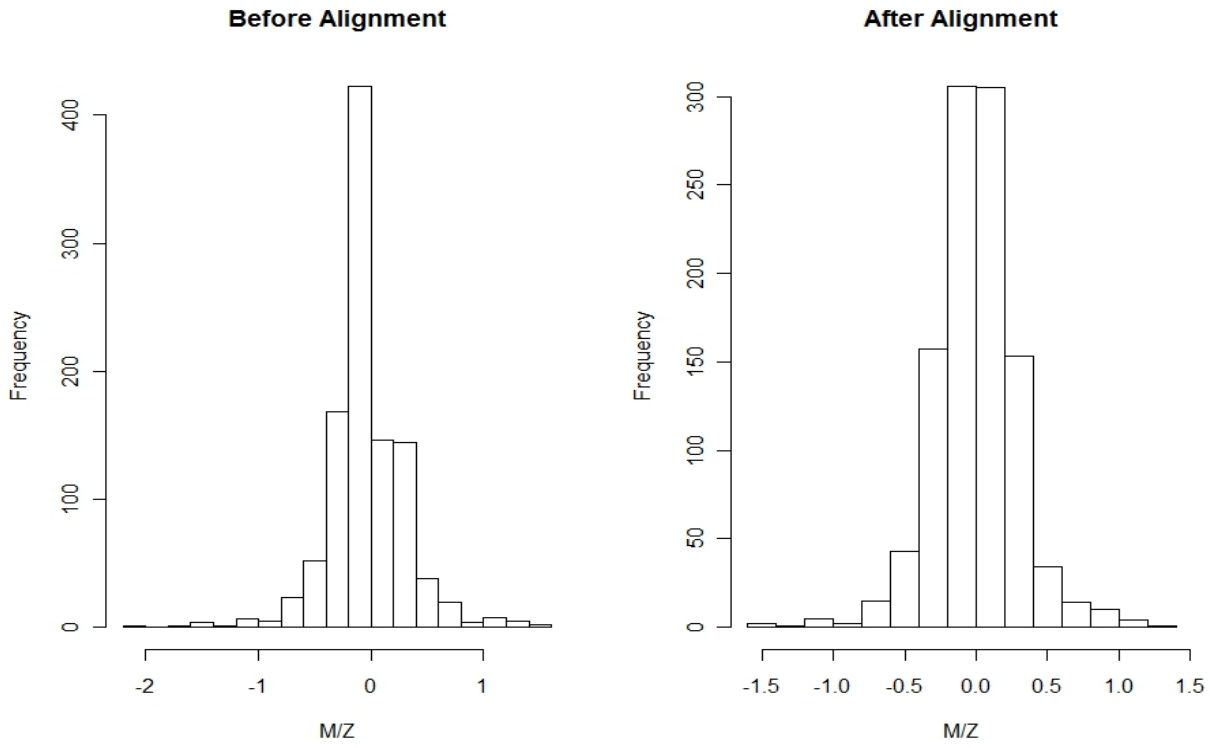

Fig. 3.3. Histograms of M/Z of anchor points. Left panel is for before alignment and right panel for after alignment.

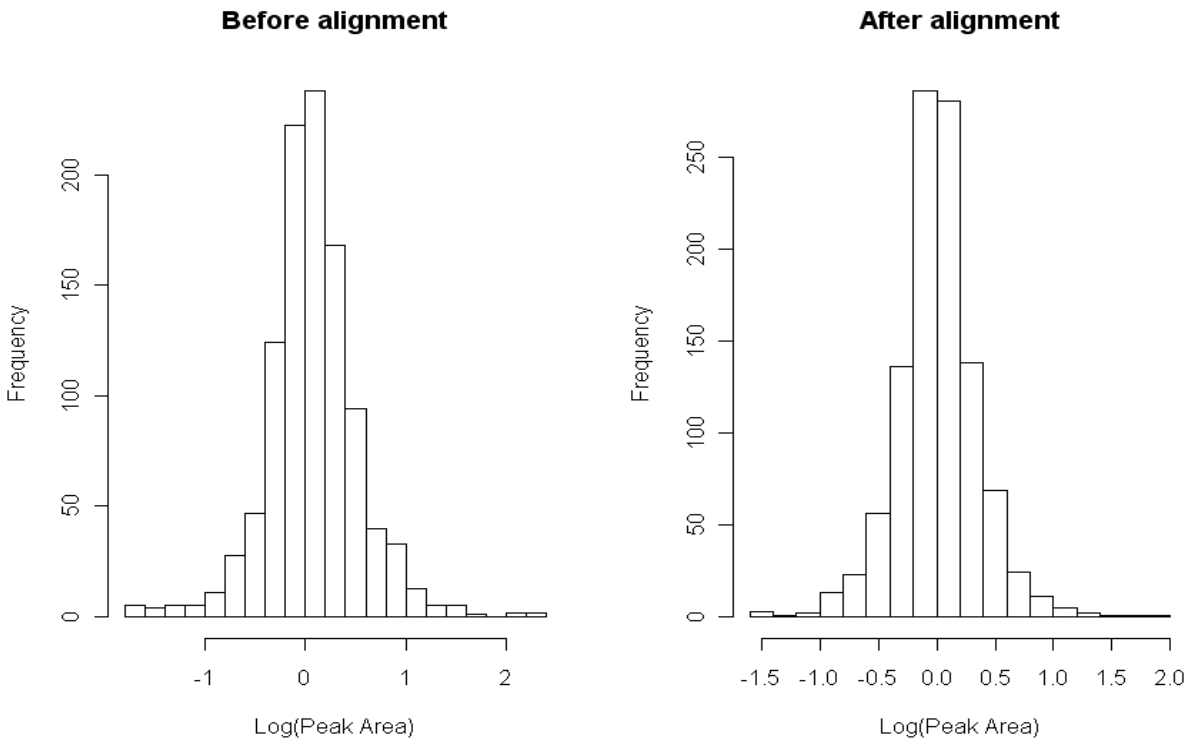

Fig. 3.4. Histograms of Log(Peak Area) of anchor points. Left panel is for before alignment and right panel for after alignment.

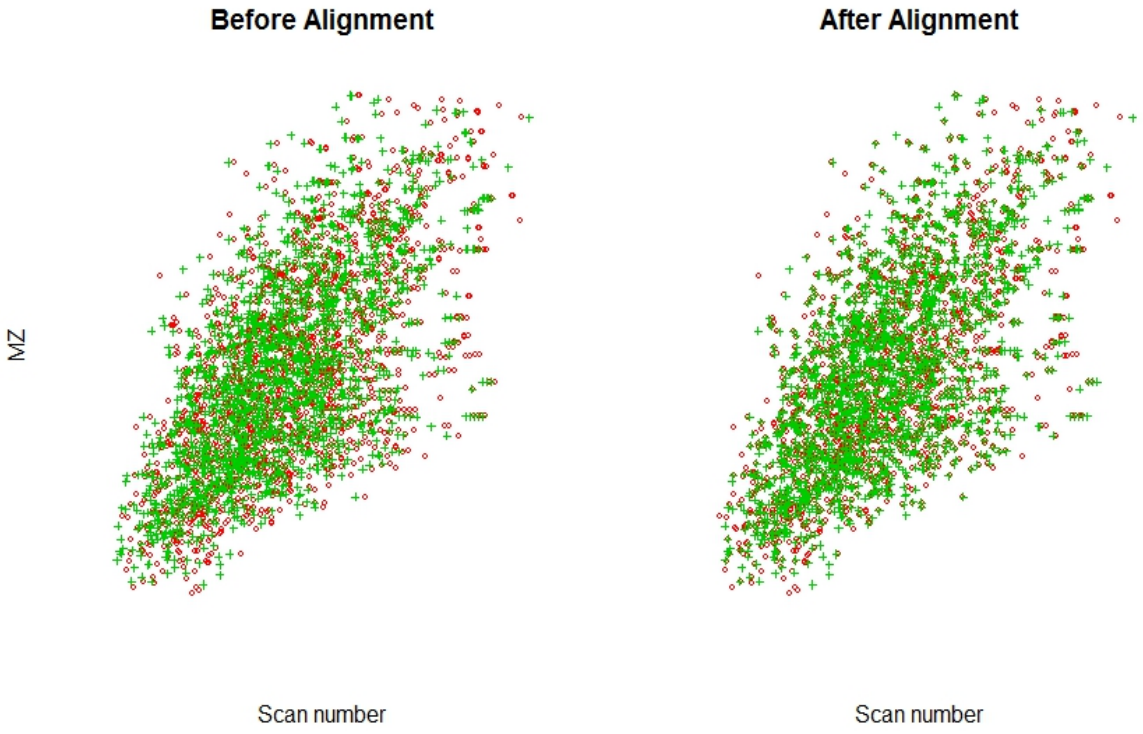

Fig. 3.5. Scatter plot of Scan Number vs. M/Z on all data points. Left panel is for before alignment and right panel for after alignment.

To visualize the alignment, scatters plots of Scan Number versus M/Z and Scan Number log (Peak Area) on "Anchor points" before and after alignment are given as Figures 3.5 and 3.6.

3.3.2 Global P-value for Alignment Performance

It is important to perform a statistical test to justify that after alignment, there is no systematic difference across all the "Anchor points" between the two samples. Wilcoxon signed-rank test is applied on the differences on Scan Number, M/Z and log(Peak Area) of the "Anchor points" between the two samples, before and after

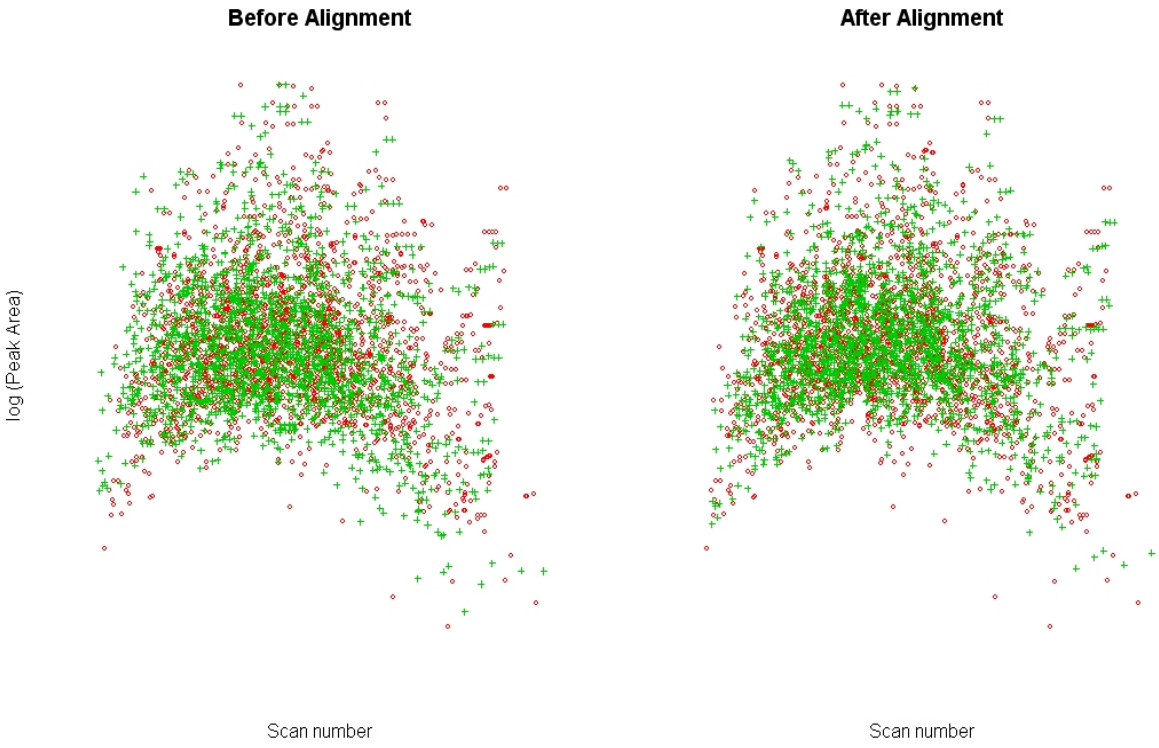

Fig. 3.6. Scatter plot of Scan Number vs Log (Peak Area) on all data points. Left panel is for before alignment and right panel for after alignment.

alignment with $K = 5$. The null hypothesis is that the true location shift is equal to 0 while the alternative hypothesis is the true location shift is not equal to 0.

We have the following results: Before alignment Wilcoxon signed-rank test on Scan Number of anchor point with continuity correction $W = 537733$, p-value $\leq 2.2e-16$. After alignment Wilcoxon signed-rank test on Scan Number of anchor point with continuity correction $W = 2932224$, p-value $= 0.0986$.

Before alignment Wilcoxon signed-rank test on M/Z of anchor points with continuity correction W $= 142212$, p-value $= 0.1091$. After alignment Wilcoxon signed-rank test on M/Z of anchor point with continuity correction W $= 265884$, p-value $= 0.9049$.

Before alignment Wilcoxon signed-rank test on log(Peak Area) of anchor point with continuity correction W $= 336362$, p-value$= 1.655e-09$. After alignment Wilcoxon signed-rank test on log(Peak Area) of anchor point with continuity correction W $= 281910$, p-value $= 0.6141$.

So we conclude that the difference on Scan Number, M/Z and log(Peak Area) of "Anchor points" after alignment in two samples has a common median 0, which indicates our alignment approach is valid and effective to fulfill the alignment task.

3.3.3 Local P-values for Alignment Performance

We then generalize the single global p-value on the whole data set to local p-values, realized by Wilcoxon sum rank test. For example, the entire span of Scan Number versus Log (Peak Area) are divided into, for example, 10 X 10 small rectangular, the cutoffs on the margins are the $10\%, 20\%, \cdots, 90\%$ quantiles of Scan Number and Log (Peak Area). Take a single rectangular for example, Wilcoxon sum rank test is applied on Scan Numbers that fall into that single rectangular before and after alignment and hence gives a p-value for that region. We repeat this on the other 99 rectangular and obtain 100 p-values in total. The p-values on Log (Peak Area) are calculated in a similar way. Figure 3.7 displays the histogram of the 100

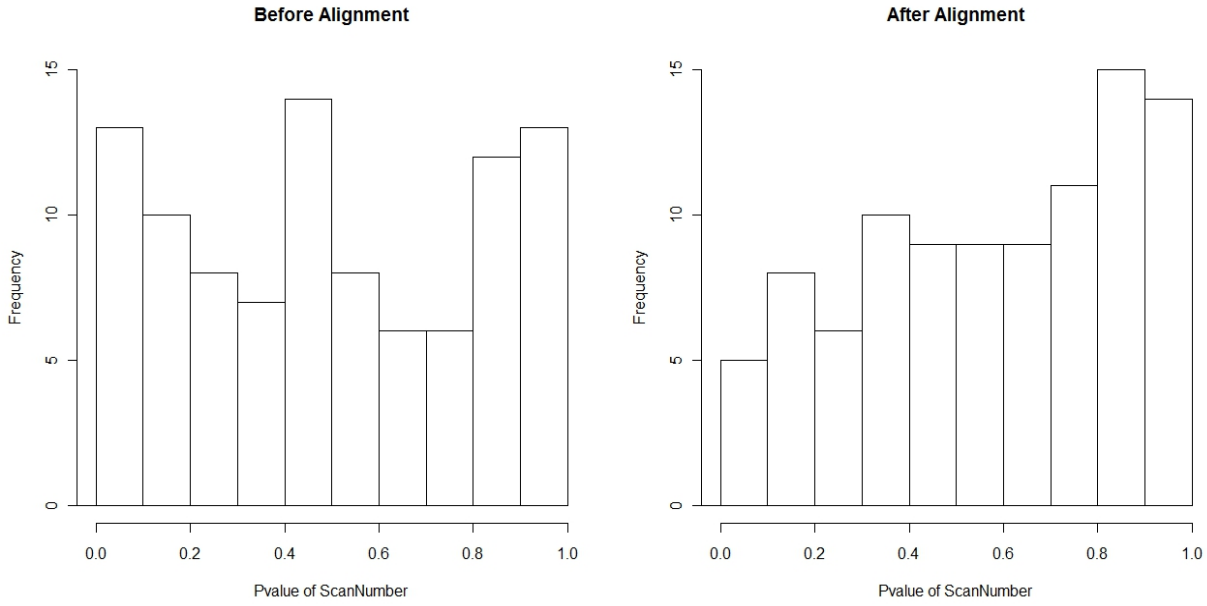

Fig. 3.7. Histogram of regional p-values on Scan number. Left panel is for before alignment and right panel for after alignment.

p-values on Scan Number and Log (Peak Area). We can see that there are fewer small p-values on Scan Number after alignment. The heat map on Scan Number confirms that since there are less red but more yellow rectangular after alignment, shown as in Figure 3.8

The regional division could be realized in three dimensional fashion, but due to the limitation of data size, there would not be enough peaks in a large proportion of three-dimensional cubes to do the Wilcoxon sum rank test. However, if the data size is sufficient, the localization of p-value would be more precise and valuable to evaluate the effectiveness of the alignment algorithm.

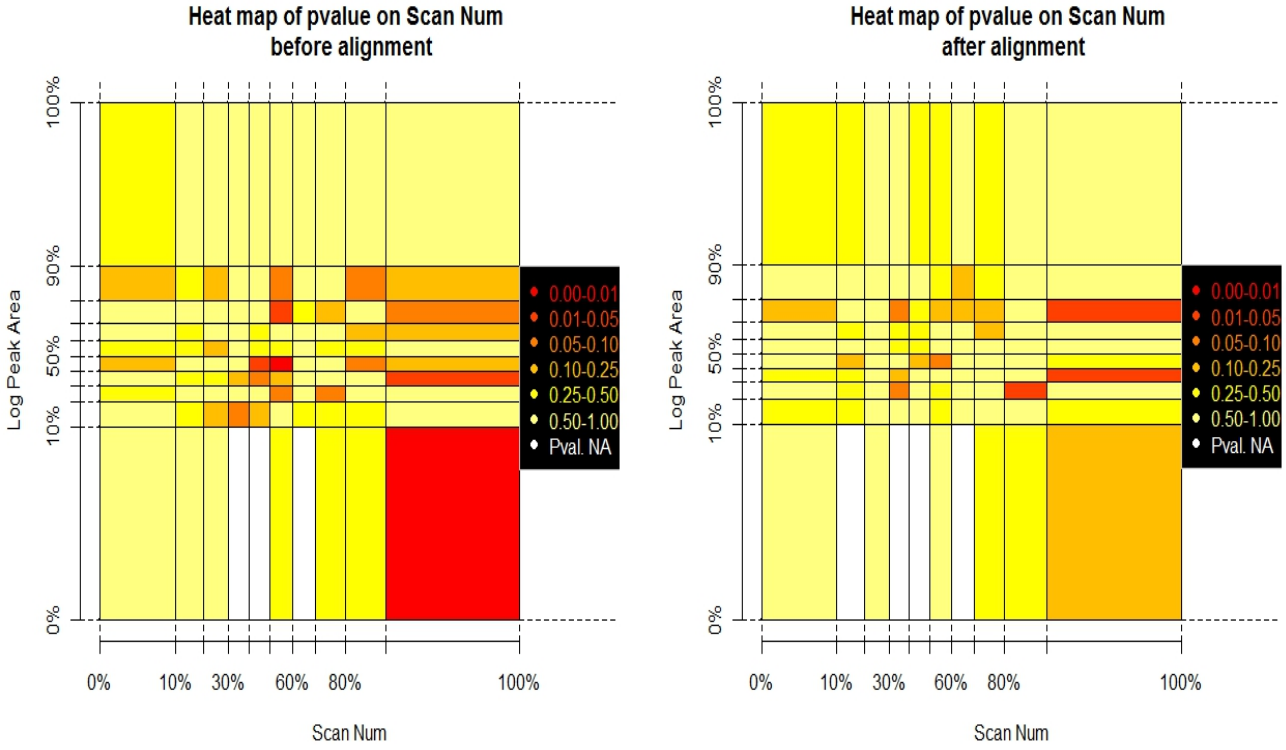

Fig. 3.8. Heat map of regional p-values on Scan number. Left panel is for before alignment and right panel for after alignment.

3.4 Discussion

One advantage of our method is that it allows statistical assessment of alignment performance. We could statistically evaluate the performance of our methodology with other alignment algorithms on some data set that has peptide identification with high confidence.

The statistical confidence measure of our method was given on sample level. We expanded it to region level and the future work would be developing peptide level statistical confidence measure and pass it to downstream quantitative analysis.

4. A HYBRID APPROACH TO PROTEIN DIFFERENTIAL EXPRESSION IN MASS SPECTROMETRY-BASED PROTEOMICS

4.1 Introduction

A key goal of quantitative mass spectrometry-basad proteomics is statistical inference on differential protein expression. Quantitative information is derived froe spectral peak intensities that are identified as having come from one of a protein's constituent peptides. Statistical procedures for differential protein expression are naturally constructed in the context of regression or ANOVA [39], or as a "rollup" problem [23].

However, intensity-based procedures are challenged by the presence of widespread missing intensities. It is typical for 20-40% of the total collection of attempted measurements to be missing; that is, in a matrix with all identified peptides in the rows, samples in the columns, 20-40% of the matrix cell entries are empty. With standard regression or ANOVA procedures, peptides with missing values must either be removed from the analysis, or their missing values must be imputed. There will typically be very few peptides with no missing values, so filtering peptides in this way results in a much less informative data set. Furthermore, previously-published reports indicate that the vast majority of missing values are the results of censoring of absent or low-abundance peptides [24]. This means that simple imputation routines are not appropriate [25].

Parametric imputation and other specialized methodology can be employed to enable intensity-based inference with lessened information loss. However, some information loss is inevitable. In particular, "one-state" (or nearly so) peptides, those for which there are many observed intensities in one comparison group but few in another comparison group, are not amenable to an intensity-based analysis; not lim-

ited to the two-class problem. As a result, such peptides are typically filtered out of an intensity-based analysis. A protein that is always present in a diseased state, say, and never in the healthy state would be of great biological interest, so it is unfortunate if our statistical methodology can not identify such a protein.

An alternative to an intensity-based analysis is a "presence / absence" analysis, in which peak intensities are digitized into binary measurements depending on whether a peak was observed or not. This is analogous to the spectral counting approach in MS/MS studies [40], where a peptide is quantified by the number of fragmentation spectra assigned to it. Data collected in our laboratory does not necessarily have MS/MS fragmentation data associated with at, instead being obtained according to the Accurate Mass and Time (AMT) tag pipeline [18]. Still, we have information on whether or not a particular peptide was measured in each sample.

While presence / absence analysis is better-suited to finding one-state proteins, it necessarily has less statistical power to detect abundance differences in proteins with little to moderate missingness. Ideally, protein differential expression analysis would simultaneously target proteins of both types, resulting in a single list of differentially expressed proteins, with a single associated false discovery rate (FDR). A hierarchical Bayesian model would be well-suited to this purpose, but such techniques are complex, computationally intensive, and hence amenable to high-throughput pipelines.

We present a hybrid analysis protocol that consists of two stages: (i) intensity-based analysis, and (ii) a presence / absence analysis. The results of each are merge‘ to create a single collection of "interesting proteins" po which we use novel methodology to apply a single FDR. This enables the researcher to detect more information from a quantitative proteomic data set than would be achievable by either approach alone, while still maintaining an interpretable measure of overall statistical confidence. For the proposed hybrid analysis protocol, we demonstrate the following: (a) Resulting FDR estimates are conservative, (b) One-state proteins are consistently selected as differentially expressed, and (c) The number of differentially expressed

proteins selected at a specified FDR exceeds that either intensity-based or presence / absence analysis alone.

4.2 Methods

4.2.1 Data

Diabetes

These data are as previously described [21]. In order to minimize the number of sibling peptides with large missingness proportion differences, we use PeptideSieve [41] to filter peptides whose amino acid sequences are unlikely to be detected by MS. 554 peptides are filtered out before carrying on any further analysis.

Simulation

We carried out simulation studies as follows, to investigate the operating characteristics of our methodology at both peptide and protein levels. Peptide-level data were generated from a Binomial model, under the same conditions as the diabetes data (two comparison groups with 10 samples in each). Presence probabilities in group one took the values on $p_1 = 0.2, 0.3, 0.4, 0.5$. Half of the group-two peptides were assigned the same presence probabilities as their group-one counterparts. In the other half, differential presence probabilities were created, with probability differences (comparing group two to group one) of $p_d = p_2 - p_1 = 0.1, 0.2, \cdots, 0.9 - p_1$. Separate simulations were carried out for each of the group-one presence probability values with even replications on different p_d settings. Similarly, for protein-level data, the number of peptides per protein was randomly selected to range between 1 and 30. Protein-level presence probabilities also took the values $p_1 = 0.2, 0.3, 0.4, 0.5$ and $p_d = p_2 - p_1 = 0.1, 0.2, \cdots, 0.9 - p_1$. For each constituent peptide, the group-one peptide-level presence probability equaled the protein-level probability multiplied

by a randomly-selected number between 0 and 1 (to allow for different levels of detectability for peptides of the same protein). Peptide-level differential presence probabilities were handled as described for the peptide-level simulation above. Finally, to simulate data for use by the hybrid method, with both peak intensities and presence / absence indicators, we randomly generated intensities from a Normal distribution with parameters chosen to mimic the diabetes data. Missingness proportions took the values 10%, 20%, 30%, and 40%, with missingness created by censoring the lowest corresponding percentages of peptide intensities. As in the above simulations, half of the peptides / proteins were given differential expression, now defined in terms of mean intensity levels. Differential intensity magnitudes took both low-magnitude values of 1, 2, as well as high-magnitude values of 5, 10, all on the log scale.

4.2.2 Logistic Model for Protein Presence / Absence

Logistic regression is a natural analysis method for presence / absence data, given their binary nature. Specifically, let Y_{ijkl} be the indicator for whether a peak was observed for peptide j of protein i in comparison group k and sample l. Then, we can say $Y_{ijkl} \sim \text{Binomial}(1, p_{ijk})$ for $l = 1, 2, \ldots, n_k$, where n_k is the number of samples in comparison group k. A simple logistic regression model would then be

$$\text{logit}(p_{ijk}) = \texttt{Prot}_i + \texttt{Pep}_{ij} + \texttt{Grp}_{ik}. \tag{4.1}$$

Here, $\texttt{Prot}_i$ represents the overall (across all comparison groups) log odds of peak presence for protein i, $\texttt{Pep}_{ij}$ is the effect of peptide j of protein i (assumed to be the same across all k comparison groups), and $\texttt{Grp}_{ik}$ is the protein-level effect of comparison group k in protein i. Usual sum-to-zero constraints apply; namely, $\sum_{j=1}^{m_i} \texttt{Pep}_{ij} = 0$ and $\sum_{k=1}^{K} \texttt{Grp}_{ik} = 0$ for $i = 1, 2, \ldots, M$, where M is the total number of proteins in the data.

For the purposes of comparing protein presence probabilities across comparison groups, the parameters of interest are the $\texttt{Grp}_{ik}$, $i = 1, 2, \ldots, M$. For example, in the diabetes data, $K = 2$, with $k = 1, 2$ corresponding to the diabetic and control groups, respectively. Hence, $\texttt{Grp}_{i1} - \texttt{Grp}_{i2}$ is the log odds ratio for protein i, comparing diabetics to controls. Testing for a difference in presence probabilities corresponds to testing the null hypothesis that $\texttt{Grp}_{i1} - \texttt{Grp}_{i2} = 0$; given the model's sum-to-zero constraints, this is equivalent to the null hypothesis that $\texttt{Grp}_{i1} = \texttt{Grp}_{i2} = 0$. Of course, the model is not restricted to the two-class case and can naturally be generalized to the K-class case, in which the corresponding null hypothesis is that $\texttt{Grp}_{i1} = \texttt{Grp}_{i2} = \cdots = \texttt{Grp}_{iK}$.

Unfortunately, logistic regression is not well-suited in practice to the analysis of presence / absence data. In particular, biologically-interesting proteins are liable to be missed entirely, due to inherent limitations of the methodology. Consider a "one-state" protein, present in all samples for one comparison group, absent in all samples for the other comparison groups. From a biological perspective, this would be a very interesting protein. However, in logistic model, the p-value for such a protein will tend to be reported as nearly equal to one, meaning that the protein would not be selected as differentially expressed under any reasonable criteria.

A simple scenario illustrates this problem. Consider a "one-state" protein with just a single peptide. In logistic regression, the assumed variance-covariance matrix for regression coefficients is $(X'WX)^{-1}$, where X is the model matrix, and W is diagonal with entries $p_k(1 - p_k)$, with p_k the presence probability in comparison group k, $k = 1, 2$. For comparison groups in which no peaks were observed, the estimated value of p_k is zero, making the corresponding entry in W equal to zero. This results in an overestimation of the standard error for the group effect model term, hence an understatement of statistical significance for that protein's group effect. In the diabetes data, for example, "one-state" proteins are assigned p-values of one.

4.2.3 Peptide-level Exact Test

In light of the logistic regression limitations, we propose an exact procedure for testing for differences in presence / absence between two comparison groups. Let $y_{jk\cdot} = \sum_{l=1}^{n} y_{jkl}$ be the number of observed peaks for peptide j in comparison k, $k = 1, 2$. We use $T_j = |y_{j1\cdot} - y_{j2\cdot}|$ as the peptide-level test statistic. Based on the Binomial probability model, the exact sampling distribution of T_j under the null hypothesis H_0 of no difference in presence probabilities can be written as

$$\begin{aligned} Pr_{H_0}(T_j = t) = &\sum_{m_1=0}^{n_1-t} \sum_{m_2=m_1+t}^{n2} B(m_1; n_1, p_{j0}) \times B(m_2; n_2, p_{j0}) \\ &+ \sum_{m_2=0}^{n_2-t} \sum_{m_1=m_2+t}^{n1} B(m_1; n_1, p_{j0}) \times B(m_2; n_2, p_{j0}) \end{aligned}$$

where $B(m; n, p)$ is the Binomial PMF at m, with n trials and probability of success p, and p_{j0} is the shared probability of peak presence for both groups. Thus, based on an observed statistic of t_j, the p-value is $\sum_{t \geq t} Pr_{H_0}(T_j = t)$. In practice, we need only to estimate the shared presence probability under the null hypothesis, p_{j0}, in order to approximate the p-value for a given peptide. We estimate p_{j0} with a pooled sample proportion, resulting in $\hat{p}_{j0} = \frac{\sum_k n_k \hat{p_{jk}}}{\sum_k n_k}$, where $\hat{p}_{jk} = y_{jk\cdot}/n_k$, $K = 2$.

As an example, consider a "one-state" peptide, present in all samples of one group but in no samples of the other group, in the diabetes data. Whereas logistic regression reports a p-value of one, the exact test correctly highlights the peptide as statistically significant. Specifically, the test statistic T_j equals 10, and $\hat{p}_{j0} = 0.5$, so the p-value is computed as $2 \times B(10; 10, 0.5) \times B(0; 10, 0.5) < 0.0001$

4.2.4 Protein-level Bootstrap Test

For inference at the protein level, there is the added challenge of multiple peptides belonging to the same peptide. To incorporate all sibling peptides into a single test for differential presence probabilities, we use the following test statistic:

$$T_{Mi} = \left| \sum_{j=1,\ldots,m_i} \kappa_{ij} \left(y_{ij1\cdot} - y_{ij2\cdot} \right) \right| \tag{4.2}$$

where i is protein index, j is peptide index, and k is comparison group index, $i = 1, 2, \ldots, M,\ j = 1, 2, \ldots, m_i,\ k = 1, \cdots, K,\ K = 2$. The statistic in (4.2) is a weighted average of observed presence difference on each sibling peptide. For the weighting term κ_{ij}, we use $\kappa_{ij} = y_{ij\cdot\cdot} / \sum_j y_{ij\cdot\cdot}$

A parametric bootstrap procedure [42] is used to approximate the sampling distribution of the T_{Mi} under null hypothesis setting as follows. Firstly the Binomial parameters p_{ijk} are estimated, for which two approaches are considered. The first approach simply uses the sample proportion for peptide j of protein i in comparison group k being present, which needs $2 \times m_i$ parameter estimation per protein. Alternatively, we approaching the problem by inducing some structure between the p_{ijk}, assuming that $p_{ijk} = p_{ik} \times d_{ij}$, where p_{ik} is the overall presence probability for protein i in comparison group k, and d_{ij} is the "detectability" probability (the probability that a particular ion species is detected by the LC-MS instrument) for peptide j of protein i. This assumption of structure translates to an assumption that the detectability of a peptide does not differ between comparison groups. Since detectability is a function of chemical composition rather than abundance [41], this seems a reasonable assumption. After introducing the structure assumption the number of parameters per protein to be estimated reduces from $K \times m_i$ to $K + m_i$.

With the second approach, the presence probability p_{ik} of protein i in group k is estimated by averaging the presence proportion of its top 10% most prevalent peptides.

$$\hat{p}_{ik} = \sum_{j \in \{top 10\%\}} \hat{p}_{ijk} / \#\{top\ 10\%\ peptides\ \}$$

The rationale here is that, for these most prevalent peptides, the detectability probability will be close to one, making $p_{ijk} \approx p_{ik}$. Then $\hat{p}_{ik}$ is used to estimate d_{ij} as $\hat{d_{ij}} = \frac{1}{K}\left(\frac{\hat{p_{ij1}}}{\hat{p_{i1}}} + \cdots + \frac{\hat{p_{ijk}}}{\hat{p_{ik}}}\right)$, where $\hat{p}_{ijk}$ and $\hat{p}_{ik}$ are the sample presence proportions.

Since we have $\hat{p_{ik}}$ and $\hat{d}_{ij}$, according to the equation $p_{ijk} = p_{ik} \times d_{ij}$, the null sampling distribution of our test statistic could be generated by parametric bootstrap. Under the null hypothesis setting, the presence probabilities of protein i across K comparison groups are the same and set to be p_{i0}. In two group case, $p_{i0} = p_{i1} = p_{i2}$ and $\hat{p_{i0}} = \frac{p_{i1}+p_{i2}}{2}$. Thus, for peptide j of protein i in group k, n_k zeroes or ones are generated from the Binomial distribution with probability $\hat{p}_{ijk} = \hat{p}_{ik} \times \hat{d}_{ij}$, $k = 1, 2$. We run B bootstrap iterations and compute the test statistic (4.2), T_{Mb} in each iteration. The p-value is then computed as the proportion of bootstrap test statistic values being as or more extreme as our observed T_{Mi} value:

$$\text{p-value} = \frac{\#\left\{T_{Mb} \geq T_{Mi}\right\}}{B}.$$

4.2.5 False Discovery Rate (FDR) Estimation

The false discovery rate (FDR) associated with a list of features selected at a p-value cutoff c_p [43] is the expected number of false positives F out of the total number of selected features S:

$$\text{FDR}(c_p) = \text{E}\left[\frac{F_{c_p}}{S_{c_p}}\right] \approx \frac{\text{E}\left[F_{c_p}\right]}{\text{E}\left[S_{c_p}\right]} \tag{4.3}$$

The denominator can be replaced simply with the observed number of selected features. The traditional approach to estimate the numerator is to exploit the expected uniform sampling distribution of the null p-values [43]. In particular, we can

estimate $\mathrm{E}\left[F_{c_p}\right]$ by $M \times \hat{\pi}_0 \times c_p$, where M is the total number of features and $\hat{\pi}_0$ is the estimated proportion of null features out of the total M features. However, as our test statistic is discrete, its null sampling distribution is not necessarily Uniform. As an example, Figure 4.1 shows a simulated null sampling distribution for peptide-level test statistics, in which the shape of the null sampling distribution is quite non-Uniform and could depend on many factors, including the number of peptides of a protein, the sample size of each comparison group and the overall number of observed peaks of a protein.

Peptide-level FDR Estimation

An alternative way to estimate the numerator in equation (4.3) is as a weighted summation over all peptides on their estimated null right-tail probabilities at p-value cutoff c_p:

$$\hat{\mathrm{E}}[F_{c_p}] = \sum_j w_j \widehat{\mathrm{CDF}}_{j0}(c_p)$$

where w_j is a weight (number between 0 and 1) constructed to be close to one for null peptides. If we assign all w_j with value 1, i.e. each peptide is given weight 1, the corresponding FDR estimation is called "unweighted" estimation. $\hat{\mathrm{CDF}}_{j0}$ is the estimated null p-value cumulative distribution function (CDF) for peptide j. The estimate of CDF_{j0} is derived from an estimate of the null probability mass function (PMF) for peptide j:

$$\begin{aligned}\widehat{\mathrm{PMF}}_{j0}(c_p) &= \hat{Pr}_0(T_j = T_j(c_p)) \\ &= \sum_{T_j = T_j(c_p)} Pr(Y_{j1}|\hat{p}_{j0}) \times Pr(Y_{j2}|\hat{p}_{j0})\end{aligned}$$

where $\hat{p}_{j0} = (Y_{j1} + Y_{j2})/(n_1 + n_2)$.

For the w_j, we use

$$w_j = \begin{cases} 1 & \text{if} \quad \widehat{\mathrm{PMF}}_{j0}(c_p) \geq \widehat{\mathrm{PMF}}_{j1}(c_p) \\ 0 & \text{if} \quad \text{otherwise} \end{cases}$$

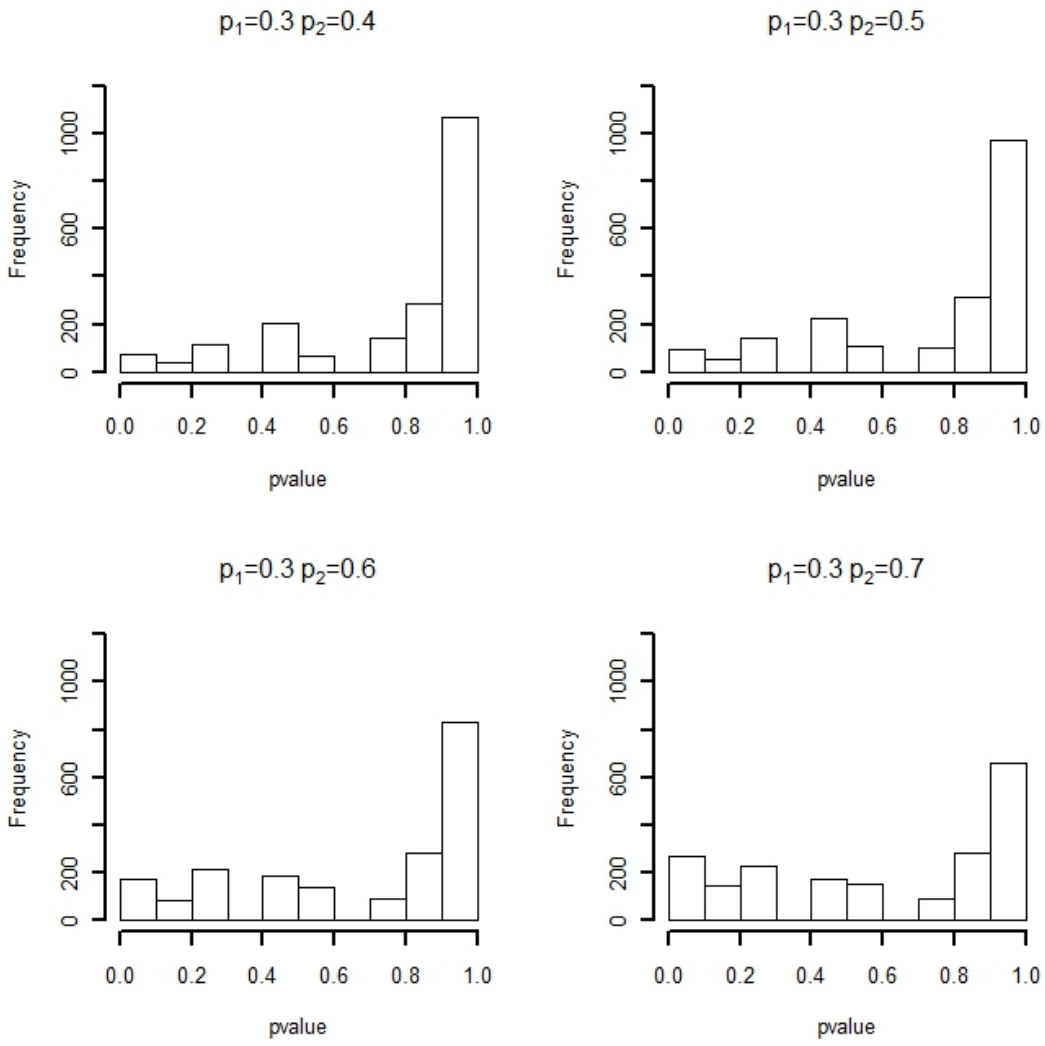

Fig. 4.1. P-value histograms of simulated null peptides with shared presence probabilities of $0.2, 0.3, 0.4, 0.5$ across each comparison group.The null sampling distribution is non-uniform, due to the discrete nature of the test statistic.

We therefore estimate the peptide-level FDR as

$$\widehat{\mathrm{FDR}}_{\mathrm{pep}}(c_p) = \frac{\sum_j w_j \widehat{\mathrm{CDF}}_{j0}(T_j(c_p))}{\#\left\{\mathrm{p-value} \leq c_p\right\}}$$

Protein-level FDR Estimation

Our simulation studies indicate that for most settings, the p-value of the test statistic for multi-peptide protein (4.2) is approximately uniformly distributed under null hypothesis setting, especially when there are moderate overall levels of presence (data not shown) and moderate number of sibling peptides in a protein. Because of this, we use the standard [43] method for FDR estimation. Namely, at p-value cutoff c_p, we estimate the FDR as

$$\widehat{\mathrm{FDR}}_{\mathrm{pro}}(c_p) = \frac{M \times \hat{\pi}_0 \times c_p}{\#\left\{\text{p-values} \leq c_p\right\}}$$

where M is the total number of proteins, and $\hat{\pi}_0$ is the estimated proportion of null proteins, which could also be seen as a uniform weight across proteins $w_j = M \times \hat{\pi}_0$ across all proteins". We estimate π_0 by fitting a smooth lowess curve to the values of $\hat{\pi}_0(\lambda) = \frac{\#\{\mathrm{p-values}>\lambda\}}{M(1-\lambda)}$ as $\lambda \to 1$ [43].

Figure 4.2 shows the histogram of p-values from bootstrap on simulated mixed null and alternative five-peptide proteins, which has a uniform shape right tail that indicates the appropriateness of using the standard method (2003) to estimate FDR.

Mixed Single-peptide and Multi-peptide Protein Level FDR Estimation

In practice, both single-peptide proteins and multi-peptide proteins are usually mixed in a data set, for which we've developed p-value and FDR estimation separately. The two sets of p-value are left as is while a unified FDR estimate needs to be generated based on the pooled set of p-values. The numerator of FDR for mixed case is given by summing up the estimation of expected number of false positive features for single-peptide protein and multi-peptide proteins, as indicated in the above two sections, and the denominator is the number of selected features based on the p-value pool.

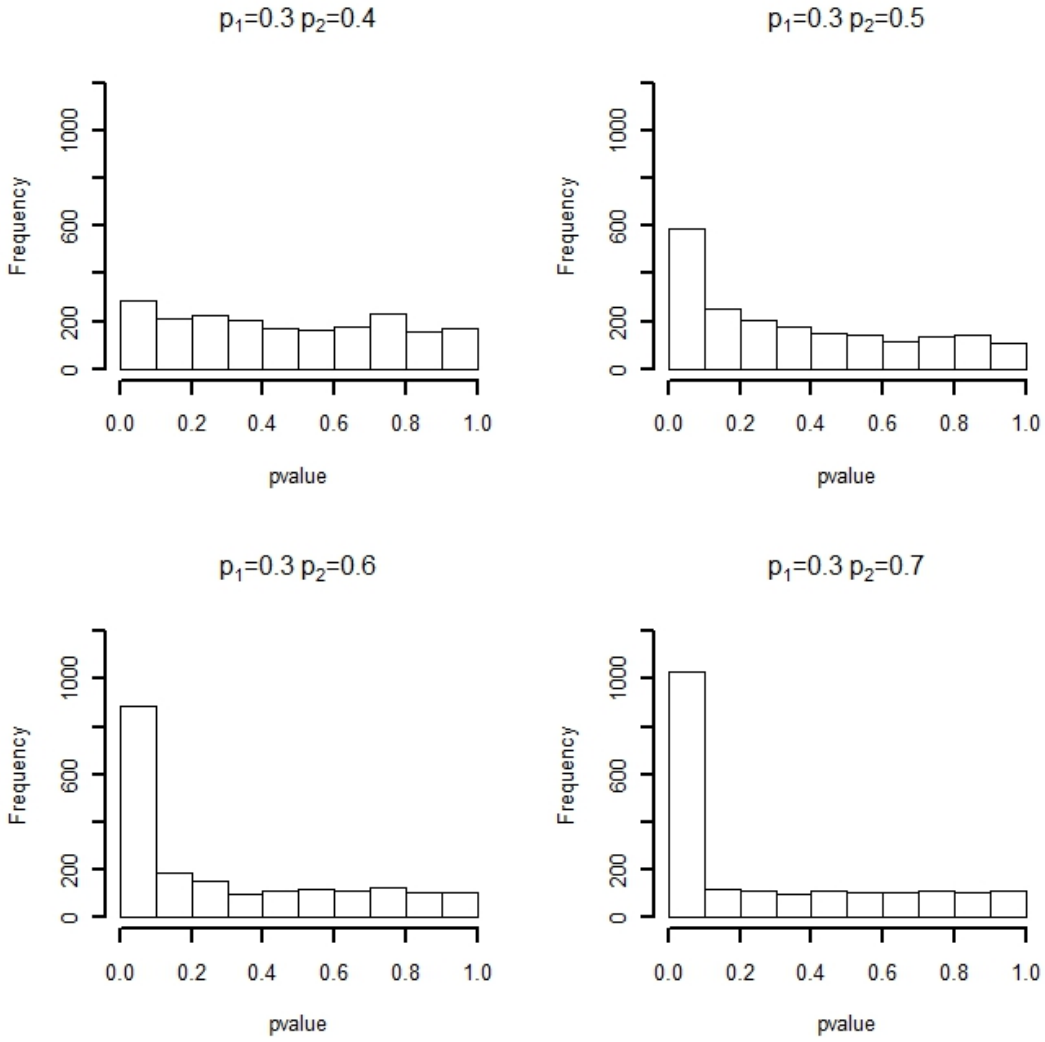

Fig. 4.2. P-value histograms of simulated null peptides with shared presence probabilities of $0.2, 0.3, 0.4, 0.5$ across each comparison group.The null sampling distribution is non-uniform, due to the discrete nature of the test statistic.

$$\widehat{\text{FDR}}_{\text{mix}}(c_p) = \frac{E(\#\{FPI_{c_p}\})}{\#\{\text{p-value} \leq c_p\}} = \frac{\sum_j w_j \widehat{\text{CDF}}_{j0}(T_j(c_p)) + M \times \hat{\pi}_0 \times c_p}{\#\{\text{p-values}_s \leq c_p \text{ or p-values}_m \leq c_p\}}$$

where p-values$_s$ are the p-values for single peptide proteins and p-values$_m$ are those for multi-peptide protein

4.2.6 Hybrid Analysis Incorporating Both Presence / Absence and Intensity Measurements

The above methodology has dealt only with presence / absence data, from which peak intensity measurements are excluded. The rationale for simplifying peak intensity measurements to presence / absence is that it better enables discovery of "one-state" (or similar) proteins. However, statistical information is lost by throwing out intensity measurements, which would translate to decreased statistical power to detect differentially expressed proteins that differ in terms of abundance but not presence / absence. Thus, we would ideally incorporate both peak intensity and presence / absence information into a differential expression analysis. One simple way to do this is to carry out separate intensity-based and presence / absence-based analysis, select proteins at a specified FDR from each analysis, then report the union of the two resulting protein lists. However, while we might intuitively expect a small FDR for the resulting list of proteins, we will not generally be able to assign an actual FDR estimate. In what follows, we derive a FDR estimate for the union list of differentially expressed proteins. Thus, taken together, the methodology presented here allows the researcher to select a list of differentially expressed proteins, some based on intensity and others based on presence / absence, to which an overall FDR estimate can be assigned.

We use a single p-value threshold c_p for both intensity measurements and presence / absence; so, a protein is selected if either of its intensity-based and presence / absence p-values are less than c_p. Intensity-based p-values are derived from regression models and censored likelihoods from our prior work [21]. Let p-value$_b$ and p-value$_p$ correspond to the binary presence / absence and peak intensity measurements, respectively. The FDR for a hybrid analysis can then be estimated by

$$\begin{aligned}\hat{\text{FDR}}_h(c_p) &= \frac{\sum_i w_i \hat{Pr}_0\left(\text{p-value}_{bi} \leq c_p \cup \text{p-value}_{pi} \leq c_p\right)}{\#\left\{\text{p-values}_b \leq c_p \text{ or p-values}_p \leq c_p\right\}} \\ &= \frac{\sum_i w_i \hat{Pr}_0\left[c_p + (1-c_p)\hat{Pr}\left(T_{M_i} \geq T_{M_i}(c_p)\right)\right]}{\#\left\{\text{p-values}_b \leq c_p \text{ or p-values}_p \leq c_p\right\}}\end{aligned}$$

We set the weight w_i equal to the average of the binary weight and the uniform weight derived from Storey and Tibishirani's FDR estimation scheme in the intensity-based method [21].

4.3 Results

4.3.1 Peptide-level Simulation Result

Figure 4.3 shows the number of significant single-peptide proteins versus FDR, based on the proposed peptide-level presence / absence methodology. The particular simulation scenario displayed in the figure has $p_1 = 0.3$, with a random mixture of differential presence / absence, ranging over $p_d = 0.1, 0.2, \ldots, 0.6$. The unweighted FDR estimate is very conservative, resulting in many fewer significant peptides at a given FDR estimate, relative to the true FDR curve. The binary weighting improves this somewhat, resulting in greater power while maintaining conservative FDR estimate. Results for other values of p_1 are similar (data not shown). Table 4.1 shows Type I error and power for the proposed presence / absence methodology applied to simulated single-peptide and 5-peptide proteins. As would be expected, we have

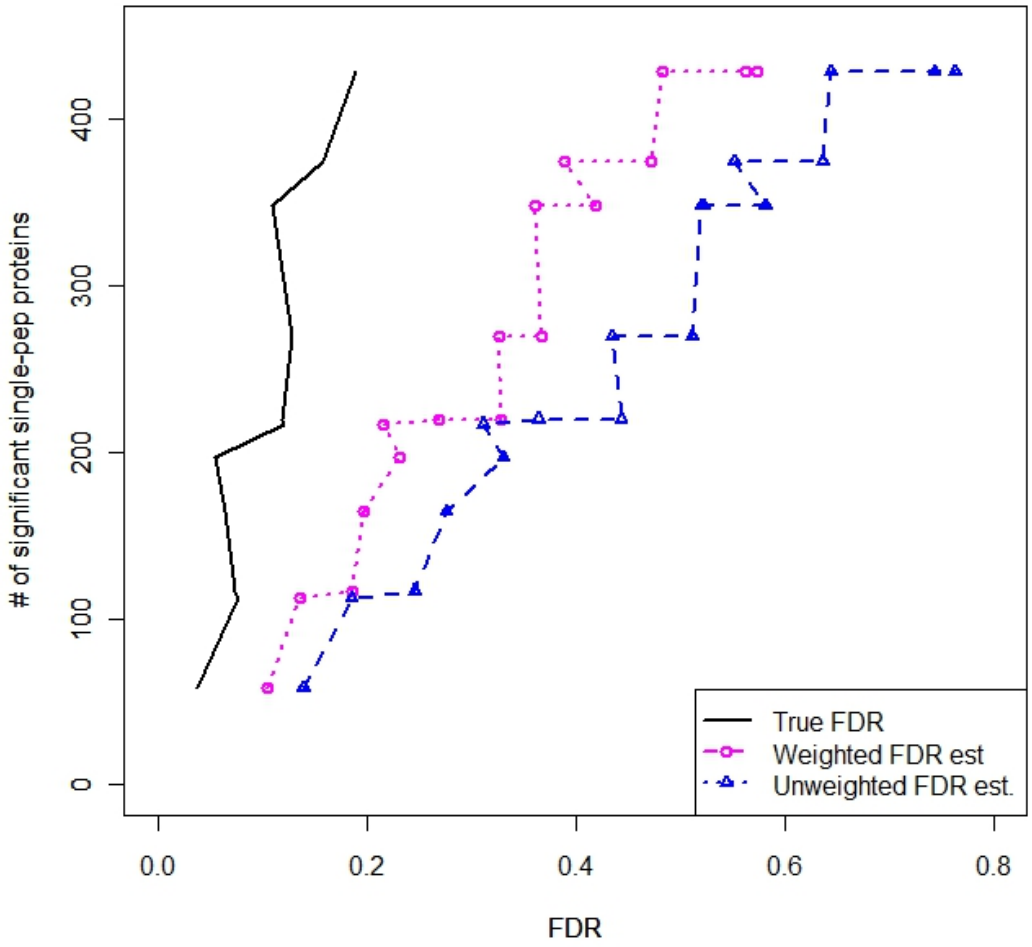

Fig. 4.3. Numbers of significant single-peptide proteins versus FDR for the proposed protein-level methodology, on simulated data with $p_1 = 0.3$ and a mixture of differential presence / absence levels. The weighted FDR estimate is conservative.

greater power to detect differential expression when there are multiple peptides in a protein.

Table 4.1
Peptide-level error rates and power with $p_1 = 0.2, 0.3, 0.4, 0.5$ and $p_d = 0.0, 0.1, \ldots, 0.7$.

$p_d = p_2 - p_1$	$p_1 = .2$	$p_1 = .3$	$p_1 = .4$	$p_1 = .5$
p_d=0.0	0.053	0.051	0.050	0.047
p_d=0.1	0.069	0.065	0.058	0.048
p_d=0.2	0.133	0.122	0.120	0.110
p_d=0.3	0.240	0.232	0.210	0.182
p_d=0.4	0.381	0.365	0.353	0.348
p_d=0.5	0.512	0.461	0.430	*
p_d=0.6	0.720	0.677	*	*
p_d=0.7	0.874	*	*	*

4.3.2 Protein-level Simulation Result

Figure 4.4 shows the number of significant five-peptide proteins versus FDR, based on the proposed protein-level presence / absence methodology. The simulation scenario in this figure is similar to that in Figure 4.3, now with each protein having 5 constituent peptides. In this case, "weighting" is carried out using the standard $\hat{\pi}_0$ estimate, again resulting in conservative FDR estimate. The pictures for different values of p_1, as well as for different numbers of constituent peptides, are not qualitatively different (data not shown).

Table 4.2 shows Type I error and power for the proposed presence / absence methodology applied to simulated five-peptide proteins. As would be expected, we have greater power to detect differential expression when there are multiple peptides

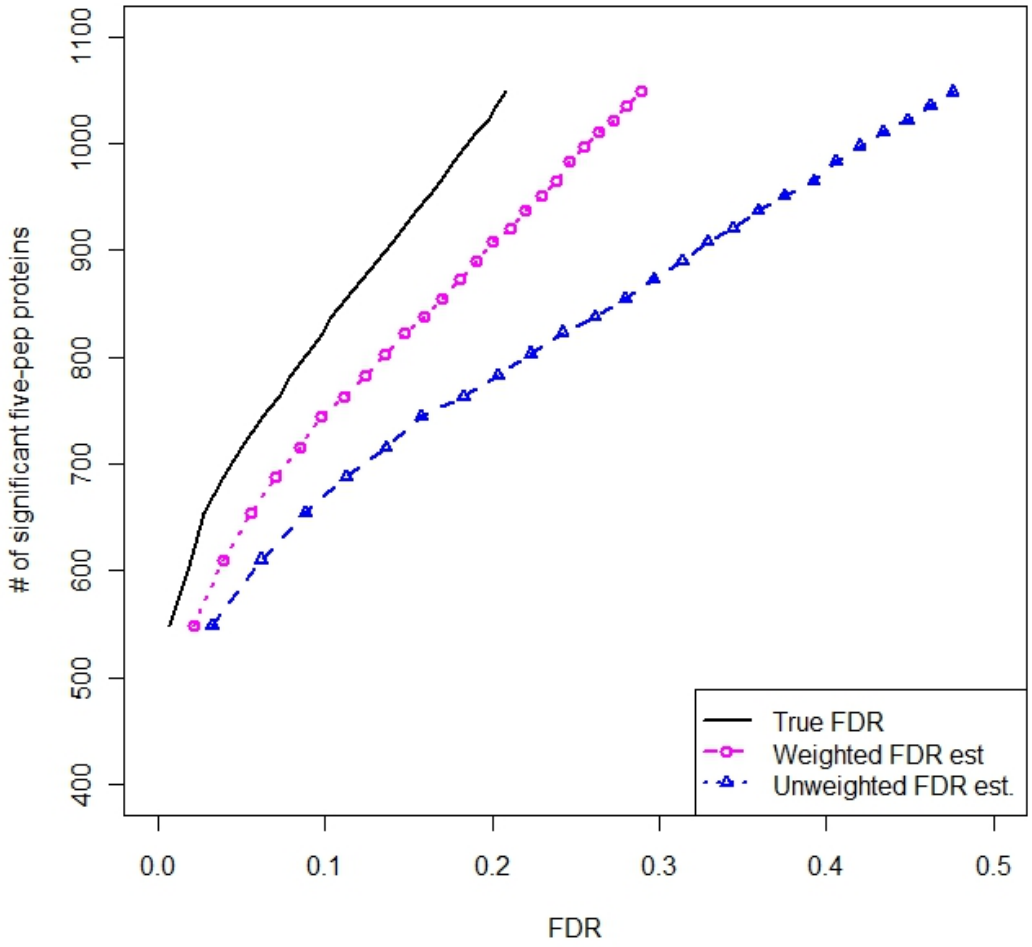

Fig. 4.4. Numbers of significant five-peptide proteins versus FDR for the proposed protein-level methodology, on simulated data with $p_1 = 0.3$ and a mixture of differential presence / absence levels. The weighted FDR estimate is conservative.

in a protein.The pool of number of peptides of a protein is $5, 10, 15, 20, 25$, the presence probability of proteins take value among $p_1 = 0.2, 0.3, 0.4, 0.5$ and $p_d = p_2 - p_1 = 0.1, \cdots, 0.9 - p_1$, peptide detectablity is set to vary among $0.9, 0.7, 0.5, 0.1, 0.01$.

Table 4.2
Protein-level error rates and power with $p_1 = 0.2, 0.3, 0.4, 0.5$ and $p_d = 0.0, 0.1, \ldots, 0.7$.

$p_d = p_2 - p_1$	$p_1 = .2$	$p_1 = .3$	$p_1 = .4$	$p_1 = .5$
	#pep=5			
p_d=0.0	0.051	0.052	0.050	0.054
p_d=0.1	0.196	0.158	0.136	0.096
p_d=0.2	0.486	0.404	0.388	0.352
p_d=0.3	0.778	0.734	0.710	0.692
p_d=0.4	0.960	0.924	0.910	0.908
p_d=0.5	0.994	0.990	0.990	*
p_d=0.6	1.000	1.000	*	*
p_d=0.7	1.000	*	*	*

4.3.3 Mixed Single-peptide Protein and Multi-peptide Protein Simulation Result

Figure 4.5 shows the number of significant mixed proteins versus FDR, based on the proposed protein-level presence / absence methodology. The simulation scenario in this figure is similar to that in Figure 4.4, now with proteins having $1, 2, \cdots, 5$ constituent peptides. In this case, "weighting" is borrowing the single-peptide weight and multi-peptide weight for the corresponding proteins, again resulting in conservative FDR estimate. The pictures for different values of p_1, as well as for different numbers of constituent peptides, are not qualitatively different (data not shown).

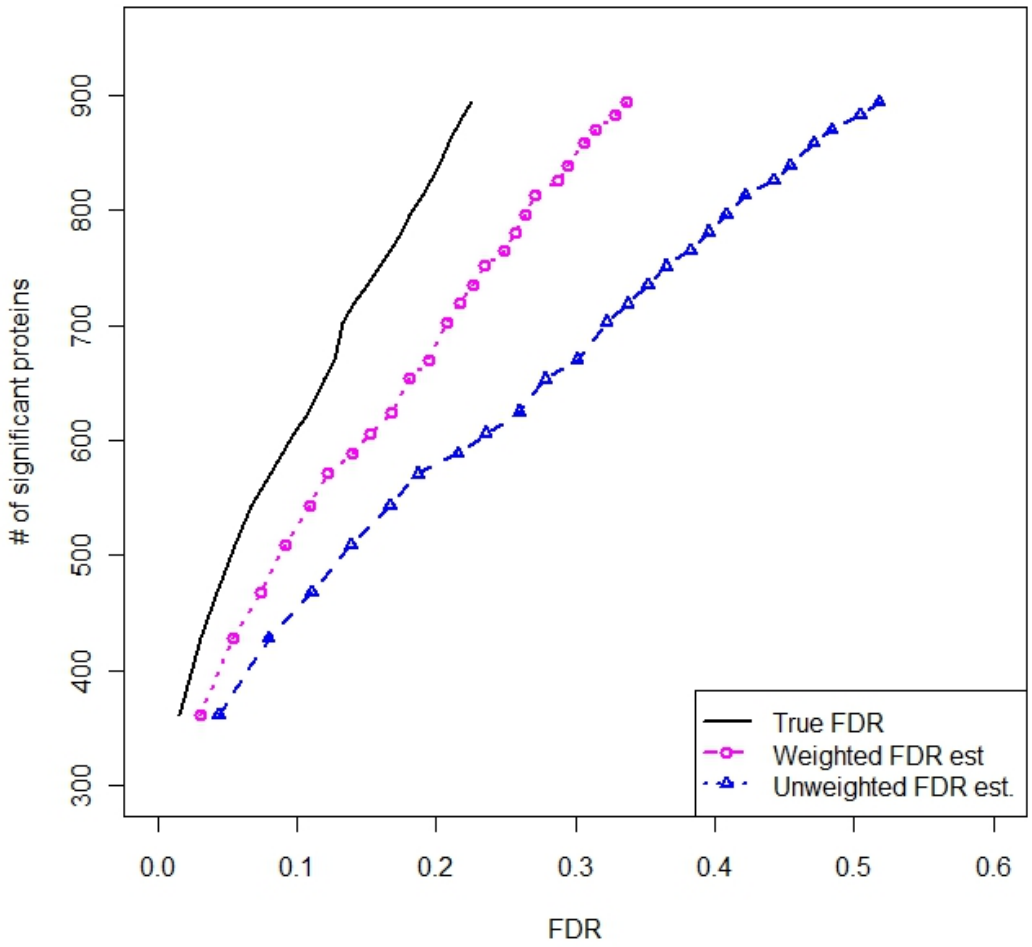

Fig. 4.5. Numbers of significant mixed proteins versus FDR for the proposed protein-level methodology, on simulated data with $p_1 = 0.3$ and a mixture of differential presence / absence levels. Constituent peptide number varies from one to five. The weighted FDR estimate is conservative.

4.3.4 Hybrid Approach Simulation Result

Our final simulation contains a combination of single- and multi-peptide proteins, a variety of differential expression magnitudes, with a sample size of 10 in each of the two comparison groups. In this case, peak intensities were simulated, from which presence / absence data were obtained, to reflect the intended real-world setting in which both intensity-based and binary presence / absence information is available. Figure 4.6 compares the proposed hybrid approach with both our "qualitative" (presence / absence-based) and previously-published "quantitative" (intensity-based) [21] methodology. All FDR estimates are conservative, and the hybrid approach results in greater numbers of significant proteins, at a given FDR, than either of the presence / absence- or intensity-based approaches. Table 4.3 lists the number of differentially expressed proteins at an estimated FDR cutoff of 0.05, for a variety of simulation settings (varying the proportion of missing data as well as the amount and magnitude of differential expression). The results are consistent with those shown in Figure 4.6. Thus, by combining a traditional intensity-based analysis with a presence / absence analysis, we are able to supplement our findings with additional proteins of interest; these would potentially include "one-state" proteins.

4.3.5 Diabetes Data

We found substantial differential expression overall in the diabetes experiment. The original data set is comprised of 177 proteins containing 1396 peptides.In binary-based method, after initial filtering, we were left with 841 peptides and all the 177 proteins, among which there are 105 multi-peptide proteins and 72 single peptide proteins. In intensity-based method, 44 proteins are filtered out. On the 133 proteins that have p-values from both methods, 76 are identified by intensity-based algorithm, 55 are identified by binary-based algorithm and 88 are identified by hybrid method, at estimated FDR 0.05. At estimated FDR 0.10, 89 are identified by intensity-based

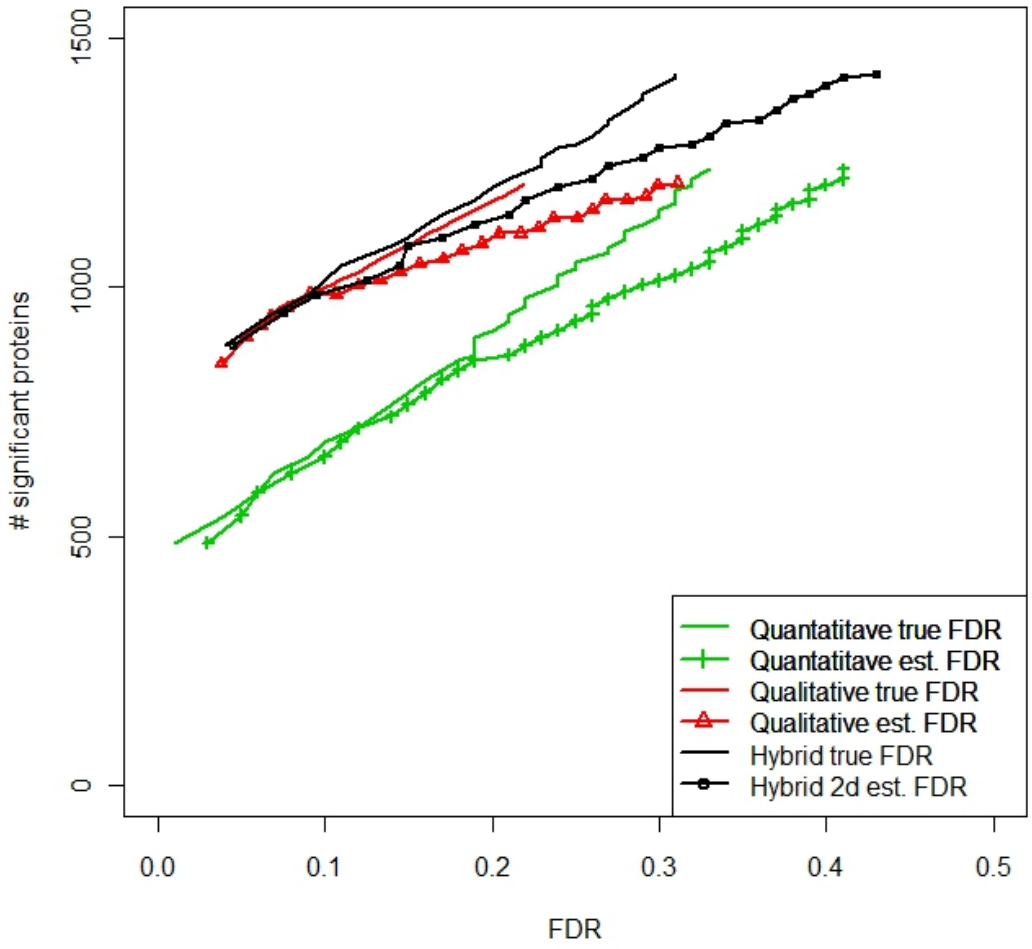

Fig. 4.6. Numbers of significant peptides versus FDR for the proposed peptide-level methodology, on simulated data with $p_1 = 0.3$ and a mixture of differential presence / absence. The weighted FDR estimate is conservative.

Table 4.3
Number of identified features at estimated FDR level of 0.05 obtained from binary-based method, intensity-based method and hybrid method under a variety of simulation settings. The hybrid approach consistently results in greater numbers of differentially expressed proteins.

Low mag.		25%	dif			50%	dif	
Method/Miss	10%	20%	30%	40%	10%	20%	30%	40%
Quantitative	313	313	166	174	697	630	488	136
Qualitative	218	263	301	299	490	587	612	641
Hybrid	358	387	326	356	743	685	657	609
High mag.								
Quantitative	422	349	281	114	1010	900	668	547
Qualitative	491	514	530	519	812	998	1011	1033
Hybrid	525	537	539	503	1036	1080	1055	1084

algorithm, 64 are identified by binary-based algorithm and 100 are identified by hybrid method

Figure 4.7 compares the proposed hybrid approach with both presence / absence-based and intensity-based methodology on the diabetes data set. The X axis is FDR estimate and the y axis is the number of identified proteins at corresponding FDR. the hybrid approach results in greater numbers of significant proteins, at a given FDR, than either of the presence / absence- or intensity-based approaches.

4.4 Discussion

The proposed presence / absence-based methodology is designed to enable the detection of "one-state" (or similar) proteins that are not amenable to traditional intensity-based methods. Furthermore, we have proposed a hybrid approach that

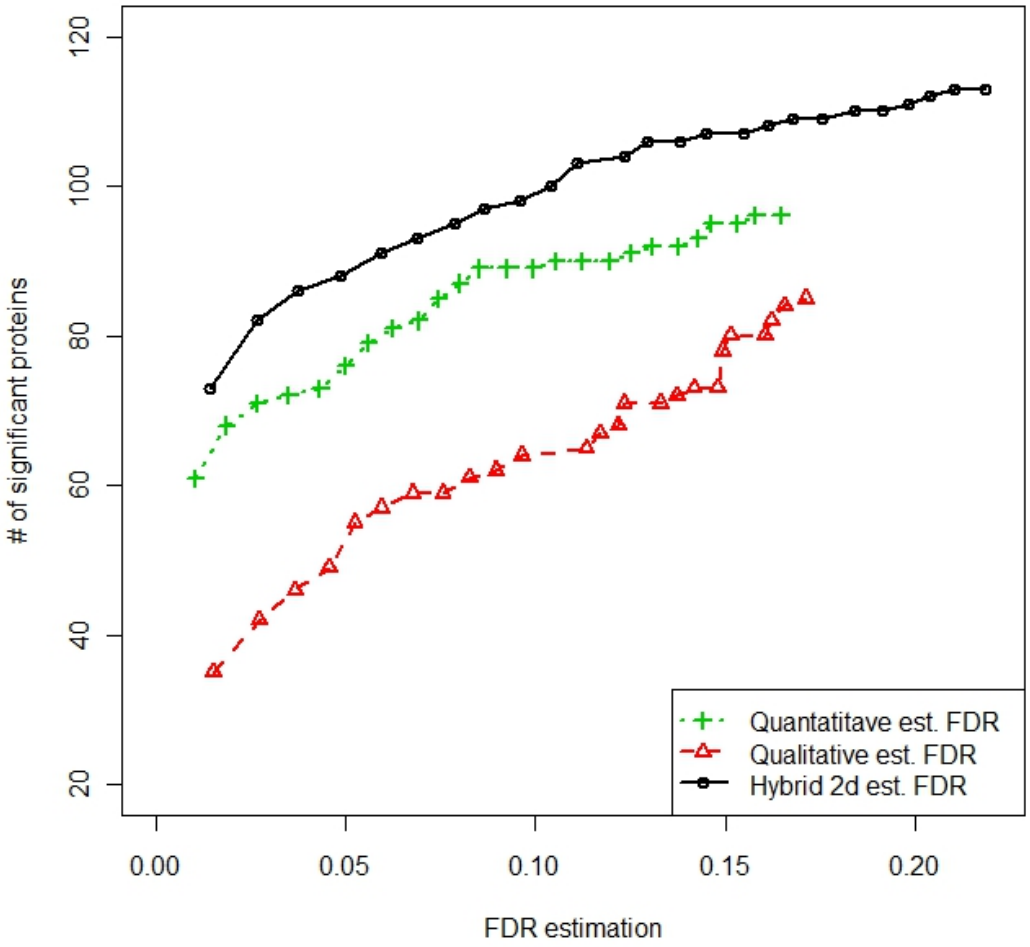

Fig. 4.7. Numbers of significant proteins versus FDR estimates on diabetes dataset by presence/absence based method, intensity based method and hybrid method

combines both intensity- and presence / absence-based analysis of a data set, together with FDR estimation of the combined list of differentially expressed proteins. The proposed hybrid approach was demonstrated to outperform either of the intensity- or presence / absence-based methods alone.

An obvious limitation to our work is its applicability to only two comparison groups. A regression-based implementation would be more generalizable, and we intend to pursue this in our future work. The choices of weights in the peptide-level and hybrid methods could undoubtedly be improved upon as well; all FDR estimates are quite conservative.

5. SUMMARY

In the Section 2, the fully Bayesian hierarchical modeling successfully identifies peptide and protein in a unified framework and detect more peptides than PeptideProphet at most of the FDR cutoff values. This is owe to modeling the top R ranked peptide spectrum matches to a MS fragmentation pattern with a collections of discriminant features, allowing for dependence among the list candidates, incorporating decoy search results to normalize target search score and selecting objective criteria for the evaluation of the FDR at peptide level.

In the Section 3, one advantage of our algorithm is that it allows statistical assessment of alignment performance on elution time, peak intensity and M/Z values utilizing highly-confidently identified peptide information located as "Anchor points". The statistical confidence of our method is measured on sample level, generalized to region level and then passed to downstream quantitative analysis. The future work is to localize statistical confidence measurement to peptide level.

In the Section 4, the proposed presence / absence-based methodology is designed to enable the detection of "one-state" (or similar) proteins that are not amenable to traditional intensity-based methods partially due to large scale of missingness that can't be solved by traditional imputation. Furthermore, we have proposed a hybrid approach that combines both intensity- and presence / absence-based analysis of a data set, together with FDR estimation of the combined list of differentially expressed proteins. The proposed hybrid approach was demonstrated to outperform either of the intensity- or presence / absence-based methods alone.

REFERENCES

[1] P. T. N. Tang and S. R. Weinberger, "Current developments in SELDI affinity technology.," *Mass Spectrometry Reviews*, vol. 23, no. 1, pp. 33–34, 2004.

[2] M. Karas, D. Bachman, U. Bahr, and F. Hillenkamp, "Matrix-assisted ultraviolet laser desorption of non-volatile compounds.," *International Journal of Mass Spectrometry and Ion Processes*, vol. 78, pp. 53–68, 1987.

[3] B. Domon and R. Aebersold, "Mass spectrometry and protein analysis.," *Science*, vol. 312, pp. 212–217, 2006.

[4] H. Zhang, E. C. Yi, X. J. Li, P. Mallick, K. S. Kelly-Spratt, C. D. Masselon, D. G. Camp, R. D. Smith, J. K. Christopher, and R. Aebersold, "High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry.," *Molecular & Cellular Proteomics*, vol. 1, pp. 144–155, 2005.

[5] J. Eng, A. Mcormack, and J. R. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society for Mass Spectrometry*, vol. 5, pp. 976–989, 1994.

[6] D. Perkins, D. Pappin, D. Creasy, and J. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, pp. 3551–3567, 1999.

[7] R. Craig and R. Beavis, "TANDEM: Matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, pp. 1466–1467, 2004.

[8] A. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, "A statistical model for identifying proteins by tandem mass spectrometry," *Analytical Chemistry*, vol. 75, pp. 4646–4658, 2003.

[9] Y. Ding, H. Choi, and A. I. Nesvizhskii, "Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics," *Journal of Proteome Research*, vol. 7, no. 11, pp. 4878–4889, 2008.

[10] J. Elias and S. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nature Methods*, vol. 4, pp. 207–214, 2007.

[11] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Analytical Chemistry*, vol. 74, no. 20, pp. 5383–5392, 2002.

[12] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble, "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases.," *Journal of Proteome Research*, vol. 7, pp. 29–34, 2008.

[13] H. Choi, D. Ghosh, and A. I. Nesvizhskii, "Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling," *Journal of Proteome Research*, vol. 7, no. 01, pp. 286–292, 2007.

[14] L. Käll, J. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss, "Semi-supervised learning for peptide identification from shotgun proteomics datasets," *Nature Methods*, vol. 4, no. 11, pp. 923–926, 2007.

[15] B. C. Searle, M. Turner, and A. Nesvizhskii, "Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies," *Journal of Proteome Research*, vol. 7, no. 1, pp. 245–253, 2008.

[16] D. Bylund, R. Danielsson, G. Malmquist, and K. Markides., "Chromatographic alignment by warping and dynamic programming as a pre-processing tool for parafac modelling of liquid chromatography-mass spectrometry data. ," *Journal of Chromatography*, vol. A 961, pp. 237–244, 2002.

[17] P. Nielsen, J. M. Carstensen, and J. Smedsgaard, "Aligning of single and multiple wave length chromatographic profiles for chemometric data analysis using correlation optimised warping.," *Bioinformatics*, vol. A 805, pp. 17–35, 1998.

[18] G. Kiebel, K. Auberry, N. Jaitly, D. Clark, M. Monroe, E. Peterson, T. Nikola, G. Anderson, and R. Smith, "PRISM: A data management system for high-throughput proteomics," *Proteomics*, vol. 6, pp. 1783–1790, 2006.

[19] D. Radulovic, S. Jelveh, S. Ryu, T. G. Hamilton, E. Foss, Y. Mao, and A. Emili, "Informatics platform for global proteomic profiling and biomarker discovery using liquid-chromatography-tandem mass spectrometry.," *Molecular & Cellular Proteomics*, vol. 3, pp. 984–97, 2004.

[20] M. Palmblad, M. Ramstrom, K. E. Markides, P. Hakansson, and J. Bergquist., "Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry.," *Analytical Chemistry*, vol. 74, pp. 5826–5830, 2002.

[21] Y. V. Karpievitch, J. S. J, T. Taverner, J. Huang, J. Adkins, C. Ansong, F. Hefron, T. Metz, W. J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney,

"A statistical framework for protein quantitation in bottom-up ms-based proteomics," *Bioinformatics*, vol. 25, pp. 2028–2034, 2009.

[22] W. M. Old, K. Meyer-Arend, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing, and N. G. Ahn, "Comparison of label-free methods for quantifying human proteins by shotgun proteomics," *Molecular & Cellular Proteomics*, vol. 4, pp. 1487–1502, 2005.

[23] A. D. Polpitiya, W. J. Qian, J. Navdeep, A. P. Vladislav, N. A. Joshua, D. G. C. II, G. A., and R. D. Smith, "DAnTE: a statistical tool for quantitative analysis of proteomics data," *Bioinformatics*, vol. 24, pp. 1556–1558, 2008.

[24] W. Wang, H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle, and C. H. Becker, "Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards," *Analytical Chemistry*, vol. 75, pp. 4818–4826, 2003.

[25] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New Jersey: John Wiley & Sons, 2002.

[26] H. Wang, W. Qian, M. Chin, V. Petyuk, R. Barry, T. Liu, M. Gritsenko, H. Mottaz, R. Moore, D. C. II, A. Khan, D. Smith, and R. Smith, "Characterization of the mouse brain proteome using global proteomic analysis complemented with cysteinyl-peptide enrichment," *Journal of Proteome Research*, vol. 5, pp. 361–369, 2006.

[27] J. Zimmer, M. Monroe, W. Qian, and R. Smith, "Advances in proteomics data analysis and display using an accurate mass and time tag approach," *Mass Spectrometry Reviews*, vol. 23, pp. 450–482, 2006.

[28] L. Pasa-Tolic, C. Masselon, R. Barry, Y.Shen, and R. Smith, "Proteomic analyses using an accurate mass and time tag strategy," *BioTechniques*, vol. 37, pp. 621–636, 2004.

[29] C. Shen, Z. Wang, G. Shankar, X. Zhang, and L. Li, "A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry," *Bioinformatics*, vol. 24, no. 2, p. 202, 2008.

[30] K. Petritis, L. Kangas, B. Yan, M. Monroe, E. Strittmatter, W. Qian, J. Adkins, R. Moore, Y. Xu, M. Lipton, D. C. II, and R. Smith, "Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information," *Analytical Chemistry*, vol. 78, pp. 5026–5039, 2006.

[31] H. Choi and A. I. Nesvizhskii, "Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics," *Journal of Proteome Research*, vol. 7, no. 01, pp. 254–265, 2007.

[32] L. Käll, J. D. Storey, M. J. MacCoss, and W. S. Noble, "Posterior error probabilities and false discovery rates: Two sides of the same coin," *Journal of Proteome Research*, vol. 7, no. 01, pp. 40–44, 2007.

[33] S. Martinez-Bartolome, P. Navarro, F. Martin-Maroto, D. Lopez-Ferrer, A. Ramos-Fernandez, M. Villar, J. P. Garcia-Ruiz, and J. Vazquez, "Properties of average score distributions of SEQUEST: the Probability Ratio method," *Molecular & Cellular Proteomics*, vol. 7, no. 6, p. 1135, 2008.

[34] A. M. Frank, N. Bandeira, Z. Shen, S. Tanner, S. P. Briggs, R. D. Smith, and P. A. Pevzner, "Clustering millions of tandem mass spectra," *Journal of Proteome Research*, vol. 7, no. 01, pp. 113–122, 2007.

[35] A. E. Gelfand and A. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990.

[36] A. Gelman, *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2004.

[37] B. Efron and R. Tibshirani, "Empirical Bayes methods and false discovery rates for microarrays," *Genetic Epidemiology*, vol. 23, no. 1, pp. 70–86, 2002.

[38] P. Wang, M. P. F. H. Tang, M. Mcintosh, M. Coram, H. Zhang, E. YI, and R. Aebersold, "A statistical method for chromatographic alignment of LC-MS data," *Biostatistics*, vol. 8 2, pp. 357–367, 2007.

[39] R. Fisher, *The Design of Experiments*. New York: Hafner Press, 1935.

[40] B. Zybailov, M. K. Coleman, L. Florens, and M. P. Washburn, "Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling," *Analytical Chemistry*, vol. 77, pp. 6218–6224, 2005.

[41] P. Mallick, M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster, and R. Aebersold, "Computational prediction of proteotypic peptides for quantitative proteomics," *Nature Biotechnology*, vol. 25, pp. 125–131, 2007.

[42] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC, 1993.

[43] J. D. Storey and R. Tibshirani, "Statistical significance for genome-wide studies," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 9440–9445, 2003.

APPENDIX A

Bayesian Implementation:

- W_i^X : a latent variable such that $P(X_i = 1) = \Phi(W_i^X > 0)$
- $W_{i,j}^{\pi}$: a latent variable such that $P(Y_j = 1|X_i = 1) = \pi_{i,j} = \Phi(W_{i,j}^{\pi} > 0)$
- $W_{k,r}^{Z}$: a latent variable such that $P(Z_{k,r} = 1|Y_{pep(k,r)} = 1) = \Phi(W_{k,r}^{Z}) > 0$
- **The first stage: a multi-component mixture model for discriminant scores**

  Sample Z_k from $1, \cdots, R_k + 1$ with probability $(p_{k,1}, \cdots, p_{k,R_k+1})$,
  where $p_{k,r} = \frac{p_{k,r}^{prior} f_1(s_{k,r}) f_0(s_{k,r})}{\sum_{r=1}^{R_k} p_{k,r}^{prior} f_1(s_{k,r}) f_0(s_{k,r}) + p_{k,0}^{prior} f_0(s_k)}$ for $r = 1, \cdots, R_k$ and $p_{k,R_k+1} = 1 - \sum_{r=1}^{R_k} p_{k,r}$. $Z_{k,r}$ is obtained directly given the sample of Z_k
- **The second stage: Pr(Z|Y)**

  If for any $r \leq R_k$, $Z_{k,r} = 1$, let $Y_{pep(k,r)} = 1$, draw a positive $W_{pro(k,r),pep(k,r)}^{\pi}$ from a truncated normal with mean μ_y and $SD = \sigma_y$ and draw a positive $W_{k,r}^{Z}$ from a truncated normal with mean μ_z and $SD = \sigma_z$.

  Else if for any $r \leq R_k$, $Z_{k,r} = 0$, sample r from $1, \cdots, R_k + 1$ with probability $Pr(pep(k,r) = 1|Z_{k,r} = 0) = \frac{Pr(Y_{pep(k,r)}=1)*(1-\tau)}{Pr(Y_{pep(k,r)}=1)*(1-\tau)+1-Pr(Y_{pep(k,r)}=1)}$ when $r \leq R_k$, and $Pr(pep(k, R_k+1) = 1|Z_{k,r} = 0) = 1 - \sum_{r=1}^{R_k} Pr(pep(k,r) = 1|Z_{k,r} = 0)$.

  - If the sampled $r \leq R_k$, draw a positive $W_{pro(k,r),pep(k,r)}^{\pi}$ from a truncated normal with mean μ_y and $SD = \sigma_y$, draw a negative $W_{k,r}^{Z}$ from the truncated normal with mean μ_z and $SD = \sigma_z$ and set the corresponding $Y_{pep(k,r)} = 1$.
  - Otherwise draw negative $W_{k,r}^{Z}$ for r in $1, \cdots, R_k + 1$ from a truncated normal with mean μ_z and $SD = \sigma_z$, set $Y_{pep(k,r)} = 0$ for all r in $1, \cdots, R_k$.

- **The third and fourth stage: Pr(Y|X)** and **Pr(X)**

  Let $j = pep(k, r)$, search for the indices of proteins I_j that could produce Y_j.

  If there is only one protein in I_j, substitute I_j with i.

  - If $Y_j = 1$, let $X_i = 1$, draw a positive W_i^X from a truncated normal with mean μ_x and $SD = \sigma_x$.
  - Else sample X_i from $0, 1$ with probability $Pr(X_i)(1 - \pi_{i,j}), 1 - Pr(X_i)$ since $P(X_i = 1|Y_j = 0) = \frac{P(X_i)(1-\pi_{i,j})}{P(Y_j=0)}$, $P(X_i = 0|Y_j = 0) = \frac{1-P(X_i)}{P(Y_j=0)}$, where $P(X_i)$ and $\pi_{i,j}$ have prior information and are updated through $W_i^{X_i}$ and $W_{i,j}^{\pi}$ in each MCMC iteration.

  If $X_i = 1$, draw a negative $W_{i,j}^{\pi}$ from a truncated normal with μ_y and σ_y, otherwise, keep $W_{i,j}^{\pi}$ un-updated

  If there are $n_j > 1$ proteins in I_j, list all the 2^{n_j} possible combinations of the proteins and their probability of generating peptide Y_j. For $1 \leq l \leq 2^{n_j}$, take C_{j_l}, the l th combination of I_j for example.

$$
\begin{aligned}
& P(X_i \epsilon C_{j_l} | Y_j = 1) \\
= \; & \frac{P(Y_j = 1, X_i \epsilon C_{j_l})}{P(Y_j = 1)} \\
= \; & \frac{P(C_{j_l}) P(Y_j = 1 | C_{j_l})}{P(Y_j = 1)} \\
\propto \; & \Big[\prod_{i \epsilon C_{j_l}} P(X_i) \prod_{i \epsilon I_j \backslash C_{j_l}} \big(1 - P(X_i)\big) - \prod_{i \epsilon C_{j_l}} P(X_i)(1 - \pi_{i,j}) \Big] \prod_{i \epsilon I_j \backslash C_{j_l}} 1 - P(X_i)
\end{aligned}
$$

$$
\begin{aligned}
& Pr(X_i \epsilon C_{j_l} | Y_j = 0) \\
= \; & \frac{Pr(Y_j = 0, X_i \epsilon C_{j_l})}{Pr(Y_j = 0)} \\
= \; & \frac{Pr(C_{j_l}) Pr(Y_j = 0 | C_{j_l})}{Pr(Y_j = 0)} \\
= \; & \frac{\prod_{i \epsilon C_{j_l}} Pr(X_i)(1 - \pi_{i,j}) \prod_{i \epsilon I_j \backslash C_{j_l}} (1 - Pr(X_i))}{Pr(Y_j = 0)} \\
\propto \; & \prod_{i \epsilon C_{j_l}} Pr(X_i)(1 - \pi_{i,j}) \prod_{i \epsilon I_j \backslash C_{j_l}} (1 - Pr(X_i))
\end{aligned}
$$

  - If $Y_j = 1$, each combination's probability conditional on peptide Y_j being present is proportional to its numerator in the above formula, based on which one combination C_{j_l} is sampled with probability $(\prod_{i \epsilon C_{j_l}} P(X_i) \prod_{i \epsilon I_j \backslash C_{j_l}} (1 - P(X_i)) - \prod_{i \epsilon C_{j_l}} P(X_i)(1 - \pi_{i,j})) \prod_{i \epsilon I_j \backslash C_{j_l}} 1 - P(X_i)$ For each protein X_i in the sampled combination C_{j_l}, set $X_i = 1$, draw positive $W_i^{X_i}$ for from truncated normal with mean μ_x and $SD = \sigma_x$. For each protein $X_i \epsilon I_j \backslash C_{j_l}$, set $X_i = 0$, draw negative $W_i^{X_i}$ for from the same distribution.
  - If $Y_j = 0$, randomly sample one combination C_{j_l} out of the 2^{n_j} combinations with probability $\prod_{i \epsilon C_{j_l}} Pr(X_i)(1 - \pi_{i,j}) \prod_{i \epsilon I_j \backslash C_{j_l}} (1 - Pr(X_i))$ For each protein X_i in the sampled combination C_{j_l}, set $X_i = 1$, draw negative $W_i^{X_i}$ for from truncated normal with μ_x and $SD = \sigma_x$, draw positive $W_{i,j}^{\pi}$ from a truncated normal with μ_y and $SD = \sigma_y$. For each protein $X_i \epsilon I_j \backslash C_{j_l}$, set $X_i = 0$, draw negative $W_i^{X_i}$ from the same truncated normal, keep $W_{i,j}^{\pi}$ unchanged.

- $q = pnorm(\mu_x)$, $\pi = pnorm(\mu_y)$, $\tau = pnorm(\mu_z)$ are assigned normal priors and are updated with the MCMC sample mean of $W_i^X, W_{i,j}^{\pi}, W_{k,r}^Z$. σ_x, σ_y, σ_z are assigned inverse-gamma prior and are updated with sample standard deviation of $W_i^X, W_{i,j}^{\pi}, W_{k,r}^Z$.

After sufficient rounds of iterations, the posterior probability of peptide j could be estimated by the sample proportion of Y_j and the posterior probability of protein i could be inferred by the sample proportion of X_i or by the posterior sample mean of $Pr(W_i^X > 0)$.

VITA

Wang was born in 1984. She majored in statistics at University of Science and Technology of China, where she obtained her Bachelor of Science in 2006. She received her Master of Science in statistics from Texas A&M University in 2008. She received her Ph.D. in statistics from Texas A&M University in May 2012. Her research interests include Quantitative Proteomics, Bayesian computing, Clinical Trial Design.

She may be reached at:
Department of Statistics
Texas A&M University
College Station, TX 77843-3143.
and her email address is wangxuan1104@gmail.com.