STATISTICAL ISSUES ON MASS SPECTROMETRY-BASED PROTEIN IDENTIFICATION AND QUANTITATION

by

SHANGBIN LIU

(Under the Direction of PAUL SCHLIEKELMAN)

ABSTRACT

The main goal of analytical proteomics is the complete and quantitative proteome analysis of species, cells, and/or tissues. Although the great success has been achieved via incremental improvements in Mass Spectrometry (MS)-based proteomics, some principal limitations make the goal of rapid, complete and quantitative proteome analysis not yet achieved. Besides further improvement in MS related machinery and technique, statistical considerations could be one of the aspects to narrow this gap, by choosing the number of replicates and analyzing the variations of factors in LC-MS/MS process. First, I propose a probability-based model that provides the probabilities of achieving a fixed coverage of sample proteins as a function of the number of replicates. With a fixed confidence level, the developed model can determine the coverage of sample proteins as a function of number of replicates. Typically, four to forty replicates are required to have a high confidence of identifying intermediate and high abundance proteins. More than 50 replicates will often be required to reliably identify low abundance proteins. Secondly, in order to analyze effects of various factors on the detection probability in LC-MS/MS process, a mathematical model was derived based on order statistics from independent non-identical normal random variables. As an approximation to the

mathematical model, a simulation approach was applied to analyze the impacts of the following factors, protein abundance, complexity of samples, proteolytic digestion efficiency, peptide separation and co-eluting peptides, scanning speed of the mass spectrometer, and dynamic exclusion efficiency, on the peptide/protein identification. The proposed simulation approach could be used as a framework for analysis of impacts of various factors on the peptide/protein detection. The simulation results provide valuable information for optimizing LC-MS/MS techniques and practical guidelines for conducting MS-based experiments. Thirdly, a methodology was developed to conduct statistical test of differential expression of proteins detected in two different samples. By combining the test results from the spectral counts and protein occurrence based methods on the basis of multiple runs of MS data, significantly differentially expressed proteins with high confidence in two different treatments can be obtained.

INDEX WORDS: Statistical application; Probability model; Mass spectrometry (MS); MS-based proteomics; Replicate; Simulation; Protein identification; Protein quantitation; Protein abundance; Protein differential expression; Proteotypic peptide; Retention time

# STATISTICAL ISSUES ON MASS SPECTROMETRY-BASED PROTEIN IDENTIFICATION AND QUANTITATION

by

SHANGBIN LIU

B.S. Northeast Forestry University, P.R. China, 1993

M.S. Northeast Forestry University, P.R. China, 1996

M.S., The University of Georgia, 2004

Ph.D, The University of Georgia, 2005

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

© 2008

Shangbin Liu

All Rights Reserved

# STATISTICAL ISSUES ON MASS SPECTROMETRY-BASED PROTEIN IDENTIFICATION AND QUANTITATION

by

SHANGBIN LIU

| | |
|---|---|
| Major Professor: | Paul Schliekelman |
| Committee: | Jaxk Reeves |
| | Lynne Seymour |
| | Yehua Li |
| | Chris Cieszewski |

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2008

# ACKNOWLEDGEMENTS

There are several to whom I owe a sincere debt of gratitude. First and foremost I would like to thank my major professor Dr. Paul Schliekelman for the constant sources of help and inspiration. His mentorship was instrumental in the successful completion of this work. Dr. Jaxk Reeves, Lynne Seymour, Yehua Li, and Chris Cieszewski were kind enough to serve on my committee and I am grateful for that. Their wisdom and critical thinking served me well. I would like to thank Dr. Orlando and colleagues for their help and advice on biological aspects of this work and kindly providing experimental data. Thanks also to Dr. Cieszewski for providing financial support during my study. Special thanks to Dr. Nico Pfeifer, Eberhard-Karls University, Germany, Dr. Edward Marcotte and colleagues, University of Texas, and Dr. Boris Zybailov, Stowers Institute for Medical Research, for their kindness of providing data for the study.

Deepest thanks also to my parents on the other side of the Earth who supported me throughout the years in my study. And finally, to my wife Shenghua, who has been there for the entirety of both of our undergraduate and graduate degrees, sharing many late nights up studying and my son, where most of my strength and encouragement come from in the latest several years.

TABLE OF CONTENTS

# LIST OF TABLES

Page

# LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION AND LITERATURE REVIEW

## 1 Introduction

Proteomics in general is a large-scale study of structures and functions of protein. The terms of proteomics and proteome were coined as analogies with genomics and genome (Wilkins *et al.* 1995). Although the great advancement has achieved in the collection and analysis of genome data, as the peak for sequencing human genome (International Human Genome Sequencing Consortium 2001, 2004), most probably the concentration of proteins and their interactions are the true causative forces in the cell. Due to the often minimal and/or limited correlation exists between gene expression measured at the mRNA level and protein expression (Gygi *et al.* 1999b; Greenbaum *et al.* 2003), the complete and quantitative proteome analysis of species, cells, and/or tissues has increasingly become the main goal of analytical proteomics (Nesvizhskii *et al.* 2007).

Mass spectrometry (MS)-based protein identification using PMF (peptide mass fingerprint) and LC-MS/MS (liquid chromatography-tandem mass spectrometry) is a widely used method in biological studies (Henzel *et al.* 1993; Mann & Wilm 1994; Yates *et al.* 1995). Over the past ten years, the protein identification has transited from analyzing one protein at a time using a technique known as Edman degradation (Edman 1950) to analyzing highly complex mixtures (Gygi *et al.* 1999a; Washburn *et al.* 2001). In common MS-based proteomic pipelines, protein samples are partially purified or separated by chromatographic or electrophoretic methods. Purified or separated protein samples are digested with trypsin (or other proteases), which results

in complex peptide mixtures. Then the peptide mixtures are separated by single or multi dimensional liquid chromatography (LC) and analyzed using tandem mass spectrometers. Peptides and proteins are subsequently identified by correlating MS spectra with a protein sequence database with the aid of specialized software (Aebersold & Mann 2003; Mallick *et al.* 2007). Over the past decade incremental improvements of MS-based proteomics have increased the resolution and reproducibility of sample separation (Heller *et al.* 2005; Malmstrom *et al.* 2006), the speed and quality of data acquisition (Domon & Aebersold 2006), and the confidence of inferring the true peptide and protein identification from MS/MS spectra (Eng *et al.* 1994; Perkins *et al.* 1999; Keller *et al.* 2002; Nesvizhskii *et al.* 2003; Haas *et al.* 2006; Elias & Gygi 2007).

In spite of the success of these approaches, some principal limitations seem difficult to overcome without new technical disciplines (Malmstrom *et al.* 2007). The limitations include extreme redundancy of LC-MS/MS spectra [meaning that the high abundant proteins are identified multiple times at the cost of missing proteins at low abundance] (Desiere *et al.* 2005; Omenn *et al.* 2005), under sampling [meaning that only a portion of peptides detectable to mass spectrometer is identified in a LC-MS/MS replicate due to the great complexity of samples and data-dependent acquisition], sample complexity [meaning that the presence of the large number of non-fully tryptic peptides after protein digestion], and saturation [meaning that the discovery rate of new proteins per spectrum added to the dataset decreases] (King *et al.* 2006). Due to these fundamental limitations, the goal of complete and quantitative proteome analysis will remain elusive (Malmstrom *et al.* 2007). To overcome at least a portion of these limitations, the target-driven quantitative proteomics is emerging on the basis of proteotypic peptides that are detectable in mass spectrometer and can uniquely identify its protein of origin (Aebersold 2003;

Kuster *et al.* 2005). The substantial influence on biomarker discovery and validation would be expected to gain from the target-driven approaches. There are two ways to obtain targeted proteotypic peptides. One is to mine the proteotypic peptides of targeted proteins from experimental proteomic databases (Desiere *et al.* 2005; Marzolf *et al.* 2006); the other is by predicting the proteotypic peptides using computational approaches (Lu *et al.* 2007; Mallick *et al.* 2007).

The rapid expansion of proteomics, whilst exciting, has brought with it many debate. Even though various technical disciplines, all of which contribute to proteomics, have been increasingly improved, MS remains a new and immature technology. Researchers have extensively applied this technology, without a consistent rule on the degree of stringency required in data generation and analysis, to significant biological studies. Consequently, some results are likely to be of questionable quality and further validation is needed (Wilkins *et al.* 2006). Replicate technique is generally used to improve the comprehensiveness and accuracy of analyses, the achievement of which to some extent was shown in Zybailov *et al.* (2006). Theoretically, fitting sampling process with a statistical model allows the number of experiments necessary to reach a reasonable level of completeness for sample proteins at a certain abundance level to be predicted (Liu *et al.* 2004). In the first part of the dissertation, I propose a probability-based model that allows predicting the number of experiments required to reach a certain level of completeness for sample proteins at a certain level of abundance. Distinct from the model used by Liu *et al.* (2004), which determined only the expected number of different protein species identified at an abundance level after several repeated runs, my probability-based model provides the probabilities of achieving a given level of completeness as a function of the number of replicates. It could also be used to determine the level of completeness for sample proteins at a

certain abundance level as a function of number of replicates at a fixed confidence level, which would provide some practical guidelines in planning MS-related experiments.

The process of spectrum detection from the LC-MS/MS depends on a multitude of factors, such as protein abundance (Aebersold & Mann 2003), proteolytic digestion efficiency (Gatlin *et al.* 2000; Kjeldsen *et al.* 2003), peptide separation and co-eluting peptides (Breci *et al.* 2003; Nielsen *et al.* 2004; Ocaña *et al.* 2005), peptide ionization efficiency, ion suppression, scanning speed of the mass spectrometer, and dynamic exclusion efficiency (Dobo & Kaltashov 2001; Peschke *et al.* 2002; Pan & McLuckey 2003; Pan *et al.* 2004), which are difficult to control (Aebersold & Mann 2003; Liu *et al.* 2004). In the second part of the dissertation, with some necessary assumptions I derive an analytical model on the basis of order statistics from independent non-identical normal random variables to capture the variations of factors in LC-MS/MS process. As an approximation to the analytical model, I use a simulation approach to analyze impacts of the above factors on the peptide/protein identification. The proposed simulation approach can be used as a framework for analysis of impacts of various factors on the peptide/protein detection. The simulation results provide valuable information for optimizing LC-MS/MS techniques and practical guidelines for conducting MS-based experiments.

Detection of complex protein mixtures is often not comprehensive. Under sampling (or analytical incompleteness in Wilkins *et al.* 2006) leads to a phenomenon where only a fraction of the peptides is identified in single replicate, which results in limited overlap of identified proteins in repeated analysis of the same or similar biological sample, i.e. differences in proteins detected per replicate (Durr *et al.* 2004; Liu *et al.* 2004). This phenomenon has a strong impact on the application of LC-MS/MS to protein differential expression and biomarker discovery (Rifai *et al.* 2006; Smit *et al.* 2007), as the presence or absence of a protein in a specific replicate may be the

result of under sampling instead of true differences between samples (Wilkins *et al.* 2006). Consequently, a simple list of proteins detected in the different states is insufficient to make the analysis of differential expression meaningful (Aebersold & Mann 2003). In the third part of the dissertation, I propose a methodology to conduct statistical test of levels of differentially expressed proteins detected in two samples using data from repeated MS experiments.

## 2 Literature review

The terms of proteomics and proteome were coined as analogies with genomics and genome (Wilkins *et al.* 1995). A main goal of analytical proteomics is the complete and quantitative proteome analysis of species, cells, and/or tissues (Nesvizhskii *et al.* 2007). The field of proteomics has grown rapidly, shows no sign of slowing. Although the great success has been achieved via incremental improvements in MS-based proteomics (Aebersold & Mann 2003), some principal limitations make the goal of rapid, complete and quantitative proteome analysis not yet achieved (Malmstrom *et al.* 2007). This review is structured to capture the advances in MS-based proteomics, in which the current limitations and the future direction likely taken discussed. Special emphases are placed on the elaboration of results supported by sound statistical arguments.

### 2.1 Overview of MS-based proteomics

Analytical protein identification is built around on the essential fact: most six or more amino acid peptides are largely unique in the proteome of any organism (Liebler 2002). Therefore, if we could obtain the accurate peptide mass or the peptide sequence, we could identify its protein of origin by finding its match in a protein sequence database.

A typical proteomics experiment consists of five stages. 1) Proteins from a specific organism are extracted and then partially purified or separated using chromatographic or electrophoretic

methods to reduce the complexity of the mixture; 2) Purified or separated protein samples are digested by proteolytic enzymes, typically trypsin. This process results in a highly complex peptide mixture; 3) Peptide mixtures are separated by LC or tandem LC prior to mass analysis; 4) Separated peptides are then analyzed using certain type of mass spectrometer and produce MS or tandem MS spectrum; 5) Peptides and proteins are subsequently identified by searching a protein sequence database with the aid of specialized software (e.g. MASCOT, SEQUEST). These five stages can be classified into three broad categories: sample preparation (stage 1, 2, and 3), data acquisition (stage 4), and data analysis (stage 5). Steen and Mann (2004) provided a review on the principles of peptide sequencing, in which they elaborated the fundamental issues on sample preparation and fractionation, protein digestion, peptide separation, peptide ionization, peptide fragmentation, how peptides are identified against database searching and how to validate the peptide hit.

**2.2 Advances in sample preparation**

In the case of analyzing highly complex protein mixture, the proteins are fractionated by 1-D gel electrophoresis to reduce the complexity of the mixture. In 1-D gel electrophoresis, the proteins are resolved into bands in order of molecular weight when the gel is subject to high voltage. Each band typically contains dozens to hundreds of different proteins due to the rather modest degree of resolution achieved by 1-D gel electrophoresis. The whole lane of the gel is excised into equally sized slices and each slice can be analyzed separately (Lasonder *et al.* 2002). The known molecular weight of the protein can be used to validate the protein identification from the database searching algorithms, which increases the confidence of protein identification (Schirle *et al.* 2003).

MS analysis of whole proteins is less sensitive than that of peptides. In addition, the mass of the intact protein by itself is insufficient for identification (Aebersold & Mann 2003). Therefore, proteins must be cleaved into peptides prior to MS analysis. This is generally done with proteolytic enzymes (protease). The most widely used protease in analytical proteomics is trypsin. Trypsin cleaves the intact protein into peptides at very specific amino acids residuals of lysine (K) or arginine (R), unless either of these is followed by a proline (P) residual on the C-terminal direction. Generally, a typical 50 kDa protein will yield about 30 tryptic peptides (Liebler 2002). The ideal length of peptide fragments for MS analysis and database comparisons is around 6 - 20 amino acids (Liebler 2002, Steen & Mann 2004).

Peptide mixtures are separated by LC prior to introducing to mass spectrometer. Peptides are eluted according to certain physicochemical properties in LC column. For example, in reverse phase (RP) LC, peptides are stuck to columns and eluted from columns with high organic mobile phase. By pumping a liquid at a linear gradient of the organic solvent at high pressure through the column, peptides are separated on the basis of their hydrophobic character. To get a better separation, peptide mixtures can be separated in the combination of LC separation modes (referred to as tandem LC or LC/LC), which facilitate the identification of proteins present at low abundance in the mixture. For example, a strong cation exchange (SCX) column back-to-back with RP column is used in a technique known as multidimensional protein identification technology [MudPIT] (Washburn *et al.* 2001). The MS can obtain data on a greater fraction of the components by further "spreading out" the peptide mixture using the tandem approach. In addition, isoelectric focusing (IEF) techniques in gels and in solutions using *p*I information (Malmstrom *et al.* 2006) and multiplexed peptide IEF (Heller *et al.* 2005) have been described to improve the resolution of peptide separation. Furthermore, the advances in microcapillary

chromatography methods using particle or monolithic columns (Premstaller *et al.* 2001; Shen *et al.* 2005) and microfluidic chips (Yin *et al.* 2005) have also significantly improved data quality. As an alternative strategy for protein quantitative analysis, the stable isotope labeling, such as chemical, enzymatic, and biosynthetic labeling (Gygi *et al.* 1999a; Ong *et al.* 2002; Ross *et al.* 2004; Schmidt *et al.* 2005), has been widely used to collect quantitative protein data on a large scale over the past several years.

**2.3 Advances in data acquisition**

While one protein at a time was analyzed by Edman degradation (Edman 1950) before early 1990s, highly complex protein mixtures can be analyzed by modern MS-based methods by now (Gygi *et al.* 1999a; Washburn *et al.* 2001). Mass spectrometers consist of three essential parts – source, mass analyzer, and detector. The source produces ions from the components of a mixture. Then the mass analyzer resolves ions on the basis of their mass-to-charge (m/z) ratio with an external magnetic or electric field. Finally the detector detects resolved ions and generates the measurable signals.

Electrospray ionization [ESI] (Fenn *et al.* 1989) and matrix assisted laser desorption/ionization [MALDI] (Karas & Hillenkamp 1988) is two techniques most widely used to volatize and ionize the peptides for MS analysis. The former technique earned its inventor a share of the Nobel Prize for chemistry in 2002. In general, three main types of mass analyzers, time of flight (TOF), quadrupole, and ion trap, are used in proteomics. In order to tackle the challenges in analytical proteomics problems, new mass analyzers providing high mass resolution and accuracy have been increasingly developed. These include TOF-TOF (Medzihradszky *et al.* 2000), Q-TOF (Morris *et al.* 1996), FT-ICR (Marshall *et al.* 1998; Martin *et al.* 2000; Syka *et al.* 2004), and orbitrap (Makarov 2000), and other multistage instruments.

Characteristics and performances of these mass spectrometers have been reviewed by Domon and Aebersold (2006). In short, MALDI-TOF and ESI-MS/MS are two commonly used mass spectrometers. MALDI-TOF provides peptide masses, while ESI-MS/MS produces peptide ion fragmentation (Henzel *et al.* 1993; Mann & Wilm 1994; Yates *et al.* 1995). Several literatures have showed that these two ionization methods ionize differently but with overlapping detectable peptides (Bodnar *et al.* 2003; Elias *et al.* 2005; Mallick *et al.* 2007). Therefore, combinations of different types of sources and mass analyzers have been adopted broadly to increase the proteome coverage in analytical proteomics (Malmstrom *et al.* 2007).

**2.4 Advances in data analysis**

Two types of MS data, peptide masses and peptide sequence, are generated using different mass spectrometers. When MS is used to measure the peptide masses, the protein identification technique is referred to as the peptide mass fingerprinting (PMF). For PMF, the peptides are identified by matching the measured peptide masses to corresponding *in silico* digested peptide masses from protein or nucleotide sequence databases. For MS/MS spectra, there are three ways to identify peptides and proteins: *de novo* sequencing, database searching, and hybrid approaches. The details of these three methods are as follows.

**Spectral identification by *de novo* sequencing**. The *de novo* sequencing approach interprets the spectrum to obtain a peptide sequence [e.g. PEAKS (Ma *et al.* 2003)] followed by BLAST searching of the sequence against protein sequence database to identify the protein. Unfortunately, the success of *de novo* interpretation/BLAST searching approach depends crucially on the quality of data, in terms of both the mass accuracy and the resolution of the instruments (Steen & Mann 2004).

**Spectral identification by database searching**. To circumvent the above problem, multiple database searching algorithms have been developed since the early 1990s. In general, the database searching approach directly correlates MS/MS spectral data with peptide sequences in protein database without explicitly interpreting MS/MS spectra. The pool of candidate peptides is subject to the user-specified criteria such as mass tolerance, proteotypic enzyme used and types of post-translational modifications allowed. The best match or matches (ranked according to the search scores) for each MS/MS scan analyzed is then reported. The algorithms include MASCOT (Perkins *et al.* 1999), SEQUEST (Eng *et al.* 1994; MacCoss *et al.* 2002), ProbID (Zhang *et al.* 2002; Zhang *et al.* 2005), TANDEM (Craig & Beavis 2004), OMSSA (Geer *et al.* 2004), PEP_PROBE (Sadygov & Yates 2003), SALSA (Hansen *et al.* 2001), and others (see Table 1 in Nesvizhskii *et al.* 2007 for a list of publicly available tools). These algorithms adopt different scoring methods, including cross-correlation method (for example, SEQUEST), shared fragment counts and dot product method (for example, TANDEM, OMSSA, MASCOT), Bayesian approach (for example, ProbID), and hypergeometric model (for example, PEP_PROBE). Regardless of the scoring method, the peptide and protein identification process remains unreliable and further validation is required. To address this problem, statistical models and strategies have been developed to validate peptide (Keller *et al.* 2002; Sadygov & Yates 2003) and protein identification (Nesvizhskii *et al.* 2003; Sadygov *et al.* 2004; Adamski *et al.* 2005; Weatherly *et al.* 2005), to assess false positive rate (Moore *et al.* 2002; Peng *et al.* 2003; Yu *et al.* 2004; Qian *et al.* 2005; Haas *et al.* 2006; Elias & Gygi 2007) by searching reversed, shuffled, or randomized sequence databases. PeptideProphet employs empirical Bayes approach in conjunction with the expectation maximization algorithm to learn and calculate probabilities for correct peptide identification based upon database search scores and the number of tryptic

termini of peptides (Keller *et al.* 2002). As its sibling, ProteinProphet computes probabilities of true protein identification on the basis of peptide probabilities calculated from PeptideProphet or other algorithms (Nesvizhskii *et al.* 2003). PEP_PROBE and PROT_PROBE are another pair of tools for peptide and protein validation. PEP_PROBE utilizes a hypergeometric distribution to model frequencies of matches between an experimental tandem mass spectrum and predicted fragment ions from a protein sequence database (Sadygov & Yates 2003). PROT_PROBE further applies two independent models, binomial and multinomial models, to generate protein identification score using the hypergeometric probabilities calculated from PEP_PROBE and cross-correlation scores (independent of PEP_PROB), respectively. The combination of the two independent methods provides a useful tool for protein identification (Sadygov *et al.* 2004). A Poisson model with two-peptide hits was used to assemble a minimal and representative set of protein identifications in Human Plasma Proteome Project, which is a part of the Human Proteome Organization (HUPO). Recently, a semi-random sampling model (Xue *et al.* 2006) has been developed to evaluate these various methods in large-scale datasets. The simulated results from human liver samples showed that PROT_ PROBE is a more efficient method with higher specificity (Xue *et al.* 2006).

Algorithms discussed above for identifying and validating peptides and proteins use information only from the final spectrum, ignoring non-mass-based information acquired routinely in LC-MS/MS analyses. The use of auxiliary information provides information independent of the information contained in the MS/MS spectrum. Therefore, the auxiliary information can be used in conjunction with the MS/MS spectrum to increase both the number of protein identifications and their confidence. Peptide chromatographic retention time is one of such auxiliary properties. Over the past several years, models have been developed to predict

retention time and then use predicted retention time to improve peptide identification (Petritis *et al.* 2003; Petritis *et al.* 2006; Krokhin 2006; Klammer *et al.* 2007; Pfeifer *et al.* 2007). Petritis *et al.* (2003) developed a peptide retention time predictor using an artificial neural network (ANN). This simpler form of predictor has further been modified by incorporating peptide sequence information such as peptide length, sequence, hydrophobicity, hydrophobic moment, and nearest-neighbor amino acid (Petritis *et al.* 2006). The ANN-based predictor is the most accurate and sophisticated peptide retention time predictor up to date. In spite of the success of this predictor, large amount of data required to train the ANN (345000 non-redundant peptides for Petritis *et al.*) limited its uses for new chromatography conditions. A sequence-specific retention calculator (SSRCalc) is another peptide retention time predictor requiring large training datasets (Krokhin 2006). It showed that amino acid composition, position of the amino acid residues (N- and C-terminal), peptide length, overall hydrophobicity, *p*I, nearest neighbor effect of charged side chains (K, R, H), and propensity to form helical structures are highly correlated with the protein retention time in RP LC. Peptide retention time predictors discussed above are all restricted to conditions (e.g., column, mobile phase, and gradient) identical to those used to train it, making a static retention time predictor impractical. A dynamically trained support vector regressor (SVR) has been described to predict peptide retention time in a given LC-MS/MS analysis, using only data generated during the current run (Klammer *et al.* 2007). The SVR is portable to new conditions, but the performance may be under the ANN and SSRCalc when models are trained and tested under highly similar conditions. An approach well-suited for the prediction of peptide retention time while requiring only very small training data (about 40 peptides instead of thousands) has been developed (Pfeifer *et al.* 2007). It is a new kernel function that can be applied in conjunction with support vector machines (SVM) to

computational proteomics problems. Both SVR and kernel-SVM predictor require a small collection of data from a single LC run for training and testing predictors. Besides the peptide retention time, mass accuracy (Strittmatter *et al.* 2003) and *p*I value (Cargile *et al.* 2004; Malmstrom *et al.* 2006) were also used to improve the confidence of peptide and protein identification.

**Spectral identification by hybrid approaches**. Hybrid approaches for spectral identification refer to the combination of components of both *de novo* sequencing and database searching, i.e. starting on an extraction of short sequence tags of 3 – 5 residuals in length by *de novo* sequencing, followed by an error-tolerant database searching (Mann & Wilm 1994; Tabb *et al.* 2003; Tanner *et al.* 2005).

To identify proteins for which the exact sequences are not present in the searched sequence database, a more effective strategy may be to start with database searching, and apply *de novo* sequencing tools to the remaining unassigned high quality spectra (Nesvizhskii *et al.* 2006).

### 2.5 Principal limitations and new emerging technology

In spite of the success of MS-based approaches on protein identification, some principal limitations seem difficult to overcome without new technical disciplines (Malmstrom *et al.* 2007). The limitations include extreme redundancy of LC-MS/MS spectra (Desiere *et al.* 2005; Omenn *et al.* 2005), under sampling, sample complexity, and saturation (King *et al.* 2006). Due to these fundamental limitations, the goal of complete and quantitative proteome analysis will remain elusive (Malmstrom *et al.* 2007).

The target-driven quantitative proteomics is emerging on the basis of proteotypic peptides that are detectable to mass spectrometer and can uniquely identify its protein of origin (Aebersold 2003; Kuster *et al.* 2005). The substantial influence on biomarker discovery and

validation would be expected to gain from the target-driven approaches. There are two ways to obtain targeted proteotypic peptides. One way mines the proteotypic peptides of targeted proteins from experimental proteomic databases (Desiere *et al.* 2005; Marzolf *et al.* 2006); the other predicts the proteotypic peptides from computational approaches (Lu *et al.* 2007; Mallick *et al.* 2007).

MS-based approaches typically assume that different peptide fragments in a protein are detected with equal probability. However, a few peptides so-called proteotypic peptides are repeatedly and consistently identified for any protein present in a mixture using a particular proteomic platform (Mallick *et al.* 2007), and these peptides are identified with much higher probabilities than other peptides in a protein (Lu *et al.* 2007). Mallick *et al.* (2007) empirically identified more than 16000 proteotypic peptides for 4030 yeast proteins starting from more than 600000 peptide identifications. The empirical rule for proteotypic classification is that a peptide was classified as proteotypic if detected in >50% of all identifications of the corresponding protein. Characteristic physicochemical properties of these peptides (most related to charge and hydrophobicity) were then used to develop a computational tool that can predict proteotypic peptides for any given organism and instrument platform. Lu *et al.* (2007) described an approach that uses statistical inference to estimate the expected number of unique peptides for any protein and to incorporate information on the repeated sampling of spectra from each protein in a shotgun proteomics experiment. One of the most attractive uses of proteotypic peptides may be to use the proteotypic-to-detected ratio (i.e. ratio between the number of such proteotypic peptides and the number of peptides detected in a proteomics experiment) as an index of protein abundance (Bergeron & Hallett 2007).

## 2.6 Future directions

Complete protein identification and its corresponding abundance level are important issues in the field of proteomics. Further direction in this field would focus on them. The emerging of proteotypic peptides annotation makes people pay more attention to the easily detected peptides that would uniquely represent their source protein. More profoundly, a list of proteotypic peptides fit a species would collectively define a proteome – the proteotypic proteome.

# CHAPTER 2

# STATISTICAL POWER IN PROTEOMICS EXPERIMENTS AND THE RELATIONSHIP BETWEEN REPLICATION AND PROTEIN COVERAGE[*]

[*] Liu S, Orlando R, and Schliekelman P. To be submitted to Journal of Proteome Research.

**Abstract**

There has been rapid technological progress in methods for large scale proteomic analyses. This has led to intense interest in the use of proteomics methods for the identification of biomarkers. Successful biomarker discovery requires that researchers can determine with confidence which proteins are present in a sample. This requires that researchers have some idea of the statistical power achieved in their experiment. That is, they must know the probability that a protein present in the sample was actually detected. However, little attention has been paid to this question in the proteomics literature. Because of the time and expense of each run, the number of replicates is typically dictated by availability of resources and not statistical considerations. However, in order for biomarker discovery to be a rigorous enterprise it must rest on sound statistical practice. In this paper we propose a probability-based model that provides the probabilities of achieving a fixed coverage of sample proteins as a function of the number of replicates. With a fixed confidence level, the model developed can determine the coverage of sample proteins as a function of number of replicates. This will provide practical guidelines in planning MS-related experiments. Typically, four to forty replicates are required to have a high confidence of identifying intermediate and high abundance proteins. More than 50 replicates will often be required to reliably identify low abundance proteins.

**Keywords:** Probability model; Mass spectrometry (MS); MS-based proteomics; Replicate; Protein identification; Protein abundance; Proteotypic peptide

## 1 Introduction

Proteomic analysis of complex protein mixtures using LC-MS/MS (liquid chromatography-tandem mass spectrometry) has become a widely used method in recent years. The ability to identify proteins in complex protein mixtures has led to intense interest in biomarker discovery

using these methods. The goal is to identify differences in the profile of expressed proteins that are characteristic of diseases or other biological states.

In order to do this, researchers must be able to reliably identify which proteins are present in a sample. It is well understood that a single LC-MS/MS run will only identify a certain percentage of the proteins in a sample (Liu *et al.* 2004). Many factors could cause the absence of peptides from the list of identified peptides. Besides the influence of protein abundance (Aebersold & Mann 2003), factors such as proteolytic digestion efficiency (Gatlin *et al.* 2000; Kjeldsen *et al.* 2003), peptide separation and co-eluting peptides (Breci *et al.* 2003; Nielsen *et al.* 2004; Ocaña *et al.* 2005), peptide ionization efficiency, ion suppression, scanning speed of the mass spectrometer, and dynamic exclusion efficiency (Dobo & Kaltashov 2001; Peschke *et al.* 2002; Pan & McLuckey 2003; Pan *et al.* 2004) influence the peptide identification. Thus, multiple replicate runs are required to achieve acceptable sample coverage.

Because each run is costly and time consuming, the amount of replication in LC-MS/MS experiments is typically very low and is determined primarily by considerations of available resources. However, sound statistical practice requires that an experiment be designed with some regard to the expected probability of detecting a protein that is present in a sample. Since the goal of biomarker discovery is finding proteins that are present in one treatment condition and absent in another, then we must be able to say with confidence whether a protein actually is present in a sample or not.

In order to improve the comprehensiveness and accuracy of protein analyses, biological replication is needed. However, there is currently little information available to guide researchers in choosing the number of replicates needed to reach a given level of coverage of proteins in a sample. Of course, we can not ignore the limitations caused by practical factors such as the cost

of experiment, the availability of samples, etc. However, statistical considerations should be the starting point in choosing the number of replicates. Fitting the sampling process with a statistical model allows the number of experiments necessary to reach a reasonable level of coverage for sample proteins at a certain abundance level to be predicted (Liu *et al.* 2004).

The capture-recapture model is used in ecology and wildlife fields to estimate population sizes (Otis *et al.* 1978; Seber 1982). Applying the capture-recapture model with a closed population and time-varying and heterogeneous individual probabilities of capture, Koziol *et al.* (2006) showed how to estimate the total number of proteins in multidimensional protein identification technology (MudPIT) experiments. They also provided some practical guidelines on selecting the necessary replicates for planning MudPIT experiments. Although this is a clever application of capture-recapture, there are three shortcomings in the application of the capture-recapture method for solving this problem: 1) It does not take advantage of the knowledge of unobserved peptides that are known to be in the sample because other peptides of the same protein are observed; 2) It ignores potential information by operating at the protein level instead of peptide level; and 3) It appears that it may not be able to handle the vast range of variation in abundances (i.e. the results seem questionable for samples with different abundant proteins).

We propose a probability-based model that allows predicting the number of experiments required to reach a certain level of coverage for sample proteins at a certain abundance level. Distinct from the model used by Liu *et al.* (2004), which determined only the expected number of different protein species identified at an abundance level after several repeated runs, the probability-based model provides the probabilities of achieving a given level of coverage for different replicates. It can be used to determine the level of coverage for sample proteins at a

certain abundance level for different replicates at a fixed confidence level. This will provide practical guidelines in planning MS-related experiments.

## 2 Model development

### 2.1 The simplest case: equal likely observed peptides

For the convenience of the model derivation, first assume that peptide fragments in a protein sample are observed with equal likelihood. Furthermore, suppose that peptide identifications in a sample are independent. Upper case letters represent random variables.

Let $n$ be the number of proteins in a sample, $m$ be the mean number of peptides per protein, and $p$ be the probability of a peptide being identified in a protein in one replicate. If $Z$ is the number of identified peptides in a protein for one replicate, then $Z \sim Bin(m, p)$, where $Bin(.,.)$ represents the binomial distribution. Suppose that a protein match is defined as at least $a$ different peptides matching that protein in one replicate. The probability of a protein being identified in one replicate is

$$p_0 = 1 - p(Z < a) = 1 - CDF(Bin(a-1, m, p)) \quad (1)$$

where CDF represents the cumulative distribution function.

If $X$ indicates the number of replicates that identify at least $a$ different peptides in a protein in $r$ replicates, then $X \sim Bin(r, p_0)$. More generally, suppose that a protein match is now defined as at least $r_1$ replicates (typically one replicate) in which at least $a$ different peptides being identified in a protein in each replicate. Then the probability of a protein being identified in $r$ replicates is

$$p_r = 1 - p(X < r_1) = 1 - CDF(Bin(r_1 - 1, r, p_0)) \quad (2)$$

If $W$ indicates the number of proteins identified in $r$ replicates, then $W \sim Bin(n, p_r)$. The probability of at least a portion of $c$ proteins in the sample being identified in $r$ replicates is

$$p_c = 1 - p(W < c * n) = 1 - CDF(Bin(c * n - 1, n, p_r) \quad (3)$$

Next, we extend this simple model by allowing the probability of detection to vary between different peptides/proteins.

**2.2 Proteotypic peptides based method (PPM)**

Recent research has shown that only a few peptides (so-called proteotypic peptides) are repeatedly and consistently identified for any protein present in a mixture (Mallick *et al.* 2007), and these peptides are identified with much higher probabilities than other peptides in a protein (Lu *et al.* 2007). It remains unknown how many proteins contain proteotypic peptides. Computational predictions integrating physicochemical properties of studied peptides showed that only a proportion of proteins contain at least one high-confidence (>0.99) proteotypic peptide for yeast and human proteins (Mallick *et al.* 2007).

In PPM, we assume that all peptides can be divided into categories of proteotypic and non-proteotypic. This allows us to consider some variation in the physiochemical properties of proteins, although allowing only two categories is still likely a considerable simplification from reality.

Let $n$ be the number of proteins in a sample. Let $m_p$ and $m_{np}$ be the mean number of proteotypic and non-proteotypic peptides per protein respectively. Let $p_p$ and $p_{np}$ be the probabilities respectively that a proteotypic and non-proteotypic peptide is identified in a protein in one replicate.

If $Z_p$ and $Z_{np}$ denote the number of identified proteotypic and non-proteotypic peptides in a protein in one replicate, then $Z_p \sim Bin(m_p, p_p)$ and $Z_{np} \sim Bin(m_{np}, p_{np})$.

In this case, the probability of a protein being identified (at least $a$ different peptides being identified in a protein) in one replicate is

$$p_0 = 1 - p(Z_p + Z_{np} < a) \quad (4)$$

For example, the probability of at least one peptides being identified in a protein in one replicate is $p_0 = 1 - p(Z_p + Z_{np} < 1) = 1 - p(Z_p + Z_{np} = 0) = 1 - p(Z_p = 0) * p(Z_{np} = 0)$. The probability of at least two peptides being identified in a protein in one replicate is

$$p_0 = 1 - p(Z_p + Z_{np} < 2) = 1 - p(Z_p + Z_{np} = 0) - p(Z_p + Z_{np} = 1)$$
$$= 1 - p(Z_p = 0) * p(Z_{np} = 0) - p(Z_p = 1) * p(Z_{np} = 0) - p(Z_p = 0) * p(Z_{np} = 1).$$

Once we have $p_0$, the probability of a protein being identified in $r$ replicates is obtained from equation (2). The probability of at least a portion of $c$ proteins in the sample being identified in $r$ replicates is calculated from equation (3).

**2.3 Protein abundance class based Method (PAC)**

Protein abundance is an important factor that might affect the probability of a peptide being observed. Peptide ion abundance level (intensity) trigger data acquisition in analyses of LC-MS/MS (Liebler 2002). Typically, high abundant proteins are identified with multiple peptides and low abundant proteins by one or two (Liu *et al.* 2004). Different methods have been developed to determine the relative protein abundance. It has been shown that the spectral counts (for example, normalized spectral abundance factors [NSAF]) is an effective measurement of the relative abundance between different proteins in a protein mixture (Blondeau *et al.* 2004; Liu *et al.* 2004; Powell *et al.* 2004; Girard *et al.* 2005; Old *et al.* 2005; Zybailov *et al.* 2005; Paoletti *et*

*al.* 2006; Zybailov *et al.* 2006). In PAC, we allow that the peptides from proteins at different abundance levels in a sample have different probabilities of being observed.

We first classify $n$ proteins into $l$ abundance classes. Let $f_i$ denote the proportion of proteins at the $i^{th}$ abundance class, $i = 1,2,..,l$, and $\sum_i f_i = 1$. Then there are $f_i * n$ proteins at the $i^{th}$ abundance class in the sample. Let $m_i$ be the mean number of peptides per protein at the $i^{th}$ abundance class and $p_i$ indicate the probability of a peptide being identified in a protein in the $i^{th}$ abundance class in one replicate.

If $Z_i$ is the number of identified peptides in a protein for one replicate, then $Z_i \sim Bin(m_i, p_i)$. Suppose that a protein match is defined as at least $a_i$ different peptides matching the protein. The probability of a protein at the $i^{th}$ abundance class being identified in one replicate is

$$p_{0i} = 1 - p(Z_i < a_i) = 1 - CDF(Bin(a_i - 1, m_i, p_i) \quad (5)$$

For $r$ replicates, if $X_i$ indicates the number of replicates in which at least $a_i$ different peptides in a protein at the $i^{th}$ abundance class are identified, then $X_i \sim Bin(r, p_{0i})$. Therefore, the probability of a protein at the $i^{th}$ abundance class being identified (at least $r_{1i}$ replicates in which at least $a_i$ peptides being identified in a protein at the $i^{th}$ abundance class in each replicate) in $r$ replicates is

$$p_{ri} = 1 - p(X_i < r_{1i}) = 1 - CDF(Bin(r_{1i} - 1, r, p_{0i}) \quad (6)$$

If $W_i$ indicates the number of proteins identified in the $i^{th}$ abundance class in $r$ replicates, then $W_i \sim Bin(f_i * n, p_{ri})$. Similarly, the probability of at least a portion of $c_i$ proteins at the $i^{th}$ abundance class being identified in $r$ replicates is

$$p_{ci} = 1 - p(W_i < c_i * f_i * n) = 1 - CDF(Bin(c_i * f_i * n - 1, f_i * n, p_{ri}) \quad (7)$$

**2.4 Applications of PPM and PAC**

Theoretically, PPM and PAC can be used individually for calculating the probabilities of achieving a given level of coverage for different replicates. For example, PPM method can be individually used for samples that contain proteins with similar abundances. However, it rarely makes sense to do PPM without doing PAC since there is variation in abundance for majority of samples in real MS experiments. One possibility, among others, is to apply PPM within each protein abundance class, and PAC is used for different abundance classes. That is, applying equation (4) to calculate $p_{0i}$ for each protein abundance class, then applying equations (6) and (7) to calculate $p_{ri}$ and $p_{ci}$ respectively.

**2.5 Protein coverage at a fixed confidence level**

When the protein coverage ($c$ in equations above) is fixed, we can obtain the probabilities of achieving that coverage as a function of the number of replicates applying PPM, PAC or their combinations described above. On the other hand a numerical search is needed in order to determine the protein coverage for a fixed confidence level. When PPM is applied, we numerically search equation (3) to obtain the maximum $c$ that meets the given confidence level $p_c$. In order to obtain the overall sample protein coverage at a given number of replicates when PAC is applied, we first search equation (7) to get maximum $c_i s$ that meet the given confidence level $p_{ci}$, then convert them into absolute values of proteins, and finally sum these absolute values up and divide by the total number of proteins. When the combination of PPM and PAC is used, the protein coverage is determined in the same way as with PAC.

## 3 Estimation of the model parameters

This model can be parameterized using output from repeated MS experiments. Protein identifications are compared between each replicate, tracking the cumulative totals, $n_{\text{det}}$.

### 3.1 Classification of abundance and peptide class

The overall detectability of a protein should be related to its abundance. Thus, quantities like the mean and median probability of detection of a protein's peptides should be positively (probably strongly) correlated with the protein's abundance. The detection probabilities for individual peptides will deviate from this mean as a function of their individual physicochemical properties. Proteotypic peptides will tend to have substantially higher detection probabilities than the protein on average.

Take $X_{ij}$ as the number of replicates in which peptide $j$ of protein $i$ is observed. $X_{ij}$ varies from 0 to $r$, where $r$ is the number of replicates. The index $j$ varies from 1 to $n_i$, where $n_i$ is the number of theoretical peptides digested from protein $i$. The index $i$ varies from 1 to $n_{\text{det}}$, the number of unique proteins observed in the sample. Take $p_{ij}$ as the probability that peptide $j$ of protein $i$ is observed in a replicate. Take $p_{i.}$ as the mean of the $p_{ij}$'s across the peptides of a protein.

Then protein abundance class and proteotypic peptide can be classified with the following procedures: 1) Calculate the mean detection probability $p_{i.}$ for each protein; 2) Divide the proteins into bins according their $p_{i.}$ value, and assume that these represent abundance classes; 3) Within each abundance class, divide the $p_{ij}$'s into bins. Assume that these bins represent peptide's different intrinsic detection probabilities. Without loss of generality, we call these peptide classes. And the classification of peptides into proteotypic and non-proteotypic is a

specification of peptide classes. Furthermore, the average probabilities that a proteotypic and non-proteotypic peptide is identified in a protein in one replicate can be calculated within each abundance and peptide class; 4) Conduct the previously described method within each abundance and peptide class. That is, using equation (4) to calculate $p_{0i}$ for each protein abundance class, then applying equations (6) and (7) to calculate $p_{ri}$ and $p_{ci}$ respectively. This method is simple, but gives a means of classifying abundance and peptide properties without relying on large amount of calculations or external sources of information.

**3.2 Estimation of mean number of peptides per protein**

We do *in silico* digestions for $n_{\text{det}}$ detected proteins and get the total theoretical peptides by proteotypicness. Within each protein abundance class, the mean number of proteotypic and non-proteotypic peptides per protein can be calculated as the total theoretical proteotypic (or non-proteotypic) peptides divided by total number of proteins.

**3.3 Estimation of total proteins in the sample**

Koziol *et al.* (2006) adapted the capture-recapture model, which has been widely used in ecology and biometry to estimate the size of populations (Otis *et al.* 1978; Seber 1982), with a closed population and time-varying and heterogeneous individual probabilities of capture to estimate the total number of proteins in typical MudPIT experimental sample (Washburn *et al.* 2001). A nonparametric approach proposed by Chao and colleagues (Chao 1989; Chao *et al.* 1992; Lee & Chao 1994) was used to obtain the estimation of the underlying population size, i.e. total proteins in the sample. For convenience, the number of estimated total proteins in the sample is denoted as $n$ hereafter.

# 4 Results

## 4.1 Estimation of the model parameters

Six LC-MS/MS runs were conducted under the same conditions of sample complexity (unknown proteins), separation resolution, instrumental and data acquisition, database searching and filtering using MASCOT. Peptides scoring with 1% FDR (false discovery rate) were retained. Protein identifications were compared between each run, tracking the cumulative totals. A total of 36 different proteins were identified in the cumulative dataset. Of these, 2 (5.6%) were found in every analysis and 17 (47.2%) were found in only one of the analyses. A total of 19 (52.8%) proteins were found in at least two or more analyses. It is not expected that we would observe all proteins in the samples. The 36 identified proteins contain 797 peptides (*in silico* digestion) with the number of amino acids larger than or equal to 6 (Note: short peptides are ignored by mass spectrometry after tryptic digestion). Overall, 139 peptides were detected in 6 runs.

We estimated parameters for our model using this data. Using the protein abundance class and peptide class classification described in Section 3.1, we nearly evenly classify proteins into 3 abundance classes according to their mean detection probabilities. Within each abundance class, we divide peptides into proteotypic and non-proteotypic according to empirical rule of "observed in >50% of all identifications of the corresponding protein" (Mallick *et al.* 2007). Then we calculate the average probability of a peptide being identified in a replicate and mean number of peptides per protein for each peptide class within each protein abundance class (Table 2.1).

The total number of proteins in the sample was estimated from a method described in Koziol (2006), which used a nonparametric approach proposed by Chao and colleagues (Chao 1989;

Chao *et al.* 1992; Lee & Chao 1994). The estimated total number of proteins is 89 ($t = 6, M_{t+1} = 36, f_1 = 17, f_2 = 3, f_3 = 7, f_4 = 4, f_5 = 3, f_6 = 2, n_1 = 17, n_2 = 8, n_3 = 13,$ . $n_4 = 17, n_5 = 12, n_6 = 20$ are used in calculations. See Koziol's paper for the formulas for calculations).

Our analysis assumes that all proteins have the average number of proteotypic peptides. However the actual number presumably varies from protein to protein. Proteins without any proteotypic peptides are probably much harder to detect. Assume that the number of proteotypic peptides per protein follows a $binomil(m, p)$ distribution, where $m$ is the number of peptides per proteins and $p$ is the probability that a peptide is proteotypic. $p$ can be estimates as the number of proteotypic peptides divide by the total number peptides, which yields 0.0263 based on our data. Then the expected proportion of proteins with no proteotypic peptides is about 56 percent. That is, more than half of proteins are with no proteotypic peptides in the sample.

**4.2 Probabilities of achieving a given level of coverage for different replicates**

Figure 2.1 shows the probability of identifying at least 95 percent of peptides as a function of the number of replicates. This is calculated as described in the methods section using the parameter estimates from above. With the definition of a protein match being that there is at least one replicate in which at least one peptide is identified in a protein in $r$ replicates applied (Fig. 2.1, left), there is very small chance (1%) to identify 95% of High abundant proteins in 1 replicate, there is moderate chance (54%) of identifying 95% of High abundant proteins in 2 replicates, but with 3 replicates, 94% of High abundant proteins could be identified with >95% confidence. The probability of identifying 95% of Medium abundant proteins within 2 replicates is very small, 7 replicates are needed to identify >95% of Medium abundant proteins with >95%

confidence. For Low abundant proteins, more than 30 replicates are needed to identify 95% of proteins with >95% confidence (Fig. 2.1, left).

With the definition of a protein match representing at least one replicate in which at least two peptides are identified in a protein if $r$ replicates applied (Fig. 2.1, right), the probability of identifying 95% of High abundant proteins within 5 replicates is very small (<1%), but with 17 replicates, 95% of High abundant proteins could be identified with >95% confidence. For Medium abundant proteins, 95% of proteins could be identified with >95% confidence in 43 replicates. Under this definition, there is essentially zero probability of identifying 95% of Low abundant proteins with realistic sample sizes (Fig. 2.1, right).

**4.3 Protein coverage in samples at a fixed confidence level**

Figure 2.2 shows the protein coverage under a fixed confidence level as a function of the number of replicates. This is obtained as described in the methods section using numerical search. With the definition of a protein match representing at least one replicate in which at least one peptide is identified in a protein in $r$ replicates (Fig. 2.2, left), with the probability 0.95, 59% of proteins at High abundance are identified in one replicate. And 4 replicates are needed to identify at least 95% of proteins at High abundance. The coverage of 95% of Medium abundant proteins requires 7 replicates. For the proteins with Low abundance level, 3% of low abundant proteins are identified in one replicate with the probability of 0.95. More than 30 replicates are needed to identify 95% of proteins at Low abundance. Overall, 30% of proteins in samples could be identified in one replicate with the probability 0.95. In 20 replicates, 95% proteins can be identified with the probability 0.95 (Fig. 2.2, left).

With the definition of a protein match representing at least one replicate in which at least two peptides are identified in a protein if $r$ replicates (Fig. 2.2, right), 9% of proteins at High

abundance are identified in 1 replicate with the probability of 0.95. Nine and 17 replicates are needed to identify at least 75% and 95% of proteins at High abundance, respectively. About 3% and 50% of Medium abundant proteins could be identified in 1 and 10 replicates with the probability 0.95, respectively. About 22% of proteins at Low abundance are identified in 50 replicates with the probability of 0.95. Overall, 3% and 50% of proteins in samples could be identified in 1 and 15 replicates with the probability 0.95, respectively. With up to 50 replicates, about 71% proteins in samples could be identified with 95% confidence (Fig. 2.2, right).

**5 Discussions**

Researchers have utilized replicate analyses as a means to increase the number of proteins identified in analyzing complex samples (Lipton *et al.* 2002; Wu *et al.* 2003). The fact observed in repeated MS runs is that not all proteins identified in the first analysis are observed again in a second run. It is also a fact that new proteins are identified with each new run (Liu *et al.* 2004). Yates and colleagues (Liu *et al.* 2004) developed a statistical model and predicted 10 analyses are required for 95% saturated sampling of a yeast-soluble cell lysate sample. The prediction of the capture-recapture model showed a dramatic increase of new proteins in the first several analyses, and with saturation level generally being reached at about 10 – 50 runs, depending on the protein abundances (Koziol *et al.* 2006). In both cases, the saturation means that the discovery rate of new proteins per spectrum added to the dataset decreases, rather than the complete coverage of all proteins in the sample. Our results are broadly consistent with these findings but show that there is considerably more complexity to the question. If the goal is to identify most (say 95%) of the medium and high abundance proteins in the sample, then between 4 and 40 replicates are required (closer to four if most proteins can be identified with a single peptide and closer to 40 if most require two peptides). If the goal is to identify low abundance

proteins also, then 50 or more replicates are required. If some important biomarker proteins occur only in low abundance, then a complete rethinking of the scale on which proteomics experiments are conducted is required. Due to these fundamental limitations, the goal of complete and quantitative proteome analysis will remain elusive (Malmstrom *et al.* 2007). To overcome these limitations, the target-driven quantitative proteomics is desired on the basis of proteotypic peptides that are detectable in mass spectrometer and can uniquely identify its protein of origin (Aebersold 2003; Kuster *et al.* 2005). It should also be note that we parameterized the model from a relatively low complexity sample. The peptide detection probabilities would presumably be lower in a more complex mixture because of the inability to detect large numbers of simultaneously eluting peptides (see Chapter 3 for a detailed analysis of this issue). Other platform-, organism-, or tissue-specific factors may also affect peptide detection probabilities. Thus, the appropriate replication number should be determined from samples similar to that used in the target study. The method proposed here can either be applied to data collected in a small study in order to estimate the replicate number required for a larger study, or it can be used *post-hoc* in order to estimate the coverage achieved in an experiment.

**5.1 Classification of protein abundance class and proteotypic peptides**

In our study, a simple method is applied for classification of protein abundance classes and proteotypic peptides. However, other methods can be used for these classifications as long as it is reliable. Various approaches can be used to classify protein abundance class: 1) Lu *et al.* (2007) developed an approach to calculating the absolute protein expression index (APEX) on the basis of MS data, which has been published as a Nature protocol (Vogel and Marcotte, 2008). We can classify proteins into abundance classes according to their APEXs. 2) Several studies have shown that the spectral counts or normalized spectral abundance factors (NSAF) are effective

measure of the relative abundance among different proteins in protein mixtures (Blondeau *et al.* 2004; Liu *et al.* 2004; Powell *et al.* 2004; Girard *et al.* 2005; Old *et al.* 2005; Zybailov *et al.* 2005; Paoletti *et al.* 2006; Zybailov *et al.* 2006). Typically, we can classify proteins into abundance classes according to the spectral count or NSAF. 3) The external data of the classification of protein abundance can be directly used if it is available. For example, Ghaemmaghami *et al.* (2003) obtained the abundances of 4500 yeast proteins using western blots, and these data can be used for protein abundance classification.

There are different ways to determine proteotypic peptides: 1) mine the proteotypic peptides of targeted proteins from experimental proteomic databases (Desiere *et al.* 2005; Marzolf *et al.* 2006); 2) predict the proteotypic peptides using computational approaches (Mallick *et al.* 2007; Lu *et al.* 2007); and 3) on the basis of peptide identifications from multiple MS runs, apply an empirical rule like a peptide being denoted as proteotypic if observed in more than 50% of all identifications of the corresponding protein (Mallick *et al.* 2007).

**5.2 Model robustness**

There are a number of simplifying assumptions in our model. First, we assumed detection probability to be independent and identically distributed between peptides within a peptide and abundance class. This assumption of independence seems reasonable as a first approximation, but is probably not true in reality. Peptides that elute from the LC at the same time "compete" for detection. Thus, their detection probabilities are not independent.

The assumption of equal detection probabilities is likely more problematic. We allowed different peptide and abundances classes in order to model some of the variability in detection probability. However, this is still a major weakness of our approach. In our data application, we considered only two classes of peptide: proteotypic and non-proteotypic. Given more data, we

could divide peptides into finer detection classes and this might substantially impact our estimates of protein detection probabilities. Likewise, we only considered only three abundance classes. In reality, peptide detectability and protein abundance both vary on continuous scales. A method using local density estimates for peptide detectability and protein abundance might give improved performance. Another possibility would be to construct a hierarchical model with binomial counts and some underlying model for peptide detection probabilities.

Another assumption is that all proteins have the average number of peptides for their abundance class. In reality there is substantial variation in peptide number and this impacts the detection probability for the protein, both because it changes the number of opportunities for detection and because it changes the probability of the protein having any proteotypic peptides. Thus certain proteins will be harder to detect than average and others will be easier to detect. It is not obvious what the net impact of this assumption is. Thus, a more advanced analysis should consider this issue.

It is worth noting that when a fixed peptide detection probability is used, the probabilities of a given percent of proteins being identified in different replicates do not significantly decrease with the increasing number of unique proteins in samples. With three samples of 89, 150, and 300 proteins, the probabilities of identifying 95% of proteins over different replicates are similar (Fig. 3). Recall that the calculation of the probability of at least a portion of $c$ proteins in the sample being identified in $r$ replicates is $p_c = 1 - p(W < c*n) = 1 - CDF(Bin(c*n-1, n, p_r)$. Obviously, for a given $p_r$ or the corresponded probability of a peptide being identified in a protein in a replicate, $p_c$ would decrease as $n$ increases when $c = 1$. That is, the probability of all proteins in samples being identified in $r$ replicates decreases with the increasing proteins in samples (Fig. A2.1). When the goal is to detect a fixed portion of proteins (say, $c$) in samples,

the number of combinations of choosing $c*n$ in $n$ increases with the increasing of $n$. This increase in the number of combinations (choosing $c*n$ in $n$) would balance the decrease of $p_c$ due to the increase of number of sample proteins. Consequently, the final $p_c$ remains nearly constant under samples containing different number of proteins when peptide detection probabilities are fixed in these samples (Fig. 3).

It must be re-emphasized that the independence of $p_c$ with the number of proteins in samples would only be valid under the assumption of a fixed probability of a peptide being identified in a protein in a replicate for samples at different complexities. This might not be the case in real experiments, in which the probability of a peptide being identified in a protein in a replicate for less complex samples is higher than that for more complex samples. The point here is that our model is robust for the parameter $n$, total number of proteins in sample. That is, the model yields similar results (e.g. $p_c$ and coverage estimates) when different $n$ estimates (within certain range) are used. The robustness of $p_c$ estimates with respect to $n$ provides great advantages of our model, because reliable estimation of *n* is difficult. The protein coverage is also robust to the total number of proteins, $n$ (Fig. A2.2).

**5.3 Effect of number of peptides per protein**

An interesting phenomenon is that the mean numbers of peptide per protein decreases with the increase of protein abundance. We observe this phenomenon in both published data (e.g. Zybailov *et al.* 2006 YPD and MIN data) and our 6-replicates data. The higher number of peptides per protein in Medium abundant proteins would compensate the decrease in detection probability due to the smaller probability of a peptide being identified.

Table 2.1: Parameter estimates for each peptide class within each protein abundance class

| Peptide class | Parameter | Protein abundance class | | |
|---|---|---|---|---|
| | | High | Medium | Low |
| Proteotypic | Probability of a peptide being identified in a replicate | 0.6111 | 0.3889 | - |
| | Mean number of peptides per protein | 1 | 1 | - |
| Non-proteotypic | Probability of a peptide being identified in a replicate | 0.0321 | 0.0113 | 0.0051 |
| | Mean number of peptides per protein | 12 | 22 | 28 |
| Proportion of proteins | | 0.25 | 0.39 | 0.36 |

-: Not applicable.

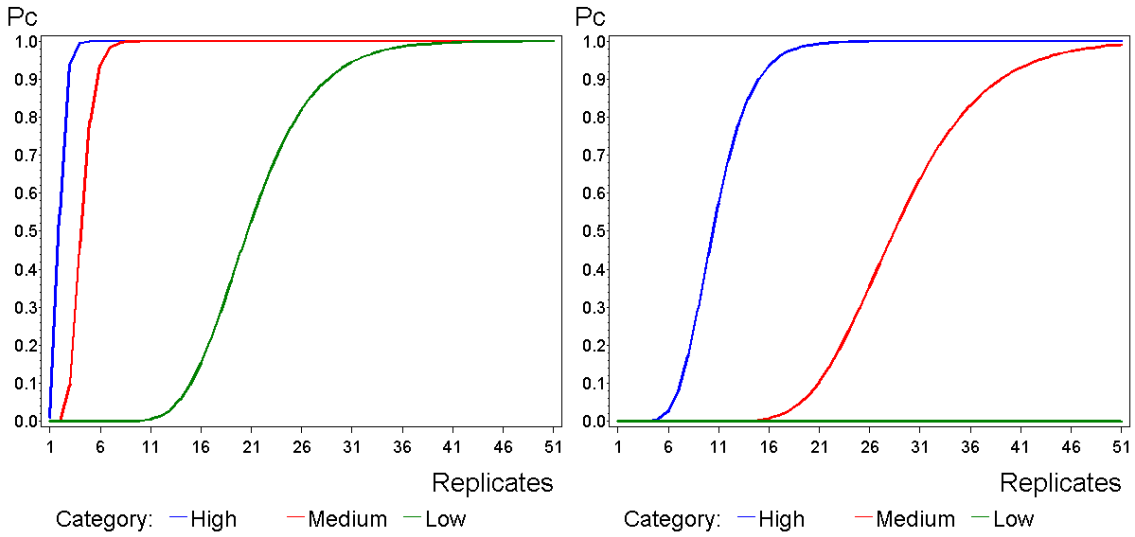



Figure 2.1: The probabilities of identifying 95% of proteins in each protein abundance class as a function of the number of replicates. A match is defined as at least one replicate in which at least one (left) and two (right) unique peptides are identified in a protein in $r$ replicates. Numbers of proteins: 89.

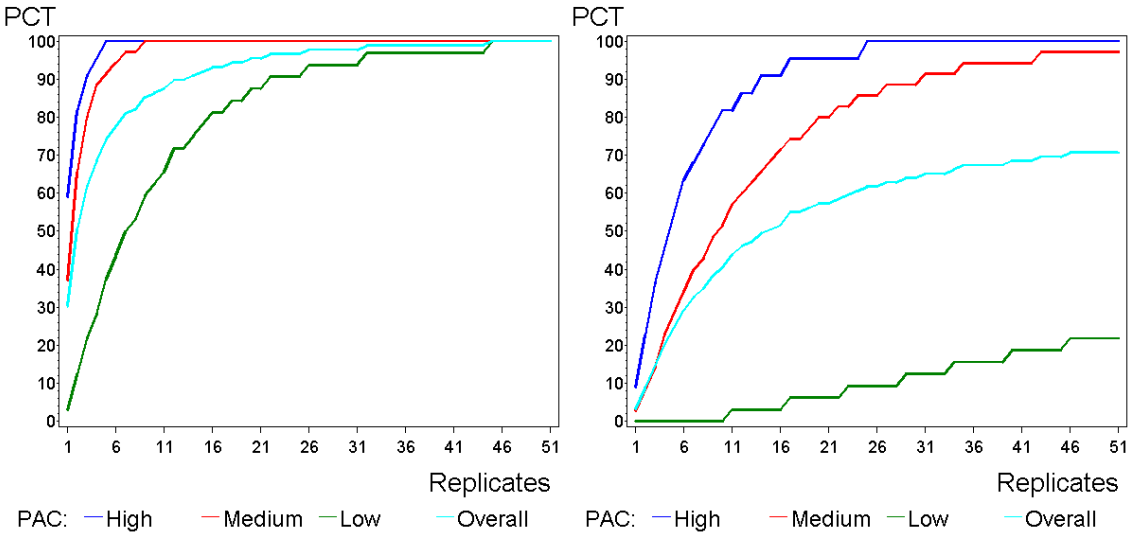



Figure 2.2: The protein coverage at each protein abundance class and overall coverage as a function of the number of replicates with >95% confidence level. A match is defined as at least one replicate in which at least one (left) and two (right) unique peptides are identified in a protein in $r$ replicates. Numbers of proteins: 89.

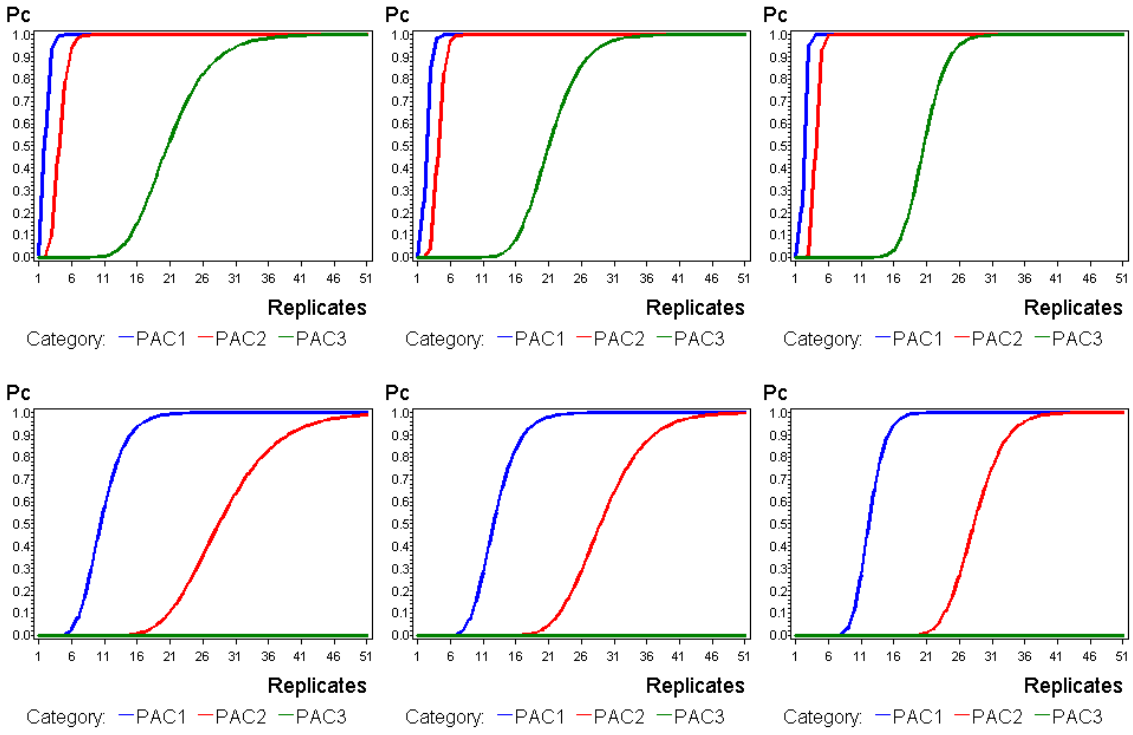



Figure 2.3: The probabilities of identifying 95% of proteins in each protein abundance class as a function of the number of replicates. A match is defined as at least one replicate in which at least one (top panel) and two (bottom panel) unique peptides are identified in a protein in $r$ replicates. Different numbers of proteins in samples are applied: left column for 89, central column for 150, and right column for 300 proteins.

# CHAPTER 3

# COMPETITION FOR DETECTION BETWEEN PEPTIDES AND VARIATION IN PROTEIN DETECTION PROBABILITIES IN MASS SPECTROMETRY BASED PROTEOMICS[*]

[*] Liu S, Orlando R, and Schliekelman P. To be submitted to Proteomics.

**Abstract**

The development of LC-MS/MS technologies has greatly expanded the ability of researchers to study proteomes on a broad scale. Although a major leap forward, these methods still suffer from serious limitations – in particular, the low protein coverage obtained in a single LC-MS/MS run. There are many factors that contribute to the variation in detection probabilities of proteins, including proteolytic digestion efficiency, peptide separation and co-eluting peptides, and peptide ionization efficiency, among others. There is currently little known about the relative importance of these factors in determining detection efficiency. Thus, it is difficult to know how to improve peptide detection probabilities. Quantifying these different sources of error would allow researchers to optimize technology, experimental protocols, and statistical procedures.

In this study, we analyze one potentially significant source of missed detections: competition for detection between co-eluting peptides. A single MS/MS scan can only detect a limited number of peptides. When the protein mixture is complex, there will be many sampling intervals with many more co-eluting peptides than can be simultaneously detected and this can result in many missed peptide detections. In order to analyze this process we developed a mathematical model based on order statistics from independent non-identical normal random variables. As an approximation to this model, a simulation approach was applied to quantify the importance of this effect. We study the impacts of protein abundance, complexity of samples, proteolytic digestion efficiency, peptide separation and co-eluting peptides, scanning speed of the mass spectrometer, and dynamic exclusion efficiency on the peptide/protein identification. We show that competition for detection between peptides is expected to be a major source of missed detections with more complex protein mixtures. We also show how peptide detection probability is expected to depend on protein abundance. The proposed simulation approach can be used as a

framework for analysis of impacts of various factors on the peptide/protein detection. The simulation results provide valuable information for optimizing LC-MS/MS techniques and practical guidelines for conducting MS-based experiments. Simulation results agree well with the findings in publications from the real MS experiments.

**Keywords:** Simulation; Mass spectrometry (MS); MS-based proteomics; Protein identification; Protein abundance; Retention time

## 1 Introduction

Proteomic analysis of complex protein mixtures using LC-MS/MS (liquid chromatography-tandem mass spectrometry) has become a widely used method in recent years. A typical MS-based proteomic discipline consists of five stages (Aebersold & Mann 2003; Mallick *et al.* 2007): 1) Protein samples are partially purified or separated by chromatographic or electrophoretic methods. 2) Purified or separated protein samples are digested with trypsin or other proteases, which results in complex peptide mixtures. 3) The peptide mixtures are separated by single or multiple dimensional liquid chromatography (LC) and 4) analyzed using tandem mass spectrometers to obtain MS/MS spectra. 5) Peptides and proteins are subsequently identified by correlating MS spectra with a protein sequence database with the aid of specialized software (e.g. SEQUEST, MASCOT). Over the past decade, rapid improvements of MS-based proteomics have increased the resolution and reproducibility of sample separation (Heller *et al.* 2005; Malmstrom *et al.* 2006), the speed and quality of data acquisition (Domon & Aebersold 2006), and the confidence of inferring the true peptide and protein identification from MS/MS spectra (Eng *et al.* 1994; Perkins *et al.* 1999; Keller *et al.* 2002; Nesvizhskii *et al.* 2003; Haas *et al.* 2006; Elias *et al.* 2007).

Despite the success of these methods, they suffer from major limitations and a typical LC-MS/MS run will often only identify a small portion of the proteins in a sample. The process of spectrum detection from the LC-MS/MS depends on a multitude of factors. Besides the influence of protein abundance (Aebersold & Mann 2003), factors such as proteolytic digestion efficiency (Gatlin *et al.* 2000; Kjeldsen *et al.* 2003), peptide separation and co-eluting peptides (Breci *et al.* 2003; Nielsen *et al.* 2004; Ocaña *et al.* 2005), peptide ionization efficiency, ion suppression, scanning speed of the mass spectrometer, and dynamic exclusion efficiency (Dobo & Kaltashov 2001; Peschke *et al.* 2002; Pan & McLuckey 2003; Pan *et al.* 2004) influence peptide identification. Researchers have extensively applied this technology without a consistent rule on the degree of stringency required in data generation and analysis. Consequently, some results are likely to be of questionable quality and further validation is needed (Wilkins *et al.* 2006). A key problem is that there is little understanding of the reliability of these methods at protein detection. While it is well known that a single LC-MS/MS run will miss many proteins in the sample, there has been little effort to quantify this. Furthermore, little is known about which of the many factors listed above are most important in determining the probability of protein identification or the extent to which the detection failures are random versus systematic. Better understanding of the relative importance of these factors would allow optimization of the technology, experimental protocols, and statistical analysis.

In this study, we attempt to quantify one potentially important source of variation in peptide detections: the "competition" for detection between peptides. A single MS scan can only detect a limited number of peptides. Typically, the peptides producing the $c$ most intense signals in an scanning interval are further analyzed and the remaining peptides are discarded ($c$=9 is a typical value). When the protein mixture is complex, then there will be greater than $c$ co-eluting peptides

in many scanning intervals. A peptide will only be detected if it is in the top *c* most intense peptides in some sampling interval. When there are substantially more than *c* co-eluting peptides, then it is probable that some peptides are never in the top *c* before their period of elution ends.

In this study, we derive a mathematical model to study this process. As an approximation to the analytical model, we use a simulation approach to analyze the impacts of factors, such as protein abundance, complexity of samples, proteolytic digestion efficiency, peptide separation and co-eluting peptides, scanning speed of the mass spectrometer, and dynamic exclusion efficiency, on the peptide/protein identification. We show that this competition for detection is expected to be a major source of missed detections when protein mixtures are complex, but of limited importance in simpler protein mixtures. We also show how peptide detection probability is expected to depend on protein abundance. The proposed simulation approach can be used as a framework for analysis of impacts of various factors on the peptide/protein detection. The simulation results provide valuable information for optimization of LC-MS/MS techniques and practical guidelines for conducting MS-based experiments.

**2 Model derivation**

Assume there are $n$ proteins in the sample. Protein $i$ consists of $b_i$ peptides and its abundance (number of copies of the protein) is $a_i$.

To simplify the analysis, first assume a complete protein digestion, i.e. trypsin cleaves each copy of protein sufficiently at the identified site. If the protein digestion is incomplete, peptides from the same protein may have different abundances in the peptide mixtures. In this case, simply replace $a_i$ with $a_{ij}$ (abundance of peptide $j$ in protein $i$) in the formulas below to get the appropriate analysis. Under the assumption of a complete digestion, each peptide in protein $i$

has $a_i$ copies. Then total number of distinct peptides $M$ in the sample is $\sum_{i=1}^{n} b_i$, and total number of copies of peptides $C$ is $\sum_{i=1}^{n} a_i b_i$.

Peptide mixtures pass into the LC. Via this process, the peptides are eluted at different times. Take the retention time $\tau_{ij}$ for peptide $j$ in protein $i$ as $\tau_{ij} \sim normal(\mu_{ij}, \sigma_{ij})$. The mean and variance for each peptide are determined by the physicochemical properties of the peptide. After the peptides are ejected from the LC apparatus, they pass directly into the mass spectrometer. A peptide is selected for fragmenting if it is ever among the top $c$ most intense peptides. Take the number of peaks in the full-scanned MS as $c_l$ at sampling point $t_l$. If $c_l$ is less than $c$, all peptides in the full-scanned MS are sequentially selected for fragmenting. In short, $\min(c_l, c)$ MS/MS spectra are generated at sampling point $t_l$.

We hypothesize that a major source of variation in peptide detection results from successful or failure of peptides to ever be among the *c* most intense ions during some sampling interval. Thus, we must determine the probability of a peptide being selected.

The probability that a given copy of peptide $j$ in protein $i$ elutes at time *t* is $f(t \mid \mu_{ij}, \sigma_{ij})$. So, we have a collection of copies of the peptide all eluting independently according to this distribution. MS sampling events happen at discrete time intervals $t_1, t_{2,} \ldots$ and last for some time interval $d$. Thus, a given ion would be detected at sampling point $t_l$ if it elutes between $t_l$ and $t_l + d$. There is presumably some time lag between elution and entering the mass spectrometry. So retention time is actually defined as the time it enters the mass spectrometer.

Now define an indicator variable $I_{ijk}(t_l)$ for each copy $k$ of peptide $j$ in protein $i$. It takes the value 1 if the copy elutes between $t_l$ and $t_l + d$. That is the peptide ion would be detected by a MS sampling event starting at $t_l$. It is reasonable to assume that the indicator variables are independent, i.e. each copy $k$ of peptide $j$ in protein $i$ independently elutes between $t_l$ and $t_l + d$. Under this assumption, the probability of a given copy $k$ of peptide $j$ in protein $i$ being detected at sampling point $t_l$ is

$$P_{ij}(t_l) = P\left(I_{ijk}(t_l) = 1\right) = \int_{t_l}^{t_l+d} f\left(t \mid \mu_{ij}, \sigma_{ij}\right) dt \tag{1}$$

where $f\left(t \mid \mu_{ij}, \sigma_{ij}\right)$ is a normal $pdf$ under the assumption that the retention times are distributed normally.

The sampled abundance (i.e. the abundance measured in the mass spectrometer) of peptide $j$ in protein $i$ at sampling point $t_l$ is $A_{ij}(t_l) = \sum_{k=1}^{a_i} I_{ijk}(t_l)$, and thus $A_{ij}(t_l)$ is distributed as a binomial distribution with $a_i$ trials and probability $P_{ij}(t_l)$ of success. Therefore,

$$E\left(A_{ij}(t_l)\right) = a_i \int_{t_l}^{t_l+d} f\left(t \mid u_{ij}, \sigma_{ij}\right) dt \tag{2}$$

and

$$Var\left(A_{ij}(t_l)\right) = a_i \left( \int_{t_l}^{t_l+d} f\left(t \mid \mu_{ij}, \sigma_{in}\right) dt \right) \left( 1 - \int_{t_l}^{t_l+d} f\left(t \mid \mu_{ij}, \sigma_{in}\right) dt \right) \tag{3}$$

Then we can apply central limit theorem (CLT) to get the distribution of $A_{ij}(t_l)$ since the protein abundance $a_i$ is usually large. There may be some peptides with low abundances (e.g. 10 copies for some proteins), but the majority of peptides have more than thousands of copies. Thus,

for these peptides (using CLT), the number of copies sampled $A_{ij}(t_l)$ has a normal distribution with mean and variance given by equation (2) and (3) respectively. Therefore, at any sampling point we have a set of normal distributions in operation (one for each peptide).

In real MS experiments, two strategies are used for sampling the top most intense peptide ions in each cycle. The difference between them is whether or not to sample peptide ions that have already been sampled in previous cycles.

**Case 1) Sample the top most abundant peptide ions in each cycle without exclusion**

At sampling point $t_1$, the instrument switches between full-scanned and tandem MS modes to generate MS/MS spectra for the $\min(c_1, c)$ top intense peptide ions. This process is repeated until all of the peptides have been analyzed from the LC step. In general, a given peptide gets sampled if its sampled abundance is in the top $\min(c_l, c)$ for sampling point $t_l$. Take $A_{11}(t_l),...,A_{1b_1}(t_l),...,A_{i1}(t_l),...,A_{ib_i}(t_l),...,A_{n1}(t_l),...,A_{nb_n}(t_l)$ as the sampled abundances at sampling point $t_l$, and $A_{(1)}(t_l), A_{(2)}(t_l),..., A_{(M)}(t_l)$ as its order statistics. Then peptide $j$ in protein $i$ is selected at time $t_l$ if $A_{ij}(t_l) > A_{(M-\min(c_l,c)-1)}(t_l)$, i.e. the top $\min(c_l, c)$ most intense ions at sampling point $t_l$.

**Case 2) Sample the top most abundant peptide ions in each cycle with exclusion**

In order to detect more distinct peptides in experiments, peptides having been detected in previous cycles could be excluded from the current sampling. In the first cycle, the instrument switches between full-scanned and tandem MS modes to generate MS/MS spectra for the $\min(c_1, c)$ top intense peptide ions. After the first cycle is completed, the selected peptide ions are put on an exclusion list (by mass). The exclusion list has a maximum size of $u$ (typically, 50) and $e_l$ at sampling point $t_l$. Peptides are removed from the list after some cycle $\gamma$ (generally a

given peptide is done eluting within some time interval). In this case, a given peptide gets sampled if its sampled abundance is in the top $\min(c_l - e_l, c)$ after excluding peptides on the exclusion list for the given sampling point. The exclusion list is updated after each cycle. Take $A_{11}(t_l),...,A_{1b_1}(t_l),...,A_{i1}(t_l),...,A_{ib_i}(t_l),...,A_{n1}(t_l),...,A_{nb_n}(t_l)$ as the sampled abundances at sampling point $t_l$. Take $I_{ij}(t_l)$ as an indicator variable that takes value 0 if peptide $j$ in protein $i$ was in the exclusion list at sampling point $t_l$ (the list contains peptides sampled in the previous γ cycles). Take $A_{(1)}(t_l), A_{(2)}(t_l),..., A_{(M)}(t_l)$ as the order statistics for $I_{11}(t_l)A_{11}(t_l),..., I_{1b_1}(t_l)A_{1b_1}(t_l),..., I_{i1}(t_l)A_{i1}(t_l),..., I_{ib_i}(t_l)A_{ib_i}(t_l),..., I_{n1}(t_l)A_{n1}(t_l),..., I_{nb_n}(t_l)A_{nb_n}(t_l)$. Then, peptide $j$ in protein $i$ is selected at time $t_l$ if $I_{ij}(t_l)A_{ij}(t_l) > A_{(M-\min(c_l-e_l,c)-1)}(t_l)$, i.e. the top $\min(c_l - e_l, c)$ most abundant peptide ions not present in the exclusion list at sampling point $t_l$.

We have to handle order statistics from independent non-identical normal random variables in the case of without exclusion, and deal with order statistics from the products of indicator variables and independent non-identical normal random variables in the case of with exclusion. Using some known properties of permanents [the definition of the permanent is equivalent to the determinant except that all signs in the expansion are positive] (Minc 1983, 1987), general recurrence relations have been established for computing moments of order statistics of non-identically random variables arising from Exponential (Balakrishnan 1994), Pareto (Childs & Balakrishnan 1998), Weibull (Barakat & Abdelkader 2000), Erlang, and Laplace (Barakat & Abdelkader 2004) distributions. To use the general recurrence relations, it is necessary to find an explicit cumulative distribution function for the specific distribution or its approximation (personal communication with Y. Abdelkader). Since the cumulative distribution function of the

normal distribution is implicitely expressed in terms of its density function, how to extend the recurrence relations to the case arising from normal distribution remains an unsolved problem. As an approximation, we use a simulation approach instead.

**3 Simulation model**

In general, the steps of the model are as follows: 1) specify the number of proteins in the sample, the number of peptides of each protein, and the abundances for each protein. 2) Generate the mean elution time ( $\mu_{ij}$ ) for each peptide from an appropriate distribution and also specify the standard deviation ( $\sigma_{ij}$ ) in elution time. The actual elution times are then randomly generated for each peptide from a normal distribution with the mean $\mu_{ij}$ and standard deviation $\sigma_{ij}$ . 3) Specify a sampling points $t_1, t_{2,}...$ and then do a loop over these intervals. Next, determine how many copies of each peptide elute in each such interval. Then, previously detected peptides are removed if applicable (with exclusion). The remaining peptides are ordered by eluted copy number and the top $\min(c_l, c)$ [or $\min(c_l - e_l, c)$ if with exclusion] most abundant peptides are detected. This continues until the final sampling interval (see Appendix A3.0 for the details). We do multiple replicates of this process and keep track of detection probabilities for individual peptides and/or proteins. This simulation was implemented as an R (R Core Development Team) program.

**3.1 Selection of distribution of mean elution time and its parameter estimations**

**1) Data**

Petritis *et al.* (2006) trained an artificial neural network for prediction of peptide retention times using ~345,000 non-redundant peptides identified from a total of 12,059 LC-MS/MS analyses of more than 20 different organisms, and chose 1303 high confident identifications for

testing the predictive capability of the model. We use these data to explore the distribution of mean retention time here.

Pfeifer *et al.* (2007) employed LC-ESI-MS methodology to analyze 19 different proteins. The proteins were divided into three artificial protein mixtures to avoid excessive overlapping of peptides in LC separations, and each mixture was analyzed in triplicate. Peptide identifications were verified by sequence comparison with the protein sequences. The peptide retention times were extracted (several retention times were obtained if multiple spectra identifying the same peptide in a replicate) for each replicate. Two standard peptides (one eluted very early and the other eluted very late) identified in all of the runs were chosen as a reference. Peptide retention times were scaled linearly so that the early eluting peptide got an NRT (normalized retention time) of 0.1 and the late eluting peptide an NRT of 0.9. The lists of identified peptides for each replicate of each mixture, together with their respective retention times, were kindly provided by N. Pfeifer (personal communication), and these data were used to generate the retention time distribution in our analysis. That is, we could determine the mean retention time ( $\mu$ ) and also specify the standard deviation in retention time ( $\sigma$ ) for each peptide. In this study, data were merged and the mean NRT for the given peptides were calculated if a peptide was measured more than once during the triplicate analyses, which resulted in 321 unique peptide identifications.

**2) Distribution of mean retention time**

The distribution of the retention time among the real tryptic peptides (called distribution-in thereafter) is hard to obtain since *in silico* peptide digestions by trypsin are uncertain due to the existence of miss cleavage, partial digestion, etc. We could only use the distribution of the retention time of identified peptides (called distribution-out thereafter) as its representation, and

would expect that when the complexity of samples is low, the distribution-out of the MS will resemble the distribution-in.

Petritis' 1303-peptide data (Fig. 3.1 left) and Pfeifer's 321-peptide data (Fig. 3.1 right) were fitted against gamma, lognormal, and normal distributions. In general, the gamma distribution fit the peptide retention times reasonably well, although the distribution of Pfeifer's data also resembles a uniform with tails. Since Pfeifer's mixtures are quite simple (several proteins per mixture), this distribution-out may represent the distribution-in.

Using Petritis's data, the parameter estimates of gamma distribution at different retention time scales were obtained. The results show that if we reduce the total sampling period and number of proteins in samples by the same factor, then simulation results are unaffected (Appendix A3.1). Furthermore, our results show that absolute abundances don't matter – only relative abundances do. Thus, we can reduce the scale of the simulations without affected the results. Since the abundance of some proteins are more than millions and sampling period generally last 60 minutes with a sampling interval of about 1 second, the reductions in simulations are necessary to perform simulations within a reasonable time period on well-equipped computers.

The best fit for the gamma distribution had parameter estimates (threshold =-3.0541, scale=6.1134, shape=6.8060) at the scale of 100 time units. This distribution is denoted as Gamma(.) hereafter. Two thousands (200 proteins and 10 peptides per protein) random numbers were generated from Gamma(.) (Fig. 3.2).

**3) Standard deviation of mean retention time**

Using Pfeifer's data, the estimated mean standard deviation (SD) of NRT is 0.55 at a scale of 100 units, and the $90^{th}$ percent quantile is 0.92 time unit. It should be noted that the estimated

average SD of NRT at a scale of 100 is based on detected spectra. However, only the top most intense peptides can be detected in each cycle. Peptide copies that elute in the tail of their elution distribution are unlikely to be among the *c* most intense and thus are unlikely to be detected. Therefore, the estimate of SD in NRT is smaller than the "real" one since we may miss detection of the tails. On the other words, the real average SD of NRT is at a scale of larger than the 0.55 in Pfeifer's data. We set a SD of 1 time unit in simulations. The simulations show that the impacts of the selection of various SDs on detection probabilities are negligible (Appendix A3.2). Figure 3.3 shows the scatter plot of mean NRT against NRT for all spectra at a scale of 1 unit. The small variation of elution times within mean NRT implies that retention times are measured with high reproducibility.

**4) Values for other parameters**

The current values of parameters were selected according to the following logic: We fix the sample interval ($t_0$) and the top most intense ions ($c$) according to the real experiments. Ideally, we do not want to any reduction of total number of peptides ($A$), total sampling period ($T$), and abundances in simulations. However, we have to do such reduction due to the limitation of computational resources. So, we do appropriate reductions in both A and T simultaneously (keeping a fixed ratio of $A/T$) by using a distribution at different scales. Typically, a MS experiment may last 1 hour. However, most elution occurs over a shorter time interval. If we assume that most elution happens in 30 minutes (instead of 60), that there is an average of 30 peptides per protein, and that the 100 seconds that we are looking at is "typical", then the implied number of proteins in the sample is on the order of 1200. That is, 1200 proteins times 30 peptides gives 36,000 peptides. If these peptides elute over 1800 seconds then the average number eluting in a period of 100 seconds is 36000*100/1800=2000 peptides.

### 3.2 Investigated factors

Several factors, which are closely related to the peptide identification, are considered in the simulations. They are protein abundance, complexity of samples, proteolytic digestion efficiency, peptide separation and co-eluting peptides, scanning speed of the mass spectrometer, and dynamic exclusion efficiency. Other factors related to peptide ionization efficiency and ion suppression, which could influence the peptide identification, are not explicitly explored in this study (see Discussion section).

**Protein abundance.** Protein abundance is one of the most important factors that could cause the absence of peptides from the list of identified peptides. One of the sources of missed peptides is that they never make it into the top $c$ most intense peptide ions because there are other peptides with higher eluted abundance during the same sampling interval. Impacts on the detection probabilities of peptides and proteins from varying protein abundance are explored. The relationship between the detection probability and abundance is investigated. In addition, various compositions of differentially abundant proteins in mixtures would result in different detection probabilities for mixtures with a fixed number of proteins. Levels to be investigated are 1) proteins evenly distributed, mainly distributed at 2) low, 3) medium, and 4) high abundant levels in samples.

**Complexity of samples.** The number of proteins in the sample is one of the factors that could affect the sample complexity. We expect that a less complex protein mixture would show a higher detection probability than a more complex mixture. Levels to be investigated are set from 50 to 1000 by an increment of 50 proteins.

**Proteolytic digestion efficiency.** The efficiency of protein digestions is another factor that affects the sample complexity. The protein digestion is generally incomplete (Gatlin *et al.* 2000;

Kjeldsen *et al.* 2003). Thus, number of real tryptic peptides for detection is smaller than that from *in silico* digestion. For example, 1/5 theoretical tryptic peptides were used as candidate peptides for LC-MS/MS analysis (Xue *et al.* 2006). Levels to be investigated change from 5 to 25 by an increment of 5 peptides per protein in the simulation analysis.

**Peptide separation and co-eluting peptides.** The distribution of peptide mean retention time and standard deviation among multiple copies mainly determine the peptide separation and co-eluting peptides. A sequence-specific retention calculator (SSRCalc) showed that amino acid composition, position of the amino acid residues, peptide length, overall hydrophobicity, *p*I, nearest neighbor effect of charged side chains, and propensity to form helical structures are highly correlated with the protein retention time (Krokhin 2006). For an unknown peptide mixture, these properties are also unknown. We could simulate different patterns for peptide retention times from uniform distribution and gamma distribution with different parameters and specify standard deviations of retention time, and then test their influence on the detection probability. The result would provide practical guidelines for the improvement of LC, i.e. finding a better way to elute peptides into MS to achieve the maximum peptide/protein detection.

**Scanning speed of the mass spectrometer.** When different sampling interval applied, different number of peptides would pass into MS. The longer the sampling interval, the more peptides pass. Two aspects of scanning speed are investigated: 1) Varying sampling interval with a fixed number of the top most intense ions sampled in each cycle; and 2) Varying sampling interval with a fixed time (on average) of detecting a peptide ion in each cycle.

**Dynamic exclusion efficiency.** Higher detection probability will be achieved when previously detected peptides are excluded from current detection, comparing with the case

without exclusion. The simulations will provide quantitative assessments of the improvement of peptide/protein detection from including an exclusion list.

**4 Results**

In simulations, we used 200 proteins, each with 10 peptides, unless otherwise specified. The mean elution times are generated from Gamma(.) [Fig. 3.2], and the standard deviation of elution time is set as 1. The number of the top most intense ions detected in each cycle is 9. The 50 most recently detected peptides are excluded from further detection and thus are ignored when determining the top nine most intense peptides. Each simulation was repeated 100 times and we report summary characteristics of these simulations.

**Effects complexity of samples.** We first consider the simple case where the abundance is the same for all proteins. Bear in mind that we are modeling only the impact on detection from competition between co-eluting peptides and not any other factors. Figure 3.4 shows how the peptide detection probability varies with sample complexity. The probability of detection is about 92% when the sample complexity is 50 proteins eluting in 100 seconds. The probability of detection decreases dramatically as the sample complexity increases, to about 33% when then there are 200 proteins, 19% when there are 400 proteins, and 10% when then there are 800 proteins eluting in 100 seconds. These protein numbers of 50, 200, 400, and 800 scale to values on the order of 300, 1200, 2400, and 4800 proteins in a full experiment (see the section on "Values for other parameters" for more detail).

The standard deviation in elution time in the simulations is one second. If we assume that most of the elution occurs in four standard deviations, then the probability that a randomly selection peptide is eluting at a given point in time is 4/100=0.04. If there are a total of $A$ peptides that elute in the 100 second period, then the expected number eluting at any given point

in time is 0.04*A*. This gives 20, 80, 160 and 320 for the expected number of eluting peptides with 50, 200, 400, and 800 proteins, respectively (assuming ten peptides per protein). If the nine most intense peptides are detected in each scanning interval, then with no exclusion the probabilities of detection in a single scanning interval are 9/20=0.45, 9/80=0.113, 9/160=0.056, and 9/320=0.028, respectively. This demonstrates the fundamental point of this paper: that is, the competition for detection between peptides in complex samples is a major source of missed detections. The calculation is more complicated when previously detected peptides are excluded from future detection (as is the case in for these simulations). However, results later in this paper show that peptide exclusion does not have a major effect.

The distribution-outs (identified peptides) for samples with different complexity (number of proteins in mixtures) are shown in Figure 3.5. The distribution-ins of these samples is similar to the one shown in Figure 3.2. As we expected, when the complexity is low, the distribution-out will resemble the distribution-in. However, when the complexity is high, the tails of the retention time distribution will be identified at a higher rate than the center. This tends to make a gamma distribution look more uniform, but with a short left tail and relative long right tail.

Figure 3.6 shows how the protein detection probability depends on the sample complexity. We use two different definitions for protein identification. For the higher curve, the protein is identified if at least one of its peptides is identified. For the lower curve, the protein is identified if at least two of its peptides are identified. As expected, the protein detection probabilities are substantially higher then the individual peptide detection probabilities.

### 4.1 Effects of proteins abundance

#### 1) Varying abundance

In order to study the impact of varying protein abundance, we simulated samples with 200 proteins with abundance increasing from 50 to 1045 by increments of 5 ( i.e. abundance varied over a range with a fold change of 20).

Figure 3.7 shows the relationship between peptide detection probability and abundance. Each point shows the detection frequency (over 100 replicates) for a peptide. The X-axis is the abundance for the protein containing the peptide. We see that most of the high abundant peptides are detected 100% of the time and most of the low abundance proteins are detected none of the time. The distribution of detection probability of peptides shows that the detection probability of 0 is most common, with many 1.0's also, but little in between (Fig. 3.8). Nearly 60% peptides have a detection probability of 0 and about 25 percent of peptides have a detection probability of 1.

Certain peptides are detected 100% of the time even with very low abundance. These peptides have abnormal elution times - either very high or very low values (Fig. 3.9). This is easy to understand in that peptides with elution times quite different from the most will not have to compete with other peptides for detection and will therefore have a high detection probability. This effect is more apparent if we look at the case with no variation in abundance (with the abundance set at a fixed level of 100 copies). All peptides that were detected 100% of the time have abnormal elution times (Fig. 3.10).

The relationship between detection probability of peptides and abundance shows that for peptides at low and medium abundances, the detection probability depends almost entirely on the elution time and very little on the abundance (Figs. 3.7, 3.9). For the most abundant peptides,

there is a steep linear relationship between abundance and detection probability (Fig. 3.7). The Spearman's rank correlation increases from 0.12 to 0.69 for peptides in two categories of peptides - below and above the middle abundance.

The number of peptides needed to identify a protein is an important issue in proteomics studies. The larger number of peptides used to identify a protein, the higher the probability of correct protein identification. Simulation results showed that the average number of peptides detected per replicate for each protein (up to 10 since we assume 10 peptides per protein in simulations) shows that about 3/4 proteins are detected with at least one peptide, about 1/2 proteins are detected with at least two peptides, and about 1/5 proteins are not detected at all (Fig. 3.11).

The relationship between number of detected peptides and abundance shows that high abundance proteins are identified with multiple peptides and low abundant proteins by one or two, which agrees with the real MS experiments (Liu *et al.* 2004). In addition, for the most abundant proteins, there is a steep linear relationship between abundance and number of detected peptides per protein (Fig. 3.12).

Figure 3.13 shows the probability that proteins are detected as a function of abundance for two definitions of protein identification – P1) at least one peptide being identified in a protein, and P2) at least two peptides being identified in a protein are investigated. Under these two definitions of protein identification, the probabilities of the protein being detected show almost (>90%) entirely 0 or 1 (Fig. 3.13).

Since the peptide identifications determine the protein of its origin, the conclusion of the relationship between detection probability of proteins and abundance is similar to that of peptides, i.e. most of high abundant proteins and certain low abundant proteins are detected

100% of the time (Fig. 3.14). The Spearman's rank correlation between the detection probability of proteins and abundance is 0.57.

**2) Compositions of differentially abundant proteins**

For mixtures with a fixed number of proteins, different compositions of differentially abundant proteins in mixtures will result in different detection probabilities. We will investigate four different abundance distributions: 1) proteins are evenly distributed between abundance classes, and the protein distribution is dominated by 2) low, 3) medium, and 4) high abundance classes. The definition of low, medium, and high abundance class and allocation of proteins in three protein abundance classes (PAC) are shown in Table 3.1.

The overall peptide detection probability of the four investigated allocations of proteins shows only slight differences (~0.4%) in detection probability (Fig. 3.15).

When the protein identification definition of P1 applied, the overall protein detection probabilities, which are defined as the average proportion of identified proteins to all proteins in samples over multiple replicates, are similar for the four investigated allocations of proteins, with a value of about 0.74. While under definition of P2, the high-abundance-dominated sample produces the highest detection probability of 0.57, comparing to the other three allocations, which have a similar mean protein detection probability of 0.47 (Fig. 3.16 red line).

Under both definitions of protein identification, the protein detection probability is positively associated with the abundance. The detection probability at high abundant proteins is significantly higher than others, and the detection probability at medium abundant proteins is significantly higher than that at low class (Fig. 3.16). Regardless of allocations, the high abundant proteins are detected with very high probability. In low-abundance-dominated samples, the medium and low abundant proteins have a higher chance of being detected comparing with

non-low-abundance-dominated (i.e. high- and medium-abundance-dominated) samples. With samples dominated by low abundant peptides, the total numbers of high and medium abundant peptides are less than that in the non-low-abundance-dominated samples, which increases the detection chance of medium and low abundant peptides. In contrast, in high-abundance-dominated samples, the medium and low abundant proteins have the lower chance being detected comparing to non-high-abundance-dominated samples. When samples dominated by high abundant peptides, the competition for detection of low and medium abundant proteins would increase, consequently reduce the detecting chance of medium and low abundant proteins. In general, more proteins at high abundance in samples, the higher the protein detection probability.

**4.2 Effects of dynamic exclusion efficiency**

As we expected, the detection probability is higher with exclusion than without (Table 3.2). 34 more peptides are detected under the case of excluding previous detected peptides in current sampling for a fixed abundance of 100 per protein in samples. With varying abundance from 50 to 1045 by 5 copies for sample proteins samples, 150 more peptides are detected when an exclusion list is applied. The new protein identifications are 15 and 47 under the fixed and varying abundance cases, respectively.

Although exclusion does have significant effect, it is smaller than might be expected. The reason for this is that under the parameter values assumed here, most peptides on the exclusion list are not currently eluting. Results that will appear in a future publication (Schliekelman, unpublished) show that the expected number of peptides eluting in a given interval that are on the exclusion list can be approximated as $c(\tau -1)/2$, where $\tau$ is the period over which most (loosely defined) elution occurs. If we take $c$=9 and $\tau$ equal to four standard deviations, then this gives 13.5 as the expected number of currently eluting peptides that are have been previously

detected and are on the exclusion list. Because this number is small compared to the number of eluting peptides for more complex sample, exclusion does not have a major effect in reducing competition for detection.

### 4.3 Effects of other factors

The investigations of impacts on the detection probability from other factors such as proteolytic digestion efficiency, peptide separation and co-eluting peptides, and scanning speed of the mass spectrometer, are based on the simulations at a fixed abundance of 100 per proteins.

**Peptide separation and co-eluting peptides.** Mean elution times for 2000 peptides (200 proteins and 10 peptides per protein) were generated from 5 different distributions: 1) Normal (50,15); 2) Gamma(.); 3) Uniform with two relative long tails, generating from Gamma(.) and replacing [12,78] part with Uniform, abbreviated as GU(12,78); 4) Uniform with two relative short tails, generating from Gamma(.) and replacing [6,84] part with Uniform, abbreviated as GU(6,84); and 5) Uniform (3,99). With this setting, we construct samples (mean elution times) with an increasing proportion of uniform part in the entire sampling period (from zero of Gamma to 100 percent of Uniform). The distribution-ins and outs for mean elution times from 5 distributions are shown in Figure 3.17. The distribution-outs from simulations provide indirect support about the distribution-ins. That is, the distribution-in is probably Gamma distributed since its corresponding distribute-out is more close to the real ones (e.g. Pfeifer's 321-peptide data). In contrast, if the distribution-in exactly follows Uniform distribution, it would result in a uniformly distributed out, which do not reflect the real experimental results.

With an increasing proportion of Uniform part in the entire sampling period, the mean peptide detection probability increases (Fig. 3.18). The mean peptide detection probability from the uniformly eluted peptides is the highest. The mean peptide detection probability from

Gamma eluted times is slightly larger than that eluted according to Normal distribution. This may be caused by the contribution of the long right tail of Gamma distribution since the peptides have a high chance being identified due to the less competition at tails.

For both definitions of protein identifications, the mean protein detection probabilities show an increasing trend as that of peptide (Fig. 3.19). However, with a more restricted definition of correct protein identification (P2), the mean protein detection probability decreases with different magnitudes for different distribution-ins – with a 17% reduction in Normal and Gamma and an about 6% decrease in Uniform.

Standard deviation of retention time measures the spread of a given peptide being eluted. While the mean retention time give the time point that a given type of peptide being eluted, the standard deviation of retention time gives the variation of retention time for each copy of that given peptide. Within each scanning cycle, the length of a given peptide being "fully" eluted may influence the results of the peptide identification. A new kernel-based model reflects the fact that retention times of later eluting peptides show a higher deviation than early eluting ones (Pfeifer *et al.* 2007). Five standard deviations are investigated here: 0.5, 1.0, 1.5, 2.0, and a mixed one with 1 for the mean elution time less than or equal to half of the entire sampling period and 2 otherwise. The results show that the peptide and protein detection probabilities for various SDs are similar and the differences are negligible (Appendix A3.2).

**Effects of proteolytic digestion efficiency.** The proteolytic digestion efficiency leads to different number of peptides per protein after the digestion. In general, the protein digestion is incomplete and number of real tryptic peptides for detection is smaller than that from theoretical digestion. With an increasing number of peptides per protein, the total number of peptides in samples increases significantly. Therefore, the mean peptide detection probability goes down. In

contrast, with an increasing number of peptides per protein, the chance of at least one or two peptides in any protein being identified increases, which further result in an increase in the protein detection probability (Appendix A3.3).

**Effects of scanning speed of the mass spectrometer.** Under varying sampling interval with a fixed number of the top most intense ions sampled in each cycle, the mean peptide detection probability decreases with an increasing sampling interval. This implies that we can increase the number of detected peptides/proteins via the improvement of scanning speed of the mass spectrometer. However, in real experiments, there is a tradeoff between the length of sampling interval and number of the top most intense ions being sampled for any cycle. The bottom line is that the time interval should be long enough for the specified number of ions being sampled. When varying sampling interval with a fixed time for detecting a peptide ion in each cycle is applied, the detection probabilities between two strategies, short sampling interval and small number of sampled ions and long sampling interval and large number of sampled ions are negligible (Appendix A3.4).

## 5 Discussions

The goal of this study was to create a framework for investigating the potential impacts of the individual factors involved in the LC-MS/MS process on peptide/protein detections. At present, little is known about how these factors contribute to variation in protein detection. Better understanding these sources of variation will allow optimization of experimental protocols and statistical procedures. The focus in this study is the dynamics of the “competition” for detection between peptides simultaneously eluting from the LC. Because of the difficulties in measuring this phenomenon experimentally, a mathematically/simulation approach is a good starting point

for clarifying the issues. However, experimental explorations will be required for full understanding and verification.

The reason for this competition is that the MS/MS scan can only detect a limited number of peptides at a time. Simple calculations (see Results) show that the number of simultaneously eluting peptides will often greatly exceed this number. A peptide will only be detected if its elution intensity is in the top *c* values in some scanning interval. Our simulations show that this will never occur for many peptides in a complex mixture. When the number of proteins in the mixture is less than about 500, then most peptides will make the top c at some point during the period in which they are eluting. However, when the number of proteins is on the order of 1000 or more, then many peptides will not be detected even if the LC-MS/MS process works with perfect efficiency in all other respects.

Our results show that the characteristic elution time for a peptide is a primary determinant of whether it will be detected. Peptides with abnormal elution times (either unusually high or low) will be detected with very high probability, while peptides that elute at the same time as many other peptides have a much lower detection probability. Our results show that for peptides of low and intermediate abundance, elution time is a more important determinant of detection probability than abundance is.

It has been observed that many proteins are repeatedly and consistently identified by a small number of peptides. That is, these peptides have much higher probability of being detected than other peptides in the protein. Mallick *et al* (2007) termed such peptides as proteotypic peptides. The existence of proteotypic peptide can be easily explained using our results: proteotypic peptides are those that have abnormal elution times and thus elute when few other peptides are doing so. This is consistent with the Mallick's prediction model in which the hydrophobicity,

which is directly related to the retention time in LC separation, is one of the most important predictors. Although the peptides present will change between different samples, the distribution of elution times is not likely to change substantially and thus an abnormal elution time in one sample will be likely to be abnormal in another sample. Thus, proteotypic properties would be expected to be consistent across different samples (as has been observed). Different experimental protocols may produce different elution time distributions and so proteotypic properties would be expected to change between experimental platforms. Of course, other factors (e.g. ionization efficiency) likely also contribute to determining whether a peptide is proteotypic. However, we conjecture that elution time is a major component.

Our results make some important predictions about the relationship between detection probability and protein abundance. Most striking is the essentially binary relationship between detection probability and abundance (Fig. 3.7). That is, proteins of lower abundance have essentially zero probability of detection unless they have a peptide of abnormal elution time. In this case, there detection probability is essentially one. High abundance proteins have a detection probability of essentially one. There is only a narrow abundance range over which there is a linear relationship between abundance and detection probability. The sizes of these different ranges will depend on the distribution of abundances in the sample and thus may be quite different from Figure 3.7. However, the basic pattern should hold for most samples.

Thus, our results predict that detection probability is a poor predictor of protein abundance. Spectral count is a better measure, but still has problems. A low abundance protein is only likely to produce spectra for proteotypic peptides, while a high abundance protein should produce spectra for multiple peptides. Thus, on average, spectral count should be higher in higher abundance proteins. However, many proteins likely have multiple proteotypic peptides. Thus, a

low abundance protein that happened to have 2-3 proteotypic peptides could produce a rather high spectral count. Likewise a high abundance protein could produce a relatively low spectral count if all or most of its peptides happened to have common elution times. Things are further complicated by other factors such as ionization efficiency. Each of these factors adds another layer of variation unrelated abundance.

Other functions of the spectral count may perform better. A protein with non-zero spectral count across many of its peptides is more likely to be high abundance than one with a high spectral count for one or two peptides. Thus, for example, the median or some other quantile of spectral count (across peptides of the protein) may be a better measure because it won't be influenced by a small number of proteotypic peptides.

**Implications for biomarker discovery.** The goal of biomarker discovery is to find proteins that are differentially expressed between different disease states or other physiological conditions. This requires being able to measure protein abundance (or at least presence/absence) accurately. Our results predict that measures of abundance such as spectral count and detection probability are only gross indicators of abundance and are subject to substantial variation due to elution time and other factors. Thus, the ability to measure differences in abundance between different treatment conditions is also limited.

Of course, the general solution to variation is replication and these problems can be at least partially overcome by sufficient replication. Because of the time and expense involved, replication has usually been limited in proteomics studies. We and others (Koziol *et al.* 2006) have shown that well over 50 replicates may be needed to obtain good coverage in establishing which proteins are present in a sample. Establishing differences in protein abundance will require far more replicates. Even then, replication will not solve all of the problems because not all of

the unwanted spectral count variation is random in nature. Characteristics such as elution time and ionization efficiency are properties of the peptides themselves and don't change between replicates. A peptide with an abnormal elution time will produce a high spectral count in every replicate.

Another issue is that the spectral count for a peptide depends not only on the abundance of peptide itself, but also on the abundances of peptides with similar elution times. Thus, if the spectral count for a peptide changes between treatment conditions, it could be either because the peptide's abundance is different or because the abundance of simultaneously eluting peptides are different. It is not clear how significant this issue should be expected to be. The probability that co-eluting peptides would both have different abundance be treatment conditions would be small unless many proteins had different abundances. This probability would be even smaller if there multiple peptides from the protein showing changes in spectral count.

It may be possible to improve statistical power by accounting for sources of systematic variation. For example, peptide elution times can be obtained from LC-MS/MS output. Taking this information into account may improve inference of protein abundance. We could, for example, adjust the spectral count for a peptide according to the average spectral count of peptides with similar elution times.

**Caveats.** As with any mathematical model, our model is based on assumptions. The results of any model are only valid to the extent that its assumptions are reasonable. The structure of our model is very simple. Each peptide $j$ has some number $A_j$ copies that elute at times sampled from a normal distribution whose mean varies from peptide to peptide. If the number of eluting copies is ever in the top $c$, then the peptide is detected.

The most important question is whether the basic structure of the model is reasonable. The first point to be made is that we are only modeling the impact of elution time distributions on peptide detection. We have not looked at other factors such as ionization efficiency, the probability that the database search correctly identifies a peptide. If these factors are independent from the elution time distribution, then it is reasonable to model them separately. However, if there are correlations then this complicates the analysis. Future work should further examine the relationship between elution time and other factors.

In our formulation the distributions are correlated because of their time dependence, but independent when conditioned on time (both between different copies of the same peptide and between different peptides). This implies that there is no interaction between the peptide copies in the LC or at any other point in the process. Such interactions could change the dynamics substantially.

With these caveats, we don't know of any reason why the basic model structure is unreasonable. However, there is no way to test the model without a designed experiment because we can't observe the elution times of peptides that aren't detected.

The next consideration is whether the specific distributions that we have used are reasonable. We have data (Petritis *et al.* 2006; Pfeifer *et al.* 2007) available for the distribution of mean retention times. The data appear to be consistent with the gamma distribution that we assumed, although they also bear some resemblance to a uniform distribution with tails. Whichever distribution is correct, our simulations show that the choice of distribution does not have a major impact on the results. We don't have any data available for the distribution of elution times for the copies of an individual peptide. We that assume that these elution times follow a normal distribution with mean sampled from the above gamma and constant variance. We don't have

any way to verify this distribution. However, we expect our results to be fairly robust to the choice of this distribution, provided that the true distribution has a single peak and two tails.

The final consideration is that of parameter values. We know the time between scanning intervals exactly and we have a good idea of the total length of the experiment (which for our purposes is the period over which most elution occurs). We were able to estimate the parameters of the distribution for mean elution times from the data of Petritis *et al.* (2006) and Pfeifer *et al.* (2007), but we don't know the extent to which these parameters very between different experimental protocols or biological sample types. The mean elution time for each peptide was sampled from this distribution. The standard deviation of elution time initially appears problematic. We could obtain only a very rough estimate of the mean standard deviation of elution time from the data of Pfeifer *et al*, and little information on the distribution of standard deviations. Thus, we assigned the same somewhat arbitrary standard deviation of elution time to all peptides. Fortunately, our simulations show that the peptide detection probability is very nearly independent of that standard deviation of elution time. This is because increasing the standard deviation has two opposing effects. First, it increases the length of the time that the peptide is eluting and thus increases the number of opportunities to be in the top *c*. However, increasing standard deviation also decreases the number of copies of the peptide eluting in any one scanning interval. Results that will appear in a future publication show that these two effects are both roughly linear in the standard deviation and thus cancel out. Because of this fortuitous effect, the only parameter that we don't have good values for doesn't impact the results.

Table 3.1: Definition of low, medium, and high class and allocation of proteins in three protein abundance classes (PAC)

| Allocation | High (1000 copies) | Medium (300 copies) | Low (50 copies) |
|---|---|---|---|
| Evenly | 1/3 | 1/3 | 1/3 |
| Low-abundance-dominated | 1/4 | 1/4 | 1/2 |
| Medium-abundance-dominated | 1/4 | 1/2 | 1/4 |
| High-abundance-dominated | 1/2 | 1/4 | 1/4 |

Table 3.2: Peptide and protein identification under the cases of with and without exclusions for fix and varying abundance of proteins in samples[1]

| Detection | Fix abundance | | | Varying abundance | | |
|---|---|---|---|---|---|---|
| | Exclusion Off | Exclusion On | Difference | Exclusion Off | Exclusion On | Difference |
| Peptides | 619 | 653 | 34 | 500 | 650 | 150 |
| Proteins[2] | 124 | 139 | 15 | 97 | 144 | 47 |

[1] 200 proteins and 10 peptides per protein are in samples.

[2] Protein detections refer to proteins identified in more than 95% of time in 100 repeated simulations.

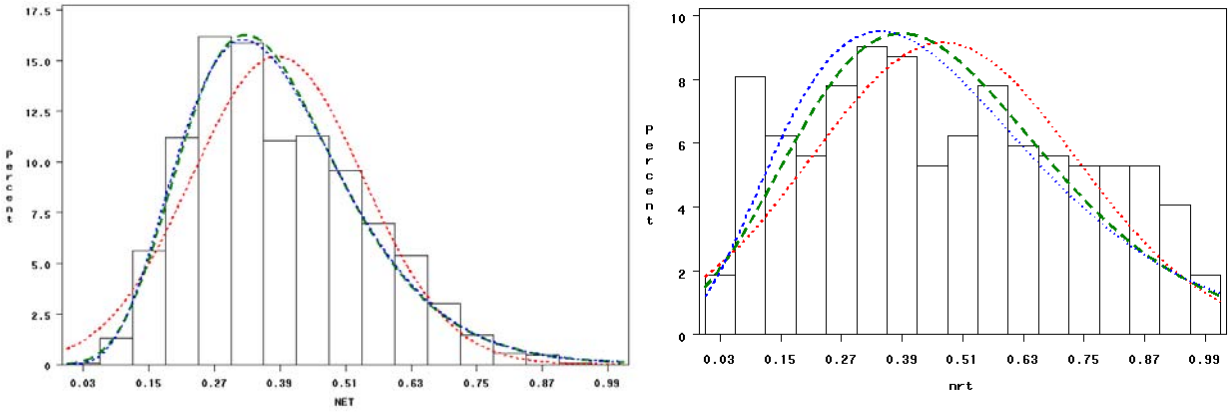



Figure 3.1: Petritis's 1303-peptide data (left) and Pfeifer's 321-peptide data (right) was fit against gamma, lognormal, and normal distributions. Red – Normal, green – Lognormal, blue – Gamma.

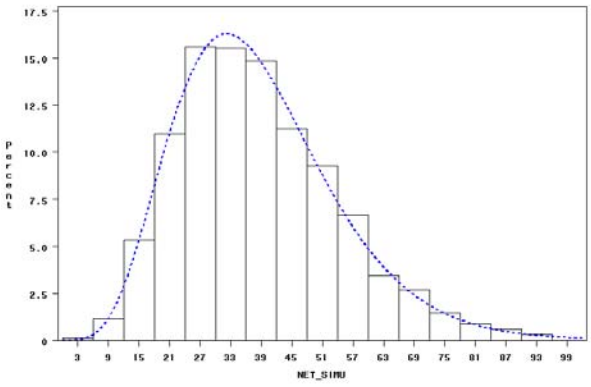



Figure 3.2: The retention time distribution of 2000 simulated peptides from a gamma distribution of threshold =-3.0541, scale=6.1134, shape=6.8060.

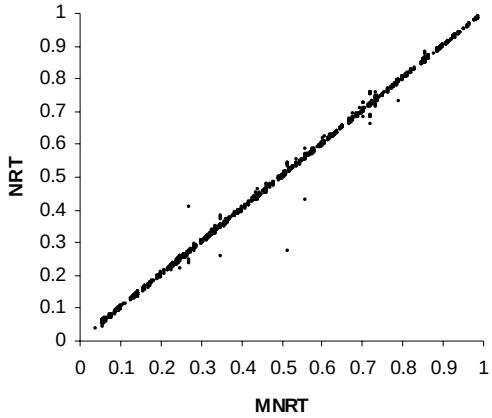



Figure 3.3: The scatter plot of mean normalized retention time (MNRT) against normalized retention time (NRT) for all spectra from Pfeifer's 321-peptide data.

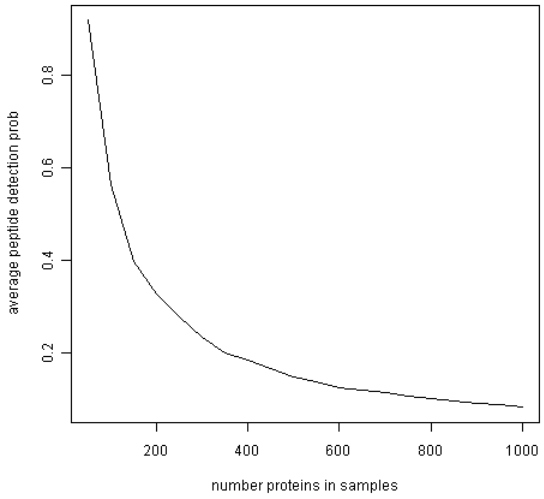



Figure 3.4: Mean peptide detection probabilities for samples with different complexity – protein number changes from 50 to 1000 by an increment of 50.

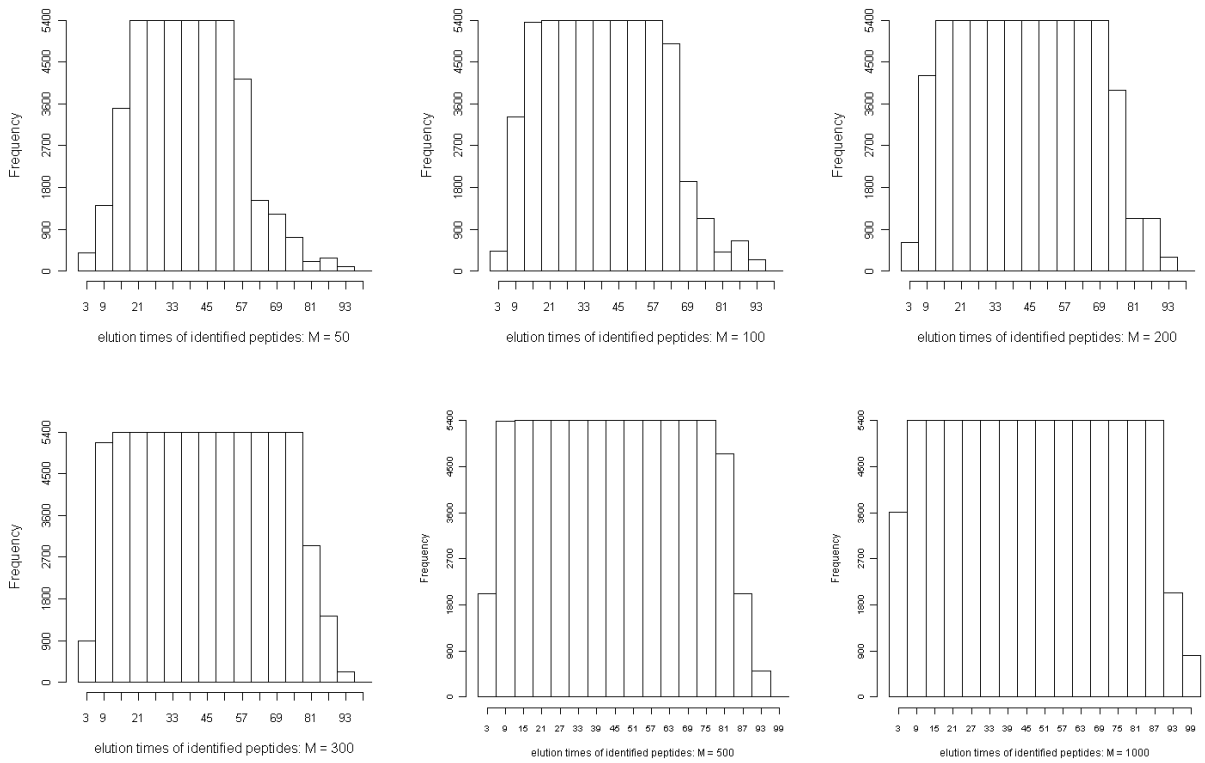



Figure 3.5: Distribution-outs (identified peptides) of samples with different complexity – protein number changing from 50 to 1000.

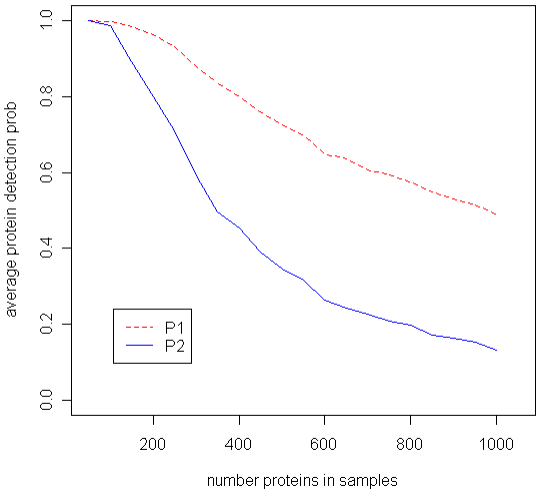



Figure 3.6: Mean protein detection probabilities for samples with different complexity under two definitions of protein identification - P1) at least one peptide being identified in a protein, and P2) at least two peptides being identified in a protein.

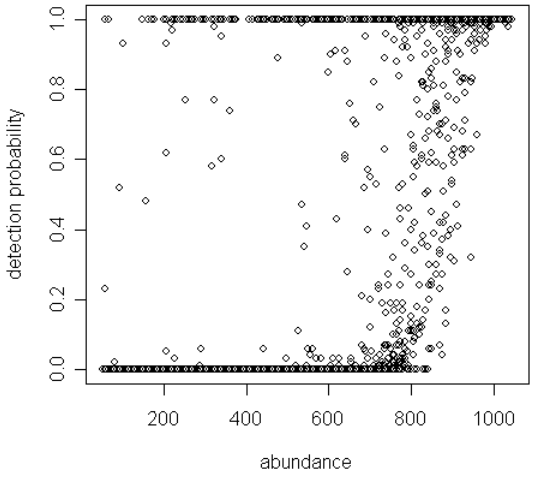



Figure 3.7: Relationship between detection probability of peptides and abundance.

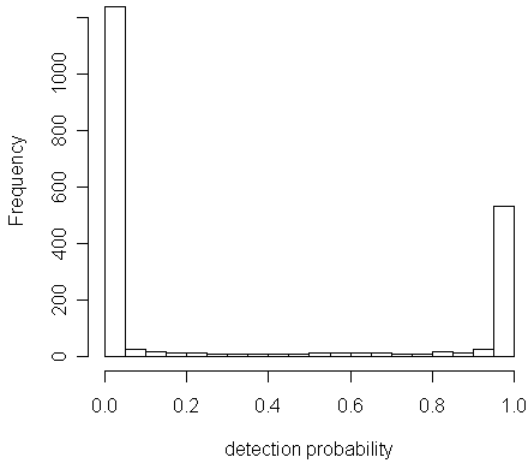



Figure 3.8: Distribution of detection probability of peptides.

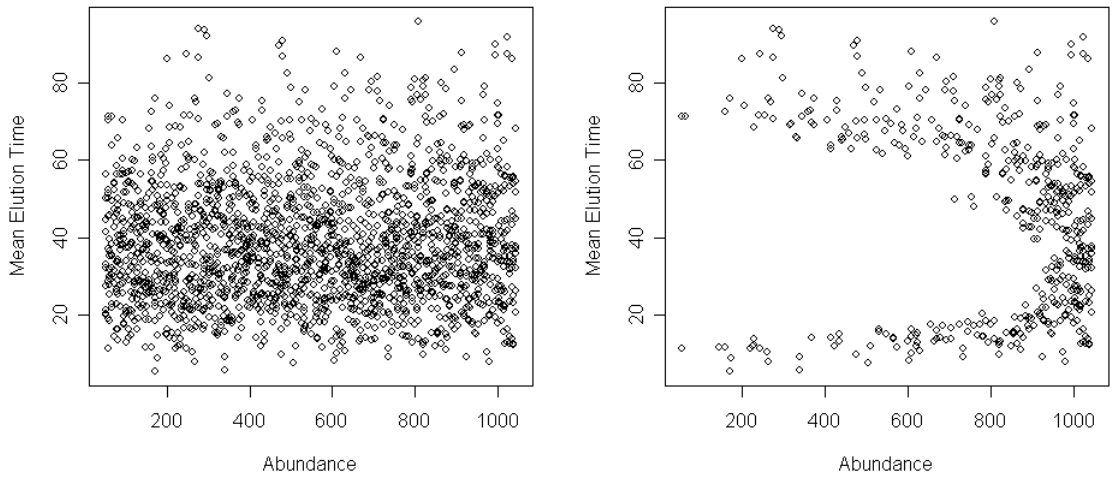



Figure 3.9: Distributions of mean elution time for all peptides in samples (left) and peptides detected 100% of the time (right) for varying protein abundance from 50 to 1045 by 5 copies in samples.

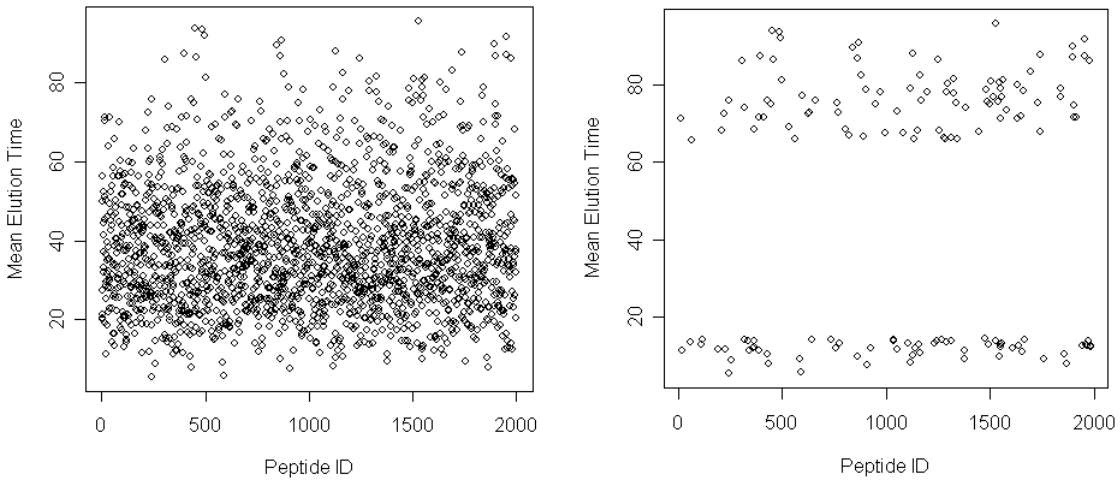



Figure 3.10: Distributions of mean elution time for all peptides in samples (left) and peptides detected 100% of the time (right) for a fixed protein abundance of 100 in samples.

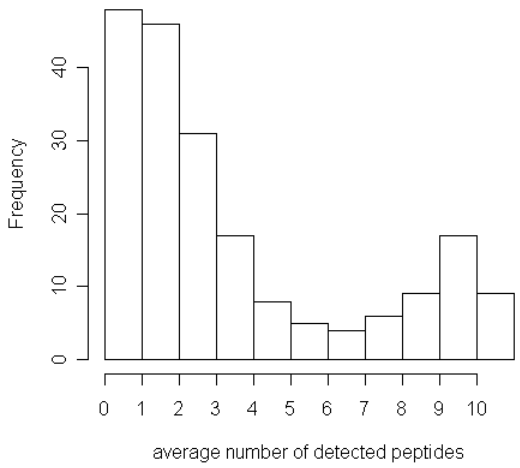



Figure 3.11: Distribution of average number of peptides detected per replicate for each protein.

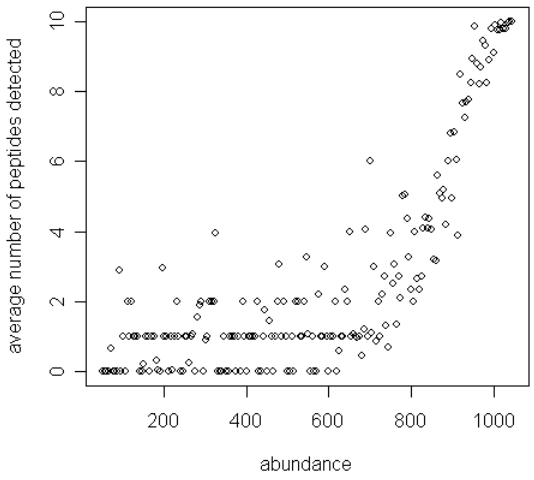
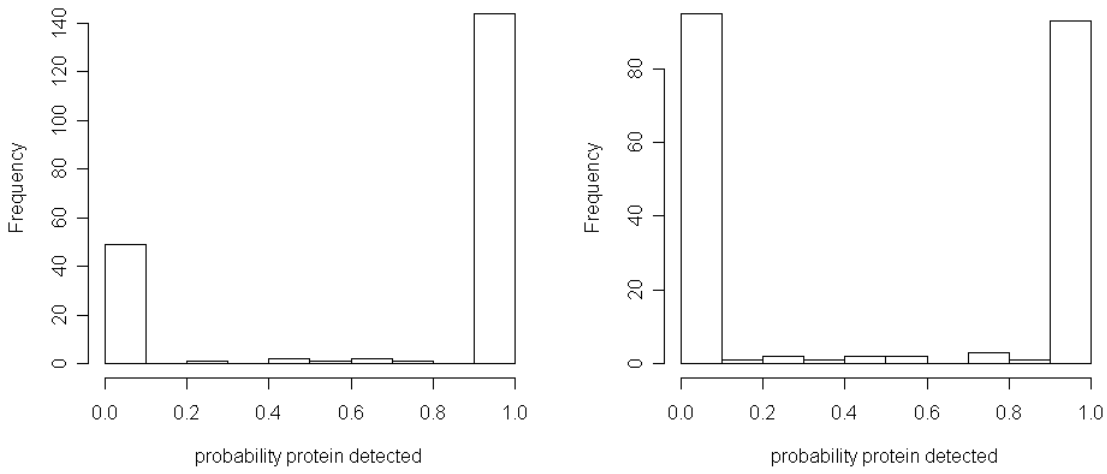
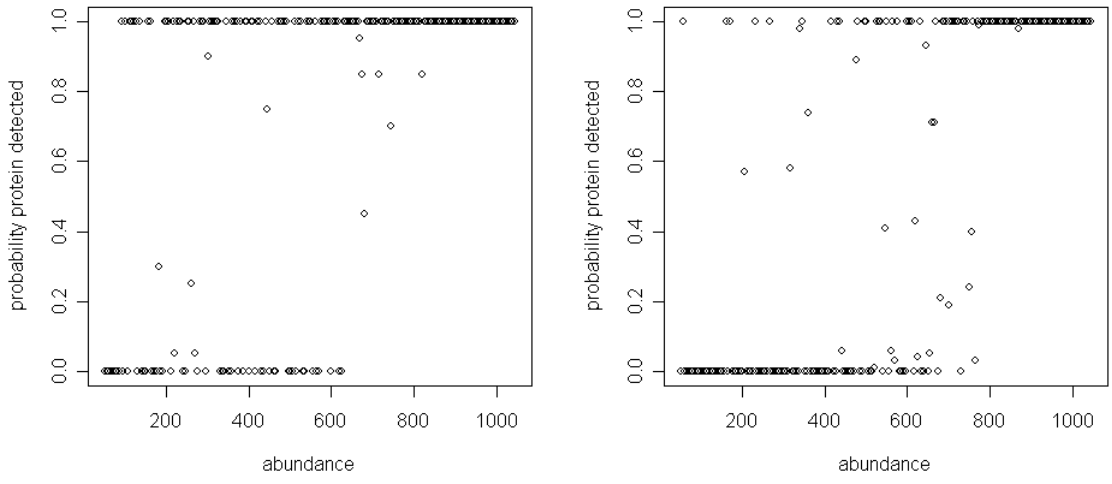



Figure 3.12: Relationship between average number of detected peptides per protein and abundance.




P1 P2

Figure 3.13: Distribution of the detection probability of proteins under two definitions of protein identification - P1) at least one peptide being identified in a protein, and P2) at least two peptides being identified in a protein.




P1 P2

Figure 3.14: Relationship between detection probability of proteins and abundance under two definitions of protein identification - P1) at least one peptide being identified in a protein, and P2) at least two peptides being identified in a protein.

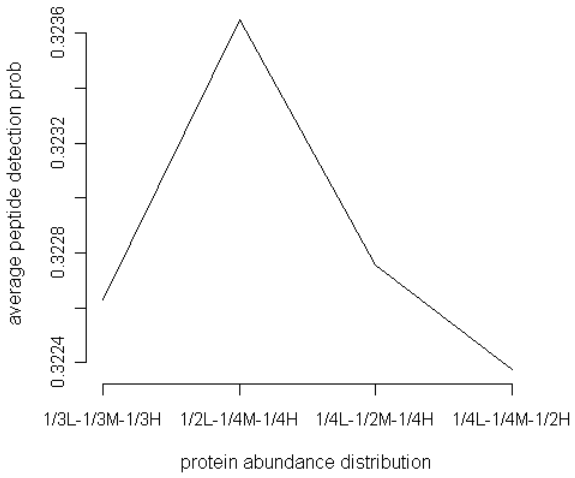
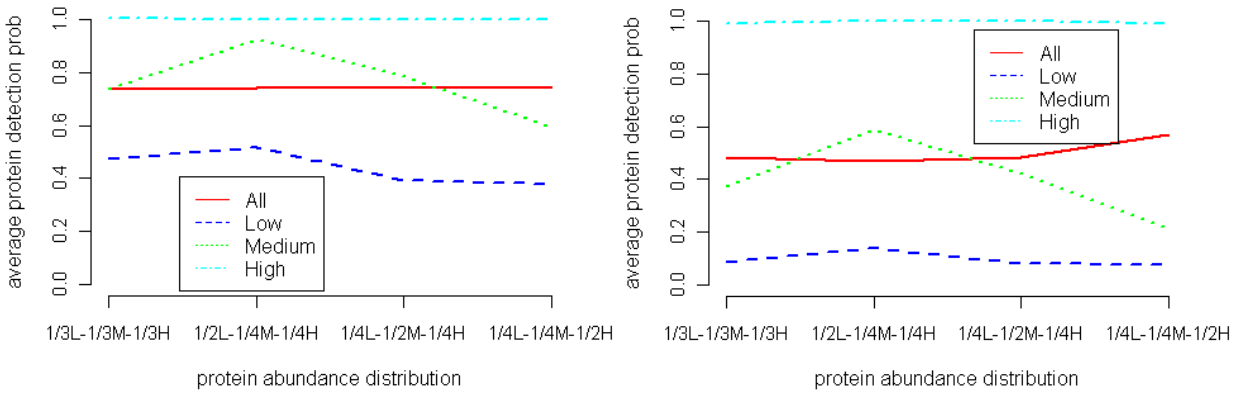



Figure 3.15: Mean peptide detection probability of different combinations of peptides with different abundance.




P1 P2

Figure 3.16: Mean protein detection probabilities by protein abundance class (PAC) under two definitions of protein identification - P1) at least one peptide being identified in a protein, and P2) at least two peptides being identified in a protein.

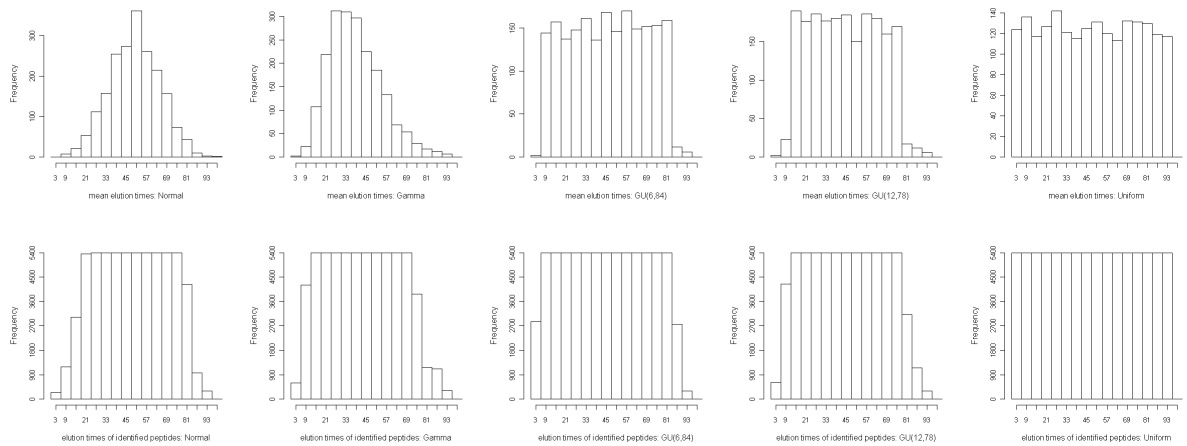



Figure 3.17: Distribution-outs (bottom panel) from their corresponded distribution-ins (top panel) – Normal (50,15); Gamma(.); Uniform with two relative long tails, generating from Gamma(.) and replacing [12,78] part with Uniform, abbreviated as GU(12,78); Uniform with two relative short tails, generating from Gamma(.) and replacing [6,84] part with Uniform, abbreviated as GU(6,84); and Uniform (3,99).

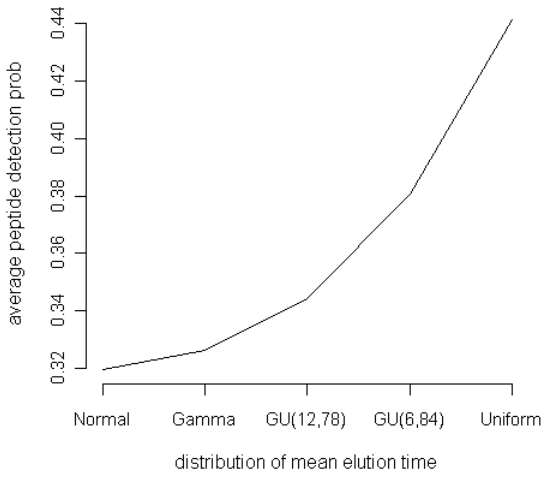



Figure 3.18: The mean peptide detection probabilities for different distribution-ins – Normal (50,15); Gamma(.); Uniform with two relative long tails, generating from Gamma(.) and replacing [12,78] part with Uniform, abbreviated as GU(12,78); Uniform with two relative short tails, generating from Gamma(.) and replacing [6,84] part with Uniform, abbreviated as GU(6,84); and Uniform (3,99).

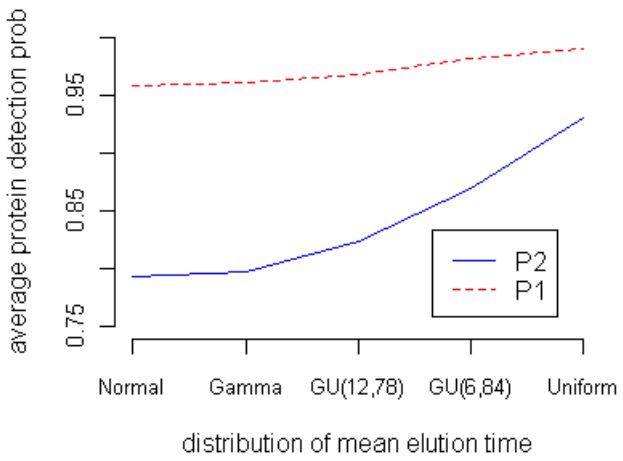



Figure 3.19: The mean protein detection probabilities for different distribution-ins under two definitions of protein identification - P1) at least one peptide being identified in a protein, and P2) at least two peptides being identified in a protein.

# CHAPTER 4

# COMPARISON OF PROTEIN IDENTIFICATIONS IN TWO SAMPLES - DIFFERENTIALLY EXPRESSED OR NOT?[*]

[*] Liu S, Orlando R, and Schliekelman P. To be submitted to Statistical Application in Genetics and Molecular Biology.

**Abstract**

One of the common and versatile uses of high-throughput mass spectrometry (MS)-based proteomics has been to compare the protein expression profiles in two different types of biological samples. Typically, the non-occurrence of a specific sequence in a list of identified peptides does not necessarily indicate that the peptide or protein was not originally present in the sample. We propose a methodology to conduct statistical test of differential expression of proteins detected in two different samples. By combining the test results from the spectral counts and protein occurrence based methods on the basis of multiple runs of MS data, significantly differentially expressed proteins with high confidence in two different treatments can be obtained.

**Keywords:** Statistical application; Mass spectrometry (MS); MS-based proteomics; Protein identification; Protein quantitation; Protein differential expression

## 1 Introduction

Using the common mass spectrometry (MS)-based proteomic pipelines (Aebersold & Mann 2003), the complex protein mixtures can now be routinely characterized in depth, with thousands of proteins detected for suitably complex samples. One of the common and versatile uses of high-throughput MS-based proteomics has been to compare the protein expression profiles in two different types of biological samples, e.g. normal vs. diseased tissues (Rifai *et al.* 2006; Smit *et al.* 2007). Analyses of complex protein mixtures are often not comprehensive (Wilkins *et al.* 2006; Malmstrom *et al.* 2007). In addition, the process of spectrum detection depends on a multitude of variables that are difficult to control (Aebersold & Mann 2003). Many factors could cause the absence of peptides from the list of identified peptides. Besides the influence of low protein abundance (Aebersold & Mann 2003), factors such as incomplete protein digestion

(Gatlin *et al.* 2000; Kjeldsen *et al.* 2003), physicochemical properties of peptides (Breci *et al.* 2003; Nielsen *et al.* 2004; Ocaña *et al.* 2005), ionization response of peptides (Dobo & Kaltashov 2001; Peschke *et al.* 2002; Pan & McLuckey 2003; Pan *et al.* 2004) and data-dependent MS/MS acquisitions of peptide ions (Liu *et al.* 2004) influence the peptide/protein identification. Therefore, the non-occurrence of a specific sequence in a list of identified peptides does not necessarily indicate that the peptide or protein was not originally present in the sample. Consequently, a simple list of proteins detected in different states is insufficient to make such analysis meaningful (Aebersold & Mann 2003).

In order to obtain the estimation of relative protein expression (comparison of two sample), Z-scores of differential expression for each protein were calculated on the basis of the fraction of interpreted peptides accounted for by each protein in the experiment under the assumption that the probability of observing each peptide in the mass spectrometer is constant between two samples and can be ignored (Lu *et al.* 2007). This method, together with the approach to calculating the absolute protein abundance, has been published as a Nature protocol for estimation of relative protein expression (Vogel and Marcotte, 2008). Their results showed that protein of high abundance in two different samples could be significantly differentially expressed even if the actual expression fold change (the ratio of absolute protein expression index [APEX] in the two mentioned samples) is small (Vogel and Marcotte, 2008). If APEX estimations are accurate and reliable, it implies that the spectral counts from single experiment are not sufficient as an indicator of protein abundance, at least for high abundant proteins. In this study, we propose a methodology to conduct statistical test of differential expression of proteins detected in two samples on the basis of repeated MS runs, which produces a high confidence of the significantly differential expression of proteins. An alternative method for the estimation of

relative protein expression is further proposed on the basis of numbers of occurrence of each protein in multiple replicates.

## 2 Data and method

Three repeated MS experiments were conducted for each of two protein complexes (strawberry data in Shah *et al.*, submitted to JPR) and proteins that are present in each sample were recorded separately. After filtering the high-confidence set of protein identifications using 5% false discovery rate [FDR] (Benjamini & Hochberg 1995) for each replicate, each identified protein is associated with the corresponding spectral counts detected in all replicates for each sample.

Comparing the identified proteins in the two samples, we obtain three types of proteins: 1) proteins present in one sample but not the other; 2) proteins present in both samples but with different frequencies; and 3) proteins present in two samples with similar frequencies. For each identified protein, we want to conduct a test of null hypothesis that the protein is actually present in both samples with equal abundance (denote as Hyp1). To do this, assume $r_1$ and $r_2$ replicates are conducted for two samples $S_1$ and $S_2$, respectively. For analytical proteomics experiments, replicates are very small in general (e.g. triplicates in this example).

### 2.1 Test of significance between two independent binomial proportions – spectral counts based method

The number of MS/MS spectra associated with a given protein ( $x$ ) out of all spectra ( $n$ ) can be treated as Bernoulli trial in which a spectrum is either associated with that protein or not. With $n$ being large enough (typical values of $n$ is more than 5000), the central limit theorem (CLT) can therefore be applied. In the case of multiple replicates, $x$ and $n$ represent the corresponding

average number. Once we obtain the fraction of spectral counts ( $f = \frac{x}{n}$ ) for each protein in a sample, the hypothesis test of Hyp1 would be equivalent to test the equality of two independent binomial proportions. Of course, one of the proportions has an estimate of zero (and therefore an estimated variance of zero) for the case of "in one but not the other". We make the assumption that the probability of observing each peptide in the mass spectrometer is different between two samples instead of constant applied in Lu *et al.* (2007). The non-constant assumption is more realistic since two samples from the same tissue but at different states (e.g. normal vs. diseased) more likely belong to two different populations. A widely used test statistic for two independent binomial proportions under normal approximation is given by

$$Z = \frac{f_{i,1} - f_{i,2}}{\sqrt{f_{i,1}(1-f_{i,1})/n_1 + f_{i,2}(1-f_{i,2})/n_2}} \quad (1)$$

where the numerator represents the difference in proportion $f$ for protein $i$ between two samples, and the denominator represents the standard error of the difference.

**2.2 Multiple testing corrections**

Regardless of how to conduct the test, we have to make a multiple testing correction because we will be doing it for each protein, while multiple proteins detected in the experiment are tested simultaneously. A high-throughput proteomic experiment usually identifies hundreds of proteins. Therefore, the analysis of these data sets also involves tests on hundreds of hypothesis, but only portion of them are significant. In such cases, traditional p-value of controlling the family-wise error rate (FWER), which guards against at least one false positive among all tests, is typically going to be too strict and will lead to many missed findings (Storey 2002). Analogous to a p-value, a q-value (Storey 2002) is a well-suited measure of significance when the goal is to identify as many significant features as possible, while incurring a relatively low proportion of

false positives. However, the q-value assigns significance in terms of the false discovery rate (FDR), while the p-value assigns significance in terms of the false positive rate (FPR). The distinction between FPR and FDR is critical. For a given significance threshold, FPR is the probability that a null statistic is significant. While FDR is the expected proportion of null statistics among all statistics called significant. i.e., the proportion of false positives among all accept alternative hypothesis (Storey & Tibshirani 2003).

**2.3 Alternative approach for testing the hypothesis that a protein is not present in one treatment – protein occurrence based method**

We conduct $r$ replicates for each sample and record proteins being present in one sample and not the other and calculate an estimate for $p$, the probability of a given protein being detected in a replicate, from the sample where the protein was detected. One approach (called protein occurrence based method) would be then to take this estimate as "correct" and then calculate the probability that the protein would fail to be detected in $r$ replicates in the other sample. This would then be the *p*-value for testing the hypothesis that a protein is not present in one treatment. If the number of replicates is moderate large, this method would produce a reliable estimation. But even under the case of several replicates, we recommend examining the results from this method and the spectral counts based method described in section 2.1 to get more reliable estimation.

**3 Results**

Comparing the identified proteins in the two samples, 70 and 150 unique proteins were identified in two samples, respectively. Totally 172 unique proteins were detected, in which 48 proteins being present in both samples, 22 proteins being present in sample 1 but not in sample 2, and 102 proteins being present in sample 2 but not in sample 1 (Table 4.1).

### 3.1 Spectral counts based method

For comparison, Z-scores based on both non-constant (Eq. 1) and constant assumptions (Lu *et al.* 2007) are calculated, following which the multiple testing corrections are conducted to get the q-values. The numbers of significantly differential expressed proteins under two FDRs, 0.01 and 0.05, are listed in Table 4.1.

As a FDR of 0.01 is applied, 28 proteins are significantly differential expressed in two samples totally, in which 14 proteins being identified in both samples, 3 proteins in sample 1 but not sample 2, and 11 proteins in sample 2 but not sample 1. The significantly differentially expressed proteins are almost identical under both non-constant and constant assumptions, except for one protein, BC1G_07315.1, which is not significantly differentially expressed under non-constant assumption (Table 4.1).

When a FDR of 0.05 is applied, 64 proteins are significantly differential expressed in two samples in total, in which 20 proteins being identified in both samples, 10 proteins in sample 1 but not sample 2, and 34 proteins in sample 2 but not sample 1. The differences of significantly differential expressed proteins under non-constant and constant assumptions are mainly in the category of sample 2 but not sample 1, in which 11 more proteins are labeled as significant under the constant assumption (Table 4.1).

Figure 4.1 shows the average spectral counts detected in samples of two treatments for significantly differentially expressed proteins using two FDRs. As a FDR of 0.01 is applied, 15 proteins are up-regulation (the average detected spectral count in sample 2 is more than that in sample 1). The rest of 13 proteins are down-regulation (the average detected spectral count in sample 2 is less than that in sample 1). As a FDR of 0.05 is applied, 40 and 24 proteins are up- and down-regulation, respectively.

### 3.2 Protein occurrence based method

Totally 124 unique proteins were detected in one but not the other, in which 21 and 57 proteins are not present in one treatment with significance levels of 0.01 and 0.05, respectively (see Appendix Table A4.1).

If the number of replicates is large enough, the method would produce a reliable estimation. But under the case of several replicates, we recommend examining the results from both this method and the one based on the spectral counts mentioned above to get more reliable estimation. In the 44 significantly differential expressed proteins (FDR of 0.05) from the spectral counts based method, 36 proteins overlapped with 57 proteins according to the protein occurrence based method (see Appendix Table A4.1).

## 4 Discussions

In this study, multiple runs of MS experiments were conducted and results were used for testing differential expression of proteins in samples from two treatments. Replicated data have at least two advantages: 1) increasing the chance of detection of proteins in samples, thus the "extra" proteins would be identified compared to the single run. Analyses of complex protein mixtures are often not comprehensive (Wilkins *et al.* 2006; Malmstrom *et al.* 2007) and peptide/protein identifications are often incomplete (Liu *et al.* 2004) in single run of MS experiment since the process of spectrum detection depends on a multitude of variables that are difficult to control (Aebersold & Mann 2003). The new proteins would be identified in suitable multiple replicates compared with the single run (Liu *et al.* 2004). One of our related studies also showed that the new proteins, commutatively, could be detected in 10 replicates (unpublished). And 2) reducing the variation of spectral counts in high abundant proteins by taking average of spectral counts, thus increase the confidence of testing significantly differentially expressed

proteins. Marcotte and colleagues (2007, 2008) showed that proteins of high abundance in both samples could be significantly differentially expressed even if the actual expression fold change is small. By combining the test results from the spectral counts and protein occurrence based methods on the basis of the multiple runs of MS data, significantly differentially expressed proteins with high confidence in two treatments can be obtained.

Table 4.1: Number of protein identification and significantly differentially expressed proteins for two false discovery rates (FDRs) under the assumptions of non-constant and constant probability of observing each peptide between two samples (Lu *et al.* 2007)

| Category | | Protein identifications | Non-constant assumption | | Constant assumption (Lu *et al.* 2007) | |
|---|---|---|---|---|---|---|
| | | | FDR<=0.01 | FDR<=0.05 | FDR<=0.01 | FDR<=0.05 |
| Proteins in both samples | | 48 | 14 | 20 | 14 | 21 |
| Proteins in one but not the other | In 1 but not 2 | 22 | 3 | 10 | 3 | 9 |
| | In 2 but not 1 | 102 | 11 | 34 | 12 | 45 |
| | Total | 124 | 14 | 44 | 15 | 54 |
| Total | | 172 | 28 | 64 | 29 | 75 |

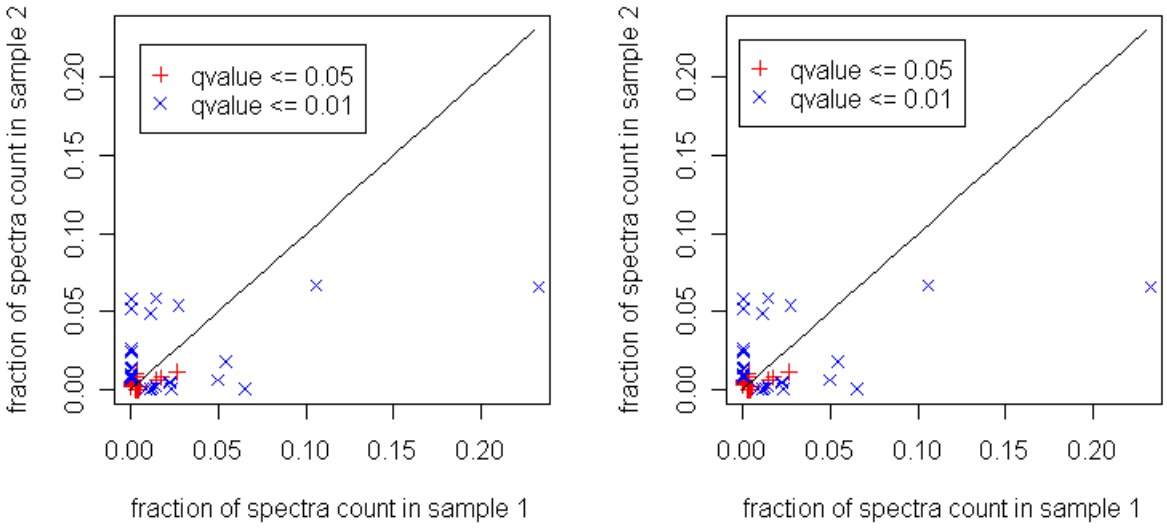



Figure 4.1. The average spectral counts detected in samples of two treatments for significantly differentially expressed proteins using two FDRs of 0.01 and 0.05. Left: constant assumption (Lu *et al.* 2007); Right: non-constant assumption.

# CHAPTER 5

# CONCLUSIONS

A main goal of analytical proteomics is the complete and quantitative proteome analysis of species, cells, and/or tissues. The field of proteomics has grown rapidly, shows no sign of slowing. Although the great success has been achieved via incremental improvements in MS-based proteomics, some principal limitations, such as extreme redundancy of LC-MS/MS spectra, under sampling, sample complexity, and saturation, make the goal of rapid, complete and quantitative proteome analysis not yet achieved. However, statistical considerations should be a starting point in choosing the number of replicates and analyzing the variations of factors in LC-MS/MS process.

The study presented here resulted in the following conclusions.

The developed probability-based model provides the probabilities of achieving a fixed coverage of sample proteins as a function of the number of replicates. In general, there is very small chance to identify 95% of sample proteins in one analysis. More than 95% of high abundant proteins can be identified with several analyses with a high confidence. However, in order to identify 95% of low abundant proteins in which at least one identified peptide per protein, more than 20 analyses are needed to achieve a moderate confidence. For a fixed confidence level, the developed model determines the coverage of sample proteins as a function of number of replicates. For a fixed confidence of >95%, most of high abundant proteins can be detected in several analyses, however, only a small portion of low abundant proteins can be detected with realistic runs.

In order to analyze effects of various factors on the detection probability in LC-MS/MS process, a mathematical model was derived on the basis of order statistics from independent non-identical normal random variables. As an approximation to the mathematical model, a simulation approach was applied to analyze impacts of different factors, such as protein abundance, complexity of samples, proteolytic digestion efficiency, peptide separation and co-eluting peptides, scanning speed of the mass spectrometer, and dynamic exclusion efficiency, on the peptide/protein identification. The proposed simulation approach can be used as a framework for analysis of impacts of various factors on the peptide/protein detection. The simulation results provide valuable information for optimization of LC-MS/MS techniques and practical guidelines for conducting MS-based experiments. The simulations show that high abundant proteins were identified with multiple peptides and low abundant proteins by one or two. A steep linear relationship between abundance and detection probability is revealed for the high abundant proteins, and the detection probability depends almost entirely on the elution time and very little on the abundance for proteins at low and medium abundances. Certain peptides are repeatedly and consistently identified, regardless of its abundance. The identified peptides at low abundances have abnormal elution times.

A methodology was developed to conduct statistical test of differential expression of proteins detected in two samples. By combining the test results from the spectral counts and protein occurrence based methods on the basis of the multiple runs of MS data, significantly differentially expressed proteins with high confidence in two treatments can be obtained.

**REFERENCES**

Adamski M, Blackwell T, Menon R, Martens L, Hermjakob H, Taylor C, Omenn GS, States DJ. 2005. Data management and preliminary data analysis in the pilot phase of the HUPO plasma proteome project. Proteomics, 5:3246–3261

Aebersold R. 2003. Constellations in a cellular universe. Nature, 422:115-116

Aebersold R, Goodlett DR. 2001. Mass spectrometry in proteomics. Chem Rev, 101:269–295

Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. Nature, 422:198-207

Balakrishnan N. 1994. Order statistics from nonidentically exponential random variables and some applications. Comput Statist Data-Anal, 18:203–225

Barakat HM, Abdelkader YH. 2000. Computing the moments of order statistics from nonidentically distributed Weibull variables. J Comp Appl Math, 117:85–90

Barakat HM, Abdelkader YH. 2004. Computing the moments of order statistics from nonidentical random variables. Stat Methods Appl, 13:13–24

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol, 57:289–300

Bergeron JJM, Hallett M. 2007. Peptides you can count on. Nat Biotechnol, 25:61-62

Blondeau F, Ritter B, Allaire PD, Wasiak S, Girard M, Hussain NK, Angers A, Legendre-Guillemin V, Roy L, Boismenu D, *et al.* 2004. Tandem MS analysis of brain clathrincoated vesicles reveals their critical involvement in synaptic vesicle recycling. Proc Natl Acad Sci USA, 101:3833–3838

Bodnar WM, Blackburn RK, Krise JM, Moseley MA. 2003. Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage. J Am Soc Mass Spectrom, 14:971-979

Breci LA, Tabb DL, Yates JR, III, Wysocki VH. 2003. Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. Anal Chem, 75:1963–1971

Cargile BJ, Bundy JL, Freeman TW, Stephenson JL. 2004. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. J Proteome Res, 3:112–119

Chao A. 1989. Estimating population size for sparse data in capture-recapture experiments. Biometrics, 45:427-438

Chao A, Lee S-M, Jeng S-L. 1992. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. Biometrics, 48:201-216

Childs A, Balakrishnan N. 1998. Generalized recurrence relations for moments of order statistics from non-identical Pareto and truncated Pareto random variables with applications to robustness. In: Balakrishnan N, Rao RC (eds) Handbook of statistics, Vol. 16, pp 403-438. North-Holland, Amsterdam

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics, 20:1466–1467

Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, *et al.* 2005. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol, 6:R9

Dobo A, Kaltashov IA. 2001. Detection of multiple protein conformational ensembles in solution via deconvolution of charge-state distributions in ESI MS. Anal Chem, 73:4763–4773

Domon B, Aebersold R. 2006. Mass spectrometry and protein analysis. Science, 312:212-217

Durr E, Yu J, Krasinska KM, Carver LA, Yates JR, III, Testa JE, Oh P, Schnitzer JE. 2004. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. Nature Biotechnol, 22:985 - 992

Edman P. 1950. Method for determination of the amino acid sequence in peptides. Acta Chem Scand, 4:283-293

Elias JE, Haas W, Faherty BK, Gygi SP. 2005. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods, 2:667-675

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods, 4:207-214

Eng JK, McCormack AL, Yates JR, III. 1994. An approach to correlate MS/MS data to amino acid sequences in a protein database. J Am Soc Mass Spectrom, 5:976–989

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for the mass spectrometry of large biomolecules. Science, 246:64–71

Gatlin CL, Eng JK, Cross ST, Detter JC, Yates JR, III. 2000. Identifying SNPS expressed in Proteins: Rapid identification of amino acid sequence variations by LC/MS/MS with SEQUEST data analysis. Anal Chem, 72:757–763

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. 2004. Open mass spectrometry search algorithm. J Proteome Res, 3:958-964

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature, 425:737-741

Girard M, Allaire PD, McPherson PS, Blondeau F. 2005. Non-stoichiometric relationship between clathrin heavy and light chains revealed by quantitative comparative proteomics of clathrin-coated vesicles from brain and liver. Mol Cell Proteomics, 4:1145–1154

Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biology, 4:117

Guzzetta A.2001. Reverse phase HPLC basics for LC/MS: an IonSource tutorial. http://www.ionsource.com/tutorial/chromatography/rphplc.htm#Introduction

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. 1999a. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, Nature Biotechnol, 17:994-999

Gygi SP, Rochon Y, Franza BR, Aebersold R. 1999b. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol, 19:1720–1730

Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP. 2006. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. Mol Cell Proteomics, 5:1326-1337

Hansen BT, Mason DE, Jones JA, Liebler DC. 2001. SALSA: An algorithm for detection of modified peptides by automated evaluation of their CID spectra in LC-tandem MS analyses. Anal Chem, 73, 1676-1683

Heller M, Ye M, Michel PE, Morier P, Stalder D, Junger MA, Aebersold R, Reymond F, Rossier JS. 2005. Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. J Proteome Res, 4:2273-2282

Henzel WJ, Billeci TM, Stults JT, Wong SC. 1993, Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc Nat Acad Sci, 90:5011-5015

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature, 409:860-921

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature, 431:931-945

Karas M, Hillenkamp F. 1988. Laser desorption ionization of proteins with molecular mass exceeding 10000 daltons. Anal Chem, 60:2299–2301

Keller A, Nesvizhskii AI, Kolker E, Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem, 4:5383-5392

King NL, Deutsch EW, Ranish JA, Nesvizhskii AI, Eddes JS, Mallick P, Eng J, Desiere F, Flory M, Martin DB *et al.* 2006. Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. Genome Biol, 7:R106

Kjeldsen F, Haselmann KF, Budnik BA, Sørensen ES, Zubarev RA. 2003. Complete characterization of posttranslational modification sites in the bovine milk protein PP3 by tandem mass spectrometry with electron capture dissociation as the last stage. Anal Chem, 75:2355-2361

Klammer AA, Yi X, MacCoss MJ, Noble WS 2007. Improving Tandem Mass Spectrum Identification Using Peptide Retention Time Prediction across Diverse Chromatography Conditions. Anal Chem, 79:6111-6118

Koziol JA, Feng AC, Schnitzer JE. 2006. Application of capture-recapture models to estimation of protein count in MudPIT experiments. Anal Chem, 78:3203-3207

Krokhin OV. 2006. Sequence-Specific Retention Calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-Å pore size C18 sorbents. Anal Chem, 78:7785-7795

Kuster B, Schirle M, Mallick P, Aebersold R. 2005. Innovation: scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol, 6:577-583

Lasonder E, Ishihama Y, Andersen JS, Vermunt AMW, Pain A, Sauerwein RW, Eling WMC, Hall N, Waters AP, Stunnenberg HG, Mann M. 2002. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature, 419:537–542

Lee S-M, Chao A. 1994. Estimating population size via sample coverage for closed capture-recapture models. Biometrics, 50:88–97

Liebler DC. 2002. Introduction to proteomics: tools for the new biology. Humana Press

Lipton MS, Pasa-Tolic L, Anderson GA, Anderson DJ, Auberry DL, Battista JR, Daly MJ, Fredrickson J, *et al.* 2002. Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. Proc Natl Acad Sci USA 99:11049–11054

Liu H, Sadygov RG, Yates JR, III. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201

Lu P, Vogel C, Wang R, Yao X, Marcotte EM. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol, 25:117–124

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 17: 2337–2342

MacCoss M, Wu C, Yates JR, III. 2002. Probability based validation of protein identifications using a modified SEQUEST algorithm, Analytical Chemistry, 74:5593-5599

Makarov A. 2000. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem, 72:1156-1162

Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. 2007. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol, 25:125–131

Malmstrom J, Lee H, Nesvizhskii AI, Shteynberg D, Mohanty S, Brunner E, Ye M, Weber G, Eckerskorn C, Aebersold R. 2006. Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. J Proteome Res, 5:2241-2249

Malmstrom J, Lee H, Aebersold R. 2007. Advances in proteomic workflows for systems biology. Current Opinion in Biotechnology, 18:378-384

Mann M, Wilm M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem, 66:4390-4399

Marshall AG, Hendrickson CL, Jackson GS. 1998. Fourier transform ion cyclotron resonance mass spectrometry: a primer. Mass Spectrom Rev, 17:1

Martin SE, Shabanowitz J, Hunt DF, Marto JA. 2000. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI Fourier transform ion cyclotron resonance mass spectrometry. Anal Chem, 72:4266-4274

Marzolf B, Deutsch EW, Moss P, Campbell D, Johnson MH, Galitski T. 2006. SBEAMS-Microarray: database software supporting genomic expression analyses for systems biology. BMC Bioinform, 7:286

Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL. 2000. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. Anal Chem, 72:552

Minc H. 1983. Theory of permanents 1978–1981. Linear and multilinear Algebra, 12:227–263

Minc H. 1987. Theory of permanents 1982–1985. Linear and multilinear Algebra, 21:109–198

Moore RE, Young MK, Lee TD. 2002. Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom, 13:378–386

Morris HR, Paxton T, Dell A, Langhorne J, Berg M, Bordoli RS, Hoyes J, Bateman RH. 1996. High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. Rapid Commun Mass Spectrom, 10:889-896

Nesvizhskii AI, Keller A, Kolker E, Aebersold R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem, 75:4646-4658

Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. 2006. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics, 5:652–670

Nesvizhskii AI, Vitek O, Aebersold R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods, 10:787-797

Nielsen ML, Savitski MM, Kjeldsen F, Zubarev RA. 2004. Physicochemical properties determining the detection probability of tryptic peptides in Fourier transform mass spectrometry. A correlation study. Anal Chem, 76:5872–5877

Ocaña MF, Jarvis J, Parker R, Bramley PM, Halket JM, Patel RKP, Neubert H. 2005. C-terminal sequencing by mass spectrometry: Application to gelatine-derived proline-rich peptides. Proteomics, 5:1209–1216

Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG. 2005. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics 4:1487–1502

Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS *et al.* 2005. Overview of the HUPO plasma proteome project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics, 5:3226-3245.

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

Otis DL, Burnham KP, White GC, Anderson DR. 1978. Wildlife monographs. No. 62. The Wildlife Society: Bethesda, Maryland

Pan P, McLuckey SA. 2003. Electrospray ionization of protein mixtures at low pH. Anal Chem, 75:1491–1499

Pan P, Gunawardena HP, Xia Y, McLuckey SA. 2004. Nanoelectrospray ionization of protein mixtures: Solution pH and protein pI. Anal Chem, 76:1165–1174

Paoletti AC, Parmely TJ, Tomomori-Sato C, Sato S, Zhu D, Conaway RC, Conaway JW, Florens L, Washburn MP. 2006. Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. Proc Natl Acad Sci USA, 103:18928-18933

Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res, 2:43–50

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20:3551-3567

Peschke M, Blades A, Kebarle P. 2002. Charged states of proteins. Reactions of doubly protonated alkyldiamines with $NH_3$: Solvation or deprotonation. Extension of two proton cases to multiply protonated globular proteins observed in the gas phase. J Am Chem Soc, 124:11519–11530

Petritis K, Kangas LJ, Ferguson PL, Anderson GA, Pasa-Tolic L, Lipton MS, Auberry KJ, Strittmatter EF, Shen Y, Zhao R, Smith RD. 2003. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. Anal Chem, 75:1039-1048

Petritis K, Kangas LJ, Yan B, Monroe ME, Strittmatter EF, Qian W-J, Adkins JN, Moore RJ, Xu Y, Lipton MS, Camp DG, II, Smith RD. 2006. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. Anal Chem, 78:5026-5039

Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O. 2007. Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. BMC Bioinformatics. 8:468

Powell DW, Weaver CM, Jennings JL, McAfee KJ, He Y, Weil PA, Link AJ. 2004. Cluster analysis of mass spectrometry data reveals a novel component of SAGA. Mol Cell Biol, 24:7249-7259

Premstaller A, Oberacher H, Walcher W, Timperio AM, Zolla L, Chervet JP, Cavusoglu N, van Dorsselaer A, Huber CG. 2001. High-performance liquid chromatography-electrospray ionization mass spectrometry using monolithic capillary columns for proteomic studies. Anal Chem, 73:2390-2396

Qian WJ, Liu T, Monroe ME, Strittmatter EF, Jacobs JM, Kangas LJ, Petritis K, Camp DG, II, Smith RD. 2005. Probability-Based Evaluation of Peptide and Protein Identifications from Tandem Mass Spectrometry and SEQUEST Analysis: The Human Proteome. J Proteome Res, 4:53–62

Rifai N, Gillette MA, Carr SA. 2006. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol, 24:971-983

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S. 2004. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics, 3:1154-1169

Sadygov RG, Yates JR, III. 2003. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem, 75:3792–3798

Sadygov RG, Liu H, Yates JR, III. 2004. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. Anal Chem, 76:1664-1671

Schirle M, Heurtier MA, Kuster B. 2003. Profiling core proteomes of human cell lines by 1D PAGE and LC–MS/MS. Mol Cell Proteomics, 2:1297–1305

Schmidt A, Kellermann J, Lottspeich F. 2005. A novel strategy for quantitative proteomics using isotope-coded protein labels. Proteomics, 5:4-15

Seber GAF. 1982. The estimation of animal abundance and related parameters, 2nd ed. Griffin: London

Shen Y, Smith RD, Unger KK, Kumar D, Lubda D. 2005. Ultrahigh-throughput proteomics using fast RPLC separations with ESI-MS/MS. Anal Chem, 77:6692-6701

Smit S, Hoefsloot HCJ, Smilde AK. 2007. Statistical data processing in clinical proteomics. J Chromatogr B, doi:10.1016/j.jchromb.2007.10.042

Steen H, Mann M. 2004. The abc's (and xyz's) of peptide sequencing. Nat Rev Mol Cell Biol, 5:699-711

Storey JD. 2002. A direct approach to false discovery rates. J Roy Statistical Society, Series B 64: 479–498

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide experiments. Proc Natl Acad Sci USA, 100:9440–9445

Strittmatter EF, Ferguson PL, Tang K, Smith RD. 2003. Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. J Am Soc Mass Spectrom, 14:980-991

Syka JE, Marto JA, Bai DL, Horning S, Senko MW, Schwartz JC, Ueberheide B, Garcia B, Busby S, Muratore T *et al.* 2004. Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. J Proteome Res, 3:621-626

Tabb DL, Saraf A, Yates JR, III. 2003. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem, 75:6415–6421

Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. 2005. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem, 77:4626–4639

Vogel C, Marcotte EM. 2008. Calculating absolute and relative protein expression levels from mass spectrometry data. Nat Protoc, 3(9):1444 - 1451

Washburn MP, Wolters D, Yates JR, III. 2001. Large scale analysis of the yeast proteome via multidimensional protein identification technology. Nat Biotechnol, 19:242-247

Weatherly DB, Atwood JA, III, Minning TA, Cavola C, Tarleton RL, Orlando R. 2005. A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. Mol Cell Proteomics, 4:762-772

Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL. 1995. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev, 13:19-50

Wilkins MR, Appel RD, Van Eyk JE, Chung MCM, Görg A, Hecker M, Huber LA, Langen H, Link AJ, *et al.* 2006. Guidelines for the next 10 years of proteomics. Proteomics, 6:4-8

Wu CC, MacCoss MJ, Howell KE, Yates III, JR. 2003. A method for the comprehensive proteomic analysis of membrane proteins. Nat. Biotechnol. 21, 532-538

Xue X, Wu S, Wang Z, Zhu Y, He F. 2006. Protein probabilities in shotgun proteomics: Evaluating different estimation methods using a semi-random sampling model. Proteomics. 6:6134-6145

Yates JR, III, Eng JK, McCormack AL. 1995. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to nucleotide sequences, Anal Chem, 67:3202-3210

Yin H, Killeen K, Brennen R, Sobek D, Werlich M, van de Goor T. 2005. Microfluidic chip for peptide analysis with an integrated HPLC column, sample enrichment column, and nanoelectrospray tip. Anal Chem, 77:527-533

Yu LR, Conrads TP, Uo T, Kinoshita Y, Morrison RS, Lucas DA, Chan KC, Blonder J, Issaq HJ, Veenstra TD. 2004. Global analysis of the cortical neuron proteome. Mol Cell Proteomics, 3:896–907

Zhang N, Aebersold R, Schwikowski B. 2002. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics, 2:1406-1412

Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R. 2005. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics, 5:4096-4106

Zybailov B, Coleman MK, Florens L, Washburn MP. 2005. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal Chem, 77:6218–6224

Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. 2006. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res 5:2339–2347

# APPENDICES

## A1. Introduction to mass spectrometry (MS)-based proteomics

Marc Wilkins and colleagues coined the terms proteomics and proteome in the early 1990s (Wilkins *et al.* 1995). In short, proteomics is the study of the proteome, the protein complement of a genome that describes the entire collection of genes in an organism. Every organism has one genome, but many proteomes. The proteome in any organism is a collection of somewhere between 30 and 80 percent of the possible gene products, and these proteins are expressed at different abundance levels, typically from $10^1$ to $10^6$ per cell (Liebler 2002).

### 1 Overview of MS-based proteomics

Analytical protein identification is built around on the essential fact: most six or more amino acid peptides are largely unique in the proteome of any organism. Therefore, if we can obtain the accurate peptide mass or the peptide sequence, we can identify its protein of origin by finding its match in a protein sequence database.

The essential elements of the common MS-based proteomic pipelines are shown in Figure A1.1. Proteins from a specific organism are extracted and then separated using 1-D gel electrophoresis to reduce the complexity of the mixture. Proteins excised from the gel are digested by proteolytic enzymes (e.g. trypsin), resulting in the highly complex peptide mixture. To analyze peptide mixtures by MS efficiently, the highly complex peptide mixture of many components must be separated (usually by High Performance Liquid Chromatography, abbreviated as HPLC or LC) into somewhat less complex mixtures containing fewer components. The peptides are then analyzed by any type of mass spectrometers, commonly Matrix Assisted Laser Desorption Ionization-Time of Flight (MALDI-TOF) and Electrospray Ionization (ESI)-tandem MS instruments, to obtain MS (or tandem MS) spectrum, which is a

recording of the signal intensity of the ion (or fragment ion) at each value of the mass-to-charge (m/z) ratio. Peptides and proteins are subsequently identified by correlating MS spectra with a protein sequence database with the aid of specialized software (e.g. SEQUEST, MASCOT).

**2 Protein separation and digestion**

The goal of the protein separation is to reduce the complexity of the mixture. In 1-D gel electrophoresis, the proteins are resolved into bands in order of molecular weight when the gel is subject to high voltage. Each band typically contains dozens to hundreds of different proteins since the degree of resolution achieved by 1-D gel electrophoresis is rather modest. Each band (containing multiple proteins) is excised from the gel for subsequent analysis.

MS analysis of whole proteins is less sensitive than peptide MS. In addition, the mass of the intact protein by itself is insufficient for identification. Therefore, excised proteins must be cleaved into peptides. This is generally done with proteolytic enzymes. The most widely used protease in analytical proteomics is trypsin. Trypsin cleaves the intact protein into peptides at the amino acids residuals of lysine (K) or arginine (R), unless either of these is followed by a proline (P) residual at the C-terminal direction. Generally, a typical 50 kDa protein will yield about 30 tryptic peptides (Liebler 2002). The ideal length of peptide fragments for MS analysis and database comparisons is about 6 - 20 amino acids. Peptides longer than 20 amino acids generally are difficult to obtain its sequence in tandem MS analyses. On the other hand, peptides shorter than 6 amino acids may produce multiple matches in database search. The objective of protein digestion would be to produce the highest yield of peptides of optimal length for MS analysis.

**3 LC for peptide separation**

In order to give the MS instruments a better chance to detect elements of peptide mixtures, highly complex and heterogeneous peptide mixtures must be separated into somewhat less

complex and more homogeneous mixtures prior to MS analysis. The LC has considerable resolving power for peptide separation. Reverse phase (RP) LC is the most common form of chromatography used in LC-MS applications. The RP LC is simple. Peptides stick to RP LC columns (usually a tube packed with small silica particles) in high aqueous mobile phase and are eluted from RP LC columns with high organic mobile phase. By pumping a liquid at a linear gradient of the organic solvent (mobile phase) at high pressure through the column, peptides are separated on the basis of their hydrophobic character. Typical scheme of the gradient starts at near 100% aqueous and ramps to 60% organic solvent in 60 minutes. As this scheme applied, the majority of peptides (10 to 30 amino acid residues in length) will elute by the time the gradient reaches 30% organic (Guzzetta 2001). The resolved peptides consequently pass directly into the MS instrument for generating spectra.

All LC systems have a gradient delay. The gradient delay is the time between when the pumps to start pumping at a certain mobile phase composition and the time it takes for that solvent composition to reach the column and have an effect. A good guess for a gradient delay is 10 minutes (Guzzetta 2001).

**4 Mass spectrometers for peptide analysis**

Mass spectrometers consist of three essential parts – source, mass analyzer, and detector. The source produces ions from the components of a mixture, then the mass analyzer resolves ions on the basis of their m/z ratio with an external magnetic or electric field, and finally the detector detects resolved ions and generates the measurable signals. This procedure of producing MS data is illustrated in Figure A1.2. Instruments commonly used in MS could distinguish ions differing in m/z values of at least one Dalton, i.e., the mass of single hydrogen atom.

Two commonly used mass spectrometers are MALDI-TOF and ESI-tandem MS instruments. In short, MALDI-TOF MS provides peptide masses and ESI-tandem MS produces peptide ion fragmentation (Aebersold & Goodlett 2001). In both terms, the first part [MALDI (Karas & Hillenkamp 1988) or ESI (Fenn *et al.* 1989)] refers to the source, whereas the second part (TOF or tandem MS) refers to the mass analyzer. Tandem MS refers to mass analyzers that are able to perform two-stage or multistage mass analyses of ions.

MALDI-TOF is best-suited to measure peptide masses, which is nevertheless the limited information for protein identifications. As its name said, TOF (Time of Flight) mass analyzer measures the time it takes for the ions to fly from one end of the analyzer to the other and strike the detector. The fly time is proportional to m/z values of ions. The greater the m/z, the faster the ions fly.

Since peptide ion fragmentation provides true sequence information, which has greater intrinsic value, ESI-tandem MS is becoming more attractive and is widely used in analytical proteomics study. The most commonly used ESI-tandem MS analyzers are quadrupole, ion trap, and TOF mass analyzers. These analyzers could be used in various combinations. The general procedure of generating mass data using ESI-tandem MS is as follows. First, peptide ions from the source are full-scanned according to m/z values of all ions coming from the source at any given time. This could be considered as a snapshot of the peptide ions entering the source over the short time period of the scan. After the full scan is complete, the instrument switches to tandem MS mode and selects the most intense ion and subjects it to collision-induced dissociation (CID), which induces fragmentation of the peptide into fragment ions and then is analyzed based on their m/z, to obtain a tandem MS spectrum. Then the instrument switches back to full-scan mode and select next most intense peptide ion and subjects it to CID. The

switching cycle is repeated to obtain tandem MS spectra of multiple peptide ions (Fig. A1.3). After the switching cycle is done, the next group of peptide ions eluted from LC step is full-scanned, which then are further analyzed by the switching cycle between full-scan and tandem MS modes to generate tandem MS spectra. This process is repeated until all of the peptides have been analyzed from the LC step. The MS and MS/MS spectra are typically acquired for about one second each and stored for matching against protein sequence databases (Aebersold & Mann 2003).

**5 Computer-assisted peptides and proteins identification**

There are two types of MS data, peptide masses (typically from MALDI-TOF MS) and peptide sequence (generally from ESI-tandem MS). When MS is used to measure the peptide masses, the protein identification technique is referred to as the peptide mass fingerprinting (PMF). For PMF, the peptides are identified by matching the measured peptide masses to corresponding *in silico* digested peptide masses from protein or nucleotide sequence databases. For the tandem MS spectra, there are two ways to identify peptides and proteins. The first is *de novo* interpretation of the spectrum to obtain a peptide sequence followed by BLAST searching of the sequence against protein sequence database to identify the protein. Unfortunately, the success of *de novo* interpretation/BLAST searching approach crucially depends on the quality of data, in terms of both the mass accuracy and the resolution of the instruments. The second approach circumvents this problem, which directly correlates tandem MS spectral data with peptide sequences, actually with hypothetical spectra predicted from candidate peptides (of equal mass), in protein database without explicitly interpreting tandem MS spectra (Eng *et al.* 1994). For both types of MS data, the accurate protein identifications usually require multiple peptide matches.

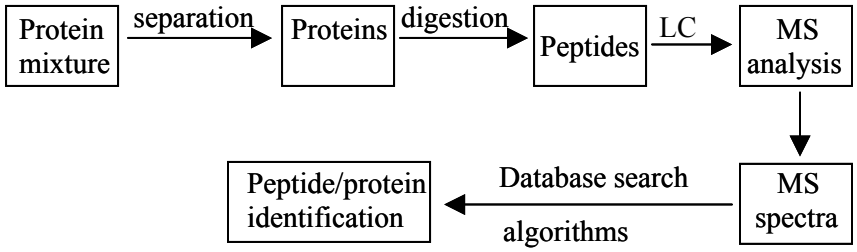



Figure A1.1: General flow scheme for proteomic analysis.

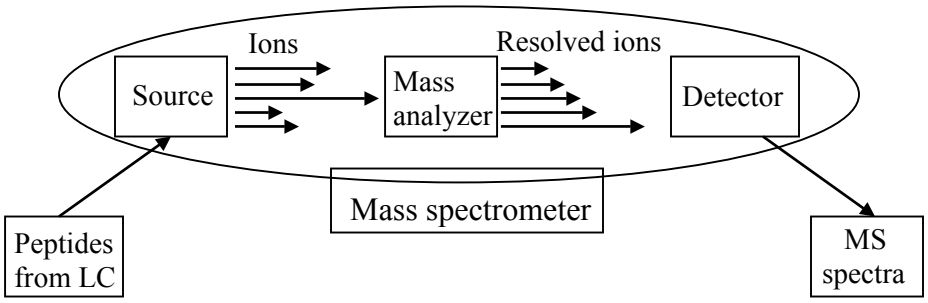



Figure A1.2: Schematic representation of a mass spectrometer.

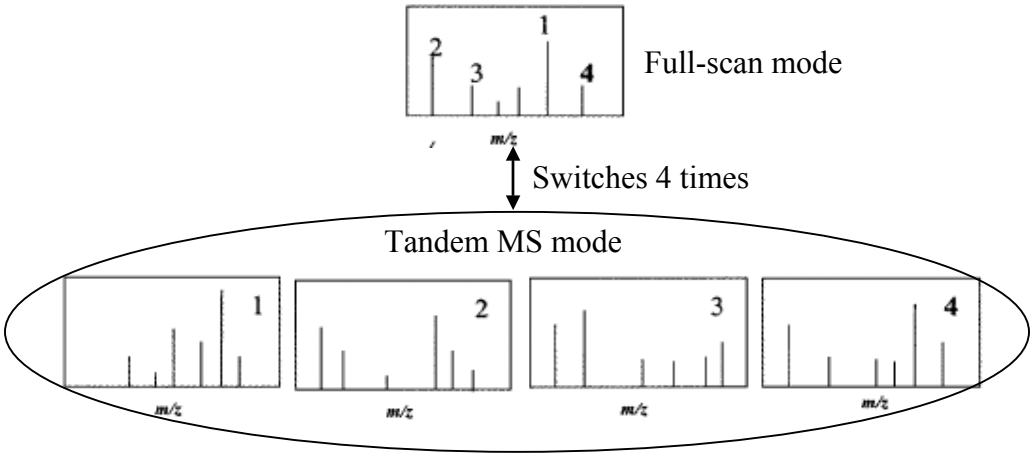



Figure A1.3: Schematic representation of the automated collection of tandem MS spectra by switching cycle between full-scan and tandem MS modes.

## A2. Supplementary information to Chapter 2

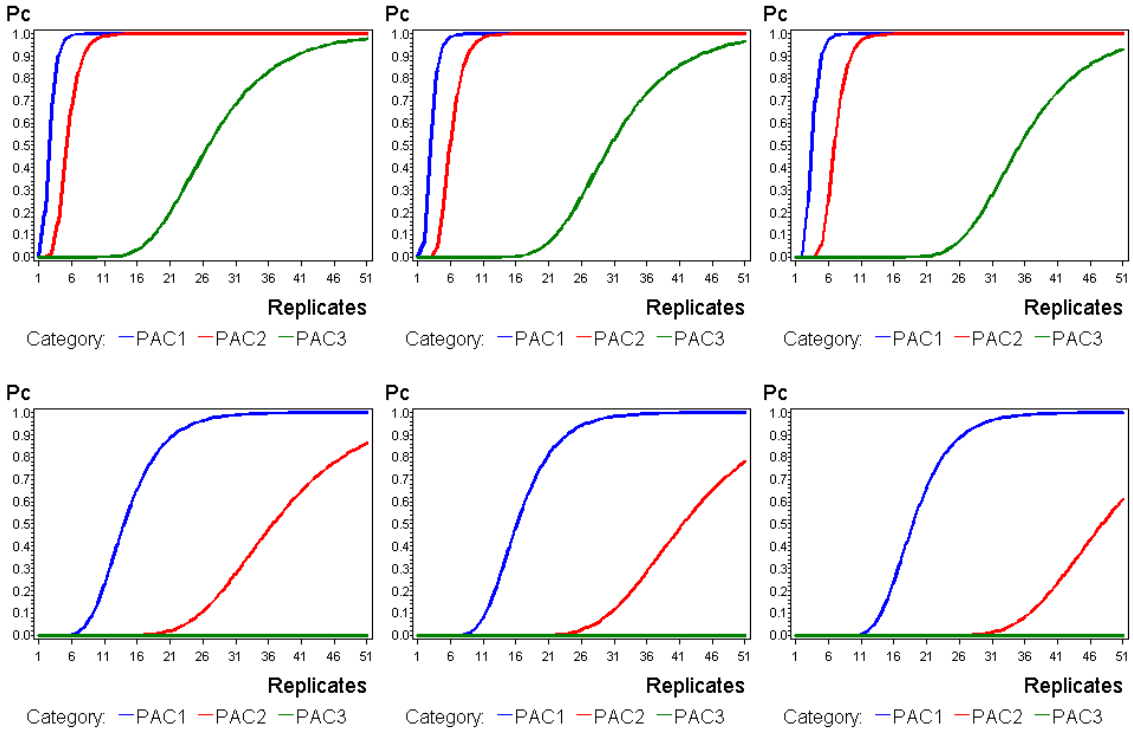



Figure A2.1: The probabilities of identifying all proteins in each protein abundance class as a function of the number of replicates. A match is defined as at least one replicate in which at least one (top panel) and two (bottom panel) unique peptides are identified in a protein in $r$ replicates. Different numbers of proteins in samples are applied: left column for 89, central column for 150, and right column for 300 proteins.

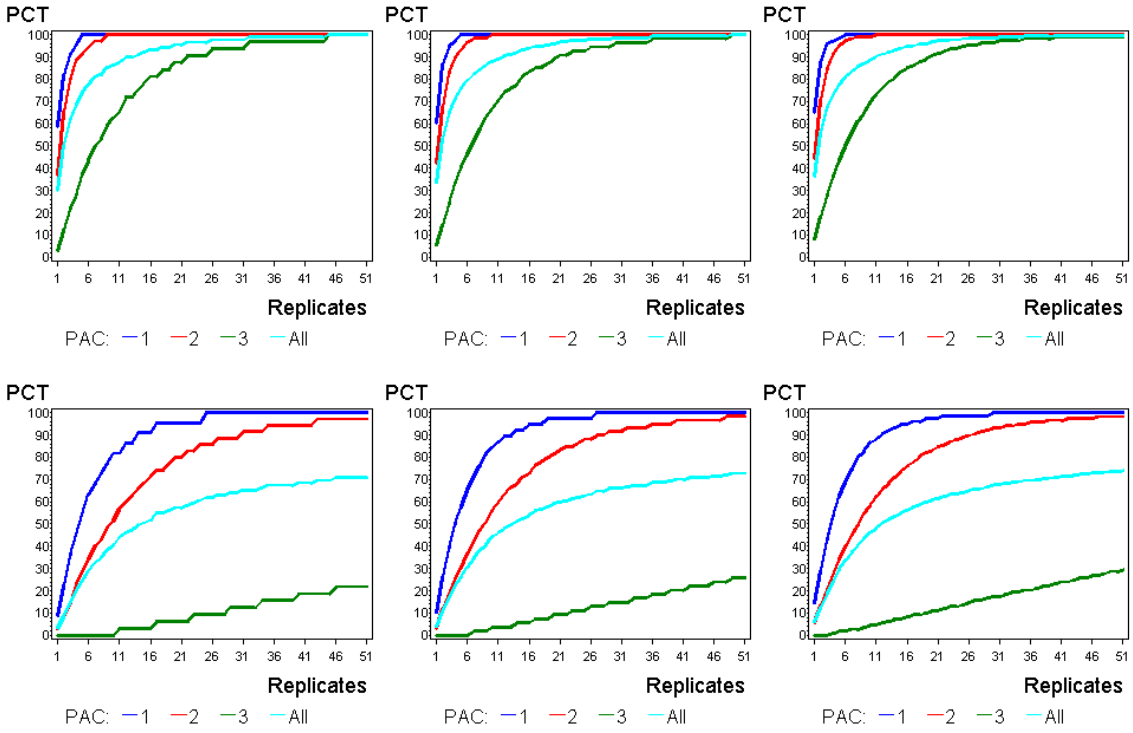



Figure A2.2: The protein coverage at each protein abundance classes and overall coverage as a function of the number of replicates with >95% confidence level. A match is defined as at least one replicate in which at least one (top panel) and two (bottom panel) unique peptides are identified in a protein in $r$ replicates. Different numbers of proteins in samples are applied: left column for 89, central column for 150, and right column for 300 proteins.

## A3. Supplementary information to Chapter 3

### A3.0 Algorithm of simulations

In general, the steps of the model are as follows: 1) specify number of proteins in samples and number of peptides of each protein. 2) Specify abundances for each protein in a vector. 3) Generate the mean elution time ( $\mu_{ij}$ ) for each peptide from an appropriate distribution and also specify the standard deviation ( $\sigma_{ij}$ ) in elution time. 4) The actual elution times are then generated for each peptide from a normal distribution with the mean and standard deviation as specified in $\mu_{ij}$ and $\sigma_{ij}$ . 5) Specify a sampling interval $d$ then do a loop over these intervals. Determine how many copies of each peptide elute in each such interval. The matrix *current* has columns corresponding to peptides and rows corresponding to peptide copies. In each sampling interval ones are assigned to peptides that elute and zeros to those that don't. Then, columns are summed to give the vector *eluted_abundance*. Each element is the number of copies of the peptide that eluted during that sampling interval. This vector is copied to another called *dela*. Then, previously detected peptides are removed (if with exclusion) and peptides with zero eluted abundance are also removed. The remaining peptides in *dela* are ordered by eluted copy number. The top $\min(c_l, c)$ [or $\min(c_l - e_l, c)$ if with exclusion] most abundant peptides are stored in the vector *detected*. This continues until the final sampling interval. Do multiple replicates and keep track of detection probabilities for individual peptides and/or proteins.

### A3.1 Parameter estimates of gamma distribution at different retention time scales

Using Petritis's 1303-peptide data, the parameter estimates of gamma distribution at different retention time scales were obtained (Table A3.1). The shape parameter estimates are the same at different scales, and scale and threshold parameter estimates change with the data scale accordingly. Due to the limitation of computational sources, it is not feasible to do simulations as

real experiments, in which the copies of high abundant proteins are more than several millions, tens to thousands proteins in samples, the number of peptides per protein after digestion is large and unknown, and the total time of experiments is 1 hour, i.e. 3600 seconds, or more. We have to do some reductions on the total sampling period, number of proteins, peptides per protein, and abundances of proteins. The logic here is to reduce the total sampling period and complexity of samples (including number of proteins, peptides per protein, and abundances) simultaneously, however, keep the same distribution of peptides (results in Table A3.1 support this point) for different time scales (e.g. 1, 10, 100, and 1000 time units above).

**A3.2 Effects of standard deviation of retention time**

Standard deviation of retention time measures the spread of a given peptide being eluted. For a fixed scan interval, the length of a given peptide being "fully" eluted might strongly influence the results of the peptide identification. The mean detection probabilities (Fig. A3.1 for peptide and Fig. A3.2 for protein) at various SDs are similar and the differences are negligible. With a more restrict definition of a correct protein identification (P2), the mean protein detection probability decreases significantly (Fig. A3.2).

**A3.3 Effects of proteolytic digestion efficiency**

The proteolytic digestion efficiency leads to different number peptide per protein after the digestion. In general, the protein digestion is incomplete and number of real tryptic peptides for detection is smaller than that from the theoretical digestion. A typical 50 kDa protein will yield about 30 tryptic peptides (Liebler 2002). The ideal length of peptide fragments for MS analysis and database comparisons is about 6 - 20 amino acids. Peptides longer than 20 amino acids generally are difficult to obtain its sequence in tandem MS analyses. On the other hand, peptides shorter than 6 amino acids may produce multiple matches in database search.

With an increasing number of peptides per protein, the total number of peptides in samples increases significantly. Therefore, the mean peptide detection probability goes down (Fig. A3.3). In contrast, with an increasing number of peptides per protein, the chance of at least one or two peptides in any protein being identified increases, which further determine an increase in the protein detection probability (Fig. A3.4). With a more restrict definition of a correct protein identification (P2), the mean protein detection probability decreases significantly, reducing from [0.90, 0.98] of P1 to [0.55. 0.88] of P2 (Fig. A3.4).

**A3.4 Effects of scanning speed of the mass spectrometer**

**1) Varying sampling interval with a fixed number of the top most intense ions sampled in each cycle**

We vary sampling interval $t_0$ from 0.1 to 1.0 by intervals of 0.1. In each cycle, the top 5 most intense ions were sampled. Not surprisingly, we see that the sampling interval has a major impact - the mean peptide detection probability, which is defined as the average proportion of identified peptides to all peptides in samples over multiple replicates, decreases with an increasing sampling interval (Fig. A3.5). The source of missed peptides/proteins is that they never make it into the top 5 most intense peptide ions because there are other peptides with higher eluted abundance. The more often you sample, the more likely you are to catch a peptide before it decreases in eluted abundance. This implies that we can increase the number of detected peptides/proteins by decreasing the time of scanning a MS and MS/MS spectrum.

**2) Varying sampling interval with a fixed time for detecting a peptide ion in each cycle**

However, in real experiments, there is a tradeoff between the length of sampling interval and number of the top most intense ions being sampled for any cycle. The bottom line is that the time interval should be long enough for the specified number of ions being sampled. Generally, the

number of ions being sampled is determined by the sampling interval. So, which one is better: short sampling interval and small number of sampled ions or long sampling interval and big number of sampled ions?

Again, we vary sampling interval $t_0$ from 0.1 to 1.0 by intervals of 0.1. In each cycle, the top $t_0/0.1$ most intense ions were sampled. With this setting, the $c$, the top most intense ions being sampled in each cycle, increases from 1 to 10 as $t_0$ increases from 0.1 to 1, respectively. When the time for detecting a peptide ion is fixed, we see very slight differences in the mean detection probabilities between two investigated strategies (Fig. A3.6).

Table A3.1: The estimation of parameters for gamma distribution based on Petritis's 1303-peptide data at different scales

| | NRT | NRT*10 | NRT*100 | NRT*1000 |
|---|---|---|---|---|
| Threshold: theta | -0.03054 | -0.30541 | -3.05406 | -30.5354 |
| Scale: sigma | 0.061134 | 0.611339 | 6.113393 | 61.13505 |
| Shape: alpha | 6.806035 | 6.806 | 6.805999 | 6.805789 |

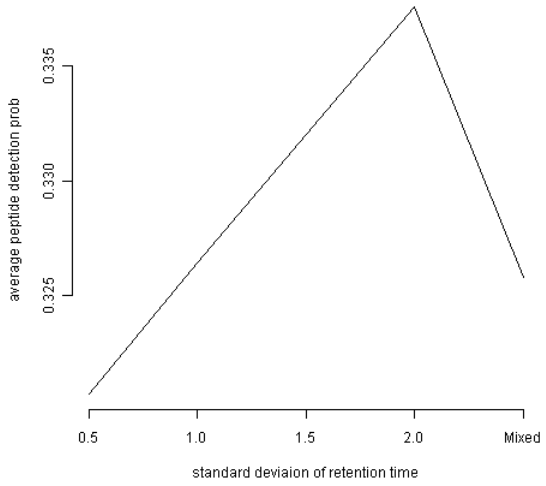



Figure A3.1: The mean peptide detection probabilities for different standard deviations of retention time. Mixed - SD of 1 for early eluted and 2 for later eluted peptides.

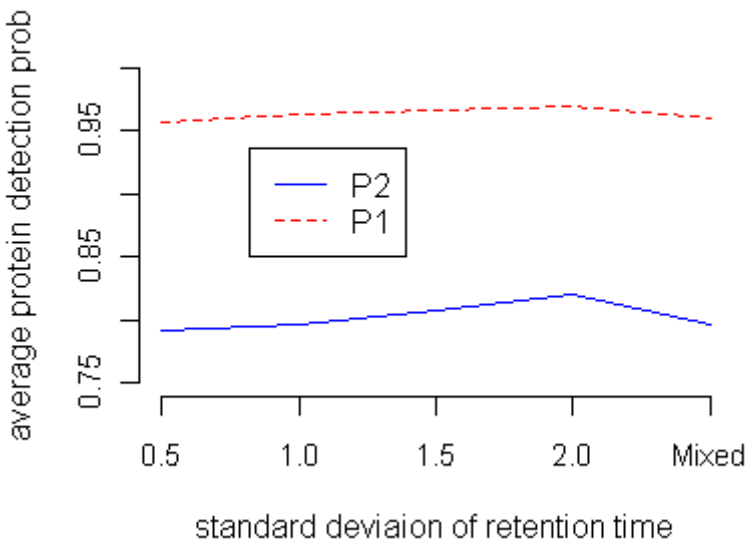



Figure A3.2: The mean protein detection probabilities for different standard deviations of retention time. Mixed - SD of 1 for early eluted and 2 for later eluted peptides.

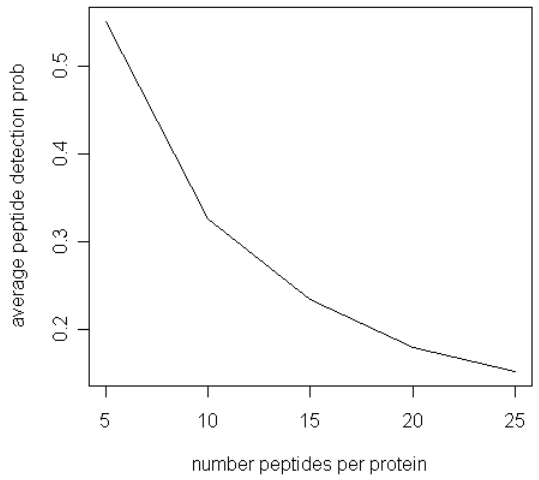



Figure A3.3: The mean peptide detection probabilities for different number of peptides per protein.

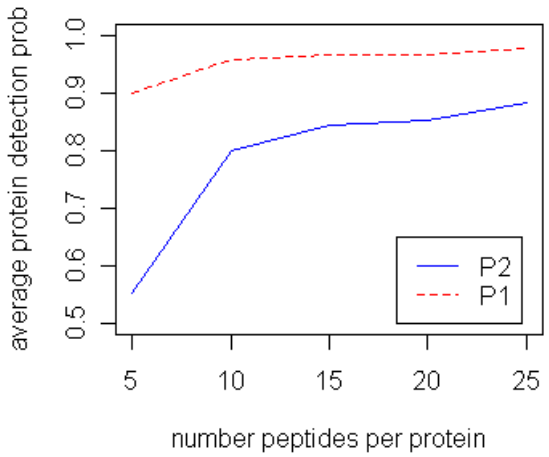



Figure A3.4: The mean protein detection probabilities for different number of peptides per protein under two definitions of protein identification - 1) at least one peptide being identified in a protein (P1), and 2) at least two peptides being identified in a protein (P2).

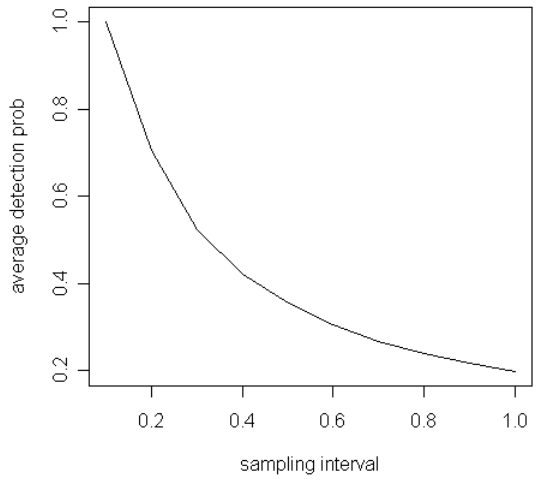



Figure A3.5: The mean peptide detection probability under various sampling interval with a fixed number of the top most intense ions sampled in each cycle.

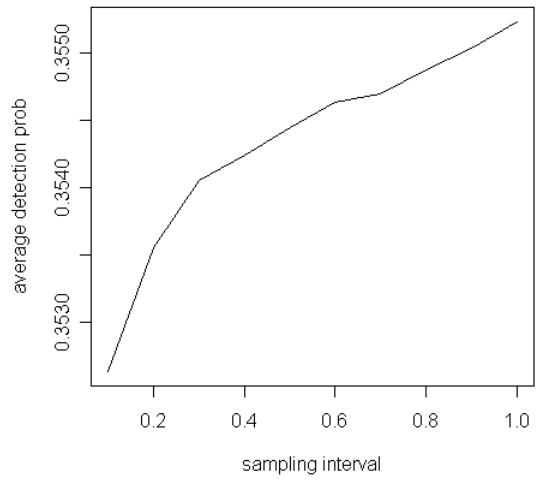



Figure A3.6: The mean peptide detection probability under various sampling interval with a fixed time of detecting a peptide ion in each cycle.

## A4. Supplementary information to Chapter 4

Table A4.1: Statistics for testing differential expression of proteins with constant and non-constant assumptions using spectral counts and protein occurrence based methods

| PNAME | PTYPE | Spectral counts based method | | | | | | | | | | Protein occurrence based method | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Constant assumption | | | | | Non-constant assumption | | | | | | |
| | | # SPECTRA | # RUN | Z score | P | QVALUE | # SPECTRA | # RUN | Z score | P | QVALUE | $P_{DET}$ | $P_{FAIL}$ |
| BC1G_02163.1 | hypothetical protein (translation) | 232.00 | 3 | 10.0442 | 0.0000 | 0.0000 | 58.00 | 3 | 10.6251 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| BC1G_10880.1 | hypothetical protein (translation) | 64.67 | 3 | 7.7183 | 0.0000 | 0.0000 | 0.00 | 0 | 8.3165 | 0.0000 | 0.0000 | 1.0000 | 0.0000 |
| NP_181159.1 | putative ubiquitin/ribosomal protein CEP52 [Arabidopsis thaliana] | 0.00 | 0 | -7.7024 | 0.0000 | 0.0000 | 51.33 | 1 | -7.3817 | 0.0000 | 0.0000 | 0.3333 | 0.2963 |
| BC1G_13715.1 | predicted protein (translation) | 0.00 | 0 | -7.2809 | 0.0000 | 0.0000 | 46.00 | 2 | -6.9657 | 0.0000 | 0.0000 | 0.6667 | 0.0370 |
| BC1G_08755.1 | hypothetical protein (translation) | 49.00 | 3 | 5.5842 | 0.0000 | 0.0000 | 5.33 | 3 | 5.8918 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| AAL25813.1 | polyubiquitin [Prunus avium] | 14.33 | 1 | -5.1717 | 0.0000 | 0.0000 | 51.67 | 1 | -5.0322 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| BC1G_02021.1 | hypothetical protein (translation) | 0.00 | 0 | -5.1164 | 0.0000 | 0.0000 | 23.00 | 3 | -4.8594 | 0.0000 | 0.0000 | 1.0000 | 0.0000 |
| BC1G_10455.1 | hypothetical protein (translation) | 0.00 | 0 | -4.9645 | 0.0000 | 0.0000 | 21.67 | 2 | -4.7131 | 0.0000 | 0.0000 | 0.6667 | 0.0370 |
| BC1G_02326.1 | predicted protein (translation) | 0.00 | 0 | -4.8863 | 0.0000 | 0.0000 | 21.00 | 1 | -4.6379 | 0.0000 | 0.0000 | 0.3333 | 0.2963 |
| BC1G_00448.1 | hypothetical protein (translation) | 11.33 | 3 | -4.8058 | 0.0000 | 0.0000 | 43.00 | 3 | -4.6665 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| BC1G_14975.1 | predicted protein (translation) | 23.00 | 1 | 4.5510 | 0.0000 | 0.0000 | 0.00 | 0 | 4.8522 | 0.0000 | 0.0000 | 0.3333 | 0.2963 |
| BC1G_12374.1 | hypothetical protein (translation) | 54.00 | 3 | 4.1372 | 0.0000 | 0.0000 | 16.00 | 3 | 4.2760 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| BC1G_15663.1 | hypothetical protein (translation) | 0.00 | 0 | -3.7354 | 0.0002 | 0.0003 | 12.33 | 3 | -3.5361 | 0.0004 | 0.0006 | 1.0000 | 0.0000 |
| BC1G_06164.1 | hypothetical protein similar to heat shock protein 70 (translation) | 0.00 | 0 | -3.6334 | 0.0003 | 0.0004 | 11.67 | 3 | -3.4389 | 0.0006 | 0.0008 | 1.0000 | 0.0000 |
| BC1G_13335.1 | predicted protein (translation) | 0.67 | 1 | -3.4627 | 0.0005 | 0.0008 | 12.33 | 1 | -3.2942 | 0.0010 | 0.0012 | 0.0000 | 0.0000 |
| BC1G_14030.1 | beta (1-3) glucanosyltransferase (translation) | 22.00 | 3 | 3.4444 | 0.0006 | 0.0008 | 3.33 | 2 | 3.6007 | 0.0003 | 0.0005 | 0.0000 | 0.0000 |
| BC1G_09079.1 | hypothetical protein (translation) | 22.00 | 3 | 3.1711 | 0.0015 | 0.0020 | 4.33 | 2 | 3.2995 | 0.0010 | 0.0012 | 0.0000 | 0.0000 |
| BC1G_05503.1 | hypothetical protein (translation) | 0.00 | 0 | -3.1293 | 0.0018 | 0.0021 | 8.67 | 2 | -2.9590 | 0.0031 | 0.0027 | 0.6667 | 0.0370 |
| BC1G_08615.1 | hypothetical protein (translation) | 10.67 | 3 | 3.0895 | 0.0020 | 0.0023 | 0.00 | 0 | 3.2841 | 0.0010 | 0.0012 | 1.0000 | 0.0000 |
| BC1G_02060.1 | predicted protein (translation) | 12.33 | 3 | 3.0387 | 0.0024 | 0.0025 | 0.67 | 1 | 3.2079 | 0.0013 | 0.0014 | 0.0000 | 0.0000 |
| BC1G_00558.1 | Superoxide dismutase (translation) | 11.33 | 1 | 3.0363 | 0.0024 | 0.0025 | 0.33 | 1 | 3.2152 | 0.0013 | 0.0014 | 0.0000 | 0.0000 |
| BC1G_08642.1 | predicted protein (translation) | 105.33 | 3 | 3.0035 | 0.0027 | 0.0027 | 59.00 | 3 | 3.0473 | 0.0023 | 0.0021 | 0.0000 | 0.0000 |
| BC1G_02986.1 | Phosphatidylglycerol / phosphatidylinositol transfer protein (translation) | 14.33 | 3 | 2.9486 | 0.0032 | 0.0031 | 1.67 | 2 | 3.0909 | 0.0020 | 0.0019 | 0.0000 | 0.0000 |
| BC1G_05297.1 | hypothetical protein (translation) | 27.00 | 3 | -2.9256 | 0.0034 | 0.0031 | 47.33 | 3 | -2.8789 | 0.0040 | 0.0034 | 0.0000 | 0.0000 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1G_00350.1 | Enolase (2-phosphoglycerate dehydratase) (translation) | 0.00 | 0 | -2.8763 | 0.0040 | 0.0034 | 7.33 | 3 | -2.7187 | 0.0066 | 0.0050 | 1.0000 | 0.0000 |
| BC1G_12307.1 | hypothetical protein similar to cobalamin-independent methionine synthase (translation) | 0.00 | 0 | -2.8763 | 0.0040 | 0.0034 | 7.33 | 3 | -2.7187 | 0.0066 | 0.0050 | 1.0000 | 0.0000 |
| BC1G_05476.1 | hypothetical protein (translation) | 0.00 | 0 | -2.7433 | 0.0061 | 0.0050 | 6.67 | 3 | -2.5924 | 0.0095 | 0.0069 | 1.0000 | 0.0000 |
| BC1G_00617.1 | hypothetical protein similar to pectin methylesterase (translation) | 8.67 | 2 | 2.6153 | 0.0089 | 0.0070 | 0.33 | 1 | 2.7646 | 0.0057 | 0.0047 | 0.0000 | 0.0000 |
| BC1G_07315.1 | hypothetical protein similar to A Chain A, The Structure Of The Complex Between Aha1 And Hsp90 (translation) | 0.00 | 0 | -2.5286 | 0.0115 | 0.0087 | 5.67 | 3 | -2.3888 | 0.0169 | 0.0115 | 1.0000 | 0.0000 |
| BC1G_11143.1 | hypothetical protein similar to endopolygalacturonase PGb (translation) | 26.33 | 3 | 2.3839 | 0.0171 | 0.0126 | 10.00 | 2 | 2.4433 | 0.0146 | 0.0102 | 0.0000 | 0.0000 |
| BC1G_01740.1 | Peptidyl-prolyl *cis*-trans isomerase (cyclophilin) (translation) | 0.00 | 0 | -2.2942 | 0.0218 | 0.0154 | 4.67 | 2 | -2.1667 | 0.0303 | 0.0193 | 0.6667 | 0.0370 |
| BC1G_08946.1 | hypothetical protein (translation) | 0.00 | 0 | -2.2089 | 0.0272 | 0.0187 | 4.33 | 3 | -2.0860 | 0.0370 | 0.0229 | 1.0000 | 0.0000 |
| BC1G_06304.1 | hypothetical protein (translation) | 0.00 | 0 | -2.1229 | 0.0338 | 0.0218 | 4.00 | 3 | -2.0045 | 0.0450 | 0.0262 | 1.0000 | 0.0000 |
| BC1G_07647.1 | hypothetical protein (translation) | 0.00 | 0 | -2.1229 | 0.0338 | 0.0218 | 4.00 | 3 | -2.0045 | 0.0450 | 0.0262 | 1.0000 | 0.0000 |
| BC1G_00459.1 | predicted protein (translation) | 4.67 | 1 | 2.0406 | 0.0413 | 0.0250 | 0.00 | 0 | 2.1661 | 0.0303 | 0.0193 | 0.3333 | 0.2963 |
| BC1G_01095.1 | hypothetical protein (translation) | 0.00 | 0 | -2.0332 | 0.0420 | 0.0250 | 3.67 | 2 | -1.9197 | 0.0549 | 0.0295 | 0.6667 | 0.0370 |
| BC1G_04945.1 | hypothetical protein (translation) | 0.00 | 0 | -2.0332 | 0.0420 | 0.0250 | 3.67 | 3 | -1.9197 | 0.0549 | 0.0295 | 1.0000 | 0.0000 |
| BC1G_00912.1 | hypothetical protein (translation) | 0.00 | 0 | -1.9366 | 0.0528 | 0.0290 | 3.33 | 3 | -1.8283 | 0.0675 | 0.0321 | 1.0000 | 0.0000 |
| BC1G_01617.1 | hypothetical protein similar to exo-polygalacturonase (translation) | 0.00 | 0 | -1.9366 | 0.0528 | 0.0290 | 3.33 | 2 | -1.8283 | 0.0675 | 0.0321 | 0.6667 | 0.0370 |
| BC1G_14012.1 | hypothetical protein (translation) | 0.00 | 0 | -1.9366 | 0.0528 | 0.0290 | 3.33 | 2 | -1.8283 | 0.0675 | 0.0321 | 0.6667 | 0.0370 |
| BC1G_04151.1 | hypothetical protein (translation) | 17.33 | 3 | 1.9049 | 0.0568 | 0.0303 | 6.67 | 1 | 1.9509 | 0.0511 | 0.0290 | 0.0000 | 0.0000 |
| BC1G_15542.1 | hypothetical protein (translation) | 3.00 | 2 | -1.8686 | 0.0617 | 0.0303 | 8.67 | 3 | -1.8135 | 0.0697 | 0.0324 | 0.0000 | 0.0000 |
| BC1G_08635.1 | predicted protein (translation) | 14.67 | 3 | 1.8399 | 0.0658 | 0.0303 | 5.33 | 3 | 1.8869 | 0.0592 | 0.0310 | 0.0000 | 0.0000 |
| BC1G_02492.1 | hypothetical protein (translation) | 0.00 | 0 | -1.8380 | 0.0661 | 0.0303 | 3.00 | 2 | -1.7350 | 0.0827 | 0.0331 | 0.6667 | 0.0370 |
| BC1G_04390.1 | dnaK-type molecular chaperone BiP (translation) | 0.00 | 0 | -1.8380 | 0.0661 | 0.0303 | 3.00 | 2 | -1.7350 | 0.0827 | 0.0331 | 0.6667 | 0.0370 |
| BC1G_06849.1 | hypothetical protein similar to aspartic proteinase precursor (translation) | 0.00 | 0 | -1.8380 | 0.0661 | 0.0303 | 3.00 | 2 | -1.7350 | 0.0827 | 0.0331 | 0.6667 | 0.0370 |
| BC1G_10724.1 | conserved hypothetical protein (translation) | 0.00 | 0 | -1.8380 | 0.0661 | 0.0303 | 3.00 | 2 | -1.7350 | 0.0827 | 0.0331 | 0.6667 | 0.0370 |
| BC1G_12319.1 | NAD-dependent formate dehydrogenase (translation) | 0.00 | 0 | -1.8380 | 0.0661 | 0.0303 | 3.00 | 2 | -1.7350 | 0.0827 | 0.0331 | 0.6667 | 0.0370 |
| BC1G_01204.1 | hypothetical protein similar to glyoxal oxidase (translation) | 0.00 | 0 | -1.7338 | 0.0830 | 0.0351 | 2.67 | 3 | -1.6365 | 0.1017 | 0.0364 | 1.0000 | 0.0000 |
| BC1G_05133.1 | hypothetical protein (translation) | 0.00 | 0 | -1.7338 | 0.0830 | 0.0351 | 2.67 | 2 | -1.6365 | 0.1017 | 0.0364 | 0.6667 | 0.0370 |
| BC1G_05327.1 | pyruvate carboxylase (translation) | 0.00 | 0 | -1.7338 | 0.0830 | 0.0351 | 2.67 | 2 | -1.6365 | 0.1017 | 0.0364 | 0.6667 | 0.0370 |
| BC1G_08294.1 | 14-3-3-like protein (translation) | 0.00 | 0 | -1.7338 | 0.0830 | 0.0351 | 2.67 | 2 | -1.6365 | 0.1017 | 0.0364 | 0.6667 | 0.0370 |
| BC1G_11950.1 | hypothetical protein (translation) | 3.33 | 2 | 1.7226 | 0.0850 | 0.0352 | 0.00 | 0 | 1.8279 | 0.0676 | 0.0321 | 0.6667 | 0.0370 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1G_08016.1 | 60S acidic ribosomal protein P2 (translation) | 2.67 | 1 | -1.6659 | 0.0957 | 0.0390 | 7.33 | 2 | -1.6182 | 0.1056 | 0.0372 | 0.0000 | 0.0000 |
| BC1G_00978.1 | hypothetical protein (translation) | 3.00 | 1 | 1.6348 | 0.1021 | 0.0401 | 0.00 | 0 | 1.7347 | 0.0828 | 0.0331 | 0.3333 | 0.2963 |
| NP_199802.1 | heat shock protein 70 [Arabidopsis thaliana] | 3.00 | 1 | 1.6348 | 0.1021 | 0.0401 | 0.00 | 0 | 1.7347 | 0.0828 | 0.0331 | 0.3333 | 0.2963 |
| BC1G_02744.1 | 6-phosphogluconate dehydrogenase (translation) | 0.00 | 0 | -1.6195 | 0.1053 | 0.0406 | 2.33 | 3 | -1.5284 | 0.1264 | 0.0437 | 1.0000 | 0.0000 |
| BC1G_00576.1 | hypothetical protein (translation) | 2.67 | 2 | 1.5422 | 0.1230 | 0.0458 | 0.00 | 0 | 1.6362 | 0.1018 | 0.0364 | 0.6667 | 0.0370 |
| BC1G_06035.1 | hypothetical protein (translation) | 2.67 | 2 | 1.5422 | 0.1230 | 0.0458 | 0.00 | 0 | 1.6362 | 0.1018 | 0.0364 | 0.6667 | 0.0370 |
| BC1G_03567.1 | hypothetical protein (translation) | 0.00 | 0 | -1.5003 | 0.1335 | 0.0473 | 2.00 | 3 | -1.4158 | 0.1568 | 0.0500 | 1.0000 | 0.0000 |
| BC1G_06885.1 | hypothetical protein similar to polyketide synthase (translation) | 0.00 | 0 | -1.5003 | 0.1335 | 0.0473 | 2.00 | 1 | -1.4158 | 0.1568 | 0.0500 | 0.3333 | 0.2963 |
| BC1G_12947.1 | hypothetical protein (translation) | 0.00 | 0 | -1.5003 | 0.1335 | 0.0473 | 2.00 | 2 | -1.4158 | 0.1568 | 0.0500 | 0.6667 | 0.0370 |
| BC1G_11018.1 | hypothetical protein (translation) | 5.33 | 2 | 1.4039 | 0.1603 | 0.0500 | 1.33 | 2 | 1.4515 | 0.1466 | 0.0499 | 0.0000 | 0.0000 |
| BC1G_11898.1 | hypothetical protein (translation) | 2.00 | 2 | -1.3934 | 0.1635 | 0.0500 | 5.33 | 3 | -1.3541 | 0.1757 | 0.0513 | 0.0000 | 0.0000 |
| BC1G_00044.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 1 | -1.2935 | 0.1958 | 0.0513 | 0.3333 | 0.2963 |
| BC1G_04092.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 1 | -1.2935 | 0.1958 | 0.0513 | 0.3333 | 0.2963 |
| BC1G_05132.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 1 | -1.2935 | 0.1958 | 0.0513 | 0.3333 | 0.2963 |
| BC1G_06038.1 | predicted protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 3 | -1.2935 | 0.1958 | 0.0513 | 1.0000 | 0.0000 |
| BC1G_08882.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 2 | -1.2935 | 0.1958 | 0.0513 | 0.6667 | 0.0370 |
| BC1G_08895.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 1 | -1.2935 | 0.1958 | 0.0513 | 0.3333 | 0.2963 |
| BC1G_09731.1 | hypothetical protein similar to EF2_NEUCR Elongation factor 2 (EF-2) (Colonial temperature-sensitive 3) (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 3 | -1.2935 | 0.1958 | 0.0513 | 1.0000 | 0.0000 |
| BC1G_09782.1 | translation initiation factor eIF-5A (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 1 | -1.2935 | 0.1958 | 0.0513 | 0.3333 | 0.2963 |
| BC1G_10503.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 2 | -1.2935 | 0.1958 | 0.0513 | 0.6667 | 0.0370 |
| BC1G_14944.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 3 | -1.2935 | 0.1958 | 0.0513 | 1.0000 | 0.0000 |
| BC1G_15343.1 | hypothetical protein (translation) | 0.00 | 0 | -1.3708 | 0.1704 | 0.0500 | 1.67 | 2 | -1.2935 | 0.1958 | 0.0513 | 0.6667 | 0.0370 |
| NP_188108.1 | hypothetical protein [Arabidopsis thaliana] | 2.00 | 1 | 1.3345 | 0.1820 | 0.0527 | 0.00 | 0 | 1.4156 | 0.1569 | 0.0500 | 0.3333 | 0.2963 |
| BC1G_16047.1 | hypothetical protein (translation) | 6.67 | 3 | 1.2764 | 0.2018 | 0.0576 | 2.33 | 2 | 1.3099 | 0.1902 | 0.0513 | 0.0000 | 0.0000 |
| BC1G_00347.1 | hypothetical protein similar to glutamine:fructose-6-phosphate amidotransferase (translation) | 0.00 | 0 | -1.2232 | 0.2212 | 0.0583 | 1.33 | 1 | -1.1541 | 0.2484 | 0.0597 | 0.3333 | 0.2963 |
| BC1G_00555.1 | Transaldolase (translation) | 0.00 | 0 | -1.2232 | 0.2212 | 0.0583 | 1.33 | 2 | -1.1541 | 0.2484 | 0.0597 | 0.6667 | 0.0370 |
| BC1G_02623.1 | hypothetical protein (translation) | 0.00 | 0 | -1.2232 | 0.2212 | 0.0583 | 1.33 | 1 | -1.1541 | 0.2484 | 0.0597 | 0.3333 | 0.2963 |
| BC1G_03430.1 | peptidyl-prolyl cis-trans isomerase (FK506 binding protein) (translation) | 0.00 | 0 | -1.2232 | 0.2212 | 0.0583 | 1.33 | 1 | -1.1541 | 0.2484 | 0.0597 | 0.3333 | 0.2963 |
| BC1G_12627.1 | hypothetical protein (translation) | 0.00 | 0 | -1.2232 | 0.2212 | 0.0583 | 1.33 | 1 | -1.1541 | 0.2484 | 0.0597 | 0.3333 | 0.2963 |
| BC1G_14129.1 | hypothetical protein (translation) | 0.00 | 0 | -1.2232 | 0.2212 | 0.0583 | 1.33 | 1 | -1.1541 | 0.2484 | 0.0597 | 0.3333 | 0.2963 |
| BC1G_00198.1 | hypothetical protein (translation) | 1.67 | 2 | 1.2193 | 0.2227 | 0.0583 | 0.00 | 0 | 1.2934 | 0.1959 | 0.0513 | 0.6667 | 0.0370 |
| BC1G_09564.1 | hypothetical protein (translation) | 1.00 | 1 | -1.1201 | 0.2627 | 0.0641 | 3.00 | 3 | -1.0855 | 0.2777 | 0.0641 | 0.0000 | 0.0000 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1G_08719.1 | hypothetical protein (translation) | 1.33 | 2 | 1.0881 | 0.2766 | 0.0641 | 0.00 | 0 | 1.1540 | 0.2485 | 0.0597 | 0.6667 | 0.0370 |
| BC1G_00939.1 | hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 2 | -1.0006 | 0.3170 | 0.0641 | 0.6667 | 0.0370 |
| BC1G_01354.1 | Saccharopine dehydrogenase, catalyzes the eighth and final step in lysine biosynthesis pathway (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 2 | -1.0006 | 0.3170 | 0.0641 | 0.6667 | 0.0370 |
| BC1G_02018.1 | hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_03241.1 | conserved hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 2 | -1.0006 | 0.3170 | 0.0641 | 0.6667 | 0.0370 |
| BC1G_08314.1 | hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_09443.1 | hypothetical protein similar to MASY_EMENI Malate synthase, glyoxysomal (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 2 | -1.0006 | 0.3170 | 0.0641 | 0.6667 | 0.0370 |
| BC1G_10247.1 | hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 2 | -1.0006 | 0.3170 | 0.0641 | 0.6667 | 0.0370 |
| BC1G_10466.1 | hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_11392.1 | hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_12950.1 | Inhibitor of mitochondrial ATPase that forms a complex with ATP synthase to inhibit enzyme activity (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_13428.1 | conserved hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_15701.1 | hypothetical protein (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_16370.1 | 60S ribosomal protein L9 (translation) | 0.00 | 0 | -1.0606 | 0.2889 | 0.0641 | 1.00 | 1 | -1.0006 | 0.3170 | 0.0641 | 0.3333 | 0.2963 |
| BC1G_00455.1 | hypothetical protein (translation) | 1.00 | 3 | 0.9434 | 0.3455 | 0.0743 | 0.00 | 0 | 1.0005 | 0.3171 | 0.0641 | 1.0000 | 0.0000 |
| BC1G_10221.1 | hypothetical protein (translation) | 1.00 | 2 | 0.9434 | 0.3455 | 0.0743 | 0.00 | 0 | 1.0005 | 0.3171 | 0.0641 | 0.6667 | 0.0370 |
| BC1G_04759.1 | conserved hypothetical protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 2 | -0.8188 | 0.4129 | 0.0720 | 0.6667 | 0.0370 |
| BC1G_05131.1 | hypothetical protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_05433.1 | hypothetical protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 2 | -0.8188 | 0.4129 | 0.0720 | 0.6667 | 0.0370 |
| BC1G_05991.1 | hypothetical protein similar to ATP citrate lyase, subunit 2 (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_08348.1 | hypothetical protein similar to G6PD_ASPNG Glucose-6-phosphate 1-dehydrogenase (G6PD) (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_08794.1 | Pyruvate dehydrogenase E1 component beta subunit (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_10581.1 | conserved hypothetical protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 2 | -0.8188 | 0.4129 | 0.0720 | 0.6667 | 0.0370 |
| BC1G_10587.1 | hypothetical protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_11454.1 | activator of heat shock protein 90 (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 2 | -0.8188 | 0.4129 | 0.0720 | 0.6667 | 0.0370 |
| BC1G_11685.1 | glucose-repressible gene protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_11968.1 | glyceraldehyde 3-phosphate dehydrogenase (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_12136.1 | hypothetical protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 2 | -0.8188 | 0.4129 | 0.0720 | 0.6667 | 0.0370 |
| BC1G_14880.1 | hypothetical protein (translation) | 0.00 | 0 | -0.8681 | 0.3854 | 0.0743 | 0.67 | 1 | -0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1G_05033.1 | hypothetical protein (translation) | 9.67 | 3 | 0.7968 | 0.4256 | 0.0806 | 5.67 | 2 | 0.8065 | 0.4199 | 0.0726 | 0.0000 | 0.0000 |
| BC1G_05299.1 | hypothetical protein (translation) | 5.33 | 3 | -0.7741 | 0.4388 | 0.0806 | 7.33 | 2 | -0.7647 | 0.4445 | 0.0754 | 0.0000 | 0.0000 |
| BC1G_04585.1 | ubiquitin-like protein (translation) | 0.67 | 1 | 0.7721 | 0.4400 | 0.0806 | 0.00 | 0 | 0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_09892.1 | hypothetical protein (translation) | 0.67 | 1 | 0.7721 | 0.4400 | 0.0806 | 0.00 | 0 | 0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_14136.1 | predicted protein (translation) | 0.67 | 1 | 0.7721 | 0.4400 | 0.0806 | 0.00 | 0 | 0.8188 | 0.4129 | 0.0720 | 0.3333 | 0.2963 |
| BC1G_05298.1 | hypothetical protein (translation) | 6.67 | 3 | 0.7482 | 0.4543 | 0.0806 | 3.67 | 2 | 0.7587 | 0.4481 | 0.0754 | 0.0000 | 0.0000 |
| BC1G_01463.1 | hypothetical protein (translation) | 1.00 | 2 | -0.6804 | 0.4963 | 0.0806 | 2.00 | 2 | -0.6653 | 0.5058 | 0.0754 | 0.0000 | 0.0000 |
| BC1G_00896.1 | predicted protein (translation) | 39.33 | 3 | -0.6491 | 0.5163 | 0.0806 | 40.33 | 3 | -0.6466 | 0.5179 | 0.0754 | 0.0000 | 0.0000 |
| BC1G_07215.1 | hypothetical protein (translation) | 1.33 | 1 | -0.6363 | 0.5246 | 0.0806 | 2.33 | 2 | -0.6244 | 0.5324 | 0.0754 | 0.0000 | 0.0000 |
| BC1G_00290.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_00567.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_00769.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_02930.1 | predicted protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_03337.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_03991.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_04443.1 | hypothetical protein similar to KETOL-ACID REDUCTOISOMERASE PRECURSOR (ACETOHYDROXY-ACID REDUCTOISOMERASE) (ALPHA-KETO-BETA-HYDROXYLACIL REDUCTOISOMERASE) (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_04836.1 | fructose-bisphosphate aldolase (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_04994.1 | hypothetical protein similar to alpha-L-arabinofuranosidase (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_05278.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_05980.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_05989.1 | hypothetical protein similar to ACL1_SORMA ATP-citrate synthase subunit 1 (ATP-citrate (pro-S-)-lyase 1) (Citrate cleavage enzyme subunit 1) (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_07448.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_07482.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_07510.1 | hypothetical protein similar to valosin-containing protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_07637.1 | hypothetical protein similar to lipase (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_07653.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_08198.1 | hypothetical protein similar to smooth muscle alpha actin (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_08318.1 | conserved hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BC1G_08393.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_11083.1 | Alcohol dehydrogenase (ADH) (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_12563.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_13215.1 | hypothetical protein (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_13641.1 | dipeptidyl aminopeptidase (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_14217.1 | acetyl-CoA hydrolase (translation) | 0.00 | 0 | -0.6092 | 0.5424 | 0.0806 | 0.33 | 1 | -0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_11823.1 | inorganic pyrophosphatase (translation) | 0.67 | 1 | -0.5506 | 0.5819 | 0.0839 | 1.33 | 2 | -0.5386 | 0.5902 | 0.0781 | 0.0000 | 0.0000 |
| BC1G_02551.1 | hypothetical protein (translation) | 0.33 | 1 | 0.5418 | 0.5879 | 0.0839 | 0.00 | 0 | 0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_02656.1 | hypothetical protein (translation) | 0.33 | 1 | 0.5418 | 0.5879 | 0.0839 | 0.00 | 0 | 0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_06726.1 | hypothetical protein (translation) | 0.33 | 1 | 0.5418 | 0.5879 | 0.0839 | 0.00 | 0 | 0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_07073.1 | hypothetical protein (translation) | 0.33 | 1 | 0.5418 | 0.5879 | 0.0839 | 0.00 | 0 | 0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_11865.1 | hypothetical protein (translation) | 0.33 | 1 | 0.5418 | 0.5879 | 0.0839 | 0.00 | 0 | 0.5746 | 0.5656 | 0.0754 | 0.3333 | 0.2963 |
| BC1G_02936.1 | hypothetical protein (translation) | 1.00 | 2 | 0.5146 | 0.6069 | 0.0855 | 0.33 | 1 | 0.5286 | 0.5971 | 0.0781 | 0.0000 | 0.0000 |
| BC1G_08030.1 | predicted protein (translation) | 1.00 | 1 | 0.5146 | 0.6069 | 0.0855 | 0.33 | 1 | 0.5286 | 0.5971 | 0.0781 | 0.0000 | 0.0000 |
| BC1G_12859.1 | hypothetical protein (translation) | 3.33 | 3 | 0.4424 | 0.6582 | 0.0922 | 2.00 | 2 | 0.4475 | 0.6545 | 0.0851 | 0.0000 | 0.0000 |
| BC1G_10191.1 | 40S ribosomal protein S3 (translation) | 3.67 | 2 | 0.4051 | 0.6854 | 0.0954 | 2.33 | 2 | 0.4091 | 0.6825 | 0.0882 | 0.0000 | 0.0000 |
| BC1G_10301.1 | predicted protein (translation) | 0.33 | 1 | -0.3993 | 0.6897 | 0.0954 | 0.67 | 1 | -0.3903 | 0.6963 | 0.0888 | 0.0000 | 0.0000 |
| BC1G_13938.1 | hypothetical protein (translation) | 1.33 | 3 | 0.3848 | 0.7004 | 0.0962 | 0.67 | 1 | 0.3911 | 0.6957 | 0.0888 | 0.0000 | 0.0000 |
| BC1G_14403.1 | thioredoxin (translation) | 10.67 | 2 | 0.3677 | 0.7131 | 0.0974 | 8.00 | 3 | 0.3695 | 0.7118 | 0.0902 | 0.0000 | 0.0000 |
| BC1G_06122.1 | Polyubiquitin (translation) | 70.67 | 2 | -0.3581 | 0.7203 | 0.0977 | 66.67 | 1 | -0.3575 | 0.7207 | 0.0908 | 0.0000 | 0.0000 |
| BC1G_02364.1 | hypothetical protein (translation) | 3.00 | 3 | 0.3174 | 0.7509 | 0.1013 | 2.00 | 3 | 0.3201 | 0.7489 | 0.0938 | 0.0000 | 0.0000 |
| BC1G_06836.1 | hypothetical protein (translation) | 1.00 | 2 | -0.3062 | 0.7595 | 0.1018 | 1.33 | 1 | -0.3027 | 0.7621 | 0.0948 | 0.0000 | 0.0000 |
| BC1G_02223.1 | hypothetical protein (translation) | 1.67 | 1 | -0.2851 | 0.7756 | 0.1030 | 2.00 | 2 | -0.2826 | 0.7775 | 0.0956 | 0.0000 | 0.0000 |
| BC1G_07780.1 | hypothetical protein similar to mitochondrial ATP synthase H+ transporting F1 complex alpha subunit isoform 1 (translation) | 3.33 | 2 | 0.2821 | 0.7779 | 0.1030 | 2.33 | 3 | 0.2841 | 0.7764 | 0.0956 | 0.0000 | 0.0000 |
| BC1G_06840.1 | hypothetical protein similar to pectin methyl esterase (translation) | 26.67 | 3 | 0.2653 | 0.7908 | 0.1041 | 22.00 | 2 | 0.2658 | 0.7904 | 0.0966 | 0.0000 | 0.0000 |
| BC1G_07825.1 | predicted protein (translation) | 6.00 | 2 | -0.2036 | 0.8387 | 0.1097 | 6.00 | 2 | -0.2029 | 0.8392 | 0.1020 | 0.0000 | 0.0000 |
| BC1G_12729.1 | Woronin body major protein (translation) | 1.67 | 2 | 0.0952 | 0.9241 | 0.1202 | 1.33 | 1 | 0.0955 | 0.9239 | 0.1116 | 0.0000 | 0.0000 |
| BC1G_13862.1 | hypothetical protein (translation) | 1.00 | 2 | -0.0829 | 0.9339 | 0.1208 | 1.00 | 1 | -0.0826 | 0.9342 | 0.1122 | 0.0000 | 0.0000 |
| BC1G_08184.1 | hypothetical protein (translation) | 0.67 | 1 | -0.0678 | 0.9459 | 0.1216 | 0.67 | 1 | -0.0676 | 0.9461 | 0.1129 | 0.0000 | 0.0000 |
| BC1G_14679.1 | hypothetical protein (translation) | 2.67 | 2 | 0.0213 | 0.9830 | 0.1256 | 2.33 | 2 | 0.0213 | 0.9830 | 0.1166 | 0.0000 | 0.0000 |